**Presented by:**

**Mahmoud Hatem Mohamed**

**Bishoy Talaat Fathy**

**Amr osama abdelfatah**

**Seif wael metwaly**

**Evan abdelmalek qaiser**

**Salah amer mohamed**

**Project Planning & Management: Sales Dataset Analysis**

**1. Project Planning & Management**

**Project Proposal**

* **Overview**: This project analyzes a U.S.-based sales dataset containing 9,800 unique transactions (after removing 119 duplicates) from 2022–2025. It examines sales patterns across regions, states, product categories, and delivery timelines to derive actionable business insights.
* **Objectives**:
  + Clean the dataset by addressing null values (City, Postal Code, Sales) and duplicates.
  + Identify high-performing regions, states, and product categories.
  + Evaluate delivery delays and their impact on sales performance.
  + Provide recommendations for optimizing inventory, logistics, and resource allocation.
* **Scope**:
  + **Inclusions**: Data preprocessing, sales analysis by region/category, delivery delay assessment, and reporting.
  + **Exclusions**: Predictive modeling, external datasets, or real-time system integration unless specified.

**Project Plan**

* **Timeline (Gantt Chart)**: A 5-week plan (visualizable in Excel/MS Project):
  + **Week 1**: Data loading and exploration (Days 1–5).
  + **Week 2**: Data cleaning (nulls, duplicates, date formatting) (Days 6–10).
  + **Week 3**: Feature engineering (e.g., Year, Delivery Delay) (Days 11–15).
  + **Week 4**: Sales analysis (by region, category, delays) (Days 16–25).
  + **Week 5**: Reporting and finalization (Days 26–30).
* **Milestones**:
  + **M1 (Day 5)**: Dataset loaded, nulls identified (26 in City, 43 in Postal Code, 2 in Sales).
  + **M2 (Day 10)**: Data cleaned (nulls filled, 119 duplicates dropped).
  + **M3 (Day 15)**: Features added (Year, Month, Delivery Delay).
  + **M4 (Day 25)**: Analysis completed (top regions, product insights).
  + **M5 (Day 30)**: Final report and project\_last.csv delivered.
* **Deliverables**:
  + Cleaned dataset (project\_last.csv).
  + Report with sales insights and recommendations.
  + Optional visuals (e.g., sales by region) if requested.
* **Resource Allocation**:
  + **Data Analyst**: Handles cleaning, analysis, and reporting (Python, pandas).
  + **Project Manager**: Oversees timeline and deliverables.
  + **Tools**: Jupyter Notebook, Python (pandas, numpy), Excel for planning.

**Task Assignment & Roles**

* **Overview**: The project team consists of six members—Mahmoud, Bishoy, Amr, Seif, Evan, and Salah—each assigned specific responsibilities to execute the sales dataset analysis and develop a supporting system. Tasks are distributed across data preprocessing, feature engineering, analysis, system design, quality assurance, and project management.
* **Assigned Roles and Responsibilities**:
  + **Mahmoud Hatem Mohamed – Project Manager**:
    - **Tasks**: Define milestones (M1–M5: e.g., data cleaned by Day 10, analysis completed by Day 25), track progress using a Gantt chart, and ensure deliverables (e.g., project\_last.csv, final report) are submitted by Day 30.
    - **Responsibilities**: Coordinate team activities, review outputs for clarity and business relevance, and liaise with stakeholders (e.g., business manager) for feedback.
  + **Bishoy Talaat Fathy – Data Analyst (Preprocessing)**:
    - **Tasks**: Load the dataset (df = pd.read\_csv('Sales\_dataset (1).csv')), clean data by removing 119 duplicates (df.drop\_duplicates()), and address nulls (e.g., fill City with mode like Los Angeles, Sales with mean ~$229).
    - **Responsibilities**: Ensure data quality, document preprocessing steps, and prepare project\_last.csv for subsequent tasks.
  + **Amr Osama Abdelfatah – Data Analyst (Feature Engineering)**:
    - **Tasks**: Convert Order Date and Ship Date to datetime (pd.to\_datetime), create features like Year, Month, Day Name, and calculate Delivery Delay (days between Ship Date and Order Date).
    - **Responsibilities**: Validate feature accuracy (e.g., cross-check Delivery Delay), ensure features support analysis requirements.
  + **Evan Abdelmalek Qaiser – Data Analyst (Sales Analysis)**:
    - **Tasks**: Analyze sales by Region, State, Category, and Year (e.g., aggregate sales for California, identify top categories like Furniture), evaluate Delivery Delay impact on sales performance.
    - **Responsibilities**: Derive insights (e.g., top states: California, Florida), recommend actions (e.g., optimize shipping in Henderson), and summarize findings.
  + **Seif Wael Metwaly – System Designer**:
    - **Tasks**: Design the system architecture (MVC: Model for database, View for UI, Controller for logic), develop database schema (e.g., Orders, Customers tables), and create UI wireframes (e.g., dashboard with sales by Region).
    - **Responsibilities**: Ensure system scalability, meet functional requirements (e.g., report generation), and support deployment planning.
  + **Salah Amer Mohamed – Quality Assurance & Testing**:
    - **Tasks**: Verify data cleaning (e.g., confirm df.duplicated().sum() == 0), test analysis results (e.g., sales totals align with raw data), and validate system functionality (e.g., UI displays accurate reports).
    - **Responsibilities**: Detect errors (e.g., date inconsistencies), maintain 99% data accuracy, and document testing outcomes for final review.

**Risk Assessment & Mitigation Plan**

* **Risks**:
  + **Incomplete Data**: Nulls in City (26), Postal Code (43), Sales (2).
    - **Mitigation**: Filled City with mode (e.g., Los Angeles), Sales with mean (~$229), Postal Code with placeholder/mode.
  + **Date Inconsistencies**: Formats like 8/11/2017 vs. 2017-11-08.
    - **Mitigation**: Standardized via pd.to\_datetime.
  + **Duplicates**: 119 duplicate rows skewing results.
    - **Mitigation**: Removed with df.drop\_duplicates() (verified 0 duplicates).
  + **Analysis Bias**: Overemphasis on top regions (e.g., California).
    - **Mitigation**: Include smaller regions (e.g., Kentucky) in analysis.
  + **Timeline Delays**: Data complexity or tool issues.
    - **Mitigation**: Buffer days per phase, use reliable tools (pandas).
* **Contingency**: Backup raw/cleaned datasets, document steps for reproducibility.

**KPIs (Key Performance Indicators)**

* **Response Time**: Data processing completed within 1 hour (achieved in notebook).
* **System Uptime**: Analysis environment (e.g., Jupyter) operational 99% of the time.
* **User Adoption Rate**: 100% of stakeholders use project\_last.csv for decisions.
* **Additional Metrics**:
  + **Data Retention**: Retained 98.8% of rows (9,800/9,919).
  + **Accuracy**: <1% error in transformations (e.g., date conversions).
  + **Insight Delivery**: At least 3 actionable recommendations.

**2. Literature Review**

**Feedback & Evaluation – Assessment of the Project**

* **Assumed Feedback**:
  + **Strengths**: Robust cleaning (nulls filled, duplicates removed) and feature engineering (Delivery Delay) provide a solid base. Insights into top regions (e.g., California) are clear.
  + **Weaknesses**: Analysis lacks depth (e.g., no statistical tests) and visualizations. Temporal trends (2022–2025) are underexplored.
* **Evaluation**: A strong start with practical preprocessing, but deeper analysis and presentation enhancements are needed.

**Suggested Improvements – Areas Where the Project Can Be Enhanced**

* **Statistical Depth**: Add correlations (e.g., Sales vs. Delivery Delay) or averages by Category.
* **Temporal Trends**: Analyze sales growth by Year and Category.
* **Visuals**: Include charts (e.g., sales by Region, category breakdown).
* **Outlier Analysis**: Investigate anomalies (e.g., Henderson’s $731 chair sale).
* **Customer Insights**: Explore Segment (Consumer, Corporate) for targeting strategies.

**Final Grading Criteria – Breakdown of Marks**

* **Documentation (30%)**: Clear cleaning steps and narrative, but lacks methodology detail.
* **Implementation (30%)**: Effective cleaning and basic analysis, limited to aggregations.
* **Testing (20%)**: Basic validation (nulls, duplicates), no statistical checks.
* **Presentation (20%)**: Text-based insights clear, but no visuals or polished summary.

**3. Requirements Gathering**

**Stakeholder Analysis – Identifying Key Stakeholders and Their Needs**

* **Business Manager**:
  + **Needs**: Top regions/products for resource allocation, logistics optimization.
* **Data Analyst**:
  + **Needs**: Clean data, efficient tools, clear objectives.
* **Customers**:
  + **Needs**: Timely delivery, product availability.
* **Logistics Team**:
  + **Needs**: Delivery delay insights for shipping improvements.

**Functional Requirements – List of Features and Functionalities**

* **Data Cleaning**:
  + Remove duplicates (df.drop\_duplicates()).
  + Fill nulls: City (mode), Sales (mean), Postal Code (placeholder).
* **Feature Engineering**:
  + Standardize dates (pd.to\_datetime).
  + Add Year, Month, Delivery Delay.
* **Analysis**:
  + Aggregate sales by Region, State, Category.
  + Assess delivery delay impacts.
* **Reporting**:
  + Summarize findings, export to project\_last.csv.

**Non-functional Requirements – Performance, Security, Usability, Reliability**

* **Performance**: Process 9,800 rows in <1 minute, complete project in 5 weeks.
* **Security**: Anonymize customer data if shared, secure project\_last.csv.
* **Usability**: Clear outputs for non-technical users, documented steps.
* **Reliability**: 99%+ data accuracy, validated calculations.

**4. System Analysis & Design**

**1. Problem Statement & Objectives**

* **Problem Statement**: Businesses lack a streamlined system to process, analyze, and visualize sales data (e.g., 9,800 transactions from 2022–2025) for actionable insights. Manual analysis is time-consuming, error-prone (e.g., 119 duplicates, 71 nulls), and fails to scale with growing data or provide real-time decision support for inventory, logistics, and regional strategies.
* **Objectives**:
  + Develop a system to clean and process sales data efficiently (e.g., handle nulls, duplicates).
  + Enable analysis of sales by region (Region, State), product (Category), and delivery performance (Delivery Delay).
  + Provide a user-friendly interface for stakeholders (e.g., business managers) to access insights.
  + Ensure scalability and reliability for future data (e.g., 2026+ sales).