Ministry of Higher Education
and Scientific Research

University of Tunis

Tunis Business School

وزارة التعليم العالي و البحث العلمي

جامعة تونس

المعهد العالي للأعمال بتونس

Project Report

ShelfSense: Smarter Decisions for Retail Success

Elaborated By :
Fatma BEN REJEB
Roua ELKAMEL
Hayfa AMARA

Supervisor :
Prof. Ameni AZZOUZ

Academic year : 2024/2025

Contents

List of Figures

GitHub Link

Google Colab Link

1 Introduction

This document talks about a small project in Business Intelligence and Database Management Systems. It mainly looks at supermarket sales data. We will use Python for ETL tasks on basic data in CSV and JSON forms, changing it into useful information. MySQL will help in storing and organizing data properly, making sure it can grow later as needed. Power BI will be used for visual analysis, allowing stakeholders to see trends and make decisions based on data. Our focus is on products, categories, sales, and customer info to improve operations, boost sales, and support ongoing growth.

1.1 Methodology

The processes related to extraction, transformation, and loading (ETL) are performed with Python, where the raw data from CSV and JSON files are transformed into structured data and loaded into MYSQL. Data is captured and stored in MYSQL as the system's database. Power BI is utilized for Relational Online Analytical Processing (ROLAP) and data visualization, which includes interactive reports and dashboards. These systems reveal trends and patterns, making it possible for stakeholders to make decisions that have a positive impact on supermarket operations and profitability.

1.2 Key Objectives:

1. **Analyzing Sales Performance:** Our primary objective is to evaluate how various factors, such as product categories, sub-categories, and regions, impact overall sales performance, including metrics like quantity sold, discount, and profit.

2. **Identifying Operational Improvements:** By examining data such as order dates, ship dates, and customer segments, we aim to provide insights into underperforming regions or customer segments. This will guide decisions on improving logistics and customer service efficiency.

3. **Supporting Business Expansion and Optimization:** Using insights gained from customer demographics, regions, and product performance, we aim to provide strategic recommendations for business growth. This includes determining areas for territorial expansion, identifying high-potential markets, and re-evaluating product offerings to optimize the catalog for better profitability.

4. **Optimizing Product Offerings:** By analyzing the relationships between products, categories, and sub-categories, we aim to identify the most profitable products and product groups, allowing the business to make informed decisions on inventory and product placement.

2 Work Explanation

2.1 Data Gathering

Our data exploration involved utilizing the StoreSales and SuperStoreData datasets, sourced from Kaggle. Through a careful selection process, we curated the necessary files to align with our Key Performance Indicators (KPIs). This combination of CSV and JSON formats provides a comprehensive source for analysis. The dataset we worked with contain this combinaison of two data files:

- StoreSales.JSON
- SuperStoreData.CSV

2.2 ETL Process

In this project, we did ETL on two datasets using Python: a JSON file (4002 rows, 23 columns) and a CSV file (2000 rows, 20 columns), with the purpose of integrating them into a single dataset.

2.2.1 Data Extraction

In Google Colab, we imported the files into Python using the Pandas package to do initial analysis and compatibility tests. We validated the data structure, ensuring that all fields were correctly recognized and matched the expected formats. This phase prepared the datasets for future processing in the ETL pipeline.

Figure 1: Screenshots Illustrating the Data Extraction Process

2.2.2 Data Transformation

- **For the CSV file:**
  - Converted 'Order Date' and 'Ship Date' to datetime.
  - Created a new column to calculate Order Processing Time.
- **For the JSON file:**
  - Converted 'Order Date' and 'Ship Date' to datetime.
  - Converted numeric object columns (e.g., Sales, Profit, Discount) to float.
  - Calculated Processing Time.
  - After checking for missing data, the 'Postal Code' column was removed because it had 3340 unnecessary missing values.
- **Shared Columns:**
  - We aligned the shared columns between the datasets to prepare for concatenation.

```python
# Convert 'Order Date' and 'Ship Date' to datetime
json_df['Order Date'] = pd.to_datetime(json_df['Order Date'], format='%d-%m-%Y')
json_df['Ship Date'] = pd.to_datetime(json_df['Ship Date'], format='%d-%m-%Y')

# Convert object columns that should be numeric to float
json_df['Sales'] = pd.to_numeric(json_df['Sales'], errors='coerce')
json_df['Quantity'] = pd.to_numeric(json_df['Quantity'], errors='coerce')
json_df['Discount'] = pd.to_numeric(json_df['Discount'], errors='coerce')
json_df['Profit'] = pd.to_numeric(json_df['Profit'], errors='coerce')
json_df['Shipping Cost'] = pd.to_numeric(json_df['Shipping Cost'], errors='coerce')

#calculate processing time
json_df['Processing Time'] = (json_df['Ship Date'] - json_df['Order Date']).dt.days
```

```python
# Check for missing or empty/whitespace values in all columns
empty_or_missing = json_df.apply(lambda x: x.isnull() | (x.str.strip() == '') if x.dtype == "object" else x.isnull()).sum()

print(empty_or_missing)
```

```
Order ID              0
Order Date            0
Ship Date             0
Ship Mode             0
Customer ID           0
Customer Name         0
Segment               0
City                  0
State                 0
Country               0
Postal Code        3340
Market                0
Region                0
Product ID            0
Category              0
Sub-Category          0
Product Name          0
Sales                 0
Quantity              0
Discount              0
Profit                0
Shipping Cost         0
Order Priority        0
Processing Time       0
```

Figure 2: Screenshots Illustrating the Data Transformation Process

2.2.3 Data Loading

For additional analysis, the converted data was stored in Excel and CSV formats. Then, we utilized MySQL Workbench to store the merged dataset. First, we created a table with a structure matching the column order of the dataset. Then, we imported the data using the Table Data Import Wizard, which facilitated the seamless transfer of the CSV file into the MySQL database. This step ensured the data was centralized and ready for further analysis and visualization.

```
# Save the concatenated DataFrame to CSV in the Colab environment
combined_df.to_csv("/content/store1_data.csv", index=False)

print("Data saved as CSV in Colab environment as '/content/store1_data.csv'")

from google.colab import files

# Download the CSV file
files.download("/content/store1_data.csv")
```

Figure 3: Loading Data into .csv Format

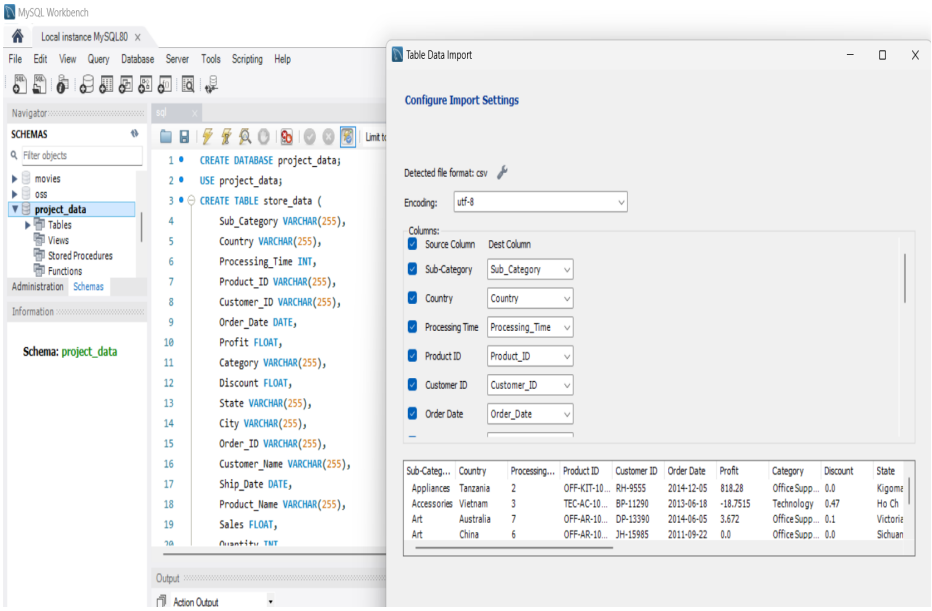

Figure 4: Loading Data into MySQL Workbench

2.3 Data Modeling

This section explores the data modeling process, emphasizing the organization of our dataset to support effective analysis and reporting. The core elements of our data model include the Fact Table, Dimension Tables, Measures, and the selected schema structure. This thoughtful design lays the groundwork for extracting meaningful insights and driving informed decisions based on the Store dataset. It was designed using Miro, a specialized application that facilitates visual collaboration and data modeling, ensuring a clear and organized representation of our dataset.

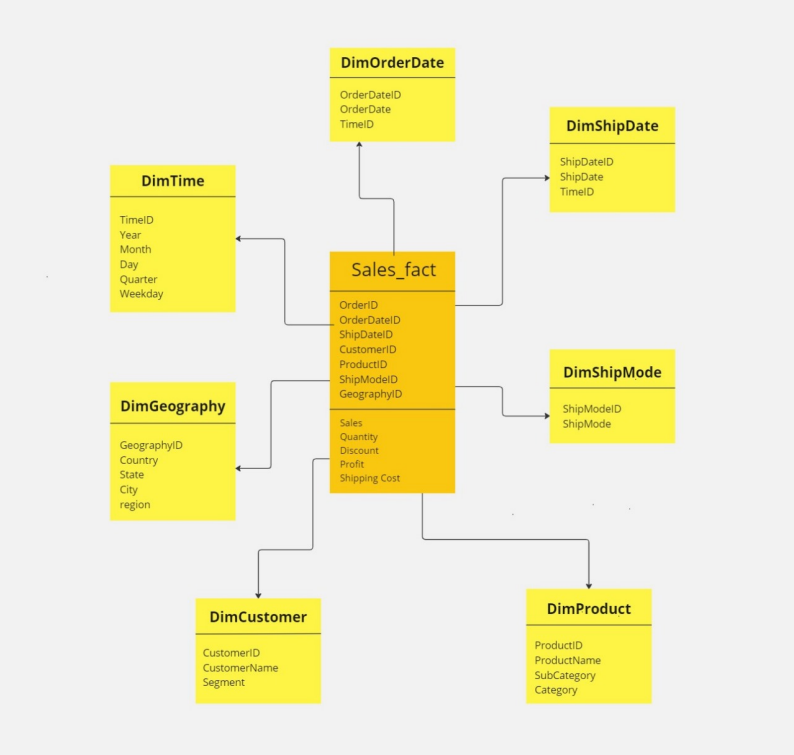

Figure 5: Star Schema

2.3.1 Fact Table

At the heart of our data model lies the Sales Fact Table, which serves as the primary repository for essential data on sales transactions. It records key identifiers such as Order ID, Customer ID, Product ID, Geography ID, Order Date ID, Ship Date ID, and Ship Mode ID, alongside many measures like Sales, Quantity, Discount, Profit, and Shipping Cost. Together, these elements form the foundation of our analysis, offering deep insights into sales performance, customer preferences, regional trends, and the overall efficiency of operations.

2.3.2 Dimensions

To enrich our analysis, we leverage multiple Dimension tables, each offering a distinct angle to interpret and understand the data comprehensively. These dimensions include:

- **DimOrderDate**: Provides a temporal perspective by identifying when orders were placed, enabling trend analysis and time-based performance evaluation.
- **DimShipDate**: Offers insights into shipping timelines by tracking when orders were shipped, facilitating the analysis of delivery performance and delays.

- **DimTime**: Adds a detailed temporal dimension by breaking down time into granular components like year, quarter, month, and weekday, supporting seasonal and time-of-day analyses.
- **DimGeography**: Provides a geographical perspective by categorizing data by location (region, country, state, city), enabling regional sales analysis and location-based strategy formulation.
- **DimCustomer**: Brings a customer-centric view by segmenting data based on customer information and classification, supporting targeted marketing and customer relationship analysis.
- **DimShipMode**: Focuses on the logistics aspect by categorizing shipping methods, aiding in evaluating shipment efficiency, cost-effectiveness, and customer preferences.
- **DimProduct**: Delivers a product-centric analysis by categorizing products into names, subcategories, and categories, enabling insights into product performance and inventory management.

2.3.3 Measures

Our analysis relies on key quantitative metrics, referred to as Measures, to evaluate performance and uncover trends. In the context of our data model, the primary Measures include:

- **Quantity**: This represents the number of units purchased in each order. It helps assess product demand and identify popular items.
- **Profit**: Calculated as the revenue minus costs (including product and shipping costs). It indicates the financial success of orders and helps pinpoint high-margin products or regions.
- **Discount**: The percentage or amount deducted from the product price during a sale. This measure provides insight into promotional strategies and their impact on sales performance.
- **Shipping Cost**: The cost incurred for delivering an order to the customer. This measure highlights logistical expenses and helps identify opportunities to optimize shipping methods.

- **Sales**: The total revenue generated from orders. This measure is essential for understanding overall performance and identifying top-performing products or categories.

2.3.4 Purpose of Schema and Model

The Star Schema was selected for its simplicity and efficiency in analyzing the SuperStore and Store datasets. With a centralized fact table and clear dimensions, it supports seamless querying and visualization. This design enables focused analysis of key metrics like sales, profit, and shipping costs, uncovering insights into customer behavior, product trends, and regional performance. The model empowers data-driven decisions and strategy optimization for business success.

2.4 OLAP and Data Visualization Process

We utilized Power BI as our OLAP and data visualization tool to unlock meaningful insights from our comprehensive data model. Power BI allowed us to transform raw data into dynamic, interactive reports that provided valuable visualizations. The key steps in this process are:

2.4.1 Data Extraction and Transformation

After establishing the connection, we imported the StoreData file from MYSQL into Power BI. Leveraging Power BI's data transformation tools, we refined the dataset to meet our project's specific analytical requirements. This process included data cleansing, filtering, and restructuring to ensure the highest quality for our visualizations.

2.4.2 Interactive Dashboards and Reports

Power BI's drag-and-drop interface enabled us to create interactive dashboards and reports effortlessly. We developed dynamic dashboards to show the most important findings from the combined dataset. These dashboards offered actionable insights for data-driven decision-making by highlighting parameters like shipping performance, profit margins, and sales trends.

Figure 6: Screenshots Illustrating the Sales Dashboard

2.4.3 Insights and decision support

By utilizing Power BI's features, we could effectively analyze our data and create interactive, insightful reports highlighting key trends and patterns, aiding in data-driven decision-making.

- **Segment Sales Overview (Pie Chart)**: The Consumer Segment emerges as the dominant revenue driver, accounting for 47.49% of total sales. The Corporate Segment follows with 30.28%, and the Home Office Segment contributes 22.23%.

While the Consumer Segment is performing well, there is untapped potential in the Corporate and Home Office Segments, which can be targeted for growth to diversify revenue streams and enhance overall profitability.

- **Bestselling Sub-categories (Bar Chart):** High demand is evident for Binders and Chairs in the Office Supplies and Furniture categories, respectively. Machines also lead in the Technology category, emphasizing their importance in the product mix. Additionally, Furnishings and Copiers show robust sales, suggesting these sub-categories should remain focal points in inventory and marketing strategies.
- **Top 10 Customers by Sales (Bar Chart):** Highlights the significant influence of key customers on revenue. Sean Miller holds the highest sales volume, followed by Tom Ashbrook and Ken Lonsdale. These customers illustrate how a small group can substantially impact overall sales, underscoring the need to nurture these high-value relationships.
- **Sales Distribution by Country (Map):** The United States dominates as the primary market, with notable contributions from France, Mexico, and Australia. Meanwhile, countries like India, Brazil, and Italy display lower sales volumes but offer potential for growth through targeted marketing and localized strategies.
- **Top Products by Quantity (Bar Chart):** Identifies Binders and Paper as the top-selling products, followed by Art Supplies and Storage Products. However, products like Fasteners and Labels show lower sales volumes, indicating insufficient demand or promotion. This suggests a need to revisit promotional strategies or evaluate their relevance in the product portfolio.
- **Orders by Shipping Mode (Donut Chart):** Standard Class accounts for 63% of orders, making it the most utilized shipping mode. Second Class contributes 19.41%, and premium options collectively represent 4.52%. The data highlights the importance of ensuring cost-effective and reliable logistics for Standard Class while promoting other shipping options to balance demand.
- **Processing Time by Region (Bar Chart):** Reveals bottlenecks in the Central region, which has the longest processing times. In contrast, Oceania exhibits the shortest processing times, showcasing operational efficiency.

This disparity underscores the need to replicate successful practices from Oceania in slower regions.

- **Top 5 Countries by Sales (Pie Chart):** The U.S. accounts for 91.16% of total sales, highlighting an over-reliance on a single market. France, Mexico, Australia, and Germany contribute to the remaining sales. The data emphasizes the need to diversify market focus while maintaining strong performance in the U.S.
- **Top Markets by Sales (Bar Chart):** Shows that the European Union (EU) leads in sales, followed by Latin America (LATAM) and the Asia-Pacific (APAC) regions, which also show strong sales performance. The United States (US) and Africa are moderate performers, while EMEA and Canada lag significantly in sales.
- **Shipping Cost by Country (Map):** Highlights high shipping costs in North America and Europe, which present opportunities for cost optimization. Conversely, low-cost regions like Africa and South America can be leveraged for promotional campaigns emphasizing competitive shipping rates.

3 Conclusion

To seize emerging possibilities and handle current challenges, we advocate the following strategic actions:

- Focus on building stronger relationships with top-performing customer groups.
- Optimize popular product sub-categories to stay competitive.
- Expand into underperforming markets like India and Brazil.
- Diversify beyond the U.S. to reduce dependence on a single region.
- Apply successful practices from high-performing regions to others.
- Enhance profitability by optimizing shipping logistics and using cost-effective solutions.
- Focus on expanding market penetration in EMEA and Canada by conducting market research to understand barriers and tailoring marketing strategies to increase sales in these underperforming regions.