Customer Shopping Behavior Analysis

1. Project Overview

The following project analyzes customer shopping behavior through transactional data from 3900 purchases of different categories of products, aiming to find insights into spending patterns, customer segments, product preferences, and subscription behavior that can influence strategic business decisions.

2. Dataset summary

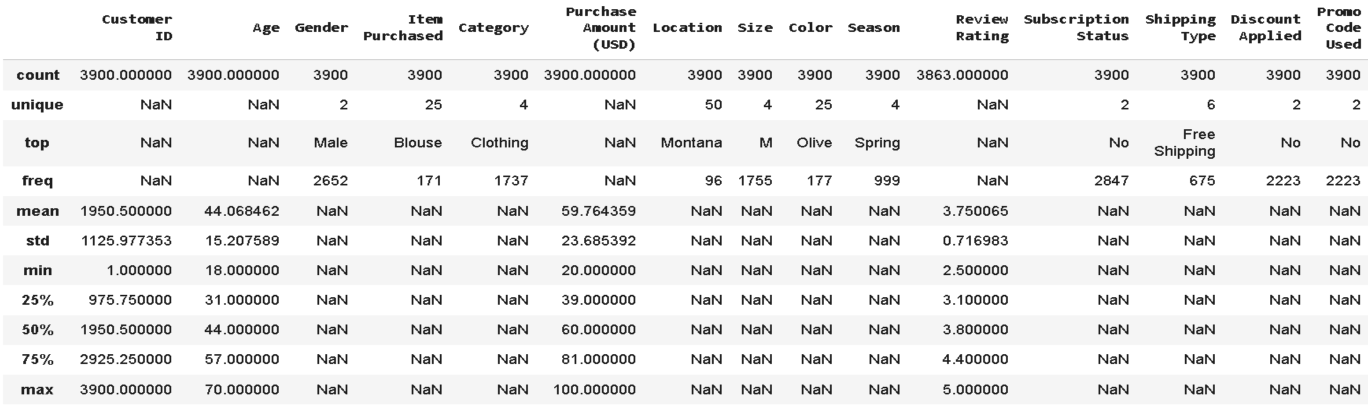
- **Rows**: 3900
- **Columns**: 18
- **Key Features**:
- **Customer demographics include:** age, gender, location, and subscription status.
- **Purchase details:** Items Purchased, Category, Purchase amount, Season, Size, Color)
- **Shopping behavior:** Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)
- **Missing Data:** 37 values in Review rating column

3. Exploratory Data Analysis using Python

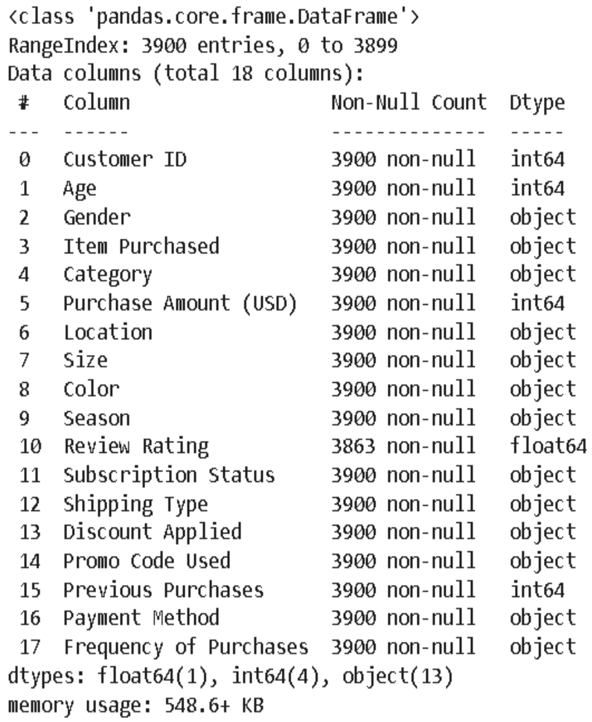
I began with data preparation and cleaning in Python:

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

Initial Exploration: Used describe( ) to check structure and df.info( ) for summary statistics.



	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
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Customer ID             3900 non-null   int64  
 1   Age                     3900 non-null   int64  
 2   Gender                  3900 non-null   object 
 3   Item Purchased          3900 non-null   object 
 4   Category                3900 non-null   object 
 5   Purchase Amount (USD)   3900 non-null   int64  
 6   Location                3900 non-null   object 
 7   Size                    3900 non-null   object 
 8   Color                   3900 non-null   object 
 9   Season                  3900 non-null   object 
 10  Review Rating           3863 non-null   float64
 11  Subscription Status     3900 non-null   object 
 12  Shipping Type           3900 non-null   object 
 13  Discount Applied        3900 non-null   object 
 14  Promo Code Used         3900 non-null   object 
 15  Previous Purchases      3900 non-null   int64  
 16  Payment Method          3900 non-null   object 
 17  Frequency of Purchases  3900 non-null   object 
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

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

- **Column standardization:** Renamed columns to **snake case** for better readability and documentation.

- **Feature Engineering:**
  - Created **age_group** column by binning customer ages.
  - Created **purchase_frequency_days** column from purchase data.

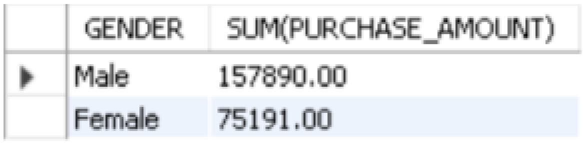
- **Data consistency:** Verified id discount_applied and promo_code_used were reductant; dropped promo_code_used.

- **Database Integration:** Connected Python script to MySQL and loaded cleaned DataFrame into the database for analysis.

4. Data Analysis using SQL:

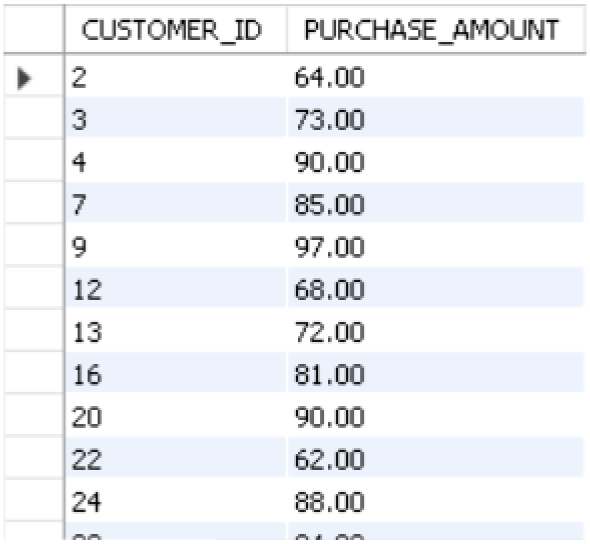
I performed structured analysis in MySQL to answer key business questions:

1. **Revenue by Gender:** Compared total revenue generated by male vs female customers.



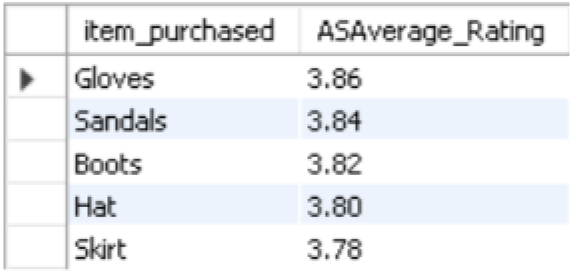
GENDER	SUM(PURCHASE_AMOUNT)
Male	157890.00
Female	75191.00

2. **High- spending Discount Users:** Identified customers who used discounts but still spent above the average purchase amount.



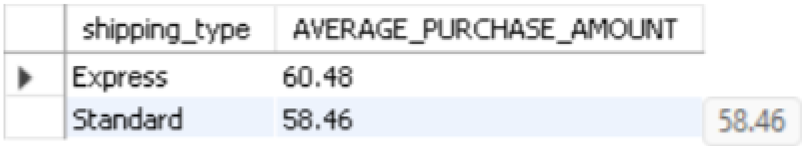
CUSTOMER_ID	PURCHASE_AMOUNT
2	64.00
3	73.00
4	90.00
7	85.00
9	97.00
12	68.00
13	72.00
16	81.00
20	90.00
22	62.00
24	88.00

3. Top 5 Products by Rating: Found products with the highest average review ratings



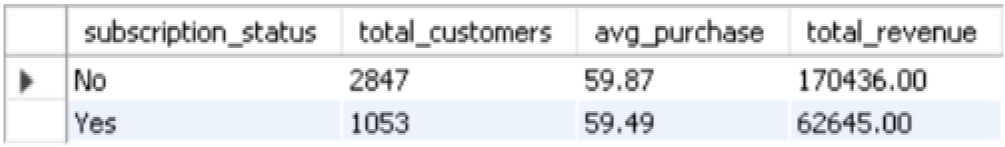
item_purchased	ASAverage_Rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.80
Skirt	3.78

4. Shipping Type Comparison: Compared the average purchase amounts between Standard and Express shipping customers.



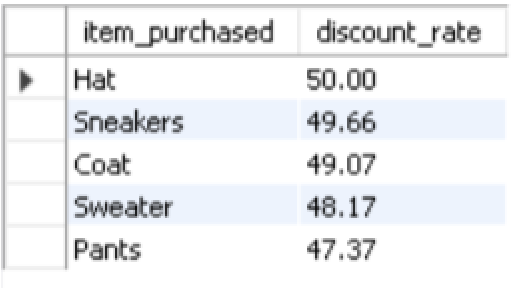
shipping_type	AVERAGE_PURCHASE_AMOUNT
Express	60.48
Standard	58.46

5. Subscriber Spending Analysis: Compared average spending and total revenue to determine whether subscribed customers spend more.



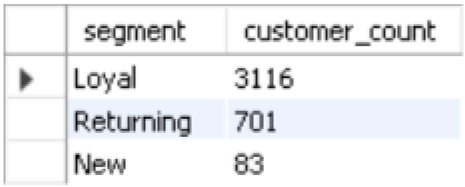
subscription_status	total_customers	avg_purchase	total_revenue
No	2847	59.87	170436.00
Yes	1053	59.49	62645.00

6. Products with the Highest Discount Usage: Identifying the top five products for which the discounts are used most.



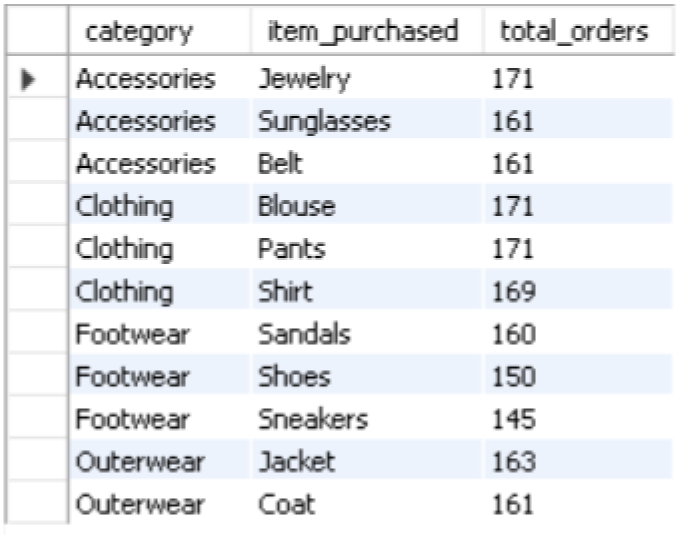
item_purchased	discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

7. Customer Segmentation: Segmented customers into New, Returning, and Loyal according to their previous purchases, counting how many are in each category.



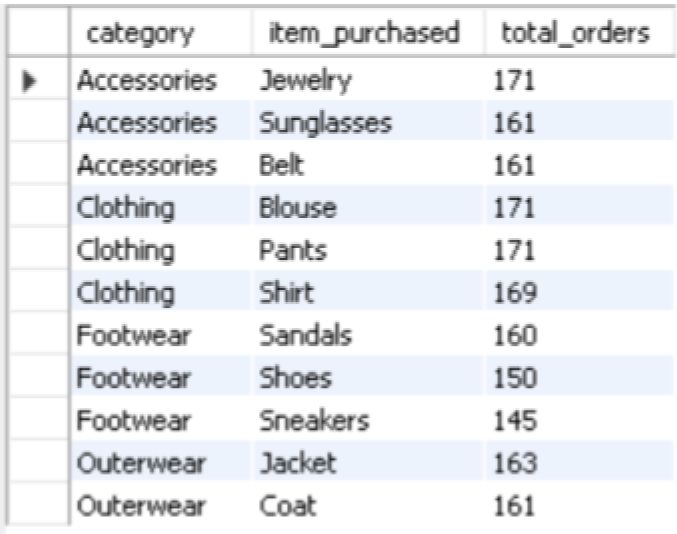
segment	customer_count
Loyal	3116
Returning	701
New	83

8. Top Products in Each Category: Determine the top three most sold products of each product category.



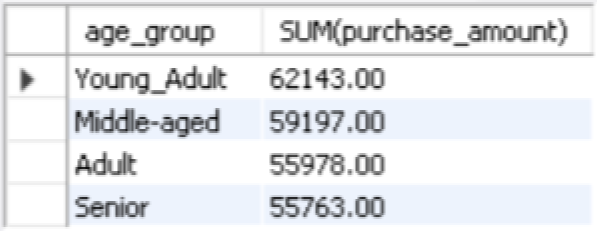
category	item_purchased	total_orders
Accessories	Jewelry	171
Accessories	Sunglasses	161
Accessories	Belt	161
Clothing	Blouse	171
Clothing	Pants	171
Clothing	Shirt	169
Footwear	Sandals	160
Footwear	Shoes	150
Footwear	Sneakers	145
Outerwear	Jacket	163
Outerwear	Coat	161

9. Subscription and Repeat Buyers: Analyzed whether customers who have already purchased more than five times are more likely to subscribe.



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

10. Revenue by Age Group: Calculated the total revenue contributed by each age group.

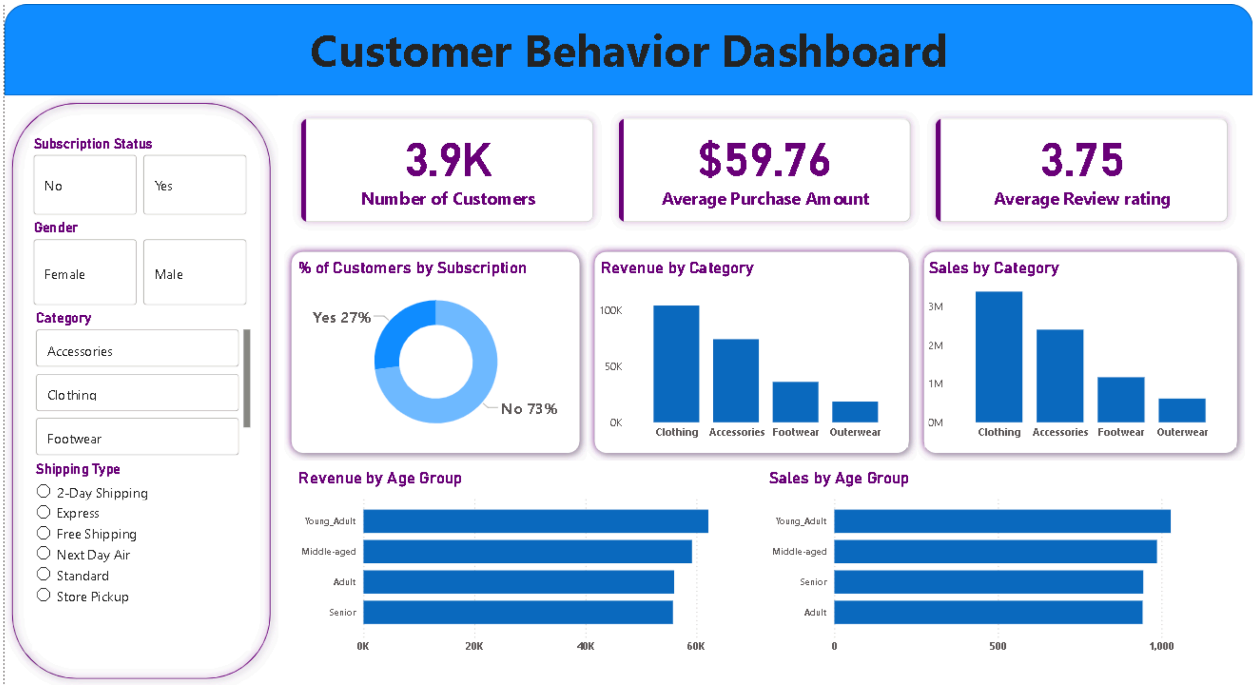


age_group	SUM(purchase_amount)
Young_Adult	62143.00
Middle-aged	59197.00
Adult	55978.00
Senior	55763.00

5. Data Visualization using Power BI:

- **Built an interactive Power BI dashboard to clearly communicate customer behavior patterns and business insights.**

- **Connected and transformed the dataset using Power Query to ensure clean and analysis-ready data.**



6. Business Recommendations:

- **Boost Subscription:** Increase subscription sign-ups by offering exclusive discounts and benefits.
- **Grow Accessories Revenue:** Use bundle offers to encourage add-on purchases.
- **Improve Ratings:** Focus on faster delivery and better product quality.
- **Drive Repeat Purchases:** Introduce loyalty rewards for frequent shoppers.
- **Optimize Inventory:** Stock more of high-demand items to avoid missed sales.