Project Report

Business Intelligence (IT 300)

Retail Store Analysis

Prepared by: Eya Trabelsi
Aya Essid
Islem Hamzaoui
Yosr Boussarsar

Table Of Content:

1. Introduction

This retail store sales analysis project is dedicated to examining data related to customer transactions. The primary objective is to investigate the impact of various factors, including customer traits like annual income, gender, status, age, and the types of products purchased, on retail store sales and revenues. The goal is to unravel the connections between these variables, providing valuable insights into customer behavior and preferences. By gaining a deeper understanding of these relationships, we aspire to assist retail stores in enhancing their offerings, thereby increasing customer satisfaction, and devising more effective strategies for the retail industry.
The primary aim is to analyze transactional trends, specifically focusing on sales, to discern and prioritize valued consumers based on their purchase history. The objective is to identify the most appealing products to these consumers and pinpoint periods with heightened sales, enabling the strategic preparation of adequate merchandise stock for peak periods. This approach provides the retail store manager with the necessary tools and insights to make informed decisions, enhancing overall operational planning and ensuring readiness for high-demand phases.

2. Project Phases

2.1 Data Gathering

The data sets crucial to our projects were sourced from Kaggle, offering a comprehensive foundation for analysis. This platform provides useful transactional details, retail store-specific information and consumer preferences, allowing us to explore factors like income, activity, purchase history, and retail store types that influence consumer behavior.
Here are the multiple data sets we worked on:

1- Customer Segmentation: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python?fbclid=IwAR2Y7TVm911KzDkkj6E4sfLcvpHXfoWJpewXbLGDhCK7_c0hQI-2UJMrPjM

2- Retail transactions dataset
https://www.kaggle.com/datasets/prasad22/retail-transactions-dataset?fbclid=IwAR0uU5ZaKUwrPFpfYXTB3TVjUCt5NpfmSnHdt-UMt9c71-OikfDJyjuln60
3- Store features
https://www.kaggle.com/datasets/surajjha101/stores-area-and-sales-data?fbclid=IwAR01APMKaiQBV5MB7TK_t7XAr2-Nww7lTPn4ttBWxZOhnlrY_45VUnZLW4k

Big data provides a vast and diverse landscape of information, allowing for a more comprehensive understanding of complex patterns, trends, and relationships. By working with large datasets, we aimed to gain a deeper understanding of consumer behavior, preferences, and transactional trends in retail sectors.

2.2 Data Preparation:

We employed Python script for a comprehensive ETL (Extract, Transform, Load) process.
The data structure was modified to facilitate easy management and analysis.

2.2.1 Data extraction

To transform the data from CSV to JSON data storage form we used two tools:
First method:
We imported the data into the MongoDB database using the "mongoimport" command to import data into a collection and then export it. We have installed the MongoDB database tools to use the Mongoimport and export commands.
To install the database tools, we visited Database Tools and downloaded the zip file from the platform.
link : https://www.tutorialsteacher.com/mongodb/import-data-using-mongoimport
command line (of the computer):

```
mongoexport --db your_database --collection
your_collection --out your_data.json
```

Second method:
We used a website to convert csv file to json file
link to the website: https://csvjson.com/csv2json

2.2.2 Data Integration

We imported the csv and json files into python using *Jupyter Notebook*. To execute this step, libraries such as *pandas*, *numpy* and *glob* were installed and imported.

2.2.3 Data Transformation

The transformation phase involved renaming columns for consistency and generating random data for customer-related metrics. Subsequently, data from different sources is manipulated, and unnecessary columns are dropped to align with the desired structure.
The actions we took to improve the data sets:

- Renaming columns for consistency (*semantic modeling*)
- Creating a function labeled “generate_random_data” in order to extend the customer dimension from 200 to 15000 to have a more relevant analysis.
- Adding column membership that takes 0 or 1 randomly.
  1 if the customer is enrolled in the store’s fidelity program and 0 if not.
- Extracting only the time (under the format: '%H') from the date column in the transaction data frame.
- Dropping unnecessary columns to align with the desired structure.
- Copying columns and inserting them into other data frames to construct the needed tables (fact table and dimensions).
- Using the merge function to assign each transaction to a customer_id and store_id randomly in order to ensure that each customer may make more than one transaction across different stores.

2.2.4 Data Loading

The script then integrates and prepares data sets for loading into a *MySQL* database. The final stage involves connecting to the database and utilizing the *SQLAlchemy* library to load transformed data into respective tables “customer_dim”, “store_dim”, “transaction_dim”, “time_dim”, and “retail_store_sales”.

The script showcases a systematic approach to data integration, transformation, and loading for effective analysis in a retail store context.

Below is attached the GITHUB link to python code used for ETL process:
https://github.com/Aya123-sys/task1/blob/main/ETL_PROCESS/ETL_process%20(2).ipynb

2.3 Data Storage & Modeling

We used the PHPMyAdmin SQL code, specifying the structure and initial data insertion for tables. This SQL code is designed to be executed in a MySQL environment, creating, and populating the specified table within the given database.
Below is the link for the SQL Script:
https://raw.githubusercontent.com/Aya123-sys/task1/main/datawarehouse/datawarehousedb.sql

Fact and dimensions :

- **Fact Table:** Retail store Sales
  Measures:
  - Sales
  - Quantity
  - Daily customer count
  - Store sales

- **Dimensions:**
  Customer Dimension:
  - Customer name
  - Gender
  - Age
  - Annual Income
  - Spending Score
  - Membership

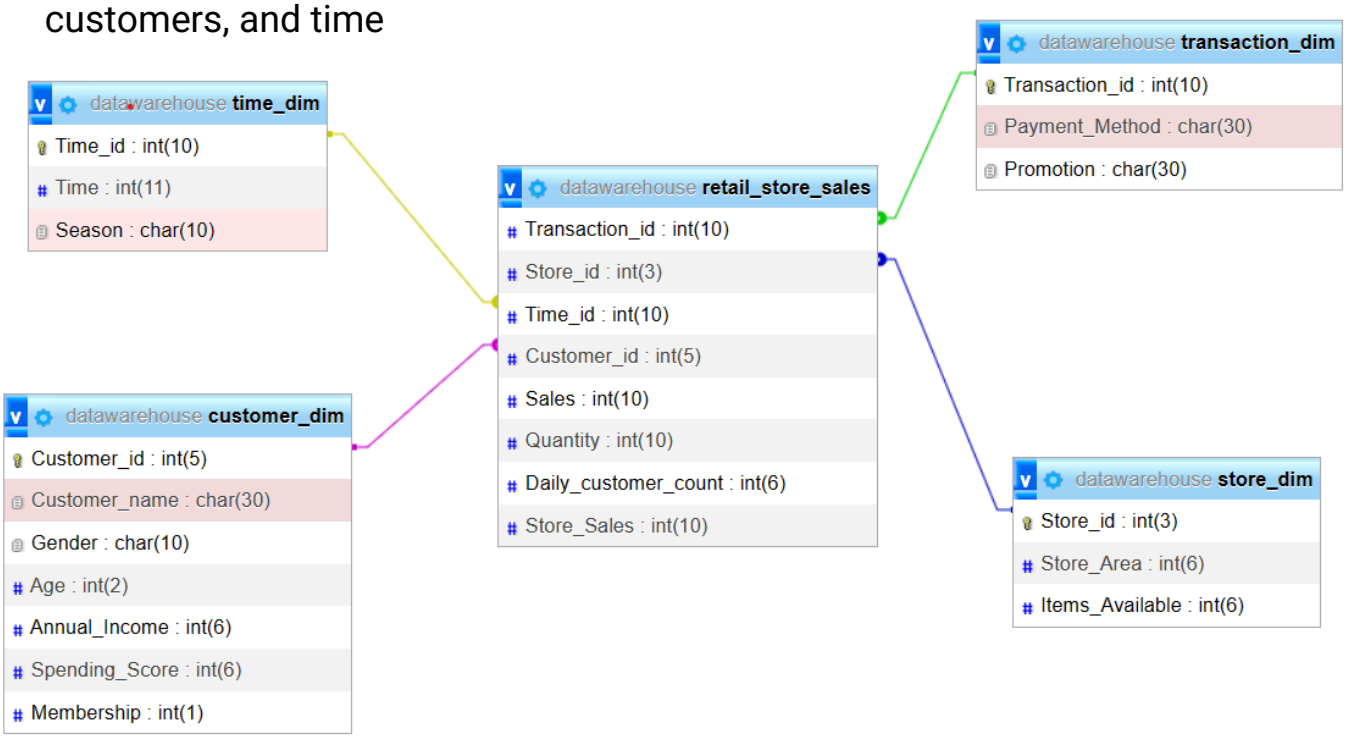
  Transaction Dimension
  - Payment Method:
  - Promotion:

  Store Dimension
  - Store area
  - Items available

  Time Dimension
  - Time
  - Season

Star Schema

The star schema organizes the data into a central fact table "Retail_Store_Sales" containing the measures of interest which are Quantity, Sales, Daily customer Count, and Store sales surrounded by dimension tables describing the attributes of the Stores, Transactions, customers, and time



SQL queries to derive some insights

Characteristics of the customers

```
SELECT
  AVG(C.Age) AS averageAge,
  AVG(C.Annual_Income) AS averageIncome,
  C.Gender,
  C.Membership,
  COUNT(*) AS customerCount
FROM Retail_Store_Sales S,Customer_dim C
WHERE  Membership IS NOT NULL  AND S.Customer_id=C.Customer_id
GROUP BY C.Gender, C.Membership
ORDER BY customerCount DESC;
```

averageAge	averageIncome	Gender	Membership	customerCount ▼ 1
51.7648	146.6647	Female	0	3797
51.9025	150.9339	Male	1	3754
51.1354	149.6653	Female	1	3744
51.7188	146.6213	Male	0	3705

Sales amount according to promotions

```
SELECT Promotion, Total_Sales, Total_Quantity
FROM (
    SELECT T.Promotion,  SUM(S.Sales) AS Total_Sales,
SUM(S.Quantity) AS Total_Quantity
    FROM Retail_Store_sales S, transaction_dim T
    WHERE S.Transaction_id = T.Transaction_id
    GROUP BY T.Promotion
) AS Subquery;
```

Promotion	Total_Sales	Total_Quantity
BOGO (Buy One Get One)	527248	55414
Discount on Selected Items	524313	54516
None	523570	54988

The most used payment method

```
SELECT
  T.Payment_Method,
  COUNT(*) AS transactionCount
FROM
  Transaction T
GROUP BY
  T.Payment_Method
ORDER BY
  transactionCount DESC;
```

Payment_Method	transactionCount ▼ 1
Debit Card	7624
Credit Card	7513
Cash	7480
Mobile Payment	7381

Sales amount according to seasons

```
SELECT Season, Total_sales, Total_Quantity
FROM (
    SELECT Tm.Season,  SUM(S.Sales) AS Total_sales, SUM(S.Quantity)
AS Total_Quantity
    FROM Retail_Store_sales S, time_dim Tm
    WHERE S.Time_id = Tm.Time_id
    GROUP BY Tm.season
) AS Subquery;
```

Season	Total_sales	Total_Quantity
Fall	394899	41205
Spring	396087	41888
Summer	397234	41533
Winter	386911	40292

The “peak” time of the sales

```
SELECT
   Tm.time,
   AVG(S.Sales) AS average_sales
FROM
   Retail_Store_sales S,
   time_dim Tm
WHERE
   S.time_id=Tm.time_id
GROUP BY
   Tm.Time
```

time	average_sales
8	52.8695
9	53.2672
10	51.3426
11	52.0945
12	52.4380
13	53.0462
14	52.2410
15	53.0137
16	52.6677
17	53.1760
18	53.4222
19	52.2719
20	52.6053
21	51.1087
22	52.2526
23	52.4867

The daily customer counts that generate the highest store sales

```
SELECT  Daily_customer_count, Store_Sales,
RANK() OVER (ORDER BY Store_Sales DESC) AS rank
FROM (
    SELECT S.Daily_customer_count,
           S.Store_Sales
    FROM Retail_Store_sales S
) AS Subquery
LIMIT 10
```

Daily_customer_count	Store_Sales	rank
860	116320	1
980	105150	2
680	102920	3
1310	102310	4
820	101820	5
700	101780	6
900	100900	7
680	99570	8
480	99480	9
1100	98260	10

Store areas that generate the largest sales amount

```
SELECT  Store_Area, Store_Sales,
RANK() OVER (ORDER BY Store_Sales DESC) AS rank
FROM (
    SELECT St.Store_Area, S.Store_Sales
    FROM Retail_Store_sales S, store_dim St
    WHERE S.Store_id = St.Store_id
) AS Subquery
LIMIT 10
```

Store_Area	Store_Sales	rank
1989	116320	1
1775	105150	2
1365	102920	3
1303	102310	4
1486	101820	5
1137	101780	6
1565	100900	7
1465	99570	8
1548	99480	9
1800	98260	10

The most bought products:

For this insight, we conducted a market basket analysis using RStudio:
Below the link to the R code
https://github.com/Aya123-sys/task1/blob/main/market_basket_analysis/market_basket_analysis.r
Below the link to the R code output:
https://github.com/Aya123-sys/task1/blob/main/market_basket_analysis/Routpout.txt

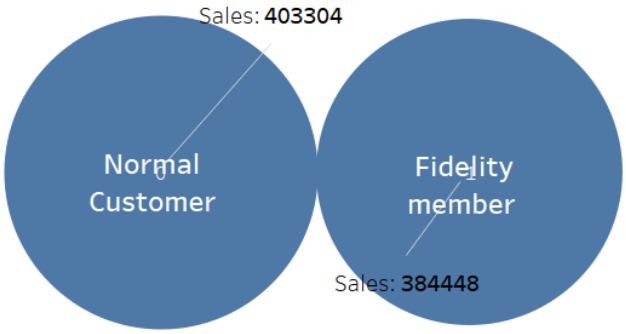
OLAP process

In the OLAP phase of our data warehouse project, we utilized *Schema Workbench* which is a tool used to create and manage *Mondrian* OLAP cube schemas based on the star schema. It allows users to visually design and test these schemas. It is essential for defining a logical model, including cubes, dimensions, hierarchies, levels, and members, and defining the necessary connections to the database. Hence, the use of Schema Workbench and star schema significantly enabled us to make informed business decisions based on insights derived from the retail store data that are well explained in the next phase which is the visualization part.
Link to the schema: https://github.com/Aya123-sys/task1/blob/main/data_visualization/workbench_schema/Schema1.xml

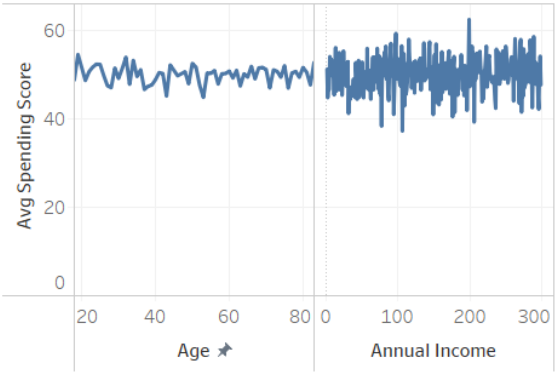
2.4 Data Visualization

In the data visualization phase, we leveraged Tableau to create a comprehensive dashboard for retail store analysis. The dashboard incorporated various visual elements to present key insights from the data stored in the star schema, allowing for intuitive and interactive analysis.
For instance, we used Tableau to develop interactive charts and graphs that visualized sales performance, product trends, and customer behavior.
Furthermore, the dashboard included trend indicators, providing a quick and comprehensive overview of the retail store data. The use of Tableau's drag-and-drop interface and compatibility with multiple sources facilitated the seamless creation of the dashboard, allowing for efficient data visualization and analysis.
Overall, the Tableau dashboard proved to be an invaluable tool for our data warehouse project, enabling stakeholders to gain actionable insights such as:

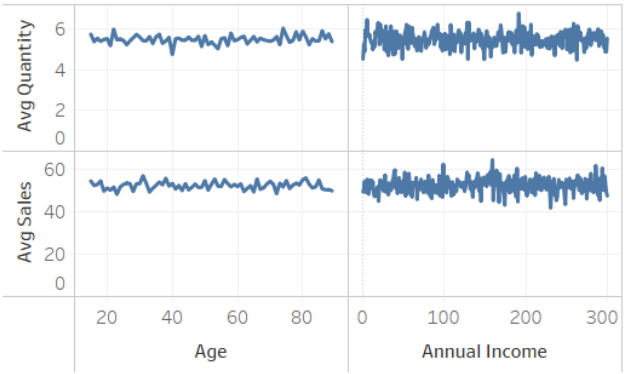
Sales amount with respect to customers membership



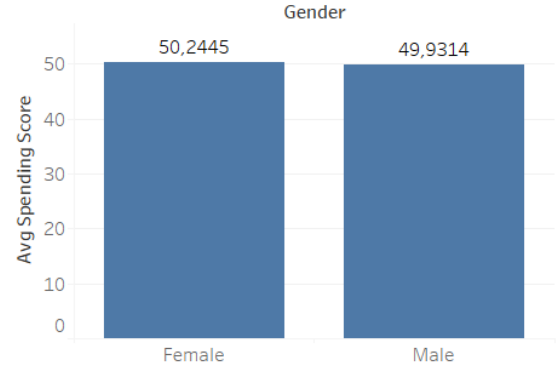
Average spending score of a customer according to the age and annual income



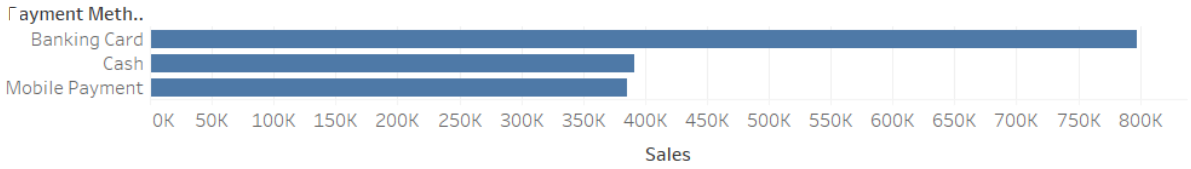
Average sales according to the age and annual income of the customers



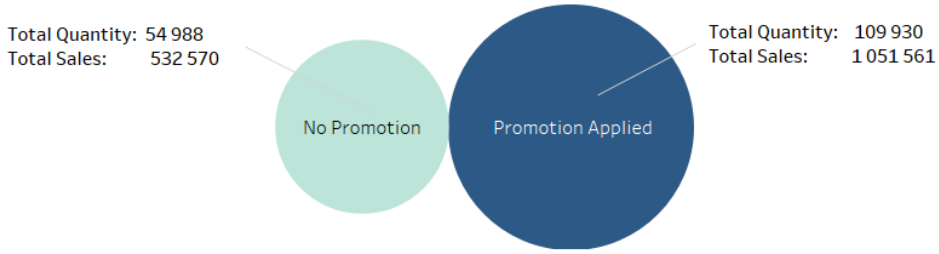
Average spending score of a customer with respect to the gender



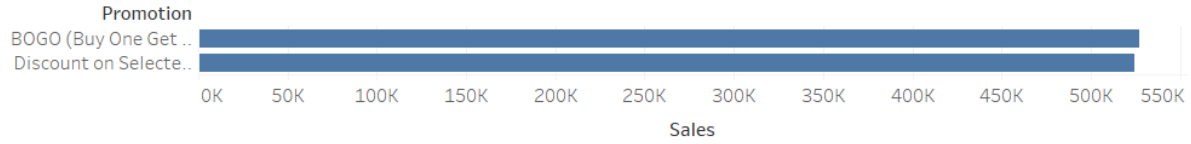
Amount of sales with respect to different payment methods used by customers



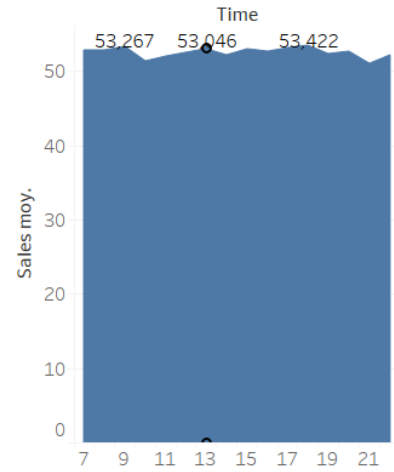
Comparison between sales performance with and without promotional activities

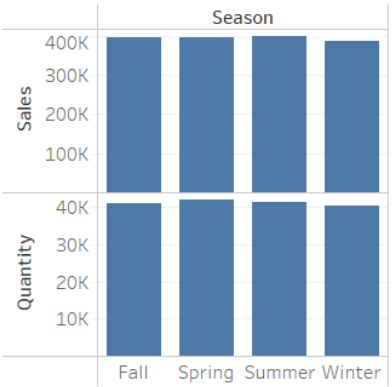


Comparison of sales performance according to different types of promotion

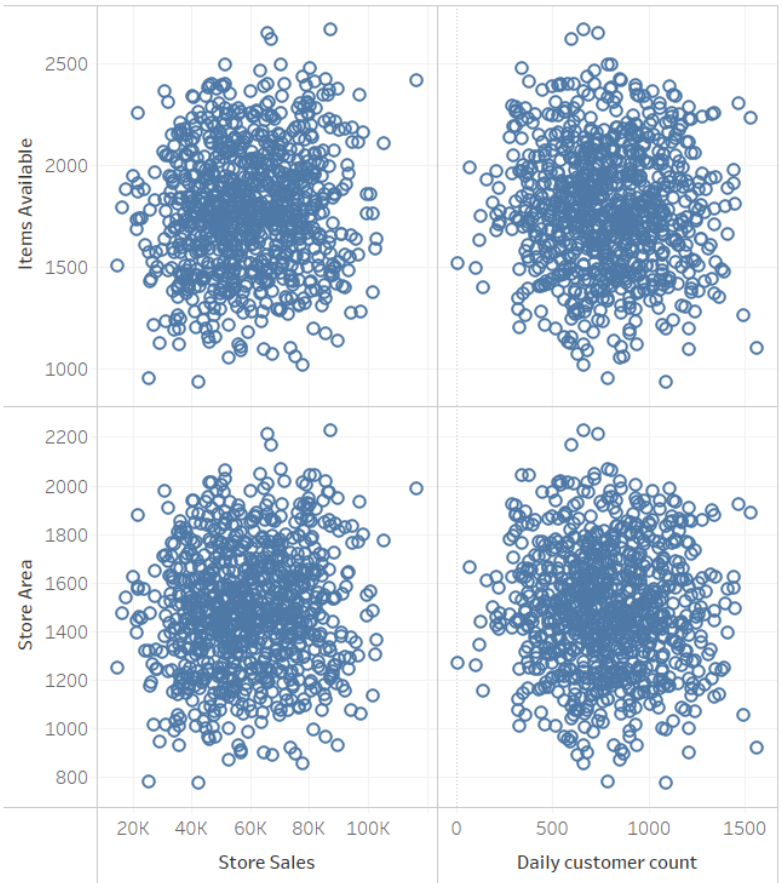


Sales Peak Time





Store Features Analysis



3. Conclusion

Based on the data analysis of the retail stores, we can conclude that:
- The customer spending score increases with the annual income, and - The average spending score for females is slightly greater than that of males.
- The store sales are divided almost equally between customers that are enrolled in fidelity programs and those who are not.
- Teenagers and young adults are highly valuable customers as they have the highest spending scores.
- Regardless of the annual income, customers have high spending scores (between 40 and 60) which leads to high sales amounts.
- The most used payment method is the banking card.
- The sales increase with promotional activities, particularly with the “Buy One Get One” (BOGO) promotion, which is the most preferred promotion type among customers.
- There is no significant difference in the store performance among seasons.
- The sales peak times are at 9 am, 1 pm, and 7 pm.
- The optimal area that generates the greatest sales performance lies between 1300 and 1800 yard square.
- Most of the customers prefer shopping from stores where there is a reasonable number of items (between 1500 and 2000)
- The average daily customer count is around 750 individuals.

=>These findings suggest that:
- The retail store should focus on optimizing promotional activities, enhancing payment processing for banking cards, and maximizing sales during peak hours.
- Retail stores should also consider offering incentives to customers who use banking cards for purchases and targeting promotions toward female customers.
- Furthermore, the store can consider segmenting customers based on their annual income and tailoring marketing efforts to each segment.
- Adequate staffing during peak hours is also crucial to cater to the high number of customers.
These insights can inform strategies to improve the performance of the retail store industry, ultimately contributing to increased revenue and customer satisfaction.