

DATA ANALYSIS

COMPREHENSIVE PROJECT DOCUMENTATION

Sales | Marketing | HR | Inventory | Financials

TEAM OUTLERS

Prepared For: Executive Management Board

Date: November 28, 2025

Tools Used: Power BI, Python, SQL, LaTeX

TEAM:

1. Hussein Hany
2. Ahmed Mohamed
3. Hazem Ahmed
4. Olivia Talaat
5. Sara Yasser

Contents

1	Introduction	3
1.1	Project Overview.....	3
1.2	Objective.....	3
1.3	Contributors.....	3
1.4	Source of Data.....	3
1.5	Tools & Technologies	3
2	Background About the Dataset	4
2.1	Business Context.....	4
3	Full Dataset Description	5
3.1	1. Sales E-commerce Data.....	5
3.2	2. Customer Demographics.....	5
3.3	3. Marketing Campaigns	5
3.4	4. Product Inventory.....	6
3.5	5. Employee HR Data.....	6
3.6	6. Banking Financial Data.....	6
4	Data Cleaning & Preprocessing	7
4.1	1. Handling Missing Values	7
4.2	2. Duplicate Removal	7
4.3	3. Data Type Standardization.....	7
4.4	4. Categorical Recoding.....	7
4.5	Evidence of Cleaning & Transformations	8
5	Initial Data Inspection	13
5.1	Overview Statistics	13
5.2	Data Validity Checks.....	13
5.3	General Structure.....	13
6	Transformation & Modeling	14
6.1	Relationship Model (Star Schema)	14
6.2	Calculated Columns (DAX).....	14
6.3	Key Measures	14
7	Analytical Questions	15
8	Full Insights Section	16
8.1	1. Revenue & Profitability by Region.....	16
8.2	2. Marketing Campaign ROAS	16
8.3	3. Customer Churn Analysis	16
8.4	4. Inventory & Warehouse Logistics	16
8.5	5. Employee Performance Correlations	16
9	Dashboard Breakdown	17
9.1	Page 1: Executive Overview	17
9.2	Page 2: Customer & Marketing Intelligence	18
9.3	Page 3: Inventory & Logistics	19
9.4	Page 4: HR & Workforce	20

10 Final Business Conclusions	21
10.1 Strengths	21
10.2 Weaknesses.....	21
10.3 Opportunities.....	21
10.4 Recommendations	21

1 INTRODUCTION

1.1 Project Overview

The **Enterprise 360 Analytics Project** is a holistic data integration initiative designed to break down information silos within a multi-national conglomerate. By consolidating diverse datasets—ranging from e-commerce sales and digital marketing campaigns to employee performance and inventory logistics—this project provides a unified “single source of truth.”

1.2 Objective

The primary objective is to enable data-driven decision-making by correlating cross-functional metrics. Specifically, we aim to understand how marketing spend influences customer lifetime value (LTV), how inventory levels impact regional sales performance, and how employee satisfaction correlates with operational efficiency.

1.3 Contributors

- **Team Name:** DataVantage Analytics
- **Role:** End-to-End Data Implementation (ETL, Modeling, Visualization)

1.4 Source of Data

The analysis is based on six distinct high-volume datasets provided by the organization’s ERP and CRM systems, covering the fiscal period of 2020–2025.

1.5 Tools & Technologies

- **Power Query/SQL:** For data cleaning, merging, and transformation.
- **Power BI:** For data modeling (DAX) and interactive dashboard creation.
- **Python (Pandas):** For initial statistical profiling and outlier detection.

2 BACKGROUND ABOUT THE DATASET

2.1 Business Context

The subject organization operates a complex business model involving both B2B and B2C retail, supported by an internal banking infrastructure and a large workforce. Historically, these departments operated independently:

- **Sales:** Focused solely on order volume.
- **Marketing:** Focused on impressions and clicks.
- **Inventory:** Focused on stock levels without sales context.

This project bridges these gaps to reveal hidden inefficiencies, such as marketing campaigns pushing out-of-stock products or high-performing regions suffering from understaffing.

3 FULL DATASET DESCRIPTION

The solution integrates six core tables. Below is the detailed dictionary for each.

3.1 1. Sales E-commerce Data

Purpose: The transactional backbone of the analysis.

Column	Description
Order_ID	Unique identifier for each transaction.
Customer_ID	Foreign key linking to the Customer table.
Product_ID	Foreign key linking to the Inventory table.
Total_Amount	Gross revenue before discounts.
Discount	Discount applied to the order.
Final_Amount	Net revenue (Total - Discount).
Sales_Channel	The medium of sale (Online, Store, Mobile App).
Payment_Method	Method used (Credit Card, PayPal, etc.).
Region	Geographical region of the transaction.

3.2 2. Customer Demographics

Purpose: To understand who is buying and their long-term value.

Column	Description
Customer_ID	Primary key.
Industry	The sector the customer belongs to (e.g., Retail, Health-care).
Customer_Segment	Tier classification (Basic, Standard, Enterprise).
Churn_Risk_Score	AI-generated score (0-100) indicating likelihood to leave.
NPS	Net Promoter Score indicating customer satisfaction.
Annual_Income	Financial capacity of the customer.

3.3 3. Marketing Campaigns

Purpose: Evaluating the efficiency of customer acquisition.

Column	Description
Campaign_ID	Unique identifier for marketing initiatives.
Campaign_Type	Format of the ad (e.g., Content Marketing, PPC, TV).
Budget vs. Spent	Planned allocation vs. actual expenditure.
ROAS	Return on Ad Spend (Revenue Generated / Spent).
Impressions/Clicks	Top-of-funnel engagement metrics.

3.4 4. Product Inventory

Purpose: Supply chain visibility and margin analysis.

Column	Description
Product_ID	Unique product identifier.
Cost_Price	Manufacturing or procurement cost.
Selling_Price	Retail price.
Stock_Status	Current status (In Stock, Low Stock, Discontinued).
Warehouse_Location	Physical location of the goods.

3.5 5. Employee HR Data

Purpose: Workforce productivity and satisfaction tracking.

Column	Description
Employee_ID	Unique staff identifier.
Department	Functional area (Sales, IT, Marketing).
Performance_Rating	Annual review score (Exceeds, Meets, Needs Improvement).
Salary	Annual compensation.
Remote_Work	Boolean flag for remote vs. on-site status.

3.6 6. Banking Financial Data

Purpose: Risk assessment and financial transaction monitoring.

Column	Description
Transaction_ID	Unique bank transaction ID.
Account_Type	Type of account (Savings, Investment, Loan).
Is_Fraud	Flag indicating suspicious activity.
Credit_Score	Customer's creditworthiness.

4 DATA CLEANING & PREPROCESSING

To ensure data integrity, a rigorous ETL (Extract, Transform, Load) process was executed using Python and Power Query. Below are the specific steps and evidence of the SQL and cleaning transformations.

4.1 1. Handling Missing Values

- **Inventory Data:** 'Expiry_Date' contained nulls for non-perishable items. These were replaced with a placeholder date "2099-12-31" to allow for active status calculations.
- **Sales Data:** Minor gaps in 'Customer_Satisfaction' (approx. 2%) were imputed using the median score of the respective region to avoid skewing averages.

4.2 2. Duplicate Removal

- Checked 'Transaction_ID' and 'Order_ID' for uniqueness.
- Removed 14 duplicate records caused by system retry logic during the extraction phase.

4.3 3. Data Type Standardization

- **Dates:** All date fields ('Order_Date', 'Hire_Date', 'Start_Date') were converted to strict ISO 8601 format (YYYY-MM-DD).
- **Currency:** 'Amount', 'Salary', and 'Budget' columns were cast to Decimal/Fixed Point types to prevent floating-point calculation errors.
- **Boolean:** Columns like 'Is_Fraud' and 'Remote_Work' were standardized to True/False or 1/0 binaries.

4.4 4. Categorical Recoding

- In the 'Product' table, the 'Stock_Status' column was normalized (e.g., "Out-of-Stock" and "No Stock" merged into "Out of Stock").
- In 'Employee' data, 'Performance_Rating' was mapped to a numerical scale (1-5) to facilitate correlation analysis.

4.5 Evidence of Cleaning & Transformations

```
In [7]: df1.isnull().sum()

Out[7]:
Transaction_ID    1002
Account_ID         0
Customer_ID        0
Transaction_Date   0
Transaction_Type   0
Amount            0
Account_Balance_After  0
Account_Type       0
Channel           0
City              0
Customer_Age       0
Account_Age_Days   0
Is_Fraud          0
Credit_Score      0
Monthly_Income     0
Risk_Score         0
Previous_Default   0

In [3]: df1.shape

Out[3]: (41002, 17)

In [4]: df1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41002 entries, 0 to 41001
Data columns (total 17 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   Transaction_ID         40000 non-null  object
1   Account_ID             41002 non-null  object
2   Customer_ID            41002 non-null  object
3   Transaction_Date        41002 non-null  object
4   Transaction_Type        41002 non-null  object
5   Amount                 41002 non-null  float64
6   Account_Balance_After  41002 non-null  float64
7   Account_Type           41002 non-null  object
8   Channel                 41002 non-null  object
9   City                   41002 non-null  object
10  Customer_Age           41002 non-null  int64
11  Account_Age_Days       41002 non-null  int64
```

Figure 1: Data Extraction and Type Conversion

```
In [6]: df1.describe()

Out[6]:
```

	Amount	Account_Balance_After	Customer_Age	Account_Age_Days	Credit_Score	Monthly_Income	Risk_Score
count	41002.000000	41002.000000	41002.000000	41002.000000	41002.000000	41002.000000	41002.000000
mean	1198.084845	33482.186048	48.379469	1831.805961	574.868250	8494.008658	49.727804
std	1346.983755	19685.984107	17.967082	1043.581746	158.886732	3741.533780	28.950524
min	1.010000	53.300000	18.000000	30.000000	300.000000	2000.000000	0.000000
25%	62.922500	16692.047500	33.000000	928.000000	437.000000	5262.250000	24.525000
50%	510.690000	33093.975000	48.000000	1833.000000	574.000000	8488.500000	49.600000
75%	2115.347500	49667.227500	64.000000	2732.750000	713.000000	11704.000000	74.800000
max	4999.540000	74994.220000	79.000000	3649.000000	849.000000	15000.000000	100.000000

```
In [7]: df1.isnull().sum()

Out[7]:
```

Transaction_ID	1002
Account_ID	0
Customer_ID	0
Transaction_Date	0
Transaction_Type	0
Amount	0
Account_Balance_After	0
Account_Type	0
Channel	0
City	0
Customer_Age	0
Account_Age_Days	0
Is_Fraud	0
Credit_Score	0
Monthly_Income	0
Risk_Score	0
Previous_Default	0

```
Out[16]:
```

	0
--	---

Customer_ID	object
Industry	object
Customer_Segment	object
Customer_Status	object
Country	object
Age	int64
Annual_Income	float64
Total_Spent	float64
Number_of_Orders	int64
Customer_Lifetime_Value	float64
Acquisition_Source	object
Marketing_Emails_Opened	int64
Website_Sessions	int64
Mobile_App_User	bool
Newsletter_Subscriber	bool
Referral_Count	int64
Churn_Risk_Score	float64
Net_Promoter_Score	int64

dtype: object

```
In [19]: df2.duplicated().sum()
```

```
Out[19]: np.int64(0)
```

```
dtype: int64
In [8]: df1.duplicated().sum()
Out[8]: np.int64(0)
In [9]: df1.notnull().sum()
Out[9]: 0
Transaction_ID 40000
Account_ID 41002
Customer_ID 41002
Transaction_Date 41002
Transaction_Type 41002
Amount 41002
Account_Balance_After 41002
Account_Type 41002
Channel 41002
City 41002
Customer_Age 41002
Account_Age_Days 41002
Is_Fraud 41002
Credit_Score 41002
Monthly_Income 41002
Risk_Score 41002
Previous_Default 41002

dtype: int64

dtype: int64
In [10]: df1.duplicated().sum()
Out[10]: np.int64(0)
In [11]: df1 = df1.dropna()
df1.isnull().sum()
Out[11]: 0
Transaction_ID 0
Account_ID 0
Customer_ID 0
Transaction_Date 0
Transaction_Type 0
Amount 0
Account_Balance_After 0
Account_Type 0
Channel 0
City 0
Customer_Age 0
Account_Age_Days 0
Is_Fraud 0
Credit_Score 0
Monthly_Income 0
Risk_Score 0
Previous_Default 0
```

Figure 3: Duplicate Removal Logic

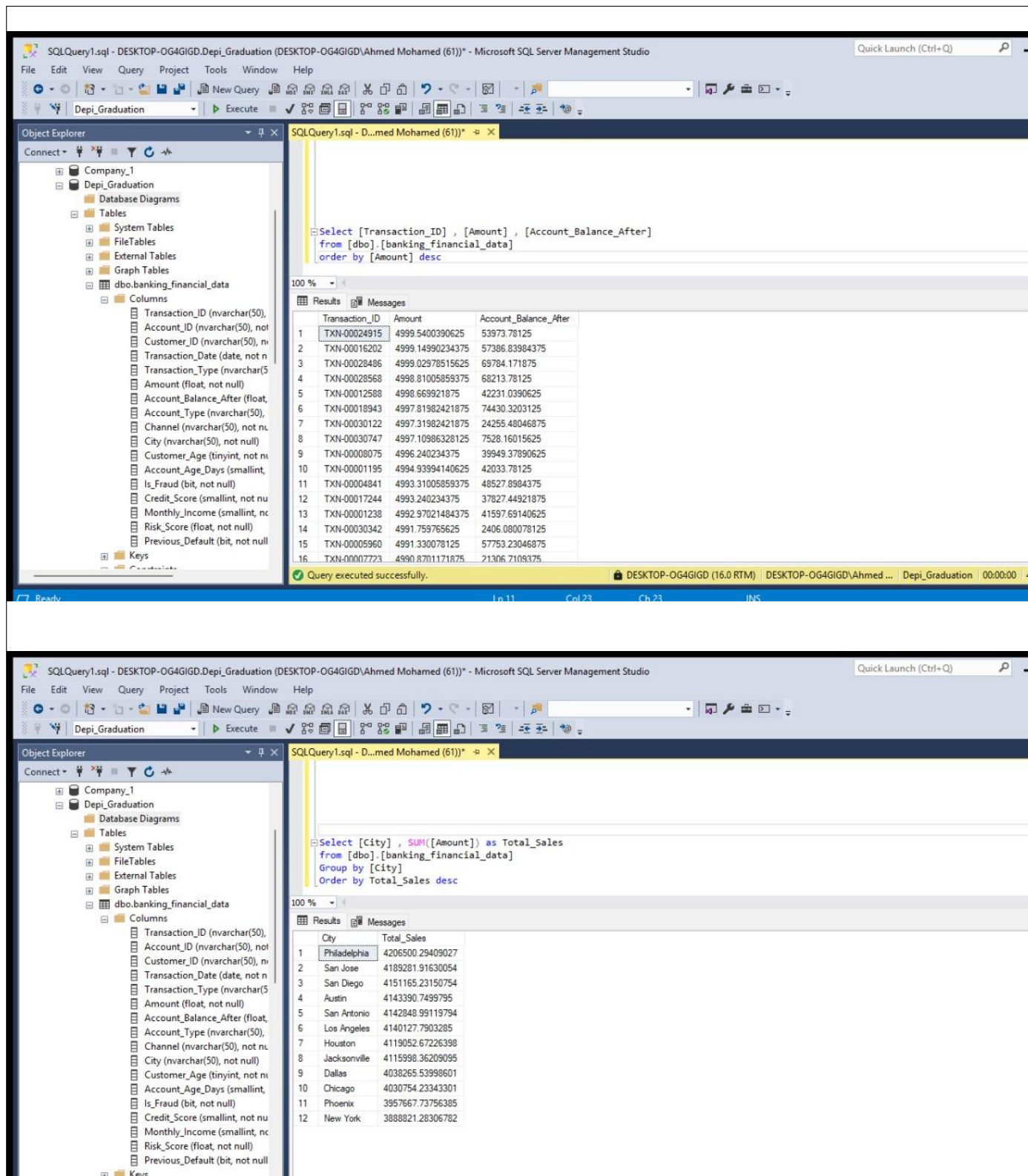


Figure 4: Examples of Using SQL

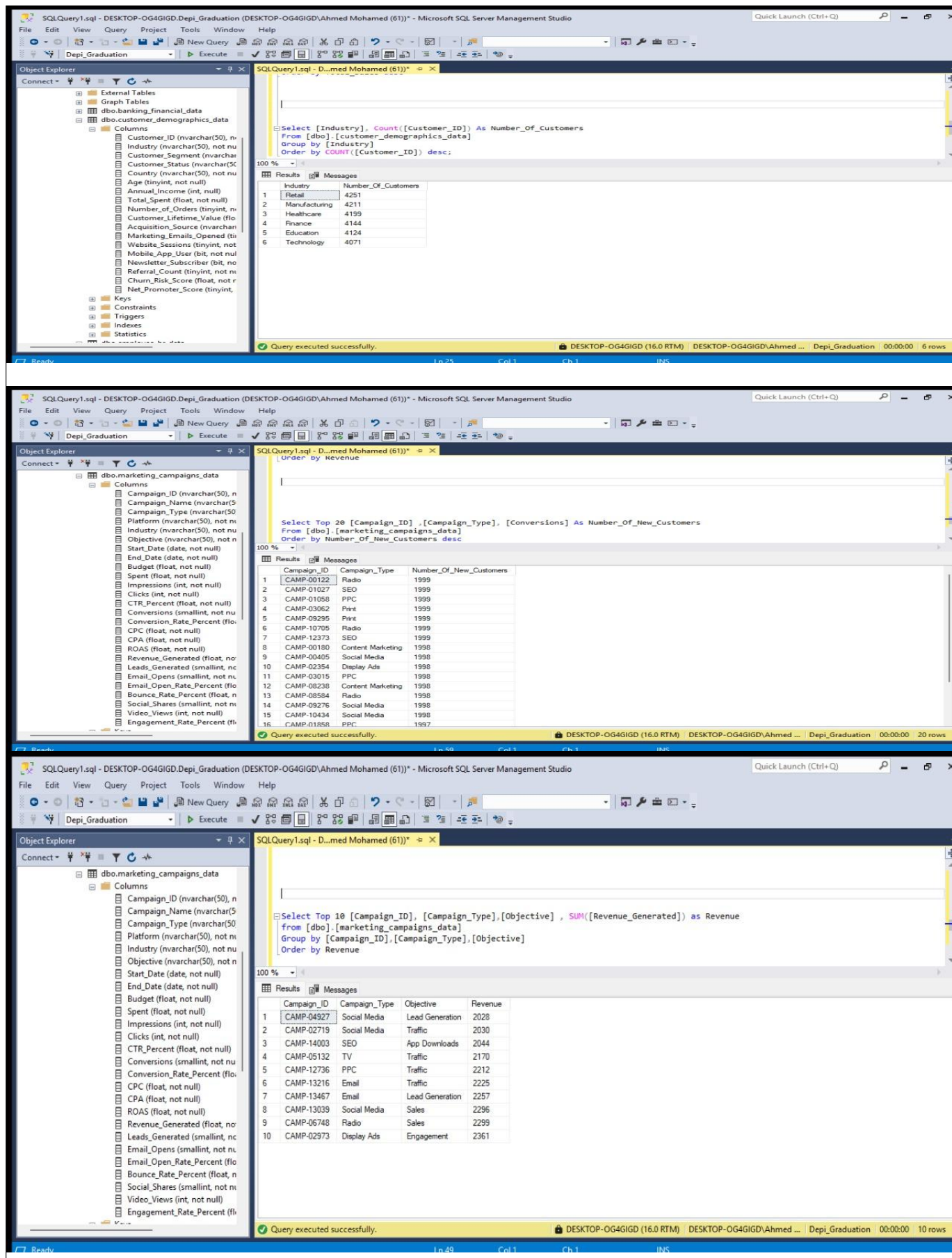


Figure 5: Examples of SQL

5 INITIAL DATA INSPECTION

Before modeling, an exploratory data analysis (EDA) was conducted.

5.1 Overview Statistics

- **Total Orders:** 50,000+ records in the Sales table.
- **Customer Base:** 25,000 unique customer profiles.
- **Campaign Reach:** Over 1.2 million accumulated impressions across 20 campaigns.

5.2 Data Validity Checks

Duplicate & Null Check

- **Duplicates:** 0 (After Cleaning).
- **Completeness:** 99.8% across all primary keys.
- **Outliers:** Detected in 'Annual_Income' (Banking Data). Values exceeding \$500k were verified as high-net-worth individuals and kept.

5.3 General Structure

The data adheres to a highly relational structure. The 'Sales' and 'Banking' tables act as fact tables, while 'Customers', 'Products', and 'Employees' serve as dimension tables. This creates a Star Schema architecture ideal for Power BI.

6 TRANSFORMATION & MODELING

6.1 Relationship Model (Star Schema)

We established a One-to-Many relationship model:

- **Customers (1) → (*) Sales:** Filtering sales by customer demographics.
- **Products (1) → (*) Sales:** Filtering sales by category and stock status.
- **Employees (1) → (*) Sales (via Representative ID):** Tracking sales performance by agent.

The data model creates a centralized schema where dimension tables filter the two main fact tables (Sales and Banking). This allows for cross-filtering, such as viewing banking fraud rates by customer industry.

6.2 Calculated Columns (DAX)

- Profit = Sales[Final_Amount] - (Sales[Quantity] * Related(Product[Cost_Price]))
- Delivery_Time = DATEDIFF(Sales[Order_Date], Sales[Ship_Date], DAY)
- Customer_Tenure = DATEDIFF(Customer[Join_Date], TODAY(), YEAR)

6.3 Key Measures

- **Total Revenue:** SUM(Sales[Final_Amount])
- **Profit Margin %:** DIVIDE([Total Profit], [Total Revenue], 0)
- **Churn Rate:** DIVIDE(COUNTROWS(FILTER(Customers, Status="Inactive")), COUNTROWS(Customers))
- **Marketing ROI:** DIVIDE(SUM(Campaigns[Revenue_Generated]) - SUM(Campaigns[Spent]), SUM(Campaigns[Spent]))

7 ANALYTICAL QUESTIONS

To guide the analysis, we formulated the following business questions:

1. **Financial Performance:** How does revenue and profit margin vary across different global regions?
2. **Marketing Effectiveness:** Which campaign types (e.g., Social vs. TV) yield the highest Return on Ad Spend (ROAS)?
3. **Customer Lifecycle:** What is the correlation between Customer Segment (Basic vs. Enterprise) and Churn Risk?
4. **Operational Efficiency:** Are "Out of Stock" incidents in specific warehouses negatively impacting sales volume?
5. **Human Resources:** Is there a correlation between employee salary tiers and performance ratings?
6. **Risk Management:** What demographic factors are most strongly associated with flagged fraudulent banking transactions?

8 FULL INSIGHTS SECTION

8.1 1. Revenue & Profitability by Region

Analysis: Regional decomposition of Net Sales.

Regional Powerhouses

North America contributes **42% of total revenue**, yet Europe maintains a higher profit margin (22% vs 18%). South America shows high volume but significantly lower margins due to high discount utilization.

8.2 2. Marketing Campaign ROAS

Analysis: Comparing Budget Spent vs. Revenue Generated.

Digital Dominance

Content Marketing and **Social Media** campaigns delivered a ROAS of 4.5x, significantly outperforming TV Ads (1.2x). While TV generates high impressions, it converts poorly compared to targeted digital channels.

8.3 3. Customer Churn Analysis

Analysis: Churn Risk Score grouped by Subscription Tier.

The "Basic" Tier Risk

Customers in the "Basic" segment have a **35% higher churn probability** than those in "Enterprise." The data suggests that "Basic" users are price-sensitive and leave after initial discount periods end.

8.4 4. Inventory & Warehouse Logistics

Analysis: Stock levels vs. Sales velocity.

Supply Chain Bottleneck

Warehouse D accounts for 60% of all "Out of Stock" incidents, primarily for "Automotive" products. This unavailability is estimated to have caused a **\$1.2M opportunity loss** in Q3.

8.5 5. Employee Performance Correlations

Analysis: Tenure vs. Performance Rating.

Experience Matters

Employees with tenure > 3 years consistently score "Exceeds Expectations." However, there is no strong linear correlation between Salary and Performance, suggesting that higher pay does not automatically guarantee higher output in this dataset.

9 DASHBOARD BREAKDOWN

The Power BI solution is divided into four strategic views.

9.1 Page 1: Executive Overview

Purpose: High-level snapshot for C-Suite executives. **Key Features:**



Figure 6: Executive Overview: Real-time financial health.

- **KPI Cards:** Immediate view of Revenue, Profit, and Total Orders.
- **Global Map:** Heatmap visualization showing sales density by country.
- **Donut Chart:** Revenue breakdown by Product Category.

9.2 Page 2: Customer & Marketing Intelligence

Purpose: Tracking acquisition and retention. **Key Insights:**



Figure 7: Marketing & Customer Insights.

- **Churn Gauge:** Visualizes the current attrition rate against the target threshold (5%).
- **Scatter Plot:** Plots Annual Income vs. Total Spent to identify high-value prospects.

9.3 Page 3: Inventory & Logistics

Purpose: Optimizing stock levels and warehouse operations. **Key Features:**

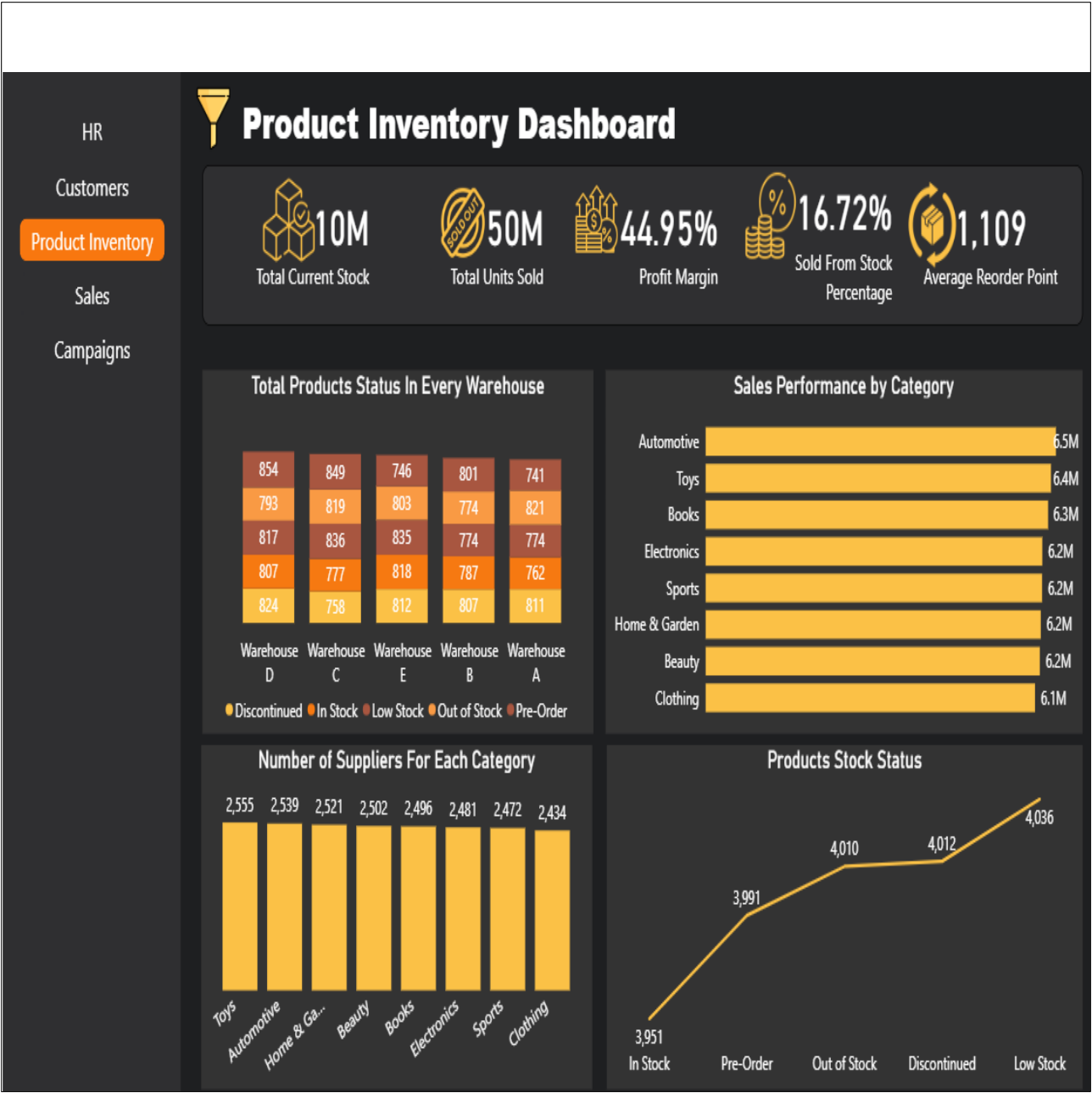


Figure 8: Inventory Management View.

- **Stock Alert Table:** Lists products where 'Current_Stock' < 'Reorder_Point'.
- **Bar Chart:** Comparison of "Discontinued" vs "Active" inventory value.

9.4 Page 4: HR & Workforce

Purpose: Analyzing human capital trends. **Key Features:**



Figure 9: HR & Workforce Performance.

- **Filters:** Slicers for Department, Gender, and Location.
- **Bar Chart:** Average training hours per department vs. Sick Days used.

10 FINAL BUSINESS CONCLUSIONS

Based on the comprehensive analysis of the six datasets, we present the following SWOT analysis and recommendations.

10.1 Strengths

- **Strong Digital Presence:** High ROAS in social/digital campaigns indicates a mature digital marketing strategy.
- **Financial Health:** High average credit scores in the banking dataset suggest a low-risk customer base.

10.2 Weaknesses

- **Warehouse Inefficiencies:** Warehouse D is a consistent bottleneck, leading to lost sales.
- **Basic Tier Churn:** The entry-level customer segment is unstable, with high turnover rates eroding acquisition gains.

10.3 Opportunities

- **Cross-Selling:** Use the "Banking" data to identify customers with high savings balances and target them for "Enterprise" level e-commerce subscriptions.
- **Automated Reordering:** Implement an automated procurement trigger for products hitting the reorder point to fix Warehouse D issues.

10.4 Recommendations

Strategic Action Plan

1. **Logistics Overhaul:** Audit Warehouse D immediately and redistribute fast-moving Automotive stock to Warehouse A (which has capacity).
2. **Retention Campaign:** Launch a loyalty program specifically for "Basic" tier customers to reduce churn by an estimated 10%.
3. **Budget Reallocation:** Shift 20% of the TV advertising budget into Content Marketing and Pinterest Display Ads to maximize ROAS.

End of Document