TUNIS BUSINESS SCHOOL
UNIVERSITY OF TUNIS

BI Project

E-shop sales analysis

IT300 DBMS BI Final Project

Malek Ben Hamed
Fady Iben Habel
Islem Jeridi
Zakaria Ayadi

Submitted to Prof. Manel Abdelkader and Prof. Amira Azzouz

January 2024

Tunis Business School
Ben Arous, TUNISIA

2023-2024

1 Introduction

This Business Intelligence project revolves around the analysis of an e-shop's performance during the first quarter of 2019 data.The primary goal is to investigate the impact of various factors, including Gender,Age,Product line,Customer type and country of origin on sales. By delving into these relationships, we seek to extract valuable insights into consumer behavior and preferences. This understanding can help owners enhance their offerings and, consequently, attract more business. Through this project, we aim to improve overall customer satisfaction and build better strategies for the e-shop.

1.1 Main Goals

The main goals of this project include:

- Conducting a broad study of 2019's first quarter sales.
- Providing insights for strategic decision-making.
- Using a BI dashboard to display key performance indicators (KPIs).
- Integrating data from multiple sources for effective analysis.

1.2 Business Insights

The business insights sought include:

- Quantity sold by country.
- Customer demographics and preferences.
- Financial metrics such as total income, cost, and profit.
- Analysis of order and product-related dimensions.
- Difference between members and normal customers behavior .

2 Implementation Stages

2.1 Data Gathering (Phase 1)

In the initial phase, raw data is collected from disparate source systems or databases. The Python code snippet below demonstrates the process of reading and merging data from CSV files.

Listing 1: Data Gathering

```
import pandas as pd

# Reading data
df = pd.read_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\Final.csv")
cd = pd.read_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\CustF.txt", delimiter
    ↪ =",", encoding='latin1')

# Checking data and data types
print(df)
print(cd)
print(cd.columns)
print("Data type in df:", df['Customer_ID'].dtype)
print("Data type in cd:", cd['Customer_ID'].dtype)

# Merging data on 'Customer_ID'
md = pd.merge(df, cd, on='Customer_ID', how='right')

# Extracting and saving the merged data
md.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\mergedV1.csv")
```

2.2 Data Preparation (Phase 2)

The data preparation phase involves cleaning and transforming the data. This includes concatenating first name and last name columns, handling date formats, and dealing with missing values.

Listing 2: Data Preparation

```
# Concatenating first name and last name columns
md['CustFN'] = md[' Customer_FirstName'] + ' ' + md[' Customer_LastName']
md = md.drop([' Customer_FirstName', ' Customer_LastName'], axis=1)

# Extracting and saving the transformed data
md.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\transV1.csv")

# Handling missing values and date formats
NaN = md.isna().sum()
md['Birth_Date'] = pd.to_datetime(md['Birth_Date'], format='%m/%d/%Y', errors='
    ↪ coerce')
md['Date'] = pd.to_datetime(md['Date'], format='%m/%d/%Y', errors='coerce')
age_in_years = (md['Date'] - md['Birth_Date']) / pd.Timedelta(days=365.25)
md['age_when_order'] = age_in_years.astype(float)
md.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\transV2.csv")
```

2.3 Data Storage and Modeling (Phase 3)

The data storage and modeling phase involve further transformations and the creation of dimensions and a fact table.

Listing 3: Data Storage and Modeling

```
# Extracting month and changing month numbers to month names
md['Date'] = pd.to_datetime(md['Date'])
md['month'] = md['Date'].dt.month
md.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\transv3.csv", index=False)
md['month'] = md['month'].apply(lambda x: calendar.month_name[x])
md.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\transV4.csv", index=False)

# Changing unit price format and calculating total income per order
md['Unit price'] = pd.to_numeric(md['Unit price'], errors='coerce')
md['Unit price'] = md['Unit price'].astype('float')
md['total_income'] = md['Unit price'] * md['Quantity']
md.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\transV5.csv", index=False)

# Changing unit cost format and calculating total cost per order
md['Unit cost'] = pd.to_numeric(md['Unit cost'], errors='coerce')
md['Unit cost'] = md['Unit cost'].astype('float')
md['total_cost'] = md['Unit cost'] * md['Quantity']
md.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\transV6.csv", index=False)

# Calculating total profit per order and changing shipping cost format
md['total_profit'] = md['total_income'] - md['total_cost']
md['shipping cost'] = pd.to_numeric(md['shipping cost'], errors
```

2.4 Extracting the fact table and dimensions (Phase 4)

Listing 4: Extracting the fact table and dimensions

```
Time_dim = md[['Date', 'Time', 'month']].copy()
Time_dim.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\F_D\\TimeDim.csv",
    ↪ index=False)
Cust_dim = md[['Customer_ID', 'Gender', 'age_when_order', 'CustFN', 'Country', '
    ↪ Customer type']].copy()
Cust_dim.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\F_D\\CustDim.csv",
    ↪ index=False)
Prod_dim = md[['Product line', 'Unit price', 'Unit cost', 'Rating']].copy()
Prod_dim.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\F_D\\ProdDim.csv",
    ↪ index=False)
Order_dim = md[['Invoice ID', 'Product line', 'Date', 'Quantity', 'Shipment_mode
    ↪ ', 'shipping cost', 'Payment', 'warehouse_name']].copy()
Order_dim.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\F_D\\OrderDim.csv",
    ↪ index=False)
Fact_tab = md[['Customer_ID', 'Invoice ID', 'month', 'total_income', 'total_cost
    ↪ ', 'total_profit']].copy()
Fact_tab.to_csv("C:\\Users\\malek\\OneDrive\\Bureau\\Bi2\\F_D\\FactTab.csv",
    ↪ index=False)
```

2.5 Star Schema

The star schema is a crucial part of our data modeling. It consists of a fact table (FactTab) surrounded by dimension tables (TimeDim, CustDim, ProdDim, OrderDim). This schema simplifies data querying and enhances the overall performance of the BI system.

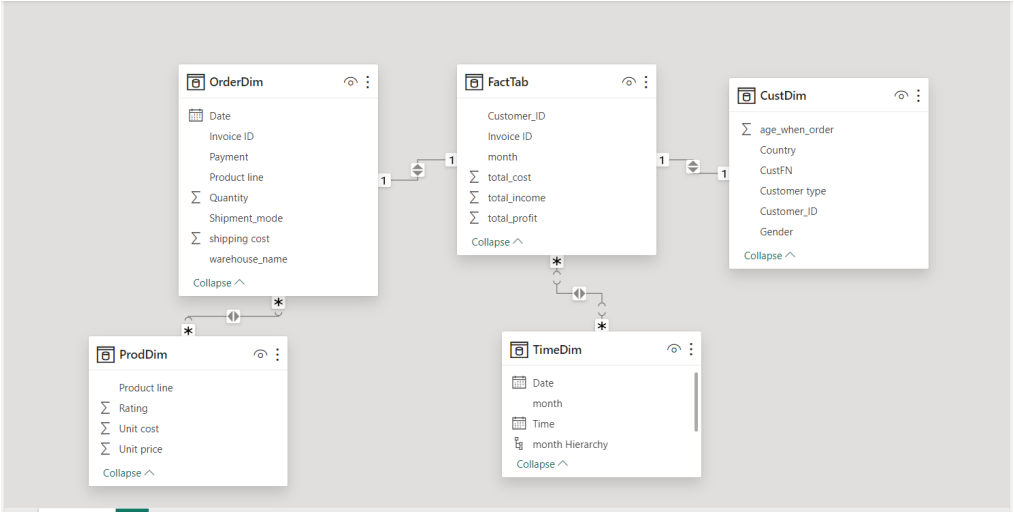


Figure 1: Star Schema

3 Power BI Dashboards

3.1 Sales and Transaction Metrics Dashboard

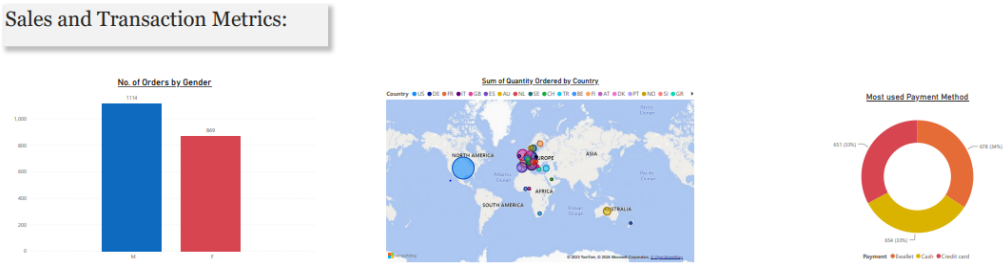


Figure 2: Sales and Transaction Metrics Dashboard

Comment: Provides insights and analysis related to sales and transaction metrics per country and per payment mode.

3.2 Customer Demographics and Behavior Dashboard

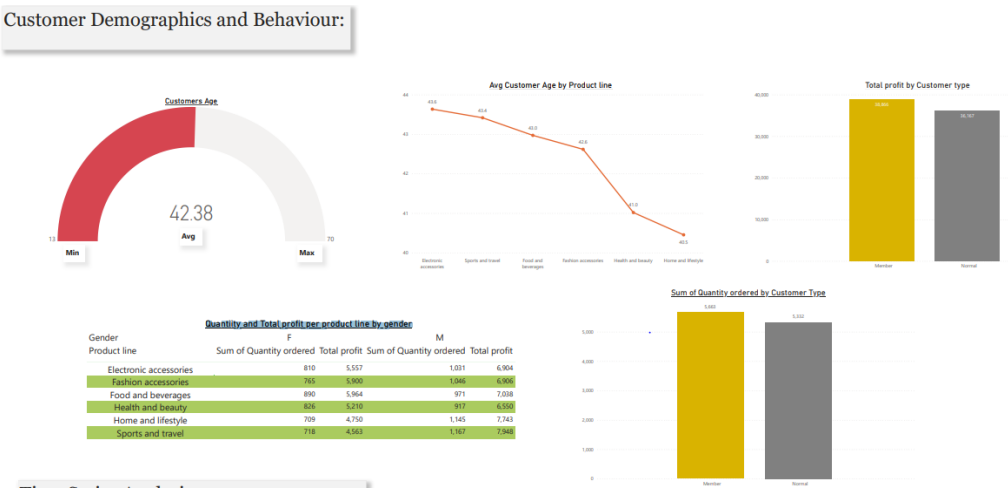


Figure 3: Customer Demographics and Behavior Dashboard

Comment: Analyzing customer demographics and behavior patterns.

3.3 Monthly and Product Line Insights Dashboard

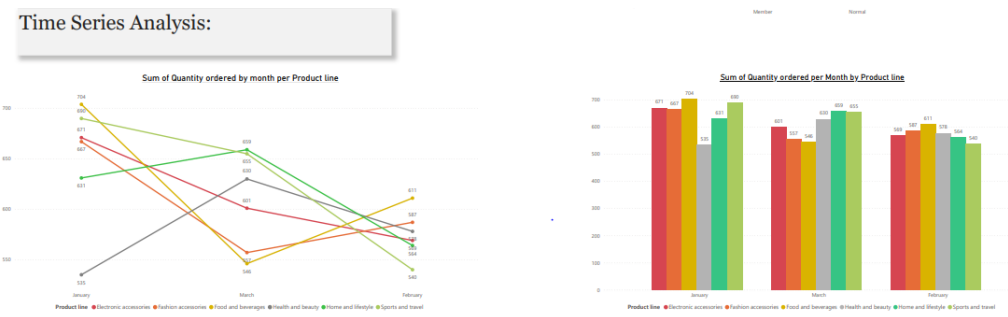


Figure 4: Monthly and Product Line Insights Dashboard

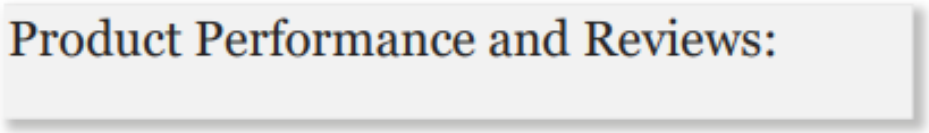
Comment: Exploring insights related to monthly trends and product line performance.

3.4 Logistics and Warehouse Metrics Dashboard

Figure 5: Logistics and Warehouse Metrics Dashboard

Comment: Evaluating logistics and warehouse-related metrics for optimization.

3.5 Product Performance and Ratings Dashboard



Min, Avg, Max Rating per Product line

Product line	Min Rating	Avg Rating	Max Rating
Electronic accessories	2.50	3.65	5.00
Fashion accessories	2.50	3.77	5.00
Food and beverages	2.50	3.78	5.00
Health and beauty	2.50	3.80	5.00
Home and lifestyle	2.50	3.72	5.00
Sports and travel	2.50	3.77	5.00

Figure 6: Product Performance and Ratings Dashboard

Comment: Assessing product performance and ratings for strategic decisions.

4 Insights

4.1 Observations

- The majority of our customer base is concentrated in the US and Western Europe.
- Remarkably, there are very few orders originating from Africa, Asia, and South America.
- Quantity-wise, there's minimal disparity between orders placed by members and those by regular customers.
- The average shipping cost remains relatively consistent across both customer types.
- Members, on average, incur higher shipping costs than regular customers.
- Our primary demographic target appears to be middle-aged individuals.
- A discernible downtrend characterizes the overall sales trajectory.
- Across various warehouses, shipments are evenly distributed, signifying the equal importance of each warehouse in fulfilling orders.

4.2 Proposed Solutions

- Explore marketing strategies to increase sales in underrepresented regions (Africa, Asia, South America).
- Investigate the reasons behind the downtrend in sales and implement strategies to boost sales.
- Analyze the shipping cost structure and explore ways to optimize it for different customer types.
- Consider targeted promotions for different age groups to attract a broader customer base.
- Evaluate the performance of warehouses and identify opportunities for improvement in logistics.

5 Conclusion

In conclusion, this Business Intelligence project has provided in-depth insights into the company's first-quarter performance in 2019. Our systematic progression through the stages of data gathering, preparation, storage, and analysis culminated in the development of insightful Power BI dashboards and the implementation of a well-structured star schema.

The observations gleaned from the data offer a nuanced understanding of various facets of the company's operations, and the proposed solutions strategically target areas for improvement. These insights, complemented by the clear visual representation in Power BI dashboards and the efficiency of the star schema, equip the company for informed decision-making and set the foundation for future success.

Working on this project has been an enriching experience. Despite encountering more challenges than initially expected, we are grateful for the valuable lessons and substantial growth, both academically and personally. We extend our gratitude to Prof. Manel Abdelkader and Prof. Amira Azouz, our professors, whose guidance has been instrumental in shaping our learning journey and bringing us to this point. We express our sincere appreciation for their significant contributions and unwavering support.