CUSTOMER SHOPPING BEHAVIOR ANALYSIS

1. Project Overview

This project focuses on performing an end-to-end analysis of customer shopping behaviour using transactional retail data. The objective of the project is to understand how customers interact with products, discounts, subscriptions, and purchasing patterns, and to translate these behaviours into actionable business insights.

The analysis begins with raw customer shopping data, which is cleaned and prepared using Python and Pandas. The processed data is then stored in a relational database and analysed using SQL to answer key business questions related to customer segments, loyalty, revenue drivers, and purchasing trends. Finally, the insights are visualised using an interactive Power BI dashboard to support data-driven decision-making.

This project simulates a real-world data analyst workflow, combining data preparation, exploratory analysis, structured querying, and business-focused visualisation to help stakeholders better understand customer behaviour and optimise marketing and product strategies.

2. Dataset Description

The dataset used in this project represents customer shopping behaviour collected from retail transactions. It contains detailed information about customer demographics, purchase characteristics, and shopping patterns, making it suitable for behavioural and business analysis.

- **Number of Records:** Approximately 3,900 customer transactions
- **Number of Features:** 18 columns

Key Data Categories

The dataset includes the following types of information:

1. Customer Demographics

- Age
- Gender
- Location
- Subscription Status

2. Purchase Details

- Item Purchased
- Product Category
- Purchase Amount
- Season

3. Shopping Behaviour Indicators

- Discount Applied
- Review Rating
- Previous Purchases
- Purchase Frequency
- Shipping Type

Data Quality Overview

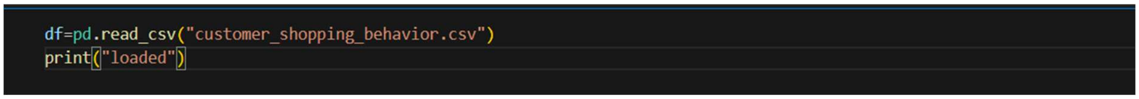
- A small number of missing values were observed in the review rating column
- No duplicate transaction records were found
- Data types require standardisation before analysis

The dataset provides a comprehensive view of how customers interact with products and promotions, enabling analysis of spending behaviour, loyalty patterns, and revenue drivers.

3. Exploratory Data Analysis (Python & Pandas

We began with data preparation and cleaning in Python:

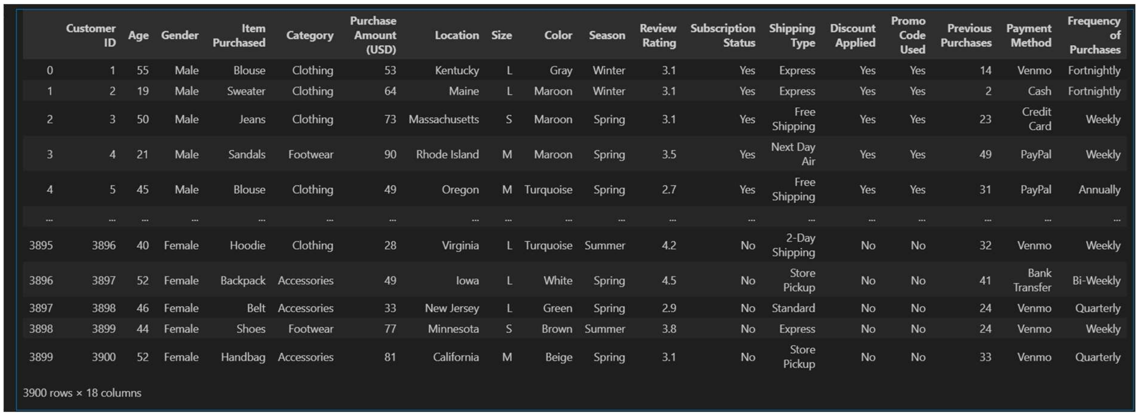
● **Data Loading:** Imported the dataset using pandas.



```
df=pd.read_csv("customer_shopping_behavior.csv")
print("loaded")
```

● **Initial Exploration:** Used `df.info()` to examine the data structure and `df.describe()` to review summary statistics.

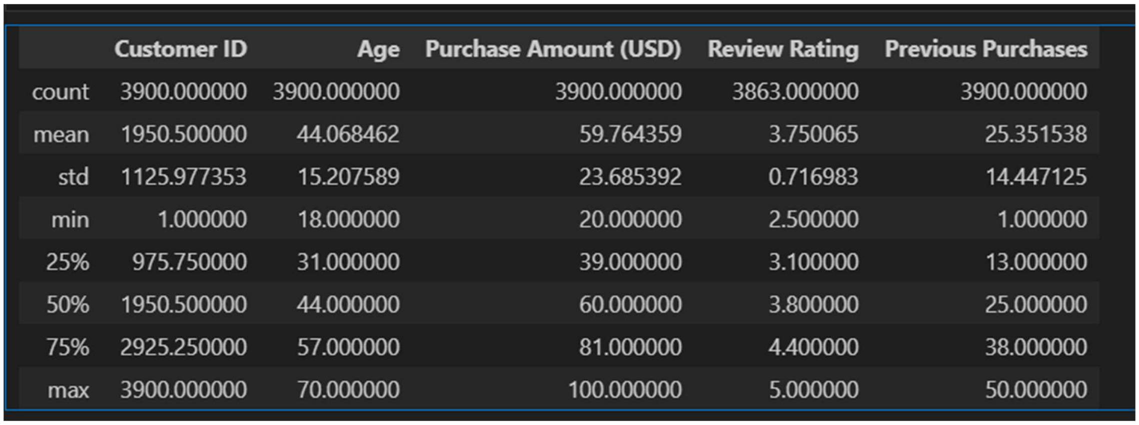
➢ df



	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	2-Day Shipping	No	No	32	Venmo	Weekly
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	Store Pickup	No	No	41	Bank Transfer	Bi-Weekly
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	Standard	No	No	24	Venmo	Quarterly
3898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No	Express	No	No	24	Venmo	Weekly
3899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No	Store Pickup	No	No	33	Venmo	Quarterly

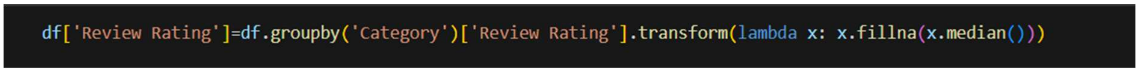
3900 rows × 18 columns

- **df.describe( )**



	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3863.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.750065	25.351538
std	1125.977353	15.207589	23.685392	0.716983	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.800000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

● **Missing Data Handling:** Checked for null values and imputed missing values in the Review Rating column using the median rating of each product category.



```
df['Review Rating']=df.groupby('Category')['Review Rating'].transform(lambda x: x.fillna(x.median()))
```

● **Column Standardization:** Renamed columns to snake case to improve readability and maintain consistency.

● **Feature Engineering:**

○ Created an **age_group** column by binning customer ages.

○ Created a **purchase_frequency_days** column by converting purchase frequency into day-based values.

● **Data Consistency Check:** Verified whether **discount_applied** and **promo_code_used** were redundant and removed **promo_code_used**.

● **Database Integration:** Connected the Python script to a **MySQL** database and loaded the cleaned DataFrame for SQL-based analysis.

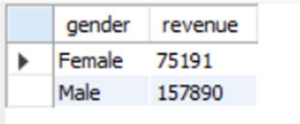
4. Data Analysis Using MySQL

After cleaning and preparing the dataset in Python, the data was loaded into a MySQL database to perform structured, query-based analysis and simulate real-world business reporting.

The SQL analysis focused on answering key business questions related to customer behavior, revenue drivers, and purchasing patterns.

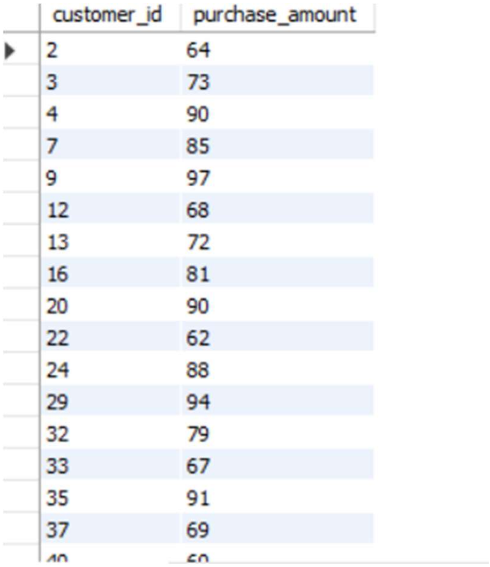
Key SQL Analyses Performed:

1. **Revenue by Gender:**
Calculated the total revenue generated by each gender to understand demographic contribution to sales.



gender	revenue
Female	75191
Male	157890

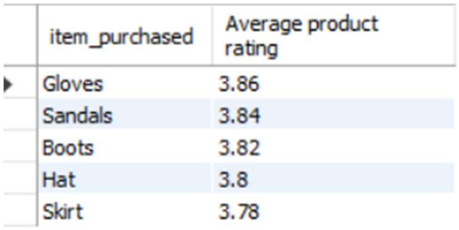
2. **High-Spending Customers Using Discounts:**
Identified customers who applied discounts but still spent more than the average purchase amount, highlighting value-driven customers.



customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90
22	62
24	88
29	94
32	79
33	67
35	91
37	69
40	60

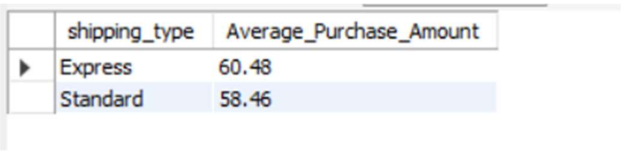
Total row : 839

3. **Top-Rated Products:**
Retrieved products with the highest average review ratings to identify well-performing items.



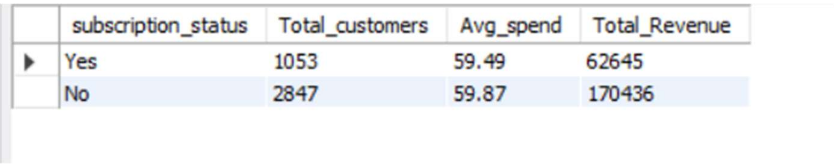
item_purchased	Average product rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Skirt	3.78

4. **Shipping Type Comparison:**
Compared average purchase amounts between standard and express shipping to evaluate the impact of shipping preferences on spending.



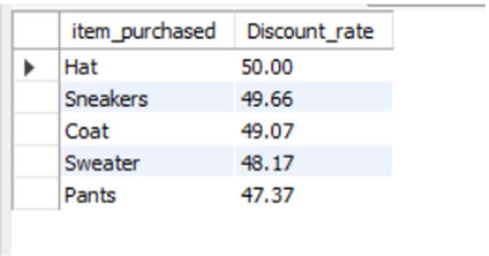
shipping_type	Average_Purchase_Amount
Express	60.48
Standard	58.46

5. **Subscription-Based Analysis:**
Analyzed differences in spending behavior between subscribers and non-subscribers, including total revenue and average purchase value.



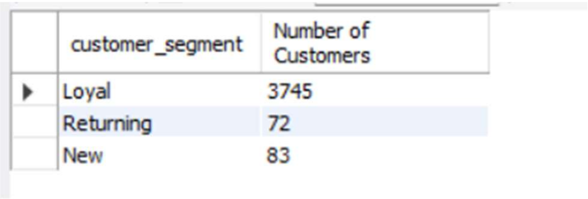
subscription_status	Total_customers	Avg_spend	Total_Revenue
Yes	1053	59.49	62645
No	2847	59.87	170436

6. **Discount-Dependent Products:**
Identified products with a high percentage of discounted purchases to assess reliance on promotional strategies.



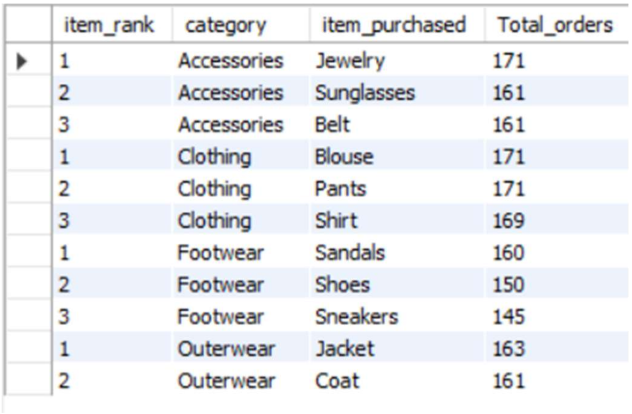
item_purchased	Discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

7. **Customer Segmentation:**
Classified customers into **New, Returning, and Loyal** segments based on purchase history.



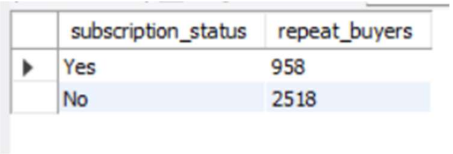
customer_segment	Number of Customers
Loyal	3745
Returning	72
New	83

8. Top Products per Category:
Ranked the most frequently purchased products within each category.



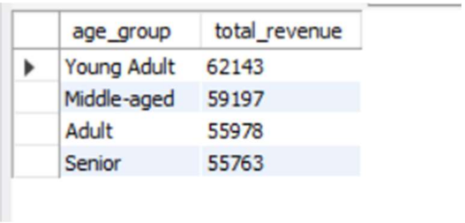
item_rank	category	item_purchased	Total_orders
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
1	Clothing	Blouse	171
2	Clothing	Pants	171
3	Clothing	Shirt	169
1	Footwear	Sandals	160
2	Footwear	Shoes	150
3	Footwear	Sneakers	145
1	Outerwear	Jacket	163
2	Outerwear	Coat	161

9. Repeat Purchase and Subscription Relationship:
Evaluated whether customers with multiple purchases were more likely to have an active subscription.



subscription_status	repeat_buyers
Yes	958
No	2518

10. Revenue by Age Group:
Calculated revenue contribution across different age groups to identify high-value demographic segments.



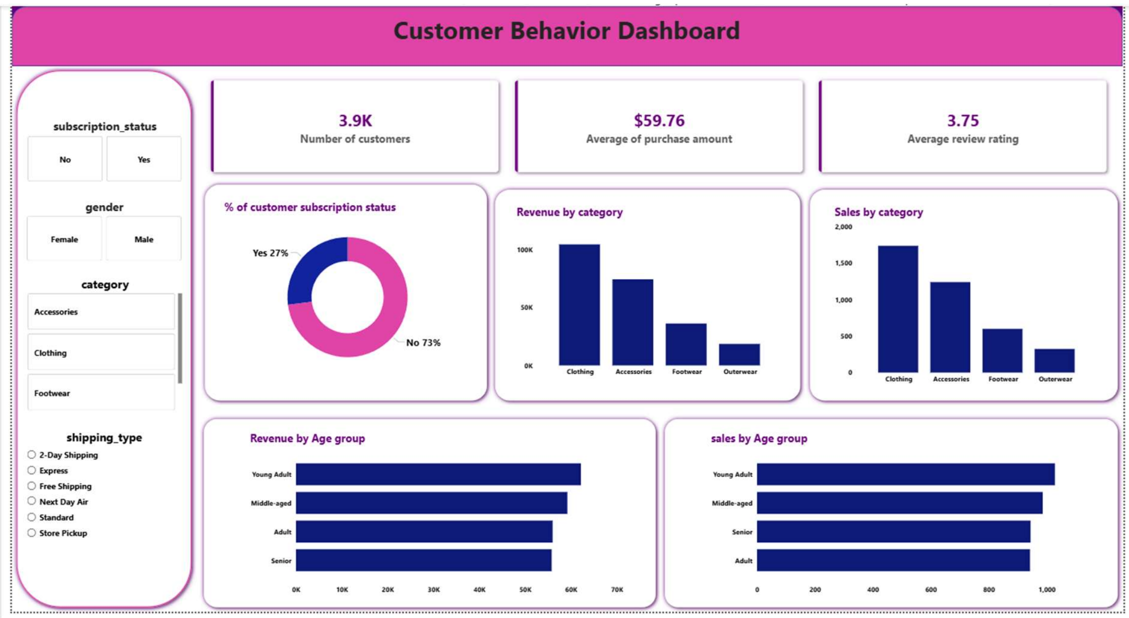
age_group	total_revenue
Young Adult	62143
Middle-aged	59197
Adult	55978
Senior	55763

`The SQL-based analysis provided:

- Clear visibility into customer segments and loyalty patterns
- Identification of high-performing products and categories
- Insights into the impact of discounts, shipping, and subscriptions on revenue.

5. Data Visualization & Dashboard (Power BI)

An interactive **Power BI dashboard** was created to present insights from the analysis in a clear and business-friendly format.



- Displayed key KPIs such as total revenue, total customers, and average purchase value.
- Visualized category-wise sales and revenue performance.
- Analyzed customer behavior across age groups and gender.
- Compared subscribers and non-subscribers to understand spending patterns.
- Evaluated the impact of discounts and shipping types on purchases.
- Used interactive filters to enable dynamic exploration of data.

6. Business Recommendations

- **Boost Subscriptions** – Offer exclusive benefits to increase customer retention and repeat purchases.
- **Customer Loyalty Programs** – Reward frequent buyers to move them into the loyal customer segment.
- **Review Discount Strategy** – Optimize discount usage to increase sales while protecting profit margins.
- **Product Positioning** – Promote top-rated and best-selling products in marketing campaigns.
- **Targeted Marketing** – Focus efforts on high-revenue age groups and customers preferring express shipping.