Datasets for Final Projects

Business Analytics for Decision Making

Group_PS_8058

Disclosure Statement

The case studies included in this compilation have been meticulously designed and adapted for the exclusive purpose of supporting practical learning among MBA students enrolled in the course *Business Analytics for Decision Making*, taught jointly by **Dr. Yasser Yakout and Hossam Daoud** at the Graduate School of Business, Arab Academy for Science, Technology and Maritime Transport (AASTMT).

These case studies are intended to simulate real-world business challenges across diverse domains, such as construction, healthcare, manufacturing, finance, and technology, in order to provide students with experiential learning opportunities in data-driven strategy, analytical decision-making, and applied problem-solving. Each case study aligns with the academic and pedagogical objectives of the MBA programme and facilitates the application of key methodologies including exploratory data analysis, dashboard development, AI-assisted modelling, and managerial insight generation relevant to contemporary business contexts.

Usage Restrictions

These case studies are intended strictly for academic use by registered MBA students within the context of coursework, assessments, and instructional activities under the *Business Analytics for Decision Making* curriculum. Any form of external use including but not limited to publication, dissemination, reproduction, or commercial exploitation, is expressly prohibited without the prior written consent of Dr. Yasser Yakout or the authorised administration of the Graduate School of Business, AASTMT.

All organisational names, characters, figures, and scenarios contained within these case studies are fictitious and created solely for educational purposes. They do not represent real individuals, organisations, or confidential business information.

For permissions or inquiries regarding appropriate use, please contact:

Dr. Yasser Aly Yakout Shehata

Dean, Quality Assurance and Accreditation Center

Email: y.yakout@aast.edu

© 2025 Graduate School of Business, AASTMT. All rights reserved.

AI-Assisted Editing and Simulation Statement

The development and refinement of these simulated datasets have been **assisted by artificial intelligence**, specifically through **OpenAI's ChatGPT platform**. AI was utilized to:

- Structure realistic variables aligned with domain-specific roles and industry standards
- Generate plausible but fictitious data distributions to reflect operational complexity
- Enhance dataset descriptions, metadata, and derived metrics for educational relevance
- Ensure consistency, clarity, and analytical depth across multiple datasets

All AI-generated content has been **reviewed and curated by faculty members** to ensure alignment with the course objectives and academic integrity of the **MBA program at the Graduate School of Business, AASTMT**.

Disclaimer: While AI tools supported the simulation and editing process, the datasets remain **entirely fictional** and are intended exclusively for **academic instruction and non-commercial use**.

For questions regarding AI-assisted dataset design, please contact:

Dr. Yasser Aly Yakout Shehata

Email: y.yakout@aast.edu

© 2025 Dr. Yasser Aly Yakout Shehata

Contents

Abstract 8
Dataset 1: Steel QA Inspection Analytics 10
Dataset 2: ElDabaa Nuclear QA Business Analytics 14
Dataset 3: Dabaa QA Insight Analytics Project 19
Dataset 4: HSE Analytics For Decision Making 23
Dataset 5: MEP Analytics For Decision Making 28
Dataset 6: Construction Project Management Analytics 33
Dataset 7: Corporate Leadership Performance Analytics 38
Dataset 8: Projects Human Resources Manager Analytics 43
Dataset 9: Construction Project Performance Analytics (CPPA-1000) 47
Dataset 10: Ezz Steel Sales Analytics Dataset 52
Dataset 11: Executive Decision Analytics for Construction Projects 56
Dataset 12: Decision Analytics for Authority Urban Planning and Infrastructure 60
Dataset 12: Construction Quality Control Analytics 65
Dataset 13: Integrated Electrical Inspection and Compliance Analytics (IEICAD) 70
Dataset 14: Smart Contracts and Commercial Insights Dataset 75
Dataset 15: Smart Project Operations Dataset for MEP Decision-Making 80
Dataset 16: Operational Site Management Analytics Dataset (OSMAD) 85
Dataset 17: Heavy Equipment Sales and Performance Intelligence 90
Dataset 18: Power Quality Insights for Energy Sector Performance 95
Dataset 19: Integrated Site Operations and Performance Analytics 100
Dataset 20: Pharmacy Administration in South Sinai 104
Dataset 21: CIB Corporate Credit Risk 109
Dataset 22: Ismailia Healthcare IT Transformation 114
Dataset 23: Healthcare Inventory El-Fayrouz Medical Complex 119
Dataset 24: Pharmacist Supervisor Analytics South Sinai 123
Dataset 25: Dental HIS Transformation South Sinai Dataset 127
Dataset 26: Substation Project Analytics (HV & EHV) 132

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 27: Health Informatics Digital Transformation Analytics 137
Dataset 28: Platinum Relationship Manager NBE 142
Dataset 29: Management Consulting Industry 146
Dataset 30: Healthcare IT Strategy and Analytics 150
Dataset 31: CIB Corporate Banking Analytics..... 154
Dataset 32: Data-Driven Procurement Optimisation..... 159
Dataset 33: Garment Quality Assurance Operational Dataset 164
Dataset 34: Country Manager Garment Operations Analytics..... 169
Dataset 35: General Dentist Operational and Clinical Analytics Dataset 174
Dataset 36: Global Product Manager Business Analytics Dataset..... 178
Dataset 37: ICU Operations and Patient Outcomes Analytics..... 183
Dataset 38: Global Corporate Relations (GCR) Strategic Analytics 187
Dataset 39: HRBP Strategic Workforce Analytics 192
Dataset 40: RD Performance 197
Dataset 41: Community Pharmacy Performance..... 201
Dataset 42: Radar Commander 205
Dataset 43: Patient Experience and Operational Efficiency 209
Dataset 44: GB Sales Performance 212
Dataset 45: CDP Performance 216
Dataset 46: DCS Operations..... 220
Dataset 47: Natural Gas Distribution..... 224
Dataset 48: Senior Medical Representative..... 229
Dataset 49: SAP Senior Manager KPI 234
Dataset 50: Senior Medical Representative Analytics 238
Dataset 51: Polyethylene Operations 243
Dataset 52: Accounts Receivable Performance Evyap 247
Dataset 53: Al Nahdi Operational Performance 252
Dataset 54: Biomedical Equipment Maintenance Analytics 257
Dataset 55: IT Project Manager QCC Analytics 262

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 56: Web Team Leader Analytics 267
Dataset 57: Technology Senior Specialist Vodafone 272
Dataset 58: Regional Talent Acquisition Edulisa 360 279
Dataset 59: Modern Endpoint Management 286
Dataset 60: Power BI Developer and Trainer 292
Dataset 61: MTS Customer Service Analytics 296
Dataset 62: Hannoun Pharmacies Analytics 300
Dataset 63: Technical Services Smart Systems Egypt 304
Dataset 64: Marketing Automation Insights Analytics 308
Dataset 65: CEG 313
Dataset 66: CARRIER 317
Dataset 67: ABP 322
Dataset 68: GHUDAI 327
Dataset 69: DARALRIYADH 332
Dataset 70: RAMACO 335
Dataset 71: ETSEC 339
Dataset 72: Dubai Municipality 344
Dataset 73: ASMK 349
Dataset 74: AVENTA 354
Dataset 75: Media Heist 359
Dataset 76: Rostom Medical 363
Dataset 77: SPIMACO 367
Dataset 78: PHARCO 372
Dataset 79: Planet Pharmacies 376
Dataset 80: Asharq News 379
Dataset 81: KAHRAMAA 384
Dataset 82: Sika 389
Dataset 83: NCSP 394
Dataset 84: Al Qaryan 399

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 85: Taawin 403
Dataset 86: Almajdouie 408
Dataset 87: Musanadah 414
Dataset 88: El Hassan & El Hussein 421
Dataset 89: Port Said Container and Cargo Handling Company 426
Dataset 90: Badawi Payroll Employee 433
Dataset 91: EHA HR Training Analytics 439
Dataset 92: Student Affairs Performance and Engagement 449
Dataset 93: E-Commerce Brand Manager Instagram 455
Dataset 94: Port Said Dental Analytics 460
Dataset 95: Petrojet Coating Finance Analytics Gold 465
Appendix: Guidance for Dataset Selection and Application 471
1. Construction Industry 471
2. Healthcare Sector 472
3. Pharmaceuticals and Retail 472
4. Financial Services 473
5. Manufacturing 474
6. Logistics and Operations 474
7. Corporate Leadership and HR 475

© 2025 Dr. Yasser Aly Yakout Shehata

Abstract

This compilation presents a suite of instructor-curated, artificially generated case studies designed for the *Business Analytics for Decision Making* course at the Graduate School of Business, Arab Academy for Science, Technology and Maritime Transport (AASTMT). The cases simulate realistic managerial contexts across construction and infrastructure, energy and utilities, healthcare delivery and informatics, financial services, and technology operations. Each narrative is paired with structured, analysis-ready data and a clearly articulated decision problem, enabling graduate students to practise end-to-end analytical reasoning: from scoping questions and shaping data to interpreting findings and communicating executive-level recommendations.

Pedagogically, the collection is grounded in contemporary learning objectives for MBA education: (i) mastery of Exploratory Data Analysis (EDA) and statistical reasoning; (ii) design of interactive dashboards and decision cockpits for executive audiences; (iii) application of AI and machine-learning techniques (classification, regression, clustering, and time-series forecasting) to business decisions; and (iv) ethical, evidence-based judgement under uncertainty. The cases incorporate explicitly defined constructs, measurable KPIs, and domain-specific constraints (e.g., regulatory compliance, quality assurance cycles, cost-of-poor-quality, CAPA resolution, stakeholder satisfaction, and risk exposure) to mirror the trade-offs faced by practising managers. Derived features and composite indices are included to encourage model specification, sensitivity analysis, and robustness checks, while temporal fields support cohort analysis and performance trending.

Methodologically, the cases are designed to bridge theory and practice through a consistent analytic workflow. Students are expected to: formulate decision hypotheses; perform data preparation and feature engineering; visualise patterns and outliers; select and justify modelling approaches; evaluate models using appropriate metrics; and translate quantitative results into clear managerial insights and action plans. Each case provides opportunities for comparative analysis (e.g., supplier or unit benchmarking), intervention design (e.g., targeted CAPA or staffing changes), and scenario planning (e.g., budget, schedule, or risk shocks), reinforcing the connection between analytic outputs and operational levers. Deliverables emphasise persuasive communication, concise executive summaries, well-structured visuals, and traceable logic to cultivate board-facing data storytelling skills.

From an **assurance-of-learning perspective**, the compilation supports transparent assessment across technical and managerial dimensions. Rubric-aligned artifacts include cleaned and annotated datasets, reproducible notebooks or models, and Power BI dashboards with interactive drill-downs and KPI cards. The fictional yet domain-plausible

© 2025 Dr. Yasser Aly Yakout Shehata

nature of the cases safeguards confidentiality while preserving complexity and nuance, allowing rigorous practice without exposing proprietary information. In parallel, the use of AI-assisted drafting and data simulation (with faculty curation) promotes consistency across cases and surfaces teachable moments about responsible AI use, model risk, and bias mitigation.

Collectively, these case studies offer a scalable, standards-aligned resource for experiential learning in analytics-driven management. They cultivate transferable competencies, problem framing, quantitative analysis, decision design, and ethical reasoning preparing graduates to operate confidently at the interface of data, technology, and strategy. Beyond classroom application, the framework is extensible to executive education and industry partnership projects, where rapid prototyping of dashboards, pilots, and "what-if" experiments can accelerate organisational learning. In sum, the compilation advances the programme's objective of developing analytical leaders who can convert data into durable competitive advantage while upholding professional and societal responsibilities.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 1: Steel QA Inspection Analytics

1. Introduction

The **Steel QA Inspection Analytics Dataset** has been meticulously designed to provide a realistic, structured, and analytically rich foundation for the final project of an MBA student specialising in ***Business Analytics for Decision Making***. It simulates operational activities undertaken by a **Quality Control Engineer** working within the Steel Division of Hassan Allam Construction, one of the leading engineering and construction organisations in the MENA region. The dataset is inspired by actual industry practices in steel fabrication, welding inspections, and materials testing, and is constructed to reflect both the technical intricacies of the role and the managerial implications of quality assurance.

The role of a Quality Control Engineer is central to maintaining compliance with engineering standards, ensuring structural safety, and mitigating risks associated with fabrication defects or procedural failures. These responsibilities require systematic inspection protocols, detailed reporting mechanisms, and the ability to respond quickly to non-conformities that may arise in high-risk construction environments. Accordingly, this dataset captures a variety of operational variables including inspection outcomes, compliance rates, material classifications, welding methods, and inspector profiles, all of which play a significant role in the quality performance of construction projects.

In an academic context, the dataset serves as an advanced teaching and learning resource. It supports the practical application of statistical and analytical techniques across the data lifecycle, from data cleaning and transformation to descriptive, diagnostic, and predictive analytics. The dataset is intentionally free of missing values and logically consistent, enabling students to focus on deriving insights, testing hypotheses, and building evidence-based recommendations without first needing to remediate data integrity issues. The inclusion of both raw and derived variables such as *Risk Level* and *Is Pass*, enhances its usability for a wide range of business analytics tasks.

Crucially, the dataset is formatted for compatibility with **Power BI**, making it particularly well-suited for students learning to construct interactive dashboards, develop performance indicators, and communicate complex operational trends to executive decision-makers. It also provides a platform for introducing fundamental concepts in **Artificial Intelligence for Business**, such as supervised learning for defect prediction, clustering for inspection profiling, and optimisation strategies for resource allocation in inspection workflows.

The **Steel QA Inspection Analytics Dataset** thus bridges the gap between academic instruction and professional application. It equips MBA students not only with the technical tools needed to perform high-quality analysis, but also with the strategic mindset required

© 2025 Dr. Yasser Aly Yakout Shehata

to address real-world challenges in quality management, operational excellence, and industrial decision-making. As such, it provides a comprehensive and authentic foundation for experiential learning, decision support modelling, and data storytelling in the context of modern construction and manufacturing industries.

2. Dataset Description

The **Steel QA Inspection Analytics Dataset** consists of 1,000 records, each representing a distinct inspection episode. It incorporates 19 variables encompassing operational, performance, and compliance-related attributes, with a mix of categorical and numerical formats. To enhance analytical capability, two derived variables have been introduced that capture risk and inspection outcomes.

3. Purpose and Application

This dataset is constructed to serve both academic and professional purposes:

- For MBA students, it supports the practical application of business analytics tools and methodologies in a simulated industrial setting.
- For professionals and educators, it provides a robust data foundation for teaching quality monitoring, risk classification, and process optimisation.
- For AI practitioners, it offers structured inputs for predictive modelling and intelligent decision support relevant to quality control operations.

4. Dataset Structure

- **Total Records**: 1,000
- **Variables**: 19
    - **7 qualitative (categorical)**
    - **10 quantitative (numerical)**
    - **2 derived variables** for enriched analysis

5. Qualitative Variables

1. **Project_ID**: Unique identifier for each construction project
2. **Inspector_Level**: Inspector seniority (Novice, Intermediate, Senior)
3. **Material_Type**: Type of steel inspected (Grade A, B, or C)
4. **Welding_Method**: Technique used during welding (MIG, TIG, Stick)

© 2025 Dr. Yasser Aly Yakout Shehata

5. **Standard_Applied**: Inspection standard (EN, Russian, ISO)
6. **Shift**: Work period (Morning, Evening, Night)
7. **Inspection_Result**: Outcome classification (Pass, Minor NCR, Major NCR)

6. Quantitative Variables

8. **Inspection_Date**: Date of the inspection
9. **Equipment_ID**: Machine or asset identifier
10. **Material_Batch_No**: Source batch number of the inspected steel
11. **Weld_ID**: Unique identifier for each weld job
12. **Inspection_Duration**: Time (in minutes) spent on inspection
13. **Test_Score**: Numeric score indicating inspection performance (0–100)
14. **NCR_Count**: Number of non-conformance reports issued
15. **Severity_Score**: Weighted score for NCR severity (0–10 scale)
16. **Time_To_Resolution**: Days required to resolve NCR
17. **Compliance_Rate**: Percentage of procedural compliance (50–100)

7. Derived Variables

18. **Risk_Level**: Categorical indicator derived from Severity_Score

- Low (Severity < 3)
- Medium (3 ≤ Severity < 7)
- High (Severity ≥ 7)

19. **Is_Pass**: Binary indicator of inspection outcome

- 1 = Pass
- 0 = NCR (Minor or Major)

8. Relevance to Power BI and AI for Business

The dataset is structured for optimal usability in Power BI. Qualitative variables can be applied in slicers, filters, and grouping functions, while quantitative fields enable trend analysis, comparisons, and KPI monitoring. The derived fields, **Risk Level** and **Is Pass**, introduce layers of segmentation and classification useful for AI-based dashboards,

© 2025 Dr. Yasser Aly Yakout Shehata

decision trees, and alerting systems. This compatibility enhances the dataset's value in developing data-driven business environments.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify quality trends by inspector level, material type, and shift
- Explore distributions of compliance rate and severity score
- Evaluate test score impact on non-conformance frequency

B. Power BI Visualisation

- Develop dashboards for risk heat maps, inspection success rates, and NCR tracking
- Use bar, line, and scatter charts to show KPI evolution across projects
- Filter and drill-down by standards, welding methods, and risk level

C. AI-Driven Business Insights

- Classify inspection outcomes using decision trees or logistic regression
- Forecast high-risk inspections using severity and compliance indicators
- Cluster similar inspections to optimise inspection strategies and scheduling

10. Conclusion

The **Steel QA Inspection Analytics Dataset** is a purpose-built educational and analytical asset that empowers MBA students to connect theory with practice. With carefully designed variables and derived indicators, it supports rigorous business analytics, dashboard development, and AI experimentation. Through its use, students gain the ability to turn raw operational data into actionable insights, contributing to better decision-making, quality control, and strategic outcomes in industrial and construction contexts.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 2: ElDabaa Nuclear QA Business Analytics

1. Introduction

The **ElDabaa Nuclear QA Business Analytics Dataset** is a high-fidelity synthetic dataset created to support the capstone project of an MBA student engaged in the ***Business Analytics for Decision Making*** **course**. This dataset simulates the responsibilities of a **Project Quality Manager** operating within the context of the El Dabaa Nuclear Power Plant construction, a landmark infrastructure initiative involving multiple contractors, complex quality control procedures, and adherence to strict regulatory standards. The dataset has been carefully structured to reflect the operational, managerial, and analytical challenges faced by professionals in large-scale industrial projects where quality assurance is not only a procedural function but also a strategic imperative.

In the domain of nuclear construction, quality management is paramount to ensuring compliance with international safety standards, minimising operational risks, and protecting long-term investment in critical infrastructure. The Project Quality Manager plays a pivotal role in coordinating inspections, managing non-conformance reports, overseeing corrective and preventive actions, and generating data-driven insights for performance optimisation. The dataset replicates these activities by providing a detailed representation of inspection events, audit findings, laboratory test results, rework metrics, supplier performance data, and project-wide quality indicators.

The dataset consists of 1,000 records, each corresponding to an inspection, lab test, or audit conducted on one of four nuclear reactor units. Key attributes include inspection type, inspector level, material grade, supplier identity, severity scores, compliance rates, and cost-of-poor-quality estimates. To enhance its analytical value, the dataset integrates several derived variables, such as binary outcome flags, audit and lab test indicators, and a composite **Quality Index** designed to summarise performance efficiency and procedural adherence.

This resource has been developed not only to support academic instruction in exploratory data analysis and visual analytics but also to introduce students to the early-stage application of artificial intelligence (AI) in quality management. It allows users to construct Power BI dashboards, perform diagnostic analytics, and model predictive outcomes using supervised learning techniques. Through this dataset, **MBA candidates** can simulate strategic decision-making in real-world industrial environments, demonstrating their ability to interpret operational data, monitor KPIs, and generate actionable recommendations.

Crucially, the **ElDabaa Nuclear QA Business Analytics Dataset** bridges the gap between classroom-based business analytics education and the complex, data-intensive

© 2025 Dr. Yasser Aly Yakout Shehata

environments encountered in engineering, infrastructure, and construction management. It prepares future managers to use data not only to describe performance, but to anticipate risk, identify patterns, and drive improvements that align with both technical standards and organisational goals.

2. Dataset Description

The **ElDabaa Nuclear QA Business Analytics Dataset** contains 1,000 synthetic records, each representing a unique inspection, lab test, or audit event conducted across one of four nuclear reactor units. Each record includes a blend of categorical and numerical data capturing core aspects of quality management in large-scale construction projects. The dataset is enriched with several derived fields, including indicators for inspection success, composite quality scoring, and operational flags, which support analytics, visualisation, and machine learning applications.

3. Purpose and Application

The dataset serves multiple educational and professional functions:

- **For MBA students**, it offers an applied learning experience in business analytics and decision support using industry-aligned data.
- **For academic instruction**, it provides a consistent, high-quality data source for coursework in data visualisation, EDA, performance reporting, and introductory AI.
- **For project-oriented learning**, it supports the analysis of supplier performance, quality trends, defect management, and cost-related KPIs in a complex, high-risk domain.

4. Dataset Structure

- **Total Records**: 1,000
- **Total Variables**: 24
    - 7 qualitative (categorical)
    - 10 quantitative (numerical)
    - 7 derived or enhanced fields (flags, scores, or indices)

5. Qualitative Variables

1. **Reactor_Unit** – The nuclear unit where the inspection was conducted
2. **Project_ID** – Project reference code

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Inspection_Type** – Type of quality event (Inspection, Lab Test, Audit)
4. **Inspector_Level** – Seniority level of the inspector (Junior, Mid, Senior)
5. **Material_Type** – Classification of steel or other materials inspected
6. **Supplier_ID** – Vendor associated with the inspection or material
7. **Inspection_Result** – Outcome of the inspection (Pass, Minor NCR, Major NCR)

6. Quantitative Variables

8. **Cost_of_Poor_Quality** – Estimated monetary value of poor-quality outcomes
9. **NCR_Count** – Number of Non-Conformance Reports raised
10. **Severity_Score** – Weighted score indicating seriousness of issues
11. **CAPA_Time** – Number of days taken to close corrective actions
12. **Compliance_Rate** – Degree of procedural or regulatory adherence
13. **Audit_Score_Filled** – Normalised audit score with gaps handled
14. **Test_Value_Filled** – Lab test result value where applicable
15. **Rework_Hours** – Time spent correcting quality defects
16. **Monthly_KPI_Score** – Composite score representing monthly performance
17. **Inspection_Date** – Date of the quality event

7. Derived Variables

18. **Cost_Category** – Classification into Low, Medium, or High cost impact
19. **Is_Pass** – Binary indicator for whether the inspection was successful
20. **Is_Audit** – Flag indicating whether the inspection was an audit
21. **Is_LabTest** – Flag indicating whether the event was a lab test
22. **Audit_Score_Norm** – Audit score set to zero when not relevant
23. **Test_Value_Norm** – Lab test value set to zero when not applicable
24. **Quality_Index** – A composite index combining audit scores, KPI scores, compliance, and cost efficiency

© 2025 Dr. Yasser Aly Yakout Shehata

8. Relevance to Power BI and AI for Business

This dataset has been engineered for direct integration with Power BI and other business intelligence platforms. Its categorical fields enable powerful filtering, grouping, and drill-through capabilities, while its numerical fields support scorecards, time-series plots, and performance dashboards. The inclusion of derived fields expands the dataset's usability for AI models such as decision trees, classifiers, or clustering algorithms.

- **Power BI-ready** for professional reporting, interactive dashboards, and executive briefings
- **AI-compatible** structure allows modelling of risk, performance, and outcome prediction
- **Analytically rich** for KPI analysis, trend detection, and supplier segmentation

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify inspection result patterns by reactor unit or supplier
- Explore cost and rework time distributions
- Analyse compliance and KPI fluctuations across project timelines

B. Power BI Visualisation

- Build dashboards showing NCR frequency, CAPA resolution speed, and supplier performance
- Map cost-of-poor-quality trends by project phase or inspection type
- Visualise Quality_Index by reactor unit or inspector level

C. AI-Driven Business Insights

- Predict inspection success using severity, compliance, and rework indicators
- Classify cost risk categories based on historical performance data
- Forecast monthly KPI scores based on inspection volume and CAPA response time

10. Conclusion

The **ElDabaa Nuclear QA Business Analytics Dataset** is a fully developed analytical resource designed to enhance the business analytics capabilities of MBA students while anchoring those skills in a real-world, high-stakes quality management environment. It

© 2025 Dr. Yasser Aly Yakout Shehata

enables comprehensive engagement with performance data, supports strategic quality decision-making, and encourages the development of insights through modern data tools and AI frameworks. It is an ideal model for training future quality leaders in complex project settings.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 3: Dabaa QA Insight Analytics Project

1. Introduction

The **Dabaa QA Insight Analytics Project Dataset** is a synthetic dataset developed to simulate the data environment encountered by a **Project Quality Manager** at the El Dabaa Nuclear Power Plant in Egypt. Designed as a capstone resource for an MBA student undertaking a final project in the course *Business Analytics for Decision Making*, this dataset enables students to explore the practical intersection between quality management, data analytics, and strategic decision-making in a large-scale nuclear infrastructure context.

The El Dabaa project is Egypt's flagship civil nuclear initiative, demanding rigorous standards for quality, safety, and regulatory compliance. The role of the Project Quality Manager is central to the oversight of inspection processes, non-conformance handling, supplier evaluations, and corrective/preventive action cycles. The dataset mirrors this professional setting by providing a structured and detailed record of 1,000 inspection events, enabling a realistic application of business analytics tools to operational and managerial quality challenges.

By simulating diverse quality events—including field inspections, lab tests, and internal audits—the dataset helps students develop technical and strategic skills. It integrates raw variables and derived analytical fields, fostering insights through Exploratory Data Analysis (EDA), dashboard construction in Power BI, and machine learning-driven business applications. The dataset is optimised for educational use, ensuring ethical data modelling while maintaining industrial realism.

2. Dataset Description

The **Dabaa QA Insight Analytics Project Dataset** consists of 1,000 records, each corresponding to a quality assurance inspection conducted during various construction stages at the El Dabaa Nuclear Power Plant. The dataset is multi-dimensional and includes:

- Reactor unit identifiers
- Project segment IDs
- Inspection types and personnel qualifications
- Supplier and material data
- Inspection results and compliance outcomes
- Non-conformance and rework metrics
- Financial implications of quality performance

© 2025 Dr. Yasser Aly Yakout Shehata

The dataset is enriched with derived variables to enhance insight generation and support analytical depth, such as normalised audit scores, cost categories, inspection flags, and a composite Quality Index.

3. Purpose and Application

This dataset supports MBA students in applying analytical thinking to quality-related decision-making in complex project environments. Its objectives include:

- Strengthening proficiency in data exploration and statistical reasoning
- Developing visual storytelling capabilities through Power BI dashboards
- Building AI and machine learning models for quality prediction and performance optimisation
- Reinforcing the connection between analytics and executive decision-making

4. Dataset Structure

The dataset includes 27 variables divided into:

5. Qualitative Variables

1. **Reactor_Unit** – The unit (1 to 4) within the nuclear facility where inspections are performed
2. **Project_ID** – Unique code for the specific construction component under inspection
3. **Inspection_Type** – Categorises the nature of inspection (Field Inspection, Lab Test, Internal Audit)
4. **Inspector_Level** – Qualification level of the inspector (Associate, Professional, Expert)
5. **Material_Type** – The material examined (Rebar, Structural Steel, Concrete Mix)
6. **Supplier_ID** – Encoded supplier information for traceability and accountability
7. **Inspection_Result** – Outcome classification (Acceptable, Minor NCR, Major NCR)

6. Quantitative Variables

8. **NCR_Count** – Number of non-conformance reports raised during the inspection
9. **Severity_Score** – Weighted score measuring severity of issues identified

© 2025 Dr. Yasser Aly Yakout Shehata

10. **CAPA_Time** – Number of days required to implement corrective and preventive actions
11. **Compliance_Rate** – Percentage of criteria met during inspection
12. **Audit_Score** – Performance score from internal audits (if applicable)
13. **Test_Value** – Result of lab testing (if applicable)
14. **Cost_of_Poor_Quality** – Financial cost incurred due to quality failures
15. **Rework_Hours** – Estimated time for reworking the non-compliant item
16. **Monthly_KPI_Score** – Quality performance metric aggregated monthly
17. **Inspection_Duration** – Time spent conducting the inspection

7. Derived Variables

18. **Cost_Category** – Categorical classification of cost levels (Low, Medium, High)
19. **Is_Pass** – Indicates if the inspection had zero non-conformance issues
20. **Is_Audit** – Flag for internal audit inspections
21. **Is_LabTest** – Flag for lab test inspections
22. **Audit_Score_Norm** – Normalised audit score (0 to 1 scale)
23. **Test_Value_Norm** – Normalised lab test value
24. **Quality_Index** – Composite indicator incorporating audit, KPI, compliance, and cost metrics
25. **Inspection_Month** – Derived date field for temporal analysis

8. Relevance to Power BI and AI for Business

The dataset is structured to integrate seamlessly with Power BI for:

- Drill-down dashboards on quality performance by unit, project, supplier, and month
- Trend visualisation for compliance and cost metrics
- Heatmaps, KPI indicators, and forecasting visuals

In the AI context, the dataset supports:

- Classification models for inspection results or supplier risk profiles
- Regression models for cost forecasting or severity prediction

© 2025 Dr. Yasser Aly Yakout Shehata

- Time-series analysis on monthly quality trends

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Uncover data distributions, outliers, and conditional relationships
- Analyse compliance trends and rework burden by inspection type

B. Power BI Visualisation

- Create layered dashboards using multiple filters and slicers
- Develop inspection cycle time KPIs and compliance dashboards

C. AI-Driven Business Insights

- Train classification models to predict inspection results
- Use clustering to segment projects by risk and cost characteristics
- Forecast audit performance based on material and supplier data

10. Conclusion

The **Dabaa QA Insight Analytics Project Dataset** offers an academically grounded, professionally inspired platform for applying business analytics in a high-stakes, quality-sensitive environment. By simulating the inspection, reporting, and corrective action workflows of a Project Quality Manager at a nuclear power plant, the dataset enables students to transition from theoretical knowledge to practical insight generation. Its compatibility with modern analytics tools and techniques—ranging from Power BI dashboards to AI-driven predictive models—makes it a vital resource for MBA candidates preparing to lead in data-informed roles across industry sectors.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 4: HSE Analytics For Decision Making

1. Introduction

The **HSE Analytics For Decision Making** dataset has been thoughtfully constructed as an experiential learning tool for MBA students enrolled in the ***Business Analytics for Decision Making* course**. It reflects the operational complexities and data challenges faced by a **Health, Safety, and Environment (HSE) Head** within high-risk industries such as construction and manufacturing. In today's rapidly evolving business landscape, the HSE function is no longer confined to compliance monitoring or post-incident reporting. It has become an integral part of strategic and operational planning, where data-driven decision-making plays a central role in ensuring safety, reducing environmental risks, and driving organizational resilience.

This dataset serves as a practical simulation of the real-world conditions encountered in industrial environments. Each record represents one operational day at a given worksite, capturing multiple dimensions of safety performance, environmental impact, workforce behavior, and compliance outcomes. Variables include key safety indicators such as incident counts, near misses, PPE usage rates, audit findings, training coverage, emissions levels, and spill volumes. By combining both qualitative and quantitative fields, the dataset supports complex analytical tasks, from categorical comparison and segmentation to trend forecasting and risk prediction.

To enrich its analytical value, the dataset includes several **derived variables** that are calculated from raw data points to support deeper insights. These derived fields include a normalized incident rate per thousand hours worked, a synthetic risk score that consolidates audit, emissions, and environmental data, and a high-risk operational flag based on key threshold breaches. These variables mimic the types of KPIs and composite indices that real-world safety leaders rely on to assess performance and prioritize action.

The structure and logic of this dataset make it particularly well-suited to a wide range of business analytics applications. It provides a meaningful foundation for **Exploratory Data Analysis (EDA)**, helping students uncover relationships, detect anomalies, and identify key drivers of safety outcomes. The categorical variables can be used in **Power BI** for the creation of interactive dashboards, heatmaps, and performance filters, while the numerical and derived variables are ideal for applying **AI and machine learning techniques**, such as classification models and predictive analytics.

In an academic context, this dataset encourages critical thinking, problem-solving, and evidence-based reasoning, core competencies for modern business leaders. Students are not only expected to manipulate and analyze data but also to derive meaningful insights and

© 2025 Dr. Yasser Aly Yakout Shehata

translate them into actionable recommendations. Whether exploring the causes of non-compliance, evaluating the impact of safety training, or predicting high-risk days using AI models, the dataset challenges students to think like strategic decision-makers.

Ultimately, **HSE Analytics For Decision Making** serves as a bridge between academic theory and practical business application. It empowers MBA students to simulate the responsibilities of an HSE Head and to understand the broader organizational impact of data-informed safety management. This makes it not just a dataset, but a dynamic educational platform for preparing future leaders in data-driven decision-making and ethical business practice.

2. Dataset Description

HSE Analytics For Decision Making contains **1000 records**, each representing a day's operational data from a high-risk work environment. The dataset includes a balanced mix of categorical and numerical fields, along with derived variables that support analytical depth, forecasting, and classification.

All data points reflect a realistic simulation of what an HSE Head would encounter, including variables related to incident reporting, safety audits, workforce hours, environmental metrics, and compliance monitoring.

3. Purpose and Application

This dataset is specifically tailored to:

- Enable MBA students to perform **Exploratory Data Analysis (EDA)**
- Build meaningful **Power BI dashboards** for decision support
- Apply **AI techniques** to predict operational risk and safety performance
- Simulate the role of an HSE Head leveraging data to drive strategic and operational actions

4. Dataset Structure

- **Total Records**: 1000 (one per operational day)
- **Total Variables**: 20
    - 7 Qualitative (Categorical)
    - 10 Quantitative (Numerical)
    - 3 Derived Variables (Computed for deeper insights)

© 2025 Dr. Yasser Aly Yakout Shehata

5. Qualitative Variables

1. **Site Risk Level** – Risk category of the worksite: Low, Medium, or High
2. **Weather Condition** – Environmental condition: Clear, Rain, Windy, or Extreme Heat
3. **Shift Type** – Operational shift timing: Morning, Evening, or Night
4. **Contractor Compliance** – Safety performance of contractors: Poor, Fair, Good, or Excellent
5. **Audit Result** – Outcome of safety audit: Pass, Minor NC, Major NC, or Failed
6. **Training Session** – Indicates whether a safety training occurred: Yes or No
7. **PPE Compliance** – Categorical safety equipment usage level: Low, Moderate, or High

6. Quantitative Variables

1. **Near Miss Count** – Number of near-miss incidents reported
2. **Incident Count** – Count of reportable safety incidents
3. **TRIR** – Total Recordable Incident Rate per 200,000 hours worked
4. **Corrective Actions Logged** – Safety corrections initiated
5. **Audit Findings Count** – Number of issues found in safety audits
6. **PPE Detection Score (%)** – Visual compliance score from PPE detection systems
7. **Environmental Spill Volume (liters)** – Volume of hazardous material spilled
8. **Emissions Level (CO2e)** – Daily environmental emissions in carbon equivalent
9. **Time to Closure (hours)** – Average time to resolve incidents
10. **Employee Hours** – Workforce hours logged across all shifts

7. Derived Variables

1. **PPE Compliance (Categorical)** – Categorized from PPE Detection Score:
    - Low (<70%)
    - Moderate (70–90%)
    - High (>90%)
2. **Incident Rate per 1000 Hours** – Normalized incident metric: Incident Count / Employee Hours × 1000

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Synthetic Risk Score** – Weighted risk index combining:
    - Audit Findings (30%)
    - Normalized Spill Volume (30%)
    - Normalized Emissions Level (40%)
4. **High Risk Flag** – Flags operational days with any of the following:
    - Incident Count > 1
    - TRIR > 3.5
    - PPE Compliance = Low

8. Relevance to Power BI and AI for Business

HSE Analytics For Decision Making is an ideal foundation for building:

- **Power BI dashboards** that filter, group, and drill into safety data
- **Machine learning models** that classify or forecast operational risk
- **Performance heatmaps** by site risk, contractor compliance, and shift type
- **Interactive reports** with real-time alerts based on derived fields

The dataset supports the development of robust, decision-ready outputs by combining traditional business metrics with modern AI-compatible features.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify patterns in incident reporting by shift or weather condition
- Evaluate impact of training sessions on PPE compliance
- Analyze audit findings across risk levels and contractor performance

B. Power BI Visualization

- Design KPI dashboards for TRIR, emissions, and training coverage
- Create time series visuals for trend analysis
- Build risk dashboards with filters for risk level, shift, and audit status

© 2025 Dr. Yasser Aly Yakout Shehata

C. AI-Driven Business Insights

- Predict High Risk Flags using classification models
- Use clustering on Synthetic Risk Scores for operational segmentation
- Recommend proactive safety interventions using predictive analytics

10. Conclusion

HSE Analytics For Decision Making empowers MBA students to engage in applied learning, combining business intuition with data science techniques. It simulates the decision environment of an HSE Head, allowing students to explore how data can transform safety management from reactive to predictive.

With relevance to both Power BI and AI applications in business, this dataset builds real-world capabilities in analytics, critical thinking, and ethical decision-making essential traits for the future leaders of data-driven organizations.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 5: MEP Analytics For Decision Making

Role Focus: Project Manager – MEP in the Construction Industry

1. Introduction

In today's fast-evolving construction industry, the role of a Project Manager – MEP (Mechanical, Electrical, and Plumbing) has expanded far beyond the traditional responsibilities of overseeing technical implementation. It now encompasses strategic decision-making, stakeholder coordination, vendor management, regulatory compliance, and risk mitigation, all areas increasingly driven by data and analytics. As construction projects grow in size, complexity, and technological integration, there is a growing demand for professionals who can transform operational data into actionable insights to ensure project efficiency, safety, cost control, and stakeholder satisfaction.

The **MEP Analytics For Decision Making** dataset has been developed specifically to support this new paradigm. It provides MBA students particularly those enrolled in the ***Business Analytics for Decision Making*** **course** with an opportunity to engage with a rich, realistic, and analytically robust dataset that simulates the daily environment of an **MEP Project Manager**. Each of the 1000 records in this dataset represents a snapshot of a construction project's day-to-day operations, offering a combination of technical metrics, management indicators, and derived insights that mirror the kinds of data that inform high-stakes decisions on the job site.

This dataset has been curated not only for its diversity and realism but also for its pedagogical value. Students using it can explore how coordination quality affects issue resolution, how vendor submission quality impacts compliance, or how manpower deviations contribute to schedule and budget risks. More importantly, the inclusion of derived fields such as *Schedule Risk Score*, *Change Order Impact Index*, and a *High Attention Flag* adds layers of analytical depth. These fields are designed to challenge students to think critically about relationships between variables, prioritize key performance indicators, and use data to guide decision-making under uncertainty.

With a balanced combination of categorical variables (such as site phase, technical compliance, and training status) and continuous variables (such as procurement lead time, issue resolution hours, and budget variance), the dataset is highly suitable for Exploratory Data Analysis (EDA), Power BI dashboard development, and AI-powered modeling techniques. It reflects how modern construction teams use analytics platforms to visualize patterns, segment risk, allocate resources, and proactively address emerging challenges.

© 2025 Dr. Yasser Aly Yakout Shehata

Furthermore, this dataset is intended to help students bridge the gap between academic learning and professional application. While the context is specific to the construction sector and the MEP function, the underlying analytical skills, data cleaning, transformation, modeling, and visualization are broadly transferable to a wide range of industries. As MBA students learn to use tools such as Power BI, Python, or machine learning platforms to extract insights from this data, they simultaneously prepare themselves for leadership roles in data-driven business environments.

In essence, **MEP Analytics For Decision Making** is more than a learning dataset, it is a simulated decision-making environment. It empowers students to assume the mindset of a modern MEP project leader: someone who balances technical knowledge with business acumen, strategic foresight with operational control, and most importantly, intuition with evidence. Whether used to build a risk monitoring dashboard, model procurement performance, or predict non-compliance, this dataset lays the foundation for a new generation of analytics-savvy professionals ready to lead complex projects through data and insight.

2. Dataset Description

MEP Analytics For Decision Making comprises 1000 records, each reflecting a daily data point during the design, procurement, or execution phase of an MEP project. The dataset blends qualitative indicators and quantitative measures, along with derived fields that support advanced business analytics and AI applications.

3. Purpose and Application

This dataset was created to help MBA students:

- Conduct Exploratory Data Analysis (EDA) using real-world project variables
- Create dynamic dashboards using Power BI
- Build AI-driven models for classification, prediction, and performance optimization
- Practice translating technical metrics into business decisions

Students can simulate the experience of managing a live MEP project with complex, interrelated variables, analyzing trends, detecting risks, and making informed

4. Dataset Structure

- **Total Records**: 1000 (one per operational day)
- **Total Variables**: 23

© 2025 Dr. Yasser Aly Yakout Shehata

- 7 Qualitative (categorical)
- 10 Quantitative (numerical)
- 6 Derived (analytical indicators)

5. Qualitative Variables

1. **Site Phase** – Indicates the phase of the project: Design, Procurement, or Execution
2. **Technical Compliance** – Reports whether the project met technical specifications: Yes, Partial, or No
3. **Stakeholder Coordination** – Qualitative rating of team coordination: Poor, Fair, Good, Excellent
4. **Design Change Occurred** – Whether a design change was introduced on that day (Yes or No)
5. **Vendor Submission Quality** – Evaluated quality of vendor documentation or material: Poor to Excellent
6. **Contractor Performance** – Overall performance rating of subcontractors on that day
7. **Training Held** – Indicates whether a safety or technical training session was conducted

6. Quantitative Variables

1. **Design Review Turnaround Days** – Time taken to complete a design review
2. **Vendor Submittals Reviewed** – Number of vendor documents reviewed in one day
3. **Change Orders Raised** – Number of MEP-related change orders initiated
4. **Change Order Value USD** – Financial value of the change orders in US dollars
5. **Manpower Hours Deviation Percent** – Percentage deviation from planned labor hours
6. **Procurement Lead Time Days** – Number of days required to procure materials or equipment
7. **Non Conformance Count** – Number of quality or compliance issues flagged
8. **Budget Variance Percent** – Percentage difference from the projected budget

© 2025 Dr. Yasser Aly Yakout Shehata

9. **Technical RFI Count** – Number of requests for information submitted to resolve uncertainties
10. **Issue Resolution Time Hours** – Time spent resolving issues on site

7. Derived Variables

1. **Compliance Rate Percent** – Numerical translation of technical compliance (Yes = 100, Partial = 70, No = 40)
2. **Coordination Score** – Numerical scale of coordination rating (1 to 4)
3. **Change Order Impact Index** – Calculated as Change Order Value × RFI Count
4. **Schedule Risk Score** – Weighted metric combining turnaround and procurement delays
5. **Performance Category** – Classification (High, Moderate, Low) based on compliance and coordination
6. **High Attention Flag** – Indicates whether a record requires urgent attention due to key risk thresholds

8. Relevance to Power BI and AI for Business

This dataset is fully compatible with modern business intelligence and AI tools. In **Power BI**, users can:

- Build KPI dashboards (e.g., average resolution time, budget variance)
- Use filters to explore trends across site phases, contractor performance, or vendor quality
- Generate heatmaps and charts to reveal insights hidden in operational data

For **AI applications**, the dataset supports:

- Supervised learning for classification (e.g., predicting high-risk days)
- Regression analysis to forecast procurement delays or budget variance
- Clustering and segmentation based on derived performance metrics

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify patterns linking vendor submission quality with non-conformance counts

© 2025 Dr. Yasser Aly Yakout Shehata

- Analyze the impact of stakeholder coordination on resolution times
- Detect early signs of project delays through procurement and turnaround metrics

B. Power BI Visualization

- Create dashboards that track design change trends and training coverage
- Visualize high-risk project days based on the High Attention Flag
- Drill down into the impact of coordination and compliance on project outcomes

C. AI-Driven Business Insights

- Predict risk-prone days using classification models
- Forecast design delays using regression
- Cluster records by performance category for targeted intervention

10. Conclusion

The **MEP Analytics For Decision Making** dataset offers MBA students a unique opportunity to immerse themselves in real-world project analytics. By working with variables that mirror the daily operations of an MEP Project Manager, students will develop both technical proficiency and strategic thinking. This dataset bridges the gap between business theory and professional practice, helping future managers make confident, data-driven decisions in complex environments.

It is not just a dataset, it is a hands-on simulation of how modern infrastructure projects are managed, optimized, and led using data.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 6: Construction Project Management Analytics

1. Introduction

This dataset, titled **Construction Project Management Analytics**, has been carefully developed to simulate realistic, multifaceted scenarios faced by **Project Managers in the construction industry**. It is intended to serve as a practical and academic tool for MBA students, particularly those undertaking final projects within the **Business Analytics for Decision Making course**. Designed to mirror the complexity and strategic considerations involved in large-scale project execution, the dataset provides a rich foundation for the application of data analytics, visualisation, and artificial intelligence in a business context.

The construction industry presents a uniquely data-intensive environment, characterised by dynamic interactions between multiple variables such as budget, time, scope, quality, risk, and stakeholder management. Project Managers operating in this domain are expected to maintain a balance between these factors while ensuring that the project is delivered successfully. This dataset reflects that reality by integrating a blend of quantitative and qualitative dimensions that are directly relevant to project outcomes. The data structure has been informed by established project management methodologies and metrics, including the ***Earned Value Management (EVM) framework*** and the ***traditional triple constraint model***.

Constructed with 1,000 synthetic records, each entry in the dataset represents a distinct construction project overseen by a Project Manager. The projects span various types, Residential, Commercial, and Infrastructure, each with different operational and managerial challenges. Variables such as planned and actual budgets, schedules, scope changes, and risk counts are logically correlated and follow plausible distribution patterns based on real-world expectations. These elements are further enriched with performance classification variables such as “Over Budget,” “Over Schedule,” and “Performance Status,” which allow for deeper diagnostic and predictive analysis.

This dataset is particularly well-suited for students wishing to conduct Exploratory Data Analysis (EDA), build interactive dashboards using tools such as Power BI, and explore AI applications such as classification and forecasting in a business setting. By simulating real-world performance metrics and decision variables, the dataset allows students to extract strategic insights, make data-driven recommendations, and understand the cause-and-effect relationships that influence project success or failure.

From an educational standpoint, this dataset also encourages the integration of descriptive, diagnostic, and predictive analytics. It empowers students to not only describe patterns and anomalies but also to anticipate risks and optimise decision-making processes. As such,

© 2025 Dr. Yasser Aly Yakout Shehata

the **Construction Project Management Analytics** dataset stands as a versatile and realistic training ground for future business leaders and analysts, bridging the gap between theoretical knowledge and practical decision-making in the construction project management field.

2. Dataset Description

The dataset contains **1,000 synthetic records**, each representing a unique construction project. It includes both raw and derived variables, offering insights into cost, schedule, team dynamics, risk management, and stakeholder satisfaction. The data has been structured for clarity, integrity, and analytical compatibility, including appropriate types and formats for visualisation tools and statistical software.

3. Purpose and Application

The dataset has been created to serve the following objectives:

- Simulate real-world construction project management environments for academic use.
- Enable MBA students to apply business analytics tools and techniques in a project-based context.
- Support a wide range of data analysis methods, including descriptive analytics, visualisation, and machine learning applications.
- Facilitate decision-making exercises based on practical variables that influence project outcomes.

4. Dataset Structure

The dataset consists of **20 variables** categorised into three main groups:

- **7 Qualitative Variables** (categorical)
- **10 Quantitative Variables** (numerical)
- **3 Derived Variables** (calculated performance indicators)

Each row represents one project managed by a Project Manager.

5. Qualitative Variables

1. **Project Type** – Category of project: Residential, Commercial, or Infrastructure.
2. **Quality Level** – Outcome assessment: High, Medium, or Low.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Communication Frequency** – Regularity of communication: Daily, Weekly, or Biweekly.
4. **Stakeholder Satisfaction** – Perceived satisfaction: High, Medium, or Low.
5. **Procurement Complexity** – Degree of difficulty in sourcing materials or services: Low, Medium, or High.
6. **Change Management Rigour** – Formality of change control processes: Informal, Standardised, or Formal.
7. **Integration Level** – Level of coordination across project subsystems: Low, Medium, or High.

6. Quantitative Variables

1. **Budget Planned USD** – Initial budget estimate for the project.
2. **Budget Actual USD** – Final actual expenditure.
3. **Schedule Planned Days** – Planned duration in calendar days.
4. **Schedule Actual Days** – Actual time taken to complete the project.
5. **EV Percent Complete** – Earned value percentage reflecting project progress.
6. **Cost Variance USD** – Difference between planned and actual cost.
7. **Schedule Variance Days** – Difference between planned and actual schedule duration.
8. **Scope Change Count** – Number of approved changes to original project scope.
9. **Risk Count** – Number of documented risks identified during project execution.
10. **Team Size** – Size of the core team working on the project.

7. Derived Variables

1. **Over Budget** – A logical indicator of cost overrun (Yes or No).
2. **Over Schedule** – A logical indicator of time overrun (Yes or No).
3. **Performance Status** – Classification based on EV percent complete:
    - **Ahead of Schedule** (EV ≥ 75%)
    - **On Track** (50% ≤ EV < 75%)

© 2025 Dr. Yasser Aly Yakout Shehata

- **Behind Schedule** (EV < 50%)

These variables enable more nuanced insights and facilitate segment-based analysis and decision-making.

8. Relevance to Power BI and AI for Business

The dataset is fully compatible with Microsoft Power BI and supports a wide range of dashboarding and visualisation functions. It includes performance indicators, status flags, and categorical dimensions that enable slicing, filtering, and dynamic report generation. Students can use this dataset to:

- Build dashboards tracking cost and schedule performance.
- Evaluate stakeholder satisfaction in relation to project complexity.
- Monitor project risk levels and change management practices.

Moreover, the dataset structure supports the application of AI techniques, including:

- Predictive analytics to forecast cost and time overruns.
- Classification models for performance segmentation.
- Clustering to identify project archetypes and behaviour patterns.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify patterns in cost and schedule variance.
- Explore relationships among risk, scope changes, and team size.
- Analyse satisfaction outcomes by project type or quality level.

B. Power BI Visualisation

- Develop KPI dashboards.
- Use filters for deep dives into project categories.
- Visualise risk and scope change distribution.

C. AI-Driven Business Insights

- Train models to predict performance classification.
- Use supervised learning to identify high-risk projects.

© 2025 Dr. Yasser Aly Yakout Shehata

- Simulate decision scenarios for future planning.

10. Conclusion

The **Construction Project Management Analytics** dataset offers MBA students a realistic, strategically designed platform for applying analytics in the context of complex project environments. It integrates critical variables, derived insights, and analytical depth to support learning and innovation in both descriptive and predictive modelling. By working with this dataset, students can bridge academic theory and managerial practice, preparing them for real-world challenges in project leadership, digital transformation, and strategic business decision-making.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 7: Corporate Leadership Performance Analytics

1. Introduction

The **Corporate Leadership Performance Analytics** dataset has been meticulously developed to serve as a practical, analytically rich, and professionally relevant resource for MBA students undertaking their final projects in the **Business Analytics for Decision Making course**. Specifically designed to reflect the operational realities and strategic responsibilities of a **Managing Director** within the **construction and infrastructure sector**, this dataset offers a comprehensive simulation of executive decision-making metrics and organisational dynamics. It presents students with the opportunity to explore, interpret, and act upon performance data that are closely aligned with real-world business leadership scenarios.

The role of a **Managing Director in construction-oriented enterprises** encompasses a broad range of complex functions. These include overseeing multi-million-dollar budgets, ensuring profitable operations, directing strategic projects, managing risk and compliance, fostering innovation, and maintaining high levels of stakeholder confidence. The position requires not only strategic vision but also the capability to align operational execution with long-term corporate goals. This dataset captures these interdependencies through a variety of both qualitative and quantitative variables that represent core business functions and leadership challenges.

With a total of 1000 synthetic records, each representing a distinct reporting period or strategic review cycle, the dataset includes performance data across diverse company types, residential, commercial, and infrastructure, providing a realistic variation in outcomes, challenges, and achievements. The data encompasses key financial metrics such as revenue, profit, and budget variance, alongside strategic indicators such as project delivery rates, employee turnover, regulatory compliance, innovation capacity, and communication cadence. These are complemented by derived variables that support richer analytical perspectives, such as *profit margin categorisation*, *strategic success indexing*, and *revenue growth trends*.

From an educational standpoint, the dataset is designed to bridge the gap between theoretical instruction and practical application. It encourages students to develop analytical fluency by applying descriptive and diagnostic methods, creating interactive dashboards using tools like Power BI, and experimenting with AI techniques including classification, clustering, and forecasting. Through structured exploration of this dataset, students will be able to identify patterns, test hypotheses, and construct data-driven narratives that simulate real executive reporting and board-level insight generation.

© 2025 Dr. Yasser Aly Yakout Shehata

The **Corporate Leadership Performance Analytics** dataset is also an ideal platform for teaching the interconnectivity between leadership behaviours, organisational performance, and stakeholder outcomes. It reflects the data-intensive nature of modern executive decision-making and promotes a systems-thinking approach to business analytics. Whether used to build performance dashboards, run exploratory data analyses, or simulate predictive business models, this dataset offers MBA students a comprehensive, relevant, and strategically meaningful opportunity to practise and refine their analytical skills in preparation for senior leadership roles.

2. Dataset Description

The dataset comprises **1000 synthetic records**, each representing a reporting period or strategic outcome overseen by a Managing Director in the construction sector. The structure combines financial metrics, strategic indicators, organisational attributes, and performance classifications to simulate multidimensional business leadership contexts. It includes:

- **7 qualitative variables** representing managerial environment and organisational status
- **10 quantitative variables** measuring financial, operational, and engagement KPIs
- **6 derived variables** constructed to enable analytical segmentation and decision-support modelling

The dataset is ready for immediate application in Power BI, Excel, or Python/R-based analytics environments.

3. Purpose and Application

This dataset supports the following educational and professional objectives:

- Enable MBA students to practise **Exploratory Data Analysis (EDA)** in a strategic business context
- Allow learners to build **interactive dashboards** and KPI reports using **Power BI**
- Facilitate **AI-driven business modelling**, such as classification and forecasting
- Encourage critical analysis of how executive decisions influence profitability, efficiency, and stakeholder outcomes
- Support data storytelling, hypothesis testing, and visual reporting aligned with board-level decision-making

© 2025 Dr. Yasser Aly Yakout Shehata

4. Dataset Structure

The dataset includes a total of **23 variables**:

- **7 Qualitative Variables**: Categorical fields indicating the strategic, cultural, and organisational context
- **10 Quantitative Variables**: Numeric KPIs reflecting core leadership performance domains
- **6 Derived Variables**: Calculated fields for enhanced modelling and interpretation

Each record reflects a single reporting cycle or scenario under the leadership of a Managing Director.

5. Qualitative Variables

1. **Company Focus Area** – Business domain emphasis: Residential, Commercial, or Infrastructure
2. **Change Management Maturity** – Classification of organisational change approach: Informal, Standardised, Formal
3. **Strategic Initiative Status** – Status of ongoing strategic projects: On Track, Delayed, Cancelled
4. **Communication Cadence** – Frequency of executive-level reporting: Weekly, Bi-weekly, Monthly
5. **Stakeholder Confidence** – Investor or board satisfaction rating: High, Medium, Low
6. **Regulatory Compliance Level** – Extent of legal and policy adherence: Full, Partial, Non-compliant
7. **Culture Transformation Stage** – Stage of organisational development: Emerging, Developing, Established

6. Quantitative Variables

1. **Annual Revenue USD** – Total company revenue in the defined period
2. **Annual Profit USD** – Net operational profit under MD leadership
3. **Budget Variance Pct** – Budget deviation expressed as a percentage
4. **Strategic Projects Completed** – Count of major initiatives successfully finalised
5. **Operational Efficiency Pct** – Ratio-based efficiency measure

© 2025 Dr. Yasser Aly Yakout Shehata

6. **Employee Turnover Rate Pct** – Annual percentage of leadership team turnover
7. **Stakeholder Meeting Count** – Number of strategic-level meetings conducted
8. **Crisis Events Logged** – Number of serious operational or compliance disruptions
9. **Innovation Index** – Number of innovations or system/process improvements implemented
10. **Team Size** – Number of executives and leads managed

7. Derived Variables

1. **Profit Margin Category** – Categorisation of profitability (High, Medium, Low)
2. **Revenue Growth Flag** – Boolean indicator of positive revenue change
3. **Strategic Success Index** – Composite metric combining innovation, efficiency, turnover, and project success
4. **High Performer** – Composite classification for entries meeting profitability and strategic benchmarks
5. **Revenue Band** – Grouped revenue tiers for filtering and comparative visualisation
6. **Innovation Level** – Grouped classification: Low, Medium, High innovation activity

8. Relevance to Power BI and AI for Business

This dataset is optimised for deployment in **Power BI dashboards**, enabling:

- Multi-level performance tracking (financial, operational, stakeholder engagement)
- Cross-filtering based on organisational characteristics and strategic maturity
- Drill-down analysis into KPIs by business focus area, compliance status, or innovation levels

For **AI applications**, the dataset serves as a foundation for:

- Predictive modelling of performance categories
- Clustering of leadership profiles or strategic outcomes
- Supervised learning based on profitability, success index, or growth trajectory

© 2025 Dr. Yasser Aly Yakout Shehata

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Analyse strategic success distribution across company types
- Correlate communication cadence with project outcomes or profit margins
- Assess how culture transformation impacts innovation and turnover

B. Power BI Visualisation

- Build executive dashboards featuring key performance trends
- Create segmented views by profit margin, compliance level, or innovation index
- Present stakeholder engagement patterns across organisational maturity levels

C. AI-Driven Business Insights

- Classify high-performing leadership profiles
- Forecast strategic risk based on crisis events and project delays
- Cluster business units with similar financial and innovation profiles

10. Conclusion

The **Corporate Leadership Performance Analytics** dataset offers MBA students a realistic, multidimensional resource for applying advanced business analytics tools and methods. Rooted in the context of Managing Director-level responsibilities within the construction sector, the dataset empowers learners to explore executive decision-making through data. It enables practice in both descriptive and predictive analysis and supports skill development in Power BI, AI modelling, and strategic business storytelling. As such, it is ideally positioned to bridge academic training with executive-level analytics applications, preparing future leaders for data-informed leadership in complex organisational environments.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 8: Projects Human Resources Manager Analytics

1. Introduction

The ***Projects Human Resources Manager Analytics*** dataset has been carefully developed to replicate the core human resources functions and performance indicators that are critical to the **management of workforce operations across construction project environments**. With its contextual relevance to the evolving needs of the construction industry, the dataset specifically mirrors the professional scope and responsibilities of a ***Projects Human Resources Manager***, a pivotal role tasked with overseeing recruitment, onboarding, employee performance management, compliance adherence, engagement initiatives, and training allocation in complex and resource-intensive project scenarios. The dataset was conceived with a strong pedagogical purpose, tailored for **MBA students** undertaking their final projects within the ***Business Analytics for Decision Making*** **course**. It offers a comprehensive and analytically rich landscape that enables learners to simulate real-world HR decision-making using structured data exploration, descriptive and predictive analytics, and dashboard-based visual storytelling.

In many construction organisations, including industry leaders such as **Rowad Modern Engineering**, the human resources function plays a critical role not only in ensuring that labour resources are efficiently mobilised and retained, but also in guaranteeing that projects comply with legal standards, align with organisational culture, and contribute to long-term strategic goals. The **Projects Human Resources Manager** thus acts as a key intermediary between organisational leadership and site-based operational teams. This dataset captures that reality, enabling students to analyse how human resources strategies at the project level can affect broader outcomes such as project cost, compliance quality, team engagement, and turnover rates. Each data record corresponds to an individual construction project, aggregating key HR-related metrics that represent the lifecycle of personnel management throughout the duration of the project.

The dataset consists of 1,000 carefully generated synthetic project records, with each record incorporating a mix of seven qualitative variables, ten quantitative variables, and six derived fields designed to enhance analytical depth. These variables span a range of categories that reflect real business processes, such as hiring channel, onboarding methodology, average time to fill vacancies, training hours per employee, engagement scores, and HR-related costs. The derived fields include efficiency scores, engagement flags, risk indicators, and compliance flags, which are essential for implementing classification models, performance benchmarking, and actionable segmentation strategies.

© 2025 Dr. Yasser Aly Yakout Shehata

Importantly, the design of this dataset supports a wide range of analytical applications. It enables students to perform exploratory data analysis (EDA), develop and interpret descriptive statistics, generate dynamic dashboards in Power BI, and apply AI tools such as clustering and classification for insight generation. In doing so, learners are equipped not only to understand the internal mechanics of HR operations but also to model strategic HR decisions that are rooted in data and aligned with business priorities. In a professional context where HR data is increasingly leveraged for strategic forecasting, talent analytics, and risk assessment, this dataset provides a unique, hands-on learning opportunity.

Ultimately, the ***Projects Human Resources Manager Analytics*** dataset prepares MBA students to act as analytically competent, insight-driven professionals capable of transforming complex HR data into meaningful business outcomes. It fosters an integrated understanding of human capital management, operational efficiency, and analytics-based leadership in the construction sector and beyond.

2. Dataset Description

This dataset comprises 1,000 records representing human resources operations at the project level within a construction firm. Each record aggregates the activities managed by a Projects HR Manager, ranging from recruitment metrics to engagement scores and compliance tracking.

3. Purpose and Application

The dataset was designed for MBA students to:

- Apply data analytics techniques to real-world HR scenarios
- Practise visualising HR KPIs through Power BI dashboards
- Conduct descriptive and exploratory analyses
- Implement AI techniques for classification and forecasting in HR operations

4. Dataset Structure

- **Total records**: 1,000
- **Variables**: 23
    - 7 qualitative (categorical)
    - 10 quantitative (numerical)
    - 6 derived variables for analysis and modelling

© 2025 Dr. Yasser Aly Yakout Shehata

5. Qualitative Variables

1. **Project_Type** – Type of construction project (Residential, Commercial, Infrastructure)
2. **Hiring_Channel** – Recruitment source (Internal Referral, Agency, Online Job Board)
3. **Onboarding_Methodology** – Integration approach for new hires (Standardised, Informal, Structured Programme)
4. **Performance_Management** – Performance review type (Annual Review, Continuous Feedback, None)
5. **Compliance_Status** – HR compliance level (Compliant, Partial, Non-compliant)
6. **Engagement_Programme** – Level of employee engagement (Active, Minimal, None)
7. **HR_System_Level** – HR system use (Full ERP, Partial, Manual)

6. Quantitative Variables

1. **Positions_Opened** – Number of roles advertised
2. **Positions_Filled** – Number of positions successfully filled
3. **Avg_Time_to_Fill_Days** – Average days to fill roles
4. **Avg_Onboarding_Duration_Days** – Onboarding duration in days
5. **Training_Hours_Per_Employee** – Training hours per person
6. **Avg_Performance_Rating** – Mean post-hire performance rating
7. **Turnover_Percentage** – Proportion of staff leaving during the project
8. **Compliance_Issues** – Number of recorded HR compliance breaches
9. **Avg_Engagement_Score** – Employee survey-based engagement level
10. **HR_Cost_USD** – HR-related cost per project

7. Derived Variables

1. **Fill_Rate_Percentage** – Positions filled as a proportion of those opened
2. **Compliance_Flag** – Boolean indicator for zero compliance issues
3. **High_Engagement_Flag** – Boolean indicator for high engagement (≥ 4.0)
4. **HR_Efficiency_Score** – Composite indicator of fill rate, time-to-fill, and performance

© 2025 Dr. Yasser Aly Yakout Shehata

5. **Turnover_Risk_Flag** – Flag if turnover exceeds 20 percent
6. **Training_Intensity_Level** – Categorised training volume (Low, Medium, High)

8. Relevance to Power BI and AI for Business

- **Power BI**: Enables development of interactive dashboards to monitor HR KPIs such as engagement, efficiency, and risk indicators by project characteristics
- **AI for Business**: Supports implementation of machine learning models to classify high-turnover projects, predict compliance outcomes, and segment projects based on training or onboarding performance

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify key patterns in compliance, training, and engagement
- Detect anomalies in HR cost or turnover
- Establish relationships between hiring channel and time to fill

B. Power BI Visualisation

- Build executive dashboards using engagement and performance data
- Visualise trends across project types or HR system levels
- Compare cost-efficiency of various onboarding approaches

C. AI-Driven Business Insights

- Train classification models on turnover risk or compliance flag
- Predict performance rating based on training and onboarding duration
- Cluster projects by HR Efficiency Score and engagement level

10. Conclusion

The ***Projects Human Resources Manager Analytics*** dataset delivers a robust, multidimensional platform for applying business analytics in the context of construction project HR management. It enables MBA students to develop and refine skills in EDA, dashboarding, and AI modelling. By closely simulating the professional realities of a Projects HR Manager, the dataset allows students to produce insight-driven strategies, demonstrate evidence-based decision-making, and align academic learning with industry challenges in human capital optimisation.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 9: Construction Project Performance Analytics (CPPA-1000)

1. Introduction

The **Construction Project Performance Analytics Dataset (CPPA-1000)** has been purposefully designed as a simulation-based analytical resource for graduate-level business students, particularly those enrolled in the **MBA course *Business Analytics for Decision Making***. This dataset replicates real-world construction project dynamics and is inspired by the operational environment of large engineering and construction firms, including examples like **Hassan Allam Construction**. It focuses specifically on the critical and often complex role of the **Technical Office Team Leader**, who acts as a strategic intermediary between design teams, procurement officers, on-site engineers, and executive management.

In the construction industry, the **Technical Office Team Leader** is responsible for a wide spectrum of tasks ranging from structural design reviews and drawing coordination to cost estimation, procurement planning, and schedule control. These professionals face ongoing challenges such as managing RFIs (Requests for Information), handling approval delays, minimizing design errors, and navigating complex mechanical, electrical, and plumbing (MEP) integrations. The performance of the Technical Office Team directly impacts project timelines, cost outcomes, and operational risks, making it a rich area for analytical inquiry.

This dataset encapsulates these challenges by simulating 1,000 real-world-like project records. Each record reflects project-level details across variables such as estimated and actual costs, procurement delays, shop drawing timelines, team size, communication efficiency, design complexity, and approval bottlenecks. The data includes both categorical descriptors and measurable quantitative indicators, allowing for multi-dimensional analysis.

What sets **CPPA-1000** apart is the inclusion of derived variables such as *Cost Performance*, *Risk Score*, *Risk Category*, and *Efficiency Score*. These indicators are engineered to mirror how advanced analytics and AI systems derive insight from raw operational data, helping students move from data reporting to data-driven strategy.

By working with this dataset, **MBA students** and **analytics professionals** are invited to perform Exploratory Data Analysis (EDA), design interactive Power BI dashboards, and apply AI-driven techniques such as regression modeling, classification, and clustering. The CPPA-1000 dataset is also ideally structured for storytelling through visualization, decision simulation, and process improvement modeling, making it a powerful tool for both academic development and professional readiness.

© 2025 Dr. Yasser Aly Yakout Shehata

In essence, this dataset serves as a bridge between academic theory and industry practice. It not only allows learners to apply business analytics techniques to a realistic and complex domain but also prepares them to engage in cross-functional decision-making processes where technical insight and business strategy intersect. Through structured data analysis, students can explore how strategic coordination, communication effectiveness, and technical decisions converge to influence construction project success.

2. Dataset Description

The dataset comprises **1,000 records**, each representing a distinct construction project overseen by a Technical Office Team Leader. It contains **7 qualitative variables**, **10 quantitative metrics**, and **4 derived variables** that offer a realistic framework for simulating key decisions in construction planning, cost control, and project execution.

All data is synthetically generated and structured to preserve internal consistency and realistic statistical patterns, making it suitable for hands-on analytical exercises without privacy concerns.

3. Purpose and Application

This dataset is intended to:

- Provide a practical learning resource for MBA students to practice business analytics techniques.
- Simulate decision-making scenarios for roles in project management and technical oversight.
- Support analysis and visual storytelling through Power BI dashboards.
- Enable the application of AI methods to predict cost overruns, categorize project risk, and optimize communication and resource planning.

4. Dataset Structure

The dataset is organized into three categories of variables:

- **Qualitative Variables**: Categorical attributes describing project context and team dynamics.
- **Quantitative Variables**: Measurable factors related to costs, delays, errors, and workload.
- **Derived Variables**: Additional indicators generated from original data to enable deeper insights.

© 2025 Dr. Yasser Aly Yakout Shehata

5. Qualitative Variables

1. **Project_Type** – Type of construction project: Hotel, Residential, Industrial, Mosque.
2. **Structural_System** – Type of structure used: Framed, Shear Wall, Composite.
3. **Foundation_Type** – Foundation type: Pile, Raft, Shallow.
4. **Architectural_Complexity** – Rated as Low, Medium, or High based on design sophistication.
5. **MEP_Integration_Level** – Integration level of mechanical, electrical, and plumbing systems: Low, Medium, High.
6. **Team_Communication_Efficiency** – Rating (1 to 5) indicating how effectively the technical team communicates.
7. **Approval_Delay_Category** – Time delay in obtaining approvals: None, Minor, Major.

6. Quantitative Variables

8. **Team_Size** – Number of engineers and coordinators under the team leader's supervision.
9. **Estimated_Cost_EUR** – Budgeted cost of the project in euros.
10. **Actual_Cost_EUR** – Final actual cost incurred in euros.
11. **Cost_Overrun_Pct** – Percentage difference between estimated and actual cost.
12. **Num_RFIs** – Number of Requests for Information submitted during the project.
13. **Num_Change_Orders** – Number of formal changes requested after design finalization.
14. **Design_Error_Count** – Number of significant design mismatches or inconsistencies.
15. **Shop_Drawing_Delay_Days** – Total days delayed in submitting shop drawings.
16. **Procurement_Delay_Days** – Delay in procurement and delivery of critical materials.
17. **Days_Planning_Deviation** – Cumulative deviation from the original planned schedule.

7. Derived Variables

18. **Cost_Performance** – Categorized as Under Budget, On Budget, or Over Budget.

© 2025 Dr. Yasser Aly Yakout Shehata

19. **Risk_Score** – Composite risk indicator calculated from cost overruns, errors, and delays.
20. **Risk_Category** – Categorization of risk score as Low, Medium, or High.
21. **Efficiency_Score** – Efficiency metric derived from communication score and cost performance.

8. Relevance to Power BI and AI for Business

This dataset supports advanced business reporting and intelligent decision support using Power BI. Its categorical and numerical variables are designed to:

- Enable dynamic filtering, slicing, and drill-through in dashboards.
- Provide meaningful KPIs and trend analyses relevant to project planning and cost management.
- Serve as inputs for machine learning models and AI algorithms to extract actionable business intelligence.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

Students can explore the relationships between architectural complexity and design errors, the impact of communication efficiency on delays, and correlations between procurement delays and cost overruns.

B. Power BI Visualization

The dataset can be visualized through:

- Risk heatmaps by project type and system
- Cost deviation dashboards
- Delay tracking by approval and MEP integration levels
- Efficiency and risk segmentation reports

C. AI-Driven Business Insights

Using AI and statistical models, students can:

- Build predictive models for cost overrun percentages
- Classify projects into risk categories using decision trees

© 2025 Dr. Yasser Aly Yakout Shehata

- Optimize team communication strategies to reduce schedule deviation
- Use clustering to profile projects by risk and performance attributes

10. Conclusion

The **CPPA-1000** dataset bridges academic learning and professional application. It enables MBA students to understand, visualize, and model real-world project data using modern analytical tools. Through EDA, dashboards, and AI insights, students develop the skills to interpret and influence data-driven decisions. With a specific focus on the Technical Office Team Leader, the dataset prepares students to take on roles that demand both technical knowledge and strategic thinking in the evolving landscape of construction and infrastructure management.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 10: Ezz Steel Sales Analytics Dataset

1. Introduction

The **Ezz Steel Sales Analytics Dataset** is a comprehensive and realistic simulation designed for **MBA students specializing in Business Analytics**, particularly those undertaking their final project in the course **"Business Analytics for Decision Making."** This dataset places students in the role of a **Sales Project Analyst** tasked with evaluating and presenting data-driven insights on steel sales operations across domestic and international markets during the years 2022 and 2023.

Ezz Steel, one of the leading steel producers in the Middle East and North Africa region, operates in a complex environment influenced by macroeconomic shifts, currency fluctuations, pricing volatility, and export market dynamics. The dataset captures this complexity by presenting diverse variables tied to quarterly performance, manufacturing sites, currency regimes, and sales channels.

Students engaging with this dataset will simulate the analytical workflow of a business professional responsible for generating executive-level insights. This includes calculating profitability margins, segmenting markets, modeling financial risk, and forecasting future trends. With both raw and derived variables, students are equipped to explore deep relationships among operational and financial metrics.

The dataset is also purpose-built for integration with Power BI and AI applications. Students can build interactive dashboards, conduct exploratory data analysis, and apply regression, classification, and clustering algorithms to extract strategic recommendations. This hands-on approach bridges business theory with practical analytics skills, empowering students to transform raw data into compelling stories that support executive decision-making.

2. Dataset Description

The dataset consists of 1,000 synthetic records, each representing a unique sales period transaction defined by a combination of quarter, product type, sales channel, market region, and plant. It includes:

- 7 qualitative variables
- 10 quantitative variables
- 4 derived variables

All values are synthetically generated based on trends reported in Ezz Steel's 2022 and 2023 investor reports, ensuring realism without the use of any private or confidential company data.

© 2025 Dr. Yasser Aly Yakout Shehata

3. Purpose and Application

This dataset is intended to:

- Serve as a hands-on resource for students to practice EDA, visualization, and AI modeling.
- Demonstrate how analytics can enhance sales performance, margin visibility, and pricing strategy.
- Provide a foundation for applying descriptive statistics, dashboard design, and predictive modeling.
- Support data storytelling and business communication through charts, filters, KPIs, and simulation tools.

4. Dataset Structure

The dataset is divided into three types of variables:

- **Qualitative Variables**: Used for grouping, filtering, and segmentation
- **Quantitative Variables**: Used for measurement, trend analysis, and modeling
- **Derived Variables**: Calculated fields to enrich insights and support decision-making frameworks

5. Qualitative Variables

1. **Year**: Calendar year of the sales record (2022 or 2023)
2. **Quarter**: Fiscal quarter in which the sale occurred (Q1 to Q4)
3. **Product Type**: Type of steel product sold (Rebar or Coil)
4. **Market Region**: Indicates whether the sale was Domestic or Export
5. **Plant**: Manufacturing plant responsible for production (Alexandria, Suez, Sadat City, 10th of Ramadan)
6. **Sales Channel**: Type of sales relationship (Direct or Distributor)
7. **Currency Regime**: Indicates monetary policy (Fixed or Floating) affecting FX exposure

6. Quantitative Variables

8. **Sales Volume (tonnes)**: Quantity of product sold in metric tonnes

© 2025 Dr. Yasser Aly Yakout Shehata

9. **Unit Price (EGP/tonne)**: Price per tonne applied to the product type
10. **Sales Value (EGP)**: Total revenue generated from the transaction
11. **Cost of Goods Sold (EGP)**: Manufacturing cost of the sold items
12. **Gross Profit (EGP)**: Revenue minus cost of goods sold
13. **EBITDA (EGP)**: Earnings before interest, tax, depreciation, and amortization
14. **FX Loss (EGP)**: Losses due to currency devaluation or exchange differences
15. **Net Profit (EGP)**: Net earnings after FX losses
16. **Export Share (%)**: Proportion of the sale tied to exports
17. **Sales Growth (%)**: Growth rate compared to the previous period

7. Derived Variables

18. **Gross Margin (%)**: Gross Profit divided by Sales Value
19. **EBITDA Margin (%)**: EBITDA divided by Sales Value
20. **FX Impact (%)**: FX Loss divided by Sales Value
21. **Profit Category**: Categorized into "Losing," "Breakeven," or "Profitable" based on Net Profit

8. Relevance to Power BI and AI for Business

The Ezz Steel Sales Analytics Dataset is highly compatible with Power BI for building interactive dashboards, trend visualizations, and executive summaries. Categorical variables allow for dynamic filtering, while derived financial metrics provide insightful key performance indicators. AI models can also be trained on this data to forecast performance and recommend business actions.

Students will be able to:

- Build charts that compare profit margins across products, plants, and regions
- Slice time-series data by quarter and currency regime
- Design drill-down views to explore factors influencing profitability
- Train predictive models to estimate growth or classify risk segments

© 2025 Dr. Yasser Aly Yakout Shehata

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Compare product-level performance across years and plants
- Explore the relationship between pricing and profit margins
- Analyze FX losses in relation to market region and currency regime

B. Power BI Visualization

- Design dashboards with KPIs like Gross Margin and Net Profit
- Use slicers for Product Type, Region, Quarter, and Profit Category
- Build charts showing YoY sales value trends and growth patterns

C. AI-Driven Business Insights

- Build regression models to forecast sales growth or margin improvement
- Classify profit risk segments using decision trees or clustering
- Simulate the impact of FX losses on profitability across scenarios

10. Conclusion

The **Ezz Steel Sales Analytics Dataset** offers a practical and data-rich simulation for developing real-world business analytics skills. With its combination of transactional, financial, and market variables, the dataset empowers MBA students to apply analytical tools in context, develop insight-driven dashboards, and use AI to generate strategic recommendations. It serves as a strong foundation for practicing the integration of business knowledge and data science to support smart, evidence-based decision-making.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 11: Executive Decision Analytics for Construction Projects

1. Introduction

In today's data-driven business environment, **senior executive decisions** have a profound impact on organisational outcomes, especially within high-stakes, high-complexity industries such as construction. The role of the **Chief Executive Officer (CEO)** and Founder is not only central to strategic planning but also critical to day-to-day operational success, stakeholder engagement, and project delivery. This dataset, titled ***Executive Decision Analytics for Construction Projects (EDACP)***, has been meticulously designed to provide MBA students with a real-world-inspired foundation for applying business analytics, with a specific focus on leadership behaviours, decision-making patterns, and their measurable impact on construction project performance.

The **EDACP** dataset was created to support MBA final projects within the **course *Business Analytics for Decision Making***. It enables students to explore how variations in executive decisions such as meeting frequency, communication style, risk appetite, and negotiation strategy, translate into quantifiable business outcomes like budget adherence, on-time delivery, quality compliance, and financial claims. Each record in the dataset simulates a distinct construction project led by a **CEO/Founder**, embedding a wide range of realistic leadership, operational, and project-level variables. These include both objective metrics (e.g., project budget, duration, and variance) and subjective traits (e.g., leadership style, stakeholder engagement), providing a multi-dimensional view of executive impact.

The **EDACP** dataset serves as an ideal vehicle for developing key competencies in exploratory data analysis (EDA), data visualisation, and applied artificial intelligence (AI) for business. With 1,000 synthetically generated records and 17 well-defined variables spanning quantitative, qualitative, and derived categories, it is structured to facilitate deep analytical inquiry. Students can apply statistical techniques to uncover patterns, build dashboards in Power BI to visualise performance trends, and implement predictive models to simulate decision outcomes under varying leadership conditions.

The dataset also integrates derived fields that add analytical depth, such as *Decision Efficiency*, *Cost Efficiency*, and *On-Time Completion*. These composite indicators link executive behaviour to project success in meaningful ways, encouraging students to move beyond surface-level observations and into more strategic interpretations of data. For instance, students can explore how a participative leadership style correlates with project timeliness, or how communication skills and stakeholder engagement influence budget performance.

© 2025 Dr. Yasser Aly Yakout Shehata

Ultimately, the **EDACP dataset** bridges the gap between business theory and data-informed practice. It contextualises the analytical process within a realistic executive environment, reflecting the complexity of decisions made at the highest levels of leadership. By engaging with this dataset, MBA students are not only able to practice technical analytics skills but also to derive actionable business insights that mirror the responsibilities and challenges faced by real-world CEOs. This makes EDACP not only a pedagogical tool, but a practical model for preparing future business leaders to thrive in data-centric roles.

2. Dataset Description

The ***Executive Decision Analytics for Construction Projects (EDACP)* dataset** simulates 1,000 construction projects managed by a CEO/Founder. Each record captures leadership decisions, operational dynamics, and project outcomes using structured quantitative and qualitative attributes. It includes a total of 17 variables:

- 10 Quantitative Variables
- 7 Qualitative Variables
- 3 Derived Variables

The data format is suitable for Excel, Power BI, R, Python, and all major analytics platforms.

3. Purpose and Application

This dataset was developed to serve the following objectives:

- Facilitate hands-on experience in descriptive and diagnostic analytics
- Enable dashboard development in Power BI for executive insight delivery
- Provide structured inputs for AI models in business decision-making
- Encourage critical thinking on how leadership attributes influence performance
- Support predictive and prescriptive analysis to simulate business outcomes

4. Dataset Structure

- **Number of Records**: 1,000
- **Data Categories**: Quantitative, Qualitative, Derived
- **Format**: UTF-8 CSV
- **Context**: Construction industry projects led by a CEO/Founder
- **Use Case**: MBA-level analytics coursework and executive decision modelling

© 2025 Dr. Yasser Aly Yakout Shehata

5. Qualitative Variables

1. **Leadership Style** – Coaching, Autocratic, or Participative
2. **Communication Skills** – Low, Medium, or High
3. **Time Management** – Poor, Average, or Excellent
4. **Negotiation Strategy** – Win-win, Firm, or Flexible
5. **Risk Appetite** – Conservative, Balanced, or Aggressive
6. **Stakeholder Engagement** – Low, Medium, or High
7. **Decision-Making Style** – Analytical, Intuitive, or Collaborative

6. Quantitative Variables

1. **Weekly Meetings** – Number of executive meetings held per week
2. **Decision Latency (hours)** – Time to resolve critical issues
3. **Project Budget (thousands USD)** – Estimated cost of the project
4. **Budget Variance (percent)** – Over or under-spend relative to budget
5. **Project Duration (weeks)** – Actual time to complete the project
6. **Completion Delay (weeks)** – Delay beyond scheduled timeline
7. **Supplier Coordination Count** – Engagements with suppliers and partners
8. **Legal or Licensing Events** – Number of regulatory actions taken
9. **Quality Incidents** – Defects or compliance failures
10. **Financial Claim Value (thousands USD)** – Cost of claims issued

7. Derived Variables

1. **Decision Efficiency** – Weekly Meetings divided by Decision Latency
2. **Cost Efficient** – Binary flag: project under budget and zero quality incidents
3. **On-Time Completion** – Binary flag: project delivered without delay

8. Relevance to Power BI and AI for Business

The EDACP dataset is fully compatible with Power BI for developing dynamic dashboards. Students can create visualisations that filter and segment data based on leadership behaviour, risk profile, or project efficiency. Through DAX formulas, trend analysis, and KPI

© 2025 Dr. Yasser Aly Yakout Shehata

visualisation, this dataset encourages interpretation of multi-dimensional business dynamics.

For AI applications, the dataset supports machine learning workflows such as classification (e.g. predicting cost efficiency) and regression (e.g. estimating project delays). With categorical and continuous features clearly defined, it provides an ideal training ground for AI-driven business insight generation.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Use of descriptive statistics to summarise project performance
- Identification of correlations between executive behaviour and delays or overspending
- Segmentation by leadership attributes or decision-making style

B. Power BI Visualisation

- KPI tracking dashboards for Decision Efficiency and Budget Performance
- Cross-sectional comparisons using filters for qualitative variables
- Visual storytelling through bar charts, scatter plots, and matrix views

C. AI-Driven Business Insights

- Predictive models to identify high-risk projects based on executive variables
- Classification models for on-time or cost-efficient delivery
- Prescriptive analytics for strategy optimisation based on scenario analysis

10. Conclusion

The ***Executive Decision Analytics for Construction Projects (EDACP)*** dataset provides a realistic, structured, and analytically rich platform for business analytics education. It allows MBA students to engage in meaningful decision modelling from the perspective of a CEO, using data to drive insight. By combining EDA, dashboarding, and AI, the dataset fosters a deep understanding of how leadership decisions influence project-level outcomes. As such, EDACP is more than a teaching tool, it is a bridge between the analytical competencies taught in business schools and the real-world decisions awaiting future executives.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 12: Decision Analytics for Authority Urban Planning and Infrastructure

Introduction

The 21st century has witnessed an unprecedented wave of urbanisation, prompting national and municipal authorities to reimagine how cities are planned, managed, and governed. Within this transformative context, the rise of smart cities like New Alamein in Egypt symbolises a national commitment to future-ready urban development driven by technology, innovation, and strategic planning. Managing such complex and large-scale urban projects necessitates leadership that is both operationally competent and analytically empowered. The role of the Deputy of the New Alamein Authority Head embodies this demand, as it requires cross-functional oversight of construction, infrastructure, and administrative functions while ensuring alignment with broader national development goals.

To support the educational objectives of business students pursuing an MBA with a focus on analytics and decision-making, this project introduces a synthetic yet contextually realistic dataset titled **DAAUPI – Decision Analytics for Authority Urban Planning and Infrastructure**. The dataset has been designed to reflect the multifaceted operational environment encountered by senior urban planners and executives. It enables students to explore the use of business analytics tools in assessing project efficiency, cost control, risk mitigation, stakeholder engagement, and interdepartmental coordination.

The dataset provides one thousand records simulating project-specific operations and decision variables extracted from the working model of a Deputy Authority Head in an urban planning context. Each record encapsulates both qualitative and quantitative data that represent strategic choices and operational outcomes. These include coordination styles, contract strategies, risk management approaches, and decision-making patterns, as well as project costs, delays, readiness scores, and area measurements. In addition, the dataset includes several derived variables, such as cost and time efficiency, to provide enhanced analytical depth and realism.

This dataset is particularly suitable for students enrolled in the course **"Business Analytics for Decision Making"**, as it allows them to apply core analytical methods in a high-stakes real-world simulation. It serves as a training tool for Exploratory Data Analysis (EDA), supports the development of interactive dashboards in Power BI, and provides fertile ground for the implementation of AI-driven models aimed at improving business and policy outcomes. The dataset's structure and logic reflect best practices in public infrastructure project management and urban governance, enabling learners to integrate data science with leadership scenarios.

© 2025 Dr. Yasser Aly Yakout Shehata

By working with the DAAUPI dataset, students are invited not merely to manipulate data, but to engage with the strategic imperatives that drive smart city development and to evaluate the quality of decision-making in complex governmental contexts. The dataset thereby offers a unique pedagogical value, serving both academic goals and professional development within the fields of business analytics, urban planning, and public administration.

Dataset Description

The DAAUPI dataset contains one thousand records representing zone-level urban infrastructure planning and coordination tasks. Each record corresponds to a specific infrastructure initiative under the supervision of the Deputy of the New Alamein Authority Head. The dataset includes a mix of qualitative variables (capturing strategic intent and managerial preferences), quantitative variables (capturing performance indicators and cost metrics), and derived variables (enhancing analytical insights through calculated relationships).

All records are synthetically generated based on realistic operational parameters and follow logical patterns consistent with real-world urban governance practices.

Purpose and Application

This dataset has been developed to:

- Support MBA students in applying business analytics tools to public sector scenarios.
- Facilitate hands-on practice in data preparation, modelling, and dashboard design.
- Enable EDA, statistical analysis, and performance assessment in decision-making contexts.
- Serve as the foundation for predictive modelling, clustering, and segmentation tasks using AI.

It can be used in class projects, independent research, or decision support simulations for smart city case studies.

Dataset Structure

The dataset comprises:

- **Seven Qualitative Variables**
- **Ten Quantitative Variables**
- **Three Derived Variables**

© 2025 Dr. Yasser Aly Yakout Shehata

Each record reflects a full operational profile of one zone-level project cycle, including both inputs and outcomes.

Qualitative Variables

1. **Coordination Style** – Coordination philosophy: Centralised, Collaborative, or Delegative.
2. **Report Quality** – Quality standard of project reports: Poor, Adequate, or Excellent.
3. **Contract Strategy** – Budget control strategy: Conservative, Balanced, or Aggressive.
4. **Risk Approach** – Method of addressing project risks: Reactive, Proactive, or Adaptive.
5. **Stakeholder Engagement** – Degree of stakeholder participation: Low, Moderate, or High.
6. **Decision Style** – Decision-making method: Analytical, Intuitive, or Consultative.
7. **Zone Type** – Functional classification of the zone: Residential, Commercial, or Infrastructure.

Quantitative Variables

8. **Coordination Meetings** – Number of coordination meetings held.
9. **Report Time (days)** – Duration required for preparing the project report.
10. **Contract Estimate (Million EGP)** – Estimated financial cost of the zone project.
11. **Cost Deviation (percentage)** – Cost variation compared to the initial estimate.
12. **Zone Area (hectares)** – Total surface area of the project zone in hectares.
13. **Zone Area (m^2)** – Converted area into square meters.
14. **Zone Area (km^2)** – Converted area into square kilometers.
15. **Infrastructure Delay (weeks)** – Number of weeks by which the project was delayed.
16. **Complaints** – Total number of complaints received regarding the project.
17. **Permits Processed** – Number of processed permits during implementation.
18. **Readiness Score** – Evaluation score of infrastructural and operational readiness.
19. **Coordination Complexity** – A scale measuring complexity of coordination across entities.

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Variables

20. **Cost Efficiency** – Indicates whether the project stayed within budget and scored highly on readiness.
21. **Time Efficiency** – Indicates whether the project avoided delays while achieving high readiness.
22. **Coordination Efficiency** – Ratio of coordination meetings to complexity, measuring management efficiency.

Relevance to Power BI and AI for Business

This dataset is ideal for:

- **Power BI Dashboarding**: Visualising relationships between managerial variables and project performance.
- **Data Segmentation**: Categorising projects by strategy or efficiency outcomes.
- **Predictive Modelling**: Using machine learning to forecast readiness or complaints based on early decisions.

Students can construct interactive dashboards that provide city-wide insights into zone performance, simulate reporting tools used by government leaders, and communicate analytical results in strategic presentations.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Investigate the impact of contract strategy on cost deviation.
- Identify zones with consistent high readiness despite complexity.
- Analyse stakeholder engagement against complaints received.

B. Power BI Visualisation

- Compare coordination styles by region, risk approach, or decision outcomes.
- Plot readiness scores across zone types and time frames.
- Build interactive maps integrating spatial data with infrastructure KPIs.

C. AI-Driven Business Insights

- Train classifiers to predict project efficiency.

© 2025 Dr. Yasser Aly Yakout Shehata

- Use clustering algorithms to identify patterns in underperforming zones.
- Integrate cost and time variables into predictive dashboards for early risk flagging.

Conclusion

The **DAAUPI** dataset is a strategically designed educational tool that places MBA students at the heart of public decision-making within a smart city framework. Through the lens of a senior authority role, students gain exposure to the intricate interplay between planning strategies, performance measurement, and leadership decisions. It enables the practice of essential analytical skills such as EDA, visualisation, and predictive modelling while fostering deeper understanding of urban governance and digital transformation.

The dataset is not only a technical resource but also a conceptual bridge connecting classroom knowledge with real-world public administration, empowering future leaders to drive data-informed policy and infrastructure development.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 12: Construction Quality Control Analytics

Introduction

The **Construction Quality Control Analytics Dataset for MBA Business Decision-Making** is a purpose-built synthetic dataset designed to support high-impact learning and practical engagement in the field of Business Analytics, with specific relevance to operational quality control within the construction industry. This dataset was developed with a clear pedagogical vision in mind: to simulate the data environment and analytical challenges faced by a **Quality Control Section Head**, whose role is central to ensuring compliance, managing risk, and sustaining quality standards across complex, multi-site construction projects.

In modern construction environments, quality control is no longer a purely procedural function. It has evolved into a strategic pillar, closely tied to project performance, stakeholder satisfaction, and organisational reputation. A Quality Control Section Head must therefore not only execute inspections and issue certifications but also continuously monitor process efficiency, analyse failure trends, assess contractor performance, and ensure that corrective actions are timely and cost-effective. This multidimensional role requires mastery of data interpretation, visualisation, and evidence-based decision-making, skills that are equally critical for today's MBA students preparing for leadership in data-intensive sectors.

This dataset reflects the real-world complexity and structure of quality control activities, capturing detailed records of inspection events, associated documentation, and performance metrics. The design process incorporates domain-specific logic, such as higher rework costs following failed inspections, differences in inspection duration based on type, and variations in on-time submission rates among contractors. Additionally, derived variables such as **Performance Score** and **Site Pass Rate** offer aggregate indicators that simulate the kinds of composite metrics typically used by managers for benchmarking and performance evaluation.

The dataset was constructed not merely as a static collection of values, but as a dynamic tool for business exploration. It invites students to use **Exploratory Data Analysis (EDA)** to uncover hidden trends, deploy **Power BI dashboards** to visualise quality performance, and apply **AI-driven models** to predict risk and optimise outcomes. Its structure facilitates both descriptive and inferential analysis, offering practical opportunities for regression, classification, clustering, and time-based visual insights. This makes it particularly valuable for MBA students enrolled in courses such as ***Business Analytics for Decision Making***,

© 2025 Dr. Yasser Aly Yakout Shehata

where the ability to translate data into actionable business strategies is a core learning outcome.

In essence, the **Construction Quality Control Analytics Dataset** bridges the gap between classroom theory and professional application. It enables students to develop analytical fluency, experiment with business intelligence tools, and explore how AI can transform traditional functions like quality control into strategic assets for organisational competitiveness. Through structured practice with this dataset, MBA students are empowered to think critically, visualise operational data, and generate insights that mirror the expectations of real-world decision-makers in today's data-driven economy.

Dataset Description

This dataset contains **1,000 inspection records** performed across multiple construction sites. Each record represents an individual inspection event managed by a Quality Control Section Head. The dataset comprises **seven qualitative variables**, **ten quantitative variables**, and **two derived variables**. These elements capture multi-dimensional aspects of quality assurance operations, including inspection activities, contractor performance, compliance tracking, material approvals, and documentation timelines.

Purpose and Application

The dataset has been designed with dual academic and practical intent:

- **For MBA students**: to develop competence in data analysis, dashboard creation, insight generation, and AI-enabled decision support systems.
- **For business simulation**: to explore how a Quality Control Section Head monitors quality metrics, manages performance, and implements continuous improvement strategies using analytical tools.

It serves as a foundational asset for assignments, case studies, or capstone projects related to operational analytics, quality control, and digital transformation in construction.

Dataset Structure

The dataset is structured into three primary components:

- **Qualitative (Categorical) Variables** – for classification, segmentation, and filter-based analysis.
- **Quantitative (Numerical) Variables** – for aggregation, correlation, and performance tracking.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Derived Variables** – composite indicators enabling advanced performance measurement and benchmarking.

Each row corresponds to a unique inspection event, and the dataset spans across multiple sites, contractors, and inspection types.

Qualitative Variables

1. **SiteLocation** – Name of the project site (e.g. Project_1 to Project_10).
2. **ContractorID** – Unique identifier for subcontractors (e.g. C001 to C020).
3. **InspectionType** – Type of inspection: Initial, Follow-up, or Certificate Issuance.
4. **PassFailOutcome** – Outcome of the inspection: Pass or Fail.
5. **NCRStatus** – Status of the non-conformance report: Open, Closed, or Pending.
6. **MaterialType** – Material inspected: Concrete, Steel, Finishings, or Electrical.
7. **CertificateType** – Type of certificate issued: Material Approval, Acceptance Certificate, Test Certificate.

Quantitative Variables

1. **InspectionsConducted** – Number of inspections completed in the instance.
2. **InspectionDuration_h** – Duration of the inspection in hours.
3. **NumberOfNCRs** – Count of non-conformance reports raised.
4. **ReworkCost_USD** – Estimated cost for rework due to failed inspections.
5. **DaysToCloseNCR** – Days required to resolve and close a non-conformance.
6. **MaterialsApproved** – Quantity of materials that passed inspection.
7. **CertificatesIssued** – Number of documentation approvals provided.
8. **CostPerTest_USD** – Cost per inspection/test for the material.
9. **InspectionCoverage_pct** – Percentage of scheduled inspections completed.
10. **OnTimeCertificate_pct** – Percentage of certificates submitted within deadline.

Derived Variables

1. **PerformanceScore** – Composite score (0–100) calculated based on inspection coverage, NCR rate, timeliness, duration, and approval volume.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **SitePassRate** – Aggregated pass rate per site, indicating quality compliance performance.

These derived fields enhance the dataset's suitability for comparative benchmarking, AI model training, and quality performance evaluation.

Relevance to Power BI and AI for Business

The dataset has been intentionally structured for seamless integration into **Microsoft Power BI**, enabling interactive dashboard development, real-time quality tracking, and multi-layered performance analysis. It also provides a solid foundation for the application of **AI techniques in business**, such as pattern recognition, risk scoring, predictive modelling, and decision automation.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Assess distribution of inspection outcomes across sites and contractors.
- Identify factors contributing to high rework costs.
- Examine correlations between material types and inspection durations.

B. Power BI Visualization

- Develop dashboards tracking inspection KPIs by project, contractor, or material.
- Visualise NCR trends, cost dynamics, and on-time documentation rates.
- Implement slicers and drill-downs for site-specific quality reports.

C. AI-Driven Business Insights

- Predict likelihood of inspection failure using classification models.
- Cluster contractors by performance indicators to inform vendor management.
- Detect early warning signs of quality deterioration using trend analytics.

Conclusion

The **Construction Quality Control Analytics** serves as a pedagogical and practical instrument for training future business leaders in data-driven decision-making. Specifically contextualised around the role of a Quality Control Section Head, it equips MBA students with the analytical skills necessary to tackle real-world quality challenges through Power BI, statistical reasoning, and AI-enabled methodologies. Its comprehensive structure, logical

© 2025 Dr. Yasser Aly Yakout Shehata

integrity, and business relevance make it a valuable asset in developing critical thinking, operational insight, and strategic foresight within the context of quality management and digital transformation in construction.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 13: Integrated Electrical Inspection and Compliance Analytics (IEICAD)

1. Introduction

The ***Integrated Electrical Inspection and Compliance Analytics Dataset (IEICAD)*** has been carefully designed as a high-quality, synthetic dataset to serve the educational and practical needs of **MBA students** pursuing advanced coursework in **Business Analytics**. Developed for the final project in the **course *Business Analytics for Decision Making***, this dataset represents a realistic operational scenario in which a **Construction Manager – Electrical** is responsible for ensuring the quality, safety, and regulatory compliance of electrical installations in large-scale construction projects. With the increasing emphasis on data-driven decision-making in the engineering and infrastructure sectors, this dataset provides an applied framework through which students can examine how complex project environments are monitored, evaluated, and improved using analytical methods.

Electrical systems are among the most critical and risk-sensitive components of infrastructure projects. Faults in design, installation, or materials can lead to operational failures, safety violations, financial losses, and delays in project completion. The role of the **Construction Manager – Electrical** involves navigating a complex web of inspection schedules, vendor relationships, technical quality assurance, and safety protocols. In this context, timely and accurate decision-making requires the ability to analyse large volumes of operational data, assess performance indicators, and predict areas of risk or inefficiency. The **IEICAD dataset** simulates this environment by offering 1,000 detailed inspection records, enriched with both raw and derived metrics that mirror those used in real-world construction monitoring systems.

Each inspection record in the dataset includes information on the site, subcontractor, material type, and the outcome of the inspection. Additionally, the dataset contains precise measurements such as voltage deviation, power factor, and total harmonic distortion, which are key indicators of electrical quality. These are complemented by performance-related metrics such as the number of non-conformance reports (NCRs), safety incidents, and delays attributable to vendors. Beyond these core variables, the dataset integrates carefully constructed derived variables, including a Performance Score, an Inspection Risk Score, a Site Pass Rate, an Efficiency Index, and a Risk Category. These composite indicators are designed to reflect the underlying logic used by construction managers to evaluate and benchmark operational effectiveness.

The richness of the **IEICAD dataset** lies not only in its technical depth but also in its analytical versatility. It is suitable for a broad spectrum of academic exercises and

© 2025 Dr. Yasser Aly Yakout Shehata

professional applications, including Exploratory Data Analysis (EDA), visualisation with Microsoft Power BI, and the use of Artificial Intelligence tools to build predictive models and simulate decision-making scenarios. Students can employ this dataset to investigate patterns in inspection results, identify correlations between vendor delays and rework costs, assess site performance, and forecast risk levels using machine learning algorithms. This hands-on exposure to data from a technically rigorous field prepares MBA students to translate business analytics knowledge into actionable insights in real-world industrial contexts.

By working with the IEICAD dataset, students not only deepen their understanding of data analytics but also develop critical thinking skills, problem-solving abilities, and a systems-oriented mindset. These capabilities are essential for future professionals who will lead multidisciplinary teams, manage complex projects, and implement data-centric strategies in industries such as construction, manufacturing, energy, and infrastructure development. Ultimately, this dataset stands as a comprehensive academic resource that bridges the technical demands of engineering management with the strategic imperatives of modern business analytics.

2. Dataset Description

IEICAD comprises **1,000 records**, each representing an electrical inspection event within a construction project. These records span multiple sites, contractors, and materials, and are enriched with key performance, safety, and compliance indicators. Each record combines operational variables with derived metrics, making the dataset versatile for both descriptive and predictive analysis.

Data is categorised into **qualitative variables** (categorical attributes for grouping and segmentation), **quantitative variables** (numeric measures of performance and risk), and **derived variables** (composite indicators created through embedded business logic).

3. Purpose and Application

This dataset is intended to serve as a realistic and educational tool for MBA students and practitioners who aim to:

- Explore the relationships between electrical quality, safety, and project outcomes.
- Build visual dashboards that support project monitoring and strategic reporting.
- Apply AI and statistical methods for classification, regression, and risk prediction.
- Engage in scenario modelling for decision support in construction project environments.

© 2025 Dr. Yasser Aly Yakout Shehata

- Translate engineering and technical data into business-relevant insights.

4. Dataset Structure

- **Records**: 1,000 inspection events
- **Total Variables**: 22
    - Qualitative: 7
    - Quantitative: 10
    - Derived: 5
- **File Format**: CSV
- **File Name**: *Integrated_Electrical_Inspection_Compliance_Analytics_Dataset.csv*

5. Qualitative Variables

1. **SiteLocation** – Construction site or project name
2. **ContractorID** – Unique identifier for subcontractors
3. **InspectionType** – Stage of inspection (Initial, Follow-up, Certificate Issuance)
4. **MaterialType** – Type of material inspected (Electrical, Concrete, Steel, Finishings)
5. **VoltageAnomalyType** – Anomaly category (Normal, Sag, Swell, Harmonic Distortion)
6. **CertificateType** – Type of documentation issued post-inspection
7. **PassFailOutcome** – Result of inspection based on predefined criteria

6. Quantitative Variables

1. **VoltageDeviation_pct** – Percent deviation from nominal voltage
2. **PowerFactor** – Measure of power quality efficiency
3. **THD_pct** – Total Harmonic Distortion as a quality metric
4. **InspectionsConducted** – Number of inspections performed
5. **InspectionDuration_h** – Time spent on the inspection
6. **NumberOfNCRs** – Count of non-conformance reports raised
7. **VendorDelayDays** – Days of delay attributable to vendors
8. **SafetyIncidents** – Count of incidents logged per inspection

© 2025 Dr. Yasser Aly Yakout Shehata

9. **DaysToCloseNCR** – Time taken to resolve NCRs
10. **ReworkCost_USD** – Estimated financial impact of rework required

7. Derived Variables

1. **PerformanceScore** – Composite score incorporating quality, compliance, and vendor metrics
2. **SitePassRate** – Percent of inspections passed for each site
3. **InspectionRiskScore** – Composite index assessing risk based on anomalies, delays, and safety
4. **EfficiencyIndex** – Inverse function of duration and NCRs indicating operational efficiency
5. **RiskCategory** – Categorical label (Low, Medium, High) based on the risk score

8. Relevance to Power BI and AI for Business

IEICAD is purpose-built for use in Microsoft Power BI, allowing students and analysts to visualise performance indicators across multiple dimensions. The dataset enables cross-filtering of outcomes by contractor, site, or inspection type and supports the creation of KPI dashboards, trend visualisations, and predictive panels.

In AI applications, the dataset supports supervised and unsupervised modelling techniques, such as classification models to predict failure outcomes, clustering of risk profiles, and regression models to forecast rework costs or safety incident probability.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

Students can perform outlier detection, correlation analysis, and variable distribution assessments. The dataset allows discovery of patterns such as how electrical quality affects inspection outcomes or how vendor delays impact cost.

B. Power BI Visualisation

The data supports interactive dashboards with filters for site, contractor, material type, and inspection outcome. KPIs such as Performance Score and Site Pass Rate can be visually tracked to simulate a project management report.

C. AI-Driven Business Insights

Advanced use cases include:

© 2025 Dr. Yasser Aly Yakout Shehata

- Predictive modelling for PassFailOutcome using inspection and vendor variables
- Clustering sites based on InspectionRiskScore and EfficiencyIndex
- Forecasting ReworkCost based on THD, NCRs, and vendor performance

10. Conclusion

The **Integrated Electrical Inspection and Compliance Analytics Dataset (IEICAD)** represents a realistic and strategically structured data resource for graduate-level education in business analytics. By integrating real-world construction site dynamics with managerial performance indicators, the dataset enables a broad spectrum of analytical activities from descriptive exploration to advanced AI modelling. It offers an immersive learning opportunity for MBA students, particularly those examining data-driven decision-making in engineering and construction settings. Through this dataset, students can acquire practical competencies in handling operational data, visualising performance trends, and deriving actionable insights essential to modern project and risk management.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 14: Smart Contracts and Commercial Insights Dataset

1. Introduction

The **Smart Contracts and Commercial Insights Dataset for MBA Projects (SCCI)** is a purposefully constructed synthetic dataset designed to simulate the operational and financial environment encountered by professionals in contract management and commercial roles within the construction industry. Developed with academic rigor and practical application in mind, the dataset serves as a rich analytical resource for MBA students pursuing coursework in **Business Analytics for Decision Making**, **AI for Business**, and related disciplines. It specifically reflects the work context and analytical challenges faced by a **Senior Contracts and Commercial Manager**, a critical role responsible for ensuring the financial viability, legal compliance, and performance optimization of high-value construction projects.

In today's data-driven economy, construction firms operate under increasingly complex regulatory environments and heightened performance expectations. Commercial managers must balance contract obligations, financial forecasting, risk mitigation, and client communication, all while ensuring profitability. This dataset captures that complexity by embedding realistic variables such as contract values, work-in-progress balances, variations, receivables, risk assessments, and compliance indicators. It mirrors the decision-making process that takes place in companies like **Orascom Construction**, a global leader in infrastructure and industrial development.

By simulating contract-level data across diverse geographies, legal frameworks, and commercial models, the dataset enables students to develop and apply analytical thinking to real-world business challenges. Through both quantitative and qualitative variables, learners can explore a wide range of use cases, from evaluating billing efficiency and forecasting financial margins to detecting high-risk projects and recommending compliance interventions.

The **SCCI dataset** also encourages students to adopt a holistic approach to analytics by integrating descriptive, diagnostic, and predictive techniques. It supports **Exploratory Data Analysis (EDA)** to understand variable distributions and relationships, facilitates **Power BI dashboard creation** for visual storytelling, and enables **AI-driven insight generation** through classification, regression, and clustering models. Furthermore, the dataset is structured to align with the tools and workflows commonly used in enterprise settings, making it directly transferable to professional practice.

Most importantly, this dataset provides a platform for MBA students to experience the analytical responsibilities of senior-level managers who are expected to turn complex

© 2025 Dr. Yasser Aly Yakout Shehata

datasets into actionable business strategies. It bridges academic theory with industry application, ensuring that graduates are well-equipped to tackle data-centric roles in construction, project management, consulting, or general management.

In sum, the **SCCI dataset** is not only a technical resource but also a pedagogical tool that fosters critical thinking, analytical precision, and strategic problem-solving. It empowers MBA candidates to derive meaningful insights from data and present those insights in a compelling and business-relevant manner. As such, it is ideally suited for capstone projects, group analytics assignments, and individual explorations of the intersection between data and decision-making in business.

2. Dataset Description

The dataset includes **1,000 records**, each representing a monthly or contract-level summary within a commercial construction environment. It combines contract metadata, financial data, operational status, and performance indicators to simulate the full cycle of commercial contract management.

With **7 qualitative**, **10 quantitative**, and **6 derived variables**, the dataset is structured to allow users to perform meaningful **Exploratory Data Analysis (EDA)**, build **interactive dashboards in Power BI**, and develop **AI-driven decision models**.

3. Purpose and Application

The primary purpose of this dataset is to empower MBA students to:

- Analyze real-world business scenarios faced by commercial managers.
- Develop dashboards and visual analytics in Power BI.
- Apply machine learning models for classification, prediction, and segmentation.
- Draw business insights and make informed, data-driven recommendations.

By grounding the dataset in the role of a **Senior Contracts and Commercial Manager**, students gain contextual understanding of industry challenges such as cost overruns, delayed invoicing, compliance issues, and margin optimization.

4. Dataset Structure

Each row in the dataset represents a snapshot of a single contract instance or reporting period, characterized by its geographic, sectoral, and legal attributes, alongside financial and performance metrics.

- **Records**: 1,000

© 2025 Dr. Yasser Aly Yakout Shehata

- **Attributes**: 23 columns
- **Time Dimension**: Contract award date and months since award
- **Entity**: Construction contracts

5. Qualitative Variables (Categorical)

1. **region** – Project location (e.g., Egypt, KSA, UAE, USA, Other)
2. **sector** – Type of project (Infrastructure, Industrial, Commercial)
3. **contract_type** – Contract model (Lump-Sum, Cost-Plus, Unit Price)
4. **law_type** – Governing legal framework (Law 182, Law 89, Other)
5. **variation_status** – Status of change requests (None, Pending, Approved, Rejected)
6. **invoice_status** – Payment condition (Paid, Overdue, Disputed)
7. **compliance_flag** – Whether contract requires compliance review (Yes, No)

6. Quantitative Variables (Continuous/Numeric)

1. **contract_value_usd** – Total awarded contract value
2. **awarded_date** – Date of contract award
3. **months_since_award** – Time elapsed since contract start
4. **wip_balance_usd** – Work-in-progress cost
5. **billed_amount_usd** – Cumulative billing to client
6. **invoiced_amount_usd** – Amount invoiced during the period
7. **variation_value_usd** – Net variation adjustment to contract
8. **receivables_usd** – Outstanding client receivables
9. **ebitda_margin_pct** – Earnings before interest, tax, depreciation, and amortization margin (%)
10. **contract_risk_score** – Risk index (1 = Low, 5 = High)

7. Derived Variables (For Deeper Analysis)

1. **invoice_efficiency_ratio** – Ratio of invoiced to billed amount
2. **variation_pct** – Variations as a percent of contract value

© 2025 Dr. Yasser Aly Yakout Shehata

3. **contract_age_group** – Age bracket of the contract (<1yr, 1–3yrs, etc.)
4. **is_overdue** – Boolean flag for overdue invoices
5. **high_risk_flag** – Flag for contracts with risk score ≥ 4
6. **low_margin_flag** – Flag for EBITDA margin < 5%

These variables introduce advanced analytical depth, allowing segmentation, performance benchmarking, and risk stratification across multiple business dimensions.

8. Relevance to Power BI and AI for Business

This dataset is designed with the analytical demands of modern business decision-making in mind. It is highly relevant for:

- **Power BI**: Creating dashboards that track KPIs such as revenue realization, contract aging, margin performance, and billing efficiency.
- **AI for Business**: Modeling risk, predicting overdue invoices, and segmenting contracts based on performance indicators.

The structure and diversity of variables mirror those found in enterprise resource planning (ERP) systems and construction project management software, offering high-fidelity simulation for real-world business applications.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify trends in contract risk by region or contract type
- Explore variation frequency and magnitude across sectors
- Detect relationships between receivables and invoicing practices

B. Power BI Visualization

- Develop dashboards showing:
    - Risk distribution by law type
    - Invoice vs. billed ratio by client or region
    - EBITDA margin performance across sectors and contract types

C. AI-Driven Business Insights

- Use classification models to predict invoice delinquency

© 2025 Dr. Yasser Aly Yakout Shehata

- Apply regression to estimate EBITDA margin based on contract and compliance variables
- Cluster contracts into risk profiles for targeted mitigation strategies

10. Conclusion

The **Smart Contracts and Commercial Insights Dataset for MBA Projects (SCCI)** serves as a comprehensive, industry-relevant resource for MBA students pursuing advanced skills in business analytics, Power BI, and AI-enhanced decision making. By reflecting the realities of a Senior Contracts and Commercial Manager's role in the construction industry, this dataset bridges the gap between theoretical learning and applied analytics.

It empowers future business leaders to transform raw data into strategic insights, preparing them for leadership roles in complex and data-rich environments.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 15: Smart Project Operations Dataset for MEP Decision-Making

1. Introduction

The **Smart Project Operations Dataset for MEP Decision-Making** has been developed to support postgraduate-level education in business analytics with an emphasis on practical, industry-aligned applications. It is particularly crafted for MBA students undertaking the **course *Business Analytics for Decision Making***, providing them with a realistic and analytically rich dataset anchored in the operational environment of the construction industry. The dataset places a specific focus on the activities and decision-making processes of a **Technical Office Manager – MEP (Mechanical, Electrical, and Plumbing)**, a pivotal role in modern construction project management.

In large-scale infrastructure, industrial, residential, and commercial construction projects, the **Technical Office Manager – MEP** is responsible for managing the technical review of engineering documents, coordinating procurement activities, overseeing subcontractor performance, and ensuring timely delivery of MEP systems in accordance with client specifications and international standards. This professional must handle a wide array of tasks including drawing reviews, clash detection, material take-offs, coordination with design and civil teams, documentation, commissioning planning, and budget tracking. These responsibilities generate significant volumes of data that are crucial for effective project control, risk mitigation, and performance optimisation.

In recognition of this complexity, the dataset has been synthetically generated and enhanced using business logic that mirrors real-world project scenarios. The values across all variables are not randomly assigned but are informed by operational dependencies and conditional logic that reflect actual business processes. For example, a higher number of technical drawing reviews often correlates with a greater probability of design clashes, which in turn may result in project delays and increased costs. Similarly, poor vendor performance is likely to be associated with schedule deviations and a higher number of RFIs. By embedding such interdependencies, the dataset offers a realistic platform for analytical modelling and managerial decision-making simulations.

The dataset comprises 1,000 records, each representing a discrete project or technical milestone under the purview of an MEP manager. It includes both categorical and numerical variables, along with derived metrics that offer deeper analytical value. The design of the dataset allows for the application of a wide range of analytical techniques including Exploratory Data Analysis (EDA), trend analysis, dashboarding, predictive modelling, and AI-driven business insights.

© 2025 Dr. Yasser Aly Yakout Shehata

Moreover, the structure and format of the dataset make it particularly suitable for visualisation using Power BI. Users can create interactive dashboards to monitor procurement efficiency, track budget variances, identify high-risk projects, and compare vendor performance across different project types. Additionally, the dataset lends itself to AI and machine learning techniques by providing patterns and relationships that can be leveraged for classification, clustering, and forecasting exercises.

By simulating a complex operational domain with high analytical relevance, this dataset not only bridges the gap between theory and practice but also prepares students to manage data-intensive decision environments in the real world. It serves as a pedagogically robust and practically relevant tool to build competencies in data literacy, business intelligence, and strategic use of AI in the context of construction and project management.

2. Dataset Description

The dataset contains **1,000 records**, each representing a construction project or operational scenario managed by a Technical Office Manager – MEP. The data is composed of **7 qualitative variables**, **10 quantitative variables**, and **2 derived variables**, all embedded with logical relationships to simulate real-world dependencies among project dynamics, cost structures, technical coordination, and performance outcomes.

The values are generated to reflect typical conditions found in large-scale projects, including procurement strategies, vendor reliability, RFIs, delays, and budgeting deviations, enabling learners to perform a wide range of analytical and visual tasks.

3. Purpose and Application

This dataset was created with three core objectives:

- To provide **realistic, business-contextual data** for MBA students to apply analytical techniques.
- To support practical use of **Power BI and data visualisation tools** in developing performance dashboards.
- To allow the use of **AI-based methods** for deriving predictive insights, risk scores, and optimisation recommendations.

4. Dataset Structure

Each record in the dataset includes information on project classification, technical drawing reviews, procurement details, team coordination metrics, and financial indicators. The variables are logically consistent and interrelated to simulate operational scenarios encountered by MEP managers, enhancing interpretability and analytical value.

© 2025 Dr. Yasser Aly Yakout Shehata

5. Qualitative Variables (Categorical)

1. **Project Type**: Nature of the construction project (Residential, Commercial, Infrastructure, Industrial)
2. **Procurement Strategy**: Method used for material and contractor acquisition (Tender, Negotiation, Framework Agreement, Direct Purchase)
3. **Clash Detection Tool**: Software/tool used for detecting design conflicts (Revit, Navisworks, Civil 3D, Manual)
4. **Commissioning Phase**: Project stage (Pre-Commissioning, Main Commissioning, Post-Commissioning)
5. **Vendor Performance Rating**: Quality of vendor execution (Low, Medium, High, Excellent)
6. **Design Change Type**: Origin of changes in design (Scope, Design Spec, Code Update, Client Request)
7. **Reporting Frequency**: Frequency of internal or external reporting (Weekly, Bi-Weekly, Monthly)

6. Quantitative Variables (Numerical)

1. **Team Size**: Number of team members involved
2. **Number of Drawings Reviewed**: Volume of technical documents examined
3. **Procurement Cost (USD)**: Total spending on MEP-related procurement
4. **Clash Detected Count**: Number of conflicts identified in technical plans
5. **Number of RFIs Issued**: Requests for Information during execution
6. **Subcontractor Delay (Days)**: Time lost due to subcontractor performance
7. **Material Takeoff Variance (%)**: Deviation between estimated and actual materials
8. **Commissioning Hours**: Time required to conduct system testing and commissioning
9. **Invoice Amount (USD)**: Amount billed to clients or partners
10. **Budget Deviation (%)**: Percentage of deviation from the project budget

© 2025 Dr. Yasser Aly Yakout Shehata

7. Derived Variables

1. **Cost per Drawing (USD)**: Calculated by dividing procurement cost by number of drawings reviewed; reflects cost efficiency of technical design execution
2. **Delay Risk Index**: Composite score computed using subcontractor delays, clash counts, and team size to assess the risk of schedule disruption

8. Relevance to Power BI and AI for Business

The dataset is structured to maximise compatibility with Power BI for advanced data visualisation. Users can group data by project type, filter by procurement strategy or vendor rating, and visualise key metrics such as budget deviation, delay risks, and procurement costs.
In parallel, the embedded logic and derived indicators support AI-driven analyses such as risk prediction, vendor performance classification, and budget forecasting, making this dataset highly relevant for exploring real-world applications of AI in business operations.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identification of patterns, trends, and outliers across project types or vendor strategies
- Correlation analysis between design complexities, procurement cost, and delays
- Analysis of how different clash detection tools impact budget and commissioning timelines

B. Power BI Visualisation

- Construction of interactive dashboards for project monitoring
- Visual representation of procurement vs. invoice variance
- Dynamic filtering by reporting frequency, vendor rating, and delay risk

C. AI-Driven Business Insights

- Predictive modelling of budget deviation based on operational variables
- Classification of vendor performance using quantitative indicators
- Segmentation of projects by risk level using unsupervised learning techniques

© 2025 Dr. Yasser Aly Yakout Shehata

10. Conclusion

The **Smart Project Operations Dataset for MEP Decision-Making** provides a rich, realistic, and logically consistent foundation for applying modern data analytics and AI tools within the MBA curriculum. Rooted in the specific duties of a Technical Office Manager – MEP, the dataset enables learners to explore and practice decision-making in a data-intensive environment. Its relevance spans descriptive analytics, visual storytelling through Power BI, and predictive modelling, equipping students with the skills to transform data into actionable business insights in the construction sector and beyond.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 16: Operational Site Management Analytics Dataset (OSMAD)

1. Introduction

The **Operational Site Management Analytics Dataset (OSMAD)** is a strategically designed, synthetically generated dataset developed for advanced academic and professional use in the field of business analytics. Specifically crafted for MBA students enrolled in the ***Business Analytics for Decision Making*** **course**, this dataset enables learners to engage with realistic, operationally rich data representative of construction project sites. The primary aim of the dataset is to serve as a training and assessment tool, allowing students to apply exploratory data analysis, build visual analytics through Power BI, and leverage artificial intelligence for informed decision-making.

At the heart of this dataset is the role of the **Site Manager**, a critical figure in the execution and daily supervision of construction projects. The Site Manager oversees all on-site activities, ensuring that construction operations proceed according to design specifications, regulatory standards, and safety protocols. This position requires close coordination with subcontractors, labour teams, engineers, consultants, and quality control officers. It also involves managing materials, responding to design queries through RFIs (Requests for Information), handling delays due to environmental or permitting conditions, and ensuring cost efficiency. Each of these responsibilities contributes to a data-rich environment in which performance can and should be measured to support more effective management and strategic oversight.

Recognising the value of these data points, **OSMAD** was designed to simulate 1,000 discrete operational site-day records, each capturing critical variables such as labour utilisation, material wastage, cost dynamics, nonconformities, safety incidents, and schedule performance. These variables are logically interrelated to reflect realistic dependencies that occur in day-to-day construction management. For example, extreme weather conditions tend to reduce daily labour hours and lower productivity rates, while permitting delays often lead to a rise in critical-path days and overall schedule deviation. Subcontractor invoices are proportionate to site cost and are influenced by project type, while safety incidents and nonconformities directly affect a calculated Risk Index, which serves as an indicator of operational volatility.

The dataset incorporates both **qualitative** (categorical) and **quantitative** (numerical) data, making it ideal for multidimensional analysis. Additionally, it features **derived variables** such as Cost per Labor Hour and Risk Index, which add analytical depth and enable the application of statistical models, forecasting techniques, and AI-driven insights. These

© 2025 Dr. Yasser Aly Yakout Shehata

calculated fields are particularly useful for students seeking to move beyond descriptive analytics to prescriptive and predictive decision support.

Moreover, **OSMAD** is intentionally structured for seamless use with data visualisation tools such as Microsoft Power BI. Students can build dashboards to track key performance indicators, compare project types, monitor safety compliance trends, and analyse resource allocation effectiveness. The dataset also supports AI and machine learning applications, such as clustering site-days by risk profile, classifying operational efficiency levels, or predicting cost overruns based on historical site data.

In summary, **OSMAD** bridges the gap between academic instruction and industry application. It challenges **MBA students** to think critically, reason through complex operational variables, and use data as a central tool in managing and improving site performance. By offering a simulation of real-world construction site dynamics, the dataset provides a high-value platform for developing business analytics competencies that are increasingly essential in today's data-driven management landscape.

2. Dataset Description

OSMAD includes 1,000 records, each representing a full day of operational data from a construction project site under the supervision of a Site Manager. The data captures site-level dynamics, operational challenges, safety indicators, quality performance, and cost metrics. The variables are logically linked through a structure designed to simulate plausible field conditions, with embedded causality among weather, delays, cost fluctuations, productivity, and incident rates.

The dataset is enriched with **derived variables**, which provide additional insight into cost efficiency and operational risk, further enabling meaningful analysis and modelling.

3. Purpose and Application

The primary purpose of **OSMAD** is to provide MBA students and educators with a realistic dataset to:

- Practise **Exploratory Data Analysis (EDA)** and pattern identification.
- Create **interactive dashboards** and performance reports using Power BI.
- Apply **descriptive and predictive analytics** for site management decisions.
- Simulate **AI-based models** for classification, clustering, and forecasting in a construction context.

© 2025 Dr. Yasser Aly Yakout Shehata

This dataset supports practical skill development in interpreting data, identifying trends, diagnosing performance issues, and evaluating risk in construction project operations.

4. Dataset Structure

The dataset is composed of:

- **7 Qualitative (categorical) variables** describing context and status.
- **10 Quantitative (numerical) variables** capturing measurable performance and cost indicators.
- **2 Derived variables** calculated to enhance decision support.

Each record includes detailed information about a specific day on-site, structured to support time-based or project-based aggregation and segmentation.

5. Qualitative Variables

1. **Project Type**: Type of construction project (Residential, Commercial, Infrastructure, Industrial).
2. **Weather Condition**: Site weather on the day (Sunny, Rainy, Windy, Extreme).
3. **Safety Briefing Frequency**: Frequency of conducted safety sessions (Daily, Weekly, Bi-Weekly).
4. **RFI Source**: Origin of Requests for Information (Consultant, Contractor, Design Team).
5. **Equipment Status**: Availability of site equipment (Available, UnderMaintenance, Delayed).
6. **Permitting Status**: Status of site regulatory compliance (In-Progress, Approved, PendingInspection).
7. **Material Wastage Control**: Effectiveness of material management (Good, Moderate, Poor).

6. Quantitative Variables

1. **Labor Hours Daily**: Total labour hours utilised.
2. **Productivity Rate**: Ratio of completed to planned work.
3. **Subcontractor Invoice (USD)**: Invoice value issued by subcontractors.
4. **Material Wastage Percentage**: Percentage of material lost or overused.

© 2025 Dr. Yasser Aly Yakout Shehata

5. **Nonconformities Count**: Quality non-compliance issues recorded.
6. **Safety Incidents Count**: Reported safety events or hazards.
7. **RFIs Issued Count**: Number of RFIs initiated.
8. **Critical Path Days**: Number of days directly impacting schedule.
9. **Daily Schedule Deviation Percentage**: Discrepancy between planned and actual progress.
10. **Site Cost (USD)**: Total operational cost for the day.

7. Derived Variables

1. **Cost per Labor Hour**: Reflects the cost-efficiency of labour deployment.
    - Calculated as: Site Cost divided by Labour Hours plus one
2. **Risk Index**: Composite metric indicating operational risk exposure.
    - Computed from safety incidents, nonconformities, and schedule delays normalised by productivity

These variables allow deeper insight into cost optimisation and risk mitigation strategies.

8. Relevance to Power BI and AI for Business

The structure of OSMAD allows direct integration into Power BI to produce:

- Real-time dashboards tracking safety, productivity, and cost trends
- Filtering by qualitative categories (e.g., weather, equipment status, permitting)
- Aggregations for weekly, monthly, or project-type summaries

For AI applications, the dataset supports:

- Predictive modelling of cost overruns and incident risks
- Clustering of project types based on performance patterns
- Classification of material control effectiveness or risk exposure levels

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify correlations between weather and productivity
- Compare cost-efficiency across project types and material control quality

© 2025 Dr. Yasser Aly Yakout Shehata

- Evaluate how safety briefings influence incident rates

B. Power BI Visualisation

- Build dashboards tracking daily KPIs: labour use, costs, delays
- Interactive segmentation by equipment status or permitting bottlenecks
- Visual alerts for high-risk days based on the Risk Index

C. AI-Driven Business Insights

- Predict high-risk scenarios using decision trees or logistic regression
- Cluster days by similarity in cost, risk, or productivity using unsupervised learning
- Recommend optimal safety briefing frequency or equipment readiness strategies

10. Conclusion

The **Operational Site Management Analytics Dataset (OSMAD)** is a powerful educational tool for developing strategic analytical capabilities in business decision-making. Grounded in real-world construction operations and structured for analytical integrity, it offers **MBA students** a practical environment for applying business intelligence, visualisation, and AI techniques. By simulating the complexities and interdependencies faced by Site Managers, the dataset bridges theoretical learning with operational decision-making, fostering actionable insights in project management and operational excellence.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 17: Heavy Equipment Sales and Performance Intelligence

1. Introduction

The ***Heavy Equipment Sales and Performance Intelligence Dataset (HESPID)*** is an academically curated, synthetically generated dataset designed to replicate the complex, high-stakes environment of industrial equipment sales. Developed for MBA students enrolled in the ***Business Analytics for Decision Making*** **course**, this dataset provides a robust platform for applying core analytical skills in real-world business contexts. It is particularly focused on the role of a **Senior Sales Engineer**, a position that bridges technical knowledge with commercial strategy. This role exists at the intersection of customer needs assessment, technical equipment specification, deal negotiation, and post-sale client support, making it ideal for multi-dimensional analysis and decision-making scenarios.

The dataset simulates 1,000 individual transaction records within a leading heavy equipment distribution firm, modelled on the operational ecosystem of entities like Mantrac Egypt. Each record represents a sales interaction that includes variables related to the type of equipment offered, client sector, contract model, quotation value, discount applied, and client feedback after the transaction. These dimensions reflect the multi-phase nature of industrial sales, which often involve complex proposals, technical site visits, negotiation cycles, equipment demonstrations, and long-term service agreements.

What distinguishes this dataset is the integration of **embedded business logic**, which ensures that the values and interactions reflect plausible, cause-effect relationships. For example, sales cycle duration is meaningfully extended for tender-based sales and rental agreements, while higher revenue is statistically associated with customers from sectors such as mining and energy. Discounts vary based on the payment model and sales channel, while customer satisfaction correlates with the quality of post-sale support. These logical structures offer a meaningful learning experience that mirrors professional decision-making, making the dataset suitable not only for statistical exploration but also for strategic analysis.

Another key feature of **HESPID** is the inclusion of **derived variables** such as *Revenue per Demo Day* and *Sales Efficiency Index*. These fields enable a more nuanced understanding of performance metrics and allow learners to construct composite indicators that are often used in executive dashboards and AI-driven recommendations. Such variables encourage exploration of cost-benefit trade-offs and temporal efficiency, which are crucial in B2B sales contexts.

Designed for use with platforms such as **Microsoft Power BI**, the dataset supports the creation of interactive visualisations, real-time dashboards, and what-if analysis scenarios.

© 2025 Dr. Yasser Aly Yakout Shehata

It also opens pathways for applying **machine learning models**, such as predicting the likelihood of a successful deal or clustering clients based on equipment needs and support preferences. The format and logic of the dataset make it particularly suitable for both descriptive analytics and predictive modelling, bridging traditional business intelligence with modern AI tools.

In essence, **HESPID** goes beyond conventional academic datasets by offering contextual authenticity, logical consistency, and analytical versatility. It encourages students to not only apply statistical and visualisation tools, but also to interpret patterns, diagnose performance gaps, and propose data-informed recommendations. By simulating the decision environment of a Senior Sales Engineer, the dataset fosters the development of competencies in analytical reasoning, storytelling through data, and the strategic use of analytics in industrial sales management. This aligns directly with the learning outcomes of MBA-level business analytics education, particularly in preparing future professionals to navigate data-rich decision environments with clarity, confidence, and insight.

2. Dataset Description

The dataset consists of one thousand synthetic records that represent sales interactions and operational performance metrics. Each record reflects a distinct sales case involving a piece of heavy equipment, linked to customer sector, region, contract type, and other dimensions. The data has been generated using business logic and statistical principles to mirror plausible real-world distributions and relationships. This includes variations in sales cycle duration, quotation values, customer satisfaction, and revenue outcomes.

3. Purpose and Application

This dataset serves three primary educational objectives:

- To enable hands-on experience with **Exploratory Data Analysis (EDA)** techniques.
- To allow students to create **interactive dashboards and KPIs in Power BI**, promoting data storytelling and performance reporting.
- To provide a basis for **predictive modelling and clustering techniques** within the context of AI for business.

It is particularly suited for final projects in MBA-level Business Analytics modules, offering ample opportunity to explore sales forecasting, customer segmentation, sales effectiveness, and decision support through machine learning.

4. Dataset Structure

The dataset is structured with:

© 2025 Dr. Yasser Aly Yakout Shehata

- **1,000 records**
- **7 qualitative (categorical) variables**
- **10 quantitative (numerical) variables**
- **2 derived variables** enhancing analytical depth

Each observation is uniquely identified and reflects a simulated case of heavy equipment sales and its associated attributes and outcomes.

5. Qualitative Variables

1. **Equipment_Type**: Type of heavy equipment (e.g., Bulldozer, Excavator)
2. **Sales_Channel**: Method through which the sale was initiated (e.g., Direct Sale, Dealer Tender)
3. **Region**: Geographical sales region (e.g., North Egypt, West Egypt)
4. **Customer_Sector**: Sector to which the client belongs (e.g., Mining, Construction)
5. **Contract_Type**: Type of commercial agreement (e.g., Rental, Purchase)
6. **Payment_Terms**: Payment model used (e.g., Net 30, Upfront)
7. **Lead_Source**: Origin of sales lead (e.g., Referral, Cold Call)

6. Quantitative Variables

1. **Quote_Value_USD**: Quoted value of the equipment or service
2. **Discount_Percentage**: Discount offered on the quotation
3. **Sales_Cycle_Days**: Duration from lead generation to deal closure
4. **Technical_Visits_Count**: Number of technical consultations prior to closure
5. **Proposal_Revisions_Count**: Frequency of proposal modifications
6. **Equipment_Demo_Days**: Days spent on client-site demonstrations
7. **Post_Sale_Support_Hours**: Hours of support provided after sale
8. **Revenue_USD**: Realised revenue upon deal closure
9. **Customer_Satisfaction_Score**: Rated feedback on a 1 to 10 scale
10. **Equipment_Utilization_Hours**: Average operational hours post-sale

© 2025 Dr. Yasser Aly Yakout Shehata

7. Derived Variables

1. **Revenue_per_Demo_Day**: Indicates operational value returned per demo day
2. **Sales_Efficiency_Index**: Measures efficiency as a function of revenue, deal success, and time-to-close

These variables support advanced analytics tasks and enable comparisons across sales strategies and operational scenarios.

8. Relevance to Power BI and AI for Business

HESPID is highly compatible with Power BI dashboards, enabling:

- Multi-layered slicing by customer sector, region, and contract type
- Visualisation of KPI trends across time and channel
- Correlation matrices to explore drivers of satisfaction and revenue

In AI contexts, this dataset supports the development of:

- Classification models to predict deal success
- Regression models for revenue forecasting
- Clustering algorithms to define high-value client profiles

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

Students can explore central tendencies, outliers, and distributions. Key relationships such as the link between equipment type and revenue, or between satisfaction and support hours, are well suited for deep analysis.

B. Power BI Visualisation

Through data slicing, filtering, and drill-down capabilities, learners can build dashboards that offer real-time business intelligence to support strategic decision-making.

C. AI-Driven Business Insights

The dataset can be used to train models that forecast sales outcomes, recommend customer engagement strategies, and optimise resource allocation.

© 2025 Dr. Yasser Aly Yakout Shehata

10. Conclusion

The ***Heavy Equipment Sales and Performance Intelligence Dataset (HESPID)*** offers a uniquely crafted, industry-contextualised data environment for **MBA students**. By emulating the operational domain of a Senior Sales Engineer, it serves as a bridge between theoretical coursework and applied business decision-making. With a balanced mix of structured and derived variables, the dataset encourages data-driven thinking and fosters skill development in analytics, reporting, and artificial intelligence within the context of real-world enterprise performance.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 18: Power Quality Insights for Energy Sector Performance

1. Introduction

In today's energy sector, where large-scale infrastructure projects must meet stringent standards for quality, efficiency, safety, and compliance, data has become a critical tool for driving strategic decision-making. The role of the **Power Sector Quality Director** has evolved into one that is increasingly data-driven—requiring real-time insights into project performance, audit compliance, contractor reliability, and digital monitoring effectiveness. Against this backdrop, the **Power Quality Insights Dataset for Energy Sector Performance** has been developed as a synthetic but realistic simulation of the complex data landscape encountered in modern power generation projects.

This dataset is designed with the academic and professional needs of **MBA students** in mind, particularly those enrolled in **Business Analytics for Decision Making** or related courses. It enables learners to step into the analytical role of a Power Sector Quality Director and practice the process of transforming operational data into actionable insights. Drawing from the structure of real energy sector operations, the dataset covers a broad spectrum of project-related variables, including quality audit outcomes, nonconformity tracking, sensor deployment, certification statuses, and performance metrics. This structure allows students to explore how data is collected, interpreted, and applied in real-world quality management scenarios.

The **Power Quality Insights Dataset** features 1,000 unique project records encompassing diverse energy sources such as solar, wind, gas, hydro, and nuclear. Each record captures the operational phase of the project, the nature and frequency of quality inspections, corrective actions taken, and the technological maturity of the project site in terms of sensor implementation. It also includes calculated fields that assess project risk levels and operational efficiency, mirroring how quality directors would synthesize various data points to prioritize action and allocate resources.

As digitalization and artificial intelligence (AI) become embedded in the future of infrastructure management, the ability to analyze data using tools like **Power BI** and **machine learning models** becomes not just useful, but essential. This dataset provides a foundation for students to practice **Exploratory Data Analysis (EDA)**, develop **Power BI dashboards**, and apply **AI techniques** such as classification, regression, and clustering, all within the highly relevant and high-stakes context of power infrastructure.

What sets this dataset apart is its ability to simulate **strategic, operational, and technological considerations** faced by decision-makers. It challenges students to go beyond surface-level reporting and engage with advanced analytics to identify patterns,

© 2025 Dr. Yasser Aly Yakout Shehata

uncover root causes, and support executive-level recommendations. By bridging data with business context, the Power Quality Insights Dataset empowers future business leaders to become not only data-literate, but also insight-driven.

Whether used for coursework, capstone projects, or professional development, this dataset presents a compelling and comprehensive resource for anyone aiming to understand the power of analytics in managing complex, quality-focused operations in the energy sector.

2. Dataset Description

The dataset includes **1,000 records**, each representing a distinct energy project. Each record captures detailed information about project type, phase, regional location, audit outcomes, quality issues, and sensor monitoring. In addition to foundational qualitative and quantitative data, the dataset includes derived metrics that assess risk and operational efficiency, key perspectives for a Power Sector Quality Director managing diverse portfolios.

The data is logically structured and statistically consistent to enable meaningful pattern discovery, visual reporting, and AI-based modeling.

3. Purpose and Application

This dataset is created to:

- Support student projects in descriptive analytics, dashboard building, and AI applications
- Simulate real-world data interpretation tasks for quality and operations professionals
- Help students develop skills in drawing actionable insights from complex datasets
- Provide a foundation for decision-making scenarios across certification, contractor performance, and audit results
- Serve as a capstone resource for learning tools such as Power BI, Python, and Excel

4. Dataset Structure

- Total records: 1,000 project observations
- Total variables: 19 (7 qualitative, 10 quantitative, 2 derived)
- Format: CSV, compatible with Excel, Power BI, Python, and R
- Balance: Includes operational, performance, and compliance metrics

© 2025 Dr. Yasser Aly Yakout Shehata

5. Qualitative Variables (7)

1. **Project_Type**: Specifies the type of energy project, including Solar, Wind, Gas, Hydro, and Nuclear.
2. **Location_Region**: Identifies the project's geographic location within defined regions—North, South, East, West, or Central.
3. **Phase**: Indicates the current stage of project development, categorized as Design, Procurement, Construction, Commissioning, or Operation.
4. **QC_Audit_Result**: Captures the outcome of a quality control audit, which may be classified as Pass, Minor NC (nonconformity), Major NC, or Critical NC.
5. **Contractor_Type**: Identifies whether the main contractor is In-house, Local, or International, influencing project oversight and collaboration.
6. **Certification_Status**: Represents the status of ISO or equivalent quality certifications—Certified, Pending, or Expired.
7. **Sensor_Monitoring**: Indicates whether the project uses real-time digital monitoring systems on-site (Yes or No).

6. Quantitative Variables (10)

1. **Project_Capacity_MW**: Denotes the total installed power capacity of the project, measured in megawatts.
2. **Inspection_Count**: Counts the number of formal quality inspections conducted on the project.
3. **Nonconformities_Count**: Tracks the number of identified quality issues or nonconformities during inspections.
4. **CAPA_Count**: Reflects the number of corrective and preventive actions issued in response to quality findings.
5. **Onsite_Sensors**: Indicates the number of functioning digital sensors installed at the project site.
6. **Defects_per_Inspection**: Shows the average number of defects discovered per inspection event.
7. **Audit_Score_%**: Reports the project's audit result as a percentage score, scaled from 0 to 100.

© 2025 Dr. Yasser Aly Yakout Shehata

8. **Calibration_Delay_Days**: Measures the delay, in days, for overdue calibration of quality control equipment.
9. **Time_to_Resolve_NCR_days**: Captures the average number of days taken to resolve nonconformities.
10. **Quality_KPI_Score**: A composite score aggregating quality performance metrics, reflecting audit results, issue resolution, and overall compliance.

7. Derived Variables (2)

1. **Risk_Level**: Categorizes each project as Low, Medium, or High risk based on a combination of audit score and the number of nonconformities. This variable simulates the kind of qualitative risk classification used in executive dashboards.
2. **Efficiency_Score**: A calculated metric combining audit performance and the presence of digital monitoring (sensor count), offering an estimate of operational maturity and oversight capability.

8. Relevance to Power BI and AI for Business

The **Power Quality Insights Dataset** is highly suitable for Power BI due to its blend of categorical filters, measurable KPIs, and logically segmented phases and locations. Students and professionals can use it to:

- Build dynamic dashboards with filters for region, contractor type, or certification status
- Visualize relationships between performance indicators and compliance measures
- Create visual heatmaps, risk matrices, and time-trend reports
- Integrate AI predictions into dashboards to enhance decision support

It is also structured for seamless application in AI modeling, including training classification models for risk level, clustering similar project profiles, and forecasting audit performance using regression.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA):

Students can perform statistical summaries, detect outliers, and evaluate correlations among variables such as sensor count, audit scores, and resolution delays. This enables strategic diagnosis of high-risk projects and ineffective contractors.

© 2025 Dr. Yasser Aly Yakout Shehata

B. Power BI Visualization:

Students can design dashboards that track audit outcomes, visualize regional or phase-specific trends, and present performance benchmarks. Interactive elements like slicers and charts can illustrate the impact of certification or sensor monitoring on quality outcomes.

C. AI-Driven Business Insights:

By applying AI methods such as logistic regression, decision trees, or clustering, students can generate actionable insights like predicting high-risk projects, grouping similar performance profiles, or identifying underperforming contractors based on structured data.

10. Conclusion

The **Power Quality Insights Dataset for Energy Sector Performance** is a strategically designed, pedagogically robust resource that helps bridge the gap between business theory and real-world analytics. Rooted in the functional needs of a **Power Sector Quality Director**, it reflects the type of high-stakes, multi-variable decision-making that professionals must master in infrastructure, energy, and operations roles.

By integrating realistic business logic, performance measurement, and AI-readiness, the dataset offers a complete platform for mastering EDA, Power BI storytelling, and AI for Business, all in one cohesive, domain-relevant environment.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 19: Integrated Site Operations and Performance Analytics

1. Introduction

The modern construction industry is undergoing a transformative shift driven by the integration of digital technologies, data analytics, and artificial intelligence (AI) into **traditional project management and site operations**. Within this evolving landscape, the role of the **Section Head Site Engineer** is becoming increasingly complex and analytically demanding. These professionals are no longer solely responsible for overseeing on-site activities; they are also required to synthesise diverse data streams, evaluate performance metrics, anticipate project risks, and contribute to **strategic decision-making processes**. Consequently, there is a growing need for data-driven competencies within this role, as well as an urgent demand for realistic, structured datasets that can support educational programmes designed to prepare future managers and analysts for the challenges of the field.

The **Integrated Site Operations and Performance Analytics (ISOPA) Dataset** has been developed in direct response to this need. It is a synthetically generated, yet structurally authentic dataset that replicates the types of data collected and analysed by a **Section Head Site Engineer** during the execution phase of a large-scale infrastructure project modelled in this case on the **El Dabaa Nuclear Power Plant**. This dataset is intended specifically for use by **MBA students** enrolled in **Business Analytics courses**, particularly those focusing on decision-making, data visualisation, and the application of AI tools in business contexts.

What sets the **ISOPA Dataset** apart is its careful embedding of business logic and inter-variable relationships that reflect actual site operations. The dataset incorporates both environmental and managerial factors, such as weather conditions, shift types, and supervision structures, as well as operational metrics including manpower deployment, equipment utilisation, material consumption, and progress tracking. By doing so, it provides a high-fidelity simulation of a real-world data environment, making it ideally suited for training in exploratory data analysis (EDA), dashboard design using Microsoft Power BI, and the application of machine learning models for performance prediction and resource optimisation.

In addition to capturing a wide spectrum of standard metrics, the **ISOPA Dataset** also introduces derived variables such as Efficiency Score and Performance Category. These fields are calculated through composite logic that integrates multiple performance dimensions, allowing students to practise synthesising operational indicators into business-relevant insights. This feature enhances the pedagogical value of the dataset by facilitating

© 2025 Dr. Yasser Aly Yakout Shehata

advanced analytical tasks, such as classification, clustering, and predictive modelling, which are central to AI-driven business strategies.

Through this dataset, MBA students will gain hands-on experience with interpreting complex, multivariate data in a structured and purposeful manner. They will learn how to extract insights that inform strategic decisions and improve operational outcomes. Furthermore, they will develop familiarity with the types of analytical questions and challenges encountered in engineering project environments, equipping them with the competencies required to operate effectively at the intersection of construction management and business analytics. Ultimately, the **ISOPA Dataset** is not just a tool for technical skill development, it is a strategic educational asset that fosters critical thinking, analytical reasoning, and business acumen among future leaders in data-informed decision-making.

2. Dataset Description

The ISOPA Dataset consists of 1,000 weekly records encompassing both operational and contextual attributes related to construction site management. Each record represents a unique observation drawn from site activities, integrating qualitative assessments with measurable performance indicators. The dataset has been enhanced through embedded business logic to simulate the decision-making environment of engineering professionals, particularly those involved in site execution, monitoring, and resource optimisation.

3. Purpose and Application

This dataset is intended to support the academic and practical development of MBA students specialising in Business Analytics, with direct applicability to:

- Performance monitoring and KPI analysis in construction projects
- Business intelligence through dashboards and reports
- AI-driven pattern recognition and predictive modelling
- Practical case studies and simulation of operational decisions

4. Dataset Structure

The ISOPA Dataset comprises 17 variables:

- **7 qualitative variables** that provide categorical context (e.g., weather, work phase)
- **10 quantitative variables** that capture numerical performance metrics (e.g., manpower, equipment use, variance indicators)

© 2025 Dr. Yasser Aly Yakout Shehata

Records span multiple simulated weeks across diverse site zones, phases of work, and supervisory roles, offering heterogeneity suitable for segmentation and trend analysis.

5. Qualitative Variables

1. **Week ID** – Numeric week indicator (1 to 52)
2. **Shift Type** – Type of work shift: Day, Night, or Mixed
3. **Work Phase** – Current stage: Foundation, Structure, Finishing, Commissioning
4. **Site Zone** – Operational zones: Zone A to Zone D
5. **Supervisor ID** – Encoded ID representing different supervisors
6. **Weather Condition** – Environmental factor: Clear, Rainy, Hot, Windy
7. **Safety Level** – Categorical safety score: Green (safe), Yellow (moderate), Red (unsafe)

6. Quantitative Variables

1. **Progress Percentage** – Percentage of planned weekly work completed
2. **Manpower Count** – Number of workers deployed per shift
3. **Material Usage (Tons)** – Volume of materials consumed
4. **Equipment Utilisation (%)** – Use rate of site machinery
5. **Safety Incidents** – Number of safety-related issues reported
6. **Quality Findings** – Count of quality-related anomalies identified
7. **Cost Variance (%)** – Percentage difference from budgeted cost
8. **Schedule Variance (Days)** – Number of days ahead or behind schedule
9. **EV/AC Ratio** – Earned Value to Actual Cost ratio, a cost performance index
10. **Digital Drawings Processed** – BIM or CAD files reviewed weekly

7. Derived Variables

1. **Efficiency Score** – A composite performance indicator integrating progress, equipment utilisation, and safety metrics
2. **Performance Category** – Categorical output based on Efficiency Score: High, Medium, Low

© 2025 Dr. Yasser Aly Yakout Shehata

These derived variables allow for greater analytical sophistication, supporting classification and scoring frameworks essential in AI modelling and business decision-making.

8. Relevance to Power BI and AI for Business

The ISOPA Dataset is designed to align seamlessly with Microsoft Power BI for dynamic dashboards and real-time analytics. Its well-structured combination of categorical and continuous variables facilitates filtering, drill-down analysis, conditional formatting, and data storytelling. Additionally, the dataset's inter-variable logic supports AI applications including regression, clustering, and classification.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identification of patterns in safety incidents across weather conditions
- Correlation of manpower and equipment utilisation with progress percentages
- Variance analysis across site zones and work phases

B. Power BI Visualisation

- Construction of interactive dashboards to monitor weekly KPIs
- Visual representation of cost and schedule performance trends
- Heat maps for site zone performance and resource allocation

C. AI-Driven Business Insights

- Predictive modelling to estimate project delays or cost overruns
- Classification of site operations into performance categories
- Clustering site conditions to recommend optimal operational setups

10. Conclusion

The **Integrated Site Operations and Performance Analytics (ISOPA) Dataset** offers a robust platform for experiential learning and applied research in business analytics. By integrating authentic variables with derived analytical constructs, it enables MBA students to explore the intersections of engineering management, business intelligence, and AI. The dataset mirrors the operational scope and challenges of a **Section Head Site Engineer,** thereby grounding academic exercises in industry-relevant complexity. As such, it equips learners with both technical proficiency and strategic insight, fostering the analytical mindset required to lead data-driven decision-making in modern infrastructure projects.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 20: Pharmacy Administration in South Sinai

Introduction

In the evolving landscape of healthcare delivery, the integration of data analytics and strategic decision-making is becoming a cornerstone of effective administration. This synthetic dataset has been meticulously crafted to simulate the operational environment of the **Director of Pharmacy Administration** within the **Egypt Healthcare Authority (EHA)**, specifically focusing on the **South Sinai branch**. The role of the Director entails balancing critical functions such as pharmaceutical inventory management, workforce planning, clinical service deployment, and cost control, each of which presents complex, data-rich challenges that are well-suited to analytical inquiry.

The dataset is designed to serve as a practical resource for **MBA students enrolled in Business Analytics for Decision Making**, offering a realistic and context-specific framework for the application of analytical tools. By emulating real-world dynamics such as stock-out events, pharmacist deployment levels, accreditation of pharmacy staff, and variations in patient volumes, the dataset enables students to practice the end-to-end analytics cycle. This includes data cleaning, exploratory data analysis (EDA), visualisation, and the use of advanced techniques such as forecasting and scenario modelling.

Notably, the dataset supports a dual focus: first, it prepares students to extract and interpret insights relevant to **strategic healthcare management**, and second, it provides a foundation for developing **AI-assisted business applications**, such as predictive models and decision support systems. The integration of **derived variables** such as stockout impact scores and categorised inventory risk levels further enhances its analytical depth, enabling more sophisticated interpretations and business recommendations.

From a pedagogical perspective, the dataset encourages experiential learning by simulating the real-life tasks of a senior healthcare administrator. For instance, students may be asked to present a dashboard to senior health officials showing trends in drug stock-outs, explain how pharmacist staffing affects patient service levels, or forecast the impact of clinical pharmacy services on cost savings. Each of these exercises promotes the development of essential skills in critical thinking, data storytelling, and strategic planning.

In summary, this dataset provides a comprehensive, realistic, and analytically rich environment through which MBA students can apply business analytics techniques in a meaningful context. It bridges the gap between theoretical knowledge and real-world application, while also introducing students to the role of AI and data-driven insights in shaping the future of healthcare operations.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Description

The dataset consists of **1,000 records**, each representing a monthly snapshot from public hospitals and primary care facilities in South Sinai. Each record combines operational metrics (e.g., staffing, inventory levels), clinical service indicators (e.g., presence of a clinical pharmacy unit), and financial data (e.g., drug expenditure). The data structure is designed to reflect real-world variation and patterns while ensuring logical consistency and analytical integrity.

Purpose and Application

This dataset was created for MBA students pursuing business analytics with applications in healthcare and public administration. It enables:

- Strategic decision-making simulations for roles such as the Director of Pharmacy Administration.
- Visual analytics development in Power BI for reporting, monitoring, and policy formulation.
- AI-driven exploration of optimisation opportunities in pharmaceutical resource planning and clinical service deployment.

Dataset Structure

- **Total Records**: 1,000
- **Time Range**: Multi-year monthly data points (generated synthetically)
- **Facilities Covered**: 20 facilities (5 hospitals and 15 primary care clinics)
- **Variable Types**:
    - **Qualitative (Categorical)**: 8
    - **Quantitative (Numerical)**: 9
    - **Derived Variables**: 3

Qualitative Variables (Categorical)

1. **Facility_ID**: Unique identifier for each hospital (H1–H5) or primary care unit (PC1–PC15).
2. **Facility_Type**: Categorised as "Hospital" or "Primary Care".

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Therapeutic_Category**: Type of medication managed (Chronic, Emergency, Contraceptive, Vaccine, Antibiotic).
4. **Clinical_Pharmacy_Unit**: Indicates whether a facility operates a clinical pharmacy (Yes/No).
5. **Stockout_Severity**: Classified as None, Low, Moderate, or High.
6. **Region_Comparator**: Designates the geographic comparison group (South Sinai, North Sinai, Red Sea).
7. **Staff_Accreditation_Status**: Indicates whether pharmacy staff are officially accredited (Accredited/Not Accredited).
8. **Inventory_Coverage_Category**: Derived classification of inventory sufficiency (Low, Adequate, High).

Quantitative Variables (Numerical)

1. **Month**: Observation date, monthly format.
2. **Number_of_Pharmacists**: Total pharmacy staff employed at the facility that month.
3. **Pharmacist_to_Patient_Ratio**: Ratio of pharmacists to 1,000 patient visits.
4. **Stockouts_Count**: Total number of stock-out events experienced during the month.
5. **Avg_Stockout_Duration**: Average number of days drugs were out of stock.
6. **Total_Drug_Inventory_Value**: Monetary value of the available drug inventory in Egyptian pounds.
7. **Percent_Inventory_Coverage**: Proportion of monthly demand covered by available stock.
8. **Monthly_Drug_Expenditure**: Total pharmaceutical spending in that month.
9. **Monthly_Patient_Visits**: Number of patients served by the facility during the month.

Derived Variables (Computed from Base Variables)

1. **High_Stockout_Flag**: Boolean flag (True/False) for cases where stock-out duration exceeds 10 days.
2. **Stockout_Impact_Score**: Composite indicator calculated as Stockouts_Count multiplied by Avg_Stockout_Duration.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Inventory_Coverage_Category**: Categorisation of Percent_Inventory_Coverage into Low (<90), Adequate (90–150), or High (>150).

Relevance to Power BI and AI for Business

The dataset is designed to seamlessly integrate into **Power BI**, supporting:

- Rich, interactive visualisations
- Drill-down reporting by region, facility, and severity level
- Time-series performance tracking
- KPI dashboards for pharmacy leadership

From an **AI for Business** perspective, the dataset includes structures that support:

- Predictive modelling of stock-outs
- Classification of high-risk facilities
- Clustering of service utilisation profiles
- Prescriptive insights for staff deployment or inventory budgeting

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify patterns in stock-outs across facility types and regions
- Correlation between accreditation and cost savings
- Distributions of patient volume and inventory levels

B. Power BI Visualisation

- Build dashboards comparing hospitals and clinics
- Monitor seasonal variation in drug stock availability
- Visualise staffing efficiency and financial indicators over time

C. AI-Driven Business Insights

- Predict facilities at risk of high stock-out impact
- Optimise pharmacist deployment based on service demand
- Recommend inventory coverage strategies for specific therapeutic categories

© 2025 Dr. Yasser Aly Yakout Shehata

Conclusion

This dataset provides a rich foundation for exploring key concepts in **business analytics**, **healthcare management**, and **artificial intelligence**. It enables MBA students to bridge academic knowledge with practical applications by simulating the decision-making environment of a Director of Pharmacy Administration. With a combination of realistic scenarios, derived insights, and compatibility with modern analytical tools such as Power BI, the dataset serves as a valuable resource for cultivating data-literate business leaders prepared to innovate in healthcare and beyond.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 21: CIB Corporate Credit Risk

Introduction

In contemporary banking practice, decision-making is increasingly shaped by the availability of high-quality data and the ability to derive actionable insights from it. This dataset has been meticulously constructed to simulate the working environment of a **Corporate Banking Senior Credit Officer**, a strategic role situated at the heart of credit risk evaluation, corporate client relationship management, and regulatory compliance. The Senior Credit Officer's responsibilities are multifaceted, encompassing assessment of large-scale credit exposures, monitoring of financial deterioration signals, and ensuring that credit decisions are aligned with both internal risk frameworks and external standards, particularly those stipulated under **IFRS 9**.

This dataset has been developed to support the **MBA course in Business Analytics for Decision Making**, which emphasises the application of data analytics in real-world strategic contexts. It allows students to step into the analytical shoes of a banking professional and tackle the types of questions that arise in the day-to-day management of corporate credit portfolios. From quantifying Expected Credit Losses to evaluating the impact of macroeconomic volatility on borrower performance, the dataset presents rich opportunities for practical learning and business problem-solving.

Furthermore, the dataset bridges **descriptive analytics**, **exploratory data analysis (EDA)**, and **AI-enhanced modelling**, enabling users to explore and visualise patterns in data, understand relationships among variables, and simulate advanced predictive scenarios. It integrates both **qualitative variables**, such as sectoral classification and risk stage, and **quantitative financial metrics**, such as Probability of Default (PD), Exposure at Default (EAD), and Financed Emissions, to reflect the multifactorial nature of credit assessment.

What makes this dataset particularly valuable is its incorporation of **derived and composite variables**, such as Collateral Coverage Ratios and Stress-ECL Shifts, which introduce depth and realism into the analysis. These fields enable students to investigate not only what the credit risk is, but also **why** it exists, **how** it evolves, and **what** strategic decisions might mitigate it.

Ultimately, the dataset is not just an academic artefact; it is a pedagogical simulation of a critical business function. By engaging with this dataset, MBA students will enhance their ability to interpret complex financial data, construct meaningful visualisations using tools like Power BI, and formulate AI-supported insights that align with institutional goals. In doing so, they gain both technical proficiency and domain-specific expertise that are essential for careers in finance, analytics, and strategic decision-making.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Description

The dataset comprises 1,000 records representing corporate borrower profiles at specific points in time. It captures a wide range of variables that influence credit risk decisions, structured into qualitative, quantitative, and derived fields. These variables are sourced from realistic financial assumptions and regulatory frameworks applicable to corporate banking operations.

Purpose and Application

The dataset has been curated with the following objectives:

- To enable **MBA students** to conduct realistic and context-specific **descriptive and exploratory analysis**
- To support the **development of dashboards** and visual reports using **Power BI**
- To allow students to build and test **predictive models** that inform business decision-making
- To simulate tasks carried out by a **Corporate Banking Senior Credit Officer**, such as risk assessment, credit decisioning, and sensitivity analysis

It also aligns with broader themes in business education, including the application of AI in decision support systems, the role of environmental risk in financial analysis, and data-informed regulatory compliance.

Dataset Structure

- **Number of Records**: 1,000
- **Observation Unit**: Borrower-level credit snapshot per time period
- **Total Variables**: 20
    - **7 Qualitative (Categorical) Variables**
    - **10 Quantitative (Numerical) Variables**
    - **3 Derived and Enhanced Variables**

Qualitative Variables

1. **Sector**: The industry classification of the borrowing entity (e.g. Energy, Manufacturing, Real Estate).
2. **Credit Rating Category**: Internal credit risk assessment label (Low, Moderate, High).

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Stage of Credit**: Credit impairment classification under IFRS 9 (Stage 1, Stage 2, Stage 3).
4. **Has Structured Deal**: Indicates the presence of structured or syndicated lending (Yes, No).
5. **SICR Flag**: Signals a Significant Increase in Credit Risk based on internal triggers (Yes, No).
6. **Restructured**: Denotes if the borrower's facility has undergone contractual restructuring (Yes, No).
7. **Climate Risk Exposure**: Qualitative categorisation of the borrower's vulnerability to environmental or climate-related risks (Low, Medium, High).

Quantitative Variables

1. **Time Period**: The timestamp or date of the observation (monthly or quarterly).
2. **Exposure at Default (EAD)**: The gross exposure amount outstanding, at risk of default.
3. **Probability of Default (PD)**: The likelihood, expressed as a proportion, that the borrower will default.
4. **Loss Given Default (LGD)**: The percentage of exposure expected to be lost if default occurs.
5. **Expected Credit Loss (ECL)**: The calculated expected loss based on PD, LGD, and EAD.
6. **Days Past Due**: The number of days the payment is overdue at the time of observation.
7. **Macroeconomic Index**: A simulated economic indicator representing external macro conditions (e.g. GDP, interest rate levels).
8. **Change in ECL Under Stress**: The projected ECL value after applying macroeconomic stress assumptions.
9. **Collateral Value**: Estimated value of recoverable collateral pledged against the facility.
10. **Financed Emissions**: A measure of environmental impact based on carbon exposure attributable to financing.

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Variables

1. **PD × LGD Product**: A simple risk interaction metric used for segmentation and risk ranking.
2. **Collateral Coverage Ratio**: A ratio indicating the level of protection provided by collateral relative to exposure.
3. **Stress ECL Percentage Change**: The percentage increase in ECL under macroeconomic stress, used to assess sensitivity.
4. **Risk Classification**: Categorical segmentation into High, Moderate, or Low risk based on thresholds for PD and ECL.
5. **Undercollateralized Flag**: Boolean indicator for facilities where collateral is deemed insufficient (< 0.8 coverage ratio).

Relevance to Power BI and AI for Business

This dataset is well-suited for Power BI dashboards, supporting layered analysis through slicers, trend visualisation, and conditional formatting. Students can build interactive reports that track key metrics over time, visualise borrower risk across sectors, and simulate credit deterioration under adverse economic scenarios.

In AI applications, the dataset supports training and validation of classification and regression models that predict:

- Movement between credit risk stages
- Future ECL under different scenarios
- Flags for undercollateralisation or SICR

Its structure enables supervised learning, correlation analysis, and feature engineering necessary for business AI models.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify sectors with the highest ECL values
- Analyse distribution of PD and LGD across credit rating groups
- Detect outliers in collateral valuation and financed emissions

© 2025 Dr. Yasser Aly Yakout Shehata

B. Power BI Visualisation

- Construct dashboards displaying portfolio risk health by time period
- Develop risk heatmaps filtered by SICR, IFRS stage, and sector
- Visualise sensitivity of ECL to macroeconomic conditions

C. AI-Driven Business Insights

- Build classification models to predict SICR or downgrade to Stage 2/3
- Use regression to forecast credit loss or emission-based lending risk
- Apply clustering to segment borrowers by risk and collateral profiles

Conclusion

The **Enhanced Corporate Credit Risk Dataset** offers a realistic, structured, and analytically rich simulation of the data environment faced by a Corporate Banking Senior Credit Officer. Designed specifically for MBA students, it encourages the practical application of business analytics tools, promotes understanding of financial risk frameworks, and fosters the development of data-driven decision-making competencies. By enabling students to work with complex, interrelated financial data, the dataset serves as an effective bridge between academic study and professional banking practice, particularly in the age of AI and data-centric strategy.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 22: Ismailia Healthcare IT Transformation

Introduction

In an era marked by rapid digitalisation and rising demand for efficient, high-quality healthcare services, the role of digital transformation in public health institutions has grown from a peripheral concern to a core strategic function. Nowhere is this more evident than in the work undertaken by national health authorities aiming to modernise service delivery through the integration of advanced information technologies. The **Egypt Healthcare Authority**, particularly through its regional branches such as the one in **Ismailia**, has become a pivotal site for piloting and implementing digital health innovations. Central to these efforts is the **Head of IT and Digital Transformation**, a role that requires both technical expertise and strategic vision to steer institutional change, enhance operational workflows, and ensure alignment with national health objectives.

The **Healthcare IT Transformation Dataset** has been purposefully designed to support graduate-level education in business analytics by simulating the analytical environment of this strategic leadership role. This synthetic dataset presents a structured yet realistic depiction of monthly digital performance across hospitals and primary care centres under the jurisdiction of the Ismailia branch. It captures key transformation metrics such as the implementation status of electronic health records (EHRs), utilisation of telemedicine services, integration of enterprise resource planning (ERP) modules, and deployment of artificial intelligence for diagnostic purposes.

The dataset provides MBA students, particularly those undertaking the course ***Business Analytics for Decision Making***, with a practical tool to understand how digital technologies are adopted and scaled in complex healthcare ecosystems. It facilitates the exploration of how different organisational characteristics such as facility type, system adoption phase, or compliance with international standards correlate with digital performance outcomes. Furthermore, by incorporating derived variables such as EHR penetration rate, digital efficiency, and AI usage intensity, the dataset allows for multi-layered analysis that reflects the depth of insight expected at the strategic level.

This introduction sets the foundation for a sophisticated analytical exercise in which students are not merely consumers of data, but decision-makers in training. Through this dataset, they are encouraged to analyse, visualise, and derive strategic recommendations that could realistically inform the decisions of a Head of IT and Digital Transformation. Whether applied through Power BI dashboards, AI-enabled models, or descriptive analytics, the dataset empowers learners to bridge academic concepts with real-world digital health transformation practices.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Description

This dataset comprises **1,000 monthly observations** drawn from eight representative facilities (four hospitals and four primary care centres) across a three-year period from **January 2022 to December 2024**. Each record encapsulates a facility's digital adoption status, operational output, and key transformation indicators. Variables include both **qualitative attributes** and **quantitative performance metrics**, supplemented with analytically rich **derived variables**.

The dataset offers a comprehensive simulation of the kind of data routinely monitored by digital transformation leaders within healthcare systems, and is tailored for exploratory, visual, and predictive analysis.

Purpose and Application

The primary purpose of this dataset is to enable students to:

- Explore and evaluate the performance of digital transformation initiatives across healthcare settings
- Practise exploratory data analysis (EDA) to detect trends, anomalies, and correlations
- Construct interactive dashboards using Power BI to support decision-making
- Apply business analytics and AI techniques to generate operational insights and forecast outcomes

Applications include performance benchmarking, risk identification, digital maturity assessment, and strategy formulation based on data evidence.

Dataset Structure

- **Records**: 1,000 facility-month observations
- **Time Frame**: January 2022 – December 2024
- **Entities**: 8 healthcare facilities
- **Variables**: 7 qualitative, 10 quantitative, 3 derived

Qualitative Variables (7)

1. **Facility_ID**: Unique code identifying each facility (e.g. HS01, PC03)
2. **Facility_Type**: Type of healthcare facility (Hospital or Primary Care)

© 2025 Dr. Yasser Aly Yakout Shehata

3. **EHR_Implementation_Status**: Stage of electronic health record deployment (Not started, Pilot, Live, Optimised)
4. **Telemedicine_Enabled**: Indicates whether telemedicine services are active (Yes or No)
5. **ERP_Module_Adoption**: Degree of enterprise resource platform adoption (None, Finance, HR, Full-stack)
6. **AI_Diagnostic_Module**: Status of AI diagnostic tools in use (None, Partial, Full)
7. **ICD11_Compliant**: Denotes whether the facility uses the ICD-11 coding standard (Yes or No)

Quantitative Variables (10)

8. **Month**: Date of observation, formatted to the first day of each month
9. **EHR_Records_Created**: Number of new electronic health records generated in the month
10. **E_Prescriptions_Issued**: Number of digital prescriptions issued
11. **Telemedicine_Sessions**: Total number of remote consultations conducted
12. **Radiology_Images_Archived**: Count of images stored in the digital archive
13. **ERP_Transactions**: Total ERP transactions performed (e.g. HR, finance)
14. **RCM_Cycle_Time**: Average time (in days) to complete a revenue cycle
15. **Dashboard_KPI_Count**: Number of KPIs displayed on real-time dashboards
16. **AI_Diagnostics_Processed**: Number of diagnostic cases analysed by AI systems
17. **Staff_Time_Saved_Hours**: Estimated hours saved through digital efficiency mechanisms

Derived Variables (3)

18. **EHR_Penetration_Rate**: Proportion of EHRs created relative to estimated patient visits
19. **Digital_Efficiency_Factor**: Staff time saved per ERP transaction, representing digital workflow productivity
20. **AI_Usage_Intensity**: Ratio of AI diagnostics to telemedicine sessions, indicating integration depth of AI into digital services

© 2025 Dr. Yasser Aly Yakout Shehata

Relevance to Power BI and AI for Business

This dataset is highly suited for Power BI and AI applications due to its:

- Temporal structure, which supports time-series trend visualisations
- Categorical filters for slicing dashboards by facility, digital maturity, or AI adoption
- Derived metrics that facilitate benchmarking, efficiency analysis, and resource planning
- Built-in complexity for practising storytelling, visual analytics, and strategic insight development

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Examine monthly and yearly trends in digital records and AI usage
- Analyse the relationship between ERP transactions and time savings
- Explore performance differences across facility types and digital maturity stages

B. Power BI Visualisation

- Construct dashboards with KPIs for EHR usage, AI penetration, and ERP efficiency
- Use filters to compare hospitals versus primary care centres
- Apply conditional formatting to flag underperforming or high-performing sites

C. AI-Driven Business Insights

- Train regression models to predict staff time savings from ERP activity
- Develop classification models to anticipate AI adoption levels based on current metrics
- Segment facilities into performance clusters for targeted strategic interventions

Conclusion

The **Ismailia Healthcare IT Transformation Dataset** offers a robust platform for simulating real-world decision-making by a Head of IT and Digital Transformation in a public health authority. Through a well-rounded combination of descriptive, categorical, and derived variables, it enables MBA students to conduct meaningful analysis that bridges theory and application. Whether visualising adoption trends, exploring the efficiency of ERP systems, or

© 2025 Dr. Yasser Aly Yakout Shehata

forecasting AI expansion, this dataset offers rich opportunities for skill development in Power BI, business analytics, and AI for healthcare strategy.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 23: Healthcare Inventory El-Fayrouz Medical Complex

Introduction

Healthcare systems globally face increasing complexity in ensuring the uninterrupted availability of medical supplies, consumables, and equipment, particularly in decentralised and resource-sensitive regions. Within this context, the role of the **Stock Controller** emerges as a strategic function, bridging operational efficiency, clinical safety, and financial stewardship. In the South Sinai region of Egypt, the El-Fayrouz Medical Complex, governed by the **General Authority of Healthcare**, represents one of the key institutions navigating these supply chain challenges. The Stock Controller at this complex is tasked not only with managing day-to-day inventory movements but also with safeguarding against critical stockouts, mitigating overstocking, monitoring expiry risks, and ensuring compliance with reporting and documentation protocols.

As Egypt's healthcare sector accelerates its digital transformation agenda, supported by national frameworks and regional modernisation programmes, the integration of data analytics into healthcare logistics becomes imperative. Stock controllers are now expected to harness structured data and employ analytical tools to optimise inventory flow, improve coordination with suppliers, and align procurement with clinical demand patterns. In this light, developing business analytics capabilities among future healthcare leaders, particularly MBA students, is essential for sustainable system improvement.

The **Healthcare Inventory El-Fayrouz Medical Complex Dataset** has been designed to serve this purpose. It offers a structured and operationally realistic representation of inventory data reflective of the environment in which the Stock Controller at El-Fayrouz Medical Complex operates. Developed specifically for MBA students enrolled in the ***Business Analytics for Decision Making*** **course**, the dataset functions as a practical sandbox for applying analytical techniques across domains such as Exploratory Data Analysis (EDA), Power BI visualisation, and artificial intelligence in business settings.

This dataset not only encapsulates historical supply and consumption records but also integrates calculated risk indicators such as stockout flags and waste ratios. These features allow learners to simulate real decision-making scenarios, evaluate supplier performance, identify operational inefficiencies, and develop data-driven strategies to improve healthcare logistics. It thus facilitates the convergence of analytical theory and practical application, preparing students to take on data-oriented leadership roles within complex, high-stakes environments such as public healthcare.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Description

This dataset comprises 1,000 monthly inventory records from January 2022 through December 2024. Each entry represents a transactional snapshot of a healthcare department's inventory state, including stock levels, supply inflows and outflows, waste, and supplier performance indicators. In doing so, it reflects the complexity and variability of real-time inventory control within a public medical facility.

The dataset contains:

- **7 Qualitative Variables** describing operational context
- **10 Quantitative Variables** measuring consumption, storage, and logistics
- **3 Derived Variables** to enhance decision-making depth

Purpose and Application

This dataset supports practical learning across three primary domains:

- **Descriptive Analytics**: Summarising and explaining historical inventory performance
- **Power BI Visualisation**: Developing interactive dashboards to monitor stock trends, expiry risks, and supplier efficiency
- **AI-Driven Insights**: Training predictive models to forecast stockouts, overstocking likelihood, or waste ratios

It is ideal for coursework, capstone projects, and simulated decision-making exercises at the intersection of business analytics and healthcare logistics.

Dataset Structure

Each row represents a department-month record of inventory management activities. The dataset structure allows for aggregation by time, department, or item type, enabling comparative analysis and visual storytelling.

Qualitative Variables

1. **Item_Category**: Type of supply (Medicine, Medical Consumable, Durable Equipment)
2. **Department**: Department responsible for inventory usage (Pharmacy, Surgery, Emergency, Outpatient, ICU)
3. **Supplier**: Vendor or company delivering the item (e.g. PharmaPlus, HealthLine)

© 2025 Dr. Yasser Aly Yakout Shehata

4. **E_Inventory_Logged**: Whether the inventory was electronically recorded (Yes or No)
5. **Reports_Timely**: Indicates if monthly reports were submitted on time (Yes or No)
6. **Month**: The reporting month for each record
7. **Record_ID**: Unique identifier for each entry

Quantitative Variables

8. **Stock_Level**: Total number of units in stock at the end of the month
9. **Min_Threshold**: Minimum acceptable inventory level to avoid stockout
10. **Max_Threshold**: Upper acceptable inventory level to avoid overstocking
11. **Units_Consumed**: Number of units used during the month
12. **Units_Received**: Number of units added through deliveries
13. **Expiry_Months_Remaining**: Months remaining before the batch expires
14. **Expired_Units**: Quantity of stock expired during the month
15. **Disposal_Units**: Quantity of stock disposed due to damage or protocol
16. **Temp_Excursions**: Number of temperature violations affecting stored items
17. **Lead_Time_Days**: Time (in days) between order and delivery
18. **Shortage_Incidents**: Count of stockout episodes during the month
19. **Count_Accuracy_Percent**: Accuracy of physical stock count vs system records

Derived Variables

20. **Stockout_Flag**: Boolean flag indicating whether stock fell below minimum threshold
21. **Overstock_Flag**: Boolean flag indicating whether stock exceeded the maximum threshold
22. **Waste_Percentage**: Combined expired and disposed units as a percentage of total stock

Relevance to Power BI and AI for Business

This dataset is structured to facilitate interactive and automated analysis. Its time-series structure, categorical groupings, and logically coherent numerical values allow:

- Effective filtering, grouping, and aggregation in Power BI

© 2025 Dr. Yasser Aly Yakout Shehata

- KPI tracking with slicers and performance dashboards
- Development of AI models for classification and regression (e.g., forecasting stockout risk based on historical trends)

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

Students can investigate:

- Seasonal consumption trends
- Supplier lead time variances
- Accuracy and reliability of physical counts

B. Power BI Visualisation

Possible visual outputs include:

- Stockout heatmaps by department
- Time-based line graphs of inventory levels
- Bar charts comparing supplier reliability

C. AI-Driven Business Insights

Model applications include:

- Predicting likelihood of stockout events
- Estimating waste probability based on expiry and storage factors
- Clustering suppliers or departments by performance metrics

Conclusion

The **Healthcare Inventory El-Fayrouz Medical Complex Dataset** provides an academically rich and practically grounded resource for business analytics education. It offers students the opportunity to translate theory into practice, mimic executive decision-making, and derive strategic insights from complex data environments. With its realistic structure and multifaceted variables, the dataset acts as a learning bridge between data analysis, visual communication, and real-world healthcare management.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 24: Pharmacist Supervisor Analytics South Sinai

Empowering Data-Driven Decision Making in Healthcare Operations

1. Introduction

The modern healthcare landscape demands rapid, informed decision-making driven by data. In this context, the role of the **Pharmacist Supervisor** has evolved far beyond managing stockrooms and handling medication requests. Today, it encompasses a wide array of strategic responsibilities ranging from inventory optimization and staff training to regulatory compliance, digital reporting, and cost control. These expanded responsibilities make the Pharmacist Supervisor a critical node in healthcare logistics and patient safety.

This dataset was developed as part of a final MBA project under the course **Business Analytics for Decision Making**, simulating real-world conditions faced by a Pharmacist Supervisor operating within the **Egypt Healthcare Authority (EHA), South Sinai Branch**. The dataset is inspired by actual operational structures and decision-making frameworks described on official platforms and national policy documents governing medication management, audit, and pharmaceutical care across regional health institutions in Egypt.

The core purpose of this dataset is to **bridge academic learning with real-world application**. It provides MBA students with a sandbox environment where they can apply analytical techniques and digital tools to solve practical business challenges in healthcare. From ensuring drug availability to improving compliance with national standards, each variable in the dataset reflects a specific operational concern that Pharmacist Supervisors navigate daily.

Moreover, the dataset is designed to foster **multi-layered learning** by integrating a combination of:

- Operational metrics (such as stock levels, reallocation volumes, and consumption rates)
- Behavioral indicators (like training hours and digital reporting activity)
- Compliance data (such as audit scores and regulatory status)
- Derived insights (including calculated risk flags and efficiency bands)

By focusing on **Exploratory Data Analysis (EDA)** and **Visualization in Power BI**, the dataset encourages students to interact with data meaningfully transforming rows of figures into patterns, dashboards, and insights. It also serves as a launchpad for **AI-driven applications in business**, such as predicting supply chain disruptions, identifying non-compliance trends, or recommending targeted training interventions.

© 2025 Dr. Yasser Aly Yakout Shehata

In summary, this dataset presents a unique, hands-on opportunity for MBA students to step into the role of a data-driven Pharmacist Supervisor, making informed decisions that affect both operational efficiency and public health outcomes. It encourages not just technical fluency, but also critical thinking, ethical awareness, and strategic foresight, all essential skills for leaders in the AI-powered, data-driven economy.

2. Dataset Description

The dataset includes **1,000 records**, each representing one monthly operational snapshot from a healthcare facility. It incorporates **7 qualitative**, **10 quantitative**, and **4 derived variables**, offering rich ground for segmentation, benchmarking, and forecasting.

It provides insight into the Pharmacist Supervisor's role in managing risks, costs, training initiatives, and operational continuity across varied facilities and compliance environments.

3. Purpose and Application

This dataset is tailored for:

- **Teaching exploratory and visual analytics** using Power BI.
- **Applying AI-driven decision-making tools** in healthcare logistics.
- **Training MBA students** on real-world, data-supported problem-solving.
- **Enhancing competency in pharmaceutical inventory management, policy compliance, and resource optimization.**

4. Dataset Structure

- **Total Records**: 1,000
- **Variables**: 21
    - 7 Qualitative
    - 10 Quantitative
    - 4 Derived for deeper analysis

5. Qualitative Variables

1. **Facility_Type** – Type of healthcare institution: Hospital, Clinic, or Medical Center.
2. **Branch** – Facility location: Sharm El-Sheikh, Dahab, Nuweiba, or Taba.
3. **Regulation_Compliance** – Indicates whether the facility met regulatory standards during the month.

© 2025 Dr. Yasser Aly Yakout Shehata

4. **Training_Type** – The nature of professional training completed: GAHAR, EDA Antimicrobial, Clinical Pharmacy ICU, or Oncology.
5. **Inventory_Tool_Used** – System used for digital stock reporting: Excel, Google Sheets, or Custom Software.
6. **Supply_Shortage_Response** – Strategy applied when facing stock shortages: Reorder, Reallocate, or Emergency Order.
7. **Accreditation_Status** – Facility's national accreditation status: Pre-Accreditation, Accredited, or At Risk.

6. Quantitative Variables

8. **Stock_Level** – Total inventory units available.
9. **Avg_Daily_Consumption** – Average daily usage of medications.
10. **Days_To_Expiry** – Shelf life remaining on stocked medication, in days.
11. **Stockout_Events** – Number of times stock reached zero.
12. **Monthly_Training_Hours** – Total hours spent on staff development.
13. **Audit_Score** – Percentage compliance score from internal and external reviews.
14. **Complaints_Reported** – Number of reported medication errors or incidents.
15. **Reallocated_Units** – Number of units moved across facilities.
16. **Digital_Reporting_Entries** – Log count from digital inventory systems.
17. **Medication_Cost_Saved** – Estimated financial savings through smarter inventory or training actions.

7. Derived Variables

18. **Stock_Coverage_Days** – How long stock would last at current consumption levels.
19. **Critical_Inventory_Risk** – Flags urgent risk if stock is low or expiry is near.
20. **Audit_Score_Band** – Categorizes audit scores into performance bands: Low, Medium, High, Excellent.
21. **Training_Hours_Level** – Groups training efforts into levels: Very Low to High.

8. Relevance to Power BI and AI for Business

This dataset is structured for intuitive use in Power BI, allowing:

© 2025 Dr. Yasser Aly Yakout Shehata

- Quick **pivoting, filtering, and dashboarding**.
- **Interactive storytelling** through risk, compliance, and cost-saving visuals.
- Integration into **machine learning models**, especially for predictive risk alerts or performance segmentation.

It also allows AI applications like trend detection, anomaly spotting, and natural language generation of executive summaries.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Examine patterns in stockout events across facility types.
- Analyze cost savings versus training intensity or audit score.
- Study geographic differences in compliance and performance.

B. Power BI Visualization

- Create dashboards comparing training types with compliance scores.
- Use heatmaps to flag high-risk locations.
- Visualize stock coverage and expiry trends using KPI cards and slicers.

C. AI-Driven Business Insights

- Predict critical inventory risks based on past performance.
- Use clustering to profile facilities by risk, training, and digital maturity.
- Forecast medication cost savings by training and tool usage patterns.

10. Conclusion

The **Pharmacist Supervisor Analytics South Sinai Dataset** provides a versatile, high-fidelity simulation of healthcare supply chain and compliance decision-making. It empowers MBA students to practice the full cycle of business analytics from cleaning and exploration to visual storytelling and AI-enhanced forecasting.

It bridges classroom learning with real-world challenges in public health, and stands as a valuable resource for students, instructors, and future decision-makers in healthcare operations and business intelligence.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 25: Dental HIS Transformation South Sinai Dataset

Driving Clinical Excellence and Digital Innovation in Healthcare

1. Introduction

In an era where healthcare systems are increasingly shaped by data and technology, clinical professionals are being called upon not only to treat patients but also to lead innovation. The convergence of medicine, information systems, and strategic decision-making has given rise to a new hybrid role in the healthcare landscape: the clinician-technologist. One such role is the **Dental Surgeon who also serves as a Health Information Systems and Digital Transformation Specialist**, a professional who embodies both clinical expertise and a vision for digital innovation.

This dataset was developed to simulate the analytical and operational responsibilities of such a professional within the **Egypt Healthcare Authority**, particularly in the **South Sinai** region, where significant investments in digital dental labs and electronic health infrastructure are underway. The Dental Surgeon in this context is tasked not only with performing procedures and consultations but also with leading digital adoption, managing system performance, analyzing patient feedback, and translating data into better healthcare outcomes.

As part of a final project for MBA students enrolled in the **Business Analytics for Decision Making** course, this dataset offers a real-world learning experience grounded in public health transformation. It mirrors the complexity of managing digital tools like EMRs, PACS, and AI dashboards while tracking operational metrics such as case volumes, consultation times, lab turnaround, and patient satisfaction.

Unlike traditional clinical datasets that focus solely on patient records, this dataset integrates a wide spectrum of variables that reflect the broader responsibilities of a digital healthcare leader. These include training status, interoperability, system uptime, dashboard usage, and derived variables such as digital maturity and efficiency ratios. This structure enables students to not only perform descriptive analytics but also develop strategic insights, build visual dashboards in Power BI, and apply AI methods to forecast risk or segment digital maturity.

This dataset serves as a bridge between theory and practice, enabling MBA students to step into the role of a data-driven decision-maker within a digital healthcare ecosystem. It fosters critical thinking around healthcare delivery, resource optimization, user behavior, and performance management, skills that are increasingly vital for future leaders in healthcare and beyond.

© 2025 Dr. Yasser Aly Yakout Shehata

Through hands-on analysis, students are encouraged to uncover patterns, propose data-backed improvements, and understand how digital transformation can elevate both clinical practice and institutional performance. In doing so, they prepare to contribute meaningfully to the evolving landscape of smart healthcare systems.

2. Dataset Description

The **Dental HIS Transformation South Sinai Dataset** contains **1,000 records**, each representing a monthly observation from a dental or healthcare facility. The dataset balances clinical, operational, and digital performance dimensions to simulate the analytical environment in which a Dental Surgeon–Digital Transformation Specialist operates.

The data was synthetically generated but follows realistic distributions and inter-variable relationships to support educational use cases. It is fully cleaned, enhanced, and optimized for use in Power BI, Excel, Python, and other analytics platforms.

3. Purpose and Application

The dataset supports a variety of learning and strategic development objectives:

- Empower MBA students to practice analytical techniques on healthcare transformation data.
- Enable the creation of dashboards and visual summaries to support executive decision-making.
- Facilitate AI-based modeling to detect patterns, risks, and improvement opportunities.
- Illustrate the impact of digital maturity, training, and interoperability on patient satisfaction and operational performance.

The dataset is particularly useful in courses or modules on:

- Business Analytics for Decision Making
- Data-Driven Healthcare Strategy
- Digital Transformation in Public Services
- Applied AI in Healthcare

© 2025 Dr. Yasser Aly Yakout Shehata

4. Dataset Structure

Item	Count
Total Records	1,000
Qualitative Variables	7
Quantitative Variables	14
Derived Variables	3
Total Variables	24

5. Qualitative Variables

1. **Facility_Type** – Type of healthcare facility: Hospital, Clinic, or Dental Center.
2. **Region** – Facility location: Sharm El-Sheikh, Dahab, Nuweiba, or Taba.
3. **HIS_Module_Used** – Active health IT module: EMR, PACS, Tele-dentistry, or Lab-Integration.
4. **Digital_Training_Status** – Whether staff completed digital system training: Completed or Incomplete.
5. **AI_Dashboard_Adoption** – Indicates use of AI-powered analytics dashboards: Yes or No.
6. **Data_Interoperability** – Whether the system exchanges data across platforms: Yes or No.
7. **Clinical_Workflow_Type** – Patient interaction mode: In-person, Remote, or Hybrid.

6. Quantitative Variables

8. **Dental_Cases_Handled** – Number of dental procedures conducted per month.
9. **TeleDentistry_Sessions** – Count of remote consultations.
10. **EMR_Entries** – Number of digital patient records created.
11. **Imaging_Studies** – Diagnostic imaging requests uploaded to the system.
12. **Average_Consultation_Time** – Average time (in minutes) spent per consultation.
13. **Turnaround_Time_Lab** – Time (in days) from lab order to result delivery.

© 2025 Dr. Yasser Aly Yakout Shehata

14. **User_Login_Frequency** – Number of logins by staff into HIS modules.
15. **Cases_per_Login** – Efficiency metric: cases handled per system login.
16. **Dashboard_Queries_Executed** – Number of analytical queries run on dashboards.
17. **System_Uptime** – HIS system availability percentage.
18. **Patient_Satisfaction_Score** – Average patient satisfaction on a scale from 0 to 100.
19. **Remote_Care_Percentage** – Percentage of dental care delivered remotely.
20. **Digital_Maturity_Index** – Composite score (0–5) based on training, dashboard use, interoperability, and other metrics.

7. Derived Variables

21. **Consultation_Time_Band** – Categorized consultation durations: Short, Moderate, or Extended.
22. **System_Uptime_Level** – Classified uptime levels: Unstable, Fair, Stable, or Excellent.
23. **Satisfaction_Level** – Grouped satisfaction scores: Low, Medium, High, or Excellent.
24. **Risk_Flag** – Binary indicator (Yes or No) triggered by low uptime, long lab turnaround, or low satisfaction.

8. Relevance to Power BI and AI for Business

This dataset is specifically designed for seamless use in **Power BI**, where students can create:

- Slicers and filters by region, facility, or workflow.
- KPIs for system uptime, satisfaction, and lab turnaround.
- Heatmaps for risk flags and maturity profiles.
- Trend lines and comparisons of digital transformation progress.

In AI contexts, the dataset supports classification, clustering, and regression tasks. Students can predict patient satisfaction, model digital maturity, or detect early signals of operational risk using machine learning tools.

© 2025 Dr. Yasser Aly Yakout Shehata

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify patterns across training, dashboard usage, and performance.
- Compare satisfaction levels across digital maturity segments.
- Assess the impact of HIS module use on efficiency and turnaround.

B. Power BI Visualization

- Build dynamic dashboards with drilldowns by workflow and region.
- Create scorecards for system uptime and satisfaction levels.
- Use bar and line charts to visualize monthly changes and trends.

C. AI-Driven Business Insights

- Predict low satisfaction using lab turnaround and consultation time.
- Use decision trees to classify risk conditions.
- Cluster facilities by interoperability and maturity index to guide strategic planning.

10. Conclusion

The **Dental HIS Transformation South Sinai Dataset** offers a realistic, multidimensional resource for business analytics education. It reflects the operational and strategic concerns of a Dental Surgeon engaged in digital transformation, giving students the opportunity to work on a case grounded in real healthcare challenges.

By interacting with this dataset, MBA students gain hands-on experience in visual analytics, predictive modeling, and decision-making under uncertainty, skills essential for leading transformation initiatives in healthcare or any complex service environment.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 26: Substation Project Analytics (HV & EHV)

1. Introduction

In the global push toward modernizing energy infrastructure, electrical power transmission projects especially those involving High Voltage (HV) and Extra High Voltage (EHV) substations, stand at the forefront of national development and industrial resilience. These projects are not only technically complex but also critically dependent on timely execution, regulatory compliance, and strategic decision-making. As a result, the professionals tasked with overseeing such implementations must blend deep engineering knowledge with sharp analytical acumen.

One such professional is the **Electrical Projects Management Engineer**. Far more than a site supervisor or technical coordinator, this role represents a linchpin in the successful delivery of large-scale energy infrastructure. Electrical Projects Management Engineers oversee the entire lifecycle of substation projects: from reviewing and approving design documentation, to coordinating Factory Acceptance Tests (FAT) and Site Acceptance Tests (SAT), to tracking installation milestones and managing non-conformance reports. In this capacity, they act as both technical authorities and operational analysts navigating multi-vendor ecosystems, strict deadlines, evolving safety standards, and the complexities of public-sector infrastructure planning.

As this role increasingly intersects with data, the ability to extract actionable insights from project information becomes critical. Electrical Projects Management Engineers now rely on dashboards, test reports, progress logs, and performance indicators to guide decision-making, mitigate risk, and ensure accountability. This transformation aligns directly with the objectives of MBA programs that seek to produce leaders who can translate technical data into strategic advantage.

To support this pedagogical goal, the **Substation Project Analytics (HV & EHV) Dataset** has been developed as a realistic and structured resource for MBA students pursuing the **Business Analytics for Decision Making** course. The dataset simulates 1,000 records of monthly project activity across substations in multiple regions. It incorporates both operational and quality-related variables, such as the number of design drawings submitted, test outcomes, approval timelines, commissioning delays, and incident reports.

Beyond raw data, the dataset includes calculated fields and binned variables that represent performance tiers, delay classifications, and pass-rate evaluations. These derived variables allow students to go beyond description, enabling diagnostic and predictive analysis that mirrors the decision-making challenges faced by professionals in real energy infrastructure environments.

© 2025 Dr. Yasser Aly Yakout Shehata

This introduction of infrastructure analytics into the academic setting offers students a valuable opportunity to develop technical fluency in tools like Power BI and Python while also deepening their understanding of operational excellence, project risk, and stakeholder communication. Whether analyzing trends in testing outcomes, visualizing commissioning delays across regions, or modeling project risk flags, students are encouraged to think critically, act strategically, and deliver data-driven recommendations that could realistically improve substation project outcomes.

The *Substation Project Analytics (HV & EHV) Dataset* therefore serves not only as an educational exercise, but also as a bridge between classroom theory and real-world engineering practice equipping the next generation of managers with the skills to lead digital transformation in mission-critical infrastructure domains.

2. Dataset Description

The dataset consists of **1,000 records**, each representing a monthly snapshot of a substation component or project segment. These records capture project status indicators from design approval through testing and commissioning.

Each record integrates:

- Core operational data such as drawing submissions, test counts, and delay durations
- Quality and compliance signals such as failure counts, risk flags, and incident reports
- Derived fields designed to support advanced segmentation and predictive modeling

The dataset includes a mix of **qualitative**, **quantitative**, and **derived variables** to enable meaningful pattern recognition, visual storytelling, and data modeling.

3. Purpose and Application

This dataset was built to help MBA students:

- Practice exploratory data analysis (EDA) on real project management variables
- Build visual dashboards in Power BI for strategic decision support
- Train AI models to predict risk, delay, and quality outcomes
- Apply descriptive, diagnostic, and predictive analytics to engineering operations

It supports hands-on learning in domains such as:

- Infrastructure project analytics
- Operations management

© 2025 Dr. Yasser Aly Yakout Shehata

- Quality control
- Digital transformation in engineering

4. Dataset Structure

Attribute	Details
Total Records	1,000
Qualitative Variables	7
Quantitative Variables	10
Derived/Binned Variables	7
Total Variables	24

5. Qualitative Variables

1. **Project_Region** – Geographic area of the project (e.g., Cairo, Giza, Assiut, Hurghada)
2. **Substation_Type** – Voltage class of the substation: HV or EHV
3. **Component_Type** – Category of installed equipment (Transformer, Switchgear, etc.)
4. **Design_Status** – Status of design documentation: Approved, Under Review, Issued for Revision
5. **FAT_Result** – Result of Factory Acceptance Test: Pass, Fail, or Re-Test Required
6. **SAT_Result** – Result of Site Acceptance Test: Pass, Fail, or Re-Test Required
7. **Compliance_Status** – Overall compliance: On-Time, Delayed, At-Risk

6. Quantitative Variables

8. **Drawings_Submitted** – Number of engineering drawings submitted during the period
9. **Design_Approval_Days** – Days between drawing submission and design approval
10. **FAT_Count** – Total number of Factory Acceptance Tests conducted
11. **FAT_Failure_Count** – Number of FATs that failed
12. **SAT_Count** – Total number of Site Acceptance Tests executed
13. **SAT_Failure_Count** – Number of SATs that failed
14. **Commissioning_Delay_Days** – Days delayed beyond scheduled commissioning

© 2025 Dr. Yasser Aly Yakout Shehata

15. **Components_Installed** – Count of installed electrical components
16. **NonConformance_Reports** – Number of incident or quality reports filed
17. **Progress_Log_Entries** – Number of log entries made for project progress tracking

7. Derived and Binned Variables

18. **FAT_Pass_Rate** – Percentage of passed FATs out of total tests
19. **SAT_Pass_Rate** – Percentage of passed SATs out of total tests
20. **Risk_Flag** – Flags risk based on multiple conditions (Yes or No)
21. **Approval_Speed_Category** – Categorizes design approval speed: Fast, Moderate, or Slow
22. **Delay_Category** – Categorizes commissioning delay: On-Time, Slight Delay, or Major Delay
23. **FAT_Quality_Level** – Qualitative tiering of FAT pass rates: Low, Moderate, or High
24. **SAT_Quality_Level** – Qualitative tiering of SAT pass rates: Low, Moderate, or High

8. Relevance to Power BI and AI for Business

The dataset structure and variable composition make it especially well-suited for:

- **Power BI Dashboards**: Enables creation of KPI cards, trend graphs, heatmaps, and dynamic filters by region, component, or delay category
- **Segmentation Analysis**: Facilitates grouping by performance level, delay severity, or compliance status
- **AI Modeling**: Inputs support classification models to predict risk, pass rates, or probable project delay

Students can leverage this dataset to build interactive dashboards, develop machine learning pipelines, and simulate executive briefings based on real-time operational insights.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify correlations between delay and design approval speed
- Analyze failure rates across equipment types or substations
- Understand geographic patterns in project risk

© 2025 Dr. Yasser Aly Yakout Shehata

B. Power BI Visualization

- Build regional performance dashboards
- Create visual timelines of approval-to-commissioning
- Use filters to drill down into incident-prone project types

C. AI-Driven Business Insights

- Predict project risk using logistic regression or decision trees
- Cluster components or projects based on delay, quality, and compliance
- Recommend intervention strategies based on high-failure trends

10. Conclusion

The **Substation Project Analytics (HV & EHV)** Dataset is a powerful learning tool that integrates engineering project data with business intelligence techniques. It places students in a realistic role, simulating the operational landscape of an Electrical Projects Management Engineer working under regulatory, technical, and strategic pressures.

By exploring, visualizing, and modeling this dataset, MBA students gain essential experience in transforming raw infrastructure data into business-ready insights. This project not only enhances technical fluency with analytical tools but also deepens strategic thinking around performance, quality, and delivery in the context of large-scale infrastructure projects.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 27: Health Informatics Digital Transformation Analytics

1. Introduction

The global healthcare landscape is undergoing a profound transformation driven by advancements in digital technologies, data integration, and intelligent decision support systems. Across both developed and developing health systems, there is a growing emphasis on leveraging digital infrastructure to improve the efficiency, transparency, and responsiveness of care delivery. Within this context, the integration of electronic medical records (EMRs), diagnostic imaging systems, telemedicine platforms, and health analytics has become central to the strategic objectives of modern healthcare authorities.

In Egypt, the **Health Care Authority** has spearheaded national efforts toward digitalisation, particularly in regional healthcare branches such as **South Sinai**. Leading these efforts is the **Health Informatics and IT Director**, a role that demands a nuanced blend of clinical awareness, technological expertise, and strategic oversight. This individual is responsible for orchestrating the implementation and monitoring of digital systems across a wide network of facilities, ensuring that data flows efficiently between systems, that healthcare providers are equipped with the tools and training required to utilise them, and that patients benefit from more coordinated, informed, and timely care.

This dataset has been developed to simulate the analytical ecosystem in which such a Health Informatics and IT Director operates. It captures the critical data points used to monitor digital adoption, assess system uptime, evaluate staff engagement, and respond to technical support issues. These are the metrics upon which decisions are made regarding further investment in digital infrastructure, changes to training policies, or interventions to improve system usability and reliability.

Designed for use by MBA students enrolled in the **Business Analytics for Decision Making course**, this dataset serves as an applied learning resource that mirrors real-world complexity. It allows students to explore how quantitative and qualitative indicators interact, how derived metrics can be used to summarise performance, and how visualisation tools and AI methods can support executive decision-making in healthcare informatics. Beyond technical analysis, the dataset invites students to engage with the strategic dimension of data, asking critical questions about what the data implies for healthcare delivery, policy formulation, and the sustainability of digital health systems.

By engaging with this data, students are afforded a unique opportunity to not only refine their skills in exploratory data analysis, visualisation, and modelling, but also to appreciate the broader organisational and societal impact of health information technologies. As such, the

© 2025 Dr. Yasser Aly Yakout Shehata

dataset provides a robust foundation for integrating analytics with strategic thinking, policy development, and systems leadership within the healthcare sector.

2. Dataset Description

The **Health Informatics Digital Transformation Analytics** dataset contains **1,000 records**, each representing performance data from a healthcare facility within the South Sinai region. The data spans operational, technical, and behavioural indicators necessary for evaluating the success of digital health implementations. This multi-dimensional dataset captures input from electronic medical record systems, PACS platforms, support tickets, e-prescription systems, and user training programmes.

Variables are carefully structured to balance simplicity with analytical richness. Seven qualitative variables capture categorical attributes such as facility type and user sentiment, while ten quantitative variables describe continuous or count-based metrics such as adoption rates and consultation volumes. Five additional derived variables are included to simulate more complex business intelligence constructs such as risk indicators, efficiency scores, and performance tiers.

3. Purpose and Application

The primary purpose of this dataset is to provide MBA students with a realistic, analytically rich environment for applying techniques in business analytics, specifically within the healthcare informatics domain. The dataset enables:

- Assessment of digital maturity across healthcare facilities
- Exploration of operational bottlenecks in system usage or training
- Development of data visualisations and executive dashboards
- Application of AI methods such as clustering, risk classification, and regression analysis
- Formulation of strategic recommendations for health IT performance improvement

4. Dataset Structure

- **Total Records**: 1,000
- **Qualitative Variables**: 7
- **Quantitative Variables**: 10
- **Derived Variables**: 5

© 2025 Dr. Yasser Aly Yakout Shehata

- **Data Format**: Tabular (CSV-compatible for Power BI, Python, Excel)

5. Qualitative Variables (7)

1. **Facility_Type** – Identifies whether the unit is a hospital or primary care centre
2. **Region** – Specifies the geographic location of the facility
3. **EMR_Vendor** – Indicates the provider of the electronic medical record system
4. **Integration_Status** – Describes the extent of digital integration (Full, Partial, Isolated)
5. **PACS_Connected** – Denotes whether the Picture Archiving and Communication System is integrated
6. **Training_Completed** – Indicates whether the clinical team completed the digital training programme
7. **User_Feedback_Level** – Reflects user-reported satisfaction levels (High, Medium, Low)

6. Quantitative Variables (10)

8. **EMR_Adoption_Rate** – Percentage of patients recorded through the EMR
9. **Avg_Login_Sessions_Per_User** – Average monthly system logins per user
10. **E_Prescriptions_Issued** – Number of electronically issued prescriptions per month
11. **Lab_Reports_Digitized** – Count of laboratory test results digitised
12. **PACS_Uptime_Percent** – System uptime percentage for radiology imaging systems
13. **Support_Tickets_Opened** – Number of technical support tickets logged
14. **Telemedicine_Consults** – Total teleconsultation sessions delivered
15. **Avg_Training_Hours_Per_User** – Number of training hours completed per user
16. **Complaint_Resolution_Time_Days** – Average time to resolve user complaints
17. **Automation_Projects_Implemented** – Count of digital automation deployments

7. Derived Variables (5)

18. **Digital_Efficiency_Score** – Composite index measuring digital health effectiveness based on key performance inputs

© 2025 Dr. Yasser Aly Yakout Shehata

19. **Risk_Flag** – Indicator of operational risk based on PACS downtime, feedback levels, and support ticket load
20. **Engagement_Level** – Categorical segmentation of user participation in digital health systems
21. **Satisfaction_Index** – Quantified score combining user sentiment and complaint resolution efficiency
22. **Facility_Performance_Tier** – Categorical performance grouping (Gold, Silver, Bronze) based on efficiency score rankings

8. Relevance to Power BI and AI for Business

This dataset has been constructed to support multiple modes of business analytics learning and application. It is compatible with Power BI for visualisation, Python and R for data manipulation, and Excel for descriptive statistics. With the inclusion of derived variables, students can simulate executive-level metrics and construct strategic performance dashboards.

It enables:

- Dynamic filtering by region or facility type
- Visual heatmaps of performance metrics
- Hierarchical clustering of risk categories
- KPI benchmarking across vendors and integration statuses
- Predictive modelling for facility performance forecasting

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

Students can explore variable distributions, identify correlations, and uncover hidden trends affecting system usage, training, and feedback.

B. Power BI Visualisation

Interactive dashboards may include slicers for region and facility type, graphs comparing digital adoption and user satisfaction, and matrix visuals depicting support responsiveness.

© 2025 Dr. Yasser Aly Yakout Shehata

C. AI-Driven Business Insights

With the structured inclusion of target variables, students may build machine learning models for risk prediction, satisfaction forecasting, or classification of performance tiers using decision trees or logistic regression.

10. Conclusion

The **Health Informatics Digital Transformation Analytics** dataset offers a rigorously designed, practice-oriented foundation for MBA students seeking to develop competence in data-driven healthcare management. With its multidimensional structure and realistic performance indicators, the dataset bridges the gap between academic theory and operational analytics. It prepares students to critically analyse digital health projects, design data dashboards, and generate actionable insights aligned with the strategic responsibilities of a Health Informatics and IT Director. Through its use, learners gain not only technical skills but also a contextual understanding of digital transformation challenges in healthcare systems.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 28: Platinum Relationship Manager NBE

Introduction

In contemporary banking ecosystems, the strategic role of the **Platinum Relationship Manager** (PRM) has emerged as a cornerstone in the management of high-net-worth individual (HNWI) clients. These professionals operate within a highly competitive environment, wherein client retention, service personalisation, and portfolio optimisation are paramount. As banks intensify their reliance on data-driven strategies to meet these objectives, the ability to extract actionable insights from complex customer datasets becomes increasingly vital. This necessity is particularly significant for postgraduate students in **Business Analytics programmes**, whose academic development must align with the analytical competencies required in such high-stakes roles.

This dataset has been meticulously curated to simulate the operational and analytical environment of a **PRM**, with a focus on enabling rigorous business analysis and informed decision-making. It encapsulates a comprehensive array of client-level attributes including behavioural, financial, and service-interaction metrics, rendered in both categorical and numerical formats. The inclusion of derived variables such as Total Asset Value, Deposit to Loan Ratio, and Engagement Score further enhances the analytical granularity of the dataset, offering a more nuanced understanding of customer profiles.

Designed for application within postgraduate modules such as **Business Analytics for Decision Making**, this dataset facilitates the exploration of practical scenarios that PRMs encounter, such as assessing client risk, segmenting customer bases, evaluating service engagement, and identifying opportunities for cross-selling. Moreover, the dataset aligns with contemporary analytical tools and frameworks, making it particularly suitable for exploratory data analysis (EDA), data visualisation in Power BI, and the development of predictive models within artificial intelligence (AI)-driven business contexts.

Ultimately, this dataset serves as a pedagogical bridge between theoretical instruction and the application of advanced analytical techniques in real-world banking scenarios. It enables students to engage with the core analytical questions faced by relationship managers, while simultaneously cultivating the methodological rigour and technical proficiency necessary for future leadership roles in data-intensive financial institutions.

Dataset Description

The dataset contains 1000 records, each representing a unique Platinum banking customer profile. It includes 7 qualitative variables, 10 quantitative variables, and 4 analytically

© 2025 Dr. Yasser Aly Yakout Shehata

derived variables. These span client demographics, product ownership, digital and in-person engagement, financial value, service quality, and risk assessment.

Purpose and Application

The primary goal of this dataset is to empower students and practitioners to:

- Analyze customer behavior and financial metrics
- Create visual dashboards using Power BI
- Perform Exploratory Data Analysis (EDA)
- Build predictive models for customer segmentation, risk classification, and relationship optimization

It offers an ideal platform for final MBA projects where students apply theory to simulated real-world banking scenarios under the scope of business analytics and AI.

Dataset Structure

- **Total Records**: 1000
- **Total Variables**: 21
    - **Qualitative Variables**: 7
    - **Quantitative Variables**: 10
    - **Derived Variables**: 4

Qualitative Variables (Categorical – Description)

1. **Client_ID**: Unique alphanumeric identifier for each client
2. **Region**: Geographical location of the client (e.g., Cairo, Giza, Alexandria, Upper Egypt)
3. **Primary_Product_Type**: Main banking product used by the client (Deposit, Loan, Investment, Foreign Exchange)
4. **RM_Engagement_Level**: Frequency and depth of Relationship Manager interaction (Low, Medium, High)
5. **Digital_Channel_Usage**: Digital engagement level via mobile or online banking (None, Mobile, Online, Both)

© 2025 Dr. Yasser Aly Yakout Shehata

6. **Complaint_Status**: Service complaint outcome (NoComplaint, Resolved, Unresolved)
7. **Satisfaction_Rating**: Overall satisfaction with banking experience (Poor, Fair, Good, VeryGood, Excellent)

Quantitative Variables (Numerical – Description)

8. **Num_Branch_Visits**: Number of times the client visited bank branches in a year
9. **Num_Digital_Transactions**: Number of digital transactions performed annually
10. **Deposit_Balance**: Total amount held in deposit accounts (in local currency)
11. **Loan_Balance**: Total outstanding loans of the client
12. **Investment_Portfolio_Value**: Current value of investment accounts
13. **RM_Response_Time**: Average time taken (in hours) by the Relationship Manager to respond to queries
14. **Num_Products_Held**: Number of distinct financial products held by the client
15. **Platinum_Segment_Seniority**: Number of years the client has been in the Platinum segment
16. **Avg_Monthly_Fee_Revenue**: Average monthly fee revenue earned from the client
17. **Risk_Score**: A numeric score (0–100) representing client financial or behavioral risk

Derived Variables (Analytically Enriched)

18. **Deposit_to_Loan_Ratio**: Ratio of deposit balance to loan balance; indicates financial leverage
19. **Total_Asset_Value**: Aggregate of deposits, loans, and investment value; represents total client worth
20. **Engagement_Score**: Normalized value based on total in-person and digital interactions
21. **Risk_Level**: Categorized version of Risk_Score (Low, Medium, High) for classification modeling

Relevance to Power BI and AI for Business

This dataset is ideal for Power BI because it contains a balanced mix of categorical and numerical data suitable for slicing, filtering, and dynamic visualization. The inclusion of

© 2025 Dr. Yasser Aly Yakout Shehata

derived variables enhances its value for advanced analytics, including segmentation, forecasting, and predictive modeling—critical pillars of AI in business.

- **Dashboard Possibilities**:
    - Risk-Level Distribution by Region
    - Customer Satisfaction vs Product Type
    - Portfolio Value Trends by Engagement Level
    - Relationship Manager Performance Indicators

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify patterns between satisfaction and engagement
- Discover outliers in financial behavior
- Correlation mapping between risk, revenue, and product usage

B. Power BI Visualization

- Design interactive dashboards to visualize customer value
- Track key metrics like engagement score and total assets
- Compare complaint resolution outcomes across branches

C. AI-Driven Business Insights

- Train models to predict risk levels using behavior and finance metrics
- Use clustering algorithms to profile high-value clients
- Apply decision trees or logistic regression to recommend personalized actions

Conclusion

This dataset offers a comprehensive and realistic foundation for MBA students to explore the power of data-driven decision-making. By simulating the client portfolio of a **Platinum Relationship Manager NBE**, it allows students to engage deeply with banking analytics, explore visualization techniques, and build AI solutions that reflect real business challenges. Whether through EDA, Power BI dashboards, or AI modeling, this dataset serves as a bridge between academic learning and professional application in the financial services industry.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 29: Management Consulting Industry

Introduction

In today's dynamic business landscape, organizations increasingly rely on data-driven insights to enhance productivity, elevate employee engagement, and manage workforce risk. As a result, **Business Consultants** are required to harness analytical tools to evaluate both quantitative metrics and qualitative feedback, thereby generating strategic recommendations that are both evidence-based and actionable.

This dataset has been thoughtfully crafted to simulate a real-world consulting scenario within the **Management Consulting Industry**. It is specifically designed for MBA students pursuing their final project in the "**Business Analytics for Decision Making**" course. The core narrative centers around the role of a Business Consultant who is tasked with analyzing internal organizational data to assess engagement trends, performance effectiveness, workload distribution, and potential burnout across departments and regions.

The structure of the dataset allows for robust exploratory data analysis (EDA), intuitive visualizations through Power BI, and introductory-level applications of artificial intelligence (AI) in a business context. It reflects realistic interdependencies among key workforce variables such as coaching, overtime, career advancement, and satisfaction, providing fertile ground for deriving meaningful business insights. Moreover, the inclusion of carefully engineered derived variables enhances analytical depth, enabling students to move beyond surface-level patterns and uncover underlying trends and predictive indicators.

Whether used to construct dashboards, build statistical models, or simulate AI-powered decision systems, this dataset empowers students to blend theoretical knowledge with practical application. It serves not only as a capstone exercise for academic evaluation but also as a professional toolkit for real-world business analytics and consulting scenarios.

Dataset Description

The dataset consists of 998 records representing individual employees. Each record contains a mix of qualitative and quantitative variables, capturing essential information about the employee's demographics, job role, performance ratings, engagement scores, workload perception, and other key attributes. Several derived variables are included to enable deeper analysis and support advanced analytics.

Purpose and Application

The purpose of this dataset is to:

- Provide MBA students with practical experience in data analysis and storytelling.

© 2025 Dr. Yasser Aly Yakout Shehata

- Support the application of descriptive and diagnostic analytics.
- Enable dashboard creation and insight generation through Power BI.
- Serve as a foundational dataset for predictive modeling and AI-driven business intelligence.

Dataset Structure

- Total Records: 998
- Total Variables: 23
    - Qualitative (Categorical): 10
    - Quantitative (Numerical): 10
    - Derived Variables: 3

Qualitative Variables

1. **Employee_ID** – Unique identifier for each employee.
2. **Department** – The business function the employee belongs to (e.g., Strategy, HR).
3. **Region** – Geographical region of operation (Americas, EMEA, APAC).
4. **Job_Level** – Role designation within the firm (Analyst to Director).
5. **Education** – Highest academic qualification (Bachelor, Master, MBA, PhD).
6. **Managerial_Status** – Indicates whether the employee is in a managerial position (Yes or No).
7. **Tenure_Bucket** – Categorized tenure (Less than 1 year, 1-3 years, 3-5 years, 5+ years).
8. **Engagement_Category** – Broad classification of employee engagement (Low, Medium, High).
9. **Survey_Level** – Grouped representation of survey score (Low, Mid, High).
10. **Risk_Group** – Segment based on burnout and engagement index (High Risk, Moderate Risk, Low Risk).

Quantitative Variables

1. **Age** – Age of the employee in years.
2. **Tenure_Months** – Total tenure of the employee in months.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Survey_Score** – Satisfaction or engagement score on a 1.0 to 5.0 scale.
4. **Weekly_Meeting_Hours** – Hours spent in meetings per week.
5. **1to1_Coaching_Hours** – Time spent in individual coaching or mentoring sessions.
6. **Project_Count** – Number of ongoing projects assigned to the employee.
7. **Overtime_Hours** – Average weekly overtime hours.
8. **Rating_Performance** – Manager-assigned performance score (1 to 10).
9. **Workload_Perception** – Employee's perceived workload intensity (1 to 10).
10. **Career_Advancement_Score** – Score reflecting perceived growth and promotion opportunities (1 to 10).

Derived Variables

1. **Tenure_Years** – Tenure converted from months to years for better interpretability.
2. **Engagement_Index** – Composite score from survey score, performance rating, and advancement score (normalized).
3. **Burnout_Potential** – Weighted combination of overtime and workload perception, indicating risk of burnout.

Relevance to Power BI and AI for Business

This dataset is highly suitable for building Power BI dashboards that visualize relationships between engagement, performance, workload, and tenure. Its structure supports KPI tracking, segmentation analysis, and automated report generation. With AI tools, students can explore predictive modeling and behavior clustering, making it a strong foundation for workforce analytics in a business context.

Applications in Business Analytics and AI

A. **Exploratory Data Analysis (EDA)**

- Identify trends in employee satisfaction and workload.
- Analyze distribution of performance scores across regions and departments.
- Compare tenure against perceived career growth.

B. **Power BI Visualization**

- Build dynamic dashboards with slicers for region, department, and risk group.

© 2025 Dr. Yasser Aly Yakout Shehata

- Visualize burnout potential over time or by job level.
- Create scorecards showing key metrics by employee category.

C. **AI-Driven Business Insights**

- Use classification models to predict high-risk employees.
- Apply clustering to segment workforce based on engagement index.
- Generate recommendations for coaching or workload redistribution based on data patterns.

Conclusion

This dataset enables MBA students to practice core data analysis techniques in a realistic and structured context. It strengthens their ability to uncover actionable insights using both visual and algorithmic tools. With well-defined variables, built-in segmentation, and derived indicators, it serves as a practical and powerful resource for bridging business knowledge with analytical skills and AI applications in the consulting industry.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 30: Healthcare IT Strategy and Analytics

Introduction

In today's healthcare landscape, digital transformation is not a mere innovation, it is a necessity. From electronic health records to AI-assisted diagnostics, healthcare systems around the world are transitioning from paper-based, siloed operations to interconnected, data-driven ecosystems. This shift is driven by the need to improve patient outcomes, reduce operational inefficiencies, ensure compliance with international standards, and enhance system resilience.

At the forefront of this transformation is the role of the **Managing Director of Information Technology and Digital Transformation**, who serves as both a strategist and a change agent. This leader is responsible for aligning health IT systems with national objectives, implementing robust digital infrastructures, and managing projects that span electronic medical records, telehealth, data analytics, cybersecurity, and artificial intelligence.

To mirror the challenges and opportunities faced by such professionals, this dataset has been thoughtfully designed for MBA students undertaking their final project in the **"Business Analytics for Decision Making" course**. It provides a simulated, yet realistic, digital health environment where students can explore the strategic implications of IT implementation in healthcare.

The Healthcare IT Strategy and Analytics dataset includes a rich mix of qualitative and quantitative variables sourced from common digital transformation initiatives, ranging from EMR adoption levels and AI readiness to satisfaction scores and operational KPIs. It also integrates derived variables that offer deeper insights into performance and risk. These enhancements make the dataset particularly valuable for conducting exploratory data analysis, building dynamic dashboards in Power BI, and applying AI models for predictive insights.

This hands-on experience equips students with practical skills required in today's data-centric business world, especially in sectors like healthcare that are undergoing rapid technological evolution. The dataset not only supports the learning of analytical techniques but also cultivates a strategic mindset, preparing students to make informed decisions in complex, technology-driven environments.

Dataset Description

The dataset contains 1000 observations representing digital health projects or departments participating in national transformation programs. Each observation includes operational metrics, satisfaction scores, system performance data, and contextual variables. It enables

© 2025 Dr. Yasser Aly Yakout Shehata

comprehensive data-driven analysis to support project evaluation, staff engagement, digital adoption, and risk identification.

Purpose and Application

The primary purpose of this dataset is to:

1. Provide students with hands-on experience in managing and analyzing complex data relevant to digital transformation in healthcare.
2. Enable exploration of relationships between technology implementation, system efficiency, and stakeholder satisfaction.
3. Serve as a foundation for creating interactive Power BI dashboards and conducting AI-based predictive analytics.

Dataset Structure

- Total Records: 1000
- Total Variables: 20
    - Qualitative (Categorical): 7
    - Quantitative (Numerical): 10
    - Derived (Calculated): 3

Qualitative Variables

1. **Hospital_Region** – Geographical location of the hospital (e.g., Cairo, Alexandria, Aswan).
2. **Dept_Type** – Type of hospital department or function (IT, Clinical, Finance, Admin).
3. **Project_Type** – Classification of the digital transformation project (EMR, Telehealth, AI Diagnostic, Infrastructure).
4. **Staff_Role** – Role designation within the digital transformation team (Director, Manager, Specialist, Technician).
5. **EMR_Stage** – Level of EMR adoption, based on international EMRAM standards (Stage 0 to Stage 7).
6. **Training_Status** – Indicates if the involved staff completed digital systems training (Yes or No).

© 2025 Dr. Yasser Aly Yakout Shehata

7. **Stakeholder_Engagement_Plan** – Quality of stakeholder engagement strategy (Low, Medium, High).

Quantitative Variables

1. **System_Uptime** – Weekly system availability percentage.
2. **Incident_Count** – Monthly count of reported system or IT issues.
3. **Resolution_Time** – Average hours needed to resolve reported incidents.
4. **Patient_Sat_Score** – Average patient satisfaction with digital services on a scale of 1 to 10.
5. **Staff_Sat_Score** – Staff satisfaction score with digital systems on a scale of 1 to 10.
6. **AI_Readiness_Index** – A score (0–100) representing preparedness to implement AI technologies.
7. **Data_Quality_Score** – A numeric measure (1–10) of data completeness and accuracy.
8. **Training_Hours** – Average number of training hours completed per staff.
9. **Cost_Savings** – Monthly cost savings achieved through digital transformation (in thousands).
10. **EMR_Use_Rate** – Percentage of clinical staff actively using the EMR system.

Derived Variables

1. **Operational_Efficiency** – A calculated ratio of system uptime to resolution time. Reflects service reliability.
2. **Engagement_Index** – The average of patient and staff satisfaction scores. Measures overall stakeholder satisfaction.
3. **Risk_Level** – A classification of project or departmental risk based on thresholds for uptime, incident count, and engagement.

Relevance to Power BI and AI for Business

The dataset has been structured to support data modeling, DAX calculations, and dynamic visualization in Power BI. Fields such as Risk_Level, EMR_Stage, and Project_Type can be used as filters and slicers, enabling targeted drilldowns. Numerical metrics are suitable for KPI cards, line charts, and variance analysis. Additionally, the dataset can be exported into

© 2025 Dr. Yasser Aly Yakout Shehata

machine learning platforms to build classification or clustering models for AI-assisted strategic planning.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Analyze regional disparities in EMR usage and system uptime.
- Explore the relationship between training and system satisfaction.
- Identify departments with high incident frequency and low satisfaction.

B. Power BI Visualization

- Build dashboards displaying operational KPIs segmented by hospital region or department.
- Visualize trends in cost savings across project types.
- Compare AI readiness across EMR adoption stages.

C. AI-Driven Business Insights

- Predict high-risk projects using classification models.
- Cluster departments based on digital maturity and satisfaction metrics.
- Forecast system performance or cost savings based on data quality and training indicators.

Conclusion

This **Healthcare IT Strategy and Analytics dataset** offers a rich and realistic context for MBA students to practice core analytics skills while simulating the role of a strategic decision-maker in healthcare digital transformation. Its design enables exploration from multiple analytical perspectives and is highly relevant to modern business challenges in the health sector. Students can leverage it to produce professional-grade dashboards, analytical reports, and AI-based insights aligned with organizational goals.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 31: CIB Corporate Banking Analytics

Strategic Insights from Portfolio Performance and Risk Intelligence

Introduction

In the era of rapid financial globalization and regulatory complexity, the corporate banking sector has undergone a profound transformation. Institutions that once relied heavily on intuition and relationship-driven decision-making now turn to data as the foundation for managing risk, improving profitability, and ensuring client retention. In this evolving ecosystem, the integration of data analytics into core banking functions is not merely a strategic advantage, it is a business imperative.

Corporate Banking Senior Officers operate at the intersection of financial stewardship and strategic advisory. They are entrusted with managing multi-million-dollar portfolios, tailoring financial solutions to clients, detecting early risk signals, and aligning operations with evolving regulatory frameworks. To perform these roles effectively, these professionals must access, interpret, and act upon vast volumes of structured and unstructured data. From loan disbursement patterns and asset-liability balances to client risk profiles and compliance metrics, data is the lens through which every decision must be validated.

This dataset is meticulously crafted to simulate the working environment of a **Corporate Banking Senior Officer** within a major financial institution. It reflects real-world complexities such as client segmentation, portfolio diversification, transaction dynamics, and performance benchmarking. Created as a capstone learning tool for MBA students enrolled in ***Business Analytics for Decision Making***, it empowers learners to translate theoretical knowledge into practical insights.

The structure of the dataset allows students to ask and answer questions that matter: Which clients pose the greatest financial risk? What factors contribute most to profitability? How do compliance scores correlate with early warning indicators? And, more broadly, how can banking operations be optimized using data?

With the rise of artificial intelligence and business intelligence tools like Power BI, today's MBA graduates are expected to go beyond spreadsheets. They must tell compelling stories with data, build dynamic dashboards, and propose data-backed strategies that drive measurable value. This dataset supports that learning journey by offering an authentic, analysis-ready foundation for hands-on exploration.

In short, this project bridges the gap between business education and financial practice. It situates MBA students in the shoes of corporate banking professionals and gives them the tools to make data-informed decisions that shape the future of banking.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Description

The **CIB Corporate Banking Analytics Dataset** contains 1,000 records of anonymized, synthetically generated client portfolio data. It integrates categorical descriptors with numerical measures representing financial performance, risk levels, and compliance scores. Designed for analytical rigor, the dataset supports correlation analysis, risk modeling, business intelligence visualization, and early-warning signal detection.

The dataset contains:

- **7 qualitative (categorical) variables**
- **10 quantitative (numerical) variables**
- **3 derived variables** to support advanced insight generation

Each record represents a distinct corporate client managed within a senior officer's portfolio.

Purpose and Application

This dataset has been curated with the following learning and analytical objectives:

- Support **exploratory data analysis** for uncovering relationships between operational and financial performance metrics.
- Serve as a foundation for **Power BI dashboards** to visualize client segmentation, risk scoring, and KPI monitoring.
- Enable **AI-enhanced analytics** such as clustering, predictive modeling, and decision tree analysis.
- Reflect key decision areas handled by a Corporate Banking Senior Officer, including credit risk management, profitability analysis, and compliance oversight.

It aligns closely with the business context of MBA students working in financial services or preparing for leadership roles in data-centric industries.

Dataset Structure

Variable Type	Count	Description
Qualitative	7	Descriptive fields useful for segmentation and filtering
Quantitative	10	Numeric indicators used in performance and risk analysis
Derived	3	Synthesized fields that provide deeper insights

© 2025 Dr. Yasser Aly Yakout Shehata

Qualitative Variables (Categorical)

1. **Client_ID** – A unique identifier representing each corporate client
2. **Industry** – The sector to which the client belongs (e.g., Manufacturing, Technology)
3. **Client_Segment** – Classification based on client size or affiliation (e.g., SME, Government)
4. **Risk_Category** – Assessed risk level: Low, Moderate, or High
5. **Portfolio_Type** – Loan product type associated with the account (e.g., Term Loans, Working Capital)
6. **Account_Manager** – Name of the portfolio's responsible banking officer (e.g., Ahmed Hassan)
7. **Region_Branch** – Geographic location of the managing branch (e.g., Cairo, Alexandria)

Quantitative Variables (Numeric)

8. **Total_Assets_Million** – Client's reported total assets in millions
9. **Total_Liabilities_Million** – Client's total financial obligations
10. **Net_Interest_Income_Million** – Earnings from core lending operations
11. **Fee_Income_Million** – Income from non-interest services (e.g., advisory, transactions)
12. **Transaction_Volume_Thousand** – Monthly transaction count in thousands
13. **Number_of_Loans** – Active loan accounts within the client's portfolio
14. **Portfolio_ROAE_Percent** – Return on average equity as a profitability measure
15. **Efficiency_Ratio_Percent** – Operational efficiency metric (lower is better)
16. **Early_Warning_Indicators** – Number of triggered alerts based on risk signals
17. **Compliance_Score** – Rating of adherence to internal/external regulatory policies

Derived Variables

18. **Loan_Utilization_Rate** – Proportion of liabilities to assets, reflecting financial leverage

© 2025 Dr. Yasser Aly Yakout Shehata

19. **Risk_Score** – Composite score derived from efficiency ratio, compliance, and early warnings
20. **Client_Profitability_Index** – A weighted score based on net income and transaction volume, indicating value contribution

Relevance to Power BI and AI for Business

This dataset offers direct integration opportunities with Power BI to build:

- Interactive dashboards for **portfolio performance tracking**
- KPI indicators for **client risk and compliance monitoring**
- Geo-mapping visuals for **regional analysis of banking operations**

From an AI perspective, it enables:

- Classification of **high-risk clients** using decision trees
- Segmentation of **client value tiers** with clustering techniques
- Forecasting of compliance or profitability issues via regression models

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

Students can identify correlations, detect outliers, analyze variance across client segments, and evaluate performance across sectors and regions.

B. Power BI Visualization

The dataset supports slicing and filtering by qualitative dimensions, charting financial trends, building scorecards, and creating scenario-based stress testing models.

C. AI-Driven Business Insights

It can be used to train supervised or unsupervised models for predicting profitability, identifying client churn risk, and simulating policy changes on loan performance.

Conclusion

This **Corporate Banking Analytics Dataset** bridges academic learning and industry practice by giving MBA students a rich, practical environment to apply their business analytics skills. It reflects the real-world analytical challenges faced by corporate bankers, especially those in strategic leadership roles like the **Corporate Banking Senior Officer**. Whether used for

© 2025 Dr. Yasser Aly Yakout Shehata

visualization, diagnostics, or predictive modeling, it provides the ideal foundation for developing data-driven decision-making competencies.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 32: Data-Driven Procurement Optimisation

A Business Analytics Simulation for Electrical Engineering Operations

1. Introduction

In the contemporary landscape of engineering project management, procurement functions play a pivotal role in ensuring cost efficiency, timely delivery, and strategic supplier relationships. Nowhere is this more critical than in the **electrical engineering procurement sector**, where the timely acquisition of technical components such as cables, switchgear, transformers, and control systems directly influences project execution timelines, budget adherence, and regulatory compliance.

This **Data-Driven Procurement Optimisation** dataset has been **strategically designed to simulate the operational responsibilities** of a **Procurement Manager** working in such an environment. It reflects the workflow of a mid-sized engineering firm modelled after **Mega Engineering**, which operates without an enterprise resource planning (ERP) system and instead relies on spreadsheet-based procurement management. This practical context allows for the exploration of procurement practices using common tools such as Microsoft Excel and Power BI, providing a bridge between real-world operations and academic learning objectives.

The project is situated within an **MBA final project framework** for the **course *Business Analytics for Decision Making***, where students are tasked with analysing procurement data to identify inefficiencies, assess supplier performance, manage risk, and offer data-driven recommendations to improve decision-making processes. The dataset introduces key business dimensions including supplier behaviour, budget control, delivery delays, and project maturity, allowing students to engage in a structured investigation using both traditional analytics and artificial intelligence methods.

Importantly, the dataset incorporates **realistic and interrelated variables**, mimicking actual business constraints such as delayed deliveries, budget variance, and negotiation outcomes. Derived metrics such as **Procurement Efficiency Score**, **Delay Categories**, and **Budget Compliance**, further deepen the analytical scope, offering students the opportunity to practice feature engineering, build predictive models, and visualise insights with clarity and purpose.

By using this dataset, students develop a dual competency: mastering technical skills in data handling and visualisation while also cultivating strategic thinking aligned with procurement roles in engineering-driven industries. The result is a highly relevant

© 2025 Dr. Yasser Aly Yakout Shehata

educational experience that prepares future business professionals to make informed, evidence-based decisions in complex organisational contexts.

2. Dataset Description

The **Data-Driven Procurement Optimisation dataset** consists of **1,000 procurement records**, representing transactions initiated by a procurement manager across various suppliers, material categories, and project phases. It features a mix of **qualitative** and **quantitative** variables, including **derived fields** created to enhance the analytical depth and real-world relevance of the dataset.

This dataset is structured to simulate realistic procurement processes, highlighting the interconnectedness between supplier performance, budget constraints, delivery timelines, and engineering project maturity.

3. Purpose and Application

This dataset serves as a multi-purpose educational tool intended for the following applications:

- **For MBA students**: To apply business analytics and visualisation techniques to real-world procurement data.
- **For Power BI training**: To create interactive dashboards reflecting cost trends, supplier performance, and project-level insights.
- **For AI for Business modules**: To enable predictive modelling and optimisation in procurement through machine learning.

It offers learners the opportunity to explore operational challenges and propose data-driven recommendations within the context of engineering procurement.

4. Dataset Structure

The dataset includes **24 variables**, grouped into three categories:

- **Qualitative Variables (7)**
- **Quantitative Variables (10)**
- **Derived Variables (7)**

Each record corresponds to a single procurement transaction.

© 2025 Dr. Yasser Aly Yakout Shehata

5. Qualitative Variables (7)

1. **Supplier Name** – The name of the supplier fulfilling the order (e.g., Supplier_1 to Supplier_20).
2. **Material Category** – Type of electrical material ordered, such as Cable, Panel, or Switchgear.
3. **Order Status** – Indicates whether the procurement is pending, ordered, delivered, delayed, or cancelled.
4. **Negotiation Outcome** – Outcome of price negotiation: Successful, Partially Successful, or Unsuccessful.
5. **Project Phase** – The stage of the engineering project: FEED, Engineering, Procurement, Construction, or Commissioning.
6. **Region** – The regional location of the procurement or project: Cairo, Alexandria, Giza, Luxor, Sharm, or Aswan.
7. **Payment Term** – Contractual agreement for payment timing: Net30, Net45, Net60, or Advance50%.

6. Quantitative Variables (10)

8. **Unit Price (USD)** – The cost of one unit of the ordered material.
9. **Quantity** – The number of units ordered.
10. **Total Order Value (USD)** – Product of unit price and quantity, representing the full cost of the order.
11. **Budgeted Value (USD)** – The estimated or approved cost for the order prior to execution.
12. **Cost Variance (%)** – The deviation of the actual cost from the budgeted cost.
13. **Delivery Lead Time (days)** – Expected delivery time between order and receipt.
14. **Delayed Days** – Number of days the order was delayed beyond the expected lead time.
15. **Supplier Rating** – A performance rating on a scale of 1.0 to 5.0, reflecting quality, service, and reliability.
16. **Time to Pay (days)** – The actual time taken to fulfil payment obligations.

© 2025 Dr. Yasser Aly Yakout Shehata

17. **DECRIS Score** – Project engineering maturity index (0–100), reflecting readiness and specification clarity.

7. Derived Variables (7)

18. **Procurement Efficiency Score** – A composite index based on supplier rating, DECRIS score, and delayed days.
19. **Delay Category** – Categorical classification of delay (On Time, Slight Delay, Moderate Delay, Severe Delay).
20. **Budget Compliance** – Indicates whether the order was under budget or over budget.
21. **Efficiency Band** – Grouped performance levels based on Procurement Efficiency Score.
22. **Cost Overrun Risk** – Binary flag identifying high-risk procurements (e.g., when Cost Variance exceeds 5%).
23. **Delivery Risk Level** – Categorisation of supplier reliability based on delays and lead time patterns.
24. **Strategic Procurement Flag** – Flags high-value orders (e.g., Total Order Value above a critical threshold) for prioritisation.

8. Relevance to Power BI and AI for Business

This dataset is structured for seamless integration into Power BI to support:

- Interactive dashboards highlighting supplier trends, budget variances, and delivery timelines.
- Time-based analyses across project phases and regional procurement activities.
- Visualisation of procurement efficiency and risk profiles through gauges, scatter plots, and matrix charts.

In the domain of AI for Business, the dataset is ideal for feature engineering, risk prediction, and classification tasks, providing students with meaningful opportunities to model procurement performance and propose optimisation strategies.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Frequency distributions of material categories and suppliers.

© 2025 Dr. Yasser Aly Yakout Shehata

- Analysis of regional procurement volume and delays.
- Budget variance trends across project phases.

B. Power BI Visualisation

- Supplier performance dashboards (ratings, delays, efficiency scores).
- Geographic visualisations of procurement activities.
- Budget compliance and risk visualisation by category and project phase.

C. AI-Driven Business Insights

- Predictive models for delivery delays and cost overruns.
- Clustering of suppliers based on performance and risk metrics.
- Prescriptive analytics using budget compliance and procurement flags for strategic planning.

10. Conclusion

The **Data-Driven Procurement Optimisation dataset** is a high-fidelity simulation of procurement operations in the electrical engineering domain. Designed specifically for MBA students focusing on business analytics, it enables deep engagement with procurement performance metrics and decision-making challenges. The inclusion of derived variables, coupled with its readiness for Power BI and AI applications, positions it as an ideal resource for experiential learning in EDA, dashboarding, and data-driven strategy formulation.

Through this dataset, students not only practice technical skills but also gain a nuanced understanding of the managerial complexities faced by procurement professionals in project-driven environments.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 33: Garment Quality Assurance Operational Dataset

1. Introduction

In today's globalised and fast-moving apparel market, **garment manufacturers** face intense pressure to deliver consistent product quality while maintaining operational efficiency and cost control. Quality failures whether due to material flaws, inaccurate measurements, or stitching defects can lead to significant consequences, including high return rates, lost contracts, reputational damage, and increased cost of rework or wastage. In this complex production environment, the **Quality Assurance (QA) Manager** plays a pivotal role. More than just identifying defects, the QA Manager is responsible for embedding quality control mechanisms within each stage of the manufacturing lifecycle, ensuring that quality becomes a strategic differentiator for the organisation.

This project is situated within that real-world operational context. It simulates the daily responsibilities and decision-making challenges of a QA Manager working in a **mid-sized garment manufacturing company**. The simulated organisation operates across multiple regional production facilities (e.g., Cairo, Giza, Alexandria) and sources materials from a network of diverse suppliers. The QA Manager is tasked with monitoring inspection outcomes, supplier performance, process quality, and associated costs across all production batches. Data is collected using a digital SAP system, with exports prepared for analytical processing through tools like Microsoft Excel and Power BI.

To support **MBA students** undertaking the final project for the **Business Analytics for Decision Making course**, this dataset has been designed to offer not only realism but also analytical richness. It captures a wide array of quality inspection metrics ranging from Defects per Hundred Units (DHU) and Fabric Inspection Scores to Cost of Poor Quality (COPQ) and Return Rates, while also integrating **derived analytical fields** that enable deeper exploration. These derived variables, including *Inspection Efficiency Score*, *Supplier Risk Level*, and *Defect Severity Category*, are engineered to mimic the complexity and multi-dimensionality of real-world decision-making in quality assurance.

The **Garment Quality Assurance Operational Dataset** thus functions as a bridge between **academic theory and applied business practice**. Students can explore defect trends by production stage, assess the financial impact of poor quality, compare suppliers across performance dimensions, and even build machine learning models to predict inspection outcomes or identify high-risk batches. Furthermore, the dataset's structure allows for the development of executive dashboards, operational KPIs, and predictive insights using **Power BI and AI-based analytics platforms**.

© 2025 Dr. Yasser Aly Yakout Shehata

Ultimately, the purpose of this dataset is not just to teach data handling or statistical methods, but to **cultivate managerial thinking grounded in evidence-based decision-making**. In a business world increasingly shaped by data, quality assurance is no longer just an operational checkpoint, it is a source of competitive advantage. This dataset places students in the shoes of a QA Manager, preparing them to lead with insight, act with precision, and make decisions that balance cost, quality, and customer expectations.

2. Dataset Description

The **Garment Quality Assurance Operational Dataset** comprises **1,000 records**, each representing a unique quality inspection event associated with a production batch. These records contain detailed metrics spanning defect detection, inspection efficiency, supplier performance, and cost-related outcomes. The dataset includes both core operational metrics and derived indicators that reflect industry best practices in quality measurement and strategic analysis.

The data has been pre-validated and enhanced for direct use in business analytics platforms, offering a seamless experience for building insights through dashboards, statistical exploration, and predictive modelling.

3. Purpose and Application

This dataset serves as a practice-based learning resource for MBA students to:

- Conduct data-driven quality performance evaluations
- Explore defect patterns and operational bottlenecks
- Build business intelligence dashboards in Power BI
- Apply machine learning algorithms to forecast quality risks
- Translate analytical findings into actionable business decisions

It is particularly valuable for learners aiming to bridge the gap between analytics theory and real-world manufacturing practice.

4. Dataset Structure

The dataset is composed of **24 variables**, grouped into:

- **7 Qualitative Variables**
- **10 Quantitative Variables**
- **7 Derived Variables**

© 2025 Dr. Yasser Aly Yakout Shehata

Each record captures a distinct inspection instance within the quality control lifecycle.

5. Qualitative Variables (7)

1. **BatchID** – Unique identifier for the inspected production batch
2. **SupplierName** – Name of the vendor delivering materials or components
3. **InspectorID** – Identifier of the quality inspector assigned to the batch
4. **ProcessStage** – Stage of production: Cutting, Sewing, or Finishing
5. **DefectType** – Classification of detected defect (e.g., Stitching, Hole, Color)
6. **Region** – Production facility location: Cairo, Giza, or Alexandria
7. **InspectionOutcome** – Indicates whether the batch passed or failed quality inspection

6. Quantitative Variables (10)

8. **UnitsInspected** – Number of units evaluated during inspection
9. **DefectsCount** – Number of defective units identified
10. **DHU (Defects per Hundred Units)** – Standard industry measure of defect frequency
11. **FabricInspectionScore** – Score derived from the 4-Point Fabric Inspection System
12. **SamplePassRate (%)** – Percentage of sample units that passed preliminary checks
13. **FinalInspectionPassRate (%)** – Final pass rate of the inspected batch
14. **ReturnsCount** – Number of units returned post-distribution due to quality issues
15. **CostOfPoorQuality (USD)** – Total cost associated with defects, rework, and returns
16. **InspectionTime (minutes)** – Time taken to complete the quality inspection
17. **SupplierQualityRating** – Performance score of the supplier (scale 1.0–5.0)

7. Derived Variables (7)

18. **DefectSeverityCategory** – Categorical label (Low, Medium, High) based on DHU
19. **NonConformanceFlag** – Binary flag indicating a failed inspection
20. **ReturnRate (%)** – Ratio of returns to total units inspected
21. **CostPerUnitDefect (USD)** – Average cost per defective unit

© 2025 Dr. Yasser Aly Yakout Shehata

22. **InspectionEfficiencyScore** – Composite index combining pass rates, defect rates, and time efficiency
23. **SupplierRiskLevel** – Supplier risk classification (High or Low) based on historical quality ratings
24. **HighCostBatchFlag** – Binary flag for batches where Cost of Poor Quality exceeds a statistical threshold

8. Relevance to Power BI and AI for Business

This dataset has been meticulously structured for direct integration with **Power BI** and **AI-powered business analytics platforms**. It enables:

- Dynamic dashboards to monitor quality KPIs in real time
- Region-wise and supplier-wise comparisons of performance metrics
- Feature engineering for training machine learning models
- Scenario simulation for proactive risk management

Its structure supports a wide range of visual and statistical applications, from drill-down dashboards to AI-based recommendations.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Uncover defect concentration across process stages
- Analyse quality trends by supplier and region
- Detect anomalies in inspection duration and cost

B. Power BI Visualisation

- Build real-time dashboards displaying DHU, cost impact, and return rates
- Develop supplier scorecards and efficiency trendlines
- Create interactive visuals to support quality management decisions

C. AI-Driven Business Insights

- Predict batch outcomes using classification models
- Forecast high-cost defect scenarios through regression analysis

© 2025 Dr. Yasser Aly Yakout Shehata

- Cluster suppliers based on risk and consistency for strategic sourcing

10. Conclusion

The **Garment Quality Assurance Operational Dataset** is a comprehensive, pedagogically rich resource for MBA students exploring the intersection of data analytics and quality management. By replicating the operational responsibilities of a Quality Assurance Manager, it allows learners to apply analytical techniques in a realistic and industry-specific context.

Its structure supports multi-level analysis from basic performance reporting to advanced AI-based forecasting, making it a highly relevant tool for students aiming to master business analytics and decision-making in complex, quality-driven environments. The dataset ultimately empowers learners to move beyond theoretical frameworks and engage with the analytical challenges that define modern operations management.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 34: Country Manager Garment Operations Analytics

1. Introduction

In the increasingly complex and competitive global garment manufacturing industry, the role of a **Country Manager** is central to operational excellence and strategic alignment. Acting as the key liaison between multinational retail corporations and domestic production ecosystems, the Country Manager oversees the end-to-end lifecycle of production batches, from procurement and quality control to on-time delivery and export compliance. This requires not only deep domain knowledge but also advanced data-driven decision-making capabilities to manage suppliers, factories, logistics, and customer expectations simultaneously.

This **Country Manager Garment Operations Analytics dataset** has been carefully crafted to simulate the real-world decision environment of a **Country Manager responsible for garment production in Egypt**, working on behalf of a leading American apparel brand. The operational context is shaped by complex variables such as fluctuating quality metrics, supplier reliability, factory throughput, lead times, and customer delivery requirements, factors that collectively determine the success or failure of strategic production goals.

Designed specifically for MBA students enrolled in the ***Business Analytics for Decision Making*** **course**, this dataset serves as a capstone resource for applying advanced analytical techniques to authentic business challenges. The simulated data reflects the flow of information typically extracted from enterprise systems like PLM (Product Lifecycle Management), RLM (Retail Lifecycle Management), and LogNet (Logistics Network Tools). These platforms are widely used in the apparel industry to monitor production schedules, cost structures, inventory levels, and compliance status.

By working with this dataset, students are given the opportunity to explore the multifaceted responsibilities of a Country Manager through a data lens. The dataset supports learning across multiple analytics competencies, including **Exploratory Data Analysis (EDA)**, **interactive visualisation using Power BI**, and **AI-driven business modelling**. In particular, it enables students to investigate patterns in defects and delays, evaluate cost drivers, segment supplier performance, and simulate compliance-related risks.

Furthermore, the inclusion of derived variables such as *Perfect Order Flag*, *Supplier Performance Index*, and *Gross Margin Return on Inventory*, extends the analytical depth, allowing students to move beyond basic reporting and into strategic insight generation. This mirrors real-world analytical workflows where operational data must be enriched and interpreted to inform managerial decisions.

© 2025 Dr. Yasser Aly Yakout Shehata

Ultimately, the **Country Manager Garment Operations Analytics Dataset** bridges the gap between theoretical learning and practical application, equipping students with the skills and mindset needed to lead data-informed decision-making in high-stakes, global supply chain environments.

2. Dataset Description

The **Country Manager Garment Operations Analytics Dataset** contains **1,000 synthetic records**, each representing a production batch handled by a Country Manager. It draws from realistic operations involving supplier performance, quality metrics, financial costings, and logistics timelines. The dataset includes **7 qualitative**, **10 quantitative**, and **7 derived variables**.

This structure enables learners to engage in Exploratory Data Analysis (EDA), dashboard design in Power BI, and predictive modelling through machine learning. The dataset mimics data exported from common industry platforms like PLM, RLM, and LogNet, making it analytically rich and business-relevant.

3. Purpose and Application

This dataset aims to:

- Provide MBA students with a simulation of operational decision-making using real-world KPIs
- Enable EDA using Python, Excel, or Power BI
- Support visual storytelling and KPI dashboard creation
- Introduce AI-based tools for forecasting defects, delivery delays, or cost overruns
- Bridge academic learning with industry-standard practices in manufacturing analytics

4. Dataset Structure

Type of Variable	Count
Qualitative Variables	7
Quantitative Variables	10
Derived Variables	7

© 2025 Dr. Yasser Aly Yakout Shehata

Each record is uniquely identified and contains batch-level data across all three dimensions. Variables were constructed with attention to distributional realism and statistical variance, ensuring that patterns, anomalies, and relationships are analytically valuable.

5. Qualitative Variables (7)

1. **BatchID**: Unique identifier for the production batch
2. **SupplierName**: Name of the external supplier providing raw materials or semi-finished goods
3. **FactoryID**: Designation of the local Egyptian factory responsible for production
4. **ProductCategory**: Category of garment manufactured (e.g. Jeans, Shirts, Jackets)
5. **OrderStatus**: Current status of the batch in the production-export lifecycle
6. **Region**: Geographic region in Egypt where the batch is being produced
7. **RetailPartner**: Destination retail chain (e.g. Costco, Walmart, Kohl's)

6. Quantitative Variables (10)

8. **QuantityOrdered**: Number of units requested in the customer order
9. **QuantityProduced**: Actual number of units produced
10. **SAM (Standard Allowed Minute)**: Benchmark time per unit for production
11. **ActualProductionTime**: Total time used for producing the batch
12. **CycleTime**: Average time required to produce one unit in the batch
13. **DHU (Defects per Hundred Units)**: Quality performance metric
14. **FirstPassYield**: Percentage of units passing inspection on the first attempt
15. **LeadTime**: Time taken (in days) from order placement to shipment
16. **CostPerUnit**: Total production cost per unit
17. **CostOfPoorQuality**: Total cost incurred due to defects, rework, and returns

7. Derived Variables (7)

18. **PerfectOrderFlag**: Indicates whether a batch meets quality and delivery benchmarks
19. **InventoryTurnoverRatio**: Efficiency ratio indicating how often inventory is cycled

© 2025 Dr. Yasser Aly Yakout Shehata

20. **GMROI (Gross Margin Return on Inventory)**: Profitability indicator relative to inventory investment
21. **TaktTime**: Required production rate to meet customer demand
22. **SupplierPerformanceIndex**: Composite score evaluating supplier effectiveness
23. **HighRiskSupplierFlag**: Flags batches from suppliers showing repeated quality or delivery issues
24. **ExportComplianceFlag**: Identifies batches facing compliance-related shipment delays

8. Relevance to Power BI and AI for Business

This dataset is structured to facilitate:

- **Custom Power BI dashboards**: Supporting drill-downs by factory, region, supplier, or retailer
- **Dynamic visual storytelling**: Using filters and slicers for cost, quality, and delivery metrics
- **Predictive analytics models**: Training AI algorithms to forecast delivery delays or high-risk suppliers
- **Descriptive analytics**: Generating executive summaries and KPIs based on actual vs. planned performance

It allows for rapid implementation in both Excel and Power BI environments and is formatted to integrate with platforms supporting machine learning pipelines.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Evaluate variations in lead time by supplier or region
- Identify trends in defect rates by product category
- Assess production efficiency using cycle time and SAM metrics

B. Power BI Visualisation

- Create dashboards comparing supplier performance across factories
- Build KPIs for perfect order delivery and quality compliance

© 2025 Dr. Yasser Aly Yakout Shehata

- Visualise cost of poor quality across different production scenarios

C. AI-Driven Business Insights

- Train classification models to predict high-risk batches
- Build regression models for cost forecasting
- Develop clustering solutions to segment factories by performance profile

10. Conclusion

The **Country Manager Garment Operations Analytics Dataset** provides a robust and analytically rich environment for MBA students to apply business analytics and AI concepts in a practical, industry-aligned context. By simulating the responsibilities of a Country Manager in Egypt, the dataset integrates operational, financial, and compliance dimensions enabling a full-stack learning experience.

This resource not only equips learners with analytical techniques but also cultivates strategic thinking aligned with the realities of global apparel production and supply chain management.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 35: General Dentist Operational and Clinical Analytics Dataset

Introduction

In the age of data-driven decision-making, healthcare service providers are increasingly turning to analytics to enhance clinical outcomes, operational efficiency, and financial sustainability. This project introduces the **General Dentist Operational and Clinical Analytics Dataset (GDOCA Dataset)**, a synthetically constructed yet methodologically rigorous dataset designed to reflect the complex, interrelated dynamics of modern dental practice. It has been specifically developed to support the educational needs of postgraduate students pursuing an **MBA in Business Analytics**, particularly those enrolled in the **course *Business Analytics for Decision Making***.

The **GDOCA** dataset centres on the clinical and administrative activities of a **General Dentist** working at Sharmin International Hospital, a representative setting for outpatient dental care. It captures a wide array of data elements that are typically found in Electronic Dental Records (EDR), Clinic Management Systems (CMS), and Customer Relationship Management (CRM) tools. These include patient demographics, service frequency, diagnostic imaging, procedure counts, financial transactions, and appointment logistics.

Moreover, this dataset goes beyond raw data collection by integrating analytically valuable **derived variables** that simulate the types of insights healthcare professionals and business analysts would generate in practice. These include calculations such as revenue per treatment, patient compliance indicators, and classifications of high-value patients, offering a deeper lens into service utilisation and business performance.

By engaging with the **GDOCA Dataset**, **MBA students** are given the opportunity to work within a realistic healthcare context while applying foundational and advanced business analytics techniques. The dataset has been curated to support **Exploratory Data Analysis (EDA)**, **interactive dashboard construction using Power BI**, and the integration of **AI-driven business insight generation**. Its structure allows for scalable exploration from basic descriptive statistics to the development of predictive models, providing a valuable educational platform that bridges academic theory with the real-world challenges of managing and optimising healthcare services.

In summary, the GDOCA Dataset enables a rich, experiential learning environment wherein students can analyse, visualise, and model healthcare operations from the lens of a General Dentist, thereby reinforcing their analytical competencies and strategic thinking within the rapidly evolving field of business analytics.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Description

The GDOCA Dataset comprises **1,000 anonymised patient records**, each representing a unique dental patient encounter or history. It includes a balanced mixture of **qualitative** (categorical) and **quantitative** (numerical) variables, with additional **derived fields** that add analytical depth and enable predictive modelling. Designed for flexibility, the dataset supports a wide array of business analytics exercises, including performance monitoring, patient segmentation, resource planning, and revenue optimisation.

Purpose and Application

The primary purpose of this dataset is to equip MBA students with a realistic data scenario for practising EDA techniques, building interactive dashboards in Power BI, and engaging with AI-driven insights. It mirrors a real-world healthcare setting and highlights the importance of data in optimising patient care, operational efficiency, and strategic planning. Through this dataset, students can:

- Understand the performance drivers in a clinical environment
- Identify behavioural patterns among patients
- Develop predictive models and visual tools for business decision-making

Dataset Structure

- **Total Records**: 1,000
- **Total Variables**: 21
    - **Qualitative Variables**: 7
    - **Quantitative Variables**: 10
    - **Derived Variables**: 4
- **Format**: Tabular (CSV), Power BI-compatible
- **Target Entity**: Patients treated by a General Dentist
- **Contextual Scope**: Clinical, operational, financial, and behavioural indicators

Qualitative Variables (7)

1. **Gender** – Biological sex of the patient (Male, Female, Other)
2. **Insurance Type** – Type of insurance coverage (Public, Private, None)
3. **Smoking Status** – Patient's smoking behaviour (Yes, No)

© 2025 Dr. Yasser Aly Yakout Shehata

4. **Periodontal Disease** – Diagnosis of gum disease (Yes, No)
5. **Imaging Type Used** – Diagnostic imaging employed (None, X-ray, Panoramic, CBCT)
6. **Reminder Sent** – Indicates if a visit reminder was issued (Yes, No)
7. **Payment Status** – Financial payment completion status (Paid, In Process, Overdue)

Quantitative Variables (10)

1. **Age** – Age of the patient in years
2. **Number of Visits Last Year** – Frequency of dental visits in the previous 12 months
3. **Number of Cavities Filled** – Total cavities treated through fillings
4. **Number of Crown or Bridge Procedures** – Number of crown or bridge installations
5. **Number of Endodontic Treatments** – Root canal or related treatments performed
6. **Average Appointment Duration** – Mean time spent per visit in minutes
7. **Case Acceptance Rate** – Percentage of treatment plans accepted by the patient
8. **Monthly Revenue** – Revenue generated from the patient in USD
9. **Inventory Stockout Count** – Number of times necessary items were out of stock
10. **Total Treatments** – Sum of all procedures performed (derived from cavity, crown, and endodontic counts)

Derived Variables (4)

1. **Total Treatments** – Aggregated count of all procedures conducted per patient
2. **Revenue per Treatment** – Revenue generated divided by number of treatments
3. **High Value Patient** – Classification flag for patients with monthly revenue > 400 USD
4. **Compliant Patient** – Flag indicating patients who received reminders and made timely payments

Relevance to Power BI and AI for Business

This dataset is structured to support **interactive dashboarding** and **AI-readiness**. Power BI users can create slicers, filters, and drill-down visualisations for operational KPIs, while AI-driven tools can mine insights from complex variable interactions. The clean and semantically labelled structure ensures seamless integration with business intelligence software and machine learning platforms.

© 2025 Dr. Yasser Aly Yakout Shehata

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify trends in treatment volumes by demographic segments
- Explore correlations between revenue and treatment types
- Assess service usage patterns by insurance or payment status

B. Power BI Visualisation

- Create dashboards highlighting high-value patients, stockouts, and compliance rates
- Visualise geographic or temporal variations in patient behaviour (with added extensions)
- Monitor revenue and appointment KPIs through interactive charts

C. AI-Driven Business Insights

- Predict likelihood of overdue payments based on behavioural and demographic factors
- Classify patients likely to accept treatment plans using decision trees
- Segment patients for targeted communication or loyalty programmes

Conclusion

This **GDOCA Dataset** provides a robust, context-specific foundation for MBA students and researchers aiming to apply analytical tools in a healthcare service environment. With a focus on the role of a **General Dentist**, it not only simulates real-world operational complexity but also encourages the use of business analytics and AI for strategic healthcare decision-making. Its comprehensive structure, enriched by derived variables, makes it ideal for EDA, Power BI visualisation, and applied machine learning in business contexts.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 36: Global Product Manager Business Analytics Dataset

Introduction

In today's data-intensive business landscape, the role of a Global Product Manager extends far beyond traditional oversight of product features and timelines. Modern product leaders must integrate a wide range of quantitative insights and stakeholder inputs into a unified strategy that aligns with global market dynamics, user behaviour, and organisational objectives. As organisations increasingly rely on Software as a Service (SaaS) solutions to drive digital transformation, the strategic responsibilities of product managers have grown in complexity and scope. This includes the need to harness data analytics not only for operational tracking but also for prioritisation, stakeholder alignment, and decision-making at scale.

The **Global Product Manager Business Analytics Dataset (GPM-BA Dataset)** was created to meet the analytical training needs of MBA students, particularly those specialising in **Business Analytics for Decision Making**. This dataset simulates the analytical responsibilities of a Global Product Manager at **PetroApp**, a SaaS firm delivering platform-based services to clients in the logistics sector. It integrates a broad set of metrics that reflect both the technical and strategic dimensions of global product management, including performance indicators such as user reach, feature adoption, stakeholder satisfaction, and prioritisation metrics derived from the RICE framework.

The **GPM-BA Dataset** is grounded in the authentic decision-making environment of product managers operating across regional boundaries, stakeholder hierarchies, and agile development cycles. It enables students to explore how structured data can inform prioritisation, measure delivery efficiency, forecast adoption, and evaluate customer and stakeholder sentiment. Furthermore, by including derived metrics such as feature efficiency and adoption efficiency, the dataset supports deeper analytical investigations and predictive modelling opportunities.

Through this dataset, **MBA students** will engage in hands-on learning that simulates the realities of managing a global digital product portfolio. It encourages learners to apply tools such as Power BI, exploratory data analysis, and AI-driven modelling to generate actionable insights. The **GPM-BA Dataset** thereby serves as an experiential platform through which students can transition from academic theory to the practical analytical competencies demanded in contemporary SaaS product leadership.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Description

The **GPM-BA Dataset** contains **1,000 records**, each representing a product feature evaluated within a specific business context. These features are deployed across three global regions, Americas, EMEA, and APAC, and span a range of product categories including mobile interfaces, dashboards, analytics modules, APIs, and automation tools.

Each record contains detailed data reflecting both development metrics and business outcomes. The dataset includes components from the RICE prioritisation framework (Reach, Impact, Confidence, Effort), performance indicators (velocity, satisfaction, adoption), and qualitative feedback (region, sentiment, product phase). Derived fields have been introduced to support advanced analysis, including feature efficiency and categorised satisfaction ratings.

Purpose and Application

This dataset has been designed for educational use in business analytics training. It enables students to:

- Analyse patterns in product feature performance across global markets
- Prioritise development efforts using RICE scores and stakeholder satisfaction
- Visualise operational metrics and strategic KPIs using Power BI
- Apply AI and machine learning techniques to classify features, predict adoption, and estimate satisfaction levels
- Understand the data-driven nature of product leadership roles in SaaS environments

Dataset Structure

- **Total Records**: 1,000
- **Variables**: 23
    - **Qualitative Variables**: 7
    - **Quantitative Variables**: 10
    - **Derived Variables**: 6
- **File Format**: CSV
- **Unit of Observation**: Feature-level performance instance
- **Business Context**: SaaS Product Management in logistics

© 2025 Dr. Yasser Aly Yakout Shehata

Qualitative Variables (7)

1. **region** – The global region where the feature is evaluated (Americas, EMEA, APAC)
2. **feature_name** – The name of the feature (e.g. Feature_1, Feature_2)
3. **feature_category** – Grouping of features (Mobile, Dashboard, Analytics, API, Automation)
4. **product_phase** – Stage in the product lifecycle (Discovery, Development, Release, Maintenance)
5. **stakeholder_sentiment** – General sentiment feedback from internal stakeholders (Positive, Neutral, Negative)
6. **global_release** – Indicates whether the feature was globally released (Yes, No)
7. **priority_status** – Derived from RICE score (High, Medium, Low)

Quantitative Variables (10)

1. **reach** – Estimated number of users affected by the feature
2. **impact** – Business impact score of the feature on a scale from 1.0 to 3.0
3. **confidence** – Confidence in the estimates provided (0.4 to 1.0)
4. **effort** – Estimated development effort in person-days
5. **rice_score** – A computed metric from the RICE formula: (Reach × Impact × Confidence) / Effort
6. **sprint_velocity** – Number of story points completed per sprint related to the feature
7. **stakeholder_satisfaction** – Numeric feedback from stakeholders (1 = very low, 5 = very high)
8. **feature_usage_rate** – Proportion of users actively using the feature (0.1 to 0.9)
9. **nps_score** – Net Promoter Score for the feature (ranging from -100 to 100)
10. **monthly_active_users** – Actual number of monthly users of the feature

Derived Variables (6)

1. **feature_efficiency** – Ratio of sprint velocity to effort, indicating delivery productivity
2. **adoption_efficiency** – Product of feature usage rate and reach, reflecting actual user conversion

© 2025 Dr. Yasser Aly Yakout Shehata

3. **high_impact_flag** – Boolean flag for features with a RICE score exceeding 1000
4. **satisfaction_category** – Categorical breakdown of stakeholder satisfaction (Low, Medium, High)
5. **priority_status** – Labelled priority based on rice_score thresholds
6. **record_id** – Unique identifier for each observation

Relevance to Power BI and AI for Business

This dataset is optimised for integration with Power BI and other data visualisation tools. Its structured format supports the creation of dashboards that track key performance indicators, segment features by region or category, and monitor feature delivery across lifecycle phases. Additionally, the dataset is AI-ready: numerical fields are scaled appropriately, categorical fields are cleanly encoded, and derived fields offer clear targets for classification, clustering, and regression modelling.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Examine the distribution of RICE scores across regions and product phases
- Compare stakeholder satisfaction across feature categories
- Identify outliers in usage rate or development effort

B. Power BI Visualisation

- Create KPI dashboards for product teams
- Filter and slice data by region, phase, or sentiment
- Display trends in adoption and efficiency over time

C. AI-Driven Business Insights

- Use clustering to group features based on performance and adoption
- Train models to predict stakeholder satisfaction from development metrics
- Forecast monthly active users based on early feature indicators

Conclusion

The **Global Product Manager Business Analytics Dataset (GPM-BA Dataset)** provides a realistic, multidimensional environment for students to practice business analytics in a

© 2025 Dr. Yasser Aly Yakout Shehata

strategic SaaS setting. By analysing operational, behavioural, and stakeholder data, MBA students can develop the skills needed to manage complex product portfolios. This dataset bridges academic learning with practical decision-making and reflects the analytical rigour expected of modern product leaders in data.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 37: ICU Operations and Patient Outcomes Analytics

1. Introduction

In today's data-driven healthcare landscape, the ability to make informed decisions based on clinical and operational data has become a cornerstone of quality care and organizational performance. Nowhere is this more evident than in **Intensive Care Units (ICUs)**, where patient conditions can change rapidly and resource constraints are a daily reality. As healthcare systems strive for greater efficiency, safety, and patient-centered care, the integration of business analytics into frontline medical environments has become both essential and transformative.

This project is developed in the context of a practicing **ICU Staff Nurse and Acting Charge Nurse** who is also pursuing an **MBA in Business Analytics**. Her dual role in hands-on clinical care and managerial oversight uniquely positions her to explore how data can improve not only individual patient outcomes but also the overall functioning of critical care units. The ICU-OPA dataset was designed as part of her final MBA project in the **Business Analytics for Decision Making** course, with the goal of applying advanced analytical methods to solve real-world challenges in hospital operations.

The **ICU Operations and Patient Outcomes Analytics Dataset (ICU-OPA Dataset)** captures a comprehensive view of 1,000 simulated ICU patient encounters, integrating variables related to demographics, vital signs, clinical severity, length of stay, infection status, resource utilization, and discharge outcomes. In addition to commonly recorded variables within hospital information systems such as OASIS Plus and Oracle-based HIS platforms, it includes specially engineered fields like **Severity Level** and **Complexity Index** to support deeper insight generation and more nuanced analyses.

This project serves multiple academic and professional purposes. For MBA students, it provides a realistic environment to apply Exploratory Data Analysis (EDA), visualize trends using Power BI, and prototype AI models for decision support. For clinicians and hospital administrators, it demonstrates the value of integrated data systems and analytics in improving triage accuracy, resource allocation, workload forecasting, and overall ICU effectiveness.

Through this project, the nurse not only fulfills an academic requirement but also takes a critical step toward bridging the gap between clinical intuition and data-backed decision-making. It reflects a broader shift in healthcare where leadership increasingly requires a blend of domain expertise and analytical competence. The **ICU-OPA** dataset stands as a tool to explore this intersection and promote innovation at the front lines of patient care.

© 2025 Dr. Yasser Aly Yakout Shehata

2. Dataset Description

The ICU-OPA Dataset contains **1,000 ICU patient records**, capturing both clinical indicators and operational metrics. It includes:

- 7 **qualitative variables**
- 10 **quantitative variables**
- 2 **derived variables** for enriched analytical capability

All variables are logically connected to ICU workflows and are aligned with business objectives in healthcare management. The data is fully cleaned and formatted for immediate use in Excel, Power BI, Python, or R environments.

3. Purpose and Application

The dataset is designed to help MBA students:

- Practice **Exploratory Data Analysis (EDA)**
- Develop **interactive dashboards** in Power BI
- Train **AI models** for prediction and risk assessment
- Explore how analytics can support clinical and business decision-making

This project is directly aligned with the **Business Analytics for Decision Making** course, where students bridge the gap between frontline operations and strategic planning using data.

4. Dataset Structure

Each row in the dataset represents a **unique ICU patient case**. The dataset is divided into three major variable categories:

- **Qualitative Variables**: Used for filtering, grouping, and segmentation
- **Quantitative Variables**: Used for statistical analysis and modeling
- **Derived Variables**: Computed to support advanced insight generation

5. Qualitative Variables (Categorical)

1. **Gender**: Male or Female
2. **Nationality**: Saudi or Expatriate
3. **Admission Source**: Entry path to ICU (ER, Referral, Transfer)

© 2025 Dr. Yasser Aly Yakout Shehata

4. **Ventilator Use**: Whether the patient required mechanical ventilation (Yes, No)
5. **Infection Status**: Type of infection during ICU stay (No infection, Nosocomial pneumonia, Bloodstream infection)
6. **Triage Category**: Assigned criticality on admission (Critical, High, Medium, Low)
7. **Outcome**: Final status after ICU stay (Survived, Deceased)

6. Quantitative Variables (Continuous/Numerical)

1. **Age**: Patient age in years (18 to 90)
2. **Heart Rate**: Beats per minute on ICU admission
3. **Systolic Blood Pressure**: Systolic reading in mmHg
4. **Respiratory Rate**: Breaths per minute
5. **Creatinine**: Blood creatinine level in mg/dL
6. **White Blood Cell Count (WBC)**: Measured in 10^3/µL
7. **APACHE II Score**: Severity-of-illness index (0 to 71)
8. **OASIS Score**: Risk-based scoring for ICU patients (0 to 48)
9. **Length of Stay (LOS)**: Total days spent in ICU
10. **ICU Bed Occupancy Rate**: Bed usage percentage on day of admission

7. Derived Variables

1. **Severity Level**: Categorized as Low, Medium, or High based on APACHE II score bands. Used to stratify patients by clinical risk level.
2. **Complexity Index**: Composite score combining LOS, Bed Occupancy, and OASIS Score. Represents operational complexity and resource demand per patient.

8. Relevance to Power BI and AI for Business

This dataset is intentionally structured for **Power BI integration**, with categories ideal for slicers, filters, and KPI cards. The inclusion of both static (qualitative) and dynamic (quantitative and derived) data allows for real-time dashboards that can inform:

- ICU staffing efficiency
- Resource allocation

© 2025 Dr. Yasser Aly Yakout Shehata

- Clinical outcomes analysis

In **AI for Business**, the dataset supports model training for:

- Risk scoring
- Outcome prediction
- Workflow optimization using unsupervised learning techniques

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Distribution of patient outcomes by age, gender, or infection status
- Boxplots and histograms for severity levels and LOS
- Correlation analysis of vitals and risk scores

B. Power BI Visualization

- Interactive dashboards with KPIs like average LOS, survival rates, occupancy trends
- Dynamic filters by triage, admission source, or ventilator use
- Charts visualizing workload complexity and ICU capacity over time

C. AI-Driven Business Insights

- Logistic Regression to predict survival based on early indicators
- Clustering patients by Complexity Index for targeted resource planning
- Anomaly detection in vital signs to identify high-risk profiles early

10. Conclusion

This **ICU-OPA** Dataset empowers business analytics students and healthcare professionals to engage in realistic, high-impact data analysis projects. By combining operational insight with clinical understanding, it bridges the gap between care delivery and strategic business planning. For the ICU nurse completing her MBA, this project exemplifies the potential of data to transform decision-making in critical care.

Whether used for training AI models or creating real-time dashboards, this dataset equips students with the tools and mindset needed to lead data-driven change in healthcare and beyond.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 38: Global Corporate Relations (GCR) Strategic Analytics

Introduction

In an increasingly data-driven global economy, corporate banking professionals are expected to leverage analytics to inform strategic decisions, manage risk, and drive client engagement. The role of the **Assistant Vice President of Global Corporate Relations (AVP-GCR)** sits at the nexus of these demands, requiring both a broad understanding of macroeconomic trends and a granular view of client-level performance. This role involves leading complex banking relationships with large corporate clients, analyzing their credit behavior, aligning them with digital transformation initiatives, and ensuring profitability across deposit, loan, and fee-based activities.

The **GCR Strategic Analytics Dataset** has been developed as a learning tool for **MBA students** who are preparing to operate in such high-stakes, analytically intensive environments. It represents a comprehensive simulation of corporate banking data, designed to mirror the decision-making landscape managed by AVP-GCRs in premier banking institutions such as the Commercial International Bank (CIB). The dataset includes a rich blend of financial metrics, strategic indicators, and behavioral signals across multiple regions and sectors, making it ideally suited for coursework in **Business Analytics for Decision Making**, data visualization, and applied AI.

What sets this dataset apart is its intentional design for **real-world applicability** and **academic depth**. Students can explore key variables such as credit exposure, deposit growth, loan-to-deposit ratios, non-performing loan ratios, and fee income, all within the context of region, sector, digital engagement, and ESG (Environmental, Social, and Governance) positioning. Moreover, two derived variables, **Portfolio Quality Index** and **Strategic Engagement Score**, offer multidimensional insights that bridge quantitative rigor with strategic interpretation.

This simulation enables students to adopt the analytical lens of a banking executive. They can answer questions such as:

- Which clients pose the highest credit risk, and how does that correlate with their ESG compliance?
- How does digital adoption influence fee income and overall relationship profitability?
- What regional macroeconomic indicators are linked to deposit growth or credit performance?

© 2025 Dr. Yasser Aly Yakout Shehata

The dataset is purpose-built to allow for end-to-end analytics from **Exploratory Data Analysis (EDA)** to **dashboard creation in Power BI**, and **AI-driven business insights**. It not only supports academic learning but also prepares students for real-world business challenges where data is central to strategic decision-making.

In summary, the **GCR Strategic Analytics Dataset** is more than a collection of records; it is a dynamic, multidimensional learning environment. It empowers future business leaders to think critically, act strategically, and communicate insights effectively, skills essential for any professional navigating the complex world of global corporate banking.

Dataset Description

The **GCR Strategic Analytics dataset** contains **1,000 records**, each representing a quarterly observation of a corporate client relationship. The data blends financial, strategic, and operational dimensions to emulate real-world corporate banking dashboards and relationship management systems. It contains **7 qualitative variables**, **10 quantitative metrics**, and **2 derived variables**, making it rich for exploratory analysis, predictive modeling, and AI-enhanced decision support.

Purpose and Application

The main goal of this dataset is to empower MBA students to:

- Analyze financial and relationship data through descriptive and exploratory techniques.
- Create professional-grade Power BI dashboards for strategic business reporting.
- Apply AI models for predictive analytics, segmentation, and KPI forecasting.
- Understand the relationship between client behavior, financial exposure, and strategic banking priorities.

The dataset also aligns with practical decision areas for a Global Corporate Relations AVP, such as:

- Client segmentation by sector, region, and credit rating
- Portfolio risk assessment using NPL and LDR
- Strategic planning using ESG scores and macroeconomic indicators
- Tracking digital transformation and engagement levels

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Structure

Variable Type	Count	Description
Qualitative Variables	7	Categorical attributes that describe segmentation and strategic characteristics of clients.
Quantitative Variables	10	Continuous and discrete financial and macroeconomic indicators.
Derived Variables	2	Composite variables that summarize client portfolio quality and engagement.

Qualitative Variables (Categorical)

1. **Region**: Geographic location of the client (e.g., Egypt, GCC, Africa, Europe, North America).
2. **Sector**: Industry segment the client operates in (e.g., Energy, Manufacturing, Trade, Tech, Services).
3. **RelationshipType**: Classification based on client-bank interaction style (Deposit-led, Loan-led, Fee-led, Balanced).
4. **CreditRating**: Internal or external rating of the client's creditworthiness (AAA/AA, A/B, BB/B, Below B).
5. **DigitalAdoption**: Level of digital platform use by the client (High, Medium, Low).
6. **ESGTier**: Tier of ESG compliance or sustainability performance (Top, Mid, Low).
7. **KPIAchieved**: Indicates whether key performance indicators for the relationship were met during the period (Met, Not Met).

Quantitative Variables (Numerical)

8. **TotalCredit_EGPmn**: Total credit extended to the client, in Egyptian Pounds (millions).
9. **TotalDeposit_EGPmn**: Client's total deposits with the bank, in Egyptian Pounds (millions).
10. **FeeIncome_EGPmn**: Non-interest income earned from the client relationship.
11. **LDR_Percent**: Loan-to-Deposit Ratio, reflecting the leverage of credit against deposit base.
12. **NPL_Ratio_Percent**: Non-Performing Loan ratio as an indicator of credit risk.

© 2025 Dr. Yasser Aly Yakout Shehata

13. **ProvisionCoverage_Percent**: Ratio showing how well bad loans are provisioned for.
14. **GDP_Growth_Percent**: Quarterly GDP growth rate of the client's operating country.
15. **FDI_Inflows_USDbn**: Foreign direct investment inflow into the client's country.
16. **DigitalTxnVolume**: Monthly volume of digital transactions conducted by the client.
17. **YoY_DepositGrowth_Percent**: Year-over-year deposit growth by the client.

Derived Variables (Strategic Indices)

18. **PortfolioQualityIndex**: Composite score derived from credit volume, NPL ratio, and ESG tier. Reflects the health and sustainability of the client's portfolio.
19. **StrategicEngagementScore**: Weighted score based on digital adoption, fee income, and KPI achievement. Used to assess the strength and potential of the client relationship.

Relevance to Power BI and AI for Business

This dataset is specifically constructed to be highly usable in Power BI for business visualization and dashboarding. Power BI users can:

- Use slicers for **Region**, **Sector**, and **Relationship Type** to compare performance.
- Visualize **Fee Income**, **Credit Exposure**, and **Digital Transactions** using interactive KPIs.
- Create heatmaps and filters for **Portfolio Quality Index** and **Strategic Engagement Score**.

From an AI for Business perspective, the dataset can:

- Train models to **predict KPI achievement** or **segment clients** into strategic groups.
- Enable scenario simulations and what-if analysis (e.g., impact of digital transformation on SES).
- Build interpretable models such as decision trees, logistic regression, or clustering algorithms.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify performance drivers across sectors and regions.

© 2025 Dr. Yasser Aly Yakout Shehata

- Examine risk metrics distribution and their association with credit ratings.
- Analyze digital adoption trends by sector and relationship type.

B. Power BI Visualization

- Build executive dashboards for AVP-level reporting.
- Use trend lines, drill-down filters, and KPI cards to create data stories.
- Monitor strategic indices and macroeconomic impacts visually.

C. AI-Driven Business Insights

- Predict which clients are likely to **achieve KPIs** based on financial and behavioral signals.
- Segment clients into clusters for targeted engagement and upselling strategies.
- Forecast deposit growth based on FDI and GDP trends using regression or time-series models.

Conclusion

The **GCR Strategic Analytics Dataset** is a dynamic, realistic, and analytically rich dataset that captures the complexity of managing strategic banking relationships. It serves as a practical capstone environment for MBA students to develop skills in data analysis, business intelligence, and AI-enhanced decision-making. With applications spanning EDA, Power BI, and predictive analytics, this dataset provides a unique opportunity to simulate the challenges and opportunities faced by an Assistant Vice President in Global Corporate Relations, one of the most data-reliant roles in modern commercial banking.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 39: HRBP Strategic Workforce Analytics

1. Introduction

In today's complex and rapidly evolving business landscape, organizations are increasingly relying on data to drive every aspect of their operations. Human Resources (HR) has become a key player in this transformation, evolving from a traditionally administrative function into a strategic powerhouse that shapes organizational performance, employee experience, and long-term growth. At the center of this evolution is the **Human Resources Business Partner (HRBP)**, a role that bridges people management and business strategy.

Within large-scale manufacturing environments like **Canal Sugar**, the HRBP assumes a pivotal role in managing human capital across diverse functions such as production, quality, maintenance, and sales. These environments are characterized by high operational complexity, dynamic workforce demands, and performance-driven goals. The HRBP must not only ensure regulatory compliance and policy implementation but also align workforce strategies with broader business objectives ensuring the right talent is in the right place, performing at optimal levels, and supported by relevant training and engagement initiatives.

This HRBP Strategic Workforce Analytics dataset was created to simulate the real-world analytical environment encountered by an HRBP in such a manufacturing context. It is designed specifically for **MBA students** enrolled in **Business Analytics for Decision Making**, providing a hands-on, data-rich experience where learners can apply advanced analytical techniques to practical HR challenges. The dataset enables students to explore critical dimensions of workforce management, such as recruitment efficiency, performance evaluation, engagement tracking, training effectiveness, and turnover mitigation.

What sets this dataset apart is its comprehensive scope and strategic depth. It goes beyond simple employee records to include meaningful performance indicators and **derived composite indices**, namely, the **Talent Efficiency Index** and the **Workforce Stability Index**. These indices are crafted to help future HR leaders understand not only what is happening in the organization, but why, and what actions can be taken. They support the development of key insights that can influence hiring strategies, improve employee retention, and enhance overall organizational productivity.

Moreover, this dataset has been carefully structured to support interactive data visualization and AI-driven analysis using tools such as Power BI, Python, and Excel. It empowers students and professionals to perform Exploratory Data Analysis (EDA), build compelling dashboards, conduct forecasting, and generate predictive insights mirroring the kind of work expected in real HR analytics functions.

© 2025 Dr. Yasser Aly Yakout Shehata

In summary, the **HRBP Strategic Workforce Analytics Dataset** is both a learning instrument and a strategic tool. It enables MBA students to bridge theory and practice by applying analytical methods to authentic, high-impact HR problems. Through this dataset, learners can build the confidence, skills, and strategic mindset needed to become influential business partners in the digital era.

2. Dataset Description

The **HRBP Strategic Workforce Analytics Dataset** contains 1,000 synthetic employee records. Each record represents an individual employee in a mid-to-large manufacturing organization, capturing both operational HR metrics and strategic performance indicators. The dataset combines **qualitative attributes**, **quantitative KPIs**, and **derived composite indices**. It has been optimized for use in tools like Power BI, Python, and Excel, supporting a wide range of descriptive, diagnostic, and predictive analyses.

3. Purpose and Application

The purpose of this dataset is to:

- Simulate real-world workforce analytics for MBA students and HR professionals.
- Provide a foundation for practicing key business analytics skills: EDA, dashboarding, and modeling.
- Explore complex relationships between HR metrics, performance outcomes, and workforce dynamics.
- Enable strategic recommendations related to recruitment, retention, training, and productivity.

Applications include:

- Building dashboards to track performance and attrition
- Identifying efficiency gaps in the recruitment pipeline
- Using AI models to segment employees or forecast turnover

4. Dataset Structure

Variable Type	Count	Description
Qualitative	7	Categorical employee and organizational attributes
Quantitative	10	Key performance and engagement metrics
Derived	2	Composite strategic indices for decision-making

© 2025 Dr. Yasser Aly Yakout Shehata

5. Qualitative Variables

1. **Department** – Employee's functional unit (Production, Quality, Maintenance, Sales, Admin)
2. **EmploymentType** – Contract type: Permanent, Temporary, or Contractor
3. **Gender** – Employee gender: Male or Female
4. **EducationLevel** – Educational attainment (High School, Diploma, Bachelor, Postgraduate)
5. **JobLevel** – Position grade: Junior, Mid-level, Senior, or Managerial
6. **TrainingCompleted** – Indicates whether the employee has completed all required training (Yes or No)
7. **SurveyParticipation** – Whether the employee participated in the engagement survey (Yes or No)

6. Quantitative Variables

8. **TimeToHire_Days** – Days required to fill a vacancy for the employee's position
9. **CostPerHire_USD** – Total cost incurred to hire the employee
10. **EarlyTurnoverPercent** – Proportion of new hires leaving within the first year
11. **TurnoverPercent** – Overall annual turnover rate for the department or unit
12. **AbsenceDays_PerEmployee** – Average annual unplanned absence days
13. **EngagementScore** – Employee engagement score based on internal survey (scale 1 to 10)
14. **TrainingCompletionPercent** – Percent of required training programs completed
15. **PerformanceScore** – Overall performance rating (scale 1 to 5)
16. **RevenuePerEmployee_USD** – Average revenue contribution per employee
17. **HR_to_Employee_Ratio** – Ratio of HR staff to total employees

7. Derived Variables

18. **TalentEfficiencyIndex** – A composite score combining Time to Hire, Cost per Hire, and Training Completion to evaluate recruitment and onboarding effectiveness

© 2025 Dr. Yasser Aly Yakout Shehata

19. **WorkforceStabilityIndex** – A synthesized measure of turnover, absence, and performance levels, indicating the reliability and strength of the workforce

8. Relevance to Power BI and AI for Business

The dataset has been optimized for visualization in Power BI and integration into AI workflows:

- **Dimension Tables**: All categorical variables are suitable for slicers, filters, and hierarchical breakdowns in Power BI.
- **KPI Cards and Trend Visuals**: Quantitative and derived fields can be visualized as KPIs or trend lines.
- **AI Enablement**: The dataset includes both individual-level and aggregated metrics, ideal for training classification, regression, and clustering models.
- **Scenarios and What-Ifs**: Can be used in Power BI to create HR policy simulations and data storytelling dashboards.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify patterns in attrition across departments and job levels
- Correlate training completion with performance outcomes
- Analyze gender and education trends in engagement or absenteeism

B. Power BI Visualization

- Build interactive dashboards to monitor key HR metrics
- Use filters to drill into specific departments or job levels
- Display KPI cards for Talent Efficiency and Workforce Stability

C. AI-Driven Business Insights

- Predict high-risk turnover segments based on engagement, training, and performance
- Segment employees using clustering models to tailor engagement strategies
- Forecast absenteeism and its operational impact using regression models

© 2025 Dr. Yasser Aly Yakout Shehata

10. Conclusion

The **HRBP Strategic Workforce Analytics Dataset** serves as a realistic and comprehensive foundation for learning and applying business analytics techniques. It empowers MBA students and HR professionals to explore complex organizational questions through data, develop meaningful insights, and translate findings into strategic workforce actions. Whether through descriptive dashboards or AI-powered forecasts, this dataset supports a future-ready approach to human capital analytics in the manufacturing sector.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 40: RD Performance

Introduction

The **Research and Development Performance Dataset** is designed as a foundational resource for final projects in the MBA course ***Business Analytics for Decision Making***. This dataset simulates institutional data relevant to academic performance, research innovation, and educational quality. It centers around the role of a professor, a leading academic at the Arab Academy for Science, Technology, and Maritime Transport—whose scholarly profile serves as the conceptual basis for the dataset.

The dataset provides a holistic view of R&D activities in higher education settings, focusing on three critical dimensions: **research infrastructure**, **curriculum integration**, and **faculty productivity**. It enables MBA students to apply a range of analytics techniques including Exploratory Data Analysis (EDA), data visualization using Power BI, and the development of AI-driven business insights. This structured data simulates realistic academic operations and is tailored for modeling decision-making scenarios in AI- educational planning and institutional management.

Dataset Description

1. Purpose and Application

This dataset is specifically developed to allow students to:

- Conduct EDA to uncover trends in research funding, academic outcomes, and publication output.
- Build Power BI dashboards to visualize faculty productivity, lab infrastructure utilization, and program-level performance.
- Implement descriptive analytics to measure key performance indicators related to institutional effectiveness.
- Apply AI business concepts such as predictive modeling, segmentation, and optimization in an academic R&D context.

2. Dataset Structure

The dataset contains **1,000 records** representing departments, academic programs, and faculty profiles. Each record contains **7 qualitative (categorical)** and **10 quantitative (numerical)** variables, along with **2 derived variables** that provide deeper insights into funding and research impact efficiency.

© 2025 Dr. Yasser Aly Yakout Shehata

3. Variables in the Dataset

A. Qualitative (Categorical) Variables

Variable Name	Description	Values (Examples)
Dept_Name	Name of the academic department	ECE, CS, BioMed, IT, Mechatronics
Research_Lab_Type	Primary lab specialization in department	AI, Microwave, Bio-Signal, RF, Photonics
Program_Level	Academic level of the program	Undergraduate, Masters, PhD
Course_Type	Classification of course content	Core, Elective, Special Topics
Collab_Status	Nature of external research collaboration	National, International, Industry
Faculty_Rank	Academic designation of the instructor	Assistant, Associate, Professor
Accreditation_Status	Program's accreditation standing	ABET, National, None

B. Quantitative (Numerical) Variables

Variable Name	Description	Value Range (Typical)
Lab_Count	Total number of active research labs in the department	0 – 15
Research_Funding_USD	Annual external and internal research funding (in USD)	10,000 – 100,000
Courses_AI_Integrated	Number of courses incorporating AI models and technologies	0 – 10
Student_Satisfaction_Score	Average satisfaction rating for academic programs (scale 0–100)	50 – 100
Papers_Published	Total peer-reviewed publications by faculty	0 – 30
Avg_Citations	Average citations per publication	0 – 100
H_Index	Composite research output index (based on publications and citations)	5 – 30
Graduation_Rate	Percentage of students completing programs on time	60 – 100%
Employment_Rate	Percentage of graduates employed within 6 months	50 – 100%
Industry_Projects_Count	Number of industry-sponsored R&D projects annually	0 – 20

© 2025 Dr. Yasser Aly Yakout Shehata

C. Derived Variables for Analytical Depth

Derived Variable	Formula and Description
Funding_per_Lab	Research_Funding_USD divided by (Lab_Count + 1); measures resource distribution
Citations_per_Paper	Avg_Citations divided by (Papers_Published + 1); evaluates research impact per paper

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Segment departments based on lab specialization, AI adoption, and student outcomes.
- Correlate research funding levels with H-index, publications, and citations.
- Evaluate curriculum modernization through AI integration metrics.

B. Power BI Visualization

- Develop dashboards for academic unit comparison across satisfaction, funding, and publication metrics.
- Use filters by Program Level, Faculty Rank, and Lab Type to explore departmental performance.
- Display trendlines of graduation and employment rates over simulated academic years.

C. AI-Driven Business Insights

- Forecast academic success using predictive models based on AI course integration and lab counts.
- Cluster departments by research productivity and funding efficiency.
- Optimize research investment strategies using Funding_per_Lab and Citations_per_Paper metrics.

Conclusion

The **Research and Development Performance Dataset** equips MBA students with a realistic and versatile resource for analyzing educational data in an AI business context. It enables the development of dashboards, statistical insights, and predictive tools that

© 2025 Dr. Yasser Aly Yakout Shehata

simulate the challenges and opportunities in managing research-intensive academic institutions.

Designed around the scholarly environment of a professor, this dataset represents a confluence of education, innovation, and analytical rigor. It fosters strategic thinking and data literacy, preparing future business leaders to make informed, AI-driven decisions in the realm of academic and institutional development.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 41: Community Pharmacy Performance

Introduction

The **Community Pharmacy Performance Dataset** is developed to support MBA students enrolled in the *Business Analytics for Decision Making* course. It simulates retail pharmacy operations within the context of **Al-Dawaa Medical Services Co.**, one of Saudi Arabia's largest pharmaceutical chains. The dataset is structured to reflect the operational scope and business dynamics experienced by a **Community Pharmacist**, specifically modeled on the professional setting of a practicing pharmacist at Al-Dawaa.

This dataset is intended to provide students with practical exposure to business analytics techniques that support data-driven decision-making in a healthcare retail environment. It mirrors real-world pharmacy workflows, digital ordering platforms, and customer interaction models, while integrating variables that align with key performance indicators for pharmacy success.

The dataset supports a wide range of analytical tasks including exploratory data analysis (EDA), performance visualization using Power BI, and the development of AI-supported strategies in business operations. It enables MBA students to interpret data in a way that enhances operational efficiency, improves customer service delivery, and strengthens digital engagement strategies—key components of the pharmaceutical retail sector under Saudi Arabia's Vision 2030.

Dataset Description

1. Purpose and Application

The dataset is designed to allow MBA students to:

- Conduct data-driven evaluations of pharmacy performance metrics.
- Visualize customer, sales, and inventory trends across multiple locations and operational shifts.
- Apply AI business tools to forecast customer demand, identify sales optimization strategies, and predict satisfaction drivers.
- Analyze pharmacist productivity, shift-level staffing impact, and platform-based order volume.

© 2025 Dr. Yasser Aly Yakout Shehata

2. Dataset Structure

The dataset contains **1,000 records**, each representing a summary of pharmacy operations across individual shifts, transactions, and customer interactions. It includes **7 qualitative** (categorical) variables, **10 quantitative** (numerical) variables, and **2 derived** variables to enhance the depth of analysis.

3. Variables in the Dataset

A. Qualitative (Categorical) Variables

Variable Name	Description	Sample Values
Branch_ID	Identifier for the pharmacy branch	B101, B202, B303
Shift_Type	Operating shift of the pharmacist	Morning, Afternoon, Night
Staff_Level	Pharmacist staffing level during shift	Low, Medium, High
Customer_Type	Type of visiting customer	New, Returning, Chronic
Payment_Method	Mode of payment used at point of sale	Cash, Credit Card, Insurance
Platform_Used	Ordering or engagement channel	In-store, Mobile App, Call Center
Promo_Engaged	Whether a promotional offer was applied	Yes, No

B. Quantitative (Numerical) Variables

Variable Name	Description	Typical Range
Total_Sales_USD	Total transaction value in USD	100 – 5,000
Items_Sold	Number of items sold during the shift	5 – 100
Prescription_Count	Number of prescriptions fulfilled	1 – 50
OTC_Sales_Ratio	Ratio of over-the-counter to total sales	0.2 – 0.9
Avg_Consultation_Time_Min	Average time spent in consultation per customer	2 – 20 minutes
Customer_Wait_Time_Min	Average customer wait time before receiving service	0 – 30 minutes
Inventory_Turnover	Frequency of stock replenishment	1 – 10
Stockout_Incidents	Number of stockout events per shift	0 – 5
App_Orders_Count	Number of orders placed via Al-Dawaa mobile platform	0 – 60
Satisfaction_Score	Customer satisfaction score based on survey (0–100 scale)	50 – 100

© 2025 Dr. Yasser Aly Yakout Shehata

C. Derived Variables for Analytical Depth

Derived Variable	Description
Sales_per_Item	Total_Sales_USD divided by Items_Sold plus one; indicates revenue efficiency
Consult_Efficiency	Prescription_Count divided by Avg_Consultation_Time_Min plus one

These derived variables support advanced analytics on pharmacist performance and transactional effectiveness.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify patterns in sales volume and customer flow across different shifts and staffing levels.
- Analyze wait time and satisfaction scores to detect bottlenecks in service delivery.
- Segment branches by promotional effectiveness and digital platform usage.

B. Power BI Visualization

- Build dashboards that track sales, prescription counts, and customer ratings by branch or shift.
- Develop interactive filters for platform type, customer segments, and payment modes.
- Visualize digital adoption trends and evaluate app engagement metrics.

C. AI-Driven Business Insights

- Predict sales performance based on promo engagement and customer type.
- Forecast app usage trends and satisfaction fluctuations across operational shifts.
- Classify pharmacists' consultation efficiency to support training and resource allocation.

Conclusion

The **Community Pharmacy Performance Dataset** offers a high-fidelity simulation of retail healthcare operations, making it an ideal resource for MBA students undertaking capstone projects in business analytics. Rooted in the operational realities of a **Community**

© 2025 Dr. Yasser Aly Yakout Shehata

Pharmacist, this dataset aligns practical pharmacy workflows with the strategic goals of AI-powered business insights and continuous service improvement.

It enables students to apply rigorous analytical techniques to real-world business scenarios, bridging academic learning with operational excellence in one of the most dynamic sectors of Saudi Arabia's healthcare system.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 42: Radar Commander

Introduction

The *Radar Surveillance and Performance Dataset* has been purposefully developed to serve as a comprehensive analytical resource for final projects in the MBA course *Business Analytics for Decision Making*. This dataset is grounded in the operational context of a **Radar Commander** in the Armed Forces and reflects real-world complexities associated with military surveillance, radar tracking, operator performance, and system uptime.

The role of a Radar Commander is pivotal in ensuring the effective deployment and interpretation of radar systems for defence, detection, and national security. This dataset captures the breadth of responsibilities inherent to this role by simulating mission-level data on tracked objects, radar system configurations, operator decisions, and alert outcomes. It provides a rich basis for applying a range of analytical techniques, including Exploratory Data Analysis, Power BI visualisation, and AI- business modelling.

By integrating technical, operational, and human performance variables, the dataset fosters strategic thinking and data fluency among MBA students. It encourages the exploration of military analytics through the lens of business intelligence, supporting the development of actionable insights that align with both defence requirements and advanced decision-making paradigms in the AI business landscape.

Dataset Description

1. Purpose and Application

This dataset has been tailored to allow MBA students to:

- Conduct exploratory data analysis to uncover trends in radar tracking, system performance, and operator efficiency
- Build interactive dashboards using Power BI to visualise alert types, response behaviours, and operational readiness
- Apply descriptive and diagnostic analytics to monitor key indicators such as false alarm rate, track confidence, and system uptime
- Leverage AI for Business methodologies, including predictive modelling and clustering, in the context of military surveillance operations

© 2025 Dr. Yasser Aly Yakout Shehata

2. Dataset Structure

The dataset consists of 1000 records representing radar tracking events and operator responses across multiple operational shifts. Each record contains 7 qualitative and 10 quantitative variables. To enhance analytical depth, 3 derived variables have been incorporated to reflect system performance, target altitude classification, and temporal analysis.

3. Variables in the Dataset

A. Qualitative Variables

- **Radar Type**: Operational category of radar in use such as 3D TWS, Ground MTI, or Airborne Mapping
- **Sector ID**: Geographic surveillance sector being monitored
- **Operator ID**: Unique identifier for each on-duty radar operator
- **Shift**: Time frame of operation classified as Day, Night, or Swing
- **Target Classification**: Nature of the detected object such as Aircraft, Missile, Drone, or Civil
- **Alert Type**: Type of alert raised including False Alarm, Potential Threat, or Confirmed Threat
- **Engagement Action**: Response selected by the operator such as Track, Ignore, Engage, or Escalate

B. Quantitative Variables

- **Track ID**: Unique identifier for each detected object
- **Timestamp**: Time of detection in UNIX format
- **Latitude and Longitude**: Geographic coordinates of the tracked object
- **Altitude in Metres**: Height of the object above ground level
- **Radial Velocity in Metres per Second**: Speed of the object in relation to the radar station
- **Track Confidence**: Confidence score assigned by the radar system
- **False Alarm Rate**: Number of false alarms recorded within a shift
- **System Uptime Percentage**: Proportion of operational time during the shift

© 2025 Dr. Yasser Aly Yakout Shehata

- **Operator Response Time in Seconds**: Time taken by the operator to respond to the alert

C. Derived Variables

- **Datetime**: Human-readable format of timestamp to support temporal analysis
- **Hour of Day**: Extracted hour from timestamp to evaluate peak performance or workload
- **Altitude Level**: Categorisation of altitude as Low, Medium, High, or Very High for strategic evaluation
- **System Performance Score**: Composite metric combining Track Confidence and System Uptime to evaluate overall effectiveness

Applications in Business Analytics and AI

Exploratory Data Analysis

- Investigate alert patterns and false alarm distribution across radar types and shifts
- Identify peak response hours and operator performance trends
- Evaluate relationships between system uptime and confidence levels

Power BI Visualisation

- Construct dashboards for operational monitoring, highlighting performance per sector, shift, and radar system
- Visualise geospatial tracking activity and response times
- Create comparative views of engagement outcomes by operator or target classification

AI-Driven Business Insights

- Develop predictive models to estimate alert outcomes based on input characteristics
- Segment radar operations by efficiency using clustering techniques
- Optimise operator shift planning based on response behaviour and workload distribution

© 2025 Dr. Yasser Aly Yakout Shehata

Conclusion

The *Radar Surveillance and Performance Dataset* provides a unique and relevant platform for MBA students to apply business analytics in a high-impact defence environment. It enables the development of data-driven strategies, actionable insights, and AI-supported decision frameworks that mirror the challenges faced by military leaders and business analysts alike. Through this dataset, students can gain experience in managing real-world complexity using the tools and methodologies that define modern analytics.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 43: Patient Experience and Operational Efficiency

Introduction

The ***Patient Experience and Operational Efficiency*** *Dataset* has been specifically curated to support final projects within the MBA course *Business Analytics for Decision Making*. This dataset draws its conceptual foundation from the professional scope of a Senior Consultant in patient experience, as exemplified by Dr Ahmed, who leads national healthcare transformation initiatives in collaboration with the Saudi Ministry of Health and Ascend Solutions. His work focuses on designing patient experience frameworks, leading healthcare command centre analytics, and optimising service delivery in clinical and public health contexts.

In alignment with these responsibilities, the dataset reflects simulated operational data from healthcare facilities, capturing interactions between service efficiency, resource utilisation, staff performance, and patient satisfaction. It provides a realistic environment in which MBA students can apply core analytical techniques, including Exploratory Data Analysis, Power BI visualisation, and AI-driven modelling. The dataset is designed not only to emulate real-world healthcare challenges but also to support the development of actionable insights through structured, high-quality data that facilitates decision making.

Dataset Description

Purpose and Application

This dataset enables students to explore key dimensions of healthcare service delivery, including:

- Monitoring patient flow and satisfaction
- Evaluating the performance of healthcare personnel and facilities
- Analysing operational efficiency indicators such as wait time and resource utilisation
- Applying machine learning techniques to classify, predict, and segment healthcare service outcomes

It is suitable for conducting descriptive and diagnostic analytics, building interactive dashboards in Power BI, and applying AI techniques such as clustering and predictive modelling.

Dataset Structure

The dataset consists of one thousand records, each representing a single patient-related service event in a healthcare facility. These events are distributed across multiple service

© 2025 Dr. Yasser Aly Yakout Shehata

types and regions. Each record includes seven qualitative variables, ten quantitative variables, and five derived variables that enhance analytical depth and interpretability.

Qualitative Variables

1. **Facility Type** – Category of healthcare setting such as Primary Clinic, Hospital Emergency Department, Vaccination Centre, or Follow-up Unit
2. **Region ID** – Coded identifier for the geographical region of the healthcare facility
3. **Consultant ID** – Identifier for the responsible healthcare consultant
4. **Service Type** – Nature of the service provided, such as Consultation, Vaccination, Emergency, or Follow-up
5. **Survey Method** – Mode of patient feedback collection, including Digital, Paper, or Call Centre
6. **Patient Relationship** – Classification of the patient's engagement with the facility as Outpatient, Inpatient, Walk-in, or Referral
7. **Outcome Action** – Action taken following the service, including No Action, Escalated, Followed up, or Resolved

Quantitative Variables

1. **Event ID** – Unique identifier for each service event
2. **Timestamp** – UNIX time format capturing the moment of the event
3. **Wait Time in Minutes** – Duration before the patient received care
4. **Service Duration in Minutes** – Length of the consultation or care episode
5. **Patient Satisfaction** – Satisfaction rating on a scale from one to five
6. **Staff Rating** – Professionalism and performance rating of staff involved
7. **Throughput per Hour** – Number of patients served per hour at the facility
8. **Resource Utilisation Percentage** – Capacity usage during the service event
9. **Vaccination Yield Percentage** – Effective usage of vaccine doses during vaccination services
10. **No Show Rate Percentage** – Proportion of patients who failed to attend their appointments

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Variables

1. **Datetime** – Readable datetime conversion of the timestamp
2. **Hour of Day** – Extracted hour to support time-based analysis of demand and performance
3. **Performance Score** – Composite metric calculated by combining patient satisfaction, staff rating, and throughput to assess event-level service quality
4. **Utilisation Level** – Categorical variable segmenting utilisation into Low, Medium, High, and Full
5. **Flag Variables** – Indicators for High Throughput, Long Wait, and Vaccination Events to support segmentation and scenario analysis

Relevance to Power BI and AI for Business

This dataset is structured to facilitate direct import into Power BI, allowing for the creation of dynamic dashboards that explore performance metrics by facility type, region, and service category. Users can filter, drill down, and compare results across multiple dimensions. Time-based visualisation enables the tracking of demand peaks and response efficiency.

From an AI for Business perspective, the dataset enables supervised learning applications such as satisfaction prediction and no-show forecasting, as well as unsupervised learning through clustering of events by performance profiles. The inclusion of composite metrics and categorical flags supports deeper interpretation, segmentation, and decision support scenarios.

Conclusion

This dataset serves as a practical tool for MBA students simulating the work of a healthcare analytics consultant. It provides the structure and depth required to build meaningful visual insights and predictive models that reflect real-world healthcare service improvement initiatives. Through this work, students can contribute analytically grounded strategies to enhance patient experience and operational outcomes in complex healthcare environments.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 44: GB Sales Performance

Introduction

The *GB Sales Performance* dataset has been developed as an integrated analytical resource for MBA students enrolled in the *Business Analytics for Decision Making* course. It is modelled on the operational responsibilities and strategic oversight of Mr Ali Ahmed Ali Abousaied, a Sales Manager at GB Auto. His role involves managing automotive sales across multiple brands and regions in Egypt and neighbouring Middle Eastern markets, balancing inventory optimisation, sales forecasting, and revenue generation.

This dataset simulates a full year of transactional activity across key sales channels and customer types, providing students with a rich and realistic environment for developing skills in data analysis, interactive visualisation, and AI-supported business decision-making.

Dataset Description

The dataset consists of one thousand simulated sales records, each reflecting a distinct commercial transaction. It includes a balanced mix of categorical and numerical variables designed to reflect the multidimensional nature of sales performance management in the automotive sector.

Through the integration of both original and derived fields, the dataset enables a detailed examination of profitability, inventory efficiency, brand positioning, and promotional effectiveness. The simulated data adheres to realistic business logic, enabling students to draw meaningful insights while applying core analytical techniques.

Purpose and Application

The primary objective of this dataset is to equip MBA students with a platform for practising business analytics in realistic and role-relevant contexts. Specifically, the dataset is designed to:

- Foster proficiency in Exploratory Data Analysis using spreadsheet tools, statistical software, and visual analytics platforms.
- Support the development of interactive dashboards in Power BI for dynamic business monitoring.
- Facilitate the application of machine learning and AI techniques for segmentation, forecasting, and performance optimisation.
- Provide a domain-specific foundation for interpreting sales data through the lens of strategic business roles.

© 2025 Dr. Yasser Aly Yakout Shehata

This dataset directly supports experiential learning by enabling students to replicate and critique decisions made in real-world sales and operations environments.

Dataset Structure

The dataset is composed of twenty-two variables in total, divided into three main categories:

- Seven qualitative variables for categorical analysis.
- Ten core quantitative variables representing business measures.
- Five derived variables calculated from primary metrics to support deeper analysis.

All records are time-stamped and ordered chronologically across a twelve-month period, supporting trend analysis and seasonal comparison. Each variable has been formatted to ensure consistency and readiness for Power BI and statistical software use.

Qualitative Variables

The dataset includes seven categorical attributes that allow users to filter, group, and segment the data according to business-relevant dimensions:

1. **Brand** – The automotive brand involved in the transaction, such as Hyundai, Chery, or Haval.
2. **Region** – The geographical area where the sale took place, including Cairo, Alexandria, Iraq, and Jordan.
3. **Sales Channel** – The type of sales medium used, including Retail, Fleet, Online, and Dealer.
4. **Vehicle Type** – The classification of the sold product, including Passenger Car, Motorcycle, Commercial Vehicle, and Spare Parts.
5. **Promotion Type** – The nature of any promotional strategy used, such as Discount, Bundle, or Financing.
6. **Customer Type** – The type of buyer involved, including New, Returning, Fleet, or Corporate.
7. **Month Category** – A seasonal categorisation of the month as Peak, Off Peak, or Normal.

Quantitative Variables

Ten numerical variables capture the core sales, financial, and operational performance metrics:

© 2025 Dr. Yasser Aly Yakout Shehata

1. **Event ID** – A unique numeric identifier for each sales transaction.
2. **Date** – The recorded transaction date in the YYYY-MM-DD format.
3. **Units Sold** – The number of units sold in the transaction.
4. **Revenue (EGP)** – The total transaction revenue in Egyptian Pounds.
5. **Gross Profit (EGP)** – The profit realised after deducting direct costs.
6. **Conversion Rate (%)** – The percentage of leads converted into successful transactions.
7. **Inventory Days** – The number of days the sold item remained in stock.
8. **Market Share (%)** – The proportion of total market sales represented by this transaction.
9. **After-Sales Revenue (EGP)** – Revenue generated through post-sale services and support.
10. **Price Premium (EGP)** – The difference between the sale price and the market benchmark.

Derived Variables

To support advanced analysis and enhance analytical depth, five derived variables have been added:

1. **Profit Margin (%)** – Calculated as Gross Profit divided by Revenue, expressing profitability efficiency.
2. **Sales Intensity Score** – A composite index combining Units Sold and Conversion Rate to reflect sales effectiveness.
3. **Inventory Turnover Rate** – Calculated as Units Sold divided by Inventory Days, representing inventory management performance.
4. **Month** – Extracted from the Date field to support month-wise trend analysis.
5. **High Performance Flag** – A Boolean indicator classifying a transaction as high-performing if it falls within the top quartile for both Units Sold and Profit Margin.

Relevance to Power BI and AI for Business

This dataset is fully structured for compatibility with Power BI, enabling the creation of dynamic reports, drill-down views, and KPI dashboards. It also supports the application of

© 2025 Dr. Yasser Aly Yakout Shehata

AI tools through features amenable to classification, clustering, and forecasting algorithms. With time-based, categorical, and numerical variables, it offers a balanced structure for both business intelligence visualisation and applied machine learning experimentation.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA):

Students can apply statistical tools to understand variable distributions, correlations between revenue and margin, seasonal patterns in sales volume, and performance differences across brands and regions. The data structure also supports hypothesis testing and segment comparisons.

B. Power BI Visualisation:

Interactive dashboards can be developed to monitor KPIs by month, region, or product category. Visual tools such as heatmaps, time-series graphs, decomposition trees, and profit funnel analyses can be applied to interpret trends and discover root causes of performance variance.

C. AI-Driven Business Insights:

The dataset enables the construction of clustering models to segment customers or regions, regression models to predict revenue or profitability, and classification models to identify high-performing sales strategies. The inclusion of derived metrics strengthens the training of AI models by offering deeper behavioural features.

Conclusion

The *GB Sales Performance* dataset offers a comprehensive and practical tool for developing business analytics competencies among MBA students. It replicates the analytical environment encountered by a senior sales manager in the automotive sector and supports the development of strategic insights using Power BI and AI methodologies. Through structured engagement with this dataset, students will gain expertise in data interpretation, insight communication, and evidence-based decision-making, thereby preparing them for leadership roles in data-driven business environments.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 45: CDP Performance

Introduction

The *CDP Performance for Ali Swissi MBA* dataset is a purpose-built synthetic dataset developed to support experiential learning in Business Analytics for Decision Making. It simulates the analytical environment faced by an IT Project Manager operating within a Customer Data Platform (CDP), a domain that combines customer behavioural data, campaign performance, and system-level metrics. The dataset is designed to reflect realistic business scenarios and enable the application of statistical, visual, and machine learning techniques for deriving insights that guide data-driven decisions.

Dataset Description

This dataset comprises 1,000 interaction-level records from a simulated digital platform spanning multiple customer engagement channels. Each record represents a unique event, capturing information such as customer segment, campaign type, transaction characteristics, and platform performance indicators. The data includes timestamps converted to a readable format, categorical descriptors, and multiple numeric fields capturing behavioural, financial, and technical dimensions. Importantly, the dataset includes engineered fields that provide additional depth and predictive potential for AI applications.

Purpose and Application

The dataset is intended for MBA students studying business analytics, particularly those focused on applying Exploratory Data Analysis (EDA), creating dynamic dashboards in Power BI, and developing AI-driven decision support systems. The dataset enables learners to practice identifying trends, uncovering anomalies, segmenting customer behaviours, and modelling key outcomes such as customer churn and engagement. It aligns with the responsibilities of an IT Project Manager working in a CDP context, such as monitoring data ingestion performance, optimising customer touchpoints, and supporting campaign effectiveness analysis.

Dataset Structure

The dataset includes 24 variables: 7 qualitative, 10 core quantitative, and 7 derived features. It is structured as a flat CSV file suitable for import into data analytics platforms including Python, R, Excel, and Power BI. Each row corresponds to one discrete customer interaction event.

© 2025 Dr. Yasser Aly Yakout Shehata

Qualitative Variables

1. **Channel** – The platform through which the customer interacted (e.g. Web, Mobile App, Email, InStore)
2. **Customer Segment** – Classification of customers based on value and engagement level (e.g. High Value, At Risk)
3. **Campaign Type** – Type of marketing campaign linked to the interaction (e.g. Promotion, Retargeting)
4. **Subscription Status** – Current customer engagement state (e.g. Active, Churned, Dormant)
5. **Geo Region** – Geographical region of the customer (e.g. North, South, East, West)
6. **Device Type** – Device used for accessing the service (e.g. Desktop, Mobile, Tablet)
7. **Data Quality Flag** – Indicator of the reliability of the data record (e.g. Complete, Incomplete)

Quantitative Variables

8. **Session Duration (sec)** – Duration of the session in seconds
9. **Click Count** – Number of clicks performed by the customer during the session
10. **Transaction Count** – Number of transactions completed during the session
11. **Revenue EGP** – Revenue generated during the session in Egyptian Pounds
12. **Average Order Value** – Mean value of orders made during the session
13. **Churn Flag** – Binary indicator showing whether the customer has churned (1) or not (0)
14. **Customer LTV EGP** – Lifetime Value of the customer, in Egyptian Pounds
15. **Data Ingestion Latency (sec)** – Time delay in seconds for the data to be ingested into the system
16. **Datetime** – Date and time of the customer interaction, derived from UNIX timestamp
17. **Hour of Day** – Hour extracted from the datetime value to support time-based analysis

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Variables

18. **Click Through Rate (CTR percentage)** – Engagement rate calculated as clicks per second of session duration
19. **Revenue per Session** – Revenue divided by session duration, measuring profitability per second
20. **High Value Flag** – Indicator for customers whose LTV is in the top 25 percent
21. **High Latency Flag** – Flag for data records with ingestion latency in the top 10 percent
22. **High Engagement Flag** – Marks sessions with click-through rate in the top 25 percent
23. **Stable Customer Flag** – Identifies customers who are both active and have above-median LTV
24. **Datetime (reformatted)** – Readable timestamp used for sorting and visualisation

Relevance to Power BI and AI for Business

The dataset's design facilitates immediate integration into Power BI dashboards, supporting the use of dynamic visuals, filters, and drill-down capabilities. The categorical and time-based variables enable custom visualisation of customer trends, revenue performance, and campaign effects. Derived flags and ratio metrics provide KPIs and benchmarks for dashboard users. For AI integration, the dataset's clean structure and labelled fields support classification, regression, and clustering algorithms that can be trained to predict churn, customer value, or data quality anomalies.

Applications in Business Analytics and AI

A. **Exploratory Data Analysis (EDA)**

Users can explore the data distribution, detect outliers, examine correlations, and segment customers by region, campaign, or device. Summary statistics and histograms reveal usage patterns and financial performance.

B. **Power BI Visualisation**

Custom dashboards can display interactive visuals such as funnel plots for campaign impact, LTV segmentation charts, latency heatmaps, and customer engagement flows across devices and regions.

© 2025 Dr. Yasser Aly Yakout Shehata

C. **AI-Driven Business Insights**

Predictive models can be developed using churn and high-value flags. Clustering algorithms can group customers based on behavioural and value features. Anomaly detection models can identify outliers in system latency or customer engagement.

Conclusion

The *CDP Performance for Ali Swissi MBA* dataset serves as a comprehensive learning tool that bridges theory and practice in the domains of business analytics and AI. It reflects the complexity and richness of data managed by IT Project Managers in customer intelligence platforms. With its structured format, engineered features, and domain relevance, the dataset empowers MBA students to practise core analytical skills and derive actionable insights using both visualisation and machine learning approaches.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 46: DCS Operations

1. Introduction

In today's rapidly evolving industrial landscape, data is a strategic asset that drives operational efficiency, safety, and profitability. Nowhere is this more apparent than in the **petrochemical sector**, where complex processes run continuously, and any deviation in conditions can lead to costly downtime, safety risks, or suboptimal product quality. At the heart of this high-stakes environment is the **Distributed Control System (DCS)**, a sophisticated digital platform used to monitor and control plant operations in real time.

Among the professionals navigating this system, the **Operations Engineer (DCS Operator)** plays a pivotal role. These engineers oversee the stability of chemical reactions, ensure adherence to environmental and safety standards, and respond to real-time alarms and system events. Their decision-making is driven by an intricate web of sensor data, process parameters, maintenance schedules, and operational incidents. However, the growing complexity of these systems and the vast volume of data they generate demand more than traditional oversight. This is where **Business Analytics and Artificial Intelligence (AI)** step in providing advanced tools to interpret data patterns, predict risks, and support more informed and proactive decisions.

This dataset was developed by an experienced DCS Operator pursuing an MBA in Business Analytics. It is inspired by real operational scenarios from a leading Egyptian petrochemical company and serves as the foundation for a final project in the course **Business Analytics for Decision Making**. The **DCS Operations Dataset** simulates a realistic operating environment where MBA students can apply analytical techniques to explore patterns, diagnose inefficiencies, and create performance dashboards using platforms like **Power BI**.

More than just a spreadsheet, this dataset reflects a professional's perspective on how data flows through an industrial control environment. It allows students to step into the shoes of a plant engineer, analyze the implications of their operational choices, and uncover the stories that the data tells. By integrating derived performance metrics such as efficiency scores and alarm risk levels, the dataset enables deeper and more meaningful analysis beyond surface-level KPIs.

Whether the goal is to visualize alarm trends across shifts, evaluate how operator experience impacts energy consumption, or forecast production yield under varying feedstock types, this dataset provides the depth and variety needed to simulate real-world business analytics in an engineering context. It serves as both a practical learning tool and a bridge between the technical realities of industrial operations and the strategic frameworks of modern business analytics and AI.

© 2025 Dr. Yasser Aly Yakout Shehata

2. Dataset Description

The **DCS Operations Dataset** includes **1000 observations**, each representing one operational snapshot from a petrochemical unit during a specific shift. The dataset contains a mix of **qualitative (categorical)** and **quantitative (continuous)** variables relevant to real-world plant operations. Derived variables have been included to enhance the analytical depth and support insights beyond basic metrics.

3. Purpose and Application

The dataset was designed for:

- Practicing **Exploratory Data Analysis (EDA)**
- Creating compelling **Power BI dashboards**
- Developing **AI-based models** for operational efficiency and risk assessment
- Supporting business decision-making in **industrial process optimization**, **energy management**, and **incident response**

It offers a realistic learning platform for MBA students to apply analytics tools in a meaningful, industry-based context.

4. Dataset Structure

- **Total Records**: 1000
- **Variables**: 19 total
    - **7 qualitative variables**
    - **10 quantitative variables**
    - **2 derived variables**
- **Format**: CSV, compatible with Excel, Power BI, Python, R

5. Qualitative Variables

1. **Shift** – Work shift: Morning, Afternoon, or Night
2. **Unit_ID** – Identifies the processing unit: Unit_A, Unit_B, Unit_C
3. **Operator_Experience_Level** – Level of the operator: Junior, Intermediate, Senior
4. **Control_Mode** – System mode: Manual, Semi-Auto, Auto
5. **Maintenance_Status** – Maintenance condition: None, Scheduled, Unscheduled

© 2025 Dr. Yasser Aly Yakout Shehata

6. **Incident_Type** – Operational incident: None, Adjustment, Alarm, Shutdown
7. **Feedstock_Type** – Type of feed material: High_API, Medium_API, Low_API

6. Quantitative Variables

1. **Riser_Temp_C** – Temperature at the riser outlet (°C)
2. **Regenerator_Temp_C** – Temperature at the regenerator outlet (°C)
3. **Reactor_Pressure_bar** – Reactor pressure (bar)
4. **Feed_Flow_tph** – Feedstock flow rate (tons per hour)
5. **Cat_Oil_Ratio** – Catalyst-to-oil ratio (unitless)
6. **Cycle_Time_min** – Residence time in reactor (minutes)
7. **Product_Yield_Percent** – Percentage of product yield
8. **Coke_Content_Percent** – Percentage of coke in output
9. **Energy_Consumption_kWh** – Hourly energy usage (kWh)
10. **Alarm_Count** – Number of alarms during the shift

7. Derived Variables

1. **Efficiency_Score** – Calculated as Product_Yield_Percent divided by Energy_Consumption_kWh; this score estimates energy efficiency
2. **Alarm_Risk_Level** – Categorical risk banding based on Alarm_Count: Low, Moderate, High

8. Relevance to Power BI and AI for Business

This dataset has been optimized for use in **Power BI dashboards** and **AI-enabled analytics platforms**. Its well-balanced mix of structured, interpretable variables and derived fields supports:

- Effective slicing and filtering for dynamic dashboards
- Visual storytelling and operational performance monitoring
- Training predictive models and scenario analysis tools

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

© 2025 Dr. Yasser Aly Yakout Shehata

- Investigate correlations between control mode, operator experience, and yield
- Detect patterns in alarm counts across shifts or maintenance conditions
- Analyze feedstock type impact on energy efficiency and production yield

B. Power BI Visualization

- Build dashboards to monitor:
  - Efficiency trends over shifts or units
  - Incident frequency across operator experience levels
  - Energy consumption vs. product yield
- Use slicers and filters for:
  - Alarm risk levels
  - Control mode transitions
  - Maintenance scheduling

C. AI-Driven Business Insights

- Train regression models to predict:
  - Product yield based on process conditions and operator variables
- Develop classification models to:
  - Predict high-risk alarm conditions
  - Recommend optimal operating parameters
- Use clustering for:
  - Identifying high-performance operator or shift combinations
  - Segmenting operational profiles for maintenance planning

10. Conclusion

The DCS Operations Dataset is a dynamic, business-relevant resource designed for MBA students aiming to master business analytics and AI in industrial contexts. By focusing on the decision-making environment of a DCS Operator, it bridges technical process control with strategic business thinking. Students can apply EDA, visualization, and AI tools to uncover patterns, forecast outcomes, and drive operational improvements. This dataset prepares future business leaders to transform raw data into meaningful insights in complex, high-impact industries.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 47: Natural Gas Distribution

1. Introduction

The strategic management of natural gas distribution plays an increasingly vital role in addressing the intersecting challenges of energy security, environmental sustainability, and economic development. In the context of expanding urbanisation, rising energy demand, and evolving regulatory frameworks, the effective oversight of gas infrastructure, service delivery, and environmental compliance has become a central concern for executive leadership within the energy sector. The position of Managing Director within natural gas distribution entities entails significant responsibility in coordinating district-level infrastructure expansion, stakeholder engagement, maintenance planning, and financial oversight. These responsibilities demand a data-informed approach to decision-making, one that is grounded in operational transparency, predictive capability, and performance benchmarking.

This dataset has been developed as a capstone project by an experienced Managing Director in the natural gas industry, currently pursuing a Master of Business Administration with a concentration in Business Analytics. It is designed to simulate a real-world data environment, enabling postgraduate students to engage in advanced analytical techniques that are directly applicable to energy distribution management. The dataset reflects the multifaceted nature of operational management, incorporating variables that span infrastructure development, customer service metrics, financial indicators, environmental performance, and stakeholder engagement.

Furthermore, the dataset is structured to facilitate rigorous exploratory data analysis (EDA), visual analytics through platforms such as Microsoft Power BI, and the application of artificial intelligence (AI) methodologies for forecasting and strategic modelling. It integrates both categorical and continuous variables across 1,000 district-level records, capturing daily operational snapshots that mirror the temporal dynamics of natural gas service delivery. The inclusion of derived metrics, such as operational efficiency ratios and environmental intensity scores, further enhances the analytical depth and relevance of the dataset for managerial and academic use.

In sum, this dataset provides an empirically grounded foundation for applying business analytics to real-world energy challenges. It supports the development of analytical competencies essential for contemporary leadership in utility management, offering a structured and contextually rich platform for students to explore the intersection of data, strategy, and operational excellence.

© 2025 Dr. Yasser Aly Yakout Shehata

2. Dataset Description

The Natural Gas Distribution Dataset comprises one thousand records representing district-level daily observations. Each entry includes geographic identifiers, infrastructure status, customer service metrics, financial indicators, and environmental measurements. The dataset spans a mix of urban and rural regions and reflects a range of pipeline coverage and PRS (Pressure Reduction Station) capacities, simulating realistic variations in resource allocation and operational maturity across districts.

This dataset is clean, logically structured, and designed for analysis using Excel, Power BI, or Python-based platforms. Key business performance metrics have been embedded as derived variables to support advanced insight generation. This structure makes the dataset particularly useful for students and professionals who wish to understand or simulate strategic challenges in natural gas operations.

3. Purpose and Application

This dataset is built to:

- Train MBA students in **exploratory data analysis (EDA)** using real-world energy sector scenarios
- Support the creation of **Power BI dashboards** that simulate executive decision-making tools
- Serve as a foundation for **AI-powered insights**, including forecasting, segmentation, and scenario modeling
- Bridge technical operations and business performance for those aiming to lead in utility or infrastructure sectors

It encourages students to extract actionable insights, benchmark district performance, and support evidence-based decisions, echoing the real-world responsibilities of a Managing Director overseeing gas distribution.

4. Dataset Structure

- **Total Observations**: 1000
- **Time Frame**: Daily records starting January 2023
- **Coverage**: 50 simulated districts (urban and rural)
- **Variables**: 20 total
    - 7 qualitative

© 2025 Dr. Yasser Aly Yakout Shehata

- 10 quantitative
- 3 derived

5. Qualitative Variables (7)

1. **District_Name** – Name of the administrative district
2. **Month** – Reporting date for the operational entry
3. **Region_Type** – Urban or rural classification
4. **Grid_Expansion_Status** – Phase of grid infrastructure: None, Planning, In-Progress, Completed
5. **Stakeholder_Engagement_Level** – Community and stakeholder interaction level: Low, Medium, High
6. **Environmental_Incident** – Type of environmental incident reported: None, Minor, Major
7. **Maintenance_Type** – Classification of maintenance activity: Routine, Upgrade, Emergency

6. Quantitative Variables (10)

1. **Pipeline_Length_km** – Kilometers of gas pipeline in the district
2. **Household_Connections_Count** – Number of new household connections
3. **PRS_Capacity_m3ph** – Throughput capacity of the PRS (m^3/hour)
4. **Total_Customers** – Total number of active service connections
5. **Network_Coverage_Percent** – District coverage as a percentage of households
6. **Annual_NG_Consumption_MTOE** – Annual gas consumption in MTOE
7. **Revenue_EGP_M** – Monthly revenue in million Egyptian pounds
8. **Subsidy_Expenditure_EGP_M** – Government subsidy expenditure (in million EGP)
9. **Flared_Gas_Volume_MCM** – Gas flared (in million cubic meters)
10. **Emission_NOx_tonnes** – Emissions of nitrogen oxides in tonnes

© 2025 Dr. Yasser Aly Yakout Shehata

7. Derived Variables (3)

1. **Connections_per_km** – Operational efficiency ratio of household connections per pipeline kilometer
2. **Revenue_per_Customer_EGP** – Monthly revenue generated per customer
3. **Emissions_per_MTOE** – Emission intensity per unit of gas consumed

These variables help in identifying performance benchmarks, environmental efficiency, and return on infrastructure investment, important for executive-level analytics.

8. Relevance to Power BI and AI for Business

This dataset is ideal for use in **Power BI dashboards** and **AI-driven models** due to its well-structured, clean format and business-aligned variable design. It allows analysts and students to simulate complex decisions through interactive visualizations and predictive models. Slicers, filters, time-series views, and KPI dashboards can be easily created to reflect operational priorities such as:

- Growth in customer coverage
- Revenue and subsidy alignment
- Environmental impact monitoring
- Efficiency trends by region type or infrastructure level

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Analyze correlations between infrastructure expansion and customer growth
- Identify outliers in revenue or emission patterns
- Understand regional variation in network coverage or stakeholder engagement

B. Power BI Visualization

- Create dashboards displaying district-wise efficiency, revenue, and incident trends
- Use filters to compare performance across urban vs rural regions
- Visualize PRS capacity vs consumption or subsidy across timelines

C. AI-Driven Business Insights

- Predict revenue or customer growth based on pipeline and PRS data

© 2025 Dr. Yasser Aly Yakout Shehata

- Cluster districts into high- and low-performance zones
- Build models to forecast emission intensity or risk of environmental incidents
- Simulate subsidy reduction scenarios under expanded network coverage

10. Conclusion

The Natural Gas Distribution Dataset is a practical, high-value resource for MBA students and analysts seeking to bridge operational knowledge with strategic business insight. It simulates the data environment of a Managing Director responsible for natural gas distribution and creates opportunities to explore EDA, build dashboards in Power BI, and develop predictive models using AI. By integrating technical metrics with financial and environmental indicators, this dataset supports a modern, decision-oriented approach to utility management and infrastructure leadership.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 48: Senior Medical Representative

Introduction

In today's data-intensive business environment, the ability to transform raw data into actionable insights is fundamental to organisational success. This skill is especially critical in sectors where decision-making is closely tied to dynamic market conditions, customer engagement, and operational efficiency. The pharmaceutical industry exemplifies such a context characterised by high competition, regulatory pressures, and the need for evidence-based sales strategies. Within this sector, **Senior Medical Representatives** serve as frontline agents of value delivery, bridging pharmaceutical innovation and clinical practice. They operate in a complex ecosystem that demands not only strong interpersonal and medical knowledge but also sharp analytical acumen to manage territories, optimise outreach, and exceed performance benchmarks.

This dataset has been developed to emulate the multifaceted responsibilities and performance metrics associated with the role of a Senior Medical Representative. It is intended as a practical and pedagogically sound resource for MBA students undertaking final projects in the field of ***Business Analytics for Decision Making***. It represents a synthesis of realistic operational data, curated to reflect the challenges and opportunities facing pharmaceutical sales professionals operating in competitive markets such as Saudi Arabia.

The simulation is grounded in the professional profile of an experienced licensed pharmacist who currently works as a Senior Medical Representative for a prominent healthcare distribution group. These individual's responsibilities span sales forecasting, customer relationship management, product promotion, and territory development across multiple geographic regions. Inspired by these real-world dynamics, the dataset captures both contextual and performance-related variables necessary for strategic analysis.

The primary objective of this dataset is to support the development of core competencies in exploratory data analysis (EDA), visualisation using Power BI, and the application of AI methodologies to extract business insights. It provides students with the opportunity to engage with structured data that reflects organisational realities, encouraging the formulation of analytical questions, hypotheses, and business strategies. Moreover, it supports the transition from descriptive to predictive and prescriptive analytics, enabling learners to progress from understanding what is happening in the data to what actions should be taken.

The inclusion of both original and derived variables enhances the dataset's analytical richness, allowing students to explore cause-effect relationships, segmentation logic, and

© 2025 Dr. Yasser Aly Yakout Shehata

efficiency metrics. It serves not only as a learning tool for technical skills but also as a model of how data can inform strategic business decisions whether in human resource allocation, customer segmentation, or sales optimisation.

In summary, this dataset stands at the intersection of professional realism and academic utility. It offers MBA students a unique opportunity to apply analytical frameworks to a domain-specific challenge, developing insights that are both rigorous and practically relevant. Whether the goal is to visualise performance trends, build predictive models, or simulate business scenarios, this dataset provides a structured, engaging, and high-fidelity environment for applied learning in business analytics.

Dataset Description

The dataset consists of **1,000 records**, each representing a unique Senior Medical Representative's performance over a defined period. It combines **qualitative variables** (such as education, CRM usage, and therapeutic area) with **quantitative metrics** (such as monthly revenue, client feedback scores, and conversion rates), alongside several **derived variables** for deeper analysis.

The variables have been selected to reflect core components of pharmaceutical sales operations and to enable rich analytical use cases. They facilitate multiple types of analysis, including performance benchmarking, segmentation, efficiency assessments, and trend monitoring.

Purpose and Application

This dataset serves multiple academic and professional development objectives:

- To support **MBA-level education** in business analytics and data-driven decision-making.
- To provide a **realistic simulation** of operational data in pharmaceutical sales and marketing.
- To enable students to apply data analysis tools, techniques, and visualisation platforms (particularly **Power BI**) in a strategic context.
- To create a foundation for **AI-driven analysis**, such as predictive modelling and pattern recognition, within business environments.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Structure

The dataset is structured as a tabular file with each row representing one instance (a Senior Medical Representative's performance profile) and columns corresponding to the following variable categories:

- **Qualitative Variables**: 7
- **Quantitative Variables**: 10
- **Derived Variables**: 3

This structure supports robust data exploration, model building, and decision-support dashboards.

Qualitative Variables

1. **Region**: Assigned geographic area (North, South, East, West)
2. **Education**: Highest academic qualification (Diploma, Bachelor, Master, PhD)
3. **Experience**: Experience category (Junior 1-3 yr, Mid 4-6 yr, Senior 7-10 yr, Expert 10+ yr)
4. **Therapeutic Area**: Specialisation (Oncology, Cardiology, Endocrinology, General)
5. **Training Completed**: Indicates whether training was undertaken (Yes, No)
6. **CRM Usage**: Level of CRM adoption (None, Basic, Advanced)
7. **Doctor Satisfaction**: Ordinal classification of satisfaction (Low, Medium, High)

Quantitative Variables

1. **Monthly Revenue**: Total sales revenue per month (in US dollars)
2. **Calls Made**: Number of professional calls placed to healthcare providers
3. **In-Person Visits**: Number of visits to hospitals, clinics, and physicians
4. **Deals Closed**: Number of successfully completed sales agreements
5. **Average Deal Size**: Mean revenue per closed deal
6. **Conversion Rate**: Ratio of deals to total calls, expressed as a percentage
7. **Sales Cycle Days**: Average number of days taken to close a sale
8. **Client Feedback**: Score (1–5) based on customer evaluations

© 2025 Dr. Yasser Aly Yakout Shehata

9. **Retention Rate**: Percentage of customers retained over time
10. **Geographic Coverage**: Size of the representative's sales territory (in square kilometres)

Derived Variables

1. **Revenue Per Call**: Total monthly revenue divided by number of calls made; a productivity metric.
2. **Performance Band**: Categorical ranking of representatives based on monthly revenue (Low, Moderate, High, Top).
3. **Efficiency Score**: Ratio of deals closed to in-person visits, reflecting representative effectiveness.

Relevance to Power BI and AI for Business

This dataset is optimally structured for data ingestion and visualisation in Microsoft Power BI. It enables users to build interactive dashboards, define key performance indicators (KPIs), and generate dynamic filters based on experience level, geography, and performance tiers. Moreover, the inclusion of derived metrics and categorical segments allows for the deployment of AI models and business intelligence frameworks.

In the broader context of **AI for Business**, this dataset enables students and professionals to model and simulate decision-making scenarios, identify leading and lagging indicators, and assess strategic interventions like training or CRM adoption.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Distribution analysis of revenue and deal sizes
- Identification of high-performing regions or education levels
- Correlation analysis between CRM usage, training, and sales outcomes
- Experience level comparisons across performance metrics

B. Power BI Visualisation

- Sales funnel representation (Calls → Visits → Deals → Revenue)
- Performance dashboards segmented by therapeutic area and geography
- Trend analysis and forecasting based on retention or revenue

© 2025 Dr. Yasser Aly Yakout Shehata

- Drill-downs by CRM usage and doctor satisfaction levels

C. AI-Driven Business Insights

- Predictive modelling for monthly revenue using feedback and experience variables
- Classification models for performance band prediction
- Clustering representatives based on performance and engagement patterns
- Simulation of ROI for training programmes or CRM upgrades

Conclusion

This dataset provides a robust platform for applying advanced business analytics concepts within a structured, professionally relevant context. By simulating the sales and operational data of Senior Medical Representatives in the pharmaceutical industry, it bridges academic instruction with real-world practice. It equips MBA students with the opportunity to perform exploratory and visual analytics, build data narratives in Power BI, and develop AI-driven insights that support strategic business decisions. Through the combination of thoughtfully curated variables and realistic inter-variable relationships, the dataset enhances applied learning and analytical proficiency.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 49: SAP Senior Manager KPI

Introduction

In the evolving landscape of data-driven decision-making, the ability to transform operational data into actionable insights has become an essential competency for business leaders. This is especially true in complex, highly regulated industries such as pharmaceutical manufacturing, where enterprise resource planning (ERP) systems like SAP are not only foundational to operations but central to compliance, quality assurance, and strategic performance management. At the nexus of this environment stands the **SAP Senior Manager**, a role that demands both technological fluency and strategic oversight.

This dataset has been purposefully developed to replicate the real-world conditions and responsibilities faced by SAP Senior Managers in the pharmaceutical manufacturing sector. It serves as a core resource for MBA students enrolled in the ***Business Analytics for Decision Making*** **course**, specifically those undertaking final projects that require integration of analytical theory with practical enterprise applications. By drawing upon the actual domains in which SAP Senior Managers operate such as inventory control, production planning, regulatory compliance, and quality management, the dataset provides students with a realistic and academically rigorous context in which to apply analytical tools and frameworks.

Pharmaceutical manufacturing represents a particularly rich context for analytics due to its stringent quality demands, dynamic supply chains, and dependence on interdependent data systems. The SAP Senior Manager navigates this complexity through a coordinated use of various SAP modules, including Materials Management (MM), Production Planning (PP), Warehouse Management (WM), Plant Maintenance (PM), and Quality Management (QM). This dataset mirrors that multifaceted environment by capturing variables that reflect operational efficiency, compliance outcomes, system performance, and user engagement.

Designed with analytical integrity and educational utility in mind, the dataset supports a variety of business analytics methodologies ranging from exploratory data analysis and data visualisation to predictive modelling and KPI forecasting. It is formatted to be compatible with Microsoft Power BI and AI platforms, enabling students to build dynamic dashboards and derive AI-powered insights grounded in enterprise data.

What sets this dataset apart is not only its structural richness and variable diversity but also the inclusion of **derived metrics** that simulate the kinds of performance indices used in real organisations. These include composite indicators such as **Compliance Efficiency**, **Training Impact Index**, and **System Performance Index**, metrics that challenge students to consider interdependencies, causal relationships, and strategic decision levers.

© 2025 Dr. Yasser Aly Yakout Shehata

In sum, this dataset offers a sophisticated yet accessible platform for MBA students to engage with realistic enterprise data, develop key analytical competencies, and prepare for leadership roles in data-driven business environments. It reflects both the technical rigor and strategic depth required of professionals operating at the intersection of information systems, regulatory oversight, and operational excellence.

Dataset Description

The dataset contains **1,000 records**, each representing a monthly performance snapshot from a pharmaceutical manufacturing site managed under the responsibility of an SAP Senior Manager. It combines **7 qualitative variables**, **10 quantitative metrics**, and **3 derived variables**, allowing for comprehensive business analytics across multiple operational and compliance dimensions.

Purpose and Application

The dataset is intended to:

- Simulate enterprise data relevant to ERP-driven pharmaceutical operations.
- Support MBA students in applying analytics techniques to realistic management problems.
- Enable hands-on practice in exploratory data analysis (EDA), KPI evaluation, performance benchmarking, and strategic insight generation.
- Encourage the integration of Power BI visualisation and AI-based modelling into business decision-making workflows.

By working with this dataset, students will gain practical exposure to data modelling, performance evaluation, and transformation initiatives central to the SAP Senior Manager's role.

Dataset Structure

Each row in the dataset corresponds to one monthly operational report from a site. The dataset is organised as follows:

- **7 Qualitative Variables**: Categorical descriptions of site characteristics, user roles, and audit readiness.
- **10 Quantitative Variables**: Measurable operational and compliance KPIs.
- **3 Derived Variables**: Calculated indicators designed to enrich analysis and model deeper relationships.

© 2025 Dr. Yasser Aly Yakout Shehata

Qualitative Variables

1. **Region** – Site location: North, South, East, West
2. **Plant Type** – Type of facility: API Plant, Packaging Plant, Formulation Plant
3. **SAP Module Focus** – Core module in use: MM, PP, WM, PM, QM
4. **Training Completed** – Whether staff completed SAP training: Yes or No
5. **Process Complexity Level** – Site's process complexity: Low, Medium, High
6. **Audit Readiness** – Preparedness for audit: Prepared, Pending, Delayed
7. **System Access Level** – Role-based access: Operator, Supervisor, Admin

Quantitative Variables

1. **Inventory Turnover** – Number of stock cycles per month
2. **Order Accuracy (%)** – Proportion of accurately fulfilled orders
3. **On-Time Delivery (%)** – Timely completion of deliveries
4. **Inspection Pass Rate (%)** – Quality inspections passed successfully
5. **Compliance Score (%)** – SFDA and GMP compliance assessment
6. **Cycle Time (days)** – Time required to complete the production process
7. **Master Data Quality (%)** – Accuracy and completeness of system master data
8. **System Uptime (%)** – SAP system availability
9. **Training Hours** – Monthly hours spent on training initiatives
10. **User Satisfaction Score (1–5)** – Internal users' system usability rating

Derived Variables

1. **Compliance Efficiency** – Ratio of compliance score to cycle time
2. **Training Impact Index** – Change in master data quality per training hour
3. **System Performance Index** – Composite measure of system uptime, order accuracy, and user satisfaction

These derived metrics offer deeper analytical opportunities for evaluating performance drivers and strategic levers.

© 2025 Dr. Yasser Aly Yakout Shehata

Relevance to Power BI and AI for Business

This dataset is well-suited for Power BI visualisation, as it includes structured categorical variables for slicers and filters, and KPI-style numerical metrics for dashboards. The relationships between variables enable pattern recognition and performance tracking through heatmaps, trendlines, and drill-throughs.

Moreover, the inclusion of derived variables supports AI-driven business applications, such as predictive modelling and clustering for performance forecasting or strategic grouping.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identification of variable distributions, outliers, and central trends
- Correlation analysis between compliance, training, and performance metrics
- Segmentation of performance by plant type or region

B. Power BI Visualisation

- KPI dashboards tracking operational health and system uptime
- Training versus performance charts across modules or access levels
- Audit readiness reports and compliance trend visualisations

C. AI-Driven Business Insights

- Predictive models to forecast audit readiness based on training and compliance
- Classification models to identify high versus low-performing sites
- Clustering analysis to group sites based on system performance or operational complexity

Conclusion

This dataset represents a robust, realistic, and pedagogically valuable simulation of enterprise-level operations within a pharmaceutical manufacturing context. Centred on the strategic role of the SAP Senior Manager, it enables MBA students to practise advanced business analytics, visual storytelling, and AI-driven decision support. The combination of qualitative descriptors, operational KPIs, and enriched derived variables ensures that learners can explore, analyse, and generate insights that mirror the challenges and opportunities found in actual business environments. Through this dataset, students develop both the technical capabilities and strategic mindset necessary for leading data-informed initiatives in regulated, process-intensive industries.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 50: Senior Medical Representative Analytics

Introduction

The contemporary pharmaceutical industry operates within a dynamic ecosystem that demands not only clinical excellence but also operational agility and data-informed decision-making. In this environment, the role of the **Senior Medical Representative** has expanded beyond traditional sales functions to encompass strategic responsibilities including territory optimisation, digital engagement, regulatory compliance, and market intelligence. No longer limited to transactional interactions, Senior Medical Representatives are increasingly expected to serve as analytical professionals who leverage enterprise data to improve outreach, enhance compliance, and drive performance outcomes.

This evolution reflects a broader transformation in pharmaceutical field operations, driven by enterprise systems such as SAP, the integration of CRM analytics, and the rising importance of customer-centric strategies. These shifts require professionals who are capable of synthesising large volumes of performance data and translating them into actionable business insights. For MBA students, especially those pursuing studies in ***Business Analytics for Decision Making***, it is essential to develop a practical understanding of how these roles and responsibilities are quantified, assessed, and improved through data.

The **Senior Medical Representative Analytics Dataset** has been crafted to meet this pedagogical and professional need. It simulates a realistic pharmaceutical sales and compliance environment in which students can explore the multifaceted performance indicators that govern strategic and operational success. Drawing on realistic benchmarks from the pharmaceutical industry and informed by actual key performance indicators (KPIs), the dataset presents a robust foundation for learning how analytics supports decision-making in regulated and competitive sectors.

The design of this dataset allows for cross-functional learning, enabling students to explore not only sales and marketing performance but also supply chain efficiency, training effectiveness, customer satisfaction, and digital transformation readiness. With a combination of qualitative and quantitative variables, along with analytically valuable derived metrics, the dataset supports the development of competencies in exploratory data analysis (EDA), dashboard visualisation using Power BI, and the application of AI models in business contexts.

By situating the Senior Medical Representative as the central analytical agent in this dataset, MBA students are encouraged to think critically about the roles, responsibilities, and performance challenges of professionals who must bridge clinical knowledge with

© 2025 Dr. Yasser Aly Yakout Shehata

commercial execution. In this way, the dataset contributes not only to academic assessment but also to the cultivation of analytical mindsets capable of navigating complex business ecosystems in the pharmaceutical and life sciences industries.

Dataset Description

The **Senior Medical Representative Analytics Dataset** comprises **1,000 individual records**, each simulating the performance profile of a Senior Medical Representative operating in a regional pharmaceutical context. Each record includes both categorical attributes and continuous performance indicators, modelled using plausible distributions grounded in pharmaceutical sales and operational benchmarks.

In addition to primary variables, the dataset includes **derived metrics** to enable multidimensional analysis. These metrics reflect inter-variable relationships, such as the influence of training on customer engagement or the relationship between compliance and cycle efficiency. The dataset is structured to facilitate immediate use in EDA tools and dashboard platforms such as Microsoft Power BI.

Purpose and Application

The primary purpose of this dataset is to enable MBA students to:

- Practise **exploratory data analysis** using realistic, multidimensional data.
- Construct **interactive dashboards and KPIs** using Power BI.
- Apply **descriptive and predictive analytics** to derive business insights.
- Examine **inter-variable relationships**, operational trade-offs, and strategic levers.

It is designed to replicate the analytical challenges encountered in pharmaceutical field operations, with an emphasis on territory performance, digital transformation, customer engagement, and training impact.

Dataset Structure

- **Total Records**: 1,000
- **Qualitative Variables**: 7 categorical fields representing regional, operational, and managerial conditions.
- **Quantitative Variables**: 10 continuous or count-based fields capturing operational KPIs.
- **Derived Variables**: 3 analytical metrics created from the interaction of core variables.

© 2025 Dr. Yasser Aly Yakout Shehata

Qualitative Variables

1. **Region**: Geographic allocation of the representative (North, South, East, West).
2. **ProductLine**: Portfolio specialisation (Chronic Care, Acute Care, Specialty Medicines).
3. **ExperienceLevel**: Classification of tenure and role maturity (Junior, Mid, Senior).
4. **TrainingStatus**: Completion status of mandatory or elective training (Completed, Pending).
5. **SalesChannel**: Dominant sales outlet (Private, Government, International, CMO, Other).
6. **AuditPreparedness**: Current audit status (Ready, Under Review, Not Ready).
7. **TerritorySize**: Scale of coverage area (Small, Medium, Large).

Quantitative Variables

1. **Monthly_Sales_Volume**: Average monthly sales in local currency.
2. **Customer_Satisfaction_Score**: Mean satisfaction rating on a scale of one to five.
3. **KPI_Achievement_Rate**: Percentage of internal KPIs achieved.
4. **Training_Hours**: Hours spent on formal training in the last cycle.
5. **DSO_Days**: Days Sales Outstanding, a measure of collection cycle efficiency.
6. **Inventory_Days**: Days of stock turnover.
7. **Digital_Adoption_Index**: Proxy measure of digital system usage.
8. **EBITDA_Margin**: Earnings Before Interest, Tax, Depreciation and Amortisation, as a percentage.
9. **Revenue_Growth_Rate**: Year-on-year sales growth.
10. **Doctor_Engagements_Per_Month**: Count of field engagements with healthcare professionals.

Derived Variables

1. **Cash_Conversion_Cycle**: Derived from inventory days and DSO, minus fixed payables cycle.
2. **SalesEfficiency**: Sales performance per engagement with doctors.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **EngagementToTrainingRatio**: Measure of return on training hours invested.
4. **ComplianceEfficiency**: KPI achievement rate relative to financial compliance cycle.

Relevance to Power BI and AI for Business

This dataset is formatted and labelled for seamless integration with **Power BI**, where variables can be immediately used in dashboards, charts, and filters. Metrics such as revenue growth, customer satisfaction, and KPI achievements can be visualised through gauges, timelines, and comparative bar charts. The derived variables also lend themselves to multi-layered analysis such as **performance clustering**, **trend segmentation**, and **forecasting**.

In the context of **AI for Business**, this dataset provides a robust foundation for:

- **Predictive modelling** (e.g., forecasting sales volume).
- **Classification** (e.g., audit readiness prediction).
- **Clustering** (e.g., territory or product segmentation).

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identification of outliers and performance gaps.
- Correlation mapping across KPIs and operational inputs.
- Disaggregation of metrics by sales channel, product line, or region.

B. Power BI Visualisation

- Creation of role-specific dashboards for senior medical representatives.
- Interactive filters by training status, audit preparedness, and product line.
- Dynamic visuals for strategic insights and operational tracking.

C. AI-Driven Business Insights

- Feature engineering using derived metrics.
- Model training for identifying high-performing profiles.
- Simulation of digital transformation impact on financial efficiency.

© 2025 Dr. Yasser Aly Yakout Shehata

Conclusion

The **Senior Medical Representative Analytics Dataset** presents an academically rigorous and professionally realistic platform for developing business analytics competencies in the pharmaceutical domain. With a structured combination of raw and derived metrics, this dataset facilitates both foundational and advanced learning. It supports key MBA course outcomes by enabling students to explore, visualise, and extract insights from data in ways that mirror real-world pharmaceutical operations and strategic planning.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 51: Polyethylene Operations

Introduction

In the evolving landscape of industrial manufacturing, data-driven decision-making has become a cornerstone of operational excellence. As organisations embrace the integration of business intelligence and artificial intelligence within their core processes, the role of analytical professionals capable of interpreting complex operational data has grown in significance. The ***Polyethylene Operations Dataset*** has been carefully constructed to serve as a comprehensive, realistic, and pedagogically valuable resource for MBA students enrolled in courses such as *Business Analytics for Decision Making*. It aims to bridge the gap between academic theory and practical industrial application by offering a rich dataset derived from a simulated polyethylene production environment.

This dataset is contextualised within the daily operational responsibilities of a **Sector Manager, Polyethylene Department / Shift Supervisor Assistant**, a role that encapsulates the challenges of overseeing plant operations, managing reactor stability, ensuring product quality, and maintaining efficiency under varying conditions. The dataset emulates production records typically monitored by professionals in such positions, capturing a range of process variables that influence quality control outcomes and production performance.

Unlike synthetic datasets that are disconnected from real-world scenarios, this resource is rooted in the manufacturing sector's realities, making it highly relevant for students seeking to apply analytical techniques to real business problems. It enables learners to explore interrelationships among process parameters, operator characteristics, environmental factors, and product quality, thereby facilitating a holistic view of data-driven management in a manufacturing setting.

Moreover, the dataset is specifically tailored to support the development of core competencies in **Exploratory Data Analysis (EDA)**, **interactive visualisation using Power BI**, and the application of **artificial intelligence techniques for business insight generation**. These learning outcomes align directly with the skills required of contemporary business analysts, operations managers, and AI strategy consultants working in technology-enhanced industrial contexts.

By working with this dataset, students will not only develop technical proficiency in handling, interpreting, and visualising data but will also gain insight into how analytics can inform strategic and operational decisions in high-stakes industrial environments. This ensures that the analytical learning process is not only technical but also contextual, purpose-driven, and aligned with the realities of business leadership in technologically advanced sectors.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Description

The dataset contains **1,000 production records**, each representing a unique operational event during polyethylene production. It combines both **qualitative descriptors** of industrial conditions and **quantitative metrics** linked to production performance, quality control, and energy efficiency.

Designed for practical use in data analytics education, it includes variables commonly tracked in process industries such as chemical concentration, reactor stability, and product output measures. Two **derived variables** add analytical depth and simulate real-world performance indicators used in manufacturing analytics.

Purpose and Application

This dataset serves several key purposes:

- **Educational Use**: For MBA students to apply descriptive analytics, visualisation, and AI techniques in real-world operational contexts.
- **Operational Insight**: Enables simulation of managerial decision-making using empirical production and quality data.
- **Tool Integration**: Supports implementation within Power BI and AI frameworks for business insights, forecasting, and root-cause analysis.

Dataset Structure

- **Records**: 1,000 individual production logs
- **Variables**: 19 total (7 qualitative, 10 quantitative, 2 derived)
- **Format**: Tabular CSV, suitable for use in Excel, Power BI, Python, and R

Qualitative Variables

1. **Shift** – Type of operational shift (Morning, Evening, Night)
2. **Operator_Level** – Experience level of shift operator (Junior, Mid Level, Senior)
3. **Reactor_Status** – Operational status of the reactor (Stable, Unstable, Maintenance)
4. **Product_Grade** – Designated commercial grade of the polyethylene produced
5. **Catalyst_Type** – Type of catalyst used (Ziegler Natta, Chromium)
6. **Quality_Check_Result** – Outcome of the quality inspection (Pass, Fail, Retest)
7. **Weather_Condition** – Ambient environmental condition during operation

© 2025 Dr. Yasser Aly Yakout Shehata

Quantitative Variables

1. **Production_Volume_MT** – Polyethylene output per batch in metric tonnes
2. **Reactor_Temperature_C** – Operating temperature of the reactor in Celsius
3. **Reactor_Pressure_bar** – Pressure inside the reactor measured in bars
4. **Ethylene_Concentration_pct** – Percentage of ethylene in the reactor mixture
5. **Hydrogen_Ratio_ppm** – Parts per million of hydrogen in the reaction process
6. **Energy_Consumption_kWh** – Kilowatt-hours used during the production cycle
7. **Equipment_Downtime_min** – Minutes of unplanned downtime
8. **Melt_Flow_Index** – Flowability of the polymer melt (product property)
9. **Tensile_Strength_MPa** – Tensile strength of the final product in megapascals
10. **Elongation_pct** – Percent elongation at break, indicating product flexibility

Derived Variables

1. **Efficiency_Ratio** – A measure of operational efficiency, calculated as Production Volume divided by Energy Consumption
2. **Quality_Category** – Categorised as High, Moderate, or Low, based on thresholds for Melt Flow Index, Tensile Strength, and Elongation

Relevance to Power BI and AI for Business

This dataset is designed to be directly importable into **Power BI**, where users can:

- Build interactive dashboards tracking real-time performance
- Monitor KPIs such as efficiency, quality trends, and reactor stability
- Segment data by shift, operator level, or weather for managerial insights

In the context of **AI for Business**, the dataset can be used to:

- Train classification models to predict quality categories
- Detect anomalies in operational conditions
- Optimise decision-making through predictive analytics and simulations

© 2025 Dr. Yasser Aly Yakout Shehata

Applications in Business Analytics and AI

A. **Exploratory Data Analysis (EDA)**
   Students can perform EDA to uncover relationships between reactor conditions, catalyst types, operator experience, and resulting product quality. Descriptive statistics, distribution plots, and correlation matrices can be used to interpret key performance drivers.
B. **Power BI Visualisation**

With integrated qualitative and quantitative fields, the dataset supports:

- Time-series visualisation of energy and volume metrics
- Bar and line charts for shift-based efficiency comparisons
- Matrix visuals to evaluate catalyst impact on product performance
- Slicers and filters for multi-dimensional analysis by category

C. AI-Driven Business Insights

The dataset allows students and practitioners to:

- Apply supervised learning (e.g., decision trees) for quality prediction
- Use clustering to group similar production scenarios
- Develop prescriptive insights by modelling what-if scenarios related to efficiency or quality

Conclusion

The *Polyethylene Operations Dataset* exemplifies the convergence of industrial process data with business intelligence. Crafted for business analytics education, it contextualises operational decision-making in the role of a sector manager overseeing polyethylene production. With its structured design, derived insights, and multidimensional features, the dataset serves as a powerful resource for applying EDA, Power BI visualisation, and AI models in solving real-world business challenges.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 52: Accounts Receivable Performance Evyap

Introduction

In today's data-driven business environment, financial decision-making increasingly relies on the ability to extract, analyse, and interpret operational data in a way that leads to measurable improvements in performance. Nowhere is this more evident than in the management of accounts receivable, where timely collections, credit risk assessment, and process optimisation directly affect an organisation's liquidity, working capital efficiency, and customer relationships. The **Accounts Receivable Performance Dataset** has been developed with these realities in mind, serving as an applied learning resource for MBA students pursuing their capstone project in the ***Business Analytics for Decision Making*** **course**.

This dataset specifically reflects the operational responsibilities and analytical requirements of an **Accounts Receivable Supervisor at Evyap Egypt**, a multinational company engaged in personal care and hygiene products. Supervisors in this role are tasked with overseeing credit sales, managing dispute resolution processes, liaising with internal and external stakeholders, and ensuring timely cash collections from domestic and international clients. The dataset emulates this real-world context by offering granular, transaction-level data that mirrors the daily operational touchpoints and key performance indicators encountered in a modern accounts receivable function.

Importantly, the dataset is not merely a collection of static financial records. It has been carefully engineered to include logical relationships, industry-aligned variables, and derived performance indicators that promote deeper analytical reasoning. Each record encapsulates a complex set of business conditions ranging from regional market dynamics and payment methods to collector assignments and dispute flags, thereby providing a rich foundation for analysis. This structure supports a wide range of analytical techniques, from basic descriptive statistics and trend analysis to predictive modelling and AI-based decision support.

Designed for compatibility with tools such as **Power BI**, **Python**, and **machine learning frameworks**, the dataset enables students to build dynamic dashboards, uncover hidden patterns, and simulate data-informed management decisions. It also includes derived metrics such as *Best Possible DSO* (BPDSO), *Delinquent DSO* (DDSO), and *Efficiency Index*, which elevate the dataset from a transactional record to a performance evaluation framework. Additionally, a derived classification variable, *AR_Performance_Level*, offers learners an accessible way to perform segmentation, risk analysis, and policy prioritisation.

© 2025 Dr. Yasser Aly Yakout Shehata

In essence, this dataset empowers students to translate financial data into business insight. It allows them to adopt the mindset of an operational decision-maker, using analytics not just to report outcomes but to shape them. By working within the framework of a real organisational role and sector, students will develop skills that are both academically rigorous and professionally relevant, equipping them to drive data-enabled financial performance in diverse business contexts.

Dataset Description

This dataset consists of **1,000 records**, each reflecting a discrete credit transaction or receivable event. It includes a mix of **qualitative descriptors** and **quantitative performance metrics** typically monitored by receivables departments in medium to large enterprises. Simulating daily or weekly performance logs, the data captures critical financial indicators, customer-level characteristics, and operational costs associated with receivables management.

To enhance interpretability and analytical scope, the dataset incorporates **derived variables** that combine multiple inputs to reflect strategic insights—such as collection effectiveness relative to cost and optimal collection timing. Additionally, a tiered classification system ranks account performance levels to support segmentation, decision support, and forecasting exercises.

Purpose and Application

The dataset is specifically designed to:

- Support MBA students in practising **Exploratory Data Analysis (EDA)**, **dashboard development**, and **descriptive modelling**.
- Provide a platform for applying **Power BI** in a financial operations context.
- Enable the application of **AI techniques** for performance prediction, risk assessment, and pattern recognition.
- Simulate real-world scenarios encountered by an Accounts Receivable Supervisor, fostering strategic thinking through data.

Dataset Structure

Feature Type	Count	Purpose
Qualitative Variables	7	Capture operational context and categorical drivers
Quantitative Variables	10	Measure financial performance and risk indicators

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Variables	3	Integrate logic-based insights for enhanced analytical interpretation
Total Variables	20	Provide a balanced, interpretable structure for business analytics use

Qualitative Variables

1. **Customer_Segment** – Indicates client type (Retail, Wholesale, Distributor)
2. **Invoice_Type** – Differentiates between domestic and export transactions
3. **Region** – Identifies geographic market (e.g. Cairo, Alexandria, Delta)
4. **Credit_Terms** – Payment terms (30 days, 60 days)
5. **Collector_Level** – Assigned collector seniority (Junior, Senior)
6. **Payment_Method** – Mode of payment (Bank Transfer, Credit Card, Cash)
7. **Dispute_Flag** – Whether the invoice is flagged for dispute (Yes, No)

Quantitative Variables

1. **Credit_Sales** – Sales amount under credit terms (in local currency)
2. **AR_Beginning** – Opening accounts receivable balance
3. **AR_Ending** – Closing accounts receivable balance
4. **DSO (Days Sales Outstanding)** – Time taken to collect receivables
5. **ADD (Average Days Delinquent)** – Payment delays beyond expected terms
6. **CEI (Collection Effectiveness Index)** – Percent of successful collection
7. **AR_Turnover** – Number of receivable cycles completed per period
8. **Num_Revised_Invoices** – Frequency of invoice amendments
9. **Bad_Debt_Amount** – Value of receivables written off
10. **Collection_Cost** – Operational cost of collecting receivables

Derived Variables

1. **BPDSO (Best Possible DSO)** – Hypothetical minimum DSO assuming perfect collections
2. **DDSO (Delinquent DSO)** – Difference between actual and best-case DSO

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Efficiency_Index** – CEI adjusted by cost, indicating net operational effectiveness
4. **AR_Performance_Level** – Categorical performance ranking (High, Moderate, Low)

Relevance to Power BI and AI for Business

This dataset is optimally structured for integration into **Power BI** dashboards and **AI applications** in business decision-making. Key strengths include:

- Compatibility with data visualisation tools (e.g. slicers, charts, KPI cards)
- Structured relationships between variables for easy modelling
- Balanced variable distribution for clean analytics pipelines

The data format supports time-series analysis, heatmap creation, customer segmentation, and regional performance comparisons, aligning with modern data-driven business intelligence practices.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)
Students can uncover patterns in payment performance, detect anomalies in delinquency behaviour, and analyse how invoice characteristics affect collection rates.

B. Power BI Visualisation
Using the dataset in Power BI, users can:

- Build interactive dashboards showing DSO trends across regions and segments
- Visualise collection efficiency relative to operational cost
- Segment customers by risk level or performance tier

C. AI-Driven Business Insights
The dataset supports:

- **Classification models** predicting high-risk accounts
- **Clustering algorithms** for grouping by behaviour patterns
- **Regression models** to forecast collection delays or bad debt probabilities

Conclusion

The **Accounts Receivable Performance Dataset** offers a comprehensive, well-structured platform for MBA students and business professionals to apply critical analytical skills in a

© 2025 Dr. Yasser Aly Yakout Shehata

financial operations setting. Its combination of categorical and numeric data, performance-based derivations, and alignment with real-world job responsibilities ensures practical relevance. Whether used for dashboard design, AI modelling, or strategic financial analysis, the dataset serves as an invaluable tool in preparing learners for data-informed leadership in finance and operations.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 53: Al Nahdi Operational Performance

Introduction

In an era defined by rapid digital transformation and evolving consumer expectations, the role of data analytics has become integral to effective business leadership. Nowhere is this more evident than in large-scale retail healthcare operations, where success hinges on the ability to derive actionable insights from complex and high-volume datasets. As organisations like **Al Nahdi Medical Company**, the largest pharmacy chain in the Kingdom of Saudi Arabia continue to scale across geographies and product verticals, the demand for analytically empowered managers has intensified. The **Operations Assistant Area Manager** plays a vital role in this structure, tasked with driving regional performance across hundreds of outlets through strategic oversight, operational efficiency, and customer-centric decision-making.

The **Al Nahdi Operational Performance Dataset** has been purposefully developed as a synthetic but realistic analytical tool tailored for MBA students enrolled in the ***Business Analytics for Decision Making*** **course**. It reflects the multifaceted challenges and decision variables that an **Operations Assistant Area Manager** must confront on a weekly basis, spanning areas such as inventory turnover, customer satisfaction, promotional effectiveness, revenue optimisation, staff allocation, and logistics performance. Each record in the dataset represents a realistic aggregation of operational metrics at the regional level, offering both raw performance indicators and analytically derived fields to enable advanced interpretation.

What distinguishes this dataset is not merely the diversity of its variables, but the depth of its design. In addition to capturing standard operational metrics, the dataset integrates calculated indices such as **Revenue per Store**, **Cost Efficiency Index**, and **Service Level Percentage**, offering students a meaningful opportunity to investigate performance differentials and derive key insights from relational patterns. Moreover, the inclusion of a composite **Performance Tier** variable supports segmentation and classification tasks, reinforcing the learning outcomes associated with AI-enhanced business decision-making.

Designed for compatibility with contemporary analytical tools such as Power BI, Python, and Excel, this dataset provides a foundation for interactive visual exploration, hypothesis testing, and strategic simulation. It encourages students to think not just as data analysts, but as decision-makers, individuals who must interpret data in real time to optimise operations, maximise value, and maintain alignment with broader corporate objectives.

Through this dataset, students will gain critical experience in managing the complexity and ambiguity that characterise real-world business environments. By working within a scenario

© 2025 Dr. Yasser Aly Yakout Shehata

that mirrors the actual responsibilities of a mid-level operations leader at a major healthcare retailer, they will be better prepared to apply business analytics in professional contexts where insight, speed, and accuracy are essential.

Dataset Description

The dataset comprises **1,000 weekly records**, each representing a regional operational summary managed by an Assistant Area Manager. The data simulates diverse operational scenarios across multiple cities and store formats in Saudi Arabia. It includes:

- Qualitative dimensions such as store type, promotion type, inventory system, and staff levels
- Quantitative indicators of sales performance, inventory dynamics, customer satisfaction, and logistics efficiency
- Derived metrics to enable deeper interpretation and benchmarking

All values are logically coherent and represent realistic variability expected in a complex, multi-branch retail organisation.

Purpose and Application

This dataset was created to fulfil three primary educational objectives:

1. **Enable hands-on application of business analytics tools**

   Students will learn to prepare, clean, explore, and analyse a complex dataset using EDA principles and visualisation techniques.

2. **Simulate real-world decision-making scenarios**

   The data reflects the challenges of balancing customer service, revenue growth, inventory control, and staff turnover across multiple retail locations.

3. **Prepare students for AI-assisted managerial roles**

   By incorporating derived fields and classification labels, students can test predictive models and develop AI-driven decision frameworks relevant to operations management.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset Structure

Type of Variable	Count	Description
Qualitative Variables	7	Describe structural and categorical attributes of the operational units
Quantitative Variables	10	Represent measurable KPIs and operational performance metrics
Derived Variables	4	Enhance analytical depth and enable KPI benchmarking and classification

Qualitative Variables

1. **Area Region** – Operational zone (e.g., Riyadh, Jeddah, Makkah)
2. **Store Type** – Type of outlet (Pharmacy, Polyclinic, Omni-health)
3. **Promotion Type** – Marketing strategy in effect (None, Pharma discount, FMCG campaign)
4. **Staff Level** – Managerial role (Junior Manager, Senior Manager)
5. **Inventory System** – Technology platform used (Manual, Oracle-based, Infor-based)
6. **Delivery Mode** – Fulfilment channel (Central Distribution Centre, Direct supplier)
7. **Stockout Flag** – Binary indicator of whether stockouts occurred (Yes/No)

Quantitative Variables

1. **Number of Stores** – Total outlets managed in a region
2. **Weekly Revenue (SAR)** – Aggregate sales revenue
3. **Sales Growth Percentage** – Week-on-week sales growth
4. **Customer Satisfaction Score** – Feedback index (0 to 100)
5. **Staff Turnover Percentage** – Proportion of employees replaced weekly
6. **Inventory Turnover Ratio** – Inventory sold versus held
7. **Stockout Count** – Frequency of out-of-stock incidents
8. **Promotional Spend (SAR)** – Marketing cost incurred
9. **Average Delivery Days** – Logistics efficiency metric
10. **Shrinkage (SAR)** – Inventory loss due to waste or theft

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Variables

1. **Revenue per Store** – Weekly revenue divided by number of stores
2. **Cost Efficiency Index** – (Revenue – Promotions – Shrinkage) divided by promotional spend
3. **Service Level Percentage** – Measure of product availability based on stockout frequency
4. **Performance Tier** – Categorisation of overall area performance (High, Moderate, Low) based on combined KPIs

Relevance to Power BI and AI for Business

This dataset is engineered to support the full lifecycle of analytics and decision-making tasks taught in modern MBA programmes. It is particularly suited to the following applications:

- **Data transformation and visualisation using Power BI**
- **KPI dashboard design for operational monitoring**
- **Trend analysis, outlier detection, and business segmentation**
- **Predictive modelling and AI-based recommendations**

The mix of descriptive, diagnostic, and predictive features offers extensive scope for analysis using AI tools such as regression, clustering, classification, and decision trees.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

Students can explore relationships between promotional spending, revenue growth, and customer satisfaction. They can identify patterns in store performance across regions and examine operational trade-offs.

B. Power BI Visualisation

The dataset enables construction of region-based heatmaps, KPI dashboards, dynamic slicers, and drill-down visualisations for performance comparison across managerial levels and promotional strategies.

C. AI-Driven Business Insights

Students can build machine learning models to:

- Predict low performance areas based on staffing, inventory, and delivery variables

© 2025 Dr. Yasser Aly Yakout Shehata

- Segment store types for targeted marketing
- Forecast future revenue growth or stockout risk

Conclusion

The **Al Nahdi Operational Performance Dataset** offers a comprehensive, realistic, and educationally rich environment for MBA students aiming to master the intersection of business operations and data analytics. By simulating the decisions faced by an **Operations Assistant Area Manager at Al Nahdi Medical Company**, this dataset provides a dynamic platform for skill development in descriptive analytics, data visualisation, and AI-enhanced business strategy. It enables learners to go beyond theory and engage in meaningful, actionable insights that mirror professional responsibilities in high-performance retail operations.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 54: Biomedical Equipment Maintenance Analytics

Introduction

The *Biomedical Equipment Maintenance Analytics* dataset represents a high-fidelity simulation of operations within a military hospital's biomedical engineering department. It has been carefully crafted to serve as a capstone project for MBA students pursuing the ***Business Analytics for Decision Making*** **course**. Set within a realistic organisational context, this dataset places the *Chief Biomedical Engineer – Head of Lab and Technical Supervisor* at the centre of a complex ecosystem of healthcare technology management, offering students the opportunity to step into a decision-making role where engineering meets analytics, and operational data informs strategic choices.

Biomedical engineering departments in healthcare institutions play a critical role in ensuring the safety, reliability, and availability of life-saving medical equipment. In such environments, poorly maintained or inefficiently managed devices can result not only in operational downtime but also in significant patient risk and financial inefficiencies. The Chief Biomedical Engineer is tasked with navigating these challenges while maintaining compliance with stringent regulatory standards, coordinating maintenance teams, and managing equipment life cycles within strict budgetary constraints. This dataset captures these dimensions, providing a nuanced and multifaceted view of biomedical equipment management through structured, analysable data.

Designed specifically for academic and applied learning, this dataset allows students to simulate the decision environment of technical leadership in healthcare. It incorporates a blend of categorical and continuous variables that reflect common performance indicators used in real-world maintenance management systems. The inclusion of derived metrics, such as *Repair Efficiency* and *Ownership Cost Per Year*, further strengthens its relevance to data-driven analysis and supports more sophisticated modelling, including cost optimisation, risk analysis, and predictive maintenance.

Moreover, the dataset is fully compatible with modern business intelligence tools, including Power BI, and is aligned with emerging applications of artificial intelligence in business. Students can use the dataset to practise exploratory data analysis (EDA), build interactive dashboards, and apply AI algorithms to identify inefficiencies, predict failures, and propose evidence-based recommendations. Through this hands-on engagement, the dataset bridges technical insights with business strategies, offering a comprehensive foundation for decision-making in asset-intensive, high-reliability sectors such as healthcare.

By exploring this dataset, MBA students are not only exposed to technical data management and analytics but are also encouraged to think critically about the organisational context,

© 2025 Dr. Yasser Aly Yakout Shehata

operational constraints, and strategic trade-offs faced by biomedical engineering leaders. This approach promotes holistic, systems-level thinking and prepares students to use data not simply for insight, but for action.

Dataset Description

The dataset comprises 1,000 records representing individual biomedical devices maintained within a hospital setting. Each record contains both operational and contextual information, including device age, maintenance intervals, calibration history, cost metrics, and reliability scores. In addition to raw observations, the dataset features derived variables that allow for enhanced analytical interpretation, enabling students to explore performance drivers, optimise maintenance planning, and predict future outcomes.

All values are statistically generated based on plausible industry patterns, ensuring internal consistency and realistic distribution. The dataset is free from missing data and duplicates, and categorical variables have been cleaned and standardised for effective use in analytics tools.

Purpose and Application

The dataset is intended for educational and applied use in the following areas:

- Teaching and assessment in postgraduate Business Analytics courses
- Development of data storytelling and dashboard design skills using Power BI
- Application of AI and statistical methods for equipment failure prediction, cost forecasting, and performance optimisation
- Strategic decision support in asset-intensive environments such as hospitals, labs, and healthcare organisations

It also serves as a practical simulation for students aspiring to careers in data-driven healthcare management, operations analytics, or technical supervision.

Dataset Structure

- **Records**: 1,000
- **Variables**: 19 total
    - 7 qualitative (categorical)
    - 10 quantitative (numerical)
    - 2 derived (calculated from primary variables)

© 2025 Dr. Yasser Aly Yakout Shehata

- **Granularity**: Each row corresponds to a unique biomedical device under hospital supervision

Qualitative Variables

These variables represent device characteristics, operational environment, and management practices:

- **Device Type**: Category of medical device (e.g. Ventilator, Radiology)
- **Manufacturer Reputation**: Perceived quality tier (High, Medium, Low)
- **Spare Parts Availability**: Ease of obtaining parts (High, Medium, Low)
- **Training Completion**: Whether technicians are formally trained (Yes, No)
- **Compliance Status**: Regulatory adherence (Compliant, Non-compliant)
- **Maintenance Team**: Responsible party (In-house, External Vendor)
- **Criticality Level**: Operational importance (High, Medium, Low)

Quantitative Variables

These numerical values support cost, performance, and operational analysis:

- **Device Age**: Years since acquisition
- **Maintenance Interval**: Days between preventive maintenance actions
- **Time to Repair**: Hours required to complete corrective maintenance
- **Calibration Frequency**: Times per year the device is calibrated
- **Pass Rate Pre-Repair**: Initial performance rating (percentage)
- **Pass Rate Post-Repair**: Final performance rating (percentage)
- **Downtime Percentage**: Portion of time a device is unavailable
- **Spare Parts Cost**: Cost of spare parts per incident (USD)
- **Total Ownership Cost**: Annualised operational cost per device (USD)
- **MTBF (Mean Time Between Failures)**: Average operational uptime in hours

Derived Variables

These are calculated fields designed to enhance analytical depth:

- **Pass Rate Improvement**: Difference between post- and pre-repair performance

© 2025 Dr. Yasser Aly Yakout Shehata

- **Ownership Cost Per Year**: Total cost divided by device age
- **Repair Efficiency**: Improvement per hour of repair time

These derived metrics help assess cost-effectiveness, repair quality, and asset longevity.

Relevance to Power BI and AI for Business

This dataset is well-suited for visualisation and analytical modelling within Power BI and AI platforms. The presence of both categorical and numerical variables allows for the construction of dynamic dashboards, KPI visualisations, performance trend charts, and resource allocation heatmaps. The dataset also supports predictive modelling tasks such as estimating failure likelihood, predicting repair costs, and clustering devices by risk profile.

Applications in Business Analytics and AI

A. **Exploratory Data Analysis (EDA)**
   Students can perform data profiling, identify patterns in device performance and cost, and explore variability in maintenance efficiency across device types, vendors, and teams.
B. **Power BI Visualisation**
   Users can build dashboards to:

- Compare compliance rates across device categories
- Visualise cost versus reliability by manufacturer
- Track cumulative ownership costs by maintenance strategy

D. **AI-Driven Business Insights**
   Advanced learners may apply machine learning or AI techniques to:

- Predict next failure date or expected downtime
- Optimise maintenance intervals for high-criticality assets
- Segment devices for targeted maintenance investment

These applications mirror real-world decision-making practices in hospital management and technical infrastructure planning.

Conclusion

The *Biomedical Equipment Maintenance Analytics* dataset provides a rich, realistic, and structured foundation for learning and applying business analytics in a high-impact domain. Focused on the role of a Chief Biomedical Engineer managing complex hospital assets, it

© 2025 Dr. Yasser Aly Yakout Shehata

enables MBA students to integrate technical and business reasoning. Through EDA, visualisation, and AI modelling, learners develop the skills required to drive data-informed decisions in the evolving field of healthcare operations and intelligent maintenance management.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 55: IT Project Manager QCC Analytics

Introduction

The *IT Project Manager QCC Analytics* dataset has been meticulously developed to serve as a capstone learning resource for MBA students undertaking the course ***Business Analytics for Decision Making***. It immerses students in a simulated project management environment representative of the Information Technology department within the Abu Dhabi Quality and Conformity Council (QCC), a governmental body responsible for ensuring quality, safety, and compliance across various sectors. At the heart of this simulation lies the role of the *Project Manager*, a strategic professional responsible for overseeing the planning, execution, monitoring, and completion of digital transformation initiatives aligned with QCC's operational and regulatory goals.

The dataset has been conceptualised to reflect the multifaceted challenges faced by IT project managers in large public institutions. These include managing diverse project portfolios, balancing stakeholder interests, mitigating risks, adhering to strict compliance frameworks, and delivering projects within scope, time, and budget constraints. The simulated data captures the nuances of project governance by incorporating a wide range of performance and contextual variables such as project phase, stakeholder feedback, compliance status, resource allocation, and change request management.

This dataset enables students to approach complex project-related questions through a data-driven lens. They are invited to engage with real-world inspired challenges: What conditions contribute most to project delays? How can the effectiveness of change management be quantified? What patterns emerge across high-risk projects, and how might these be predicted or mitigated? By addressing such questions through structured analysis, students not only develop proficiency in analytical tools and methods but also gain insights into the strategic role of analytics in project leadership.

In particular, the dataset is tailored for use in Exploratory Data Analysis (EDA), descriptive and diagnostic analytics, and visualisation techniques using Power BI. Moreover, the data schema supports advanced applications in artificial intelligence, such as predictive modelling and pattern recognition. The inclusion of derived metrics, *Schedule Variance*, *Budget Variance*, *Performance Index*, and *Change Effectiveness*—further enriches the analytical potential, offering deeper perspectives into performance drivers and outcomes.

In today's increasingly data-dependent organisations, project managers are expected not only to manage tasks and teams but to translate performance data into actionable insights. This dataset is designed to cultivate those competencies, preparing MBA students to become future leaders who are capable of making evidence-based decisions, optimising

© 2025 Dr. Yasser Aly Yakout Shehata

project outcomes, and contributing to institutional effectiveness through the intelligent use of data. As such, it represents more than a static collection of numbers; it is a dynamic learning tool that bridges theory and practice in the field of data-informed project management.

Dataset Description

The dataset comprises 1,000 records, each representing an IT project phase, milestone, or discrete task completed under the supervision of the QCC IT department. It includes a balanced mixture of qualitative descriptors and quantitative performance indicators, capturing the diversity and complexity of projects in a regulatory context. Data points cover project timelines, financial performance, stakeholder engagement, risk activity, and change control outcomes. All records have been validated, cleaned, and enhanced to ensure analytical accuracy and suitability for visualisation and modelling.

The dataset is entirely synthetic yet structured to mimic authentic public sector project data. It includes no missing values or duplicates, and all categorical values are standardised for effective grouping and filtering in business intelligence tools. The addition of four derived variables—*Schedule Variance*, *Budget Variance*, *Performance Index*, and *Change Effectiveness*—supports diagnostic and prescriptive analytical exercises.

Purpose and Application

This dataset serves multiple instructional and practical purposes:

1. **Pedagogical**: Supports the development of MBA student competencies in data interpretation, performance monitoring, and strategic communication of insights.
2. **Applied Analytics**: Enables real-world simulation of project governance scenarios within a quality-focused public institution.
3. **Decision-Making Practice**: Allows users to evaluate project success based on multiple performance dimensions, including schedule adherence, budget control, risk mitigation, and stakeholder sentiment.
4. **AI and Predictive Modelling**: Facilitates experimentation with algorithms that classify high-risk projects, predict variance, and identify patterns associated with performance improvement.

Dataset Structure

- **Total Records**: 1,000
- **Variable Count**: 22

© 2025 Dr. Yasser Aly Yakout Shehata

  - 7 qualitative variables (categorical)
  - 10 quantitative variables (numeric)
  - 4 derived metrics to deepen analysis
- **Granularity**: Each record represents a single IT project instance or project-phase snapshot

Qualitative Variables

1. **Project Phase** – Initiation, Planning, Execution, Monitoring, Closure
2. **Stakeholder Feedback** – Positive, Neutral, Negative
3. **Compliance Status** – Compliant, Non-compliant
4. **Team Type** – In-house, External Vendor
5. **Risk Level** – Low, Medium, High
6. **Project Region** – National, Local, International
7. **Tool Usage** – JIRA, MS Project, Power BI, Other

Quantitative Variables

1. **Planned Duration** – Estimated project duration (in days)
2. **Actual Duration** – Actual time taken to complete project work
3. **Planned Budget** – Budget allocated for project completion
4. **Actual Spend** – Realised expenditure
5. **Num Milestones** – Total defined milestones in the project
6. **Milestones Completed** – Number of successfully completed milestones
7. **Risks Logged** – Total risks identified
8. **Risks Resolved** – Risks successfully mitigated or closed
9. **Change Requests Raised** – Number of project changes proposed
10. **Change Requests Approved** – Changes formally accepted by the governance team

Derived Variables

1. **Schedule Variance** – Calculated as percentage deviation from planned timeline

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Budget Variance** – Percentage over- or under-spend relative to budget
3. **Performance Index** – Combined efficiency measure of milestone completion and risk resolution
4. **Change Effectiveness** – Proportion of approved change requests relative to total raised

Schedule Variance (SV)

Definition: Schedule Variance is typically calculated as:

$$Schedule\ Variance\ (\%) = \frac{\text{Actual Duration} - \text{Planned Duration}}{\text{Planned Duration}} \times 100$$

- A **positive** value indicates the project took **longer than planned** (i.e., a delay or schedule overrun).
- A **zero** value indicates the project was completed **on time**.
- A **negative** value means the project was completed **ahead of schedule**.

Budget Variance (BV)

Definition: Budget Variance is calculated as:

$$Budget\ Variance\ (\%) = \frac{\text{Actual Spend} - \text{Planned Budget}}{\text{Planned Budget}} \times 100$$

- A **positive** value indicates **over-budget** (more was spent than planned).
- A **zero** value indicates the budget was **met exactly**.
- A **negative** value indicates the project came **under budget**.

These indicators allow students to benchmark project performance and assess the effectiveness of execution strategies under varying levels of complexity and risk.

Relevance to Power BI and AI for Business

This dataset has been structured for seamless integration with Power BI, supporting the development of dynamic dashboards, decision-making visuals, and project health indicators. Its variety of variable types makes it suitable for trend analysis, key performance index visualisation, heatmaps, and segmented filtering. Furthermore, its structure aligns with AI applications in business, enabling training of models that can predict risk outcomes, performance patterns, and budgetary variance across different project conditions.

© 2025 Dr. Yasser Aly Yakout Shehata

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify trends in project duration and cost variance
- Analyse the impact of risk level or compliance status on project outcomes
- Explore stakeholder sentiment by tool usage or team type

B. Power BI Visualisation

- Construct dashboards that track performance metrics across regions or project phases
- Visualise schedule and budget variance through line and bar charts
- Use filters to examine change request trends by risk level or compliance status

C. AI-Driven Business Insights

- Build classification models to predict project overrun risk
- Train regression models to forecast budget deviation
- Apply clustering to identify project typologies for future planning recommendations

Conclusion

The *IT Project Manager QCC Analytics* dataset is a comprehensive and adaptable educational resource designed to support the learning outcomes of MBA students. It reflects a nuanced understanding of IT project dynamics in a public sector quality assurance context, led by a Project Manager committed to data-informed performance improvement. Through its carefully designed structure, realistic variable relationships, and inclusion of derived performance indicators, the dataset serves as an ideal foundation for practising the full spectrum of business analytics and AI applications. It invites students to transition from technical exploration to strategic decision-making, equipping them with the insight-driven mindset essential for leadership roles in the digital age.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 56: Web Team Leader Analytics

Introduction

In an increasingly data-driven business environment, the ability to analyse complex operational data and derive actionable insights has become a core competency for future business leaders. This dataset has been purposefully designed to serve as a practical and pedagogical tool for MBA students specialising in Business Analytics, particularly those enrolled in the ***Business Analytics for Decision Making*** **course**. It offers a realistic, multidimensional representation of operational and performance data as encountered by a **Web Team Leader**, a pivotal role at the intersection of digital innovation, software engineering, and enterprise resource planning (ERP) systems.

The role of the Web Team Leader has evolved significantly in recent years. No longer limited to overseeing website design or backend systems, this role now encompasses the orchestration of integrated digital platforms, cross-functional team coordination, agile project delivery, and optimisation of ERP modules across various business functions. As digital transformation becomes a strategic imperative, Web Team Leaders are tasked with ensuring performance efficiency, system stability, and user-centric delivery across increasingly complex and distributed technical ecosystems.

This dataset simulates the operational environment of such a Web Team Leader who is also undertaking a final MBA project focused on enhancing decision-making through business analytics. It includes detailed records of project configurations, team dynamics, ERP integration, and performance outcomes across 1,000 unique project instances. These records reflect both internal variables such as team engagement, development methodology, and defect rates and external conditions such as ERP module complexity and support ticket volumes. Additionally, the inclusion of calculated fields like the *Efficiency Index*, *Stability Score*, and *Performance Category* adds analytical depth, enabling students to go beyond surface-level interpretation and engage in strategic analysis.

The dataset is particularly suitable for developing skills in **Exploratory Data Analysis (EDA)**, **dashboard visualisation using Power BI**, and the application of **Artificial Intelligence (AI) techniques** in business contexts. Its structure enables MBA students to apply statistical methods, build predictive models, and conduct data-driven evaluations of performance efficiency, system resilience, and leadership effectiveness. Furthermore, it supports scenario analysis, trend forecasting, and segmentation, all key capabilities in managerial analytics.

By engaging with this dataset, students will gain hands-on experience in transforming raw data into business insights. They will develop the ability to construct meaningful narratives

© 2025 Dr. Yasser Aly Yakout Shehata

around performance, diagnose operational inefficiencies, and recommend strategic interventions based on quantitative evidence. Ultimately, this dataset not only supports technical skill development but also cultivates the critical thinking and analytical reasoning required for effective leadership in technology-intensive business environments.

Dataset Description

The dataset contains **1,000 synthetic records**, each representing a Web Team Leader's quarterly project performance. The data combines contextual, operational, technical, and strategic dimensions to simulate the complexities encountered in enterprise-level digital and ERP projects. This includes variables that describe project scope, development methodologies, team composition, ERP modules, and various technical performance indicators such as uptime, ticket response times, and code churn.

The dataset also includes derived fields calculated through logical and statistical relationships to deepen the scope of analysis. These derived metrics allow students to explore operational efficiency, system stability, and leadership outcomes across different project environments.

Purpose and Application

The dataset is intended to:

- Provide MBA students with a practical tool for conducting structured and unstructured data analysis.
- Enable critical learning of relationships between business performance metrics and operational practices.
- Facilitate dashboard creation, visual storytelling, and performance reporting in Power BI.
- Support AI techniques such as clustering, regression, classification, and composite scoring.
- Simulate real-world decision-making scenarios using data analytics and business intelligence.

Dataset Structure

The dataset is a flat, single-table structure ideal for use in Power BI, Excel, and statistical software. It contains:

- **Seven qualitative (categorical) variables** describing project context and leadership environment.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Ten quantitative (continuous) variables** capturing operational, financial, and technical metrics.
- **Five derived variables** enhancing analytical depth by reflecting composite indicators and efficiency scores.

Qualitative Variables (7)

1. **Project_Type**: Nature of the project (Customer_Interface, Sales_Integration, ERP_Modules)
2. **Team_Size**: Size of the development team (Small, Medium, Large)
3. **Dev_Methodology**: Development methodology employed (Agile, Waterfall, Hybrid)
4. **ERP_Module**: ERP domain integrated into the project (CRM, Inventory, Accounting, HR, Multi)
5. **Leader_Experience**: Experience level of the team leader (Junior, Mid, Senior)
6. **Support_Ticket_Level**: Operational intensity based on ticket frequency (Low, Moderate, High)
7. **Team_Engagement**: Overall engagement level of the development team (Low, Medium, High)

Quantitative Variables (10)

1. **SPI (Schedule Performance Index)**: Ratio indicating schedule adherence
2. **CPI (Cost Performance Index)**: Ratio indicating cost efficiency
3. **Code_Churn_%**: Proportion of code changed per iteration
4. **Defect_Rate**: Number of defects per thousand lines of code
5. **Uptime_%**: System availability percentage
6. **P99_Latency_ms**: Time to serve the slowest 1 percent of user requests
7. **Ticket_Response_hr**: Time taken to respond to support tickets
8. **Ticket_Resolution_hr**: Time taken to resolve support tickets
9. **Sprint_Velocity**: Productivity measure in story points per sprint
10. **ROI_%**: Return on investment at the project level

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Variables (5)

1. **Latency_Category**: Categorisation of system latency into Low, Medium, or High
2. **Ticket_Efficiency_hr**: Aggregate measure of ticket response and resolution time
3. **Efficiency_Index**: ROI adjusted for support workload
4. **Stability_Score**: Composite score based on uptime, defect rate, and latency
5. **Performance_Category**: ROI classification into Loss, Moderate, Good, or Excellent

Relevance to Power BI and AI for Business

The dataset's structure and complexity are intentionally suited for use in Power BI and AI-enhanced analytics. The range of qualitative and quantitative data allows for:

- Data filtering, segmentation, and slicing using Power BI visuals
- Creation of calculated fields and Key Performance Indicators (KPIs)
- Data modelling using relationships among engagement, performance, and ERP complexity
- Integration with AI tools (e.g., Azure Machine Learning or Python) embedded in Power BI for advanced analytics

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Summary statistics, distribution analysis, and pattern detection across metrics
- Cross-tabulations by team size, experience, or ERP module
- Correlation analysis among SPI, CPI, and ROI
- Outlier detection and impact assessment on project outcomes

B. Power BI Visualisation

- Dashboard design for project KPIs, latency trends, and ROI outcomes
- Interactive filters to compare Agile vs Waterfall performance
- Drill-down charts to analyse performance by ERP module or engagement level
- Composite score visualisation using the Stability Score and Efficiency Index

© 2025 Dr. Yasser Aly Yakout Shehata

C. AI-Driven Business Insights

- Regression models to predict ROI based on engagement, methodology, and team size
- Classification models to predict Performance Category
- Clustering of projects to identify high-risk or high-efficiency profiles
- Feature importance ranking to support resource reallocation or risk mitigation

Conclusion

This dataset offers a powerful, structured, and educationally valuable foundation for exploring the intersection of analytics, leadership, and operational performance in technology-driven organisations. Centred on the role of a **Web Team Leader**, it encourages MBA students to apply theory to practice by simulating data-informed decision-making. Its design supports a wide spectrum of techniques from descriptive statistics and visualisation to AI-enhanced insights, making it ideal for coursework, capstone projects, and real-world simulations in Business Analytics and AI for Business.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 57: Technology Senior Specialist Vodafone

Introduction

The telecommunications sector operates at the intersection of high technology, data-intensive infrastructure, and customer-centric service delivery. Within this complex ecosystem, the role of a **Technology Senior Specialist** is both technically demanding and strategically significant. Such professionals are tasked with the ongoing assessment and optimisation of mobile network performance, ensuring reliable connectivity, superior service quality, and operational resilience. However, the importance of this role extends beyond engineering oversight, it directly influences business performance metrics such as customer satisfaction, revenue generation, and competitive differentiation.

In this context, the ability to extract, interpret, and act upon data becomes not only a technical function but a strategic imperative. This is particularly relevant for MBA students studying ***Business Analytics for Decision Making***, who are expected to apply analytical thinking to real-world business problems. The dataset presented here has been developed precisely to simulate such a scenario where data collected by a Technology Senior Specialist can be transformed into business intelligence to inform decisions at both tactical and strategic levels.

This synthetic dataset encompasses a wide spectrum of variables that reflect the operational environment of a mobile network: from performance metrics such as latency, packet loss, and throughput to business indicators like average revenue per user and data revenue growth. Additionally, it includes engineered variables such as revenue efficiency, network health score, and incident impact that deepen analytical capacity by combining multiple metrics into meaningful composite indicators.

The dataset is specifically designed for **educational and professional development** purposes, with structured features that enable both **exploratory data analysis (EDA)** and **predictive analytics using artificial intelligence (AI)** techniques. It encourages students to explore interdependencies between technical KPIs and business outcomes, simulate what-if scenarios, and draw actionable conclusions using tools like Power BI, Python, or R. By working with this dataset, MBA students will engage in hands-on learning that mirrors the responsibilities and analytical challenges faced by real-world telecom professionals.

Moreover, the dataset is enriched with qualitative variables such as customer segment, service type, and geographic classification that support segmentation, comparative analysis, and market-focused storytelling. These dimensions allow students to contextualise performance within different operational and commercial conditions, further enhancing their strategic thinking and decision-making skills.

© 2025 Dr. Yasser Aly Yakout Shehata

In summary, this dataset offers a powerful pedagogical tool for bridging theory and practice in business analytics. It enables students to step into the role of a data-driven decision maker, equipped with the skills and insights necessary to transform complex technical data into strategies that drive value, improve service delivery, and contribute to organisational success in the telecommunications industry.

Dataset Description

The dataset contains **1,000 observations** across **41 variables**, representing a multidimensional view of network operations and business performance. Each row simulates a real-time performance snapshot under various network configurations, customer segments, and geographic regions. The dataset combines:

- Qualitative descriptors of network and service environments
- Quantitative measures of technical and financial performance
- Engineered variables to enhance analytical depth
- Scaled and binned fields to support AI and machine learning applications

Purpose and Application

The dataset is intended for use in both academic coursework and professional development. It allows MBA students to:

- Apply data wrangling, descriptive statistics, and segmentation techniques
- Build insightful dashboards in Power BI
- Develop predictive and classification models using AI methodologies
- Interpret the relationship between operational conditions and financial impact

Dataset Structure

The dataset includes the following types of variables:

- **7 Qualitative Variables**
- **10 Quantitative Variables**
- **5 Derived Analytical Variables**
- **5 Binned and Segmented Variables**
- **14 Scaled Variables (standardised)**

© 2025 Dr. Yasser Aly Yakout Shehata

This structure provides layers of complexity suitable for descriptive, diagnostic, and predictive analytics.

Qualitative Variables (7)

These categorical variables define the environmental and operational context of each observation, supporting segmentation and comparison across different network scenarios.

1. **Network_Type**: Specifies the generation of mobile network technology used (e.g. 4G or 5G), which influences performance expectations.
2. **Region_Type**: Indicates the geographic area where the observation was recorded, classified as Urban, Suburban, or Rural.
3. **Equipment_Vendor**: Identifies the primary infrastructure provider responsible for the network hardware (e.g. Nokia, Ericsson, Huawei, Samsung).
4. **Test_Type**: Describes the methodology used to capture network data, such as Drive Test, Crowd-sourced monitoring, or Passive Monitoring.
5. **Incident_Level**: Categorical severity level of a network disruption or fault, rated as Low, Medium, or High.
6. **Service_Type**: Refers to the type of service being delivered or measured, including Voice, Data, or Messaging.
7. **Customer_Segment**: Classifies the end user of the service into Consumer, Small and Medium Enterprise (SME), or Enterprise categories.

Quantitative Variables (10)

These continuous variables capture technical KPIs and financial indicators relevant to network performance and customer experience.

1. **Latency_ms**: The average time, in milliseconds, that data takes to travel from source to destination and back.
2. **Jitter_ms**: Measures the variability in latency, indicating the consistency of data transmission.
3. **Packet_Loss_%**: Percentage of data packets lost during transmission, a critical measure of network reliability.
4. **Throughput_Mbps**: Actual data transfer rate achieved, measured in megabits per second.

© 2025 Dr. Yasser Aly Yakout Shehata

5. **Availability_%**: The percentage of time the network was operational and accessible to users.
6. **MOS_Score**: Mean Opinion Score, a user-based rating from 1 to 5 evaluating voice quality.
7. **Call_Setup_Time_s**: Time taken to establish a mobile call, measured at the 90th percentile to account for worst-case performance.
8. **Cell_Utilisation_%**: Indicates how much of a network cell's capacity is being used, expressed as a percentage.
9. **ARPU_USD**: Average Revenue Per User, a key financial metric reflecting user value.
10. **Data_Revenue_Growth_%**: The rate at which mobile data revenue is increasing or decreasing.

Derived Analytical Variables (5)

These variables are constructed from combinations of other fields to provide deeper business insight.

1. **QoE_Category**: Qualitative rating (High, Moderate, Low) of user experience, based on latency, jitter, and packet loss.
2. **Revenue_Efficiency**: A performance index reflecting how efficiently revenue is generated in relation to availability and throughput.
3. **Network_Health_Score**: Composite indicator that aggregates latency, jitter, packet loss, and availability into a single stability score.
4. **Incident_Impact**: Quantitative measure of disruption, calculated from incident severity and the utilisation level of affected cells.
5. **Performance_Category**: Business performance rating derived from thresholds applied to ARPU and data revenue growth.

Binned and Segmented Variables (5)

These variables group continuous metrics into discrete categories for easier interpretation and filtering in dashboard tools and machine learning applications.

1. **Latency_Level**: Categorises Latency_ms as Low, Medium, or High to indicate responsiveness.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Throughput_Level**: Segments Throughput_Mbps into performance bands for comparative analysis.
3. **ARPU_Segment**: Groups ARPU_USD into Low, Moderate, High, or Premium segments to assess financial performance.
4. **Revenue_Efficiency_Quartile**: Divides the Revenue_Efficiency metric into four quartiles for benchmarking.
5. **Health_Band**: Assigns the Network_Health_Score into four performance levels: Poor, Fair, Good, and Excellent.

Scaled Variables (14)

These variables are standardised using z-score transformation to support statistical modelling and AI applications.

1. **Latency_ms_scaled**: Standardised latency for numerical modelling.
2. **Jitter_ms_scaled**: Standardised measure of jitter for consistency in scaling.
3. **Packet_Loss_%_scaled**: Scaled version of packet loss for improved regression and classification accuracy.
4. **Throughput_Mbps_scaled**: Normalised throughput score to align units and variance.
5. **Availability_%_scaled**: Scaled system uptime variable.
6. **MOS_Score_scaled**: Standardised user experience metric.
7. **Call_Setup_Time_s_scaled**: Standardised setup time for statistical comparison.
8. **Cell_Utilisation_%_scaled**: Scaled version of network load.
9. **ARPU_USD_scaled**: Normalised ARPU to ensure equal contribution in model inputs.
10. **Data_Revenue_Growth_%_scaled**: Standardised growth rate for use in clustering and forecasting.
11. **Revenue_Efficiency_scaled**: Scaled composite efficiency score.
12. **Network_Health_Score_scaled**: Scaled network health indicator for advanced modelling.
13. **Incident_Impact_scaled**: Standardised operational impact measure.

© 2025 Dr. Yasser Aly Yakout Shehata

14. **Performance_Category_scaled**: Numerical encoding of the performance category for classification tasks.

This comprehensive set of variables provides a structured and interrelated basis for performing robust data exploration, building intuitive visualisations, and applying advanced AI methodologies, all within the context of business performance in telecommunications.

Relevance to Power BI and AI for Business

This dataset is explicitly structured for high compatibility with:

- **Power BI**: All categorical variables serve as slicers; binned fields enable filters and layered KPI tracking
- **AI Modelling**: Scaled fields enable clustering, regression, and classification tasks without preprocessing
- **Storytelling and Dashboards**: Rich context allows for dynamic dashboards comparing vendors, regions, or customer segments

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify hidden trends in performance metrics
- Evaluate distribution, central tendency, and variability
- Perform segmentation by region or service type
- Assess correlation between network reliability and ARPU

B. Power BI Visualisation

- Dashboard showing latency, revenue efficiency, and ARPU
- Heat maps by region or vendor
- Trendlines of performance over service type or customer group
- Performance band visualisations using binned variables

C. AI-Driven Business Insights

- Predict revenue growth from network performance indicators
- Classify network snapshots into performance categories
- Cluster environments with similar network and revenue profiles

© 2025 Dr. Yasser Aly Yakout Shehata

- Optimise network operations based on derived efficiency indicators

Conclusion

The **Technology Senior Specialist Analytics Dataset for Vodafone** offers a well-structured, multidimensional data environment tailored to the learning and analytical needs of MBA students. With its comprehensive scope—spanning operational KPIs, financial performance, derived analytical constructs, and AI-ready enhancements—the dataset bridges the gap between engineering operations and business intelligence. It provides a platform for students to develop high-impact insights, actionable strategies, and data-informed decision-making capabilities essential for leadership in digital, data-intensive industries.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 58: Regional Talent Acquisition Edulisa 360

Introduction

The global shift towards digitally enabled education has introduced new challenges and opportunities in the management of human capital. Within this dynamic context, the function of **talent acquisition** particularly at a regional level, has assumed strategic significance, influencing not only operational success but also long-term institutional growth. **Educational organisations such as Edulisa 360 and Monglish International Academy**, which operate across multiple geographies and business units, increasingly rely on data-informed talent acquisition practices to maintain competitiveness, foster innovation, and respond to regional staffing dynamics. In such environments, the role of the **Regional Talent Acquisition Head** becomes central to balancing local needs with institutional priorities.

The development of analytical acumen in addressing talent-related challenges is therefore critical, especially for emerging professionals in management roles. This dataset has been specifically designed to simulate the working environment and decision-making responsibilities of a Regional Talent Acquisition Head. It provides MBA students particularly those enrolled in the ***Business Analytics for Decision Making*** **course** with a robust, multidimensional dataset to explore data patterns, develop analytical narratives, and apply artificial intelligence to HR strategy.

The dataset captures the intricacies of recruitment workflows, encompassing a spectrum of variables such as time to fill, cost per hire, acceptance rates, candidate satisfaction, and turnover metrics. These variables are not merely administrative; they are strategic indicators that influence workforce planning, employer branding, and operational performance. In this context, business analytics serves as a lens through which students can interrogate organisational challenges, test hypotheses, and make data-driven recommendations that align with institutional goals.

Moreover, the dataset incorporates derived and engineered variables, such as hiring efficiency and turnover impact cost, to simulate the type of analytical reasoning required in strategic HR roles. These metrics provide depth and context beyond raw performance indicators, enabling students to engage in advanced modelling techniques, segmentation, and what-if analysis. The inclusion of scaled variables and performance categories also supports the application of machine learning and clustering methods, further preparing students for data-centric leadership roles in talent management.

By bridging human resources, analytics, and decision science, this dataset offers an authentic and pedagogically valuable foundation for business analytics education. It not

© 2025 Dr. Yasser Aly Yakout Shehata

only reflects the analytical complexity faced by talent acquisition leaders but also encourages students to apply descriptive, diagnostic, and predictive approaches to real-world HR scenarios. Through this dataset, future business leaders can develop the competencies needed to navigate data-rich environments and contribute meaningfully to organisational success in the digital era.

Dataset Description

The ***Regional Talent Acquisition Dataset*** comprises one thousand records representing aggregated recruitment outcomes across regions and business units. Each record captures both operational and contextual elements of a single hiring cycle. The dataset includes forty-one variables, classified into qualitative descriptors, quantitative measures, derived analytical metrics, segmented bins for visualisation, and standardised fields for AI readiness.

The dataset captures variation across job roles, business lines, sourcing channels, and regional performance profiles, providing a multidimensional structure ideal for descriptive and predictive analytics. It is fully synthetic, quality-assured, and logically coherent for use in educational and professional development contexts.

Dataset Structure

- **Format**: CSV (flat table)
- **Observations**: 1,000 rows (each represents a recruitment cycle)
- **Variables**: 41 columns
    - 7 qualitative
    - 10 quantitative
    - 5 derived analytical variables
    - 4 segmented variables
    - 14 scaled variables

All columns are free of missing values and ready for immediate use in BI tools and statistical analysis platforms.

Variable Descriptions

1. Qualitative Variables (7)

© 2025 Dr. Yasser Aly Yakout Shehata

These variables define the contextual environment within which recruitment processes are executed. They are useful for segmentation, filtering, and comparative analysis in dashboards and models.

1. **Region** – Denotes the geographic area where recruitment took place, such as Ulaanbaatar, Mandalgov, or Erdenet.
2. **Business_Unit** – Indicates the operational domain within the organisation, either eLearning or Academy, providing insight into recruitment by division.
3. **Recruitment_Channel** – Represents the method through which candidates were sourced, including internal referral, job boards, campus hiring, and recruitment agencies.
4. **Job_Level** – Specifies the seniority of the recruited position, categorised as Entry, Mid, or Senior.
5. **Position_Type** – Classifies the nature of the job role being filled, such as Technical, Administrative, or Teaching.
6. **Candidate_Source_Region** – Identifies the general geographic background of applicants (Urban, Suburban, or Rural), relevant for diversity and outreach analysis.
7. **Hiring_Team_Size** – Describes the size of the recruitment team handling the hiring cycle (Small, Medium, or Large), offering perspective on resource allocation.

2. Quantitative Variables (10)

These numeric indicators quantify key performance aspects of recruitment and are foundational for calculating efficiency, cost, and quality metrics.

1. **Time_to_Fill_days** – Number of days between job requisition approval and offer acceptance, indicating recruitment process length.
2. **Time_to_Hire_days** – Time elapsed from a candidate's application to their job offer acceptance, measuring operational responsiveness.
3. **Cost_per_Hire_USD** – Total cost incurred for filling a position, including advertising, agency fees, and recruiter labour, expressed in US dollars.
4. **Offer_Acceptance_Rate_%** – Percentage of candidates who accepted job offers, reflecting candidate engagement and employer attractiveness.
5. **Applicants_per_Req** – Number of applicants who responded to each job requisition, representing recruitment reach and market competitiveness.

© 2025 Dr. Yasser Aly Yakout Shehata

6. **Interview_to_Offer_Ratio** – Average number of interviews conducted per job offer made, assessing selectivity and efficiency in the screening process.
7. **New_Hire_Turnover_%** – Proportion of newly hired employees who left the organisation within a specified period, usually twelve months.
8. **Candidate_Satisfaction_Score** – Average rating (on a 1 to 10 scale) derived from candidate feedback post-application or interview, measuring candidate experience.
9. **Recruiter_Productivity** – Number of successful hires completed by an individual recruiter or team during the measurement period.
10. **Quality_of_Hire_Score** – Composite score evaluating post-hire success, combining performance assessments, retention, and hiring manager satisfaction.

3. Derived Analytical Variables (5)

These are computed fields created to deepen analytical capabilities by integrating multiple primary variables into strategic metrics.

1. **Hiring_Efficiency** – A calculated ratio indicating the productivity of recruitment efforts relative to the time taken to fill roles.
2. **Hiring_Cost_Ratio** – A cost-effectiveness measure showing the cost incurred per unit of hire quality.
3. **Channel_Effectiveness_Score** – A combined score generated by multiplying offer acceptance rate and candidate satisfaction, assessing the strategic value of each recruitment channel.
4. **Turnover_Impact_Cost** – An estimation of financial loss due to early employee turnover, derived by multiplying cost per hire by turnover percentage.
5. **Performance_Category** – A labelled variable segmenting hiring performance into categories (such as High, Moderate, Low) based on productivity and quality benchmarks.

4. Segmented Variables (4)

These variables are derived by categorising continuous values into meaningful bands or groups, enabling clearer visualisation, filtering, and comparative analysis in tools like Power BI.

1. **Time_to_Fill_Level** – Groups time to fill into categories: *Fast*, *Average*, or *Slow*, based on defined thresholds.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Cost_per_Hire_Level** – Categorises cost per hire into *Low*, *Moderate*, or *High* bands for budgeting and benchmarking.
3. **Quality_of_Hire_Level** – Segments quality of hire scores into *Low*, *Moderate*, or *High* performance tiers.
4. **Channel_Efficiency_Quartile** – Divides the channel effectiveness score into quartiles (Q1 to Q4), supporting comparative ranking and performance benchmarking.

5. Scaled Variables (14)

These variables are standardised using z-score scaling to normalise their values, making them suitable for machine learning models, clustering, and regression analysis.

1. **Time_to_Fill_days_scaled** – Standardised version of time to fill.
2. **Time_to_Hire_days_scaled** – Standardised time from application to hiring.
3. **Cost_per_Hire_USD_scaled** – Scaled cost to enable comparison across different ranges.
4. **Offer_Acceptance_Rate_%_scaled** – Standardised acceptance rate.
5. **Applicants_per_Req_scaled** – Scaled measure of applicant volume.
6. **Interview_to_Offer_Ratio_scaled** – Normalised value of interview effort per hire.
7. **New_Hire_Turnover_%_scaled** – Scaled turnover percentage for risk profiling.
8. **Candidate_Satisfaction_Score_scaled** – Standardised satisfaction metric for comparison.
9. **Recruiter_Productivity_scaled** – Scaled recruiter output.
10. **Quality_of_Hire_Score_scaled** – Normalised hire quality for advanced modelling.
11. **Hiring_Efficiency_scaled** – Scaled efficiency for correlation and AI use.
12. **Hiring_Cost_Ratio_scaled** – Standardised cost-effectiveness ratio.
13. **Channel_Effectiveness_Score_scaled** – Scaled channel metric used in benchmarking.
14. **Turnover_Impact_Cost_scaled** – Scaled financial impact of early turnover.

© 2025 Dr. Yasser Aly Yakout Shehata

Purpose and Application

This dataset was specifically designed to support the following purposes:

- Enable MBA students to conduct **exploratory data analysis (EDA)** with realistic HR data.
- Support **dashboard creation and storytelling in Power BI**, simulating executive-level reporting.
- Facilitate application of **machine learning techniques** such as regression, clustering, and classification in AI-enabled HR analytics.
- Allow for simulation of **KPI optimisation and what-if scenarios** in strategic recruitment decision-making.
- Provide a real-world case study for the role of a Regional Talent Acquisition Head working in a distributed educational ecosystem.

Relevance to Power BI and AI for Business

This dataset is tailored for analytical workflows in Power BI and AI-enabled platforms:

- Categorised and binned variables allow meaningful filters and slicers
- Quantitative KPIs enable DAX-based KPI scorecards and time comparisons
- Derived metrics simulate executive dashboards for recruitment strategy
- Scaled numeric fields are suitable for ML models such as clustering, regression, and classification

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify recruitment bottlenecks and cost inefficiencies
- Analyse patterns across regions and business units
- Assess correlations between hiring practices and retention outcomes

B. Power BI Visualisation

- Interactive dashboards showing KPIs across time, regions, and roles
- Heat maps for recruiter performance
- Trend charts for cost, satisfaction, and quality of hire

© 2025 Dr. Yasser Aly Yakout Shehata

C. AI-Driven Business Insights

- Predict high-risk hiring outcomes using turnover or cost models
- Cluster hiring teams or regions by performance profiles
- Classify quality-of-hire outcomes based on upstream variables

Conclusion

The **Regional Talent Acquisition Dataset** offers an academically rigorous yet practically grounded resource for exploring data-driven HR decision-making. Specifically aligned with the job responsibilities of a Regional Talent Acquisition Head, this dataset supports a comprehensive range of business analytics techniques suitable for MBA-level learning and professional development. Its structure, diversity of variables, and derived indicators empower students to build insights, optimise KPIs, and propose evidence-based recommendations using data visualisation, AI models, and business logic.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 59: Modern Endpoint Management

In the current digital era, the proliferation of remote workforces, hybrid cloud environments, and diverse endpoint devices has elevated the strategic importance of endpoint management within enterprise IT ecosystems. Organisations are no longer managing a fixed set of on-premise desktops but rather a dynamic and heterogeneous infrastructure composed of laptops, mobile devices, Internet of Things (IoT) nodes, and virtual endpoints spread across geographies. These endpoints serve as critical interfaces through which employees, applications, and services interact with organisational data and networks. Consequently, managing endpoint performance, security posture, compliance, and operational availability has become a foundational component of IT governance and business continuity.

The ***Modern Endpoint Management Analytics*** *Dataset* was conceived within this context to simulate the data environment in which enterprise consultants and IT infrastructure professionals operate. Specifically, this dataset is tailored for MBA students undertaking their final project in the ***Business Analytics for Decision Making*** **course**, where the role of a *Modern Endpoint Management Consultant* is used as the applied case. This consultant is responsible for advising large organisations on how to optimise their endpoint strategy using evidence-based insights, analytical performance monitoring, and intelligent automation techniques.

The dataset brings together operational KPIs, compliance indicators, system-level metrics, and financial performance measures typically available through endpoint detection and response (EDR) systems, mobile device management platforms (MDM), and enterprise IT service management (ITSM) dashboards. By embedding contextual variables such as device type, network access location, patching cadence, and user profile, the dataset facilitates a multidimensional analysis of how organisational policies and technical configurations influence endpoint health, user experience, and security outcomes.

Moreover, this dataset is not merely a collection of metrics but a pedagogical artefact designed to support experiential learning. Students are encouraged to engage with the data using a suite of analytical tools from basic descriptive statistics to advanced visualisation in Power BI and machine learning techniques in Python or R. Through this process, they will gain practical exposure to problem framing, data cleaning, variable transformation, trend identification, and decision support modelling. The inclusion of derived analytical variables and scaled features ensures that the dataset supports not only exploratory data analysis but also AI-driven business insight generation.

© 2025 Dr. Yasser Aly Yakout Shehata

Ultimately, this dataset empowers students to experience the intersection between IT operations and strategic analytics, simulating the kind of data interpretation and decision-making expected of professionals in contemporary enterprise environments. It helps bridge the gap between theoretical knowledge and applied business impact, preparing learners to contribute meaningfully in roles that demand both technical fluency and managerial insight.

Dataset Description

This dataset consists of **1,000 records** and **41 variables**, combining qualitative attributes, quantitative measurements, derived analytical variables, and standardised scaled versions. Each record represents an endpoint device contextually embedded in enterprise infrastructure, reflecting its operational performance and the organisational policies governing it.

The data structure reflects the complex nature of endpoint management, integrating metrics relevant to system security, patching, user behaviour, and IT operational efficiency. The dataset also includes categorised and scaled versions of continuous variables to support segment-based visualisation and machine learning readiness.

Purpose and Application

The primary objective of this dataset is to provide a practice environment for MBA students to:

- Perform **Exploratory Data Analysis (EDA)** using statistical and visual methods
- Create interactive **Power BI dashboards** to identify patterns and monitor performance
- Apply **machine learning models** such as clustering and classification for business-driven insights
- Develop **recommendations** for improving endpoint security, compliance, and cost efficiency

It is also suitable for assignments requiring descriptive statistics, correlation analysis, hypothesis testing, and AI-powered decision-support simulations.

Dataset Structure

The dataset includes the following categories of variables:

- **7 Qualitative Variables**
- **10 Quantitative Variables**

© 2025 Dr. Yasser Aly Yakout Shehata

- **5 Derived Analytical Variables**
- **4 Segmented Variables**
- **14 Scaled Variables**

Qualitative Variables (7)

These provide categorical context for endpoint conditions and policy configurations:

1. **Endpoint_Type** – The category of the device (Laptop, Desktop, Mobile, IoT)
2. **Operating_System** – The OS installed on the device (e.g., Windows, macOS, Linux)
3. **Security_Tool** – The security solution in place (e.g., EDR, Antivirus, SIEM)
4. **Network_Location** – Network type used by the endpoint (Corporate, VPN, Public)
5. **Patch_Schedule** – Frequency of patching policy (Weekly, Monthly, Quarterly)
6. **User_Role** – Type of user account (Admin, Standard User)
7. **Compliance_Status** – Current compliance condition (Compliant, Non-Compliant, In Review)

Quantitative Variables (10)

These metrics measure endpoint system performance, risk exposure, and IT efficiency:

1. **Threat_Detection_Rate_%** – Rate of threat detection by the endpoint's security tool
2. **Mean_Time_to_Detect_Hrs** – Average hours taken to detect a threat
3. **Mean_Time_to_Respond_Hrs** – Average hours to contain or mitigate a threat
4. **Patch_Compliance_%** – Percentage of devices patched on time
5. **Time_to_Patch_Hrs** – Hours taken from patch release to installation
6. **Failed_Patch_%** – Proportion of patches that failed to install
7. **Endpoint_Availability_%** – Proportion of time the device is operational and connected
8. **CPU_Overhead_%** – CPU load introduced by endpoint protection tools
9. **Incident_Count** – Number of security incidents detected
10. **Security_ROI** – Financial return on investment in endpoint protection measures

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Analytical Variables (5)

These calculated fields combine multiple indicators for strategic insight:

1. **Security_Responsiveness_Score** – Average of detection and response times
2. **Patch_Effectiveness** – Adjusted compliance based on patch failure rate
3. **Threat_Blocking_Efficiency** – Detected threats divided by incidents plus one
4. **Performance_Impact_Score** – CPU overhead divided by incident volume
5. **Security_Performance_Category** – ROI-based label (Low, Moderate, High)

Segmented Variables (4)

Used for grouping and filtering in Power BI and other dashboards:

1. **Detection_Efficiency_Level** – Categorisation of detection rate (Low, Medium, High)
2. **Compliance_Level** – Patch compliance grouped (Poor, Acceptable, Good)
3. **MTTx_Category** – Detection and response time banding (Fast, Average, Slow)
4. **Resource_Overhead_Level** – Classification of CPU usage (Low, Medium, High)

Scaled Variables (14)

Z-score normalised fields to support clustering, regression, and classification:

1. **Threat_Detection_Rate_%_scaled**
2. **Mean_Time_to_Detect_Hrs_scaled**
3. **Mean_Time_to_Respond_Hrs_scaled**
4. **Patch_Compliance_%_scaled**
5. **Time_to_Patch_Hrs_scaled**
6. **Failed_Patch_%_scaled**
7. **Endpoint_Availability_%_scaled**
8. **CPU_Overhead_%_scaled**
9. **Incident_Count_scaled**
10. **Security_ROI_scaled**
11. **Security_Responsiveness_Score_scaled**

© 2025 Dr. Yasser Aly Yakout Shehata

12. **Patch_Effectiveness_scaled**
13. **Threat_Blocking_Efficiency_scaled**
14. **Performance_Impact_Score_scaled**

Relevance to Power BI and AI for Business

This dataset has been crafted with visualisation and predictive modelling in mind. Its structure supports:

- Time series and KPI dashboards using Power BI cards, tables, and slicers
- Drill-down filtering using categorical and segmented variables
- Clustering of endpoint profiles for resource optimisation
- Forecasting of security or patching metrics using regression

The standardised variables ensure compatibility with Python or Power BI-integrated ML platforms such as Azure ML.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify endpoints with high failure or CPU overhead
- Assess variation in response times by endpoint type or user role
- Explore associations between compliance and ROI

B. Power BI Visualisation

- Develop compliance dashboards by region or schedule
- Map incident patterns by security tool or OS
- Show temporal trends in detection rate and patching

C. AI-Driven Business Insights

- Use clustering to segment high-risk endpoint profiles
- Apply logistic regression to predict compliance risk
- Use decision trees to estimate ROI drivers from performance metrics

© 2025 Dr. Yasser Aly Yakout Shehata

Conclusion

The ***Modern Endpoint Management Analytics Dataset*** serves as a comprehensive resource for students and professionals seeking to bridge technical IT operations and strategic decision-making. It enables a range of analytical and visual techniques while immersing learners in the operational realities of endpoint performance. With rich categorical, quantitative, and derived features, this dataset enhances the practical application of business analytics tools, data visualisation, and artificial intelligence methods in support of enterprise IT strategy and governance.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 60: Power BI Developer and Trainer

Introduction

In today's data-centric business environment, the ability to derive meaningful insights from complex datasets is an essential competency for business leaders and analysts alike. As organisations increasingly rely on data to inform strategic decision-making, the need for professionals who can interpret, visualise, and act upon data has grown exponentially. Within this context, the ***Power BI Developer and Trainer Dataset*** has been developed as a pedagogical tool to empower students pursuing a Master of Business Administration (MBA) degree, particularly those specialising in **Business Analytics and Artificial Intelligence for Business Organisations**.

This dataset is specifically tailored to support the capstone project of a seasoned **Power BI Developer and Trainer** working in the educational technology sector. The professional in focus holds both a Power BI Data Analyst certification and an MBA with a concentration in AI for Business, and is positioned at the intersection of education, technology, and analytics. Their final project seeks to transform operational and performance-related data into actionable intelligence that can improve learning outcomes, optimise system usage, and support institutional decision-making.

The dataset replicates the multifaceted and high-velocity environment typical of digital education platforms, where data is continuously generated from learning management systems, assessment tools, feedback mechanisms, and operational infrastructure. It simulates how a data analyst might integrate and model these disparate sources of information to develop dashboards, identify patterns, and anticipate stakeholder needs using **Power BI and AI-driven techniques**.

Furthermore, the dataset has been meticulously structured to reflect the types of data challenges and opportunities that analytics professionals face in real-world business contexts. It includes a wide range of variables from user engagement and platform interaction metrics to sentiment scores and system performance indicators. Derived variables such as a composite *Engagement Score* and a risk classification field (*Risk Flag*) offer additional analytical depth, enabling users to apply segmentation, predictive analytics, and decision support strategies.

Designed for use in *Exploratory Data Analysis (EDA)*, interactive dashboard development, and the application of AI methods, this dataset supports the development of practical skills in data storytelling, visualisation, and strategic insight generation. It offers MBA students a rich and realistic scenario in which they can explore the power of data analytics to drive performance, personalisation, and institutional effectiveness in educational settings, while

© 2025 Dr. Yasser Aly Yakout Shehata

simultaneously building the competencies needed to lead in AI-enabled business environments.

Dataset Description

The dataset comprises 1000 anonymised user records, each capturing a range of categorical and numerical variables that characterise user behaviour, learning engagement, platform utilisation, and system performance. These data points simulate real-world metrics collected by digital learning environments, and are structured to facilitate exploration, pattern identification, and predictive modelling. Derived variables are included to improve analytical depth and support strategic applications such as segmentation, performance tracking, and risk detection.

Purpose and Application

This dataset has been curated for MBA students to develop technical and analytical proficiency in:

- Conducting Exploratory Data Analysis (EDA)
- Building interactive visual dashboards using Power BI
- Applying AI and machine learning principles to business decision-making

It enables the practice of descriptive, diagnostic, and predictive analytics within a realistic organisational context, focused on educational data systems and performance evaluation.

Dataset Structure

The dataset includes **19 variables**, divided into:

- **7 Qualitative variables**: Used for classification, segmentation, and filtering
- **10 Quantitative variables**: Used for aggregation, correlation, and metric analysis
- **2 Derived variables**: Engineered to support composite scoring and strategic flagging

Qualitative Variables (Categorical)

1. **User_ID**: Unique identifier for each user record
2. **User_Role**: Role classification of the user (Student, Teacher, or Principal)
3. **Course_Subject**: Subject area being studied or taught (e.g. Math, Science, English)
4. **Region**: Geographical region of the user (e.g. MENA, Asia, Europe)

© 2025 Dr. Yasser Aly Yakout Shehata

5. **Sentiment_Category**: Categorised user sentiment (Positive, Neutral, Negative) based on feedback
6. **Platform_Used**: Platform type used to access the learning system (Web, Mobile, Tablet)
7. **Feedback_Type**: Feedback collection method (Survey, Open-ended, Rating, NPS)

Quantitative Variables (Numerical)

8. **Attendance_Rate**: Percentage of class sessions attended
9. **Lesson_Completion**: Percentage of lessons completed by the user
10. **Assessment_Score**: Average performance score on assessments (scale: 0–100)
11. **Time_on_Platform_min**: Total time spent on the platform (in minutes)
12. **Session_Count**: Number of platform sessions initiated by the user
13. **System_Response_Time_ms**: Average platform response time during usage (in milliseconds)
14. **Content_Clicks**: Number of content elements (videos, quizzes, links) clicked
15. **Sentiment_Score**: Numeric score representing sentiment polarity (scale: -1 to +1)
16. **RAM_Usage_MB**: Average RAM usage in megabytes during platform operation
17. **Storage_Usage_MB**: Data volume used by the user (in megabytes)

Derived Variables

18. **Engagement_Score**: Composite score calculated from attendance, lesson completion, assessments, time on platform, and session count. Measures overall user engagement (scale: 0–100)
19. **Risk_Flag**: Categorical indicator based on low engagement or negative sentiment. Flags users as *At Risk* or *Normal*
20. **Engagement_Level**: Tiered classification of engagement (Low, Moderate, High, Exceptional) based on Engagement_Score

Relevance to Power BI and AI for Business

This dataset is highly suitable for developing advanced Power BI dashboards and conducting AI-enhanced business analyses. Power BI can be used to visualise trends in engagement, identify at-risk users, and explore segmentation across user roles, regions, or course

© 2025 Dr. Yasser Aly Yakout Shehata

subjects. Through interactive filtering and drill-downs, students can create decision-support tools that simulate real business applications.

Moreover, the dataset lends itself to AI integration by supporting clustering, risk modelling, and predictive analytics. By training machine learning models on variables such as assessment scores, platform behaviour, and sentiment, students can generate automated insights and forecast academic or operational outcomes.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify relationships between lesson completion and assessment scores
- Explore how user engagement varies by role or platform type
- Detect behavioural patterns using sentiment polarity and system performance metrics

B. Power BI Visualisation

- Build dashboards with KPIs on engagement and performance
- Visualise geographical and subject-level variations
- Create drill-through reports on users flagged *At Risk*

C. AI-Driven Business Insights

- Predict student dropouts using sentiment and engagement data
- Cluster users into behavioural segments for targeted interventions
- Generate risk scores and recommend engagement strategies

Conclusion

The ***Power BI Developer and Trainer Dataset*** offers a robust and multifaceted platform for applying business analytics and AI concepts within a realistic scenario. Centred on the role of a Power BI Developer and Trainer, the dataset encourages students to engage in evidence-based decision-making and strategic insight generation. It provides the technical depth, variable diversity, and contextual relevance necessary to simulate professional analytic tasks, making it an ideal foundation for MBA-level exploration of data-driven business strategy.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 61: MTS Customer Service Analytics

Introduction

In an increasingly digital global economy, the role of customer service has evolved beyond traditional support functions to become a cornerstone of organisational performance, client satisfaction, and competitive advantage. Nowhere is this transformation more visible than in public sector digital infrastructure projects that facilitate international trade and logistics. In Egypt, **Misr Technology Services (MTS)** serves as the technological backbone of the *National Single Window for Foreign Trade Facilitation*, a platform known as **NAFEZA** integrating data exchange between importers, exporters, customs, and over 30 governmental entities. Within this environment, the **Customer Service / Client Account Specialist** is not only a point of contact but a data-informed decision agent responsible for ensuring system accessibility, responsiveness, and user satisfaction.

The ***MTS Customer Service Analytics Dataset*** was developed to reflect the realities faced by such specialists. It simulates 1000 customer service records capturing a broad range of operational interactions across Egypt's digital trade network. The dataset forms the empirical foundation for an MBA capstone project by a specialist seeking to demonstrate how business analytics, visualisation, and artificial intelligence (AI) can drive smarter decisions, reduce service friction, and proactively manage client expectations.

This dataset has been carefully engineered to address the educational needs of MBA students pursuing the course ***Business Analytics for Decision Making***. It offers a high-fidelity approximation of the data a client support team at MTS would encounter daily, ranging from initial response times and resolution durations to platform activity logs and post-service satisfaction ratings. In addition, it introduces derived variables such as a *Customer Experience Index (CEI)* and *Risk Flag* indicators that enrich the analytical framework and provide direct entry points into segmentation, prediction, and strategic decision-making.

Through this simulated environment, students are given the opportunity to explore how structured datasets can be used not only for performance monitoring and operational insights, but also to enable AI-enhanced strategies for service innovation. The dataset supports techniques such as exploratory data analysis (EDA), dashboard construction in Power BI, classification modelling, and customer profiling, all of which align with the strategic competencies expected of future data-literate business leaders.

By situating the dataset within the critical infrastructure of **Egypt's trade ecosystem** and framing it through the lens of an MBA candidate's professional challenge, this introduction sets the stage for both academic engagement and applied problem solving. It encourages

© 2025 Dr. Yasser Aly Yakout Shehata

students to think critically about the use of data in public service delivery, while offering them the tools to build intelligent, ethical, and actionable insights.

Dataset Description

The dataset contains 1000 synthetic records representing customer service cases processed by MTS through the **NAFEZA platform**. Each record simulates the interaction between a client and the support team, including details about the issue raised, the channel of communication used, platform performance indicators during the interaction, and post-service feedback.

The dataset includes:

- **7 qualitative variables** for categorisation and segmentation
- **10 quantitative variables** for metric analysis
- **3 derived variables** to enhance interpretability and decision-making potential

Purpose and Application

The dataset is specifically designed to serve the following pedagogical and analytical goals:

- **For students**: To apply EDA techniques, develop Power BI dashboards, and model decision scenarios using AI.
- **For professionals**: To understand how service metrics, client feedback, and system indicators can be integrated to enhance operational excellence.
- **For academic instructors**: To use a context-specific case study that bridges analytics theory with public-sector application.

The dataset enables real-world simulation of client support analytics and offers meaningful applications in service performance evaluation, feedback modelling, and customer segmentation.

Dataset Structure

The dataset comprises **20 variables**, grouped as follows:

Qualitative Variables

1. **Ticket ID** – Unique identifier for each support case.
2. **User Type** – Client category: Importer, Exporter, Broker, Bank, or Agent.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Issue Type** – Description of the issue: Registration, ACI Submission, Login, Technical, Invoice Conversion.
4. **Region** – The geographical area of the user: Cairo, Alexandria, Suez, Damietta, or Port Said.
5. **Channel** – Method of communication: Call, Email, In-Person, Portal.
6. **Priority** – Case urgency: Low, Medium, High, Critical.
7. **Sentiment Category** – Client's perceived experience: Positive, Neutral, or Negative.

Quantitative Variables

8. **First Response Time (min)** – Time taken to respond initially to the ticket.
9. **Resolution Time (min)** – Total duration needed to close the support case.
10. **Ticket Volume Daily** – Number of tickets submitted on the same day.
11. **Follow-ups Count** – Number of communications required to resolve the case.
12. **Satisfaction Score** – Client rating on a five-point scale.
13. **Platform Sessions** – Number of user logins on the date of the ticket.
14. **Average System Response (ms)** – Time taken by the platform to load or respond.
15. **ACI Submission Delay (hr)** – Delay in hours between support interaction and document issuance.
16. **Document Upload Size (MB)** – Size of the attachments submitted.
17. **Account Age (days)** – Time elapsed since the client created their account.

Derived Variables

18. **Customer Experience Index (CEI)** – Composite score integrating satisfaction, resolution time, and session activity; scale from 0 to 100.
19. **Risk Flag** – Identifies *At Risk* users based on low CEI or excessive resolution time.
20. **Engagement Level** – Tiered grouping of CEI: Low, Moderate, High, Exceptional.

Relevance to Power BI and AI for Business

This dataset is particularly well-suited to Power BI due to its combination of structured metrics, segmentation-ready categories, and time-based performance indicators. Key metrics like response time, satisfaction score, and CEI can be visualised through cards, heat

© 2025 Dr. Yasser Aly Yakout Shehata

maps, and line charts. Risk levels and engagement tiers allow for dashboard filtering and user profiling.

From an AI perspective, the dataset contains predictors suitable for classification and clustering, enabling predictive insights such as forecasting dissatisfaction or optimising case routing. It provides a practical context for developing AI applications in customer experience management and public service optimisation.

Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify resolution time patterns across user types and regions
- Evaluate correlations between follow-ups and satisfaction
- Analyse distribution of satisfaction scores by communication channel

B. Power BI Visualisation

- Develop KPI dashboards tracking average CEI, satisfaction, and ticket volumes
- Construct visual filters for priority levels, issue types, and sentiment
- Drill into high-risk user groups to understand failure drivers

C. AI-Driven Business Insights

- Predict customer risk using decision trees or logistic regression
- Cluster clients by engagement level to inform service segmentation
- Model the effect of resolution delays on CEI using regression or neural networks

Conclusion

The ***MTS Customer Service Analytics Dataset*** offers a rich and multifaceted foundation for developing practical expertise in business analytics, service quality evaluation, and AI-supported customer insights. Tailored to the responsibilities of a Customer Service Specialist at MTS, the dataset enables MBA students to connect data-driven problem solving with public sector realities. It equips learners to make informed, impactful decisions using visual, statistical, and intelligent analysis, core competencies for modern professionals in analytics, operations, and digital service delivery.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 62: Hannoun Pharmacies Analytics

1. Introduction

The healthcare retail industry, particularly the community pharmacy sector, is undergoing a profound transformation driven by digitalisation, customer-centric innovation, and operational efficiency. Within this evolving landscape, the role of data has become indispensable not only for monitoring sales or managing inventory, but also for enhancing patient care, tailoring marketing strategies, and supporting strategic growth decisions.

Hannoun Pharmacies, a well-established pharmacy chain in Egypt, embodies this transformation through its adoption of technology-integrated operations and data-informed business practices. The chain is led by Dr Sherihan El Araby, a licensed pharmacist and entrepreneur, who has built a strong reputation for combining pharmaceutical expertise with effective retail management. Her dual responsibilities as healthcare provider and business leader position her uniquely at the intersection of clinical service and strategic enterprise development.

As part of her MBA capstone project in ***Business Analytics for Decision Making***, this dataset simulates the operational data that a general supervisor like Dr El Araby would use to evaluate and enhance business performance. The dataset reflects real-world transactional complexity, involving diverse products, customer profiles, inventory dynamics, promotional interventions, and supplier relationships. It also incorporates derived variables that capture latent business phenomena such as inventory risk, customer lifetime value potential, and marketing return on investment.

This simulated environment enables **MBA students** to immerse themselves in practical business scenarios, engaging directly with the types of data, variables, and decisions that are increasingly central to modern pharmacy operations. By analysing this data through visual tools such as Power BI and applying artificial intelligence methods to extract predictive insights, students gain experience in translating raw data into managerial action.

This introduction therefore situates the dataset not merely as a technical resource, but as a representation of a real business setting in which data analytics is used to inform critical decisions in healthcare retail. It serves as both an academic exercise and a professional simulation, one that reflects the analytical mindset and operational challenges faced by today's business-savvy healthcare practitioners.

2. Dataset Description

The dataset contains one thousand records, each corresponding to a unique transaction within a community pharmacy setting. Variables span multiple business functions including

© 2025 Dr. Yasser Aly Yakout Shehata

inventory management, sales operations, customer service, and marketing. Alongside raw fields, the dataset includes engineered features that enhance insight generation and support AI-based applications.

3. Purpose and Application

This dataset is designed to:

- Facilitate the application of analytical methods in a retail healthcare context
- Support the development of interactive dashboards in Power BI
- Provide a realistic environment for machine learning tasks such as customer segmentation and promotion response modelling
- Strengthen business decision-making skills in areas such as inventory risk, pricing strategy, and loyalty programme evaluation

It mirrors the operational scope of a pharmacy owner who supervises transaction processes, supplier coordination, and customer relationship strategies, providing a comprehensive testbed for decision-making simulations.

4. Dataset Structure

- **Total Records**: 1,000
- **Variable Count**: 20
    - Qualitative: 7
    - Quantitative: 10
    - Derived: 3

5. Qualitative Variables

1. **Transaction_ID** – Sequential identifier for each transaction
2. **Product_Category** – Classification of item sold (Prescription, OTC, Supplement, Health Product)
3. **Customer_Type** – Type of customer (New, Returning, Loyalty Program Member)
4. **Promotional_Applied** – Indicates whether a promotion was used in the transaction (Yes or No)
5. **Supplier** – Name of the supplier providing the product

© 2025 Dr. Yasser Aly Yakout Shehata

6. **Day_of_Week** – Day the transaction occurred
7. **Payment_Method** – Method used for payment (Cash, Credit Card, Mobile Pay)

6. Quantitative Variables

8. **Quantity_Sold** – Number of units sold in the transaction
9. **Unit_Price** – Price per unit at the time of sale
10. **Transaction_Value** – Total value of the transaction (Quantity_Sold × Unit_Price)
11. **Inventory_Level_Before** – Stock level before the transaction
12. **Inventory_Level_After** – Stock level remaining after the transaction
13. **Restock_Delay_Days** – Number of days required to restock the product (if applicable)
14. **Loyalty_Points_Earned** – Points awarded to customers in loyalty programme
15. **Customer_Age_Category** – Age group classification of the customer
16. **Staff_Processing_Time_min** – Time (in minutes) taken by staff to complete the transaction
17. **Year_to_Date_Sales** – Accumulated pharmacy sales up to the date of the transaction

7. Derived Variables

18. **Stock_Risk_Flag** – Indicates whether post-transaction inventory falls below a critical threshold (Yes or No)
19. **Promotion_Effectiveness_Index** – Estimated uplift in sales value due to promotional offers
20. **Customer_Value_Score** – Customer classification based on transaction value (Low, Medium, High)

8. Relevance to Power BI and AI for Business

This dataset is highly compatible with Power BI's capabilities for real-time dashboards, KPI tracking, and drill-down analytics. Its structured format and engineered variables support:

- **Custom visuals** such as bar charts, KPIs, and trend lines
- **Segmentation and filters** by customer type, product category, and day of week
- **Dynamic metrics** like sales by region, inventory alerts, and customer value rankings

© 2025 Dr. Yasser Aly Yakout Shehata

For AI-driven business applications, the dataset enables:

- Predictive modelling (e.g. restock delays, promotion responsiveness)
- Clustering (e.g. customer segmentation based on value scores)
- Classification tasks (e.g. flagging transactions at risk of inventory depletion)

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Identify top-selling products by category
- Analyse time-based sales trends and customer flow
- Understand patterns in promotional effectiveness

B. Power BI Visualisation

- Create dashboards tracking inventory turnover and revenue
- Build interactive visuals to explore customer loyalty and purchasing behaviour
- Monitor real-time transaction values and staff efficiency metrics

C. AI-Driven Business Insights

- Predict stock-outs using inventory and sales history
- Forecast customer demand based on age and loyalty status
- Optimise promotional strategies using uplift modelling

10. Conclusion

The ***Hannoun Pharmacies Analytics Dataset*** offers a realistic, structured, and educationally relevant environment for business analytics training. Rooted in the professional practice of a pharmacist and business owner, the dataset serves as a practical tool for MBA students to bridge academic theory and managerial practice. Its design supports critical thinking, visual storytelling, and data-driven strategy formulation, making it a valuable asset in preparing future business leaders to operate in analytically intelligent and customer-focused enterprises.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 63: Technical Services Smart Systems Egypt

1. Introduction

In today's rapidly evolving landscape of educational technology and technical training, organisations delivering advanced laboratory infrastructure face increasing pressure to optimise their service operations. These operations range from the installation of robotics and mechatronic systems to commissioning programmable logic controllers and providing ongoing support and training. At the forefront of managing these complex, resource-intensive processes is the **Technical Manager**, a role that combines project management, technical oversight, client coordination, and strategic service planning.

This synthetic dataset has been meticulously designed to simulate the professional environment and operational responsibilities of a **Technical Manager** working within **Smart Systems Egypt**, a leading provider of scientific and laboratory equipment for academic institutions across Egypt and the wider Middle East. The dataset captures essential aspects of the service delivery lifecycle, encompassing project timelines, equipment deployment, technician labour, service costs, resolution times, and customer feedback mechanisms.

MBA students enrolled in the ***Business Analytics for Decision Making*** **course** will use this dataset to develop a deeper understanding of how analytics can support evidence-based decision-making in a technical service context. The dataset provides a structured platform for practising core analytical techniques, such as descriptive statistics, data cleaning, trend identification, and outlier detection. Moreover, it integrates derived variables that are critical for managerial insight, such as schedule adherence and operational efficiency, thus enabling the use of predictive models and AI-driven performance classification.

Through interaction with this dataset, students will not only develop technical fluency in analytical tools such as Power BI and Python but also gain conceptual insight into the decision frameworks used by service managers to deliver timely, cost-effective, and client-oriented solutions. This approach bridges the gap between academic theory and industrial application, providing learners with a practical, context-specific scenario that reflects the real-world challenges of managing educational lab infrastructure projects.

Ultimately, this dataset serves as both a pedagogical tool and a simulated business environment, empowering MBA students to explore the intersection of analytics, decision-making, and service excellence within one of the most critical roles in the modern technical education supply chain.

© 2025 Dr. Yasser Aly Yakout Shehata

2. Dataset Description

The dataset comprises **1,000 service project records**, each representing a unique event such as installation, maintenance, commissioning, or training related to educational laboratory equipment. It includes a mix of categorical and numerical data, allowing for comprehensive multidimensional analysis. Additionally, it contains **three analytically derived variables** that enhance interpretability and support AI applications.

Each record corresponds to the execution of a technical task delivered to a client institution, providing detailed information on project timelines, equipment types, technician effort, service costs, customer satisfaction, and risk categorisation.

3. Purpose and Application

This dataset enables students to:

- Practise business analytics techniques in a sector-specific context
- Conduct EDA using statistical tools and visual methods
- Build interactive dashboards in Power BI to track key performance indicators (KPIs)
- Apply supervised learning models to classify service risk or predict costs
- Gain managerial insight into operational bottlenecks and customer trends

It aligns theoretical learning with applied analysis, preparing students for data-driven roles in technical service operations, project management, and client success analytics.

4. Dataset Structure

- **Total Records**: 1,000
- **Total Variables**: 20
    - **Qualitative (Categorical)**: 7
    - **Quantitative (Numerical)**: 10
    - **Derived (Analytical)**: 3

5. Qualitative Variables

1. **Project_ID**: Unique identifier for each service project
2. **Equipment_Type**: Category of educational equipment (e.g. Robotics, PLC, CNC)
3. **Client_Type**: Nature of client institution (University, Technical College, etc.)

© 2025 Dr. Yasser Aly Yakout Shehata

4. **Service_Type**: Type of service performed (Installation, Maintenance, etc.)
5. **Region**: Geographic area served (Cairo, Alexandria, etc.)
6. **Technician_Team**: Assigned technician group (Team A, B, or C)
7. **Client_Satisfaction_Category**: Categorical summary of customer feedback (Positive, Neutral, Negative)

6. Quantitative Variables

8. **Schedule_Duration_days**: Number of days planned for the service project
9. **Actual_Duration_days**: Actual number of days taken for project completion
10. **Equipment_Count**: Number of equipment units installed or serviced
11. **Technician_Hours**: Total hours worked by technicians on the project
12. **First_Response_Time_hr**: Time (in hours) to acknowledge service request
13. **Resolution_Time_hr**: Time (in hours) to complete and close the service ticket
14. **Service_Cost_usd**: Financial cost of the service project in US dollars
15. **Equipment_Age_months**: Age of equipment at time of service
16. **Repeat_Service_Count**: Number of times follow-up visits were needed
17. **Customer_Feedback_Score**: Numeric satisfaction score from 1 to 5

7. Derived Variables

18. **Schedule_Adherence_Percent**: Ratio of planned to actual duration as a percentage
19. **Service_Efficiency_Index**: Composite index measuring efficiency (technician hours per resolution time normalised by cost)
20. **Risk_Flag**: Binary label indicating whether a service event was considered high risk (based on schedule delays or negative feedback)

8. Relevance to Power BI and AI for Business

This dataset is fully compatible with:

- **Power BI dashboards** featuring visual drill-downs, slicers, cards, and regional maps
- **Data modelling** through DAX expressions for creating calculated fields and measures

© 2025 Dr. Yasser Aly Yakout Shehata

- **AI integration** for risk classification, forecasting resolution time, and clustering clients or equipment types

Its structured schema and well-distributed variables support a variety of visual and algorithmic techniques for business analytics.

9. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Compare planned vs actual project durations
- Identify trends in technician workloads by region or service type
- Assess patterns in customer satisfaction and repeat service needs

B. Power BI Visualisation

- Build dashboards showing project efficiency, cost breakdowns, and service timelines
- Visualise technician team performance and regional delivery metrics
- Create client segmentation views based on satisfaction and service type

C. AI-Driven Business Insights

- Predict likelihood of project delay using logistic regression or decision trees
- Forecast service resolution time based on technician deployment and client type
- Cluster projects by efficiency and satisfaction for resource planning

10. Conclusion

The **Enhanced Technical Services Analytics Dataset** offers a rich, realistic, and analytically robust environment for MBA students to apply business analytics tools in a practical domain. Focused on the responsibilities of a Technical Manager in a high-tech educational lab provider, it merges operational realism with analytical depth. Students can simulate managerial decision-making, generate visual insights, and explore AI-driven predictions preparing them to lead data-informed projects in technology-intensive service sectors.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 64: Marketing Automation Insights Analytics

Introduction

In the contemporary digital economy, marketing automation has become a cornerstone of effective customer engagement, brand communication, and revenue optimisation. As consumers interact across multiple platforms and channels, businesses increasingly rely on data-driven strategies to personalise content, streamline outreach, and evaluate campaign performance in real time. This transformation has elevated the role of the **Content and Marketing Automation Specialist**, a professional who orchestrates these digital campaigns, harnesses behavioural insights, and leverages automation tools to maximise audience impact and marketing efficiency.

This dataset, titled **Marketing Automation Insights Analytics**, has been developed with a focus on the operational and analytical landscape experienced by such specialists. It simulates the environment in which digital marketers operate—especially in SaaS-based platforms such as Pelcro, where subscription models, user behaviour tracking, and personalised content delivery are integral to organisational growth.

Specifically tailored for MBA students pursuing the ***Business Analytics for Decision Making* course**, the dataset provides a rich and realistic framework for applying key analytical techniques. It reflects the types of variables, metrics, and derived insights that a Content and Marketing Automation Specialist would regularly examine to improve campaign outcomes, identify engagement trends, and make strategic decisions grounded in data.

Moreover, this dataset aligns with the academic objectives of developing fluency in data analytics platforms such as Power BI, conducting exploratory and descriptive analyses, and introducing AI-assisted decision-making. By engaging with this dataset, students are not only able to practice technical skills but also critically evaluate marketing strategies, assess ROI, and propose data-informed recommendations.

As such, this dataset serves as a capstone environment connecting practical industry realities with analytical competencies, and offering a realistic simulation for students aspiring to lead in the fields of marketing analytics, automation strategy, and digital business intelligence.

Dataset Description

This dataset comprises **1,000 records**, each simulating a marketing automation event involving campaign delivery, customer engagement, and performance tracking. Each row represents an individual subscriber's interaction with a campaign, reflecting a diverse array

© 2025 Dr. Yasser Aly Yakout Shehata

of conditions including trigger events, audience segmentation, campaign goals, and regional characteristics.

The data incorporates **seven qualitative variables**, **ten quantitative variables**, and **five derived variables**, making it ideal for hands-on analytical exercises in EDA, visualisation, and AI-driven modelling.

Purpose and Application

The dataset serves the following pedagogical and professional objectives:

- **Academic Training**: Enables MBA students to apply statistical and visual analysis tools in realistic scenarios.
- **Practical Simulation**: Reflects operational data encountered by marketing professionals using platforms like Pelcro.
- **Decision-Making Practice**: Offers a sandbox for exploring campaign performance metrics, churn risk, and ROI optimisation.
- **AI Applications**: Supports basic machine learning tasks such as classification and clustering relevant to customer engagement and campaign outcome prediction.

Dataset Structure

- **Total Records**: 1000
- **File Format**: CSV
- **Variables**: 22 (7 qualitative, 10 quantitative, 5 derived)

Qualitative Variables (7)

1. **Campaign_Type** – The type of marketing campaign executed (e.g., Newsletter, Promotional).
2. **Audience_Segment** – The category of users targeted (e.g., Returning Users, Premium Members).
3. **Trigger_Event** – The system event that initiated the campaign (e.g., Subscription Created, Payment Failed).
4. **Channel** – The medium through which the campaign was delivered (e.g., Email, SMS).
5. **Content_Theme** – The primary subject of the campaign message (e.g., Educational Tips, Personal Branding).

© 2025 Dr. Yasser Aly Yakout Shehata

6. **Automation_Status** – Current state of the campaign automation (e.g., Active, Completed).
7. **Customer_Region** – Geographical region of the subscriber (e.g., North America, MENA).

Quantitative Variables (10)

8. **Subscriber_ID** – An anonymised unique identifier for each subscriber.
9. **Campaign_Reach** – Total number of users targeted in the campaign.
10. **Open_Rate_Percent** – Percentage of users who opened the campaign message.
11. **Click_Through_Rate** – Percentage of users who clicked embedded links.
12. **Bounce_Rate_Percent** – Percentage of messages that were not successfully delivered.
13. **Conversion_Rate** – Percentage of users who performed a target action post-campaign.
14. **Time_to_Open_Min** – Average time taken to open the message after delivery.
15. **Time_to_Conversion_Hr** – Time taken between message open and conversion.
16. **Campaign_Cost_USD** – Cost incurred in executing the campaign.
17. **Revenue_Generated_USD** – Total revenue attributed to campaign conversion.

Derived Variables (5)

18. **Engagement_Score** – Composite metric based on weighted averages of open, click, and conversion rates.
19. **ROI_Percent** – Return on investment calculated from campaign revenue and cost.
20. **Churn_Risk_Flag** – Binary flag indicating likelihood of subscriber churn based on high bounce and low conversion.
21. **Campaign_Performance_Label** – Categorical label (High, Moderate, Low) based on Engagement_Score.
22. **Revenue_Per_Reach** – Efficiency metric reflecting revenue earned per individual targeted.

© 2025 Dr. Yasser Aly Yakout Shehata

Relevance to Power BI and AI for Business

This dataset is highly compatible with Power BI's analytical capabilities:

- **Categorical filtering** for segmentation
- **Dynamic dashboards** for tracking engagement
- **DAX functions** for ROI computation
- **Time series charts** for conversion time evaluation
- **AI insights and Copilot integration** for generating auto-summaries and predictive flags

It also provides a strong foundation for developing AI applications including customer segmentation models, campaign success predictors, and churn classification.

Applications in Business Analytics and AI

A. **Exploratory Data Analysis (EDA)** Students can use Python or Excel to:

- Identify trends in open and conversion rates
- Visualise audience responses by region
- Perform correlation analysis across performance metrics

B. **Power BI Visualisation** Visual dashboards can be created to:

- Compare campaign performance across segments
- Monitor real-time ROI and revenue contribution
- Track high-risk churn segments geographically

C. **AI-Driven Business Insights** Using machine learning, the dataset can support:

- Predictive modelling of campaign success
- Risk scoring for customer churn
- Classification of campaign performance labels based on historical attributes

Conclusion

The **Marketing Automation Insights Analytics** dataset is an essential resource for MBA students seeking to apply business analytics in real-world contexts. It simulates the daily responsibilities and decision parameters of a **Content and Marketing Automation**

© 2025 Dr. Yasser Aly Yakout Shehata

Specialist, bridging classroom theory with professional application. Its structured mix of categorical, numerical, and derived variables supports multi-layered analysis, making it a versatile dataset for academic projects, Power BI visualisation, and introductory AI implementation in marketing strategy.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 65: CEG

Introduction

The **CEG International Business Dataset** is a carefully designed dataset that simulates real-world business scenarios tailored specifically for **CEG International**, a leading international engineering consultancy firm operating in Qatar. Known for its excellence in architecture, interior design, master planning, landscape architecture, urban design, and building consultancy, CEG International has established itself as a prominent player in Qatar's construction and development sector. This dataset is designed to support **Exploratory Data Analysis (EDA)**, **Power BI Visualization**, and **Descriptive Analytics** tasks, enabling users to gain actionable insights into the company's operations and performance metrics.

The dataset comprises **1,000 records**, with **5 qualitative variables** and **8 quantitative variables**, further enhanced by **3 derived metrics** to provide deeper analytical insights. By reflecting key business performance indicators such as project revenue, R&D budgets, market share, and customer retention, the dataset enables students and professionals to analyze, visualize, and interpret data in ways that mirror CEG International's operational context. It serves as an invaluable tool for exploring business analytics applications within the architecture, engineering, and construction (AEC) industry in Qatar.

With a focus on **Qatar's unique economic environment** and the specific services offered by CEG International, the dataset is highly relevant for studying the role of AI and analytics in improving decision-making, optimizing operations, and enhancing service delivery in the AEC sector.

Dataset Description

The **CEG International Business Dataset** includes **16 variables**, categorized into **qualitative (categorical)** and **quantitative (numerical)** data types. The dataset also incorporates **derived variables** that add an additional layer of analytical depth, enabling users to benchmark performance and identify trends specific to the AEC industry. Below is a detailed description of the dataset's structure:

1. Qualitative Variables

These categorical variables provide foundational data for segmentation, comparison, and grouping of records:

© 2025 Dr. Yasser Aly Yakout Shehata

- **Service Area**: Represents the area of expertise for a given project. Categories include *Architecture, Interior Design, Master Planning, Landscape Architecture, Urban Design, Building Consultancy*.
- **Project Type**: Indicates the type of project delivered. Categories include *Residential, Commercial, Infrastructure, Healthcare, Educational, Governmental*.
- **Region**: Represents the geographic location within Qatar where the project was executed. Categories include *Doha, Al Rayyan, Al Wakrah, Umm Salal, Al Khor*.
- **Client Type**: Defines the client for whom the project was delivered. Categories include *Government, Private Sector, SMEs, International Corporations, Non-profits*.
- **AI Focus Area**: Highlights the AI application area relevant to the project, such as *Smart Infrastructure, Predictive Analytics, Computer Vision for Design, Autonomous Systems for Construction Management*.

2. Quantitative Variables

These variables capture the numerical aspects of CEG International's operations and performance:

- **Project Revenue (QAR)**: The total revenue generated by the project, measured in Qatari Riyal (QAR). This serves as a primary financial performance indicator.
- **Number of Team Members**: Represents the workforce size allocated to a project, including architects, engineers, and consultants.
- **AI R&D Budget (QAR)**: The budget allocated to AI-driven research and development for enhancing project delivery, such as AI for smart infrastructure or predictive maintenance.
- **Project Acquisition Cost (QAR)**: The cost incurred in acquiring a project, including proposal development and client outreach expenses.
- **Market Share (%)**: The percentage of the market captured by the project category within Qatar's AEC industry.
- **Client Retention Rate (%)**: The percentage of repeat clients within a given period, reflecting client satisfaction and loyalty.
- **Design Iterations**: The number of design iterations or revisions completed during the project lifecycle.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Revenue Growth Rate (%)**: The year-over-year growth in revenue generated from similar projects, highlighting business scalability and demand trends.

3. Derived Variables

To enhance the dataset's analytical potential, the following derived variables have been added:

- **Profit Margin (%)**: Measures the profitability of a project after accounting for AI R&D expenses. This is calculated as:

$$\text{Profit Margin}(\%) = \frac{\text{Project Revenue (QAR)} - \text{AI R\&D Budget (QAR)}}{\text{ProjectRevenue (QAR)}} \times 100$$

This metric evaluates the financial efficiency of CEG International's project execution.

- **Cost-to-Revenue Ratio (%)**: Indicates the percentage of project revenue spent on acquisition costs, calculated as:

$$\text{Cost} - \text{to} - \text{Revenue Ratio } (\%) = \frac{\text{Project Acquisition Cost (QAR)}}{\text{Project Revenue (QAR)}} \times 100$$

This metric helps assess the cost-effectiveness of acquiring projects.

- **Revenue Per Team Member (QAR)**: Reflects the productivity of the workforce by measuring the revenue generated per team member:

$$\text{Revenue Per Team Member (QAR)} = \frac{\text{Project Revenue (QAR)}}{\text{Number of Team Members}}$$

This variable is useful for evaluating efficiency in project delivery.

Relevance to EDA and Visualization in Power BI

The **CEG International Business Dataset** has been carefully designed to facilitate **EDA** and **Power BI visualization**, providing opportunities to uncover patterns, trends, and actionable insights within Qatar's AEC sector. Here's how this dataset is relevant to various analytical tasks:

Exploratory Data Analysis (EDA)

- **Performance Metrics**: Analyze the relationship between **AI R&D Budgets** and **Profit Margins**, or between **Client Retention Rates** and **Project Types**.
- **Market Analysis**: Examine distributions of **Market Share** and **Revenue Growth Rates** to identify high-demand services in the AEC industry.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Efficiency Metrics**: Explore how **Revenue Per Team Member** varies across **Service Areas** or **Regions**, providing insights into workforce productivity.

Power BI Visualization

- **Dashboards**:
    - Create interactive dashboards tracking KPIs like **Project Revenue**, **Profit Margin**, and **Client Retention Rates**.
    - Develop heatmaps to visualize **Revenue** or **Market Share** distribution across Qatar's regions, such as Doha or Al Rayyan.
- **Trends and Comparisons**:
    - Compare **Revenue Growth Rates** across **Service Areas** (e.g., Architecture vs. Master Planning).
    - Visualize trends in **Client Retention Rates** for Government vs. Private Sector clients.
- **Segmentation**:
    - Segment projects by **AI Focus Area** to evaluate the financial and operational impact of AI technologies like Smart Infrastructure or Predictive Analytics.

Descriptive Analytics

- Summarize average **Revenue**, **R&D Budgets**, and **Design Iterations** for different **Service Areas** and **Client Types**.
- Benchmark projects by derived metrics like **Profit Margin** and **Revenue Per Team Member** to identify top-performing categories.
- Identify growth opportunities in underutilized regions (e.g., Al Wakrah, Umm Salal).

Conclusion

The **CEG International Business Dataset** is a comprehensive and highly tailored dataset that reflects the operational realities of Qatar's leading engineering consultancy firm. It enables users to simulate real-world analytical tasks, including EDA, Power BI visualization, and descriptive analytics, while focusing on critical metrics such as project profitability, efficiency, and client satisfaction. By integrating realistic variables and derived metrics, this dataset equips students and professionals to explore the intersection of **AI applications** and business analytics in the AEC sector, empowering data-driven decision-making and operational excellence.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 66: CARRIER

Introduction

The **HVAC (Heating, Ventilation, and Air Conditioning) industry** is witnessing rapid technological advancements, with **Carrier** being a pioneer in delivering energy-efficient and AI-integrated solutions. This dataset, crafted to simulate real-world Carrier business operations, serves as a comprehensive resource for MBA students specializing in Business Analytics. It offers a unique opportunity to explore the relationships between customer behavior, product performance, and operational efficiency, emphasizing the transformative role of AI in optimizing HVAC solutions.

Designed for **Exploratory Data Analysis (EDA)** and visualization in **Power BI**, this dataset reflects practical business scenarios involving Carrier's product offerings, such as chillers, air handlers, packaged units, and control systems. The integration of **derived variables** enhances analytical depth, allowing students to uncover actionable insights and develop AI-driven strategies for business optimization. This dataset equips students with the tools to analyze the interplay between revenue, operational efficiency, and customer satisfaction, making it highly relevant for applications in AI for business.

Dataset Description

The dataset contains **1,000 records**, each representing a transaction involving Carrier products. It includes **five qualitative variables** and **eight quantitative variables**, as well as **three derived variables**, making it well-suited for data exploration and advanced analytics. Below is a detailed description of the dataset's structure:

1. Qualitative Variables

These variables offer insights into customer segmentation and transaction context:

- **Customer Segment**: Represents the type of customer (e.g., "Retail," "Corporate," "Industrial," "Educational," "Healthcare").
    - **Use Case in EDA**: Analyze which customer segments contribute most to Carrier's revenue.
- **Product Category**: Categorizes products into "Chillers," "Air Handlers," "Packaged Units," "Controls," and "Condensing Units."
    - **Use Case in Visualization**: Compare the sales performance of product categories to determine which products are driving demand.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Region**: Represents the geographical area of the transaction, including "Kuwait City," "Ahmadi," "Hawalli," "Farwaniya," and "Jahra."
    - **Use Case**: Use geographic visualizations to highlight regions with the highest sales or energy-efficient installations.
- **Sales Channel**: Describes the sales method, such as "Direct," "Distributor," or "Online."
    - **Use Case**: Analyze how sales channels impact revenue and customer satisfaction.
- **Customer Satisfaction**: Ranges from "Very Satisfied" to "Very Dissatisfied," capturing customer feedback.
    - **Use Case**: Examine how satisfaction levels correlate with product categories, sales channels, and energy efficiency.

2. Quantitative Variables

These variables provide numerical data for in-depth analysis:

- **Units Sold**: The number of units sold in each transaction (integer).
    - **Use Case**: Identify product categories with the highest sales volumes.
- **Revenue (KWD)**: Revenue generated in Kuwaiti Dinars per transaction (continuous).
    - **Use Case**: Core metric to evaluate financial performance and profitability.
- **Energy Efficiency Rating (EER)**: A measure of product energy performance, ranging from 8 to 20 (continuous).
    - **Use Case**: Analyze the relationship between energy efficiency and customer satisfaction.
- **Installation Time (hours)**: Time required for product installation per transaction (continuous).
    - **Use Case**: Evaluate the impact of installation time on customer satisfaction and revenue.
- **Maintenance Frequency (annual)**: Number of maintenance sessions required annually (integer).
    - **Use Case**: Examine maintenance trends across product categories and their operational impact.

© 2025 Dr. Yasser Aly Yakout Shehata

- **AI Integration Score**: A score from 1 to 10 measuring the degree of AI integration in Carrier's control systems (continuous).
    - **Use Case**: Analyze how AI integration enhances operational efficiency and reduces downtime.
- **Downtime Reduction (%)**: The percentage reduction in system downtime achieved through AI-based controls (continuous).
    - **Use Case**: Quantify the operational benefits of AI integration.
- **Project Duration (weeks)**: Time taken to complete a project, from order to final implementation (continuous).
    - **Use Case**: Explore how project duration affects customer satisfaction and overall performance.

3. Derived Variables

To enhance the dataset's analytical value, the following variables are derived:

- **Average Revenue per Unit (KWD)**:
    - **Formula**: $\text{Revenue} / \text{Units Sold}$
    - **Purpose**: Helps analyze pricing strategies and per-unit profitability, critical for Carrier's product-level insights.
    - **Visualization**: Compare profitability across product categories or regions in Power BI.
- **Maintenance Cost (KWD)**:
    - **Formula**: $\text{Maintenance Frequency} \times 100 \text{ (assuming 100 KWD per session)}$
    - **Purpose**: Provides insight into annual maintenance expenses for customers, enabling cost analysis and operational optimization.
    - **Visualization**: Highlight maintenance cost trends across product categories or customer segments.
- **Total Downtime Hours Saved (hours)**:
    - **Formula**: $(\text{Downtime Reduction (\%)} \times \text{Installation Time}) / 100$
    - **Purpose**: Quantifies the efficiency gains from AI integration, showcasing its value in minimizing system downtime.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Visualization**: Demonstrate the impact of AI integration on operational efficiency using Power BI dashboards.

Relevance to EDA and Power BI Visualization

This dataset is specifically designed to facilitate **Exploratory Data Analysis (EDA)** and create impactful visualizations in **Power BI**, with a strong focus on AI's role in business transformation. Here's why it is ideal for these purposes:

1. **Exploratory Data Analysis (EDA)**:
    - Perform **descriptive statistics** (e.g., mean, median, variance) on quantitative variables like Revenue, EER, and AI Integration Score.
    - Use **bivariate analysis** to uncover relationships between variables, such as how Energy Efficiency Rating (EER) impacts Revenue or Customer Satisfaction.
    - Identify **outliers and anomalies**, such as extreme values in revenue or maintenance frequency.
2. **Power BI Applications**:
    - **Trend Analysis**: Create dynamic line charts to visualize revenue trends over time.
    - **Geographic Insights**: Build heat maps to identify high-demand regions and their correlation with product categories.
    - **Category Comparisons**: Use bar and column charts to compare product categories based on revenue, units sold, or customer satisfaction.
    - **AI Impact Dashboards**: Develop AI-focused dashboards to showcase how AI Integration Score and Downtime Reduction (%) correlate with operational efficiency.
3. **AI for Business Applications**:
    - **Operational Efficiency**: Demonstrate the ROI of AI integration by linking AI Integration Score and Downtime Reduction (%) to key metrics such as Revenue and Customer Satisfaction.
    - **Customer Insights**: Segment customers based on satisfaction levels and explore how AI-enhanced systems meet their needs.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Predictive Analytics**: Use the dataset to train machine learning models for revenue prediction or customer satisfaction classification.

Conclusion

This dataset is an invaluable tool for MBA students and business analysts aiming to understand the role of data in driving AI-based innovations for companies like **Carrier**. By incorporating both core and derived variables, it enables detailed exploratory analyses and advanced visualizations in Power BI. The dataset highlights the interplay between customer preferences, product performance, and operational efficiency, equipping students with actionable insights for decision-making in the **HVAC sector**. Furthermore, its relevance to AI for business underscores the transformative impact of technology in modern enterprises.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 67: ABP

Introduction

The dataset has been designed to reflect key aspects of **Al Jomaih Bottling Plants (ABP)**, one of the largest PepsiCo bottlers in the Middle East and North Africa (MENA) region. With its comprehensive manufacturing and distribution network, ABP operates at the forefront of the beverage industry, producing a wide range of products such as carbonated drinks, water, and juices. The company is also a leader in sustainability and operational efficiency, with subsidiaries that complement its core business, such as **International Refreshment Company (IRC)** and **Al Jomaih Cans & Ends Making Plants (CMP)**.

This dataset captures operational, financial, and sustainability metrics of ABP's business to enable MBA students to explore data-driven decision-making. By focusing on key performance indicators (KPIs) such as sales, customer satisfaction, production costs, and sustainability ratings, the dataset provides a realistic foundation for **Exploratory Data Analysis (EDA)** and **Visualization in Power BI**. Additionally, derived variables such as **Profit Margin**, **Sustainability Score**, and **CO2 Emissions Per Unit** add analytical depth, enabling students to examine the relationships between business performance, environmental impact, and customer satisfaction.

By leveraging this dataset, students can simulate real-world analytical scenarios that ABP might encounter, such as understanding regional sales trends, optimizing marketing expenditures, improving delivery performance, and balancing sustainability efforts with profitability. This aligns with ABP's focus on innovation and continuous improvement in the beverage industry, making the dataset highly relevant for studies in AI-driven analytics for business.

Dataset Description

The dataset contains **1,000 records** and includes **5 qualitative variables**, **8 quantitative variables**, and **3 derived variables**. These variables reflect the key dimensions of ABP's operations, enabling students to analyze various aspects of the company's business performance. Below is a detailed description of the dataset and its specific relevance to ABP.

Qualitative Variables

1. **Region**: The geographical area served by ABP's operations (e.g., North, Central, East, West, International).

© 2025 Dr. Yasser Aly Yakout Shehata

- **Relevance**: ABP serves diverse regions in Saudi Arabia and beyond. Regional differences may influence sales, delivery performance, and customer satisfaction.

2. **Product_Type**: The category of ABP's products (e.g., Carbonated Beverage, Water, Juice, Metal Packaging, AI Service).
    - **Relevance**: Reflects ABP's portfolio of products, including PepsiCo beverages and metal packaging services.
3. **Customer_Segment**: The primary customer group (e.g., Retailer, Wholesaler, Distributor, Direct Consumer, E-Commerce).
    - **Relevance**: ABP caters to a wide range of customers, and this variable helps analyze sales distribution across different customer types.
4. **Distribution_Channel**: The primary mode of distribution (e.g., Traditional Retail, Modern Trade, Online, Direct Sales, Third-Party Logistics).
    - **Relevance**: Distribution efficiency is crucial for ABP's success in meeting customer demand and ensuring on-time delivery.
5. **Sustainability_Rating**: A rating of ABP's sustainability initiatives (e.g., Excellent, Good, Average, Poor, Not Rated).
    - **Relevance**: As part of Al Jomaih Holding Company, ABP places a strong emphasis on sustainability, making this variable critical for analyzing environmental performance.

Quantitative Variables

1. **Monthly_Sales (USD)**: The total sales revenue generated by ABP per month (range: $10,000–$500,000).
    - **Relevance**: Sales revenue is a key indicator of ABP's market performance and product demand.
2. **Production_Cost (USD)**: The monthly cost of producing beverages or packaging (range: $5,000–$300,000).
    - **Relevance**: Helps analyze cost efficiency and profitability across different products and regions.
3. **Units_Sold**: The total number of beverage units or packaging units sold per month (range: 1,000–50,000).

© 2025 Dr. Yasser Aly Yakout Shehata

- **Relevance**: Reflects ABP's production scale and market reach.

4. **Marketing_Expenditure (USD)**: The monthly spending on marketing campaigns (range: $1,000–$50,000).
    - **Relevance**: ABP's marketing efforts play a significant role in driving sales and promoting its products.
5. **Employee_Count**: The total number of employees working in ABP's facilities or regions (range: 50–500).
    - **Relevance**: Reflects the operational scale of ABP and its workforce's impact on production and distribution.
6. **On_Time_Delivery_Rate (%)**: Percentage of orders delivered on time (range: 50%–100%).
    - **Relevance**: A critical metric for ABP's distribution network, which ensures customer satisfaction.
7. **Customer_Satisfaction_Score**: A measure of customer satisfaction (range: 1–10).
    - **Relevance**: Provides insights into customer perceptions of ABP's products and services.
8. **CO2_Emissions (tons)**: Monthly CO2 emissions generated by ABP's production and logistics (range: 10–500).
    - **Relevance**: ABP's commitment to reducing environmental impact makes this a key variable for sustainability analysis.

Derived Variables

1. **Profit_Margin (%)**:
    - **Formula**: $\text{Profit Margin} = \frac{\text{Monthly Sales} - \text{Production Cost}}{\text{Monthly Sales}} \times 100$
    - **Relevance**: Measures ABP's profitability, enabling analysis of cost-effectiveness and pricing strategies.
2. **Sustainability_Score**:
    - **Formula**: Conversion of Sustainability_Rating into numerical scores (Excellent = 5, Poor = 1).

© 2025 Dr. Yasser Aly Yakout Shehata

- **Relevance**: Allows students to compare ABP's sustainability performance across regions and product types.

3. **CO2_Emissions_Per_Unit**:
   - **Formula**: $\text{CO2 Emissions Per Unit} = \frac{\text{CO2 Emissions}}{\text{Units Sold}}$
   - **Relevance**: Highlights ABP's environmental efficiency, linking production scale to its carbon footprint.

Relevance to EDA and Power BI Visualization

The dataset provides a robust foundation for performing **Exploratory Data Analysis (EDA)** and developing dynamic **Power BI dashboards** that simulate ABP's operational challenges and opportunities. Below are the specific applications of the dataset:

1. Exploratory Data Analysis (EDA)

- **Sales Performance**:
  - Analyze Monthly_Sales by region, product type, and customer segment to identify the most profitable markets and products.
- **Cost Analysis**:
  - Investigate the relationship between Production_Cost, Marketing_Expenditure, and Profit_Margin.
- **Sustainability Insights**:
  - Explore how Sustainability_Rating and CO2_Emissions_Per_Unit vary across regions and product categories.
- **Delivery Analysis**:
  - Examine the impact of On_Time_Delivery_Rate on Customer_Satisfaction_Score.

2. Power BI Visualization

- **Interactive Dashboards**:
  - Build dashboards to track KPIs such as sales, customer satisfaction, and delivery performance across regions.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Regional Analysis**:
    - Use geographical maps to visualize regional differences in Monthly_Sales, Profit_Margin, and CO2_Emissions_Per_Unit.
- **Sustainability Metrics**:
    - Create visualizations to track ABP's sustainability progress, linking Sustainability_Score to Customer_Satisfaction_Score.
- **Trend Analysis**:
    - Simulate time-series visualizations to analyze sales trends and predict future performance.

3. Application in AI for Business

- **Predictive Modeling**:
    - Use AI techniques to predict Customer_Satisfaction_Score or Monthly_Sales based on marketing and delivery metrics.
- **Optimization**:
    - Analyze how ABP can optimize Marketing_Expenditure to maximize Profit_Margin.
- **Sustainability Improvement**:
    - Explore strategies to reduce CO2_Emissions_Per_Unit without compromising production scale or profitability.

Conclusion

This dataset represents a comprehensive and realistic simulation of **Al Jomaih Bottling Plants (ABP)** operations, providing MBA students with the tools to practice EDA, visualization in Power BI, and AI-driven business analytics. By focusing on operational efficiency, sustainability, and customer satisfaction, the dataset mirrors ABP's commitment to innovation and continuous improvement. Whether analyzing sales trends, building dashboards, or applying predictive models, students can develop actionable insights to tackle real-world challenges in the beverage and packaging industry.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 68: GHUDAI

Introduction

The **GHUDAI Business Analytics Dataset** is a meticulously crafted dataset designed to empower **MBA students, data analysts, and business professionals** in exploring the application of **Artificial Intelligence (AI) in business operations** through **Exploratory Data Analysis (EDA), visualization in Power BI, and descriptive analytics.** The dataset reflects real-world business scenarios, making it an invaluable resource for hands-on learning and strategic decision-making.

With a focus on key business dimensions such as **financial performance, AI adoption, customer satisfaction, and sustainability,** the GHUDAI dataset enables users to uncover trends, analyze performance metrics, and generate actionable insights using AI-powered tools.

Designed specifically to support analytical exercises in **AI-driven business strategy**, the dataset facilitates:

- **Exploratory analysis** to identify patterns, correlations, and trends across industries and regions.
- **Visualization through Power BI**, enabling the creation of dynamic dashboards that provide actionable insights.
- **Descriptive analytics**, offering a deep dive into the key performance indicators (KPIs) that influence business success.

The dataset incorporates **derived variables** to enhance analytical depth, allowing users to explore relationships such as the impact of AI investment on profitability, sustainability performance across different industries, and customer segmentation based on satisfaction levels.

By leveraging this dataset, users will gain practical insights into AI's role in **optimizing business operations, enhancing customer experiences, and driving sustainable growth.**

Dataset Description

The **GHUDAI Business Analytics Dataset** consists of **19 well-structured variables**, categorized into **7 categorical** and **12 numerical** fields, covering key business aspects such as revenue, market share, sustainability, and AI adoption. The dataset is built to support a variety of business analyses, offering a comprehensive foundation for **EDA, visualization, and AI-driven insights.**

© 2025 Dr. Yasser Aly Yakout Shehata

1. Categorical Variables (7 variables)

These variables facilitate segmentation and comparative analysis, helping users to analyze trends based on different business characteristics.

Variable	Description	Relevance
Region	Geographic location of the company (e.g., Europe, Asia)	Enables regional trend analysis and performance benchmarking.
Industry Segment	Business sector (e.g., Retail, Healthcare)	Facilitates industry-specific performance and AI adoption trends.
AI Application Type	Type of AI technology used (e.g., Automation, NLP)	Helps analyze the impact of AI applications on business performance.
Customer Type	Classification of customers (B2B, B2C, Government)	Enables segmentation for revenue and satisfaction analysis.
Company Size	Classification by workforce size (Small, Medium, Large)	Allows comparative analysis of growth and efficiency across companies.
Quarter	Fiscal quarter of reporting (Q1, Q2, Q3, Q4)	Supports seasonal trend analysis and forecasting.
Customer Segmentation (Derived)	Categorized satisfaction levels (Low, Medium, High)	Facilitates targeted marketing and service improvement strategies.

2. Numerical Variables (12 variables)

These quantitative variables provide key business performance indicators that can be analyzed to track progress, identify trends, and develop predictive models.

Variable	Description	Relevance
Annual Revenue ($M)	Total revenue in millions of USD	Measures financial performance across regions and industries.

© 2025 Dr. Yasser Aly Yakout Shehata

Variable	Description	Relevance
Number of Employees	Workforce size	Assesses operational capacity and workforce efficiency.
R&D Investment ($M)	Investment in AI and innovation	Tracks investment in business growth and technology adoption.
Market Share (%)	Percentage of market share held	Evaluates competitive positioning and market strength.
Customer Satisfaction Score	Score measuring customer happiness (0-100)	Key indicator of customer loyalty and retention.
AI Adoption Rate (%)	Level of AI implementation in operations	Assesses AI-driven transformation and innovation adoption.
Energy Usage (MWh)	Monthly energy consumption in megawatt-hours	Enables sustainability tracking and operational efficiency.
Carbon Footprint (metric tons)	Total carbon emissions per year	Measures environmental impact and supports ESG reporting.
Profit Margin (%)	Net profit as a percentage of revenue	Analyzes financial efficiency and profitability.
Revenue per Employee ($) (Derived)	Revenue efficiency per employee	Helps understand productivity and workforce efficiency.
R&D to Revenue Ratio (%) (Derived)	R&D spending relative to revenue	Evaluates the effectiveness of innovation investments.
Year	Year of business operation (2020-2024)	Provides a timeline for historical analysis and forecasting.

Derived Variables for Enhanced Insights

To improve analytical depth and business insights, the dataset includes **derived variables** that allow users to perform in-depth analysis and benchmarking:

1. **Revenue per Employee ($):**

© 2025 Dr. Yasser Aly Yakout Shehata

- **Formula:** Annual Revenue / Number of Employees
- **Insight:** Helps measure employee productivity and operational efficiency.

2. **R&D to Revenue Ratio (%):**
    - **Formula:** (R&D Investment / Annual Revenue) * 100
    - **Insight:** Assesses how effectively companies are investing in innovation.
3. **Customer Segmentation:**
    - **Formula:** Based on Customer Satisfaction Score:
        - High Satisfaction (Score > 85)
        - Medium Satisfaction (Score 70-85)
        - Low Satisfaction (Score < 70)
    - **Insight:** Enables businesses to develop targeted strategies for different customer groups.

Relevance to EDA and Visualization in Power BI

The **GHUDAI dataset** is tailored for **Exploratory Data Analysis (EDA)** and **visual storytelling in Power BI,** offering students and professionals an opportunity to:

1. **Perform EDA:**
    - Identify patterns and trends in AI adoption across industries and regions.
    - Detect relationships between variables such as R&D investment and market share.
    - Analyze seasonality effects using quarterly financial data.
2. **Develop Power BI Dashboards:**
    - Create interactive region-specific dashboards showing revenue and AI adoption trends.
    - Build visualizations such as bar charts, heatmaps, and time-series graphs for trend analysis.
    - Compare business performance across customer segments and company sizes.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Utilize Descriptive Analytics:**
    - Summarize data using key statistical metrics (mean, median, standard deviation).
    - Segment businesses based on AI adoption levels and financial success.
    - Generate insights to guide strategic decision-making.

Potential Use Cases for AI in Business Analysis

1. **Predictive Analysis:**
    - Forecast revenue growth based on AI adoption and market trends.
2. **Customer Experience Optimization:**
    - Identify factors driving high customer satisfaction scores.
3. **Sustainability Benchmarking:**
    - Evaluate environmental impact based on energy usage and carbon footprint metrics.
4. **Investment Decision-Making:**
    - Assess how R&D investments contribute to market expansion and profitability.

Conclusion

The **GHUDAI Business Analytics Dataset** is an essential resource for gaining practical insights into AI applications in business, financial planning, and sustainability. Through EDA, visualization in Power BI, and descriptive analytics, users can develop actionable insights to drive AI-powered business strategies and enhance organizational performance.

By leveraging the dataset, students and professionals can gain valuable hands-on experience in analyzing complex business data and making informed strategic decisions in an AI-driven world.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 69: DARALRIYADH

Introduction

In an era where data-driven decision-making is central to business success, the ability to analyze and visualize complex datasets has become a critical skill for professionals, particularly in fields like Artificial Intelligence (AI) for Business. This dataset is designed specifically for MBA students specializing in Business Analytics, with a focus on Exploratory Data Analysis (EDA), visualization using Power BI, and descriptive analytics. It simulates real-world business scenarios drawn from Dar Al Riyadh's operations, a leading service provider in Saudi Arabia offering expertise in engineering, design, advisory services, construction management, project supervision, and geospatial solutions.

The dataset provides a comprehensive view of project-level data, encompassing qualitative and quantitative variables that reflect diverse aspects of project planning, execution, and evaluation. By exploring this dataset, students can gain hands-on experience with industry-relevant data while building the skills necessary to extract actionable insights through advanced analytics and visualization techniques.

Description of the Dataset

This dataset comprises **1,000 records** representing projects executed across various sectors and services in Saudi Arabia. It includes **5 qualitative variables** and **8 quantitative variables**, making it ideal for multi-dimensional analysis. The variables have been thoughtfully selected and generated to maintain realistic relationships, enabling meaningful exploration of the interplay between factors such as budgets, timelines, team sizes, and client satisfaction.

Qualitative Variables

1. **Project ID**: A unique identifier for each project, ensuring traceability.
2. **Sector**: The industry sector associated with the project (e.g., Oil and Gas, Education, Healthcare, Real Estate).
3. **Service Type**: The category of service provided (e.g., Engineering and Design, Advisory Services, Construction Management).
4. **Region**: The geographical area within Saudi Arabia where the project is executed (e.g., Riyadh, Jeddah, Mecca).
5. **Client Type**: The nature of the client, classified as either "Government" or "Private Sector."

© 2025 Dr. Yasser Aly Yakout Shehata

Quantitative Variables

1. **Project Duration (Months)**: The length of time required to complete each project.
2. **Budget (Million SAR)**: The total financial allocation for each project.
3. **Team Size**: The number of professionals assigned to the project.
4. **Subcontractors**: The count of subcontracting firms involved.
5. **Satisfaction Score**: A rating provided by the client upon project completion (scale: 1–10).
6. **Change Orders**: The number of modifications requested during the project lifecycle.
7. **Safety Incidents**: The count of safety-related occurrences reported during project execution.
8. **Environmental Impact Score**: An assessment of the project's environmental footprint (scale: 1–10).

Enhanced Features

To facilitate advanced analysis, additional calculated fields are included:

1. **Cost per Month (M SAR)**: The average monthly expenditure, derived as $Budget / Project\ Duration$
2. **Subcontractor Dependency**: The ratio of subcontractors to team size, highlighting the reliance on external resources.
3. **Safety Incident Rate**: The rate of safety incidents per project month, providing insights into risk management.

Relevance to EDA and Power BI Visualization

This dataset is specifically designed to enable students to perform:

- **Exploratory Data Analysis (EDA)**: Analyze correlations, identify trends, and uncover patterns in project performance across sectors and regions.
- **Visualization in Power BI**: Create dashboards and interactive visualizations, such as heatmaps for regional distribution, bar charts for budget allocation, and scatter plots for client satisfaction analysis.

© 2025 Dr. Yasser Aly Yakout Shehata

- **AI for Business Applications**: The dataset supports the development of predictive models, such as identifying factors influencing client satisfaction or forecasting budgets for future projects.

By working with this dataset, students will not only practice technical skills but also gain strategic insights into real-world business scenarios, bridging the gap between analytics and decision-making. This dataset is an essential tool for developing actionable insights that can drive innovation and operational excellence in AI-powered businesses.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 70: RAMACO

Introduction

The RAMACO dataset has been carefully designed to support MBA students and professionals in exploring data analytics applications within the construction and contracting sector, focusing on the integration of **Artificial Intelligence (AI)** in project management and business processes. Derived from RAMACO's operational dynamics as a "Grade A" construction company in Qatar, the dataset incorporates essential business variables to reflect real-world scenarios.

This dataset is tailored for **Exploratory Data Analysis (EDA)** and **visualization using Power BI**, providing a rich foundation for uncovering insights related to project performance, client satisfaction, and efficiency metrics. It aims to enable learners to practice critical analytical skills in line with RAMACO's mission of continuous technological and professional improvement, contributing to Qatar's **National Vision 2030**.

Dataset Description

The RAMACO dataset contains **1,000 records** with **13 variables**, capturing both qualitative and quantitative aspects of construction projects managed by the company. These variables are structured to simulate project workflows, financial performance, and client outcomes, offering a robust platform for analytics.

Qualitative Variables

1. **Company_Sector**: Represents the industry focus of RAMACO's projects, primarily in "Construction" but extending to collaborative sectors like "Healthcare" or "Technology."
    - **Purpose**: Enables segmentation for analyzing sector-specific trends in AI applications.
2. **Project_Type**: Specifies the type of AI-enhanced construction solution implemented, such as "Predictive Analytics for Resource Allocation" or "Natural Language Processing for Project Communication."
    - **Purpose**: Allows evaluation of project success rates based on AI implementation type.
3. **Team_Size_Category**: Categorizes project teams as "Small," "Medium," or "Large" based on project complexity and resource requirements.
    - **Purpose**: Facilitates analysis of performance metrics by team scale.

© 2025 Dr. Yasser Aly Yakout Shehata

4. **Client_Type**: Identifies clients such as "Government," "Enterprise," "SME," or "Non-Profit."
    - **Purpose**: Helps explore relationships between client types and factors like budget, satisfaction, and ROI.
5. **Region**: Represents the geographical area where the project was executed, including regions like "Middle East," "Asia-Pacific," or "Europe."
    - **Purpose**: Enables regional trend analysis and performance comparisons.

Quantitative Variables

1. **Project_Duration_Months**: Duration of construction projects in months, ranging from short-term renovations to long-term mega-projects.
    - **Purpose**: Useful for analyzing the relationship between time, budget, and outcomes.
2. **Budget_USD_Million**: Total budget allocated for the project, measured in millions of USD.
    - **Purpose**: Central to analyzing project cost efficiency and ROI.
3. **Revenue_Generation_USD_Million**: Revenue earned from project completion, also measured in millions of USD.
    - **Purpose**: Key metric for assessing project profitability.
4. **ROI_Percentage**: Calculated as the return on investment for each project.
    - **Purpose**: Evaluates financial success relative to the budget.
5. **Client_Satisfaction_Score**: Client feedback on a scale of 1–10.
    - **Purpose**: Helps assess project outcomes from a client-centric perspective.
6. **AI_Accuracy_Percentage**: Performance metric for AI tools employed during project execution, with accuracy scores ranging from 60% to 100%.
    - **Purpose**: Correlates AI effectiveness with client satisfaction and project ROI.
7. **Training_Hours_Per_Team_Member**: The total hours spent training team members on AI systems.
    - **Purpose**: Analyzes the impact of skill development on project outcomes.
8. **Efficiency_Score**: Measures project efficiency as revenue generated per month.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Purpose**: A derived metric for evaluating operational performance.

Relevance to EDA and Visualization in Power BI

The RAMACO dataset is explicitly structured to facilitate **EDA** and visualization in Power BI, focusing on project performance and business intelligence. Key analytical opportunities include:

1. **Segment Analysis**:
    - Compare budgets, ROIs, and satisfaction scores across **client types**, **regions**, or **team sizes**.
2. **Visualization Possibilities**:
    - Create bar charts to compare project types, line charts for temporal trends in revenue, and geospatial maps for regional performance.
3. **KPI Dashboards**:
    - Use Power BI to build interactive dashboards tracking **key performance indicators (KPIs)** like client satisfaction, efficiency scores, and ROI.
4. **Trend and Correlation Analysis**:
    - Examine relationships between AI accuracy, satisfaction scores, and ROI to identify patterns in project success.
5. **Predictive Insights**:
    - Enable scenario analysis, such as forecasting how increasing training hours might improve efficiency or satisfaction.

Applications in AI for Business at RAMACO

The RAMACO dataset bridges the gap between theoretical learning and practical business applications, allowing students to explore the **integration of AI in construction management**. Relevant business use cases include:

- **Budget and ROI Optimization**: Identify the most cost-efficient projects and suggest ways to maximize returns in future initiatives.
- **Client-Centric Insights**: Analyze satisfaction data to recommend service improvements tailored to client needs.
- **AI Impact Assessment**: Evaluate the role of AI tools in improving operational performance and reducing project delays.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Regional and Sectoral Strategies**: Develop actionable insights for RAMACO's leadership by understanding which regions or sectors offer the best growth opportunities.

This introduction and description position the dataset as a powerful learning tool for MBA students and professionals, fostering their ability to derive actionable insights from real-world data in the context of AI-driven construction business processes. Let me know if you need additional customizations.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 71: ETSEC

Introduction

The **ETS Engineering Consultancy (ETSEC) Business Dataset** is a carefully designed dataset inspired by the operations and scope of services provided by the **Engineering and Technical Services Company (ETSEC)**, a leading Saudi engineering consultancy firm established in 1976. With over four decades of experience, ETSEC has played a vital role in designing, supervising, and managing national projects across Saudi Arabia, contributing to the Kingdom's economic growth and Vision 2030 goals.

This dataset simulates key performance indicators (KPIs) and business metrics relevant to ETSEC's consultancy services, including engineering design, project supervision, and management. It reflects variables that influence business performance in engineering consultancy and offers realistic data for exploratory data analysis (EDA), Power BI visualization, and descriptive analytics. The dataset focuses on ETSEC's role in projects aligned with national infrastructure development and AI-driven project management practices.

Designed specifically for educational and analytical purposes, this dataset provides MBA students and professionals studying **Business Analytics** and **AI Applications** with an opportunity to practice advanced data analysis techniques in a consulting context.

Dataset Description

The dataset contains **1,000 records**, with each record representing a project or consulting service undertaken by ETSEC across various industries and regions in Saudi Arabia. It includes both **qualitative** and **quantitative** variables, covering details about project type, client characteristics, financial performance, and AI adoption in engineering consulting. The variables are structured to reflect ETSEC's core operations and its contributions to Saudi Arabia's infrastructure and technological advancements.

Variables Overview

1. Qualitative Variables (Categorical):

1. **Project_Type**: Reflects the type of project ETSEC is involved in:
    - "Infrastructure Design," "Commercial Building Supervision," "Industrial Project Design," "Educational Facility Design," "Urban Development."
2. **Project_Size**: Indicates the size of the project, classified as:
    - "Small," "Medium," or "Large," based on project scope and budget.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Region**: Location of the project in Saudi Arabia:
    - "Riyadh," "Eastern Province," "Makkah Province," "Madinah Province," "Asir Province."
4. **Client_Sector**: The type of client receiving ETSEC's services:
    - "Government," "Private Sector," "Public-Private Partnership (PPP)."
5. **Technology_Adoption_Level**: The level of AI and digital tools used in the project:
    - "Low," "Moderate," or "High."

2. Quantitative Variables (Numerical):

1. **Project_Budget (SR)**: Total budget allocated for the project, expressed in Saudi Riyals.
    - Range: 5M–2B SR.
2. **AI_Investment (SR)**: Proportion of the project budget dedicated to AI-enabled tools and systems.
    - Range: 500K–100M SR.
3. **Personnel_Trained_AI**: Number of ETSEC employees trained in AI systems and tools for project delivery.
    - Range: 5–300 employees.
4. **Project_Duration (Months)**: Total duration of the project in months.
    - Range: 6–60 months.
5. **AI_Efficiency_Improvement (%)**: Percentage improvement in project delivery efficiency due to AI adoption.
    - Range: 5%–40%.
6. **Revenue_Generated (SR)**: Revenue earned by ETSEC from the project.
    - Range: 2M–500M SR.
7. **R&D_Expenditure (SR)**: Research and development spending for innovation in engineering tools.
    - Range: 100K–20M SR.
8. **Profit_Margin (%)**: Percentage of profit margin achieved on the project.

© 2025 Dr. Yasser Aly Yakout Shehata

- Range: 5%–25%.

3. Derived Metrics (For Enhanced Analysis):

1. **AI_Investment_Percentage**: The proportion of the project budget spent on AI. Formula: $\text{AI_Investment_Percentage} = (\frac{\text{AI_Investment}}{\text{Project_Budget}}) * 100$
2. **Revenue_Per_Duration (SR)**: Revenue earned by ETSEC per month of project duration. Formula: $\text{Revenue_Per_Duration} = \text{Revenue_Generated} / \text{Project_Duration}$
3. **Profit_Absolute (SR)**: Total profit generated from the project. Formula: $\text{Profit_Absolute} = (\text{Revenue_Generated} * \text{Profit_Margin}) / 100$

4. Scenarios for Analysis:

To simulate real-world decision-making, each record includes a **Scenario** column with predefined conditions:

- **Expansion**: Projects aimed at scaling operations to new regions or markets.
- **Innovation**: Projects with significant investments in AI and R&D for engineering advancements.
- **Optimization**: Projects focused on cost reduction and process efficiency improvements.

Applications in EDA and Power BI Visualization

The dataset has been specifically designed to facilitate hands-on practice with **EDA** and **Power BI visualization** techniques, particularly in the context of ETSEC's consulting services and AI-driven project management.

1. Exploratory Data Analysis (EDA)

- **Identify Trends**: Analyze how AI investments influence project efficiency and profitability across different project types and regions.
- **Segment Analysis**: Segment projects by **Client_Sector** (e.g., Government vs. Private Sector) to understand variations in profitability, efficiency, and AI adoption.
- **Detect Outliers**: Identify outliers in project budgets or AI investment percentages that may indicate unique challenges or opportunities.
- **Summary Statistics**: Compute mean, median, and variance for key variables like **Revenue_Generated**, **Profit_Margin**, and **Project_Duration**.

© 2025 Dr. Yasser Aly Yakout Shehata

2. Power BI Visualization

- **Dashboards**: Create dynamic dashboards to display key metrics like:
    - AI investment by project type.
    - Profit margins by client sector or region.
    - Revenue distribution across different regions and project sizes.
- **Interactive Filters**: Use slicers to filter data by **Region**, **Client_Sector**, or **Scenario** to visualize specific subsets of the dataset.
- **Geospatial Visualization**: Develop maps to showcase regional distribution of ETSEC's projects and their corresponding budgets and profitability.
- **Trend Analysis**: Visualize how AI adoption levels have influenced project efficiency improvements over time.

3. Descriptive Analytics

- **AI Adoption Impact**: Analyze how varying levels of AI adoption (Low, Moderate, High) affect profit margins and project duration.
- **Profitability Analysis**: Calculate averages and variances in profit margins across different client sectors and project types.
- **Efficiency Metrics**: Summarize how **AI_Efficiency_Improvement** correlates with project budgets and durations.

Relevance to AI for ETSEC's Business

This dataset simulates the practical application of AI in **ETSEC's consulting projects**, demonstrating how AI-driven tools and processes can impact efficiency, profitability, and client satisfaction. Key insights from this dataset can:

1. Help assess the **ROI of AI investments** in engineering and technical consultancy services.
2. Enable segmentation of projects based on **profitability, efficiency, and client sector**.
3. Support strategic decision-making for ETSEC, such as prioritizing high-profit regions or scaling AI adoption across all project types.

© 2025 Dr. Yasser Aly Yakout Shehata

By analyzing this dataset, MBA students and professionals can gain valuable experience in applying data analytics techniques to real-world scenarios in the engineering consulting industry, particularly within the context of AI-powered business solutions.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 72: Dubai Municipality

Introduction to the Dataset

The **Dubai Municipality AI Business Analytics Dataset** is a comprehensive, synthetic dataset tailored to support advanced **Exploratory Data Analysis (EDA)**, **Visualization using Power BI**, and **Descriptive Analytics** within the context of **AI applications for business**. The dataset is specifically crafted to simulate data relevant to Dubai Municipality's operational ecosystem, focusing on AI-driven innovations in environmental monitoring, licensing automation, animal tracking, smart city projects, and geo-tagging using the **Makani system**.

This dataset provides an excellent foundation for MBA students, business analysts, and researchers to practice and develop their analytical skills by exploring business performance metrics, identifying trends, and deriving actionable insights. Designed with a clear focus on Dubai Municipality's strategic priorities, this dataset integrates qualitative and quantitative variables that reflect real-world business processes, ensuring its relevance for **AI in Business**.

The dataset incorporates **key derived variables** such as profitability, revenue-to-expense ratio, and employee-to-project ratio, adding depth to the analysis. These variables help improve understanding of financial performance, workforce allocation efficiency, and operational effectiveness, making it ideal for hands-on learning and practical projects.

Dataset Description

The **Dubai Municipality AI Business Analytics Dataset** consists of **15 variables** (5 qualitative and 10 quantitative) with **1,000 records**, representing simulated data for businesses operating in Dubai's AI-driven ecosystem. The variables have been carefully designed to simulate relationships and patterns relevant to Dubai Municipality's operational areas, making the dataset suitable for data visualization, correlation analysis, and hypothesis testing.

Qualitative Variables (Categorical)

1. **Company Type:** Categorizes businesses by size and scope (Startup, SME, Enterprise, Government-affiliated).
    - *Relevance:* Allows segmentation and comparison of business types, revealing performance trends across different company scales.
2. **AI Application Domain:** Reflects the areas of AI application relevant to Dubai Municipality, including:

© 2025 Dr. Yasser Aly Yakout Shehata

- **Environmental Monitoring, Licensing Automation, Animal Tracking, Smart City Projects, and Geo-Tagging.**
- *Relevance:* Enables analysis of business performance and AI adoption in specific domains of strategic importance to Dubai Municipality.

3. **Geographical Location:** Specific Dubai areas where businesses operate:
   - **Deira, Downtown, Al Quoz, Business Bay, and Jumeirah.**
   - *Relevance:* Supports geospatial analysis and understanding of location-based business trends and challenges.
4. **Service Model:** Defines the AI service delivery model used by businesses:
   - **SaaS (Software as a Service), PaaS (Platform as a Service), IaaS (Infrastructure as a Service), and On-Premise.**
   - *Relevance:* Helps identify popular service models and their correlation with profitability and adoption rates.
5. **Partnership Status:** Indicates whether a business has an active partnership with Dubai Municipality.
   - *Relevance:* Analyzes the impact of government partnerships on business outcomes such as revenue and AI adoption.

Quantitative Variables

1. **Monthly Revenue (AED):** Simulates the monthly revenue generated by businesses.
   - *Relevance:* Serves as a key metric for profitability analysis and revenue trend visualization.
2. **Monthly Expenses (AED):** Represents operational costs, including salaries and R&D expenses.
   - *Relevance:* Enables analysis of cost structures and financial efficiency.
3. **R&D Investment (AED):** Monthly expenditure on AI-related research and development.
   - *Relevance:* A critical driver of innovation, directly influencing AI adoption rates.
4. **AI Adoption Rate (%):** The percentage of customers adopting AI-based solutions.

© 2025 Dr. Yasser Aly Yakout Shehata

- *Relevance:* Measures AI solution penetration and its relationship with investment and business type.

5. **Customer Retention Rate (%):** The percentage of repeat customers.
    - *Relevance:* Reflects customer satisfaction and loyalty, influenced by AI adoption.
6. **Employee Count:** The number of employees working in the business.
    - *Relevance:* Indicates workforce size and its impact on operational efficiency.
7. **Project Registration Count:** The number of Municipality-registered AI projects managed by the business.
    - *Relevance:* Highlights business activity levels and engagement with Dubai Municipality initiatives.
8. **Average Project Duration (Months):** The average length of time required to complete Municipality-related AI projects.
    - *Relevance:* Supports project efficiency analysis across different domains and business types.
9. **Makani Score:** A geospatial efficiency score based on the **Makani smart geo-tagging system**.
    - *Relevance:* Provides a measure of location efficiency, vital for logistics, smart city projects, and operational optimization.
10. **Environmental Impact Score:** A sustainability metric reflecting the environmental impact of the business (e.g., carbon footprint).
    - *Relevance:* Aligned with Dubai Municipality's sustainability goals, supporting environmental performance analysis.

Derived Variables

To improve the dataset's analytical depth, several derived variables have been introduced:

1. **Revenue-to-Expense Ratio:**
    - *Formula:* $\text{Monthly Revenue (AED)} / \text{Monthly Expenses (AED)}$
    - *Relevance:* Measures financial efficiency by comparing revenue generation to operational costs.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Employee-to-Project Ratio:**
    - *Formula:* Employee Count / Project Registration Count
    - *Relevance:* Reflects workforce allocation efficiency, helping identify businesses that optimize employee utilization.
3. **Profitability (%):**
    - *Formula:* (Monthly Revenue (AED) − Monthly Expenses (AED)) / Monthly Revenue (AED) ∗ 100
    - *Relevance:* Highlights financial health and helps identify profitable and struggling businesses.
4. **Profit Margin (AED):**
    - *Formula:* (Profitability (%) / 100) * Monthly Revenue (AED)
    - *Relevance:* Represents absolute profit, providing insights into financial performance in monetary terms.

Applications for EDA and Visualization in Power BI

This dataset is ideal for **Exploratory Data Analysis (EDA)** and creating insightful dashboards in **Power BI**. Analysts can use this dataset to:

1. **Analyze Business Performance:**
    - Explore revenue trends, profitability, and expense distribution across different business types and domains.
    - Compare financial efficiency metrics like **Revenue-to-Expense Ratio** and **Profit Margins**.
2. **Visualize AI Adoption Trends:**
    - Use AI Adoption Rates and Customer Retention Rates to analyze customer behavior and satisfaction.
    - Examine the impact of **R&D Investment** on AI adoption in specific domains like Licensing Automation or Smart City Projects.
3. **Geospatial Analysis:**
    - Visualize **Makani Scores** and project activity by **Geographical Location** to identify high-performing areas within Dubai.

© 2025 Dr. Yasser Aly Yakout Shehata

4. **Monitor Sustainability Goals:**
    - Assess **Environmental Impact Scores** to identify businesses aligning with Dubai Municipality's sustainability objectives.
5. **Workforce and Project Efficiency:**
    - Analyze **Employee-to-Project Ratios** and project durations to measure operational efficiency.

Conclusion

The **Dubai Municipality AI Business Analytics Dataset** is a robust tool for practicing business analytics techniques with a focus on AI applications. It aligns with Dubai Municipality's strategic goals and provides opportunities for data exploration, visualization, and hypothesis testing. The inclusion of derived variables enhances analytical depth, making the dataset highly valuable for EDA and dashboard creation in tools like **Power BI**.

This dataset serves as an essential resource for MBA students, analysts, and researchers looking to gain practical experience in analyzing and interpreting AI business data in the context of Dubai Municipality's innovative ecosystem.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 73: ASMK

Introduction

In today's dynamic business landscape, leveraging data-driven decision-making is paramount, especially for small and medium enterprises (SMEs) looking to gain a competitive edge. **ASMK Accounting & Advisory**, a premier provider of financial consulting services for **SMEs in the UAE**, recognizes the transformative potential of data analytics and artificial intelligence (AI). To equip MBA students with the skills required to address real-world business challenges, this dataset has been carefully crafted to simulate an AI-powered business environment.

The dataset captures key financial, operational, and performance metrics related to AI implementation projects, making it particularly relevant for ASMK's focus on empowering SMEs with insights into pricing strategies, cost optimization, ROI evaluation, and VAT compliance. Designed for Exploratory Data Analysis (EDA), visualization using Power BI, and descriptive analytics, this dataset enables MBA students to explore trends, identify patterns, and derive actionable insights. Moreover, the inclusion of derived variables enhances analytical depth, helping students practice advanced techniques like efficiency analysis and success prediction.

This dataset not only provides a foundation for honing technical skills but also reflects **ASMK's commitment to using AI-driven analytics** to support SMEs in achieving sustainable growth and operational excellence.

Dataset Description

The dataset consists of **1,000 records** capturing qualitative and quantitative variables that describe various aspects of AI projects relevant to **SMEs in diverse industries**. It is tailored for analytical exercises such as EDA, visualization in Power BI, and descriptive statistics, with a focus on AI applications for business operations. The dataset is specifically contextualized for **ASMK's business model**, targeting financial performance optimization and decision-making for its SME clients in the UAE.

Variables

Qualitative Variables (Categorical)

1. **Industry**: The industry to which the SME belongs (e.g., Healthcare, Retail, Finance, Manufacturing, Education).
    - *Relevance*: Analyzing AI adoption trends across industries.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **AI Application Area**: The type of AI solution implemented (e.g., Predictive Analytics, Natural Language Processing, Robotics, Computer Vision, Automation).
    - *Relevance*: Helps students analyze which AI applications lead to better business outcomes.
3. **Region**: The geographic region of the SME (e.g., North America, Europe, Middle East, Asia-Pacific, Africa).
    - *Relevance*: Facilitates regional comparisons of AI implementation costs and ROI.
4. **Customer Segment**: The target customer group (e.g., SME, Large Enterprise, Government, Startups).
    - *Relevance*: Enables segmentation analysis for targeted business strategies.
5. **Project Outcome**: The level of success achieved in the project (e.g., High Success, Moderate Success, Low Success).
    - *Relevance*: Allows students to assess factors that contribute to project success.

Quantitative Variables (Numerical)

1. **Project Duration (Months)**: The time taken to complete the AI project.
    - *Range*: 6 to 24 months.
    - *Relevance*: Correlated with budget, team size, and project success.
2. **Budget (USD)**: The financial budget allocated for the AI project.
    - *Range*: $100K to $1.5M.
    - *Relevance*: Helps evaluate the relationship between project investment and outcomes.
3. **ROI (%)**: The return on investment achieved post-project implementation.
    - *Range*: 10% to 70%.
    - *Relevance*: Measures the financial effectiveness of the project.
4. **Revenue Increase (USD)**: Additional revenue generated due to the AI project.
    - *Range*: $100K to $2M.

© 2025 Dr. Yasser Aly Yakout Shehata

- *Relevance*: Demonstrates the tangible financial impact of AI adoption.

5. **Customer Retention (%)**: Improvement in customer retention rates as a result of AI implementation.
    - *Range*: 5% to 30%.
    - *Relevance*: Tracks the customer engagement benefits of AI solutions.
6. **Data Volume (GB)**: The volume of data analyzed during the project.
    - *Range*: 50GB to 1TB.
    - *Relevance*: Highlights the role of big data in project execution.
7. **Team Size**: The number of team members involved in the project.
    - *Range*: 5 to 25.
    - *Relevance*: Indicates the human resources allocated to the project.
8. **Software Cost (USD)**: The cost of AI software tools used in the project.
    - *Range*: $10K to $150K.
    - *Relevance*: Tracks the financial impact of AI tools on project budgets.
9. **Hardware Cost (USD)**: The cost of hardware infrastructure used in the project.
    - *Range*: $10K to $100K.
    - *Relevance*: Represents the technological investment for AI projects.
10. **Time to ROI (Months)**: The time taken to achieve ROI after project implementation.
    - *Range*: 3 to 15 months.
    - *Relevance*: Helps measure the efficiency of AI projects.

Derived Variables (Enhanced Analytical Depth)

1. **Cost Efficiency Ratio**:
    - Formula: $\text{Revenue Increase (USD)} / \text{Budget (USD)}$.
    - *Relevance*: Evaluates how efficiently the project budget was utilized to generate revenue.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Success Metric**:
    - Binary variable derived from **Project Outcome** (High/Moderate Success = 1, Low Success = 0).
    - *Relevance*: Enables predictive modeling to identify factors that drive success.
3. **Region-Specific ROI**:
    - Average ROI for each region, calculated during data analysis.
    - *Relevance*: Allows benchmarking of ROI by geographic area.
4. **Project End Date**:
    - Calculated using **Project Start Date** and **Project Duration**.
    - *Relevance*: Provides time-bound insights for project management.
5. **Normalized Customer Retention**:
    - Adjusted retention rates to reflect industry-specific differences (e.g., higher retention in Healthcare).

Relevance to EDA and Visualization in Power BI

This dataset is specifically designed to facilitate:

1. **Exploratory Data Analysis (EDA)**:
    - Students can explore relationships between variables (e.g., Budget vs. ROI, Industry vs. Project Outcome) to uncover trends and patterns.
    - Example: Analyzing whether certain industries (e.g., Healthcare) see consistently higher ROI compared to others (e.g., Manufacturing).
2. **Power BI Dashboards and Visualizations**:
    - Create dynamic dashboards with slicers for segmentation by **Industry**, **Region**, and **AI Application Area**.
    - Example: Use bar charts to compare **Customer Retention (%)** across regions or industries.
3. **Descriptive Analytics**:
    - Summarize key metrics such as mean **Budget**, median **ROI (%)**, and total **Revenue Increase** to gain actionable insights.

© 2025 Dr. Yasser Aly Yakout Shehata

- Example: Identify regions with the highest average ROI for targeted market strategies.

4. **Advanced Analytical Exercises**:
    - Predictive modeling: Students can use the **Success Metric** to build predictive models for project success.
    - Cost analysis: Use the **Cost Efficiency Ratio** to determine budget efficiency across projects.

Specificity to ASMK

This dataset aligns with ASMK's strategic objectives by providing:

- Insights into financial and operational performance metrics that SMEs in the UAE can use to improve business outcomes.
- Real-world scenarios that simulate the challenges and opportunities faced by ASMK's clients when adopting AI solutions, such as balancing project budgets with ROI.
- Data designed to improve decision-making, particularly in areas of pricing, VAT compliance, and strategic planning for AI-driven business initiatives.

By tailoring the dataset to **ASMK's mission**, MBA students can practice applying data analytics in a context that directly reflects ASMK's services, including accounting, financial advisory, and business consulting for SMEs.

This introduction and description ensure the dataset is analytically robust, relevant, and tailored to ASMK's operations, providing a comprehensive foundation for academic and professional applications in AI for business.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 74: AVENTA

Introduction

The **AVENTA Dataset** has been meticulously designed to provide MBA students and professionals with a comprehensive and structured dataset relevant for **Exploratory Data Analysis (EDA)**, **Visualization in Power BI**, and **Descriptive Analytics** with a focus on **AI for Business**. Developed with practical applications in mind, this dataset simulates the operational, financial, and customer insights of **AVENTA Laboratory**, a distinguished leader in aesthetic medicine and AI-driven business innovation.

The dataset contains 1,000 records and represents anonymized and simulated data capturing customer interactions, revenue patterns, operational performance, and customer feedback. Its variables are crafted to demonstrate relationships between key business metrics and to equip learners with the opportunity to derive actionable insights. By exploring the dataset, users can identify patterns, trends, and outliers relevant to the decision-making process in the context of AI-driven aesthetic medicine businesses.

The dataset is particularly relevant for **AI for Business** applications as it contains variables reflecting customer satisfaction, service utilization, and marketing effectiveness—key performance indicators (KPIs) for any business leveraging AI technologies. Additionally, several **derived variables** have been included to deepen the analytical potential of the dataset, allowing learners to practice advanced data modeling, visualization, and storytelling techniques.

Dataset Description

The **AVENTA Dataset** includes a combination of **qualitative (categorical)** and **quantitative (numerical)** variables, representing customer demographics, service engagement, financial data, and business-specific metrics. This combination makes the dataset highly suitable for creating dashboards in tools such as **Power BI**, while also enabling learners to conduct statistical analysis and develop machine learning models for predictive analytics.

Below is a detailed description of the variables included in the dataset, along with derived variables for deeper analysis:

Qualitative Variables (Categorical):

1. **Customer_Segment**: The customer category, reflecting the scale of their business:
    - Values: "Individual", "Small Business", "Enterprise".
    - Purpose: Helps analyze differences in service usage, satisfaction, and revenue across segments.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Region**: The geographic location of the customer:
    - Values: "North America", "Europe", "Asia-Pacific", "Middle East", "South America".
    - Purpose: Enables regional analysis of revenue patterns, service adoption rates, and customer preferences.
3. **Subscription_Type**: The subscription plan chosen by the customer:
    - Values: "Basic", "Standard", "Premium".
    - Purpose: Provides insights into how different plans correlate with revenue, adoption rates, and satisfaction.
4. **Industry**: The industry to which the customer belongs:
    - Values: "Healthcare", "Finance", "Retail", "Education", "Technology".
    - Purpose: Identifies industry-specific trends in AI adoption, marketing effectiveness, and revenue.
5. **Customer_Satisfaction**: The overall satisfaction rating given by the customer:
    - Values: "Very Dissatisfied", "Dissatisfied", "Neutral", "Satisfied", "Very Satisfied".
    - Purpose: Analyzes how satisfaction influences churn probability, referrals, and Net Promoter Score.

Quantitative Variables (Numerical):

1. **Monthly_Revenue**: The monthly revenue generated from each customer (in USD).
    - Range: $500–$50,000.
    - Purpose: Serves as a key business metric to analyze performance across segments and regions.
2. **AI_Service_Usage_Hours**: The number of hours customers use AVENTA's AI-powered services per month.
    - Range: 10–500 hours.
    - Purpose: Highlights customer engagement and efficiency of AI service offerings.
3. **Churn_Probability**: The predicted likelihood of a customer discontinuing services.

© 2025 Dr. Yasser Aly Yakout Shehata

- Range: 0.0–1.0.
- Purpose: Facilitates churn analysis and supports retention strategies.

4. **Customer_Age**: The age of the customer account (in months).
    - Range: 1–120 months.
    - Purpose: Tracks the lifecycle of customer relationships.
5. **AI_Adoption_Rate**: The percentage of available AI services utilized by the customer.
    - Range: 10%–100%.
    - Purpose: Measures how effectively customers are adopting AI services.
6. **Customer_Referrals**: The number of referrals generated by the customer.
    - Range: 0–50.
    - Purpose: Indicates customer loyalty and satisfaction, linking it to the success of referral programs.
7. **Monthly_Support_Tickets**: The number of support tickets raised by the customer each month.
    - Range: 0–20.
    - Purpose: Helps evaluate operational efficiency and customer satisfaction.
8. **Training_Sessions_Completed**: The number of AI-related training sessions completed by the customer.
    - Range: 0–15.
    - Purpose: Highlights the role of education in improving AI adoption rates.
9. **Marketing_Expense**: The monthly marketing expense associated with acquiring or retaining the customer (in USD).
    - Range: $1,000–$20,000.
    - Purpose: Supports ROI analysis on marketing campaigns.
10. **Net_Promoter_Score**: The Net Promoter Score (NPS) reflecting customer loyalty.

- Range: -100 to 100.
- Purpose: Measures customer loyalty and satisfaction.

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Variables (For Deeper Insights):

1. **Revenue_Per_Hour**:
    - Formula: $\text{Monthly_Revenue} / \text{AI_Service_Usage_Hours}$.
    - Purpose: Evaluates revenue efficiency for each customer, useful for benchmarking.
2. **Customer_Lifetime_Value (CLV)**:
    - Formula: $\text{Monthly_Revenue} \times \text{Customer_Age}$.
    - Purpose: Estimates the total monetary value contributed by a customer over their lifetime.
3. **Support_Tickets_Per_Hour**:
    - Formula: $\text{Monthly_Support_Tickets} / \text{AI_Service_Usage_Hours}$.
    - Purpose: Tracks operational efficiency and support demands relative to service usage.

Why This Dataset is Relevant for EDA, Visualization, and AI for Business

The **AVENTA Dataset** is designed to foster a holistic understanding of how data-driven decisions can optimize AI-powered business processes. Below are specific use cases:

1. **Exploratory Data Analysis (EDA):**
    - Examine revenue trends across regions, industries, and subscription types.
    - Identify key factors driving customer churn and satisfaction.
    - Detect outliers and understand their impact on overall business performance.
2. **Visualization in Power BI:**
    - Create dashboards to monitor KPIs like monthly revenue, churn probability, and customer lifetime value.
    - Visualize satisfaction levels across customer segments or industries.
    - Highlight correlations, such as the relationship between AI adoption rates and revenue.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Descriptive Analytics for AI Business:**
    - Analyze the impact of marketing expenses on referrals and Net Promoter Scores.
    - Identify the most profitable customer segments using metrics like CLV and Revenue_Per_Hour.
    - Assess the effectiveness of training programs in boosting AI adoption.

Potential Applications of the Dataset:

1. **Customer Segmentation and Targeting**:
    - Use EDA to classify customers into actionable segments based on revenue, satisfaction, and churn probability.
2. **Churn Prediction and Retention**:
    - Evaluate the factors influencing churn and design targeted strategies to retain customers.
3. **AI Adoption Analysis**:
    - Assess how different industries and customer segments are utilizing AI, providing insights into where AVENTA can improve services or tailor training programs.
4. **Revenue Optimization**:
    - Analyze revenue per service hour to identify high-performing customer profiles and focus marketing efforts accordingly.

The **AVENTA Dataset** is a powerful tool for developing skills in **EDA**, **data visualization**, and **descriptive analytics**, all while providing valuable insights into the role of AI in driving business success. Its variables and derived features offer a real-world simulation of decision-making processes for an AI-driven company in the aesthetic medicine space.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 75: Media Heist

Introduction

The **Media Heist Dataset** is a carefully crafted dataset designed to provide MBA students, data analysts, and business professionals with a robust foundation for exploring advanced analytical techniques, including **Exploratory Data Analysis (EDA)** and **Data Visualization**. This dataset is tailored to reflect the operational, financial, and marketing dynamics of AI-driven businesses, specifically within the context of **Media Heist**, a digital marketing agency headquartered in Dubai with a global footprint.

With the rapid proliferation of AI technologies in business environments, datasets like this serve as an essential tool for understanding how data analytics can optimize operations, enhance decision-making, and improve customer engagement strategies. By integrating qualitative and quantitative variables that simulate real-world business scenarios, the dataset enables learners to gain practical insights into business analytics, data visualization, and predictive modeling.

The **Media Heist Dataset** is particularly relevant for **EDA and Visualization in Power BI**, as it contains features that allow users to create interactive dashboards, develop storytelling visuals, and derive actionable insights. Additionally, derived variables included in this dataset enhance the analytical depth, enabling learners to identify meaningful patterns, correlations, and business opportunities.

Description of the Dataset

The **Media Heist Dataset** contains **1,000 records** and **19 variables** (both qualitative and quantitative) that encapsulate the key performance indicators (KPIs) of an AI-focused digital marketing agency. The dataset is designed to simulate the real-world challenges and opportunities faced by Media Heist, focusing on their branding, customer engagement, marketing campaigns, and business growth. The inclusion of both raw and derived variables allows users to practice foundational and advanced analytics techniques, ensuring a comprehensive learning experience.

Key Features of the Dataset

1. **Qualitative Variables**These variables define the categorical aspects of Media Heist's operations, enabling segmentation and comparison across different business dimensions:

© 2025 Dr. Yasser Aly Yakout Shehata

- **Business_Type**: Represents the type of business service offered (e.g., "Digital Marketing," "AI Consultancy," "FinTech AI," "Healthcare AI"). This variable reflects Media Heist's service diversity.
- **Region**: Indicates the geographical location of business operations (e.g., "Middle East," "North America," "Europe," "Asia"), enabling geographical performance analysis.
- **Customer_Segment**: Specifies the primary customer group served (e.g., "SME," "Corporate," "Individual"), which helps in understanding customer dynamics.
- **Brand_Reputation**: A qualitative indicator of the brand's reputation (e.g., "High," "Medium," "Low"), influencing retention rates and customer lifetime value.
- **Technology_Adoption_Level**: Indicates the adoption stage of AI technologies by the business (e.g., "Pioneer," "Early Adopter," "Late Adopter"), reflecting the business's innovation level.

2. **Quantitative Variables**These variables capture measurable business metrics that form the backbone of EDA and visualization:
   - **Revenue (USD)**: Annual revenue generated by Media Heist.
   - **Marketing_Expense (USD)**: Total expenditure on marketing campaigns.
   - **Employee_Count**: The total number of employees in the organization.
   - **Customer_Count**: The number of customers served during the year.
   - **Growth_Rate (%)**: Year-over-year growth in revenue, reflecting business expansion.
   - **Retention_Rate (%)**: Percentage of customers retained, providing insight into customer loyalty.
   - **SEO_Rank**: Average search engine rank for primary keywords (lower values indicate better performance).
   - **Ad_Conversion_Rate (%)**: Conversion rate from paid advertisements, critical for understanding campaign effectiveness.
   - **Customer_Lifetime_Value (USD)**: Estimated lifetime value of a customer, important for evaluating customer profitability.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Social_Media_Engagement (%)**: Percentage of users engaged on Media Heist's social media platforms.

3. **Derived Variables**To improve analytical depth, several new variables were derived from the existing ones. These variables allow for more advanced analytics and business insights:
    - **Profit (USD)**: Calculated as the difference between Revenue and Marketing_Expense. It provides a direct measure of the financial success of Media Heist's campaigns.
        - Formula: $\text{Profit} = \text{Revenue} - \text{Marketing Expense}$
    - **Revenue per Employee (USD)**: Evaluates operational efficiency by measuring revenue generated per employee.
        - Formula: $\text{Revenue per Employee} = \frac{\text{Revenue}}{\text{Employee Count}}$
    - **Customer Acquisition Cost (CAC)**: Measures marketing efficiency by calculating the cost of acquiring one customer.
        - Formula: $\text{CAC} = \frac{\text{Marketing Expense}}{\text{Customer Count}}$
    - **Retention Value (USD)**: Estimates the monetary value of customers retained annually.
        - Formula: $\text{Retention Value} = \text{Retention Rate} \times \text{Customer Lifetime Value}$

Application of the Dataset in EDA and Power BI

This dataset is specifically structured to enable learners to perform:

1. **Exploratory Data Analysis (EDA)**:
    - Identify trends and patterns in marketing expenditure, revenue, and customer retention.
    - Analyze relationships between variables, such as how brand reputation impacts retention rate or how SEO rank correlates with ad conversion rates.
    - Use statistical summaries and visualizations to generate actionable insights.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Visualization in Power BI**:
    - **Interactive Dashboards**: Create dynamic dashboards to track KPIs like revenue, profit, growth rate, and customer lifetime value.
    - **Geographical Analysis**: Visualize regional performance and identify markets where Media Heist excels or underperforms.
    - **Segmentation Analysis**: Compare customer segments (SME, Corporate, Individual) across various metrics.
    - **Performance Monitoring**: Use time-series charts to track growth rates and retention rates over time.
3. **Business Decision-Making**:
    - Perform descriptive analytics to determine the profitability of customer segments and regions.
    - Evaluate the efficiency of marketing campaigns by analyzing CAC and ad conversion rates.
    - Understand the financial impact of customer retention and lifetime value, aiding in strategic decision-making.

Relevance to AI for Business

The **Media Heist Dataset** integrates AI-related variables such as SEO rank, ad conversion rate, and social media engagement to simulate the role of AI in optimizing business operations. It provides students with a sandbox to explore how AI-driven strategies (e.g., improving ad targeting or enhancing customer retention) can directly influence business KPIs. This relevance to AI for business makes the dataset highly practical for MBA students studying **Data Analytics for AI Businesses**.

Conclusion

The **Media Heist Dataset** offers a comprehensive platform for MBA students and analysts to hone their data analytics skills. Its realistic variables, meaningful relationships, and derived metrics ensure that it provides an authentic experience for EDA, visualization in Power BI, and descriptive analytics. By using this dataset, learners can simulate real-world decision-making processes, evaluate AI's role in business optimization, and extract insights that are vital for success in the modern, data-driven business landscape.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 76: Rostom Medical

Introduction

In the contemporary healthcare landscape, data-driven decision-making has become a cornerstone for improving operational efficiency, customer satisfaction, and overall organizational success. The dataset developed for this study is specifically tailored for **Rustom Medical**, a leading provider of innovative and technologically advanced diagnostic solutions headquartered in **Riyadh, Saudi Arabia**. This dataset is meticulously designed to aid MBA students specializing in Business Analytics with a focus on Artificial Intelligence (AI) applications in the healthcare business.

The dataset serves as a foundational resource for practicing essential analytics techniques such as Exploratory Data Analysis (EDA), Visualization, and Descriptive Analytics. It captures key business operations and performance indicators at **Rustom Medical**, including customer satisfaction, diagnostic efficiency, training program participation, and revenue generation. Furthermore, the inclusion of derived variables enhances analytical depth and facilitates meaningful insights into operational dynamics and client relationships.

Given its structure and content, the dataset is highly compatible with visualization tools such as Power BI, enabling students to generate actionable insights and simulate real-world business scenarios in the context of AI-driven healthcare solutions.

Description of the Dataset

The dataset consists of **1,000 records** and **15 variables**, designed to reflect Rustom Medical's operations, client engagements, and business outcomes. It is a blend of qualitative and quantitative data points that allow users to investigate trends, patterns, and relationships critical to understanding the company's performance. Below is a detailed description of the dataset's variables:

Qualitative Variables (5):

1. **Region**: Represents the geographic region where the client is based (e.g., Central, Western, Eastern, Northern, Southern). This variable allows segmentation and regional performance analysis.
2. **Product Segment**: Indicates the type of diagnostic solutions provided (e.g., Molecular Diagnostics, Personalized Medicine, Immunodiagnostics). It facilitates analysis of product performance and customer preferences.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Client Type**: Defines the category of the client (e.g., Hospital, Laboratory, Clinic, Pharmacy, Government Institution). This variable is essential for identifying patterns in client behavior and revenue contributions.
4. **Training Program Participation**: Specifies whether the client participated in Rustom Medical's continuous medical education and training programs (Yes/No). It is a critical factor in evaluating the impact of training on outcomes.
5. **Customer Satisfaction Rating**: Reflects the client's satisfaction level with Rustom Medical's services (e.g., Very Satisfied, Satisfied, Neutral, Dissatisfied, Very Dissatisfied). This variable directly influences retention and overall business success.

Quantitative Variables (10):

1. **Revenue (USD)**: The total revenue generated from each client, reflecting the monetary value of business transactions. It ranges from $5,000 to $100,000.
2. **Number of Products Purchased**: The total number of diagnostic products purchased by the client. Higher values indicate stronger client engagement and purchasing power.
3. **Equipment Supplied (Units)**: The number of diagnostic equipment units supplied to the client. This variable is correlated with revenue and diagnostic efficiency.
4. **Technical Support Hours Provided**: The total number of hours spent on providing technical support to the client. It reflects the level of client engagement and post-sales service quality.
5. **Marketing Services Provided (Hours)**: The time allocated to marketing and promotional activities for the client. It highlights efforts made to acquire or retain customers.
6. **Number of Training Sessions Attended**: The number of training sessions attended by the client's representatives. This variable is pivotal for understanding the influence of training on client satisfaction and retention.
7. **Diagnostic Tests Performed**: The total number of diagnostic tests conducted using Rustom Medical's products. It serves as a proxy for operational efficiency and product utilization.
8. **Patient Outcomes Improved (%)**: The percentage improvement in patient outcomes attributed to Rustom Medical's diagnostic solutions. It demonstrates the effectiveness of the company's offerings in enhancing public health.

© 2025 Dr. Yasser Aly Yakout Shehata

9. **Customer Retention Rate (%)**: The percentage of clients retained over a given period. This variable reflects the company's ability to sustain long-term relationships.
10. **Average Delivery Time (Days)**: The average number of days taken to deliver products or solutions to the client. It is inversely related to customer satisfaction.

Derived Variables for Analytical Depth (3):

1. **Revenue per Product (USD)**: Calculated as Revenue / Number of Products Purchased. This variable provides insights into pricing efficiency and profitability at a product level.
2. **Delivery Speed**: Categorized as "On Time" if the **Average Delivery Time** is less than or equal to 3 days, and "Delayed" otherwise. This variable helps evaluate the company's operational performance and its influence on customer satisfaction.
3. **Training Influence Metric**: Categorized as "Significant Impact" if **Patient Outcomes Improved (%)** exceeds 90%, and "Moderate/Low Impact" otherwise. It measures the effectiveness of training programs in improving health outcomes.

Relevance to EDA, Visualization, and AI for Business

This dataset is meticulously crafted to enable MBA students to explore a range of analytical techniques using EDA and visualization tools such as Power BI. The dataset's design aligns with the following learning objectives:

1. **Exploratory Data Analysis (EDA):**
    - Students can analyze variable distributions, detect outliers, and assess inter-variable relationships.
    - For instance, correlations between **Revenue**, **Number of Products Purchased**, and **Technical Support Hours Provided** can be explored to identify key revenue drivers.
2. **Visualization in Power BI:**
    - Students can create interactive dashboards to visualize performance metrics. Examples include:
        - Regional heatmaps to analyze revenue distribution by **Region**.
        - Bar charts comparing **Revenue per Product** across **Product Segments**.

© 2025 Dr. Yasser Aly Yakout Shehata

- Scatter plots showing the relationship between **Customer Satisfaction Rating** and **Average Delivery Time**.

3. **AI for Business Applications:**
    - The dataset serves as a basis for developing AI-powered predictive models, such as:
        - Predicting **Customer Retention Rate (%)** based on factors like **Revenue**, **Satisfaction Rating**, and **Training Program Participation**.
        - Identifying patterns in **Patient Outcomes Improved (%)** using machine learning models that analyze training participation and technical support.
4. **Descriptive Analytics:**
    - Students can derive insights from key metrics such as **Revenue per Product** and **Training Influence Metric**, which enhance decision-making and operational efficiency in AI-driven healthcare businesses.

Conclusion

This dataset is a valuable asset for MBA students specializing in Business Analytics, particularly in the context of AI for Business applications. By simulating real-world data from **Rustom Medical's operations**, the dataset provides an ideal platform for students to hone their analytical skills, develop data visualization capabilities, and derive actionable insights. The inclusion of derived variables further enhances its analytical depth, making it an indispensable resource for EDA, visualization, and business intelligence projects.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 77: SPIMACO

Introduction

This dataset, named **SPIMACO Business Analytics Dataset**, has been meticulously crafted to support data-driven decision-making in alignment with the strategic operations of the **Saudi Pharmaceutical Industries & Medical Appliances Corporation (SPIMACO)**. As a leader in the pharmaceutical and medical appliances industry across the MENA region and beyond, SPIMACO's operational complexity and performance metrics provide a rich foundation for business analysis.

The dataset is specifically designed for MBA students focusing on **Business Analytics for AI Applications**, enabling them to explore advanced analytics techniques, conduct **Exploratory Data Analysis (EDA)**, and create impactful **visualizations in Power BI**. The dataset encompasses a wide range of qualitative and quantitative variables related to SPIMACO's core business functions, such as manufacturing, sales, R&D investment, and customer satisfaction. Additionally, derived variables have been included to improve analytical depth and provide actionable insights for real-world business scenarios.

Through this dataset, students will gain hands-on experience with descriptive analytics, data modeling, and visualization techniques, while also learning how AI-powered analytics can be applied to drive better decision-making in industries like pharmaceuticals and medical appliances.

Description of the Dataset

The **SPIMACO Business Analytics Dataset** contains **1000 records** and **20 variables** (including derived variables), carefully designed to represent SPIMACO's operational, financial, and customer data. The dataset simulates real-world business scenarios, providing the foundation for MBA students to practice advanced analytics techniques, discover patterns, and interpret business insights effectively.

The variables in the dataset are specifically selected to reflect SPIMACO's **pharmaceutical and biopharmaceutical operations**, with meaningful relationships that can be explored using **Power BI** and **AI-driven analytics tools**. The dataset is ideal for identifying trends, benchmarking performance, and developing AI-driven strategies.

Structure of the Dataset

1. Qualitative Variables (Categorical Data)

© 2025 Dr. Yasser Aly Yakout Shehata

1. **Region**: Geographical region of operations where SPIMACO markets its products (e.g., "MENA," "GCC," "EU," "Asia"). Regions represent diverse customer bases and market conditions.
2. **Client Type**: The type of client purchasing SPIMACO's products (e.g., "Government," "Private," "Distributor"). Client types influence revenue streams and customer satisfaction.
3. **Product Category**: Categorizes SPIMACO's offerings into three primary segments (e.g., "Pharmaceutical," "Medical Appliance," "Biopharmaceutical"). Different categories align with unique market demands and profit margins.
4. **Strategic Partner Status**: Identifies whether the product benefits from a strategic partnership (e.g., "Global Partnership," "Regional Partnership," "No Partnership"). Partnerships impact revenue, market share, and R&D intensity.
5. **Production Facility**: Indicates the manufacturing location (e.g., "Al-Qassim," "Dammam"). Production facility affects manufacturing cost, delivery timeliness, and operational efficiency.

2. Quantitative Variables (Numerical Data)

1. **Revenue (SAR)**: Total revenue generated for the product in Saudi Riyals (SAR). This is a key performance indicator for evaluating SPIMACO's financial success.
2. **Net Profit/Loss (SAR)**: Profit or loss incurred from product sales, calculated as revenue minus total costs (manufacturing and R&D). Negative values indicate losses.
3. **Units Sold**: Total number of units sold for a given product, reflecting demand in the market.
4. **Manufacturing Cost (SAR)**: Total cost incurred in producing the product, representing operational expenses.
5. **R&D Investment (SAR)**: The investment allocated to research and development, critical for developing innovative biopharmaceuticals and maintaining product quality.
6. **Digital Transformation Score**: A score (0–100) reflecting the extent to which digital transformation initiatives have impacted the product's operations. Higher scores correlate with better operational efficiency and market outcomes.
7. **Market Share (%)**: The percentage of the total market captured by the product, indicating SPIMACO's competitive position.

© 2025 Dr. Yasser Aly Yakout Shehata

8. **Customer Satisfaction Score**: An average score (1–10) provided by clients, reflecting the perceived quality of SPIMACO's products and services.
9. **Delivery Timeliness (%)**: Percentage of orders delivered on time, measuring SPIMACO's supply chain efficiency.
10. **Staff Hours Allocated**: Total staff hours devoted to the development and production of the product, indicating workforce effort.

3. Derived Variables (Calculated Fields)

To enhance the dataset's analytical depth, several **derived variables** have been included:

1. **Profit Margin (%)**:
   - Calculated as: $\text{Profit Margin} = (\frac{\text{Net Profit/Loss}}{\text{Revnue}} \times 100)$
   - This variable helps evaluate profitability efficiency.
2. **Revenue per Unit Sold**:
   - Calculated as: $\text{Revenue per Unit Sold} = \frac{\text{Revenue}}{\text{Units Sold}}$
   - This metric provides insights into pricing strategy and product value.
3. **R&D Intensity (%)**:
   - Calculated as: $\text{R\&D Intensity (\%)} = (\text{Amount Spent on R\&D} / \text{Total Revenue}) \times 100$
   - This variable measures SPIMACO's focus on innovation relative to revenue generation.
4. **Cost per Unit Sold**:
   - Calculated as: $\text{Cost per Unit Sold} = \frac{\text{Manufacturing Cost}}{\text{Unit Sold}}$
   - This field reflects cost efficiency in manufacturing.
5. **Delivery Delay Risk (%)**:
   - Calculated as: $\text{Delivery Delay Risk} = 100 - \text{Delivery Timeliness}$
   - This variable indicates the risk of delayed orders, which impacts customer satisfaction.

© 2025 Dr. Yasser Aly Yakout Shehata

Relevance for EDA and Power BI Visualization

The dataset is crafted to facilitate **Exploratory Data Analysis (EDA)** and **Visualization in Power BI**, with numerous analytical opportunities, including:

1. **EDA Techniques**:
    - Univariate Analysis: Summarize and visualize single variables like **Revenue**, **Profit Margin**, or **Customer Satisfaction**.
    - Bivariate Analysis: Explore relationships between variables such as **Revenue vs. Units Sold** or **Profit Margin vs. R&D Intensity**.
    - Multivariate Analysis: Identify trends by analyzing multiple variables simultaneously, e.g., how **Market Share** varies with **Digital Transformation Score** across regions.
2. **Visualization in Power BI**:
    - Create **bar charts** and **scatter plots** to analyze product performance across regions and client types.
    - Build **line graphs** to compare trends in profitability, R&D spending, and revenue.
    - Develop **heatmaps** to show correlations between variables such as **Customer Satisfaction**, **Delivery Timeliness**, and **Revenue**.
3. **AI for Business Applications**:
    - The dataset enables **predictive modeling** and **machine learning applications**, such as predicting **Market Share** based on **R&D Intensity** or **Revenue**.
    - Students can simulate scenarios where **AI-driven recommendations** optimize supply chain efficiency (e.g., minimizing **Delivery Delay Risk**) or increase profitability.

Conclusion

The **SPIMACO Business Analytics Dataset** is a highly relevant and realistic resource for MBA students to practice **EDA**, **data visualization**, and **descriptive analytics** using Power BI. By analyzing both core and derived variables, students will gain a comprehensive understanding of SPIMACO's operations and strategic challenges. The dataset provides an excellent opportunity to explore the intersection of business analytics and AI applications,

© 2025 Dr. Yasser Aly Yakout Shehata

ultimately equipping future business leaders with data-driven decision-making skills tailored to industries like pharmaceuticals and medical appliances.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 78: PHARCO

Introduction

The **PHARCO Analytical Dataset** has been meticulously designed to provide MBA students and business analysts with a comprehensive and realistic dataset for hands-on experience in **Exploratory Data Analysis (EDA)**, **Visualization using Power BI**, and applications of **AI-driven analytics for business decision-making**. Rooted in the context of **Pharco Pharmaceuticals**, one of the leading pharmaceutical companies in Egypt, the Middle East, and Africa, this dataset simulates critical business processes, key performance metrics, and operational relationships, offering an ideal foundation for academic and professional learning.

The dataset encapsulates Pharco's core business domains, such as **product development, production, marketing, distribution, and exportation**, while also highlighting their strategic focus on enhancing **patient quality of life** through safe and affordable pharmaceutical solutions. With this dataset, analysts can explore and visualize trends, relationships, and key insights that are directly applicable to real-world challenges faced by the pharmaceutical industry.

In particular, the inclusion of derived variables ensures deeper analytical depth, enabling advanced techniques for decision-making and AI-driven optimization in a business context.

Dataset Description

The **PHARCO Analytical Dataset** contains 1,000 records and is structured with **15 variables** that include both qualitative (categorical) and quantitative (numerical) attributes. These variables represent Pharco's operational, financial, and strategic metrics, allowing for robust analysis across key business areas. Furthermore, **derived variables** are included to enhance the dataset's analytical capabilities and enable advanced insights into Pharco's performance. The dataset is designed to emulate **real-world business scenarios** and is relevant to tasks such as KPI monitoring, market performance analysis, operational efficiency evaluation, and profitability modeling.

Qualitative Variables:

1. **Product_Category**: The therapeutic area of the product (e.g., Dermatology, Antipyretics, Nutritional Products, Chronic Diseases, Urology).
2. **Region**: The geographic region where the product is marketed/distributed (e.g., Middle East, Africa, Europe, Asia, North America).

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Brand_Type**: The branding type of the product (e.g., Generic, Branded Generic, Licensed Product).
4. **Department**: The department associated with the product (e.g., Manufacturing, Marketing, Research & Development, Distribution).
5. **R&D_Collaboration**: Whether the product was developed through an R&D collaboration (e.g., Yes, No).

Quantitative Variables:

1. **Daily_Production_Units**: The number of units produced daily (range: 10,000–500,000).
2. **Market_Share_Percentage**: The market share of the product category in the target region (range: 1–30%).
3. **Price_Per_Unit**: The price per unit of the product in USD (range: 0.50–20.00).
4. **Employee_Count**: The number of employees associated with the product's production/marketing (range: 20–300).
5. **Export_Countries**: The number of countries where the product is exported (range: 1–50).
6. **Revenue_Per_Month**: Monthly revenue generated by the product in USD (range: $10,000–$500,000).
7. **R&D_Investment_Per_Year**: Annual investment in R&D for the product in USD (range: $5,000–$200,000).
8. **Patient_Count_Treated**: The estimated number of patients treated annually by the product (range: 1,000–1,000,000).
9. **HepatitisC_Impact**: The number of Hepatitis C patients treated by relevant products (specific to certain categories, range: 0–200,000).
10. **Product_Age_Years**: The number of years the product has been in the market (range: 1–20 years).

Derived Variables (to improve analytical depth):

1. **Revenue_Per_Unit**: A derived metric calculated as Revenue_Per_Month / Daily_Production_Units, indicating the revenue generated per unit produced.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Profit_Per_Unit**: A profitability metric based on the assumption that 70% of the product's price represents production costs, calculated as Price_Per_Unit * 0.3.
3. **Export_Ratio**: A measure of export performance, calculated as Export_Countries / 50, where 50 is the total number of exportable countries.
4. **Region_Economic_Index**: An index assigned to regions to represent economic influence, with higher values for developed regions (e.g., North America = 1.2, Africa = 0.8, etc.).
5. **High_Revenue_Product**: A binary variable indicating whether the product generates high revenue (1 if Revenue_Per_Month > 250,000, otherwise 0).

Relevance to EDA, Power BI Visualization, and AI for Business

The **PHARCO Analytical Dataset** is particularly relevant to MBA students and business professionals seeking to develop skills in **data-driven decision-making**, with a focus on pharmaceutical industry challenges. Below are the specific ways in which the dataset aligns with EDA, visualization, and AI applications:

1. Exploratory Data Analysis (EDA)

- **Uncover Trends and Relationships**: Analyze how variables such as Daily_Production_Units, Market_Share_Percentage, and Revenue_Per_Month are interrelated.
- **Identify Outliers and Anomalies**: Use statistical techniques to detect and interpret deviations in production, revenue, and R&D investment.
- **Segmentation Analysis**: Examine patterns across product categories or geographic regions to understand market dynamics.

2. Visualization Using Power BI

- **Dashboards for Decision-Making**: Build interactive dashboards that visualize key insights, such as:
    - Revenue distribution across regions or product categories.
    - Export performance using the Export_Ratio.
    - Profitability analysis with Profit_Per_Unit and Revenue_Per_Unit.
- **KPI Monitoring**: Highlight high-revenue products, market share leaders, and patient impact using custom visualizations.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Dynamic Filtering**: Enable real-time filtering of data by region, product category, or R&D collaboration.

3. Application of AI for Business

- **Predictive Analytics**: Use the dataset to predict high-revenue products (High_Revenue_Product) using AI/ML models.
- **Optimization Models**: Leverage derived variables (e.g., Profit_Per_Unit, Export_Ratio) to optimize pricing, production, and export strategies.
- **Churn Analysis**: Examine which products may require higher investment or marketing focus based on their market share and revenue trends.

Significance for Pharco Pharmaceuticals

This dataset is specifically crafted to align with **Pharco Pharmaceuticals' mission** of enhancing patient quality of life through innovative and affordable products. By integrating Pharco-specific business processes and metrics into the dataset, it enables analysts to:

- Simulate decision-making scenarios faced by Pharco's leadership.
- Identify strategies to optimize Pharco's product portfolio and regional expansion efforts.
- Analyze the company's success in treating chronic diseases, including Hepatitis C, and explore opportunities for broader patient impact.

The inclusion of advanced variables such as Profit_Per_Unit and Region_Economic_Index equips users with insights that go beyond descriptive analytics, supporting strategic decision-making powered by AI and data visualization.

This dataset serves as an essential tool for developing the next generation of data-savvy business leaders, preparing them to apply analytics and AI-driven solutions in real-world business contexts, with **Pharco Pharmaceuticals** as an inspiring case study.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 79: Planet Pharmacies

Introduction

The **Planet Pharmacies Dataset** is a meticulously crafted dataset designed to provide MBA students with a practical foundation for mastering analytical techniques in Exploratory Data Analysis (EDA), visualization, and descriptive analytics using Power BI. This dataset simulates real-world business scenarios relevant to Planet Pharmacies, a leading provider of pharmaceutical products and AI-driven health solutions in the GCC region. It serves as a comprehensive resource for understanding the intricate dynamics of retail, corporate, and enterprise customer segments while emphasizing the application of Artificial Intelligence (AI) in business decision-making.

The dataset captures the core operational and financial aspects of **Planet Pharmacies**, including sales performance, customer satisfaction, product delivery, and marketing effectiveness. By incorporating derived variables that enhance analytical depth, students can uncover actionable insights, simulate real-world business challenges, and make data-driven decisions in the context of AI-driven business strategies.

Dataset Description

The **Planet Pharmacies Dataset** consists of 1,000 records and 15 variables, divided into **5 qualitative variables** and **10 quantitative variables**. It is tailored to reflect the core business operations of Planet Pharmacies in the pharmaceutical and AI health solutions sectors, enabling students to practice various analytical techniques and tools effectively.

Qualitative Variables (Categorical)

1. **Customer Segment**: Categorizes customers into four types: Retail, Corporate, Small Business, and Enterprise.
2. **Region**: Represents the geographic market, including UAE, Saudi Arabia, Oman, Other GCC countries, and Non-GCC regions.
3. **Product Category**: Highlights key product lines: Pharmaceuticals, Medical Devices, AI Health Solutions, and Supplements.
4. **Purchase Channel**: Identifies the method of purchase: Online, In-Store, or Distribution Partner.
5. **Customer Satisfaction Level**: Assesses satisfaction on a five-point scale: Very Satisfied, Satisfied, Neutral, Dissatisfied, and Very Dissatisfied.

© 2025 Dr. Yasser Aly Yakout Shehata

Quantitative Variables (Numerical)

1. **Annual Revenue**: The total revenue generated by each customer in USD.
2. **Number of Transactions**: The number of purchases made by the customer in a given year.
3. **Average Order Value**: The average monetary value of a single order in USD.
4. **Product Quantity**: The total number of products purchased by the customer.
5. **Discount Percentage**: The percentage of discounts applied to the total orders.
6. **Delivery Time**: The time taken to deliver the order, measured in days.
7. **Customer Lifetime Value (CLV)**: The predicted net revenue generated by the customer over their lifetime.
8. **Cost per Acquisition (CPA)**: The marketing cost incurred to acquire the customer.
9. **Return Rate**: The percentage of purchased products returned by the customer.
10. **Marketing Spend**: The total marketing expenditure associated with acquiring and retaining the customer.

Derived Variables

1. **Profit Margin (%)**: A calculated field that determines the percentage of profit after deducting marketing spend from revenue:

$$\text{Profit Margin (\%)} = \left(\frac{Annual\ Revenue\ -\ Marketing\ Spend}{Annual\ Revnue}\right) \times 100$$

2. **Average Discount (USD)**: Represents the monetary discount provided per order: $\text{Average Discount (USD)} = \left(\frac{\text{Discount \%}}{100}\right) \times \text{Average Order Value}$

Relevance to EDA and Visualization in Power BI

This dataset is specifically structured to support in-depth **Exploratory Data Analysis (EDA)** and visualization using **Power BI**. The diverse variables and interdependencies provide ample opportunities to apply descriptive and diagnostic analytics, enabling students to:

1. **Analyze Customer Behavior**:
    - Segment customers by region, product category, and satisfaction levels to understand trends and preferences.

© 2025 Dr. Yasser Aly Yakout Shehata

- Evaluate the impact of discounts and marketing spend on revenue and customer retention.

2. **Visualize Key Metrics**:
    - Use Power BI to create dashboards for metrics such as annual revenue, profit margin, customer lifetime value, and delivery performance.
    - Highlight regional disparities in sales and marketing effectiveness using heatmaps and geospatial visualizations.
3. **Simulate AI Business Applications**:
    - Predict high-value customers by analyzing customer lifetime value and satisfaction levels.
    - Identify areas for operational improvement, such as reducing delivery times or optimizing discounts to minimize return rates.

Application in AI for Business

The dataset facilitates AI-driven insights by simulating scenarios where machine learning and predictive analytics could be applied. For instance:

1. **Customer Segmentation**: Utilize clustering algorithms to classify customers into distinct segments for targeted marketing campaigns.
2. **Predictive Modeling**: Use regression techniques to forecast future revenue based on marketing spend, product category, and customer satisfaction.
3. **Optimization Scenarios**: Apply decision-tree analysis to identify optimal discount percentages and marketing strategies to maximize profit margins.
4. **Churn Prediction**: Analyze relationships between customer satisfaction, return rates, and transactions to predict churn and implement retention strategies.

The **Planet Pharmacies Dataset** is an invaluable resource for MBA students exploring the intersection of AI, business analytics, and strategic decision-making. Its comprehensive structure and innovative design provide a strong foundation for practical learning and professional growth.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 80: Asharq News

Introduction

The **Asharq News Dataset** has been meticulously crafted to support MBA students and business analysts in exploring the intersection of **Artificial Intelligence (AI)** and business performance. This dataset is specifically designed for **Exploratory Data Analysis (EDA)**, **Data Visualization (e.g., Power BI dashboards)**, and **Descriptive Analytics**, with a strong focus on real-world applications in AI for business. Inspired by the dynamic and data-driven environment of Asharq News, a leading Arabic-language television channel and news portal, this dataset captures critical business metrics and AI-driven performance indicators.

The dataset is contextualized for Asharq News and its operational ecosystem, reflecting key variables that influence media companies' performance in a digital and AI-driven era. It includes **qualitative data** to capture the structural characteristics of businesses (e.g., industry, customer focus, AI integration levels), **quantitative metrics** that describe performance and investment levels, and **derived variables** to provide deeper analytical insights. This combination ensures the dataset is highly relevant for evaluating AI's role in driving efficiency, revenue growth, and customer engagement in the modern business landscape.

The dataset contains **1000 records** that represent a variety of companies operating in the fields of **media, technology, finance, healthcare, retail, and manufacturing**, offering opportunities to compare performance across industries and regions. It serves as an essential tool for understanding how AI investments influence key business outcomes, enabling students to derive actionable insights through visualization and analytical techniques.

Dataset Description

The **Asharq News Dataset** is structured around 22 variables, including **core variables** that describe business performance and **derived variables** that enhance the analytical depth. This structure allows users to investigate relationships between variables, identify patterns and trends, and create compelling narratives through visualization tools like Power BI. The dataset is specifically tailored to explore how AI impacts critical aspects of a business, such as operational efficiency, market share, and customer satisfaction. Below is a detailed description of the dataset's variables:

1. Core Variables

These variables form the foundation of the dataset and represent the structural and operational characteristics of businesses:

© 2025 Dr. Yasser Aly Yakout Shehata

- **Industry Sector** (Categorical): The business sector of the company (e.g., Media, Technology, Healthcare).
- **Region** (Categorical): The geographical location of the company's headquarters (e.g., North America, Middle East, Asia).
- **AI Integration Level** (Categorical): The extent to which AI is integrated into operations (Low, Medium, High).
- **Customer Focus** (Categorical): The primary customer base of the company (B2B, B2C, or Both).
- **Sustainability Commitment** (Categorical): Indicates whether the company has formal sustainability initiatives (Yes, No).
- **Revenue (USD M)** (Numerical): The company's annual revenue, measured in millions of USD.
- **AI Investment (USD M)** (Numerical): The annual expenditure on AI technologies, measured in millions of USD.
- **Market Share (%)** (Numerical): The company's market share as a percentage of its respective industry.
- **Employee Count** (Numerical): The total number of employees in the organization.
- **Customer Satisfaction Score** (Numerical): A measure of customer satisfaction on a scale of 1 to 10.
- **AI ROI (%)** (Numerical): The return on investment (ROI) specific to AI-driven initiatives, expressed as a percentage.
- **Operational Efficiency (%)** (Numerical): The percentage improvement in operational efficiency attributed to AI.
- **R&D Expenditure (USD M)** (Numerical): Annual expenditure on research and development, measured in millions of USD.
- **Social Media Engagement (k)** (Numerical): The total number of social media interactions (likes, shares, comments) on AI-related content, measured in thousands.
- **Time to Market (months)** (Numerical): The average time it takes to launch AI-related products or services, measured in months.

© 2025 Dr. Yasser Aly Yakout Shehata

2. Derived Variables

To deepen the dataset's analytical value, derived variables have been included. These variables are calculated based on the core metrics and provide advanced insights into the efficiency and effectiveness of AI adoption:

- **Revenue per Employee (USD M)**: Measures organizational productivity by dividing revenue by the total number of employees.
- **AI Investment as % of Revenue**: Indicates how much of the company's revenue is allocated to AI investment, offering insights into resource allocation.
- **AI ROI-to-Investment Ratio**: A measure of the effectiveness of AI spending, calculated by dividing AI ROI by AI Investment.
- **Social Media Engagement Rate**: The level of social media engagement relative to revenue, highlighting the impact of online presence on business outcomes.
- **Profitability Tier**: Categorizes companies into Low, Medium, or High profitability based on their revenue performance relative to others.
- **AI Maturity Level**: Classifies companies as Low, Developing, or Advanced in AI maturity based on their AI integration level.
- **COVID-19 Impact Index**: A hypothetical index (0–10 scale) reflecting the impact of the COVID-19 pandemic on each company's operations.
- **AI-Driven Revenue (%)**: The percentage of the company's revenue directly attributable to AI initiatives.

Relevance to EDA and Visualization in Power BI

The **Asharq News Dataset** is ideal for practicing EDA and visualization in Power BI due to its rich and interconnected variables, which enable users to explore various business insights:

1. **Exploratory Data Analysis (EDA)**:
    - Students can use statistical methods to summarize data, visualize distributions, and identify trends.
    - For example:
        - Analyze how AI Investment correlates with AI ROI across industries.
        - Explore the distribution of Market Share by Region or Industry Sector.

© 2025 Dr. Yasser Aly Yakout Shehata

        - Compare Customer Satisfaction scores across companies with different AI Integration Levels.

2. **Data Visualization**:
    - **Power BI Dashboards**: The dataset enables the creation of dynamic, interactive dashboards that allow stakeholders to explore business insights at a glance.
    - Examples of possible visualizations:
        - **Regional Trends**: A heatmap showing Market Share distribution across geographical regions.
        - **AI Investment Impact**: A scatterplot illustrating the relationship between AI Investment and AI ROI.
        - **Time to Market**: A line graph comparing average Time to Market for companies with different AI Maturity Levels.
3. **Descriptive Analytics**:
    - The dataset supports descriptive analytics to summarize key performance indicators (KPIs) like average Revenue, Customer Satisfaction, and Social Media Engagement.
    - Students can calculate correlations, means, medians, and percentiles to derive actionable insights.
4. **Scenario Simulation**:
    - By analyzing the **AI ROI-to-Investment Ratio** or the **Revenue per Employee**, students can simulate different AI investment strategies and predict their outcomes.

Applications in AI for Business

This dataset is tailored to examine how AI drives business outcomes, particularly for media companies like **Asharq News**. Key applications include:

- **AI Investment Analysis**: Identifying optimal investment levels to maximize returns and operational efficiency.
- **Customer Engagement Insights**: Understanding how social media engagement and customer satisfaction impact overall performance.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Regional and Industry Comparisons**: Benchmarking companies by geography and sector to identify leaders in AI adoption.
- **AI Maturity Assessment**: Evaluating how different levels of AI integration affect profitability, efficiency, and time to market.

By leveraging this dataset, MBA students can apply advanced analytical techniques to solve real-world business challenges, making it an invaluable resource for both academic and professional growth.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 81: KAHRAMAA

Introduction

The **KAHRAMAA Dataset** has been meticulously crafted to serve as a comprehensive resource for MBA students specializing in Business Analytics with a focus on AI applications. Designed specifically to represent the operational landscape of **The Qatar General Electricity and Water Corporation (KAHRAMAA)**, this dataset provides a realistic simulation of the energy and water management sector in Qatar. KAHRAMAA, as the sole provider of electricity and water services in Qatar, emphasizes sustainability, efficiency, and innovation. This dataset embodies those principles, making it a valuable tool for analyzing AI-driven initiatives within energy and water conservation projects.

This dataset is structured to enable **Exploratory Data Analysis (EDA)**, **descriptive analytics**, and **data visualization using Power BI**, and is particularly relevant for understanding the interplay between project budgets, resource savings, customer satisfaction, and operational efficiency in energy and water utility projects. With derived variables included to enhance the analytical depth, this dataset is tailored for real-world applications in decision-making, resource optimization, and sustainability initiatives.

Dataset Description

The KAHRAMAA Dataset consists of **1000 records** representing energy and water-related projects undertaken by KAHRAMAA in various regions, client segments, and operational contexts. The dataset contains **15 original variables** and **3 derived variables** designed to improve analytical insights. These variables cover a diverse range of qualitative and quantitative aspects, enabling students and analysts to explore trends, relationships, and patterns relevant to AI applications in business.

Primary Variables

Qualitative Variables (Categorical)

1. **Business_Type**: The type of AI-driven project (e.g., "Energy Optimization," "Predictive Maintenance," "Smart Grids," or "Water Conservation"). This variable provides context for the nature of the initiative and its alignment with KAHRAMAA's goals.
2. **Region**: The geographical area where the project is executed (e.g., "Middle East," "Europe," "North America," "Asia"). This variable enables regional segmentation and trend analysis.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Client_Segment**: The type of customers targeted by the project (e.g., "Residential," "Industrial," "Commercial," "Government"). It reflects how the project's goals align with specific customer needs.
4. **Project_Status**: The current status of the project (e.g., "In Progress," "Completed," "Delayed," "On Hold"). This helps in tracking project performance and timelines.
5. **ESCO_Grade**: Accreditation grade of the Energy Services Company (ESCO) responsible for the project (e.g., "A," "B," "C"). Higher grades indicate better capabilities and resources.

Quantitative Variables (Continuous)

1. **Project_Budget_USD**: The allocated budget for each project (in USD). This variable provides insights into resource investment.
2. **Revenue_Impact_Percentage**: The percentage increase in revenue due to the implementation of the AI-driven project. This measures the economic impact.
3. **Energy_Savings_kWh**: The annual energy savings achieved by the project (in kilowatt-hours). A core metric for energy efficiency.
4. **Water_Savings_m3**: The annual water savings achieved by the project (in cubic meters). A key variable for sustainability analysis.
5. **Project_Duration_Months**: The total duration of the project (in months). This variable provides insights into project timelines.
6. **Customer_Satisfaction_Score**: The average satisfaction score from customers (on a scale of 1–100). This reflects customer perceptions of the project's success.
7. **Number_of_Clients**: The number of clients served by the project. Larger projects targeting broader client bases are reflected in this variable.
8. **Equipment_Cost_USD**: The cost of the equipment used in the project (in USD). This variable captures operational expenditures.
9. **Personnel_Training_Hours**: The total number of hours spent on personnel training for AI integration. This represents capacity-building efforts.
10. **CO2_Reduction_Tons**: The reduction in CO2 emissions achieved through the project's energy efficiency measures (in metric tons). This variable reflects environmental impact.

© 2025 Dr. Yasser Aly Yakout Shehata

Derived Variables

To deepen the analytical potential of the dataset, the following **derived variables** have been included:

1. Cost per Unit Savings

This variable calculates the cost-effectiveness of the equipment in achieving resource savings. It is the ratio of the equipment cost to the total resource savings (sum of energy and water savings).

$$\text{Cost per Unit Savings} = \frac{\text{Equipment Cost (USD)}}{\text{Energy Savings (kWh)} + \text{Water Savings (m3)}}$$

Where:

- **Equipment Cost (USD):** The total cost of equipment used in the project.
- **Energy Savings (kWh):** The amount of energy saved annually.
- **Water Savings (m^3):** The amount of water conserved annually.

2. Savings Efficiency

This variable measures the efficiency of the project in translating the allocated budget into total resource savings (energy + water).

$$\text{Savings Efficiency} = \frac{\text{Energy Savings (kWh)} + \text{Water Savings (m3)}}{\text{Project Budget (USD)}}$$

Where:

- **Energy Savings (kWh):** The amount of energy saved annually.
- **Water Savings (m^3):** The amount of water conserved annually.
- **Project Budget (USD):** The total budget allocated to the project.

3. Training Efficiency

This variable evaluates how effectively personnel training hours contribute to customer satisfaction.

$$\text{Training Efficiency} = \frac{\text{Customer Satisfaction Score}}{\text{Personnel Training Hours}}$$

Where:

© 2025 Dr. Yasser Aly Yakout Shehata

- **Customer Satisfaction Score:** The average satisfaction score (scale: 1–100) from customers impacted by the project.
- **Personnel Training Hours:** The total number of hours spent on training personnel for AI integration.

Explanation of Variables in Context

1. **Cost per Unit Savings** evaluates whether equipment costs are justified by the resource savings achieved, allowing KAHRAMAA to assess project cost-effectiveness. A lower value indicates better cost-efficiency.
2. **Savings Efficiency** highlights how efficiently the project utilizes the allocated budget to achieve energy and water savings. A higher value reflects better utilization of resources.
3. **Training Efficiency** demonstrates the relationship between training investment and customer satisfaction. A higher value indicates that training hours are effectively improving customer satisfaction.

Key Features and Analytical Relevance

Exploratory Data Analysis (EDA)

- Analyze the correlation between Project_Budget_USD and Energy_Savings_kWh to determine how budget influences energy efficiency.
- Use pair plots or heatmaps to uncover relationships between variables such as ESCO_Grade and Customer_Satisfaction_Score.
- Identify trends in Revenue_Impact_Percentage across different regions and client segments.

Visualization in Power BI

- Create dashboards showing energy and water savings by Region or Client_Segment.
- Use slicers to filter and compare completed projects by ESCO_Grade or Project_Status.
- Visualize environmental impact using bar charts or maps for CO2_Reduction_Tons.

AI for Business Applications

- Assess the **Return on Investment (ROI)** for AI projects using Savings_Efficiency and Revenue_Impact_Percentage.

© 2025 Dr. Yasser Aly Yakout Shehata

- Perform clustering to group projects by Savings_Efficiency, Customer_Satisfaction_Score, or ESCO_Grade to identify high-performing segments.
- Use predictive analytics to forecast future energy and water savings based on historical data.

Real-World Relevance to KAHRAMAA

This dataset reflects the operational challenges and innovations undertaken by KAHRAMAA in Qatar's utility sector. It aligns closely with their strategic goals:

- Promoting energy and water conservation through AI-driven projects.
- Evaluating the economic and environmental impact of sustainability initiatives.
- Improving customer satisfaction through data-driven decision-making.

By leveraging this dataset, students can explore actionable insights into how KAHRAMAA could optimize its operations, prioritize high-impact projects, and achieve its sustainability objectives. Through EDA, visualization, and analytics, students will gain practical exposure to solving real-world business challenges in the energy and water sectors.

This comprehensive **KAHRAMAA Dataset** provides an invaluable platform for MBA students studying **Data Analytics for AI Business** to explore the intersections of AI, sustainability, and business strategy, preparing them for leadership roles in the rapidly evolving utility and technology sectors.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 82: Sika

Introduction

The **Sika GCC Dataset** is a comprehensive and curated dataset designed specifically for advanced analytical purposes, with a focus on **Exploratory Data Analysis (EDA)**, **visualization using Power BI**, and **business insights through AI applications**. This dataset reflects the operations, products, and sustainability efforts of **Sika GCC**, the regional subsidiary of Sika Group, a globally renowned specialty chemicals company. It captures critical business dimensions such as revenue performance, market share, operational efficiency, customer satisfaction, and sustainability impact across six Gulf Cooperation Council (GCC) countries: the United Arab Emirates (UAE), Saudi Arabia, Bahrain, Qatar, Kuwait, and Oman.

Incorporating both qualitative and quantitative variables, this dataset simulates real-world challenges and business scenarios faced by Sika GCC in the construction and manufacturing industries. It allows MBA students and professionals to explore patterns, identify trends, and derive actionable insights, making it a perfect resource for practicing data-driven decision-making.

Additionally, the inclusion of **derived variables** and **enhanced metrics** increases the dataset's analytical depth, enabling users to explore advanced relationships such as profitability per unit sold, sustainability impact per product, and correlations between operational metrics and customer satisfaction. By leveraging the dataset in **Power BI**, students and analysts can create interactive dashboards and gain visual insights into market dynamics, product performance, and customer-centric strategies. Moreover, its alignment with AI-driven business applications allows users to explore predictive analytics and machine learning for informed decision-making.

Dataset Description

The **Sika GCC Dataset** contains 1,000 records and 18 variables, representing both core business operations and enhanced derived metrics for deeper analytical insights. The variables are categorized into **qualitative (categorical)** and **quantitative (numerical)**, with strong interrelationships that provide meaningful patterns for data analysis. Below is a detailed description of the dataset structure:

Qualitative Variables (5 variables)

1. **Region**: The GCC country where the operations occurred (UAE, Saudi Arabia, Bahrain, Qatar, Kuwait, Oman).

© 2025 Dr. Yasser Aly Yakout Shehata

- **Relevance**: Useful for regional comparison of revenue, market share, and operational performance.

2. **Product Category**: High-level categorization of Sika GCC's product portfolio (Waterproofing, Roofing, Flooring, Sealing and Bonding, Concrete, Refurbishment).
    - **Relevance**: Enables segmentation analysis of product performance and its contribution to sustainability efforts.
3. **Customer Sector**: The industry or market segment served by Sika GCC (Construction, Automotive, Marine, Manufacturing).
    - **Relevance**: Provides insights into how different sectors influence revenue and project distribution.
4. **Sustainability Rating**: A categorical rating (High, Medium, Low) representing the environmental impact and efficiency of Sika's products.
    - **Relevance**: Highlights the alignment of Sika GCC with global sustainability goals, allowing for analysis of its relationship with energy savings and carbon emission reduction.
5. **Market Segment**: The type of market being served (Residential, Commercial, Industrial, Infrastructure).
    - **Relevance**: Facilitates targeted analysis of how market segments drive business outcomes.

Quantitative Variables (10 variables)

1. **Annual Revenue (USD)**: Revenue generated by Sika GCC in USD for a given combination of region, product, and customer sector.
    - **Relevance**: Core metric to measure the financial performance of the company.
2. **Market Share (%)**: The percentage share of the market captured by Sika GCC in the respective region.
    - **Relevance**: Helps analyze competitive positioning across different GCC countries.
3. **Units Sold**: Total number of product units sold.
    - **Relevance**: Drives key performance metrics such as revenue and profitability.

© 2025 Dr. Yasser Aly Yakout Shehata

4. **Customer Satisfaction Score**: A numeric score (1 to 10) derived from customer feedback on product quality and service.
    - **Relevance**: Enables analysis of customer-centric strategies and their impact on business outcomes.
5. **R&D Expenditure (USD)**: Investment in research and development for product innovation.
    - **Relevance**: Shows the focus on innovation and its correlation with sustainability ratings and revenue.
6. **Number of Projects**: The total number of projects completed in the specified region, sector, or market.
    - **Relevance**: Indicates the operational scale and capacity of Sika GCC.
7. **Energy Savings (MWh)**: The total energy savings achieved by using Sika's products in projects.
    - **Relevance**: A sustainability metric reflecting energy efficiency.
8. **Carbon Emission Reduction (kgCO2)**: The reduction in carbon emissions achieved through the use of environmentally friendly products.
    - **Relevance**: Measures environmental impact and aligns with sustainability goals.
9. **Employee Count**: Number of employees assigned to a specific project, product category, or region.
    - **Relevance**: Useful for operational analysis and workforce planning.
10. **Delivery Timeliness (%)**: The percentage of projects delivered on or before the agreed timeline.

- **Relevance**: Reflects operational efficiency and its relationship with customer satisfaction.

Derived Variables (3 variables)

1. **Revenue per Unit Sold**:
    - Formula: Annual Revenue (USD) / Units Sold.
    - **Relevance**: Measures profitability and pricing efficiency per unit, enabling margin analysis.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Sustainability Impact Score**:
    - Formula: (Energy Savings (MWh) + Carbon Emission Reduction (kgCO2)) / Units Sold.
    - **Relevance**: Captures the environmental efficiency of Sika GCC's products on a per-unit basis.
3. **Satisfaction Level**:
    - Formula: Binning of Customer Satisfaction Score into categories ("Low", "Medium", "High").
    - **Relevance**: Simplifies the analysis of satisfaction trends across regions and market segments.

Why This Dataset is Ideal for EDA and Visualization

1. **EDA:**
    - The dataset's diverse mix of variables enables MBA students to uncover trends, patterns, and correlations through summary statistics, visual exploration, and hypothesis testing.
    - Example: Analyzing the relationship between *Sustainability Ratings* and *Revenue per Unit Sold*.
2. **Visualization in Power BI:**
    - With both regional and product-based segmentation, students can create interactive dashboards showcasing:
        - Regional market share via pie charts.
        - Product category revenue contribution through bar charts.
        - Time-series analysis of revenue and customer satisfaction using line charts.
    - Derived metrics such as *Sustainability Impact Score* provide added depth for visual storytelling.
3. **AI for Business Applications:**
    - The dataset's structure supports AI use cases like:

© 2025 Dr. Yasser Aly Yakout Shehata

- Predicting *Customer Satisfaction Scores* based on operational metrics (e.g., *Delivery Timeliness*, *R&D Expenditure*).
- Clustering regions based on revenue and market share performance for segmentation strategies.
- Forecasting *Annual Revenue* using time-series data.

Applications in Business Analytics for AI

This dataset is tailored to simulate real-world scenarios for Sika GCC's business operations. It enables students and professionals to:

- Perform **descriptive analytics** to summarize key performance metrics.
- Build **interactive dashboards** in Power BI for stakeholder communication.
- Utilize **AI techniques** like regression, clustering, and forecasting for data-driven decision-making.
- Explore **sustainability analytics** to align business goals with environmental impact.

By combining operational data, sustainability metrics, and derived variables, the **Sika GCC Dataset** serves as a rich, multi-dimensional resource for learning and applying advanced analytics techniques in the context of AI-powered business strategies.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 83: NCSP

Introduction

The dataset, named **NCSP Dataset**, is meticulously crafted to simulate operational, environmental, and financial data from the **National Company for Sulphur Products (NCSP)**, **a leading Saudi Arabian enterprise specializing in chemical production**. With over two decades of industrial expertise, NCSP operates two advanced production facilities in Riyadh and Dammam, focusing on high-demand products such as Sulphur, Sulphuric Acid, Sulphonated Products, Sodium Hypochlorite, Formaldehyde Derivatives, and Ammonia Solutions.

This dataset is specifically designed to enable **MBA students studying Business Analytics** to practice **Exploratory Data Analysis (EDA)**, **visualization techniques using Power BI**, and gain insights into operational excellence, financial performance, and process efficiencies. By including both raw metrics and derived variables, the dataset promotes deep analytical exploration, facilitates decision-making insights, and supports use cases for AI-driven predictive and prescriptive analytics in business contexts.

Dataset Description

The **NCSP Dataset** includes **15 variables** (10 quantitative and 5 qualitative) across **1,000 records**. These variables represent real-world operational, financial, and environmental metrics collected from NCSP's Riyadh and Dammam production complexes. The data has been enhanced with **derived variables**, enabling students to analyze production efficiency, energy consumption, and financial performance at a granular level.

Qualitative Variables (Categorical)

1. **Product Category**: Indicates the type of chemical product manufactured (e.g., Sulphur & Sulphuric Acid, Sulphonated Products, etc.).
2. **Facility Location**: Identifies the production complex (Riyadh or Dammam).
3. **Safety Compliance Rating**: Evaluates the safety performance of the facility (Excellent, Good, or Satisfactory).
4. **Teamwork Performance**: Rates the effectiveness of teamwork in operations (High, Moderate, or Low).
5. **Innovation Index**: Categorizes the level of innovation applied in production processes (High, Moderate, or Low).

© 2025 Dr. Yasser Aly Yakout Shehata

Quantitative Variables (Numerical)

1. **Monthly Production Volume (tons)**: Represents the total monthly output of the facility in tons.
2. **Production Downtime (hours)**: Tracks the total hours production was paused during the month.
3. **Raw Material Usage (tons)**: Reflects the amount of raw material consumed to meet production targets.
4. **Energy Consumption (kWh)**: Indicates the total energy consumption during the month.
5. **Environmental Emissions (kg CO2 equivalent)**: Measures greenhouse gas emissions during the production process.
6. **Employee Count**: Shows the total number of employees working at each facility.
7. **Safety Incidents**: Captures the number of safety-related incidents reported during the month.
8. **Customer Satisfaction (%)**: Represents the percentage of customer satisfaction, based on feedback and surveys.
9. **Product Defect Rate (%)**: Tracks the percentage of defective products in the total monthly output.
10. **Revenue (SAR)**: Represents the monthly revenue generated by the facility in Saudi Riyals.

Derived Variables (Enriching Analytical Depth)

To improve the dataset's relevance for **EDA**, **Power BI visualization**, and **AI business applications**, the following derived variables have been added:

1. **Production Efficiency**:
    - Formula: Monthly Production Volume / Raw Material Usage
    - Interpretation: Reflects how efficiently raw materials are converted into finished products. Higher values indicate better operational efficiency.
2. **Revenue per Employee**:
    - Formula: Revenue (SAR) / Employee Count

© 2025 Dr. Yasser Aly Yakout Shehata

- Interpretation: Measures the financial productivity of employees, offering insights into workforce efficiency.

3. **Energy Efficiency**:
    - Formula: Monthly Production Volume / Energy Consumption
    - Interpretation: Indicates how effectively energy is utilized to produce output, helping to assess operational sustainability.
4. **Revenue Segment**:
    - Categorization: Divides revenue into "Low," "Medium," or "High" categories based on percentile ranges.
    - Interpretation: Facilitates grouping and segmentation for analysis of revenue drivers.
5. **Emissions Segment**:
    - Categorization: Classifies facilities into "Low," "Medium," or "High" emissions categories based on percentile ranges.
    - Interpretation: Enables an environmental sustainability analysis.

Relevance for EDA and Power BI Visualization

This dataset is specifically structured to align with **EDA and visualization workflows** for MBA students. It allows students to:

1. **Explore Data Relationships**:
    - Analyze the correlation between variables such as **Production Volume**, **Raw Material Usage**, and **Revenue** to understand operational efficiency.
    - Study the impact of **Product Defect Rate** on **Customer Satisfaction** and **Revenue**.
2. **Perform Comparative Analysis**:
    - Compare operational metrics (e.g., **Production Downtime** and **Energy Consumption**) across facilities in Riyadh and Dammam.
    - Evaluate safety and teamwork performance metrics to identify areas of improvement.

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Create Interactive Dashboards in Power BI**:
    - Build dashboards for tracking KPIs such as **Revenue per Employee**, **Production Efficiency**, and **Energy Efficiency**.
    - Visualize performance by **Product Category** and **Facility Location**, enabling strategic decision-making.
4. **Derive Business Insights for AI Applications**:
    - Use AI-powered tools to build predictive models, such as forecasting **Revenue** or identifying factors driving **Customer Satisfaction**.
    - Apply clustering algorithms to segment facilities based on operational efficiency, safety performance, or emissions.

Application in AI for Business

The dataset's structure and derived variables provide a foundation for implementing **AI-driven business applications**, such as:

1. **Predictive Analytics**:
    - Forecast future **Revenue** based on variables such as **Production Volume**, **Defect Rate**, and **Customer Satisfaction**.
    - Anticipate **Energy Consumption** or **Environmental Emissions** for specific production targets.
2. **Prescriptive Analytics**:
    - Optimize **Raw Material Usage** and **Energy Efficiency** by simulating operational adjustments.
    - Suggest improvements in **Safety Compliance** or **Teamwork Performance** to minimize downtime and enhance productivity.
3. **Clustering and Segmentation**:
    - Segment facilities or production lines based on **Production Efficiency**, **Revenue per Employee**, and **Customer Satisfaction** to identify high-performing clusters.
4. **Dashboard-Driven Insights**:
    - Leverage **Power BI dashboards** to monitor KPIs, detect anomalies, and automate alerts for operational inefficiencies or safety concerns.

© 2025 Dr. Yasser Aly Yakout Shehata

Conclusion

The **NCSP Dataset** is a versatile and realistic dataset tailored to help MBA students master the art of data analytics, visualization, and AI-driven insights for business. By integrating meaningful metrics, derived variables, and industry-specific scenarios, the dataset provides a rich foundation for hands-on learning in **EDA**, **Power BI visualization**, and advanced business analytics techniques. Its focus on NCSP's operations and sustainability metrics makes it highly relevant for real-world decision-making and strategic business planning in industrial contexts.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 84: Al Qaryan

Introduction to the Dataset

The **Al Qaryan Dataset** is a synthetic yet realistic representation of business and operational data from **Al Qaryan Group**, a prominent leader in recycling and industrial sustainability in the **GCC region**. This dataset has been meticulously designed to align with the company's core activities, subsidiaries, and their contributions to recycling, waste management, and sustainability, as highlighted in Saudi Arabia's **Vision 2030**.

The dataset provides 1,000 records reflecting Al Qaryan Group's operations across different regions, subsidiaries, and industries served. It captures key quantitative and qualitative attributes that are essential for analyzing business performance, sustainability initiatives, and global outreach. Moreover, the dataset is enriched with derived variables, ensuring deeper analytical insights and making it particularly relevant for **Exploratory Data Analysis (EDA)**, **data visualization in Power BI**, and **AI-driven business applications**. This dataset bridges the gap between operational insights and strategic decision-making, providing MBA students and analysts with a robust platform to practice real-world analytics.

Description of the Dataset

The **Al Qaryan Dataset** consists of **18 variables** (5 qualitative, 10 quantitative, and 3 derived metrics), each carefully crafted to reflect the operational and sustainability-focused nature of Al Qaryan Group. The dataset supports MBA students and business analysts in applying advanced analytical techniques such as **EDA**, **visualization**, and **descriptive analytics** while uncovering relationships and patterns across business processes. Additionally, the dataset is optimized for AI applications in business, enabling predictive modeling, anomaly detection, and scenario-based simulations.

Dataset Variables and Analytical Relevance

The variables are divided into the following categories:

1. Qualitative (Categorical) Variables

These variables enable filtering, grouping, and segmentation during visualization and analysis:

- **Region**: Represents the geographic area of operations (Central, Eastern, Western, Northern, Southern). This allows regional performance analysis and segmentation.
- **Subsidiary**: Captures the specific branch of Al Qaryan Group (Recycling, Steel, Logistics, Plastic, International). Analysts can compare the operational performance of different subsidiaries.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Certification Level**: Indicates ISO certifications obtained by the company (ISO 9001, ISO 14001, ISO 45001), reflecting quality, environmental, and safety standards.
- **Client Sector**: Represents industries served by Al Qaryan (Automotive, Construction, Energy, Retail, Manufacturing). This allows sector-based revenue and project analysis.
- **Sustainability Initiative**: Denotes specific programs undertaken (Recycling Drives, Circular Economy Projects, Carbon Reduction). Analysts can evaluate the impact of these initiatives on sustainability metrics.

2. Quantitative Variables

These variables capture operational and financial performance, making them critical for descriptive and predictive analytics:

- **Year of Operation**: Captures the year in which the data was recorded (2018–2023). Useful for time-series analysis and identifying growth trends.
- **Waste Recycled (Tons)**: Indicates the total amount of waste recycled annually, a key metric for measuring the company's recycling efforts.
- **Recycled Grades**: Represents the number of different recycled material grades processed, highlighting operational diversity.
- **Employees**: The number of employees working in a subsidiary, used to assess workforce efficiency.
- **Material Recycling Facilities**: Total number of facilities in operation, enabling capacity and resource optimization analysis.
- **Global Clients**: Number of countries served through international trade, reflecting global reach and market penetration.
- **Recycled Metals (Tons)**: The total amount of metals recycled annually, a critical metric for tracking industrial performance.
- **Revenue (Million SAR)**: Annual revenue generated in millions of Saudi Riyals, allowing for financial performance evaluation.
- **Carbon Emissions Reduced (Tons)**: The amount of greenhouse gas emissions reduced annually, aligning with the company's sustainability goals.
- **Projects Completed**: Total number of recycling and sustainability projects executed, demonstrating operational impact.

© 2025 Dr. Yasser Aly Yakout Shehata

3. Derived Variables

To improve analytical depth, the dataset includes three derived metrics:

- **Recycling Efficiency**: Calculated as Waste Recycled (Tons) / Employees. This metric assesses the operational efficiency of each subsidiary by evaluating waste recycling per employee.
- **Revenue per Client**: Calculated as Revenue (Million SAR) / Global Clients. This highlights the revenue generated from each global client, providing insights into market contribution and client value.
- **Emission Reduction Rate**: Calculated as Carbon Emissions Reduced (Tons) / Waste Recycled (Tons). This metric measures the company's ability to reduce emissions relative to the waste it processes, reflecting the environmental impact of recycling efforts.

Use of the Dataset for EDA and Visualization

The dataset has been specifically designed for **EDA** and visualization in tools like **Power BI**, allowing students and analysts to extract meaningful insights and create professional dashboards. Key applications include:

1. **Exploratory Data Analysis (EDA)**:
    - Identify trends over time (e.g., growth in revenue or waste recycled by year).
    - Analyze correlations between operational metrics (e.g., Waste Recycled vs. Revenue or Employees vs. Recycling Efficiency).
    - Segment performance by region, subsidiary, or client sector to uncover patterns and disparities.
2. **Visualization in Power BI**:
    - Create interactive dashboards to monitor sustainability KPIs, such as Carbon Emissions Reduced and Recycling Efficiency.
    - Use slicers and filters to explore data by region, subsidiary, or year.
    - Generate visualizations like bar charts, scatter plots, and heatmaps to evaluate relationships between variables (e.g., Revenue vs. Global Clients).

© 2025 Dr. Yasser Aly Yakout Shehata

3. **Applications in AI for Business**:
    - Train predictive models to forecast future revenue or recycling output based on historical data.
    - Use clustering techniques to identify high-performing subsidiaries or regions.
    - Build classification models to assess the likelihood of exceeding emission reduction targets based on recycling initiatives.

Why This Dataset is Unique and Valuable

1. **Relevance to Al Qaryan Group**:
    - The dataset is tailored to reflect Al Qaryan Group's operations, subsidiaries, and sustainability focus. It integrates real-world variables such as recycling output, emissions reduction, and certifications, ensuring authenticity.
2. **Analytical Versatility**:
    - With a combination of financial, operational, and sustainability variables, the dataset supports a wide range of analytical techniques, from basic descriptive analytics to advanced machine learning models.
3. **Educational Purpose**:
    - Designed specifically for MBA students and business analysts, the dataset offers a practical platform for learning data analytics concepts and applying them to real-world business challenges.
4. **Support for EDA and Power BI**:
    - The inclusion of qualitative and quantitative variables, along with derived metrics, makes the dataset ideal for creating interactive Power BI dashboards and performing in-depth exploratory analysis.

Conclusion

The **Al Qaryan Dataset** is a comprehensive resource for studying and applying business analytics in the context of sustainability and industrial operations. By reflecting the core operations of Al Qaryan Group, it provides a unique opportunity to explore data-driven decision-making in the recycling industry. With its relevance to **EDA**, **Power BI visualization**, and **AI for business**, the dataset is a valuable tool for students, analysts, and researchers aiming to enhance their analytical skills and gain actionable insights into sustainable business practices.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 85: Taawin

Introduction to the Dataset

The **Taawin Medical Complexes Dataset** is a comprehensive, carefully curated dataset that represents simulated operational and patient data from **Taawin Medical Complexes**, a leading healthcare provider in **Al-Kharj, Saudi Arabia**. With over 40 years of experience, Taawin Medical Complexes have established themselves as pioneers in delivering exceptional healthcare services across multiple medical specialties. This dataset has been meticulously designed to reflect the operational dynamics, patient demographics, and service metrics of Taawin Medical Complexes, making it an invaluable resource for **Exploratory Data Analysis (EDA)**, **visualization in Power BI**, and the application of **AI in business analytics**.

The dataset consists of **1,000 records** and is structured to include **15 key variables**, capturing both qualitative (categorical) and quantitative (numerical) aspects of Taawin's operations. It has been enhanced with additional **derived variables** to provide greater analytical depth, enabling students and professionals to uncover insights into patient behaviors, service efficiencies, and operational performance.

Designed specifically for **MBA students studying Business Analytics for AI applications**, this dataset allows learners to apply analytical techniques such as data visualization, segmentation, trend analysis, and descriptive statistics in a healthcare context. It also demonstrates the value of leveraging data-driven insights to improve operational efficiency, patient satisfaction, and overall business performance.

Description of the Dataset

The **Taawin Medical Complexes Dataset** is composed of **qualitative (categorical)** and **quantitative (numerical)** variables, each representing a critical component of the healthcare services offered by the three branches of Taawin Medical Complexes. Below is a detailed description of the dataset variables, followed by a summary of **derived variables** introduced to enhance analytical depth.

1. Categorical Variables

The categorical variables in the dataset provide insights into patient demographics, service categories, and payment methods. These variables enable segmentation and comparative analysis across different branches and service types:

- **Branch_Location**: The branch where the service was provided, representing the three operational branches of Taawin Medical Complexes:

© 2025 Dr. Yasser Aly Yakout Shehata

  - First Taawin Complex (Al-Salam, Al-Kharj)
  - Second Taawin Complex (Al-Aziziyah, Al-Kharj)
  - Third Taawin Complex (Al-Dilam)
- **Medical_Department**: The department providing the service, spanning five key specialties:
  - Internal Medicine, Pediatrics, Obstetrics & Gynecology, Dentistry, and Auxiliary Services.
- **Patient_Type**: Indicates whether the patient is a **New Patient** or a **Returning Patient**, useful for understanding patient retention and loyalty.
- **Payment_Method**: Represents the payment method used for the service:
  - Options include "Cash", "Credit/Debit Card", and "Insurance".
- **Disability_Support_Status**: Indicates whether the patient benefited from Taawin's inclusive support services for individuals with disabilities ("Yes" or "No").

2. Numerical Variables

The numerical variables capture operational and patient data, allowing for in-depth quantitative analysis:

- **Patient_Age**: The age of the patient (0–90 years), distributed across all service departments.
- **Consultation_Fee (SAR)**: The consultation fee charged for the visit, varying by department (range: 50–500 SAR).
- **Treatment_Cost (SAR)**: Additional costs for treatments, procedures, or medications (range: 0–5,000 SAR).
- **Insurance_Coverage_Percentage (%)**: The percentage of the bill covered by insurance for insured patients.
- **Visit_Duration_Minutes**: The total duration of the visit in minutes, reflecting time spent on consultation, treatment, or waiting (range: 15–180 minutes).
- **Staff_Response_Time (Minutes)**: The average time taken by staff to respond to a patient's request during the visit (range: 1–20 minutes).
- **Lab_Tests_Ordered**: The number of laboratory tests ordered during the visit (range: 0–10).

© 2025 Dr. Yasser Aly Yakout Shehata

- **Patient_Satisfaction_Score**: A rating of the patient's overall satisfaction with the visit (scale: 1–5; 1 = Very Dissatisfied, 5 = Very Satisfied).
- **Follow_Up_Days**: The number of days until the patient's next scheduled follow-up (range: 0–90 days).
- **Staff_Attendance_Rate (%)**: The percentage of scheduled staff present at the branch on the day of the visit (range: 75%–100%).

3. Derived Variables for Enhanced Analysis

To improve the analytical depth of the dataset and enable more sophisticated insights, several derived variables have been introduced:

1. **Total_Cost (SAR)**:
    - The total monetary cost of a patient's visit, calculated as:
        - $\text{Total_Cost} = \text{Consultation_Fee} + \text{Treatment_Cost}$
    - This variable highlights the overall financial impact of healthcare services.
2. **Cost_Per_Minute (SAR/Minute)**:
    - A measure of visit efficiency, calculated as:
        - $\text{Cost_Per_Minute} = \text{Total_Cost} / \text{Visit_Duration_Minutes}$
    - Useful for analyzing cost-efficiency across departments or branches.
3. **High_Satisfaction**:
    - A binary flag (1 = Yes, 0 = No) indicating whether the patient rated their satisfaction score as 4 or higher.
    - This variable helps identify factors contributing to high satisfaction levels.
4. **Insurance_User**:
    - A binary flag (1 = Yes, 0 = No) indicating whether the patient paid via insurance.
    - Useful for analyzing the financial dynamics of insured vs. non-insured patients.
5. **Day_of_Visit**:
    - Simulates the day of the week for the patient's visit, randomly assigned to reflect peak visit days (e.g., weekends).

© 2025 Dr. Yasser Aly Yakout Shehata

- Enables temporal analysis of patient traffic patterns.

6. **Time_of_Day**:
   - Simulates whether the visit occurred in the **Morning**, **Afternoon**, or **Evening**.
   - Provides insights into operational efficiency and staffing needs during different times of the day.
7. **Service_Efficiency_Ratio**:
   - The ratio of **Lab_Tests_Ordered** to **Visit_Duration_Minutes**, measuring service efficiency and throughput.

Relevance for EDA, Power BI, and AI in Business

The Taawin dataset is highly suitable for **Exploratory Data Analysis (EDA)**, **visualization in Power BI**, and **AI applications in business analytics**. Its design allows for:

1. **EDA**:
   - Analyze trends, correlations, and distributions to uncover patterns in patient demographics, costs, and satisfaction levels.
   - Perform segmentation by branch, department, or patient type to identify key drivers of business performance.
2. **Power BI Visualization**:
   - Create dashboards to visualize patient traffic, financial metrics (e.g., consultation fees and treatment costs), and patient satisfaction scores.
   - Develop heatmaps, bar charts, and pie charts to highlight peak visit times and high-performing branches.
3. **AI in Business Analytics**:
   - Use the dataset for predictive modeling, such as predicting patient satisfaction based on operational variables (e.g., response time and visit duration).
   - Identify cost drivers and forecast revenue using machine learning techniques.
   - Optimize staffing and resource allocation through temporal and spatial analyses.

© 2025 Dr. Yasser Aly Yakout Shehata

Applications in MBA Education

This dataset is a perfect fit for MBA students specializing in **Data Analytics for AI Business**. It provides practical opportunities to:

- Perform **descriptive analytics**: Summarize trends in operational and patient data.
- Use **data visualization**: Build insightful dashboards to support decision-making.
- Explore **correlations and causal relationships**: Understand how variables like insurance coverage or response times influence patient satisfaction.
- Develop **data-driven business strategies**: Apply findings to improve efficiency, patient experience, and profitability.

The inclusion of derived variables and realistic data distributions ensures that this dataset simulates real-world business scenarios, preparing students for careers in healthcare analytics, AI-driven business strategies, and decision science.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 86: Almajdouie

Introduction to the Dataset

The **Almajdouie Motors dataset** is a carefully curated and enhanced dataset specifically designed for Exploratory Data Analysis (EDA), visualization using tools like Power BI, and applications in AI-driven business analytics. This dataset models realistic scenarios and customer interactions within the **automotive industry**, with a focus on **Almajdouie Motors—a leading automotive company in Saudi Arabia** representing prestigious brands like Hyundai, Genesis, Changan, Peugeot, and Citroën. By incorporating key customer, vehicle, and business variables, the dataset reflects market dynamics, customer preferences, and operational patterns at Almajdouie Motors, making it ideal for students and professionals studying data analytics for AI in business contexts.

The dataset bridges the gap between raw data and actionable insights, enabling users to explore correlations, identify trends, and derive meaningful patterns for decision-making. In particular, it emphasizes the relationships between customer demographics, income levels, vehicle purchasing behavior, service utilization, satisfaction levels, and loyalty scores. Furthermore, the dataset includes derived variables and logical relationships to enhance analytical depth, allowing users to apply advanced data analytics techniques, including predictive modeling, segmentation, and business intelligence dashboards.

This dataset serves as a rich resource for MBA students, data analysts, and AI enthusiasts to practice real-world business analytics and AI-driven decision-making within **a competitive automotive market like that of Almajdouie Motors**.

Dataset Description

The **Almajdouie Motors dataset** contains **1,000 records** and **15 variables**, capturing both qualitative and quantitative aspects of customer, vehicle, and service-related data. The dataset is structured to include variables that are interrelated, making it highly relevant for EDA, Power BI visualizations, and the application of AI in business. Additionally, the inclusion of **derived variables** allows for a deeper exploration of customer behavior and business performance. Below is a detailed description of the dataset's variables:

Qualitative Variables (5 Variables)

1. **Customer Region**:
    - Represents the geographical region where the customer resides.
    - Categories: Central, Western, Eastern, Northern, Southern.
    - Application: Analyze regional demand for specific brands and vehicle types.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Vehicle Brand**:
    - Indicates the brand of the vehicle purchased.
    - Categories: Hyundai, Genesis, Changan, Peugeot, Citroën.
    - Application: Identify brand preferences and segment the market by customer demographics.
3. **Vehicle Type**:
    - Specifies the type of vehicle purchased.
    - Categories: Sedan, SUV, Truck, Hatchback.
    - Application: Explore vehicle type preferences across income levels and regions.
4. **Service Package**:
    - Denotes the type of service package selected by the customer.
    - Categories: Basic, Standard, Premium.
    - Application: Analyze service package uptake and its impact on satisfaction and loyalty.
5. **Customer Satisfaction Level**:
    - Reflects the satisfaction level of the customer after their experience with Almajdouie Motors.
    - Categories: Very Satisfied, Satisfied, Neutral, Dissatisfied, Very Dissatisfied.
    - Application: Evaluate factors influencing customer satisfaction and identify areas for improvement.

Quantitative Variables (10 Variables)

1. **Age of Customer**:
    - The age of the customer in years (Range: 18–65).
    - Application: Segment customers by age group and analyze their purchasing preferences.
2. **Monthly Income (SAR)**:
    - The monthly income of the customer in Saudi Riyals (Range: 5,000–50,000).

© 2025 Dr. Yasser Aly Yakout Shehata

- Application: Assess income-based vehicle preferences and affordability trends.

3. **Vehicle Price (SAR)**:
   - The price of the vehicle purchased (Range: 50,000–250,000 SAR).
   - Application: Analyze price distribution by brand, type, and customer demographics.
4. **Loan Amount (SAR)**:
   - The financing or loan amount taken for the vehicle purchase (Range: 20,000–200,000 SAR).
   - Application: Explore financing patterns and their relationship with income and down payment.
5. **Down Payment (SAR)**:
   - The initial payment made by the customer for the vehicle (Range: 5,000–50,000 SAR).
   - Application: Analyze correlations between income levels, down payments, and loan amounts.
6. **Number of Services Availed**:
   - The total number of services availed by the customer in a year (Range: 0–10).
   - Application: Identify service usage patterns and their impact on loyalty.
7. **Mileage (km)**:
   - The distance covered by the vehicle (Range: 0–50,000 km).
   - Application: Examine vehicle usage patterns and their correlation with satisfaction.
8. **Response Time (hrs)**:
   - The average time taken by Almajdouie Motors to respond to customer queries (Range: 0.5–5 hours).
   - Application: Analyze response efficiency and its effect on customer satisfaction.

© 2025 Dr. Yasser Aly Yakout Shehata

9. **Test Drive Duration (mins)**:
    - The duration of the test drive conducted by the customer before purchasing the vehicle (Range: 10–60 minutes).
    - Application: Assess how test drives influence purchasing decisions and satisfaction.
10. **Customer Loyalty Score**:
    - A loyalty score reflecting the likelihood of a customer making repeat purchases or referring Almajdouie Motors (Range: 1–10).
    - Application: Measure loyalty across different demographics, service packages, and satisfaction levels.

Derived Variables

To improve the analytical depth of the dataset, the following derived variables were introduced:

1. **Loan-to-Value Ratio (LTV Ratio)**:
    - Formula: $\text{Loan Amount} / \text{Vehicle Price}$.
    - Indicates the proportion of the vehicle's price that is financed through a loan.
    - Application: Assess financing risks and identify customers with high dependency on loans.
2. **Satisfaction Index**:
    - Formula:
      $\text{A numeric representation of satisfaction levels (e.g., "Very Satisfied"} = 5, \text{"Very Dissatisfied"} = 1)$.
    - Application: Quantify customer satisfaction for statistical analysis.
3. **Service Utilization Rate**:
    - Formula: $\text{Number of Services Availed} / \text{Maximum Services Offered}$.
    - Reflects how extensively customers use available service packages.
    - Application: Identify service utilization patterns and their effect on loyalty.

© 2025 Dr. Yasser Aly Yakout Shehata

4. **Response Efficiency Score**:
    - Formula: $(1 \,/\, \text{Response Time}) * 10$.
    - A higher score indicates faster response times and better efficiency.
    - Application: Explore the relationship between response efficiency and customer satisfaction.
5. **Income-to-Price Ratio**:
    - Formula: $\text{Monthly Income} \,/\, \text{Vehicle Price}$.
    - Measures the affordability of vehicles based on the customer's income.
    - Application: Identify income groups that may require tailored financing options.

Application in EDA and Visualization

This dataset is specifically designed to facilitate **EDA** and **Visualization** using tools like Power BI. By exploring the relationships between variables, students and professionals can develop dashboards and visualizations that reveal **critical business insights for Almajdouie Motors**. Potential applications include:

1. **Power BI Dashboards**:
    - **Sales Analysis**: Visualize sales distribution by region, vehicle type, and brand.
    - **Satisfaction Trends**: Monitor customer satisfaction levels by region, service package, and response efficiency.
    - **Customer Segmentation**: Segment customers based on age, income, and loyalty scores.
2. **EDA Insights**:
    - Identify regional trends in vehicle preferences.
    - Explore correlations between income levels and vehicle price or loan amounts.
    - Analyze how response time impacts satisfaction and loyalty.
3. **AI for Business Applications**:
    - Use predictive models to forecast customer satisfaction and loyalty.

© 2025 Dr. Yasser Aly Yakout Shehata

- Implement clustering algorithms for customer segmentation.
- Apply machine learning to recommend optimized service packages based on customer behavior.

Conclusion

The **Almajdouie Motors dataset** is a comprehensive resource for analyzing customer behavior, sales trends, and operational efficiency within the **automotive industry**. By combining qualitative insights with quantitative relationships and derived variables, it provides a realistic platform for MBA students and data professionals to practice data analytics and AI-driven decision-making. This dataset not only supports EDA and visualization tasks but also fosters a deeper understanding of data's role in business strategy and customer-centric innovation.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 87: Musanadah

Introduction

In the rapidly evolving field of mental health and rehabilitation services, data-driven decision-making plays a critical role in enhancing patient care, optimizing resource allocation, and improving treatment outcomes. The **Musanadah Center for Psychiatry, Rehabilitation, and Addiction Treatment** is committed to leveraging data analytics to support evidence-based practices and strategic planning. The **Musanadah Dataset** has been meticulously designed to provide valuable insights into key aspects of patient demographics, clinical interventions, and treatment outcomes.

This dataset serves as an essential resource for conducting **Exploratory Data Analysis (EDA), Visualization in Power BI, and AI-driven business applications**, offering healthcare professionals, researchers, and business analysts an opportunity to analyze patterns, identify trends, and develop predictive models. By incorporating both raw and derived variables, the dataset enables deeper insights into the factors influencing patient recovery, treatment adherence, and long-term well-being.

With a focus on improving operational efficiency and patient outcomes, the dataset provides a solid foundation for applying AI-based techniques to enhance decision-making processes, personalize treatment plans, and optimize the utilization of healthcare resources.

Dataset Description

The **Musanadah Dataset** encompasses **15 variables**, including **5 qualitative (categorical) variables** and **10 quantitative (numerical) variables**, carefully selected to capture critical dimensions of psychiatric and addiction treatment services. Additionally, derived variables have been introduced to enhance analytical depth and facilitate predictive analytics in mental healthcare.

1. Qualitative (Categorical) Variables

These variables provide insights into demographic and treatment-related characteristics, enabling segmentation and pattern identification.

- **Gender:** (Male, Female, Non-Binary) – Useful for analyzing gender-based differences in treatment adherence and outcomes.
- **Marital_Status:** (Single, Married, Divorced, Widowed) – A factor influencing social support and treatment engagement.
- **Education_Level:** (None, High School, Bachelor's, Master's, Doctorate) – Reflecting patients' ability to engage with treatment programs.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Diagnosis_Category:** (Anxiety Disorder, Substance Use Disorder, Depression, PTSD, Bipolar Disorder) – Key in understanding condition-specific treatment needs.
- **Treatment_Type:** (Inpatient, Outpatient, Partial Hospitalization) – Essential for evaluating the effectiveness of different treatment settings.

2. Quantitative (Numerical) Variables

These variables provide measurable indicators of patient treatment, clinical progress, and operational efficiency.

- **Age:** (Years, 18–75) – Essential for understanding treatment needs across different age groups.
- **Length_of_Stay:** (Days, 1–365) – Measures the duration of treatment and its correlation with recovery.
- **Number_of_Sessions:** (1–100) – Indicates treatment engagement levels.
- **Severity_Score:** (Scale: 1.0–10.0) – Represents the intensity of the patient's condition at intake.
- **Relapse_Risk:** (Probability: 0.1–1.0) – Assesses the likelihood of the patient experiencing a relapse post-treatment.
- **Comorbid_Conditions:** (Count, 0–2) – Captures the presence of coexisting conditions affecting treatment outcomes.
- **Family_Visits:** (Count, 0–9) – Evaluates the role of family support in treatment success.
- **Functional_Improvement:** (Scale: 1.0–5.0) – Measures progress in daily functioning.
- **Quality_of_Life:** (Scale: 1.0–5.0) – Reflects the patient's self-reported well-being.
- **Treatment_Adherence:** (Percentage, 50–100%) – Captures the patient's compliance with the treatment plan.

3. Derived Variables for Enhanced Analytical Depth

To enrich the dataset and support deeper analysis, several derived variables have been created to provide actionable insights into patient progress and operational efficiency:

3.1 Recovery Efficiency

The **Recovery Efficiency** metric evaluates how effectively a patient's functional improvement progresses over the duration of their treatment stay.

© 2025 Dr. Yasser Aly Yakout Shehata

$$\text{Recovery Efficiency} = \frac{\text{Functional Improvement}}{\text{Length of Stay}}$$

- **Functional Improvement (FI):** Patient's improvement score (scale: 1.0 – 5.0).
- **Length of Stay (LOS):** Number of days the patient remained in treatment.
- **Interpretation:** Higher values indicate faster recovery relative to time spent in treatment.

3.2 Adherence Category

Patients are categorized based on their adherence to treatment recommendations, segmented into three levels:

$$Adherence\ Category = \begin{cases} \text{High}, & \text{if Treatment Adherence} \geq 85\% \\ \text{Moderate}, & \text{if } 70\% \leq \text{Treatment Adherence} < 85\% \\ \text{Low}, & \text{if Treatment Adherence} < 70 \end{cases}$$

- **Treatment Adherence (TA):** Percentage compliance with the prescribed treatment plan.
- **Interpretation:** Helps in identifying patients at risk of non-compliance.

3.3 Support Index

The **Support Index** quantifies the level of social support available to the patient, combining family visits and marital status as a proxy for support structure.

$$\text{Support Index} = \text{Family Visits} + \delta(\text{Marital Status} = \text{Married})$$

Where:

$$\delta(\text{Marital Status} = \text{Married}) = \begin{cases} 1, & \text{if the patient is Married} \\ 0, & \text{otherwise} \end{cases}$$

- **Family Visits (FV):** Number of visits received by the patient.
- **Marital Status (MS):** Whether the patient is married (binary indicator).
- **Interpretation:** Higher values indicate stronger social support, which may positively impact treatment adherence.

3.4 Risk-Adjusted Stay

The **Risk-Adjusted Stay** metric evaluates the effectiveness of treatment duration by factoring in the patient's relapse risk.

$$\text{Risk} - \text{Adjusted Stay} = \text{Length of Stay} \times \text{Relapse Risk}$$

© 2025 Dr. Yasser Aly Yakout Shehata

- **Length of Stay (LOS):** Total treatment duration in days.
- **Relapse Risk (RR):** Probability of the patient experiencing a relapse (range: 0.1 – 1.0).
- **Interpretation:** A higher value suggests a potentially ineffective or high-risk treatment period.

3.5 Improvement Per Session

These metric measures how much functional improvement is achieved per therapy/treatment session attended.

$$\text{Improvement Per Session} = \frac{\text{Functional Improvement}}{\text{Number of Sessions}}$$

- **Functional Improvement (FI):** Patient's improvement score (scale: 1.0 – 5.0).
- **Number of Sessions (NS):** Total therapy sessions attended.
- **Interpretation:** Provides insight into session effectiveness and helps optimize treatment plans.

3.6 Quality of Life Index

This metric provides an adjusted measure of quality of life, considering severity and adherence to treatment.

$$\text{Quality of Life Index} = \frac{\text{Quality of Life} \times \text{Treatment Adherence}}{\text{Severity Score}}$$

- **Quality of Life (QoL):** Patient's self-reported well-being score (scale: 1.0 – 5.0).
- **Treatment Adherence (TA):** Percentage compliance with prescribed treatment.
- **Severity Score (SS):** Condition severity at intake (scale: 1.0 – 10.0).
- **Interpretation:** Higher values indicate better perceived well-being relative to illness severity.

3.7 Engagement Score

The **Engagement Score** captures the extent of patient involvement in their treatment, factoring in adherence and the number of attended sessions.

$$\text{Engagement Score} = \text{Treatment Adherence} \times \log(\text{Number of Sessions} + 1)$$

- **Treatment Adherence (TA): Percentage of compliance with treatment.**
- **Number of Sessions (NS):** Total number of attended therapy sessions.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Interpretation:** A higher score reflects greater patient commitment to treatment.

3.8 Relapse Likelihood Index

A risk-adjusted metric to assess the likelihood of relapse considering multiple factors.

$$\text{Relapse Likelihood Index} = \frac{\text{Severity Score} \times (1 - \text{Treatment Adherence})}{\text{Functional Improvement}}$$

- **Severity Score (SS):** Initial condition severity (scale: 1.0 – 10.0).
- **Treatment Adherence (TA):** Compliance level with the treatment plan.
- **Functional Improvement (FI):** Measured improvement in daily functionality.
- **Interpretation:** Higher values suggest a greater risk of relapse.

3.9 Cost-Effectiveness Ratio

Evaluates the efficiency of the treatment process by considering the functional improvement per day of treatment.

$$\text{Cost} - \text{Effectiveness Ratio} = \frac{\text{Functional Improvement}}{\text{Length of Stay}}$$

- **Functional Improvement (FI):** Improvement score (scale: 1.0 – 5.0).
- **Length of Stay (LOS):** Treatment duration in days.
- **Interpretation:** Higher values indicate cost-effective treatment strategies.

3.10 Support-Adjusted Adherence

This variable adjusts the adherence rate based on available social support factors.

$$\text{Support} - \text{Adjusted Adherence} = \frac{\text{Treatment Adherence} \times (\text{Family Visits} + 1)}{1 + \text{Comorbid Conditions}}$$

- **Treatment Adherence (TA):** Compliance with prescribed treatment.
- **Family Visits (FV):** Number of family visits during treatment.
- **Comorbid Conditions (CC):** Number of additional conditions (0, 1, or 2).
- **Interpretation:** Provides insights into how social support influences adherence in patients with comorbidities.

These derived variables enrich the **Musanadah Dataset** by providing new perspectives on patient treatment outcomes, social support impact, and operational efficiency. By leveraging these equations, healthcare analysts and business professionals can conduct

© 2025 Dr. Yasser Aly Yakout Shehata

deeper analyses, enabling more informed decision-making for optimizing patient care at the **Musanadah Center for Psychiatry, Rehabilitation, and Addiction Treatment.**

Relevance to EDA and Visualization in Power BI

The **Musanadah Dataset** has been strategically designed to facilitate a range of analytical exercises suitable for Power BI and AI-driven business applications. This dataset offers opportunities to perform:

1. Exploratory Data Analysis (EDA):

- Understanding patient demographics and treatment trends.
- Assessing the distribution of severity scores across different treatment types.
- Investigating the impact of social support on treatment adherence and functional improvement.

2. Visualization in Power BI:

The dataset allows students and analysts to develop insightful dashboards, including:

- **Patient Profile Analysis:** Demographic breakdown of patients by diagnosis category and treatment type.
- **Outcome Monitoring:** Line charts showing trends in functional improvement over sessions.
- **Adherence Trends:** Bar charts comparing adherence across different demographic groups.
- **Risk Analysis:** Heatmaps to identify high-risk patients based on severity and relapse probability.
- **Treatment Cost Efficiency:** Visualization of recovery efficiency across treatment settings.

3. AI Business Applications:

The dataset is well-suited for building AI-driven models to:

- Predict patient relapse risk based on clinical and demographic factors.
- Personalize treatment plans based on adherence patterns and functional improvement rates.
- Optimize resource allocation by forecasting demand for specific treatment types.

© 2025 Dr. Yasser Aly Yakout Shehata

- Develop data-driven strategies for patient engagement and retention.

Conclusion

The **Musanadah Dataset** is a comprehensive and analytically rich resource designed for the analysis of psychiatric, rehabilitation, and addiction treatment services. By offering a mix of raw and derived variables, the dataset enables detailed exploration of factors influencing patient outcomes and operational efficiency.

MBA students and healthcare analysts can leverage this dataset to develop actionable insights, improve patient care strategies, and contribute to data-driven decision-making at the **Musanadah Center for Psychiatry, Rehabilitation, and Addiction Treatment.**

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 88: El Hassan & El Hussein

1. Introduction

This project has been developed as part of the MBA Capstone in *Business Analytics for Decision Making*, and it addresses the practical realities of workforce management within small-to-medium enterprises (SMEs). The focal case is drawn from **El Hassan & El Hussein**, a manufacturing and distribution company operating in **Port Said, Egypt**, specialising in the packaging and logistics of food and household goods. Within this organisational context, the analysis is positioned from the perspective of **Ziad Elsayed Ahmed Abowarda**, an **HR Officer** tasked with overseeing personnel records, recruitment processes, attendance monitoring, and compensation reporting. The business environment in which this organisation operates is marked by labour-intensive activities, fluctuating demand, and seasonally driven operational pressures. As such, the capacity to derive accurate, timely, and actionable workforce insights is central to both immediate operational performance and longer-term strategic planning.

The primary objective of the dataset and subsequent analysis is to enable **evidence-based workforce decision-making**. Historically, HR functions in SMEs have often relied on manual reporting or fragmented information systems, limiting their ability to provide holistic insights into staff productivity, absenteeism, or compensation patterns. This project seeks to overcome those challenges by presenting a fully integrated and coherent dataset that simulates the type of information HR managers require to evaluate employee performance, forecast workforce needs, and optimise compensation structures. By structuring the dataset around the employee–month as the unit of analysis, it becomes possible to capture the cyclical dynamics of the workforce, such as seasonal peaks in utilisation, the accumulation of tenure, and the interplay between overtime hours and fatigue-related absence.

The dataset encompasses **1,000 records across 17 variables**, with seven qualitative dimensions (e.g., department, role level, contract type) and ten quantitative measures (e.g., base pay, overtime hours, utilisation percentage, productivity index). Importantly, it incorporates an accounting identity in which **total compensation is the sum of base pay, overtime pay, and allowances**. This design feature not only reflects business logic but also serves as a built-in validation mechanism, ensuring numerical consistency across records. Furthermore, the dataset introduces derived indicators such as **absence rate**, **overtime intensity**, and **cost per productivity point**, thereby translating raw metrics into forms that are directly interpretable by managers.

© 2025 Dr. Yasser Aly Yakout Shehata

The methodological design emphasises both **realism and reproducibility**. To this end, the dataset is generated using deterministic procedures anchored by a fixed random seed. It includes benign levels of statistical noise and mild outlier mechanisms to preserve authenticity while avoiding distortions that could undermine interpretation. Winsorisation at the 1st and 99th percentiles is applied to continuous fields, reducing the influence of extreme values while maintaining the dispersion necessary for meaningful analysis. The final dataset is fully complete, containing no missing values, and is therefore suitable for seamless ingestion into Power BI or related analytics platforms.

The academic and pedagogical contribution of this work lies in its ability to simulate a realistic HR operating environment for analytical purposes. It demonstrates how **business analytics frameworks** can be applied to human resource management, a function traditionally under-utilised in data-driven decision-making. For MBA candidates, this dataset represents not merely a collection of numbers but an opportunity to engage in a full cycle of analytics: from exploratory analysis and dashboard creation through to causal reasoning and AI-driven predictive modelling. By grounding the dataset in HR decision questions relevant to SMEs, the project underscores the strategic importance of workforce data in achieving organisational resilience, efficiency, and employee well-being.

In sum, this introduction establishes the context, purpose, and design principles underpinning the dataset. It situates the work at the intersection of academic learning and managerial practice, ensuring that the analysis is not only technically rigorous but also practically meaningful for organisations like El Hassan & El Hussein operating in resource-constrained and highly competitive environments.

2. Business Questions and Challenges to be answered

1. How do tenure and role seniority influence monthly employee productivity across functions and sites?
2. What is the relationship between utilisation, overtime, and absence, and how do seasonal peaks affect workforce fatigue?
3. How do the components of compensation evolve by role and department, and what is the implied cost per productivity point?
4. Which departments or sites exhibit persistently high utilisation and absence, signalling operational or workforce risk?

3. Dataset Description

- **Population**: Employees of El Hassan & El Hussein.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Unit of Analysis**: Employee–month record.
- **Time Granularity**: Monthly (eight months).
- **Sample Size**: 1,000 records (125 employees observed over 8 months).
- **Source Context**: Simulated HRIS extracts combined with recruitment and attendance trackers.

4. Purpose and Application

The dataset is intended to support **strategic HR management** through:

- Monitoring productivity and compensation trends.
- Identifying high-risk combinations of utilisation and absence.
- Assessing the effectiveness of allowances and overtime incentives.
- Enabling scenario analysis for workforce planning in Power BI.

5. Dataset Structure

The dataset includes **17 variables**:

- **7 qualitative** (categorical attributes such as department, role, contract type).
- **10 quantitative** (continuous or derived measures including pay, overtime, productivity, absence).

6. Qualitative Variables

1. **employee_id** – Unique employee identifier.
2. **gender** – Employee gender (Male/Female).
3. **age_band** – Age group (20–29, 30–39, 40–49, 50+).
4. **department** – Functional area (Production, Logistics, Sales, Operations, Finance, HR, IT).
5. **role_level** – Job seniority (Junior, Mid, Senior, Manager).
6. **contract_type** – Employment contract (Full-time, Part-time, Contract).
7. **location** – Worksite (Port Said, Ismailia, Damietta).

7. Quantitative Variables

1. **month_index** – Month (1–8).

© 2025 Dr. Yasser Aly Yakout Shehata

2. **tenure_months** – Months since employment start.
3. **base_pay_egp** – Base salary (Egyptian Pounds).
4. **overtime_hours** – Monthly overtime hours.
5. **overtime_pay_egp** (*Derived*) – 1.25 × hourly rate × overtime hours.
6. **allowances_egp** – Monthly allowances (e.g., site premiums).
7. **total_comp_egp** (*Derived, Identity*) – Base pay + overtime pay + allowances.
8. **utilisation_pct** – Workload intensity index (0.60–1.15).
9. **absence_days** – Days absent in the month (0–5).
10. **productivity_index** – Composite score of output and quality (40–120).

8. Derived Variables

1. **Absence_Rate_%** – (absence_days ÷ 22) × 100.
2. **OT_Intensity** – overtime_hours ÷ 160.
3. **Comp_Overtime_Share_%** – (overtime_pay_egp ÷ total_comp_egp) × 100.
4. **Cost_per_Productivity_Point** – total_comp_egp ÷ productivity_index.

9. Relevance to Power BI and AI for Business

The dataset is designed for interactive visual analytics in Power BI and predictive modelling through AI.

Suggested Power BI Visuals:

- **KPI cards**: Productivity Index, Absence Rate, Utilisation, Total Compensation, Overtime Pay.
- **Line charts**: Trends over time for utilisation, overtime, absence, and productivity.
- **Clustered bars**: Department × role vs productivity and compensation.
- **100% stacked bars**: Overtime share by role and department.
- **Scatter plots**: Utilisation vs absence (bubble size = overtime hours). **Slicers**: Month, Department, Role, Location, Gender, Contract Type.

10. Applications in Business Analytics and AI

A. **Exploratory Data Analysis (EDA):**

© 2025 Dr. Yasser Aly Yakout Shehata

Validation of distributions, seasonality, class balance, and internal identities; detection of anomalies and workforce outliers.

B. **Power BI Visualisation:**
Executive dashboards with productivity KPIs, absence monitoring, and cost-to-output ratios segmented by role, department, and site.

D. **AI-Driven Business Insights:**
Predictive models for absence and productivity, scenario simulation for workforce planning, and prescriptive insights for reducing fatigue risk and optimising pay structures.

11. Conclusion

This project delivers a **synthetic yet business-plausible HR dataset** that aligns with the decision-making needs of an SME in the manufacturing and logistics sector. The dataset is structured to support both **managerial dashboards** and **AI-based workforce analytics**, with clearly defined variables, derived indicators, and embedded accounting identities. Its educational value lies in demonstrating how HR data can be transformed into actionable insights, enabling organisations to balance efficiency, cost, and employee well-being.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 89: Port Said Container and Cargo Handling Company

1. Introduction

The maritime transport and logistics industry has emerged as a cornerstone of international trade, underpinning the flow of goods across regions and connecting producers and consumers in an increasingly interdependent global economy. Container terminals, in particular, serve as vital nodes in this system, functioning not only as operational hubs but also as competitive business entities that must balance efficiency, financial sustainability, and customer satisfaction. Within this complex environment, the General Manager of Commercial Affairs plays a pivotal role, integrating operational realities with strategic vision to ensure that a terminal remains competitive and responsive to market demands.

The **Port Said Container and Cargo Handling Company** (**PSCCHC**), located at the northern entrance of the Suez Canal, represents a compelling context for this type of analysis. Its geographical position provides shipping lines with considerable cost-saving opportunities by avoiding unnecessary deviations, thereby reinforcing its strategic importance. However, PSCCHC, like all ports, faces ongoing challenges in managing yard utilisation, vessel waiting times, equipment performance, and customer perceptions, while also responding to competitive pressures from regional ports such as Alexandria, Damietta, and East Port Said. These challenges underscore the need for data-driven decision-making to strengthen competitiveness and support long-term sustainability.

This project provides MBA students with a synthetic dataset that closely mirrors the data a General Manager of Commercial Affairs would utilise when managing such a terminal. The dataset contains one thousand records and incorporates variables spanning operational efficiency, throughput, equipment utilisation, customer satisfaction, and financial outcomes. By combining quantitative variables such as container volumes, crane rates, waiting times, and revenues with qualitative dimensions including satisfaction, reputation, and competitor strength, the dataset reflects the multifaceted nature of managerial decision-making.

To enrich analytical practice, the dataset also introduces derived variables such as revenue per TEU, utilisation bands, waiting time categories, and revenue bands. These variables enable students to conduct segmentation, benchmarking, and profitability analysis with greater ease, while simultaneously reflecting the type of derived indicators used in real-world management dashboards.

The value of this dataset lies in its dual academic and practical orientation. From an academic perspective, it enables students to develop proficiency in exploratory data analysis, descriptive statistics, and visualisation techniques. From a practical perspective,

© 2025 Dr. Yasser Aly Yakout Shehata

it provides opportunities to replicate the strategic insights and operational monitoring tasks undertaken by port managers. Its design ensures compatibility with modern analytical tools such as Power BI and supports applications in artificial intelligence for business, such as forecasting, classification, and predictive modelling.

Ultimately, this dataset not only equips MBA students with the technical capacity to manipulate and analyse data but also fosters their ability to contextualise findings within a business strategy framework. In doing so, it bridges the gap between data analytics as a technical exercise and business analytics as a managerial discipline, thereby preparing future leaders to harness data as a driver of competitive advantage in dynamic industries such as maritime logistics.

2. Business Questions and Challenges to be Answered

1. How does crane productivity influence vessel waiting times, and what impact does this have on customer satisfaction?
2. What are the effects of discount utilisation on revenue per TEU and overall profitability?
3. How do service types (standard, reefer, bulk, inland) differ in throughput contribution, operational complexity, and revenue generation?
4. How does competitor strength affect port reputation and expansion potential?
5. What trade-offs exist between high storage capacity utilisation and maintaining room for expansion?
6. How can past throughput and revenue data be used to forecast future performance trends?
7. To what extent can AI techniques predict shipping line satisfaction, reputation, or expansion potential from operational and financial data?

3. Dataset Description and Structure (Data Dictionary)

The dataset comprises **1,000 records** and **21 variables**, organised into three categories: quantitative, qualitative, and derived.

Quantitative Variables

1. **Monthly_TEU_Volume**
    - Type: Quantitative (integer)

© 2025 Dr. Yasser Aly Yakout Shehata

- Definition: Total number of twenty-foot equivalent units handled in a given month.
- Role: Primary indicator of terminal throughput and operational scale.

2. **Import_TEU**
    - Type: Quantitative (integer)
    - Definition: Number of imported containers handled.
    - Role: Reflects inbound trade flows and demand characteristics.
3. **Export_TEU**
    - Type: Quantitative (integer)
    - Definition: Number of exported containers handled.
    - Role: Indicates outbound trade flows and market positioning.
4. **Avg_Crane_Rate**
    - Type: Quantitative (float)
    - Definition: Average number of containers handled per crane per hour.
    - Role: Key efficiency measure of operational productivity.
5. **Ship_Waiting_Time**
    - Type: Quantitative (float)
    - Definition: Average time vessels wait before being berthed.
    - Role: Service-level metric directly influencing customer satisfaction.
6. **Revenue_EGP**
    - Type: Quantitative (float)
    - Definition: Total revenue earned within a month.
    - Role: Financial outcome reflecting operational and commercial performance.
7. **Storage_Capacity_Utilisation**
    - Type: Quantitative (percentage)
    - Definition: Proportion of available yard space used.

© 2025 Dr. Yasser Aly Yakout Shehata

- Role: Indicator of congestion and efficiency of capacity management.

8. **Reefer_Plugs_Used**
    - Type: Quantitative (integer)
    - Definition: Number of refrigerated container plugs occupied.
    - Role: Specialised measure relevant to perishable goods logistics.
9. **Equipment_Age**
    - Type: Quantitative (float)
    - Definition: Average age of container-handling equipment.
    - Role: Proxy for reliability, maintenance requirements, and capital planning.
10. **Deviation_Cost_Savings_USD_per_TEU**
    - Type: Quantitative (float)
    - Definition: Estimated savings per container due to the port's strategic location.
    - Role: Financial advantage indicator enhancing competitive attractiveness.

Qualitative Variables

11. **ShippingLine_Satisfaction**
    - Type: Qualitative (categorical: Low, Medium, High)
    - Definition: Customer satisfaction rating based on service experience.
    - Role: Measure of customer perception influencing retention and loyalty.
12. **Port_Reputation**
    - Type: Qualitative (categorical: Poor, Fair, Good, Excellent)
    - Definition: External perception of the port's performance and reliability.
    - Role: Strategic factor affecting competitiveness and market share.
13. **Competitor_Strength**
    - Type: Qualitative (categorical: Weak, Moderate, Strong)
    - Definition: Relative market strength of competing ports.

© 2025 Dr. Yasser Aly Yakout Shehata

- Role: Contextual factor guiding commercial strategy.

14. **Equipment_Reliability**
    - Type: Qualitative (categorical: Low, Medium, High)
    - Definition: Classification of equipment dependability derived from age and usage.
    - Role: Influences operational risk and continuity of service.
15. **Service_Type**
    - Type: Qualitative (categorical: Standard, Reefer, Bulk, Inland)
    - Definition: Category of service provided to customers.
    - Role: Basis for service segmentation and revenue analysis.
16. **Discount_Utilisation**
    - Type: Qualitative (categorical: None, Partial, Full)
    - Definition: Level of commercial discounts applied to customers.
    - Role: Variable influencing customer behaviour and revenue structure.
17. **Expansion_Potential**
    - Type: Qualitative (categorical: Low, Medium, High)
    - Definition: Assessment of potential for growth based on current utilisation and satisfaction.
    - Role: Forward-looking indicator for strategic planning.

Derived Variables

18. **Revenue_per_TEU**
    - Type: Quantitative (float)
    - Definition: Average revenue generated per container handled.
    - Role: Efficiency and profitability measure.
19. **Utilisation_Band**
    - Type: Qualitative (categorical: Low, Moderate, High, Very High)
    - Definition: Banding of storage capacity utilisation.

© 2025 Dr. Yasser Aly Yakout Shehata

- Role: Simplifies benchmarking and reporting of congestion.

20. **WaitingTime_Category**
    - Type: Qualitative (categorical: Very Low, Low, Medium, High)
    - Definition: Categorisation of vessel waiting times.
    - Role: Facilitates service-level analysis.
21. **Revenue_Band**
    - Type: Qualitative (categorical: Low, Medium, High, Very High)
    - Definition: Quartile-based classification of monthly revenue.
    - Role: Enables comparative financial analysis.

4. Purpose and Application

The dataset provides MBA students with a realistic analytical platform to practise data manipulation, exploration, and interpretation. It serves both educational and professional objectives: to develop technical competencies in data analysis and to contextualise these within managerial decision-making. By enabling students to replicate the challenges faced by a General Manager of Commercial Affairs, the dataset supports applications in throughput analysis, service benchmarking, capacity planning, financial evaluation, and competitive strategy.

9. Relevance to Power BI and AI for Business

The dataset is structured for seamless integration into Power BI, allowing for the construction of dashboards that combine operational KPIs, financial performance, and qualitative dimensions such as satisfaction and reputation. It also contains sufficient complexity for AI-driven applications, including predictive models for forecasting revenue, classification models for satisfaction and expansion potential, and clustering for identifying operational or customer segments. This dual suitability ensures its relevance for both traditional BI and advanced AI contexts.

10. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

Students can explore correlations, distributions, and anomalies in the dataset. For example, investigating whether higher crane productivity consistently reduces waiting times, or whether discount utilisation strongly affects revenue per TEU.

© 2025 Dr. Yasser Aly Yakout Shehata

B. Power BI Visualisation

The dataset supports interactive dashboards with KPI scorecards, trend charts, comparative analyses, and heatmaps. These dashboards allow students to communicate findings clearly and align data insights with managerial priorities.

C. AI-Driven Business Insights

The dataset can be used to develop machine learning models that predict satisfaction levels, classify revenue bands, or forecast throughput. Clustering methods may identify distinct service segments, while regression models can provide forecasts for financial and operational outcomes.

11. Conclusion

This project provides a comprehensive and academically rigorous dataset that situates MBA students in the role of a General Manager of Commercial Affairs at PSCCHC. By integrating operational, financial, and perceptual variables into a coherent framework, it enables both exploratory and advanced analysis. The dataset is not only suitable for Power BI dashboards and descriptive analytics but is also robust enough to support AI applications such as forecasting and predictive modelling. Ultimately, this resource bridges the gap between classroom learning and real-world practice, equipping students with the analytical and strategic competencies required to thrive in complex, data-driven industries such as maritime logistics.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 90: Badawi Payroll Employee

1. Introduction

The contemporary business environment increasingly recognises Human Resources as a strategic function rather than a purely administrative one. Within this domain, the role of the Payroll and Compensation and Employee Relations Officer has gained significance, as it embodies responsibilities that directly influence both financial performance and employee satisfaction. This officer must ensure that payroll is processed accurately, compensation structures are fair and competitive, allowances and overtime are appropriately distributed, training investments are monitored, and issues of absenteeism, performance, and turnover are carefully analysed. These functions require not only operational diligence but also analytical capacity to translate raw data into insights that can support organisational decision making.

The dataset presented in this project has been developed to provide MBA students with a practical learning resource while undertaking the course *Business Analytics for Decision Making*. It consists of one thousand records, each representing an employee within a hypothetical mid to large sized company operating across multiple departments such as Human Resources, Finance, Operations, Sales, Information Technology, and Logistics. It combines demographic, professional, financial, and performance related data, reflecting the multi dimensional nature of employee management. Crucially, it also incorporates derived fields that mirror the kinds of key performance indicators typically applied in workforce analytics, including measures such as absence rate, training effectiveness, and compensation efficiency.

By simulating the environment in which a Payroll and Compensation and Employee Relations Officer operates, the dataset enables students to practise a variety of analytical tasks. These include exploratory data analysis, descriptive statistics, interactive visualisation using Power BI, and the application of artificial intelligence tools to predict outcomes such as employee turnover or promotion. The dataset is structured in a way that highlights natural relationships between variables, for example the effect of training on performance, the link between satisfaction and turnover, and the role of tenure in promotion outcomes.

The value of this dataset lies not only in its comprehensiveness but also in its educational design. It encourages students to move beyond descriptive summaries and to engage in higher order analysis. It allows them to practise designing dashboards, evaluating trends, and generating insights that can inform managerial strategy. Most importantly, it demonstrates how Human Resources can evolve into a data driven function that provides

© 2025 Dr. Yasser Aly Yakout Shehata

actionable intelligence to support organisational resilience, employee wellbeing, and sustainable growth.

2. Business Questions and Challenges to be Answered

1. How do base salaries and total compensation vary by department, job level, and employment type
2. What is the relationship between training hours and employee performance, and how effective training across different departments is?
3. How does absenteeism influence performance outcomes, satisfaction levels, and turnover risk
4. What are the patterns of promotion across tenure and performance levels
5. Which factors are most strongly associated with employee turnover, and how can these be visualised to support retention policies
6. How do employee satisfaction levels correlate with compensation, training investment, and absenteeism

3. Dataset Description and Structure

The dataset contains **1,000 records** and **21 variables**, divided into quantitative, qualitative, and derived fields.

Data Dictionary

A. Quantitative Variables (10)

1. **Employee_ID**
    - Type: Quantitative (Integer)
    - Definition: Unique identifier assigned to each employee record
    - Role: Record key for identification and dataset integrity
2. **Age**
    - Type: Quantitative (Integer)
    - Definition: Age of the employee in years
    - Role: Demographic attribute relevant to tenure and workforce planning
3. **Years_of_Service**
    - Type: Quantitative (Float)

© 2025 Dr. Yasser Aly Yakout Shehata

- Definition: Number of years an employee has been employed by the organisation
- Role: Experience indicator and predictor of promotion or turnover

4. **Base_Salary**
    - Type: Quantitative (Float, EGP)
    - Definition: Fixed monthly salary excluding allowances and overtime
    - Role: Core compensation component for benchmarking and equity
5. **Allowances**
    - Type: Quantitative (Float, EGP)
    - Definition: Fixed monthly allowances such as housing and transport
    - Role: Supplementary compensation for total pay analysis
6. **Overtime_Hours**
    - Type: Quantitative (Float)
    - Definition: Number of overtime hours worked in a given month
    - Role: Workload measure with implications for satisfaction and productivity
7. **Overtime_Pay**
    - Type: Quantitative (Float, EGP)
    - Definition: Monetary value of overtime hours
    - Role: Variable compensation and productivity cost driver
8. **Absence_Days**
    - Type: Quantitative (Float)
    - Definition: Number of days absent from work in a year
    - Role: Attendance measure with links to performance and turnover
9. **Performance_Score**
    - Type: Quantitative (Float, scale 40–100)
    - Definition: Employee appraisal score
    - Role: Performance evaluation for career and compensation decisions
10. **Training_Hours**

- Type: Quantitative (Float)
- Definition: Total number of formal training hours completed in a year

© 2025 Dr. Yasser Aly Yakout Shehata

- Role: Human capital development indicator influencing performance

B. Qualitative Variables (7)

11. **Gender**

- Type: Qualitative (Nominal)
- Definition: Gender of the employee (Male or Female)
- Role: Demographic segmentation for equity and diversity analysis

12. **Department**

- Type: Qualitative (Nominal)
- Definition: Functional area of employment (HR, Finance, Operations, Sales, IT, Logistics)
- Role: Organisational segmentation and departmental comparison

13. **Job_Level**

- Type: Qualitative (Ordinal)
- Definition: Seniority classification (Junior, Mid, Senior, Manager)
- Role: Hierarchical segmentation for compensation and promotion

14. **Employment_Type**

- Type: Qualitative (Nominal)
- Definition: Type of contractual arrangement (Full-time, Part-time, Contract)
- Role: Workforce structuring and flexibility indicator

15. **Promotion_Status**

- Type: Qualitative (Nominal)
- Definition: Whether the employee was promoted during the year
- Role: Career progression outcome linked to performance and tenure

16. **Employee_Satisfaction**

- Type: Qualitative (Ordinal)
- Definition: Overall satisfaction level (Low, Medium, High)
- Role: Engagement indicator with implications for turnover

17. **Turnover_Status**

- Type: Qualitative (Nominal)
- Definition: Whether the employee remained Active or Resigned

© 2025 Dr. Yasser Aly Yakout Shehata

- Role: Retention outcome for attrition analysis

C. Derived Variables (4)

18. **Total_Compensation**

- Type: Quantitative (Float, EGP)
- Definition: Sum of base salary, allowances, and overtime pay
- Role: Comprehensive measure of overall pay

19. **Absence_Rate**

- Type: Quantitative (Float, ratio)
- Definition: Absence days expressed as a fraction of total days in a year
- Role: KPI for attendance, allowing comparisons across employees

20. **Salary_per_Performance_Point**

- Type: Quantitative (Float, ratio)
- Definition: Total compensation divided by performance score
- Role: KPI for evaluating cost-effectiveness of compensation relative to output

21. **Training_Effectiveness**

- Type: Quantitative (Float, ratio)
- Definition: Performance score divided by training hours
- Role: KPI for assessing returns on training investment

4. Purpose and Application

The primary purpose of the dataset is to provide MBA students with an authentic simulation of the analytical tasks that a Payroll and Compensation and Employee Relations Officer would face in practice. It allows students to apply statistical and analytical methods to address questions of salary benchmarking, absenteeism monitoring, training evaluation, and turnover analysis. The dataset also supports the development of dashboards and interactive reports, enabling students to practise communicating insights effectively to managerial audiences. Finally, it provides a foundation for applying artificial intelligence methods in HR, such as predictive modelling for attrition risk or promotion likelihood, thereby bridging theoretical coursework with real world practice.

9. Relevance to Power BI and AI for Business

Power BI is particularly suited to visualising the multi dimensional relationships in this dataset. Interactive dashboards can be designed to compare compensation across

© 2025 Dr. Yasser Aly Yakout Shehata

departments, monitor satisfaction trends, track absenteeism, and highlight turnover patterns. Slicers and filters can be used to examine results by gender, job level, or employment type, thereby supporting detailed exploration. In the context of AI for Business, the dataset provides the basis for applying classification algorithms to predict turnover, regression models to estimate performance outcomes, and clustering to segment employees by satisfaction and compensation profiles.

10. Applications in Business Analytics and AI

A. **Exploratory Data Analysis (EDA)**
   Students can examine distributions, correlations, and subgroup comparisons. Examples include analysing age and years of service by department, exploring salary variation across job levels, and testing the correlation between training hours and performance.
B. **Power BI Visualisation**
   Dashboards may feature bar charts of average compensation by department, line charts of absenteeism over time, scatter plots linking performance and training hours, and pie charts showing satisfaction levels. Slicers can filter results by gender, employment type, and promotion status.
C. **AI Driven Business Insights**
   The dataset can be used to build predictive models of employee turnover based on satisfaction, promotion status, and compensation. Regression models can estimate performance scores using training and absenteeism data. AI clustering techniques may identify distinct employee profiles to guide retention and reward policies.

11. Conclusion

The dataset ***payroll_employee.csv*** represents a structured and pedagogically designed learning resource for MBA students. By simulating the professional responsibilities of a Payroll and Compensation and Employee Relations Officer, it allows students to practise exploratory analysis, create Power BI dashboards, and apply artificial intelligence tools in a human resources context. With its blend of demographic, financial, performance, and satisfaction data, it provides a comprehensive platform for developing the analytical skills required for modern business decision making. The inclusion of derived variables ensures that students not only describe but also interpret and evaluate key performance indicators. In doing so, the dataset bridges academic theory and managerial practice, supporting the development of future leaders capable of applying business analytics for organisational success.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 91: EHA HR Training Analytics

1. Introduction

The *EHA HR Training Analytics Dataset* has been carefully constructed to replicate the operational realities of training and workforce development within the Egyptian Healthcare Authority, one of the central institutions driving the modernisation of healthcare delivery in Egypt. In recent years, healthcare systems worldwide have recognised the importance of investing in human capital to meet growing demands for quality, accessibility, and efficiency of services. Within this context, the Egyptian Healthcare Authority has been a national leader in adopting digital health innovations, such as electronic health records, alongside the large-scale implementation of the Universal Health Insurance system. These reforms have created significant demand for new types of training, capacity building, and performance evaluation, all of which are central responsibilities of Human Resources and Training Specialists working in the sector.

This dataset reflects that environment by simulating training activities across a range of professional categories, including doctors, nurses, technicians, administrative staff, and managers, while also embedding information about the regions in which staff are deployed and the modes through which training is delivered. Each record represents an individual training instance and captures variables such as participant demographics, professional background, training duration, costs, digital adoption levels, and measures of competency before and after training. By combining these elements, the dataset reproduces the multifaceted decision-making context of HR and Training Specialists who must balance budgets, ensure quality improvement, and support the digital transformation of healthcare organisations.

For MBA students undertaking the course *Business Analytics for Decision Making*, the dataset provides a practical resource to bridge theoretical knowledge and applied analysis. It has been designed to allow students to practice fundamental techniques such as exploratory data analysis, descriptive statistics, and correlation studies, while also offering opportunities to develop advanced applications through Power BI dashboards and artificial intelligence modelling. The inclusion of derived variables such as cost per hour, improvement per hour, efficiency index, and return on investment per training extends the analytical depth of the dataset and enables students to move beyond simple description towards strategic evaluation and forecasting.

The dataset is particularly relevant for a final project because it integrates multiple dimensions of organisational decision-making. Students can investigate how demographic characteristics influence training outcomes, assess which training modes yield the greatest

© 2025 Dr. Yasser Aly Yakout Shehata

improvements, examine regional variations in electronic health record adoption, and analyse the relationship between investment and returns in human capital development. These exercises mirror the work of HR and Training Specialists in the healthcare sector, who are required to transform raw data into actionable insights that support strategic planning, enhance workforce performance, and improve patient care outcomes.

Ultimately, the *EHA HR Training Analytics Dataset* has been developed to serve as a comprehensive learning tool. It provides a logically consistent, realistic, and analytically rich foundation that encourages MBA students to think critically about the role of data in business decision-making. It situates analytics within a high-stakes, real-world context where investments in human resources directly influence organisational success and social impact. By engaging with this dataset, students not only practice technical skills in analytics and visualisation but also develop an appreciation for the broader strategic and ethical implications of managing people, processes, and technologies in a complex healthcare system.

2. Business Questions and Challenges to be Answered

1. **Training Effectiveness**

- Which training types deliver the highest competency improvements?
- How do training modes (in-person, virtual, blended) influence both performance improvement and satisfaction?
- Are there diminishing returns for participants with higher years of experience or more previous trainings?

2. **Financial Efficiency**

- What is the relationship between training cost and performance improvement?
- Which training types or governorates yield the best return on investment per training?
- How does cost per hour vary across different training modes and types?

3. **Digital Transformation Impact**

- Does a higher level of electronic health record adoption correlate with faster time-to-proficiency or greater performance improvement?
- How do regional differences in EHR adoption affect training outcomes?
- Can digital readiness act as a predictor of training success?

4. **Satisfaction and Perceptions**

© 2025 Dr. Yasser Aly Yakout Shehata

- What factors most strongly influence participant satisfaction?
- Do satisfaction levels align with objective measures such as performance improvement or efficiency index?
- How does training mode affect perceived quality of the learning experience?

5. **Workforce Development Dynamics**

- How does job role influence training needs, outcomes, and satisfaction?
- Do younger or less experienced staff benefit more significantly from training compared to older, more experienced participants?
- What is the relationship between the number of previous trainings and the speed of proficiency after training?

6. **Operational Performance**

- Which governorates report the shortest time-to-proficiency and why?
- How does training duration affect the balance between improvement per hour and efficiency index?
- Can training schedules be optimised to achieve faster outcomes without incurring higher costs?

7. **Strategic Insights for HR and Training Specialists**

- What training programmes should be prioritised to maximise organisational value?
- How can resources be allocated to balance cost efficiency, digital adoption, and satisfaction?
- Which combinations of variables (training type, mode, governorate, job role) provide the most effective training outcomes?

Analytical Challenges Embedded in the Dataset

- **Trade-off analysis**: Balancing cost, duration, and improvement to identify optimal strategies.
- **Predictive modelling**: Using pre-training scores, experience, and contextual factors to forecast post-training outcomes.
- **Segmentation and clustering**: Grouping staff by satisfaction, efficiency, or ROI to support targeted interventions.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Comparative benchmarking**: Evaluating performance across governorates, job roles, and training modes.
- **Visual storytelling**: Transforming complex training and HR metrics into actionable insights through Power BI dashboards.

3. Dataset Description

The dataset consists of one thousand records, each representing a healthcare staff training instance. It incorporates seven qualitative variables, ten quantitative variables, and four derived variables. This structure allows the user to explore relationships between staff demographics, training investments, digital adoption, and training outcomes. The dataset mirrors the challenges faced in healthcare organisations where HR and Training Specialists must make data-informed decisions about workforce development and performance improvement.

4. Purpose and Application

The dataset has been developed with the following purposes:

- To provide MBA students with a realistic and analytically rich dataset for their final projects.
- To enable practical exercises in exploratory data analysis, descriptive and predictive analytics, and data visualisation.
- To demonstrate how training and HR data can support strategic decision-making in healthcare organisations.
- To integrate business intelligence and artificial intelligence approaches for improving workforce outcomes.

5. Dataset Structure

The dataset contains 21 variables grouped into three categories: qualitative, quantitative, and derived. Each record captures demographic, professional, and training details of a healthcare employee, together with associated costs, outcomes, and satisfaction levels. The dataset is logically consistent, free of missing values, and suitable for advanced analytical applications.

6. Qualitative Variables (7)

1. **Employee_ID**

- *Type*: Categorical, identifier

© 2025 Dr. Yasser Aly Yakout Shehata

- *Definition*: A unique alphanumeric code assigned to each participant (e.g., EHA-0001).
- *Role*: Serves as a primary key for distinguishing records; essential for record-keeping and ensuring that each observation is uniquely identifiable.

2. **Job_Role**

- *Type*: Categorical, nominal
- *Definition*: The professional role of the participant, including Dentist, Nurse, Administrative Staff, Technician, Manager, General Practitioner, or Specialist.
- *Role*: Enables segmentation and comparison of training outcomes across different professional categories.

3. **Gender**

- *Type*: Categorical, nominal
- *Definition*: Gender identity of the participant categorised as Male, Female, or Non-binary.
- *Role*: Provides demographic insights and allows analysis of gender-based differences in training participation, satisfaction, and outcomes.

4. **Governorate**

- *Type*: Categorical, nominal
- *Definition*: The region or administrative governorate where the participant is assigned, covering Port Said, Ismailia, Luxor, South Sinai, Suez, and Aswan.
- *Role*: Facilitates regional analysis and comparisons, particularly useful for understanding variations in digital adoption, costs, and training outcomes across geographic areas.

5. **Training_Type**

- *Type*: Categorical, nominal
- *Definition*: The category of training undertaken, such as Economic Evaluation and HTA, EHR/AI Systems, Clinical Skills, Administrative Management, Leadership, Value-Based Healthcare, or Financial Sustainability.
- *Role*: Central to evaluating which training areas deliver the highest performance improvements and returns on investment.

© 2025 Dr. Yasser Aly Yakout Shehata

6. **Training_Mode**

- *Type*: Categorical, nominal
- *Definition*: The method of delivery of training sessions, defined as In-person, Virtual, or Blended.
- *Role*: Important for comparing cost efficiency, participant satisfaction, and effectiveness across different modes of training delivery.

7. **Satisfaction_Level**

- *Type*: Categorical, ordinal
- *Definition*: The self-reported level of satisfaction expressed by participants, ranging from Very Unsatisfied to Very Satisfied.
- *Role*: Provides a subjective measure of training quality and is a critical variable for linking objective improvements with participant perceptions.

7. Quantitative Variables (10)

1. **Participant_Age**

- *Type*: Numeric, integer
- *Definition*: The age of the participant at the time of training, measured in years.
- *Role*: Allows analysis of how age influences training participation, competency improvements, and satisfaction.

2. **Years_of_Experience**

- *Type*: Numeric, continuous
- *Definition*: The total number of years of professional experience of the participant.
- *Role*: Provides context for skill levels prior to training and supports analysis of how experience affects improvement and learning speed.

3. **Number_of_Previous_Trainings**

- *Type*: Numeric, integer
- *Definition*: The number of training sessions attended by the participant prior to the recorded session.

© 2025 Dr. Yasser Aly Yakout Shehata

- *Role*: Captures training history, enabling analysis of how prior exposure to training influences competency gains and efficiency.

4. **Training_Duration_Hours**

- *Type*: Numeric, continuous
- *Definition*: The total duration of the training session measured in hours.
- *Role*: Provides a key resource variable for linking time investment to outcomes such as improvement and satisfaction.

5. **Training_Cost_EGP**

- *Type*: Numeric, integer
- *Definition*: The total cost of the training programme per participant in Egyptian pounds.
- *Role*: Serves as a financial input variable for cost-effectiveness and ROI analysis.

6. **Pre_Training_Competency_Score**

- *Type*: Numeric, continuous (0–100 scale)
- *Definition*: The competency score of the participant before the training session.
- *Role*: Provides a baseline measure against which improvement and effectiveness of training can be assessed.

7. **Post_Training_Competency_Score**

- *Type*: Numeric, continuous (0–100 scale)
- *Definition*: The competency score of the participant after completing the training.
- *Role*: Measures the direct outcome of training and allows calculation of performance improvement.

8. **Performance_Improvement_Score**

- *Type*: Numeric, continuous
- *Definition*: The gain in competency measured as the difference between post-training and pre-training scores.
- *Role*: A core outcome variable used to evaluate the effectiveness of training programmes.

© 2025 Dr. Yasser Aly Yakout Shehata

9. **Hospital_EHR_Adoption_Level**

- *Type*: Numeric, continuous (0–1 scale)
- *Definition*: A measure of the level of electronic health record adoption at the participant's workplace.
- *Role*: Reflects the broader organisational digital environment and enables analysis of how digital maturity supports or moderates training effectiveness.

10. **Time_to_Proficiency_Days**

- *Type*: Numeric, integer
- *Definition*: The number of days required after training for the participant to reach operational proficiency.
- *Role*: Serves as a time-to-outcome variable for assessing training speed and efficiency.

8. Derived Variables (4)

1. **Cost_Per_Hour**

- *Type*: Numeric, continuous
- *Definition*: The average cost of training per hour, calculated as Training Cost divided by Training Duration.
- *Role*: Allows evaluation of efficiency and financial comparability across different types and modes of training.

2. **Improvement_Per_Hour**

- *Type*: Numeric, continuous
- *Definition*: The amount of performance improvement achieved per training hour.
- *Role*: Provides an efficiency measure that links time investment to competency gains.

3. **Efficiency_Index**

- *Type*: Numeric, continuous
- *Definition*: A composite indicator of training effectiveness, calculated as performance improvement relative to time-to-proficiency, scaled for comparability.

© 2025 Dr. Yasser Aly Yakout Shehata

- *Role*: Enables benchmarking of participants and training types in terms of rapid skill acquisition and learning effectiveness.

4. **ROI_Per_Training**

- *Type*: Numeric, continuous
- *Definition*: A proxy measure for return on investment, calculated as performance improvement relative to cost.
- *Role*: Provides a strategic indicator for assessing whether training investments generate proportionate improvements.

9. Relevance to Power BI and AI for Business

The dataset is particularly suited for Power BI because it combines categorical and numerical variables that enable meaningful visualisation. It allows the construction of interactive dashboards where users can filter by governorate, job role, or training type and assess outcomes such as cost efficiency, satisfaction, or improvement. The presence of derived variables enhances depth for comparative visualisations and performance metrics. Furthermore, the dataset provides a testing ground for AI applications, such as predictive modelling of post-training competency scores or clustering staff groups based on satisfaction and efficiency outcomes.

10. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA)

- Analyse distributions of demographic and professional variables.
- Identify correlations between training duration, cost, and improvement.
- Compare training modes and their impact on satisfaction and performance.

B. Power BI Visualisation

- Dashboards showing cost efficiency by training type and mode.
- Heatmaps linking governorates with adoption of electronic health records and training effectiveness.
- Drill-down reports on satisfaction and performance improvement by job role.

C. AI-Driven Business Insights

- Predictive models estimating post-training competency from pre-training scores and contextual variables.

© 2025 Dr. Yasser Aly Yakout Shehata

- Clustering of staff into groups based on improvement, cost, and satisfaction to inform targeted training strategies.
- Simulation of cost and time scenarios to optimise training resource allocation.

11. Conclusion

The *EHA HR Training Analytics Dataset* provides MBA students with a structured, realistic, and analytically rich resource to develop their skills in business analytics and decision-making. By integrating qualitative, quantitative, and derived variables, it enables exploration of training outcomes from multiple perspectives. Its relevance to healthcare HR practices, combined with its suitability for Power BI and AI-driven analysis, ensures that students can gain practical experience in developing insights, designing visualisations, and applying analytics to real-world business challenges.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 92: Student Affairs Performance and Engagement

The **Student Affairs Engagement and Performance Dataset** has been purposefully created to provide MBA students with a realistic and analytically rich resource for practicing advanced business analytics techniques. Student Affairs is one of the most critical functions in any educational institution, responsible for ensuring student well-being, monitoring academic achievement, addressing behavioral issues, and fostering engagement through extracurricular activities. This domain involves a blend of operational management and strategic decision-making, making it an ideal area for applying business analytics and AI-driven insights.

The dataset mirrors the complexity of data handled by Student Affairs professionals. It integrates demographic information, academic performance measures, attendance and behavior records, parental engagement levels, extracurricular involvement, and institutional interventions. To ensure realism, logical dependencies have been embedded. For instance, higher absenteeism is often correlated with lower GPA, weaker behavioral ratings, and a higher probability of dropout. Similarly, strong parental engagement and extracurricular participation tend to improve engagement scores, satisfaction survey results, and overall well-being.

By including both raw and derived variables, the dataset enables students to practice the full spectrum of analytical tasks. Raw variables such as age, GPA, and disciplinary actions provide the foundation for descriptive statistics and trend analysis. Derived variables such as risk scores, dropout probabilities, and wellbeing indexes simulate institutional metrics used in real decision-making processes. This structure empowers MBA students to explore not just “what happened” but also “why it happened” and “what is likely to happen in the future.”

The dataset is designed for multiple layers of learning. Through **Exploratory Data Analysis (EDA)**, students can identify distributions, correlations, and anomalies. With **Power BI dashboards**, they can transform the dataset into compelling visual stories that guide decision-making. For **AI in Business applications**, students can apply predictive modeling and machine learning techniques to forecast dropout risk, segment students, or assess the effectiveness of interventions.

Ultimately, the Student Affairs Engagement and Performance Dataset bridges the gap between theory and practice. It equips MBA students to analyze institutional challenges, derive insights that influence policy, and develop strategic recommendations rooted in data. By engaging with this dataset, students will sharpen their ability to use analytics not only to

© 2025 Dr. Yasser Aly Yakout Shehata

describe and diagnose but also to predict and prescribe, thereby strengthening their capacity as future decision-makers in data-driven organizations.

2. Business Questions and Challenges to be Answered

1. How does absenteeism impact GPA and academic standing?
2. What is the relationship between parental engagement and student performance?
3. Which factors most strongly predict dropout probability?
4. How do extracurricular activities influence student engagement and well-being?
5. What patterns emerge in disciplinary actions across different grade levels?
6. Which segments of students are at the highest academic or behavioral risk?
7. How effective are institutional interventions in reducing dropout likelihood?
8. Can high-performing students be reliably identified using engagement and behavior indicators?
9. What is the relationship between satisfaction surveys and overall academic outcomes?
10. How can resources be best allocated to support students requiring interventions?

3. Dataset Description

- **Size:** 1000 student records.
- **Coverage:** Demographic, academic, behavioral, engagement, and support-related data.
- **Composition:** 7 qualitative variables, 10 quantitative variables, and 13 derived variables.
- **Strength:** Incorporates business logic to simulate realistic institutional data and enable advanced analytics.

4. Purpose and Application

- To support MBA students in practicing **descriptive, diagnostic, predictive, and prescriptive analytics**.
- To demonstrate the value of **Power BI dashboards** in institutional decision-making.

© 2025 Dr. Yasser Aly Yakout Shehata

- To enable the application of **AI techniques in business**, such as predictive modeling and clustering.
- To foster critical thinking in connecting raw data with strategic student affairs challenges.

5. Dataset Structure

The dataset is divided into three types of variables:

- **Qualitative (categorical) variables** representing student classification and engagement characteristics.
- **Quantitative (numerical) variables** measuring attendance, performance, and satisfaction.
- **Derived variables** reflecting institutional performance indicators, risk assessments, and well-being measures.

6. Qualitative Variables (Type; Description; Role)

1. **Gender** – Nominal; identifies student gender; supports demographic segmentation.
2. **Nationality** – Nominal; country/region classification; used for diversity and equity analysis.
3. **Grade Level** – Ordinal; stage of schooling; links to age, GPA, and attendance patterns.
4. **Attendance Status** – Ordinal; regularity of attendance; highlights risk categories.
5. **Behavior Rating** – Ordinal; conduct assessment; supports behavioral analytics.
6. **Club Participation** – Nominal; extracurricular involvement; linked to engagement.
7. **Parent Engagement** – Ordinal; level of parental support; critical predictor of success.

7. Quantitative Variables (Type; Description; Role)

1. **Age** – Continuous; student age in years; demographic baseline.
2. **Days Absent** – Discrete; number of absences; key academic risk driver.
3. **Tardiness Count** – Discrete; late arrivals; correlated with absenteeism.
4. **GPA** – Continuous; academic performance (0–4 scale); central outcome metric.
5. **Disciplinary Actions** – Discrete; number of recorded offenses; indicator of conduct.

© 2025 Dr. Yasser Aly Yakout Shehata

6. **Extracurricular Hours** – Continuous; time spent in activities; engagement measure.
7. **Counseling Sessions** – Discrete; sessions attended; reflects support needs.
8. **Engagement Score** – Continuous; composite score (0–100); student involvement.
9. **Satisfaction Survey** – Ordinal; Likert 1–5; student experience measure.
10. **Support Interventions** – Discrete; count of interventions; institutional support effort.

8. Derived Variables (Type; Description; Role)

1. **Attendance Rate** – Continuous; % attendance; key operational KPI.
2. **Behavior Points** – Continuous; score (0–100); quantitative behavior measure.
3. **Engagement Level** – Ordinal; Low–Very High; segmentation of engagement.
4. **Academic Standing** – Ordinal; Honors–At Risk; outcome classification.
5. **Risk Score** – Continuous; 0–100 composite risk measure.
6. **Risk Tier** – Ordinal; Low–High; simplified risk category.
7. **Dropout Probability** – Continuous; 0–1; predictive likelihood of dropout.
8. **Intervention Needed** – Binary; Yes/No; action flag.
9. **Intervention Intensity** – Ordinal; Low–High; level of required support.
10. **Parental Contact Frequency** – Continuous; monthly contacts; family engagement.
11. **Certificates Issued** – Discrete; 0–3; recognition of achievement.
12. **Wellbeing Index** – Continuous; 0–100; holistic student wellness measure.
13. **High Performer Flag** – Binary; Yes/No; identifies top achievers.

9. Relevance to Power BI and AI for Business

- In **Power BI**, the dataset supports dynamic dashboards, KPI tracking, and drill-downs into demographics, grade levels, and engagement categories.
- For **AI in Business**, it provides predictors and target variables for models in risk prediction, segmentation, and intervention optimization.

10. Suggested Power BI Dashboard Designs

1. **Academic Performance Dashboard** – GPA distribution by grade level, gender, and nationality.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Attendance and Risk Dashboard** – Attendance rate vs. risk score, with filters for parental engagement.
3. **Engagement and Well-being Dashboard** – Engagement levels, satisfaction survey scores, and wellbeing index visualized across student groups.
4. **Discipline and Intervention Dashboard** – Disciplinary actions linked to interventions and dropout probability.
5. **Comprehensive Student Affairs Dashboard** – Executive overview combining GPA, attendance, risk tier, and engagement in a single decision-support tool.

Each dashboard should allow filtering by grade, nationality, gender, and parental engagement to create interactive, insight-rich exploration.

11. Sample Business Insights

- Students with chronic absences are three times more likely to have GPAs below 2.0 and belong to the "At Risk" academic standing category.
- High parental engagement reduces dropout probability by up to 25 percent across grade levels.
- Participation in science and sports clubs is strongly correlated with higher engagement scores and wellbeing index values.
- Disciplinary actions significantly increase risk scores, with a threshold effect beyond three incidents.
- Students flagged as "High Performers" consistently show both high GPA and very high engagement, making them prime candidates for leadership and recognition programs.

12. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA): Identify correlations, distributions, and anomalies in GPA, attendance, and satisfaction.

B. Power BI Visualization: Build dashboards for decision-makers to monitor performance, engagement, and interventions.

C. AI-Driven Business Insights: Apply predictive models to estimate dropout risk, cluster students by engagement and risk, and simulate the effect of interventions on academic outcomes.

© 2025 Dr. Yasser Aly Yakout Shehata

13. Conclusion

The **MBA Business Analytics Project Dataset: Student Affairs Engagement and Performance** provides a comprehensive, realistic, and analytically rich resource for MBA students. It integrates raw and derived variables, enabling descriptive, diagnostic, predictive, and prescriptive analytics. Through EDA, visualization, and AI-driven modeling, the dataset equips students to address key institutional challenges, derive actionable insights, and recommend strategies that improve student outcomes and optimize resource allocation. It stands as a bridge between academic training and professional practice in data-driven decision-making.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 93: E-Commerce Brand Manager Instagram

1. Introduction

The E-Commerce Brand Manager Instagram Analytics Dataset has been developed to replicate the analytical environment of a real-world fashion and apparel e-commerce business operating in the digital economy. The dataset positions MBA students in the role of a Founder and E-Commerce Brand Manager, who must oversee the management of social media campaigns, customer engagement, product pricing, and profitability while balancing growth and efficiency. It is designed to support the course *Business Analytics for Decision Making*, where students gain hands-on experience in data analysis, visualization, and the use of artificial intelligence in business contexts.

The modern e-commerce landscape is characterized by its reliance on digital platforms, especially social media, as a core driver of customer acquisition and engagement. Instagram has emerged as a central channel for building brand awareness, generating leads, and converting potential customers into loyal buyers. The dataset captures this ecosystem by modeling campaign types, customer segments, regions, and product attributes, alongside quantitative measures of impressions, reach, engagement, sales, and revenue.

The dataset is carefully engineered with embedded business logic to ensure analytical authenticity. For example, paid campaigns typically result in higher impressions and ad spend, stories generate more direct messages, and VIP customers show higher conversion rates than new customers. Similarly, material type impacts cost of goods sold, thereby influencing profit margins, while different product categories are associated with varied pricing strategies and average order values. These interdependencies allow students to explore not only what the numbers say but why they behave in certain ways.

A key strength of this dataset is its inclusion of derived variables that provide analytical depth, such as engagement rate, direct message to order conversion, average order value, return on ad spend, and customer lifetime value. These variables bridge the gap between operational data and managerial insight, enabling students to practice advanced analysis and apply their findings to real-world scenarios. The temporal dimension provided by post date, which includes both date and time components, adds further richness, allowing time-series analysis, forecasting, and exploration of time-of-day posting effectiveness.

By engaging with this dataset, MBA students will practice techniques such as exploratory data analysis, dashboard design in Power BI, and descriptive analytics, as well as AI applications such as forecasting, clustering, and predictive modeling. The dataset provides a structured but challenging platform that replicates the complexities faced by modern e-

© 2025 Dr. Yasser Aly Yakout Shehata

commerce managers. It empowers students to transform raw data into actionable business insights and to practice evidence-based decision-making.

2. Business Questions and Challenges to be Answered

1. Which content types generate the highest engagement rates and conversion outcomes
2. How do paid versus organic campaigns differ in impressions, conversions, and profitability
3. Which customer segments contribute most to revenue, profit, and lifetime value
4. How do material types influence cost of goods sold and profit margins
5. What is the efficiency of ad spend across product categories and campaign types
6. Which regions show the highest engagement and conversion rates
7. How do posting times and dates affect customer engagement and sales outcomes
8. What variables best predict customer lifetime value
9. How should campaigns be optimized to maximize profit while minimizing acquisition costs

3. Dataset Description

The dataset consists of 1,000 records, each representing a single Instagram campaign post by an e-commerce fashion brand. It integrates categorical, numerical, derived, and temporal fields to create a comprehensive analytical view of business performance. Records capture campaign activity, audience engagement, customer responses, and financial outcomes. Temporal variables provide a time dimension for studying trends, seasonal patterns, and time-of-day effects.

4. Purpose and Application

The dataset is designed for MBA students studying business analytics. It supports:

- **Exploratory Data Analysis (EDA):** Identifying relationships and patterns.
- **Visualization in Power BI:** Building dashboards for campaign monitoring and profitability tracking.
- **AI in Business:** Applying predictive models, clustering, and forecasting to generate insights for decision-making.

© 2025 Dr. Yasser Aly Yakout Shehata

5. Dataset Structure

- **Records:** 1,000
- **Qualitative Variables:** 7 categorical
- **Quantitative Variables:** 10 core numerical
- **Derived Variables:** 12 advanced business metrics
- **Temporal Variables:** Post Date as datetime (includes both date and time)

6. Qualitative Variables

1. **Brand Page –** Categorical – Brand account identifier – Role: Analytical segmentation
2. **Product Category –** Categorical – Product type (menswear, womenswear, shoes, bags) – Role: Product analysis
3. **Material Type –** Categorical – Product material (cotton, polyester, leather, etc.) – Role: Cost driver
4. **Content Type –** Categorical – Instagram post type (post, reel, story, carousel) – Role: Engagement driver
5. **Campaign Type –** Categorical – Organic or paid – Role: Marketing channel dimension
6. **Customer Segment –** Categorical – New, returning, or VIP – Role: Conversion and CLV segmentation
7. **Region –** Categorical – Local, GCC, or international – Role: Market segmentation

7. Quantitative Variables

1. **Followers –** Integer – Account audience size – Role: Reach potential
2. **Reach –** Integer – Unique viewers – Role: Campaign exposure
3. **Impressions –** Integer – Total views – Role: Visibility measure
4. **Likes –** Integer – User likes – Role: Engagement indicator
5. **Comments –** Integer – User comments – Role: Engagement indicator
6. **Saves –** Integer – Post saves – Role: Engagement indicator
7. **Direct Messages –** Integer – Inquiries – Role: Lead generation measure
8. **Orders – Integer –** Completed purchases – Role: Sales performance

© 2025 Dr. Yasser Aly Yakout Shehata

9. **Unit Price –** Float – Product price – Role: Revenue driver
10. **Revenue –** Float – Sales revenue – Role: Profitability foundation

8. Derived Variables

1. **Engagement Rate –** Float – Interactions as percentage of impressions – Role: Engagement KPI
2. **Direct Message to Order Conversion –** Float – Orders per direct message – Role: Conversion efficiency
3. **Average Order Value –** Float – Revenue per order – Role: Pricing performance
4. **Return on Ad Spend –** Float – Revenue per advertising spend – Role: ROI analysis
5. **Follower Growth Rate –** Float – Monthly growth – Role: Audience expansion metric
6. **Customer Lifetime Value Estimate –** Float – Predicted customer worth – Role: Strategic planning
7. **Cost per Acquisition –** Float – Ad spend per order – Role: Efficiency measure
8. **Cost of Goods Sold –** Float – Cost of production – Role: Profit calculation
9. **Profit –** Float – Revenue minus costs – Role: Net performance measure
10. **Profit Margin Percent –** Float – Profit percentage of revenue – Role: Sustainability measure
11. **Revenue per Follower –** Float – Monetization efficiency – Role: Audience productivity
12. **Post Date –** Datetime – Full timestamp of post (date and time) – Role: Temporal trend and time-of-day analysis

9. Relevance to Power BI and AI for Business

The dataset supports the creation of dashboards and KPIs that track campaign performance and profitability. It enables cross-sectional analysis (by segment, product, or region) and longitudinal analysis (over time). For AI applications, the dataset provides input variables that can feed predictive and classification models, helping students forecast outcomes such as conversions or profitability.

10. Suggested Power BI Dashboard Designs

1. **Marketing Funnel Dashboard** – Tracks the journey from impressions to engagement, direct messages, and orders.

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Profitability Dashboard** – Visualizes revenue, cost, and profit across categories and campaigns.
3. **Customer Insights Dashboard** – Segments customers by type and lifetime value to inform retention strategies.
4. **Content Performance Dashboard** – Compares engagement across content formats and times of day.
5. **Time-Series Dashboard** – Analyzes campaign performance across the year and by hour of posting.

11. Sample Business Insights

- Paid campaigns achieve higher impressions but profitability varies with product category and spend efficiency.
- VIP customers drive the highest lifetime value and order conversions, justifying loyalty strategies.
- Stories produce the most direct messages, making them valuable for lead generation.
- Leather products deliver higher revenue but also higher costs, creating margin trade-offs.
- Posting time influences engagement: evening posts generate higher interaction rates.

12. Applications in Business Analytics and AI

A. **Exploratory Data Analysis (EDA):** Identify patterns and outliers in engagement, conversion, and profitability.

B. **Power BI Visualization:** Build interactive dashboards to track KPIs, customer journeys, and regional performance.

C. **AI-Driven Insights:** Apply clustering for customer segmentation, regression for revenue prediction, and forecasting for time-based sales trends.

13. Conclusion

The E-Commerce Brand Manager Instagram Analytics Dataset is a comprehensive resource that combines real-world logic, diverse variable types, and advanced derived metrics. It is purpose-built to support MBA students in mastering exploratory analysis, dashboard creation, and AI-driven business decision-making. By working with this dataset, students can bridge the gap between classroom learning and professional practice, gaining the skills necessary to manage and grow data-driven e-commerce brands.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 94: Port Said Dental Analytics

1. Introduction

The **Port Said Dental Analytics Dataset** has been developed as a pedagogical resource to support MBA students in the course *Business Analytics for Decision Making*. It simulates the environment of a dentist working at the Specialised Dental Centre in Port Said, a facility operating within Egypt's Universal Health Insurance system. The dataset captures the multifaceted responsibilities of a dentist who is simultaneously a healthcare provider, service manager, and financial decision-maker. By embedding both clinical and business dimensions, the dataset enables students to investigate issues that are at the intersection of healthcare delivery, operational efficiency, and financial management.

The dataset is designed to reflect the complexity of real-world healthcare operations. It incorporates patient demographics, treatment types, insurance coverage, service utilisation, and clinical outcomes. Importantly, it includes derived variables that enhance analytical depth, such as cost-to-revenue relationships, satisfaction indicators, and operational efficiency measures. This ensures that students do not merely analyse isolated variables, but can also examine the interplay between service quality, financial sustainability, and patient experience.

Within the healthcare industry, dental clinics face persistent challenges such as managing patient demand, minimising wait times, reducing no-show appointments, optimising staff allocation, and balancing treatment costs with reimbursement policies. This dataset provides a synthetic yet realistic platform for students to engage with these challenges through structured analysis. It highlights the practical application of business analytics tools in addressing problems that have direct implications for service quality and organisational performance.

Beyond traditional descriptive analysis, the dataset is explicitly designed to be relevant for **Exploratory Data Analysis (EDA)**, **Power BI visualisation**, and the **application of artificial intelligence in business**. Students can create interactive dashboards to visualise performance trends, identify operational bottlenecks, and evaluate financial outcomes. They can also apply AI-driven methods, such as predictive modelling, to forecast patient behaviours, predict treatment outcomes, and assess the financial impact of strategic decisions.

By combining clinical, operational, and financial variables, this dataset reflects a realistic analytical environment for MBA students. It serves not only as a technical training resource but also as a conceptual bridge, illustrating how data-driven insights can improve decision-making in complex service industries such as healthcare.

© 2025 Dr. Yasser Aly Yakout Shehata

2. Business Questions and Challenges to be Answered

1. What factors influence patient satisfaction, and how can the clinic improve patient experience?
2. How do waiting times and travel distances affect no-show and cancellation rates?
3. Which treatment types generate the highest costs, revenues, and insurance reimbursements?
4. What is the relationship between treatment complexity and operational efficiency?
5. How does insurance status affect patient access, financial burden, and satisfaction?
6. What are the primary drivers of treatment outcomes and follow-up needs?
7. How can staffing and inventory usage be optimised to improve efficiency and reduce costs?
8. What operational patterns can be identified to support predictive modelling and AI applications?

3. Dataset Description

The dataset consists of **1,000 records**, representing patients who have interacted with the Specialised Dental Centre. It includes **seven qualitative variables**, **ten quantitative variables**, and a set of **derived variables**. The data integrates demographic, operational, financial, and clinical perspectives, providing a holistic view of dental healthcare services.

4. Purpose and Application

The dataset has been developed to:

- Support MBA students in practising analytical techniques in healthcare operations.
- Enable the use of Power BI for dashboard creation and interactive reporting.
- Provide a structured dataset for AI-driven applications such as predictive modelling.
- Encourage decision-making analysis based on both clinical outcomes and financial sustainability.

5. Dataset Structure

- **Records:** 1,000
- **Variable Types:** Qualitative (categorical), Quantitative (numeric), and Derived (analytical indicators)

© 2025 Dr. Yasser Aly Yakout Shehata

- **Domains Covered:** Patient demographics, treatment types, operational performance, clinical outcomes, financial indicators, satisfaction levels

6. Qualitative Variables

1. **Patient_ID** – Categorical, unique identifier; role: record tracking
2. **Gender** – Categorical, patient gender; role: demographic segmentation
3. **Treatment_Type** – Categorical, type of dental service received; role: service categorisation
4. **Insurance_Status** – Categorical, insured or uninsured; role: financial grouping
5. **Appointment_Status** – Categorical, attended, cancelled, or no-show; role: service reliability indicator
6. **Treatment_Outcome** – Categorical, successful, complication, or follow-up required; role: quality measure
7. **Satisfaction_Level** – Categorical, low, medium, or high; role: experience and service evaluation

7. Quantitative Variables

1. **Age** – Integer; role: demographic analysis
2. **Wait_Time_Minutes** – Integer; role: operational performance measure
3. **Treatment_Duration_Minutes** – Integer; role: clinical workload assessment
4. **Number_of_Visits** – Integer; role: service utilisation tracking
5. **Cost_of_Treatment** – Float; role: financial performance measurement
6. **Inventory_Usage_Units** – Integer; role: operational resource consumption
7. **Follow_Up_Days** – Integer; role: clinical outcome tracking
8. **Travel_Distance_km** – Float; role: geographical accessibility
9. **Staff_Involved** – Integer; role: resource allocation measure
10. **Reimbursement_Amount** – Float; role: insurance financial contribution

8. Derived Variables

1. **Age_Group** – Categorical; role: age segmentation

© 2025 Dr. Yasser Aly Yakout Shehata

2. **Travel_Band** – Categorical; role: distance segmentation
3. **On_Time** – Binary; role: service efficiency indicator
4. **No_Show_Flag** – Binary; role: patient behaviour classification
5. **Outcome_Success** – Binary; role: quality indicator
6. **Complexity_Score** – Integer; role: treatment intensity classification
7. **Complication_Flag** – Binary; role: risk classification
8. **Net_Revenue** – Float; role: financial sustainability measure
9. **Coverage_Ratio** – Float; role: insurance coverage indicator
10. **Inventory_Cost_Est** – Float; role: operational cost estimate
11. **Total_Clinic_Time_Min** – Integer; role: time efficiency measure
12. **Satisfaction_Score** – Integer; role: numerical representation of satisfaction

9. Relevance to Power BI and AI for Business

The dataset is particularly suited to Power BI as it enables the creation of interactive dashboards linking clinical, operational, and financial data. AI applications include predictive analytics for patient behaviour, outcome forecasting, and financial modelling. The dataset provides a training environment where students can explore how business analytics informs managerial and operational decision-making in healthcare.

10. Suggested Power BI Dashboard Designs

1. **Patient Demographics Dashboard** – Displays age groups, gender, travel bands, and insurance status distributions.
2. **Operational Performance Dashboard** – Tracks waiting times, treatment durations, staff allocation, and no-show rates.
3. **Financial Dashboard** – Highlights treatment costs, reimbursement amounts, net revenue, and coverage ratios.
4. **Clinical Outcomes Dashboard** – Visualises treatment outcomes, complication rates, and follow-up days.
5. **Satisfaction Dashboard** – Links waiting times, financial burden, and outcomes to satisfaction levels.

© 2025 Dr. Yasser Aly Yakout Shehata

6. **Integrated Executive Dashboard** – Combines key indicators across demographics, operations, finance, and outcomes for strategic decision-making.

11. Sample Business Insights

1. Long waiting times significantly increase the probability of no-shows and reduce satisfaction.
2. Insured patients show higher satisfaction due to lower financial burden.
3. Complex treatments such as implants and braces generate higher costs but also more complications and follow-up requirements.
4. Inventory usage strongly correlates with treatment complexity and financial performance.
5. Staff allocation per treatment has measurable effects on efficiency and patient outcomes.

12. Applications in Business Analytics and AI

A. Exploratory Data Analysis (EDA): Identifying patterns, distributions, and correlations among variables.

B. Power BI Visualisation: Developing interactive dashboards that track performance and highlight areas for improvement.

C. AI-Driven Business Insights: Applying predictive models to forecast no-shows, predict treatment outcomes, and simulate financial impacts of policy changes.

13. Conclusion

The **Port Said Dental Analytics Dataset** represents a unique intersection of healthcare and business analytics. It provides MBA students with a robust resource to practise EDA, create Power BI dashboards, and apply AI-driven insights to real-world challenges. By embedding business logic and derived variables, the dataset offers analytical depth and relevance, supporting the development of both technical and managerial competencies. It demonstrates how structured data, when effectively analysed, can lead to actionable strategies that improve service quality, operational efficiency, and financial sustainability in healthcare.

© 2025 Dr. Yasser Aly Yakout Shehata

Dataset 95: Petrojet Coating Finance Analytics Gold

1. Introduction

The **Petrojet Coating Finance Analytics Gold Dataset** has been developed as a comprehensive, practice-oriented resource for MBA students specializing in Business Analytics, Finance, and AI for Business. It simulates the financial and operational dynamics of Petrojet's coating plants, with a special focus on the Port Said and Suez facilities. These plants are central to Petrojet's operations, undertaking complex coating and construction projects where profitability, efficiency, and cash flow are critical for sustainable performance.

This dataset places the student in the role of a **Financial Accounts Specialist**, a professional responsible for translating engineering progress into recognised revenue, analysing cost behaviour, monitoring receivables, forecasting financial outcomes, and reporting results to management. By mirroring the responsibilities of this role, the dataset ensures that MBA students can apply analytical methods to real-world problems, linking data interpretation with strategic decision-making.

Each record represents a **monthly observation** at the plant level. Within each period, physical progress drives revenue recognition, direct and overhead costs determine margins, and invoicing and collection patterns affect cash flow. Time-series data allow for longitudinal analysis of trends, while categorical dimensions such as contract type, revenue category, and payment status provide segmentation for deeper insights. The inclusion of derived fields such as gross margin, collection rate, budget attainment, and working capital impact further enhances the dataset's analytical depth.

The dataset is particularly suited for **Exploratory Data Analysis (EDA)** and **visualization with Power BI**, as it offers multiple dimensions for slicing and dicing performance data. Its logical structure supports a clean star-schema model, making it easy to build layered dashboards that progress from executive summaries to diagnostic deep-dives.

Beyond visualization, the dataset is tailored for applications in **AI for Business**. Derived features and diagnostic flags allow students to build baseline predictive models for overdue invoices, budget overruns, or margin erosion. Thus, the dataset not only supports descriptive analytics but also enables prescriptive and predictive applications, bridging the gap between classroom exercises and industry-level analytics.

For MBA students completing their **final projects**, this dataset provides a robust foundation to demonstrate competence in financial analytics, Power BI dashboarding, and AI-driven business insights. It encourages students to tell a full managerial story: from project scope

© 2025 Dr. Yasser Aly Yakout Shehata

and operational progress, through cost and profitability analysis, to receivables management and strategic budget control.

2. Business Questions and Challenges to be Answered

1. What operational and financial factors drive margin variability across plants, clients, and categories?
2. How do invoicing and client behaviours influence receivables ageing and cash conversion?
3. What patterns explain budget deviations, and how can these be predicted or prevented?
4. How do scale and capital intensity affect cost efficiency across plants?
5. Which contract types or revenue categories carry higher risk of margin compression?
6. What leading indicators signal overdue invoices and potential collection risks?
7. How can financial performance be communicated effectively through dashboards for executives?

3. Dataset Description

The dataset includes **1,000 monthly plant-period records** from four operational sites: Port Said, Suez, Alexandria Workshop, and Ismailia Yard. Each record integrates:

- **Qualitative identifiers** such as plant, client, and contract type.
- **Quantitative measures** such as contract value, progress, revenue, costs, and receivables.
- **Derived indicators** for profitability, liquidity, efficiency, and managerial risk.

4. Purpose and Application

- To practice rigorous **EDA** on industrial financial data.
- To design **Power BI dashboards** that convey complex financial and operational insights.
- To conduct **descriptive analytics** on revenue, margin, and cash outcomes.
- To explore **AI applications in business**, including forecasting, risk scoring, and predictive modelling.

© 2025 Dr. Yasser Aly Yakout Shehata

5. Dataset Structure

- **Qualitative variables (7):** Period, Plant, Region, Client, Contract_Type, Revenue_Category, Payment_Status
- **Quantitative variables (10):** Contract_Value_MEGP, Physical_Progress, Revenue_Recognised_MEGP, Direct_Cost_MEGP, Overhead_Cost_MEGP, Capex_MEGP, Invoice_Amount_MEGP, Collected_Amount_MEGP, Receivables_Balance_MEGP, AR_Days
- **Derived variables (39):** Budget, forecast, margin, cost ratios, liquidity indicators, status flags, bandings, time intelligence fields, and surrogate keys.

6. Qualitative Variables

1. **Period** – categorical text; monthly accounting period; time slicer and join key.
2. **Plant** – categorical text; operating site; segmentation by facility.
3. **Region** – categorical text; governorate; geographic segmentation.
4. **Client** – categorical text; counterparty; commercial segmentation.
5. **Contract_Type** – categorical text; EPC, maintenance, rehabilitation, etc.; risk and margin analysis.
6. **Revenue_Category** – categorical text; pipe coating, tank lining, blasting, etc.; operational segmentation.
7. **Payment_Status** – categorical ordered; uninvoiced, invoiced unpaid, partially collected, collected; receivables risk.

7. Quantitative Variables

1. **Contract_Value_MEGP** – numeric; contracted scope in MEGP; anchor for scale.
2. **Physical_Progress** – ratio; % scope completed; driver of revenue.
3. **Revenue_Recognised_MEGP** – numeric; recognised progress-based revenue; top line measure.
4. **Direct_Cost_MEGP** – numeric; direct labour/materials; cost driver.
5. **Overhead_Cost_MEGP** – numeric; allocated overheads; cost driver.
6. **Capex_MEGP** – numeric; capital expenditure; asset intensity proxy.
7. **Invoice_Amount_MEGP** – numeric; billed value; starting receivable.

© 2025 Dr. Yasser Aly Yakout Shehata

8. **Collected_Amount_MEGP** – numeric; cash collected; conversion outcome.
9. **Receivables_Balance_MEGP** – numeric; invoice minus collections; exposure measure.
10. **AR_Days** – integer; receivable duration; liquidity measure.

8. Derived Variables

Examples (full list of 39 available):

- **Gross_Margin_MEGP** – numeric; revenue minus costs; profitability indicator.
- **Margin_Pct** – ratio; gross margin ÷ revenue; efficiency measure.
- **Collection_Rate** – ratio; collections ÷ invoices; settlement effectiveness.
- **Budget_Status** – categorical; under/on/over budget; performance classification.
- **AR_Aging_Bucket** – categorical ordered; 0–30, 31–60, 61–90, etc.; credit control segmentation.
- **High_Risk_AR_Flag** – binary; indicator of overdue receivables; risk flag.
- **Contract_Size_Band** – categorical; small/medium/large/mega; scale segmentation.
- **Margin_Delta_vs_Category** – numeric; deviation from peer average; benchmarking signal.

9. Relevance to Power BI and AI for Business

- **Power BI**: Clean star schema design, surrogate keys, time-intelligence attributes, additive/semi-additive measures, and interpretable ratios make it ideal for executive dashboards.
- **AI for Business**: Labels and derived features allow prediction of overdue invoices, classification of budget status, and forecasting of margins or cash conversion.

10. Suggested Power BI Dashboard Designs

1. **Executive Overview** – revenue, margin, collection rate cards; time series; plant contribution bar chart; payment status matrix.
2. **Revenue and Progress** – scatter of progress vs revenue; decomposition tree by plant, category, and client; budget vs actual waterfall.
3. **Cost and Margin** – stacked cost vs revenue; margin heatmap by plant; peer comparison table with deltas.

© 2025 Dr. Yasser Aly Yakout Shehata

4. **Billing and Collections** – time series of invoices, collections, receivables; client ageing matrix; funnel of settlement stages.
5. **Budget and Forecast** – attainment percentages, forecast bias distribution, list of over budget cases.
6. **Risk and Alerts** – high risk AR, overdue invoices, low margin contracts; filters by plant and client.

11. Sample Business Insights

- Port Said generates the largest revenue but its margin percentage is highly sensitive to revenue category mix.
- Rehabilitation contracts show weaker collection rates, indicating higher settlement risks.
- Budget overruns cluster in months with rapid progress but lagging invoices, highlighting operational-finance misalignment.
- Cost-to-revenue efficiency is stronger in pipe coating but weaker in blasting and painting.
- Overdue invoices correlate with payment status "invoiced unpaid" beyond 90 days, validating early-warning signals.

12. Applications in Business Analytics and AI

A. **EDA** – profiling plants, clients, and categories; testing relationships between progress, cost, margin, and cash.

B. **Power BI Visualization** – building dashboards with drill-through, hierarchies, and conditional formatting.

C. **AI-Driven Insights** – classification models for overdue invoices, regression for margin forecasting, and predictive scoring for budget adherence.

13. Conclusion

The **Petrojet Coating Finance Analytics Gold Dataset** offers MBA students a robust environment to apply analytical techniques to realistic industrial finance challenges. It enables structured **EDA**, advanced **Power BI dashboarding**, and baseline **AI-driven business applications**. By linking operational progress with revenue recognition, margin analysis, receivables management, and budget control, the dataset empowers students to deliver actionable insights directly relevant to executive decision-making.

© 2025 Dr. Yasser Aly Yakout Shehata

© 2025 Dr. Yasser Aly Yakout Shehata

Appendix: Guidance for Dataset Selection and Application

This appendix provides structured guidance for students to connect their chosen datasets with the requirements of the capstone project. Each section outlines a dataset summary, possible dashboard visuals, AI applications, and sample business recommendations. The examples serve as a reference point to inspire creative yet academically rigorous project outcomes.

1. Construction Industry

Dataset: *Construction Project Performance Analytics (CPPA-1000)*

Dataset Summary:

The dataset simulates performance metrics across multiple construction projects, incorporating cost, schedule adherence, contractor quality, safety incidents, and compliance indicators. It is modeled on real project management functions with both quantitative and categorical variables.

Possible Dashboard Visuals:

- Gantt-style chart for project milestones.
- Heatmap of cost variance across sites.
- KPI cards for safety incidents, contractor delays, and schedule adherence.
- Drill-through dashboards for contractor performance comparisons.

AI Applications:

- **Predictive Models:** Anticipate project delays based on resource allocation and contractor performance.
- **Clustering:** Group contractors or projects by quality, compliance, and cost-effectiveness.
- **Optimization:** Resource scheduling for equipment and labor to minimize time and cost overruns.

Sample Recommendations:

- Prioritize partnerships with high-performing contractors identified through clustering.
- Reallocate resources to projects predicted to face delays.
- Implement an AI-based early warning system for safety and compliance risks.

© 2025 Dr. Yasser Aly Yakout Shehata

2. Healthcare Sector

Dataset: *Ismailia Healthcare IT Transformation*

Dataset Summary:

This dataset models the digital transformation process of a regional healthcare system. It captures patient flow, adoption of digital platforms, IT system usage, and staff efficiency metrics.

Possible Dashboard Visuals:

- Sankey diagram of patient journey through IT systems.
- System usage trend lines segmented by department.
- KPIs for digital adoption rates, patient wait times, and error frequencies.
- Comparative dashboards for staff performance.

AI Applications:

- **Forecasting:** Predict peaks in system usage and patient demand.
- **Classification:** Identify patients or departments at risk of low adoption or dissatisfaction.
- **Optimization:** Scheduling staff shifts to match predicted demand.

Sample Recommendations:

- Deploy predictive analytics to allocate IT resources during peak times.
- Train departments identified as low adopters to improve digital service delivery.
- Optimize staffing patterns to reduce patient wait times.

3. Pharmaceuticals and Retail

Dataset: *Pharmacy Administration in South Sinai*

Dataset Summary:

This dataset captures pharmacy operations across branches, including sales, prescriptions, customer engagement, staffing, and promotions. It reflects both retail and healthcare elements.

Possible Dashboard Visuals:

- Branch-level sales performance maps.

© 2025 Dr. Yasser Aly Yakout Shehata

- Stacked bar charts comparing prescription vs. OTC sales.
- KPIs for average transaction value, promotion uptake, and staff productivity.
- Trend dashboards for customer retention.

AI Applications:

- **Forecasting:** Predict medicine demand by season.
- **Segmentation:** Group customers by purchasing behavior.
- **Optimization:** Inventory planning and promotional strategy.

Sample Recommendations:

- Focus marketing campaigns on customer segments with high promotion uptake.
- Use demand forecasting to optimize inventory levels, reducing waste.
- Enhance staff training in underperforming branches identified via dashboard drill-throughs.

4. Financial Services

Dataset: *CIB Corporate Credit Risk*

Dataset Summary:

The dataset models corporate lending practices, focusing on client credit profiles, repayment history, and portfolio risk. It integrates qualitative and quantitative financial metrics.

Possible Dashboard Visuals:

- Risk exposure dashboard by sector and client category.
- Heatmaps of overdue accounts by branch.
- Portfolio KPIs such as loan-to-value ratios and repayment performance.
- Trend dashboards showing non-performing loan (NPL) trajectories.

AI Applications:

- **Predictive Models:** Identify clients at high risk of default.
- **Clustering:** Segment clients by creditworthiness.
- **Optimization:** Balance portfolio risk across industries.

© 2025 Dr. Yasser Aly Yakout Shehata

Sample Recommendations:

- Adjust credit policies for sectors with consistently high default risk.
- Focus on targeted client support to reduce NPLs.
- Use predictive scoring models to prioritize credit approvals.

5. Manufacturing

Dataset: *Ezz Steel Sales Analytics*

Dataset Summary:

This dataset simulates sales and production efficiency in a large-scale steel manufacturing environment. Variables cover regional sales, production output, supply chain timelines, and cost efficiency.

Possible Dashboard Visuals:

- Regional sales comparison maps.
- KPI cards for production efficiency and fulfillment rates.
- Trend lines for steel demand vs. production capacity.
- Drill-through dashboards for supply chain analysis.

AI Applications:

- **Forecasting:** Demand prediction for steel in different regions.
- **Segmentation:** Categorize clients based on consumption behavior.
- **Optimization:** Align production scheduling with forecasted demand.

Sample Recommendations:

- Scale production capacity in regions with strong growth forecasts.
- Focus sales strategies on client segments with high repeat purchase rates.
- Use optimization models to reduce supply chain delays.

6. Logistics and Operations

Dataset: *Integrated Site Operations and Performance Analytics*

Dataset Summary:

© 2025 Dr. Yasser Aly Yakout Shehata

This dataset models operational KPIs for logistics and site management, covering fleet performance, compliance inspections, and efficiency indices.

Possible Dashboard Visuals:

- Fleet utilization KPIs with downtime tracking.
- Compliance inspection pass rates by site.
- Trend dashboards for delivery timeliness.
- Heatmaps of site performance.

AI Applications:

- **Predictive Models:** Estimate fleet breakdown probabilities.
- **Clustering:** Group sites by compliance and efficiency levels.
- **Optimization:** Maintenance scheduling to minimize downtime.

Sample Recommendations:

- Introduce predictive maintenance for high-risk fleet vehicles.
- Focus compliance training on low-performing sites.
- Optimize route planning to reduce delays and costs.

7. Corporate Leadership and HR

Dataset: *Corporate Leadership Performance Analytics*

Dataset Summary:

This dataset simulates leadership and HR performance metrics, including employee productivity, engagement, and succession planning indicators.

Possible Dashboard Visuals:

- Leadership performance dashboards with drill-down by department.
- Employee engagement KPIs segmented by role.
- Heatmaps of training effectiveness.
- Succession planning dashboards.

AI Applications:

- **Predictive Models:** Forecast employee turnover or promotion readiness.

© 2025 Dr. Yasser Aly Yakout Shehata

- **Clustering:** Identify employee groups needing targeted development.
- **Optimization:** Succession planning aligned with performance indicators.

Sample Recommendations:

- Launch training programs for clusters of employees flagged as low performers.
- Use predictive analytics to retain high-potential employees.
- Strengthen leadership pipelines through data-informed succession planning.

Concluding Note

These examples illustrate how students can connect their chosen dataset to the required deliverables: an **interactive dashboard**, a **comprehensive report**, and **AI-powered business recommendations**. By aligning the dataset's scope with analytical workflows and strategic insights, students ensure their projects meet the dual objectives of academic rigor and professional relevance.

© 2025 Dr. Yasser Aly Yakout Shehata