2023/2024

DEPARTMENT STORE ANALYSIS REPORT

BUSINESS INTELLIGENCE AND DATABASE MANAGEMENT SYSTEMS

Work done by :

Sarah Ibn El hadj - Wadie Tliche
Becher Zribi - Nour Attia

Professors:

Prof.Manel Abdelkader
Prof.Ameni Azzouz

TABLE OF CONTENTS

INTRODUCTION

In the dynamic retail landscape, our research focuses on understanding **shopping trends** in **department stores in the United States**, providing insights for strategic planning. Utilizing a **comprehensive dataset covering customer demographics**, **purchasing patterns**, **product preferences**, and **inventory levels**, we aim to enhance decision-making for businesses.

By Examining various information from the gathered datasets, our primary objective is to distil key insights to offer a nuanced understanding of consumer preferences and industry trends, aiding businesses in navigating the complexities of the local United States retail market.

INTRODUCTION

Concretely, we focused on 4 important key performance indicators(KPI):

1. **Demographic Analysis:** By understanding the relationships between marital status, income and location distribution of customers to identify our potential target markets and tailor customized marketing strategies to increase the demand.
2. **Product Preferences:** By examining the types items customers purchase, their preferred category (furniture, office suppliers, technology), their preferred sub-category, seasonal inclinations, we can optimize our inventory management and product assortment.
3. **Purchase Behaviour:** By studying the frequency of purchases and shipping mode to optimize the overall shopping experience.
4. **Inventory Optimization:** By examining the quantity on hand, the total costs and the inventory turnovers to update our management strategies and reduce expenses.

By achieving these goals, businesses can attain a more profound insight into their customer base. This, in turn, empowers them to **fine-tune marketing strategies**, **optimize inventory management**, and **elevate overall customer satisfaction**. Ultimately, these efforts contribute to sustained business growth in the competitive retail landscape.

2.IMPLEMENTATION

2.1.DATA GATHERING:

We obtained the dataset on Shopping Trends ratings in the USA from Kaggle.
This is a link to the dataset.

2.2.DATA PREPARATION:

Preparing the data , we utilized Python to streamline the manipulation and configuration of our dataset for integration into the data warehouse. The dataset existed in two formats: JSON and CSV.

Upon careful examination, a critical observation revealed the existence of duplicated data and Non-Available data. To tackle this issue, we utilized Python to identify and rectify duplicate rows within the dataset. We also converted data into the right data types.

Untitled-1

```python
# %%
import pandas as pd
import matplotlib.pyplot as plt
import json
import numpy as np
from datetime import datetime

# %% [markdown]
# Exploring the first Dataset

# %%
df= pd.read_csv('sales_data.csv')

# %%
df.info()

# %%
df.describe()

# %%
#checking for null values
df.isnull().sum()

# %% [markdown]
# Treating the first dataset
#

# %%
#tracking the null values
null_rows = df.loc[df['postal_code'].isnull(), ['postal_code', 'region','city','state']]
print(null_rows)

# %%
print(df.loc[df['city']=='Burlington',['postal_code','state','city']])

# %%
#correcting the null values
df['postal_code'] = df['postal_code'].fillna('05402')
```

```python
df.isnull().sum()

# %%
#converting dates to the date datatype
df['order_date']= pd.to_datetime(df['order_date'], format='%d/%m/%Y')
df['ship_date']= pd.to_datetime(df['ship_date'], format='%d/%m/%Y')
df.info()

# %% [markdown]
# Treating the second dataset

# %%
with open('customer_ds.json', 'r') as json_file:
    df2 = json.load(json_file)

# %%
#exploring the structure
# Check if data is a list
if isinstance(df2, list):
    # Access the first element in the list
    first_element = df2[0]

    # Display keys of the first dictionary in the list
    print("Keys of the first dictionary in the list:", first_element.keys())
else:
    print("The data is not a list of dictionaries.")

# %%
print(pd.json_normalize(df2).describe())

# %%
print(pd.json_normalize(df2).isnull().sum())

# %%
# Extract non-null incomes
incomes = [item.get('Income', 0) for item in df2]

# Calculate the mean of non-null incomes
mean_income = np.mean([income for income in incomes if income is not None])

# Replace null incomes with the mean
```

```python
for item in df2:
    if item.get('Income') is None:
        item['Income'] = mean_income

# %%
print(pd.json_normalize(df2).isnull().sum())

# %%
#convertin to the right data types
for item in df2:
    if 'Year_Birth' in item:
        try:
            year_as_int = int(item['Year_Birth'])
            item['Year_Birth'] = datetime(year_as_int, 1, 1)  # Assuming January 1st of the given year
        except ValueError:
            print(f"Invalid year format for item: {item}")

# %% [markdown]
# Building Dimensions
#

# %%
#product dimension

# %%
pdm = df.loc[:, ['product_id','product_name','sub_category','category']].copy()
pdm.head()


# %%
#eliminating the duplicates and adding the quantities :

# Group by 'id' and count occurrences
pdm['quantity'] = pdm.groupby('product_id')['product_id'].transform('count')

# Drop duplicates from the 'id' column
pdm.drop_duplicates(subset='product_id', inplace=True)

# Reset index after dropping duplicates
pdm.reset_index(drop=True, inplace=True)
```

2.IMPLEMENTATION

2.3 DATA STORAGE :

2.3.1.STORAGE:

After obtaining the refined version of the data we chose csv files to serve as a staging area.

```
# %%
pd.json_normalize(df2).to_csv('cleaned_cus')

# %%
dp.to_csv('cleaned_data.csv')

```

We used python to create the Desired Fact and dimensions , and using the same tool we exported the output to our data warehouse which is MySQL.

ESTABLISHING CONNECTION

```python
import mysql.connector

# %%
conn = mysql.connector.connect(
    host='localhost',
    user='root',
    password='Wdie.66.99',
    database='warehouse2'
)

# %%
cursor = conn.cursor()
```

DEFINED FUNCTION TO CREATE TABLES

```python
#creating a function to insert the data
def create_table_from_excel(cursor, excel_file_path, table_name):
    # Load Excel data into a Pandas DataFrame
    df = pd.read_excel(excel_file_path)

    # Create the table
    create_table_query = f"CREATE TABLE {table_name} ({', '.join([f'{col} VARCHAR(255)' for col in df.columns])})"
    cursor.execute(create_table_query)
```

DEFINED FUNCTION TO INSERT THE DATA

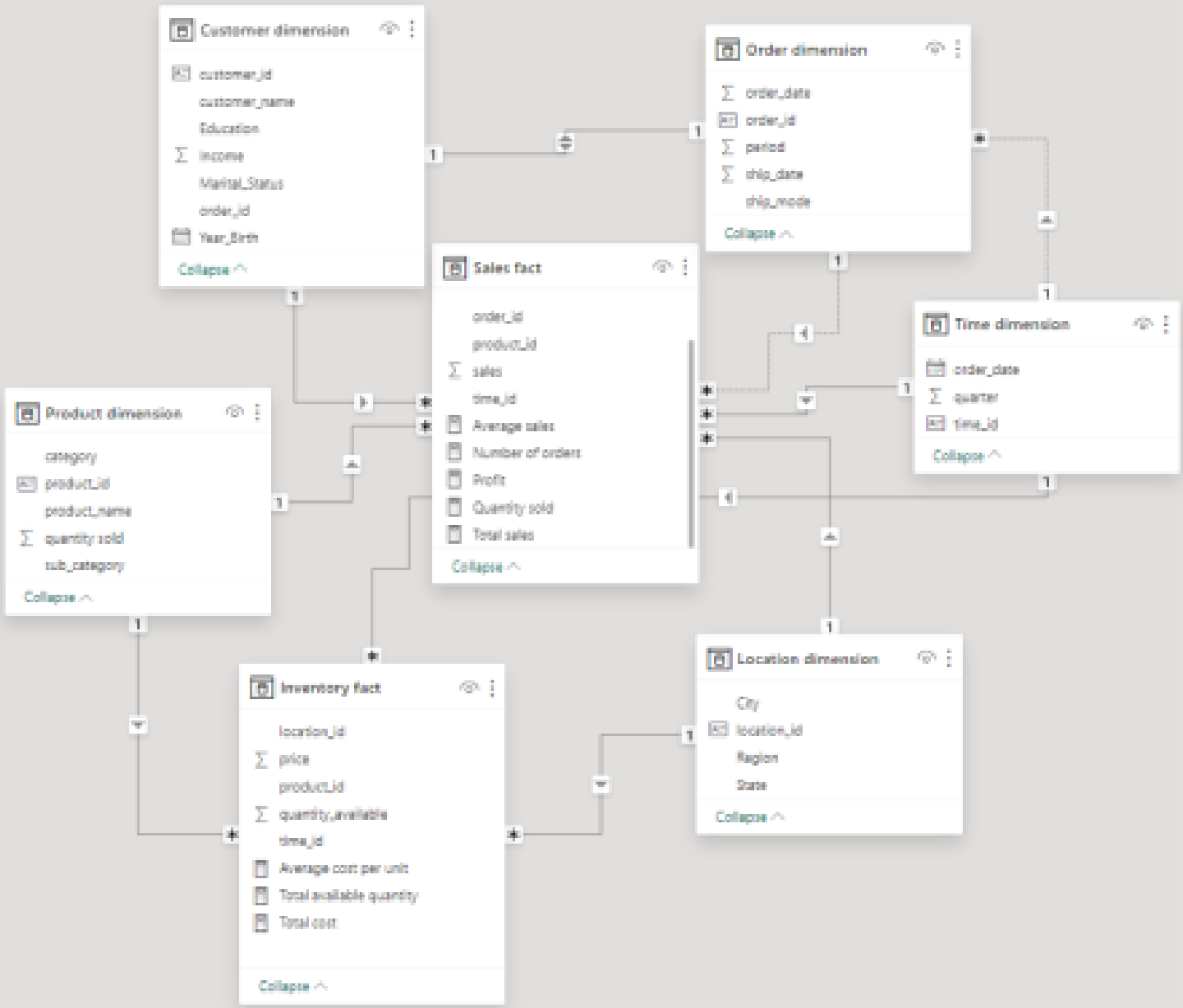
```python
def insert_data_from_excel(cursor, excel_file_path, table_name):
    # Load Excel data into a Pandas DataFrame
    df = pd.read_excel(excel_file_path)

    # Create the INSERT query
    columns = ', '.join(df.columns)
    values_placeholder = ', '.join(['%s' for _ in df.columns])
    insert_query = f"INSERT INTO {table_name} ({columns}) VALUES ({values_placeholder})"

    # Iterate through rows and insert data
    for index, row in df.iterrows():
        cursor.execute(insert_query, tuple(row))
```

Schema choice: Snowflake

We decided to implement the Snowflake Schema to optimize data organization and minimize redundancy in our database. Breaking down dimension tables into interconnected tables facilitates more efficient updates and modifications. While this may result in increased query complexity due to necessary joins, the overall benefits encompass improved scalability, simplified maintenance, and enhanced performance in specific scenarios.



2.IMPLEMENTATION

2.3 DATA STORAGE :

2.3.2.FACT:

We identified two Fact tables :

- **Sales fact** that holds the informations about the order, the product, the customer, the time, the location, and contained measures like average sales, number of orders and profit.
- **Inventory fact** that holds the details about the product, time, locations, quantity available and price.

The two fact tables allowed us to gain insight about important keys related to sales.

2.IMPLEMENTATION

2.3 DATA STORAGE :

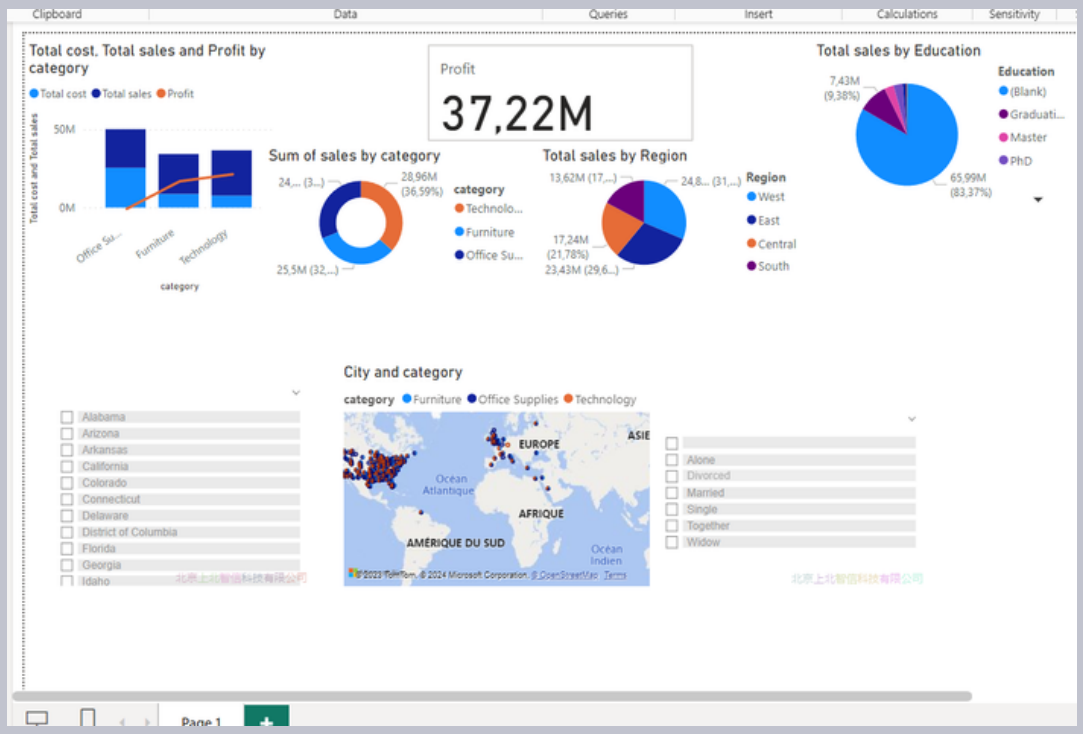
2.3.3.DIMENSIONS:

To enrich our analysis, we implied various dimension tables each providing a distinct perspective on the data. These dimensions include:

- **Customer dimension**: represents the customer infomartion.
- **Product Dimension**: represents the product information.
- **Location dimension**: represents the different locations of transactions.
- **Order Dimension**: Represents different characteristics of the order.
- **Time Dimension**: Contains information about the order time, shipment...

2.IMPLEMENTATION

2.4 DATA VISUALIZATION:



- The total profit amount is 37.22 M
- Technology prove to be the most profitable item generating 21.29M while office supplies are the least profitable with lost of 855.26 K
- California is our top selling state with 15.62M in sales followed by new york with 10.72
- Most of sales are coming from married people with 941 K and together generating 527 K

INTERPRETATION:

- the average quantity available is half about 5 M and it ranges from0 to 10 M.
- Total Sales by Product Category" provides a comprehensive view of our sales distribution across different product categories.
- Total Sales by Product Category" provides a comprehensive view of our sales distribution across different regions

- The Product Profitability Analysis, visualized through a stacked bar chart, serves as a powerful tool for businesses seeking to optimize their product offerings, pricing strategies, and overall profitability.

3. CONCLUSION

In summary Our research focuses on understanding how people shop in the ever-changing retail world of the USA. We use a big dataset covering customer demographics, buying habits, product preferences, and inventory levels to help businesses make smart decisions. We also explore global trends in office supplies, furniture, and technology retail, studying both big and new players to provide key insights for businesses navigating the dynamic global market.

CHALLENGES FACED

We Faced several Obstacles throughout our project journey:

- **Challenges in data gathering phase:** The research was not easy, we faced difficulties in tracking the ideal dataset for our project.
- **challenges in data quality:** managing numerous duplicates and navigating through misleading data introduced complexity to the project. insuring the accuracy and reliability of the data became a crucial task.
- **Challenges in using talend:** Our group encountered challenges that prevented us from utilizing talend effectively.
- **Challenges in Power BI:** one oversight has the potential to affect the entire of our progress. To overcome these challenges, innovative problem-solving and iterative methods were employed to ensure the successful execution of our data management and visualization project.