Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior using transactional data form 3,900 purchases across various product categories. The goal is to uncover insights into spending patterns, customer segments, product performance, and subscription behavior to guide strategic business decisions.

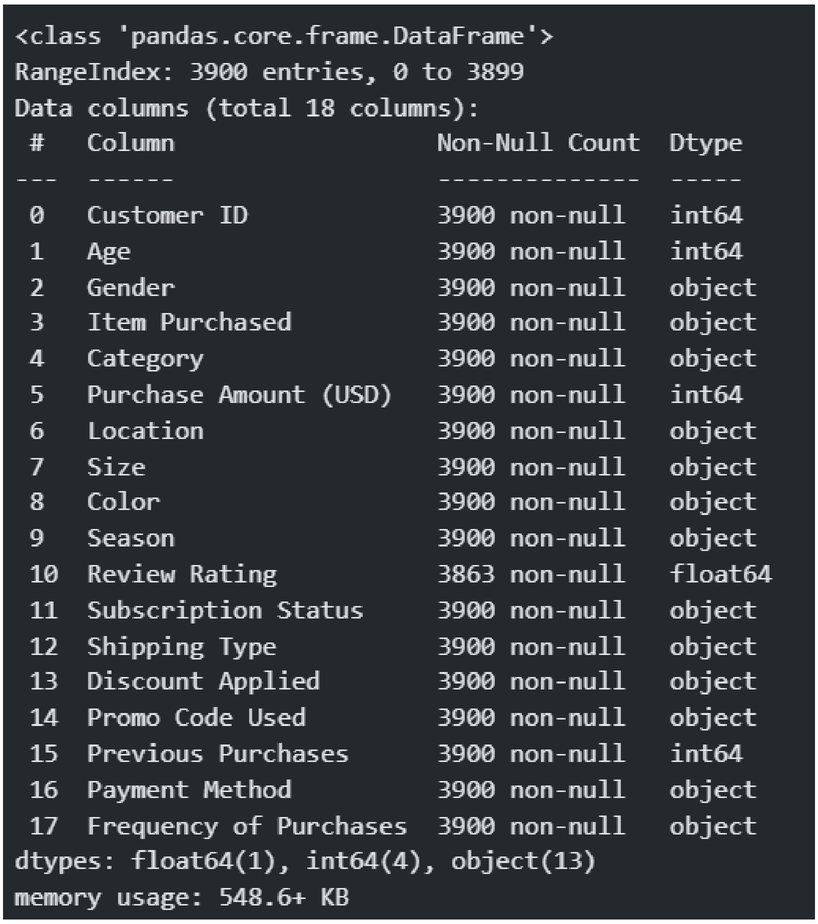
2. Dataset Summary

- Rows: 3,900
- Columns: 18
- Key Features:
  - Customer demographics (Age, Gender, Location, Subscription Status)
  - Purchase details (Item Purchased, Category, Purchase Amount, Season, Size, Color)
  - Shopping behavior (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rate, Shipping Type, Payment Method)
- Missing Data 37 values in Review Rating column

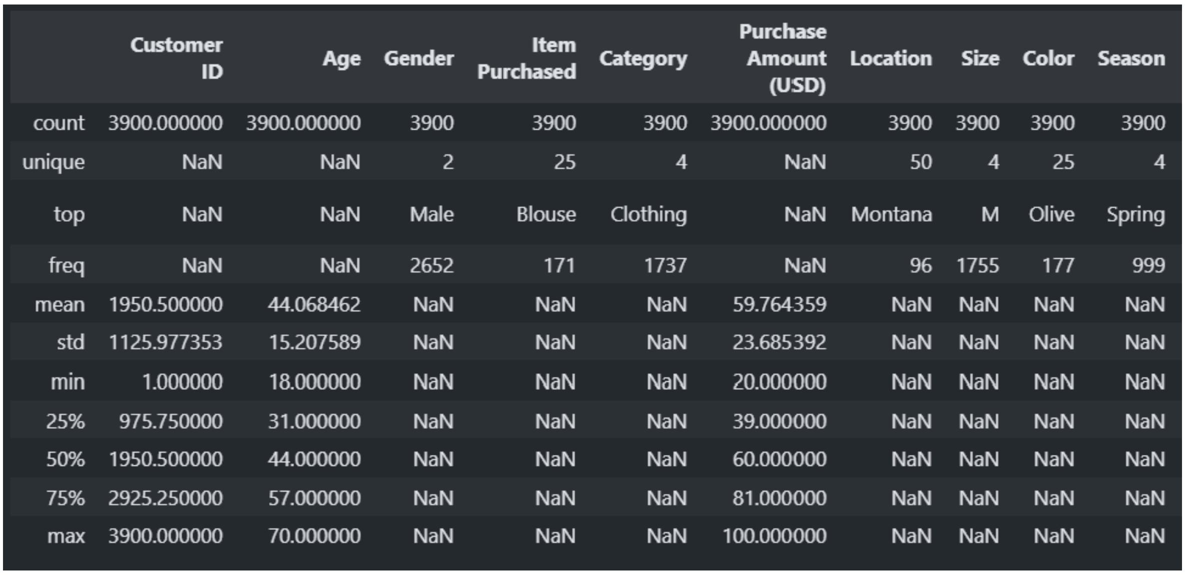
3. Exploratory Data Analysis Using Python

We began with data preparation and cleaning in Python:

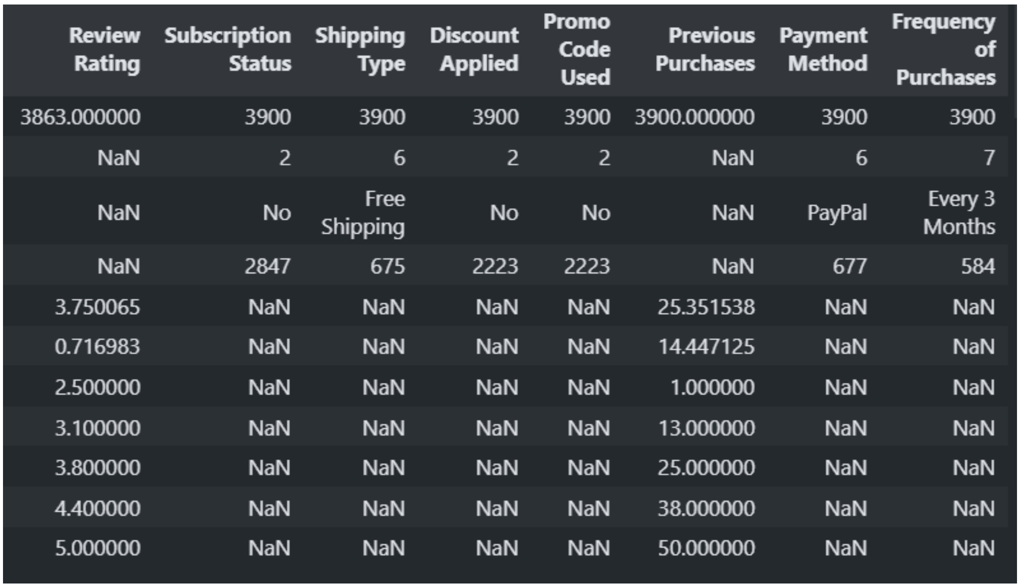
- **Data Loading:** Import the dataset using pandas.
- **Initial Exploration:** Using df.info () to check structure and .describe () for summary statistics.



```
<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
```



	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN



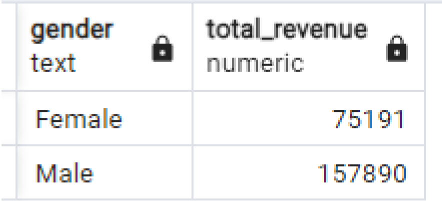
Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3863.000000	3900	3900	3900	3900	3900.000000	3900	3900
NaN	2	6	2	2	NaN	6	7
NaN	No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
NaN	2847	675	2223	2223	NaN	677	584
3.750065	NaN	NaN	NaN	NaN	25.351538	NaN	NaN
0.716983	NaN	NaN	NaN	NaN	14.447125	NaN	NaN
2.500000	NaN	NaN	NaN	NaN	1.000000	NaN	NaN
3.100000	NaN	NaN	NaN	NaN	13.000000	NaN	NaN
3.800000	NaN	NaN	NaN	NaN	25.000000	NaN	NaN
4.400000	NaN	NaN	NaN	NaN	38.000000	NaN	NaN
5.000000	NaN	NaN	NaN	NaN	50.000000	NaN	NaN

- **Missing Data Handling:** Checked for null values and imputed missing value in the Review Rating column using the median rating of each product category.
- **Column Standardization:** Renamed column to **snake case** for better readability and documents.
- **Feature Engineering:**
  - Create **age_group** column by binning customer ages.
  - Create **purchase_frequency_days** columns from purchase data.
- **Data Consistency Check:** Verified if discount_applied and promo_code_used were redundant; drop promo_code_used.
- **Data Integration:** Connected Python script to PostgreSQL and loaded the cleaned DataFrame into database for SQL analysis.

4. Data Analysis using SQL (Business Transactions)

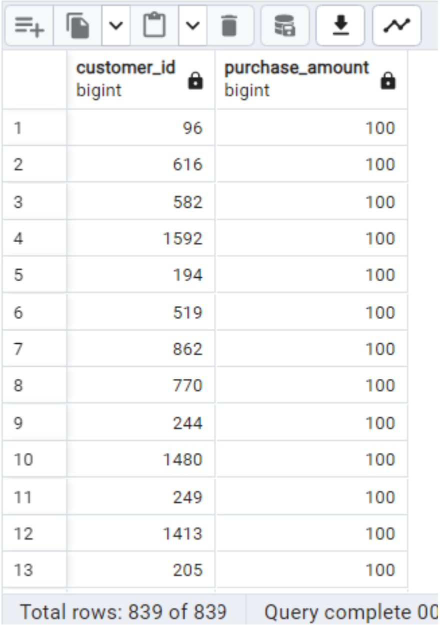
We performed structured analysis in PostgreSQL to answer key business questions.

1. **Revenue by Gender –** Compare total revenue generated by male vs. female customers.



gender text	total_revenue numeric
Female	75191
Male	157890

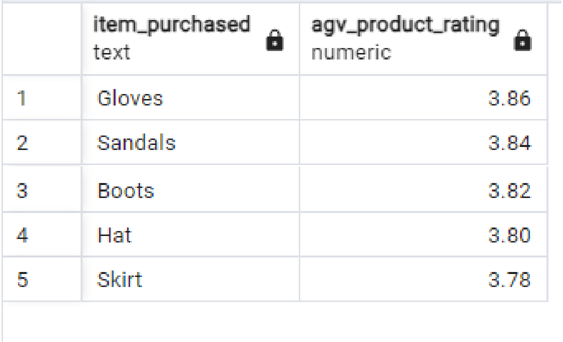
2. **High-Spending Discount User -** Identified customers who used discount but still spent above the average purchase amount.



	customer_id bigint	purchase_amount bigint
1	96	100
2	616	100
3	582	100
4	1592	100
5	194	100
6	519	100
7	862	100
8	770	100
9	244	100
10	1480	100
11	249	100
12	1413	100
13	205	100

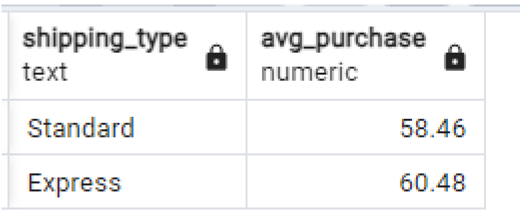
Total rows: 839 of 839 Query complete 00

3. **Top 5 Products by Rating –** Found product with the highest average review ratings.



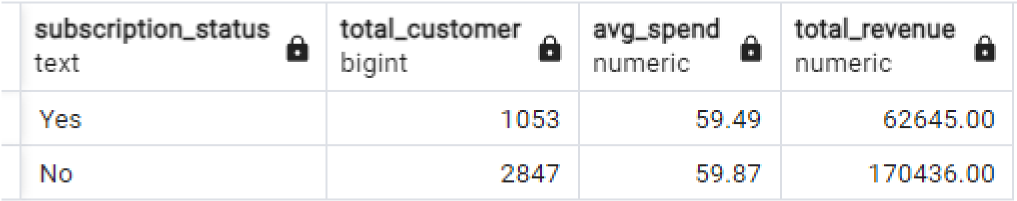
	item_purchased text	agv_product_rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

4. **Shipping Type Comparison** – Compared average purchase amount between Standard and Express shipping.



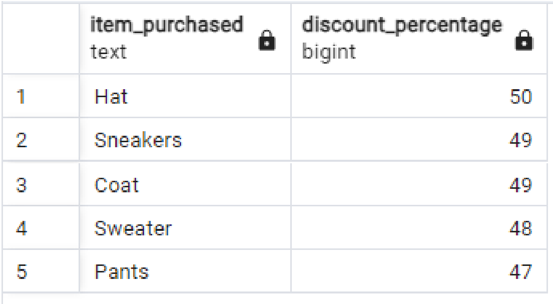
shipping_type text	avg_purchase numeric
Standard	58.46
Express	60.48

5. **Subscriber Vs. Non-Subscribers –** Compare average spend and total revenue across subscription status.



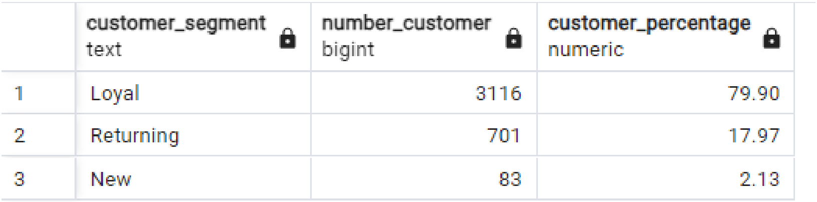
subscription_status text	total_customer bigint	avg_spend numeric	total_revenue numeric
Yes	1053	59.49	62645.00
No	2847	59.87	170436.00

6. **Discount-Dependent Products –** Identified 5 products with the highest percentage of discounted purchases.



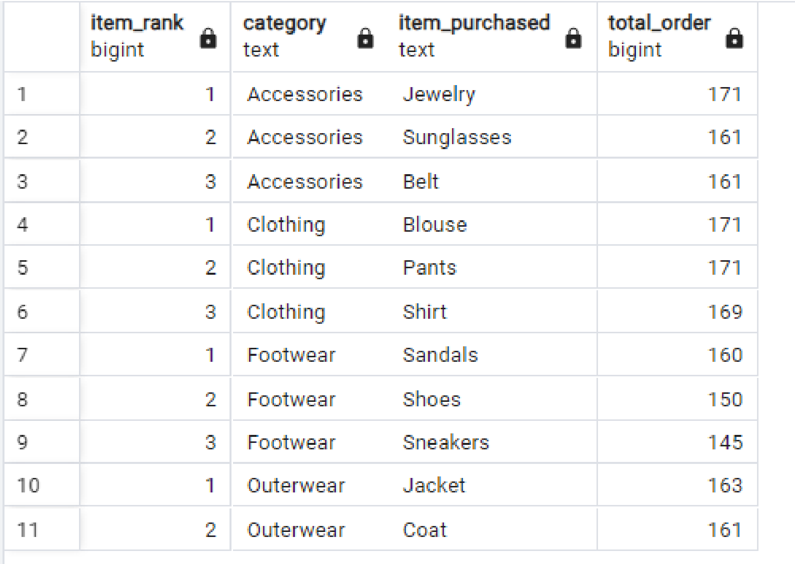
	item_purchased text	discount_percentage bigint
1	Hat	50
2	Sneakers	49
3	Coat	49
4	Sweater	48
5	Pants	47

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



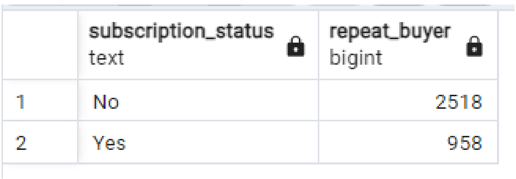
	customer_segment text	number_customer bigint	customer_percentage numeric
1	Loyal	3116	79.90
2	Returning	701	17.97
3	New	83	2.13

8. **Top 3 Products per Category -** Listed the most purchased products within each category.



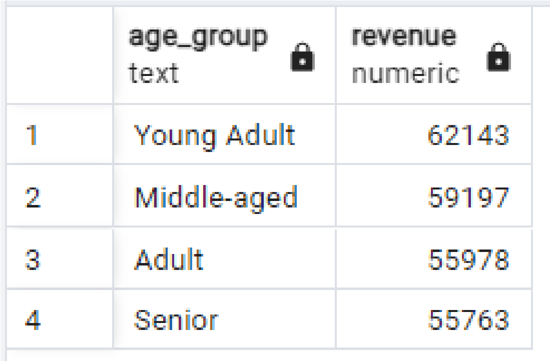
	item_rank bigint	category text	item_purchased text	total_order bigint
1	1	Accessories	Jewelry	171
2	2	Accessories	Sunglasses	161
3	3	Accessories	Belt	161
4	1	Clothing	Blouse	171
5	2	Clothing	Pants	171
6	3	Clothing	Shirt	169
7	1	Footwear	Sandals	160
8	2	Footwear	Shoes	150
9	3	Footwear	Sneakers	145
10	1	Outerwear	Jacket	163
11	2	Outerwear	Coat	161

9. **Repeat Buyer & Subscription –** Checked whether customer with 5 > purchase age more likely to subscribe.



	subscription_status text	repeat_buyer bigint
1	No	2518
2	Yes	958

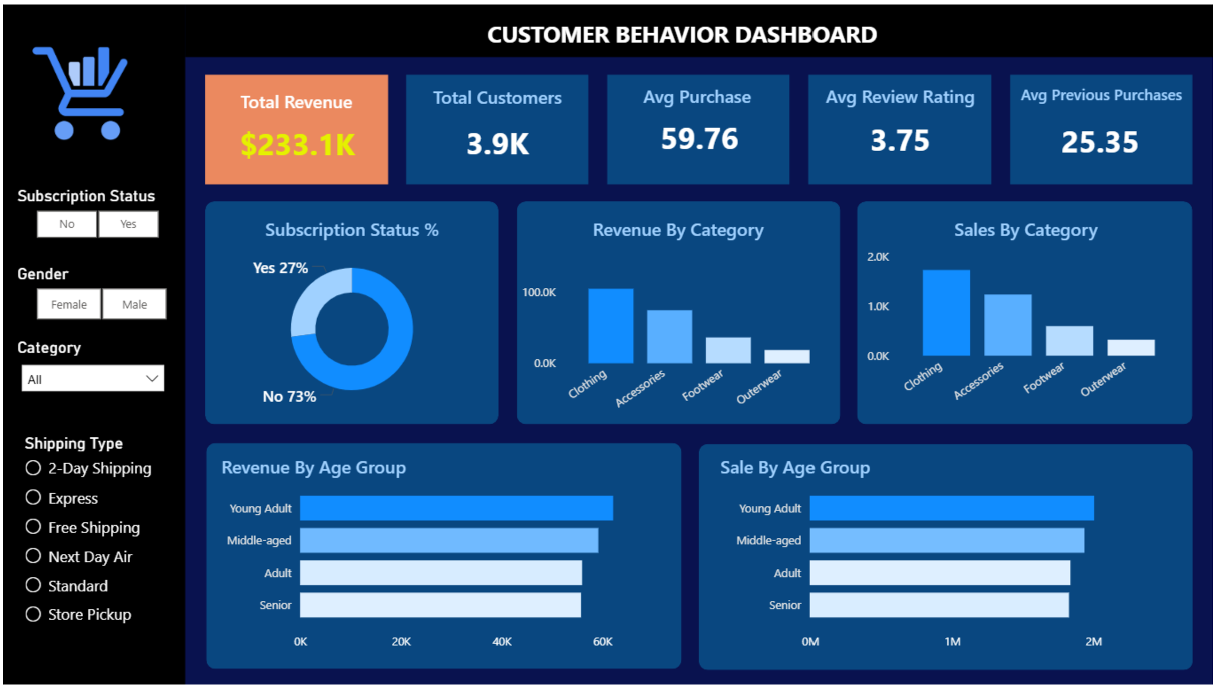
10. **Revenue by Age Group -** Calculate total revenue contribution of each group.



	age_group text	revenue numeric
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

5. Dashboard in Power BI

Finally, we built an interactive dashboard in **Power BI** to present insights visually.



6. Business Recommendations

- **Boost Subscriptions -** Promote exclusive benefits for subscribers.
- **Customer Loyalty Program -** Reward repeat buyers to move them the “Loyal” segment.
- **Review Discount Policy -** Balance sale boosts with margin control.
- **Product Positioning –** Highlight top-rated and best-selling products in campaigns.
- **Targeted Marketing –** Focus efforts on high-revenue age group and express-shipping users.