Customer Shopping Behavior Analysis

I. Project Overview

This project focuses on analyzing customer shopping behavior using transactional data.
The dataset contains information about customer demographics, purchasing patterns, discounts, and product categories.

The main objective is to extract **actionable business insights** related to:

- Customer segmentation
- Spending behavior
- Product performance
- Discount and promotion effectiveness

These insights can help businesses optimize marketing strategies, improve customer retention, and enhance product offerings.

II. Dataset Summary

- **Total Records:** 3,900 customer purchases
- **Total Columns:** 18

Key Data Categories:

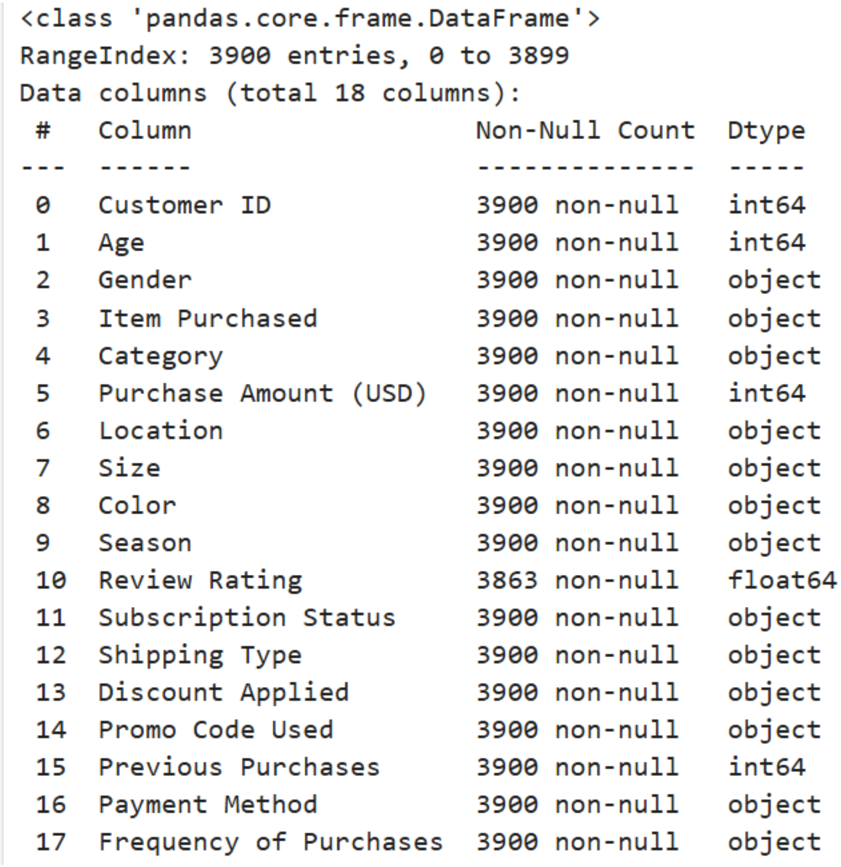
- **Customer Information:**
  Age, Gender, Location, Subscription Status
- **Purchase Details:**
  Item Purchased, Product Category, Purchase Amount, Season, Size, Color
- **Behavioral Data:**
  Discount Applied, Promo Code Used, Previous Purchases, Purchase Frequency, Review Rating, Shipping Type
- **Missing Values:**
  37 missing values were found in the *Review Rating* column.

III. Data Preparation and Exploratory Data Analysis (Python)

Python was used for initial data exploration and cleaning.

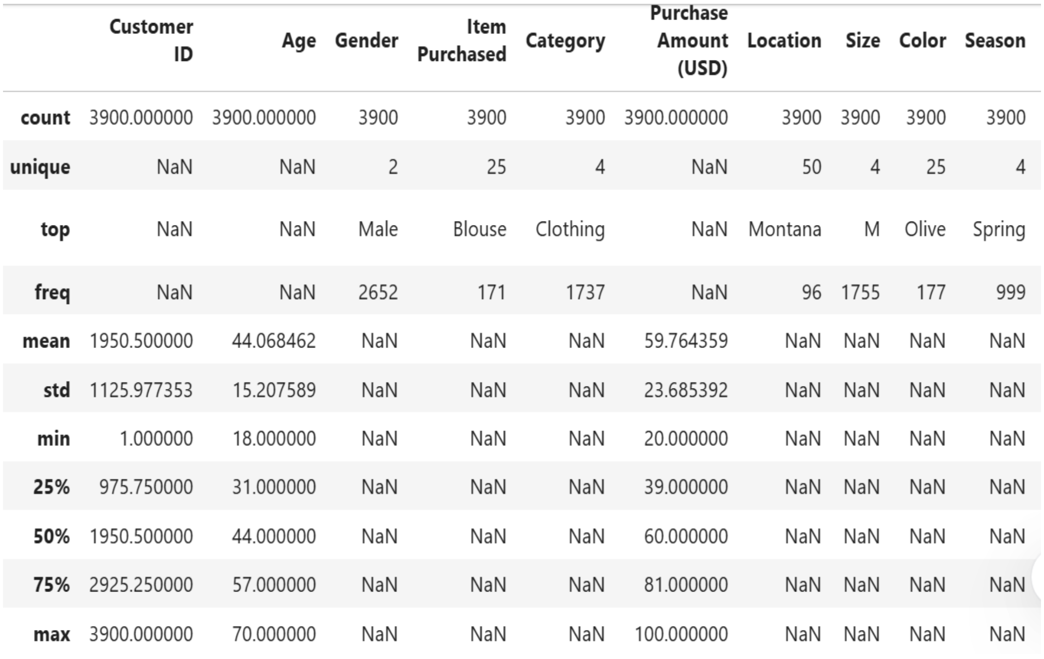
Steps Performed:

- **Data Loading:** Dataset imported using pandas
- **Structure Analysis:**
  - df.info() to inspect data types and missing values



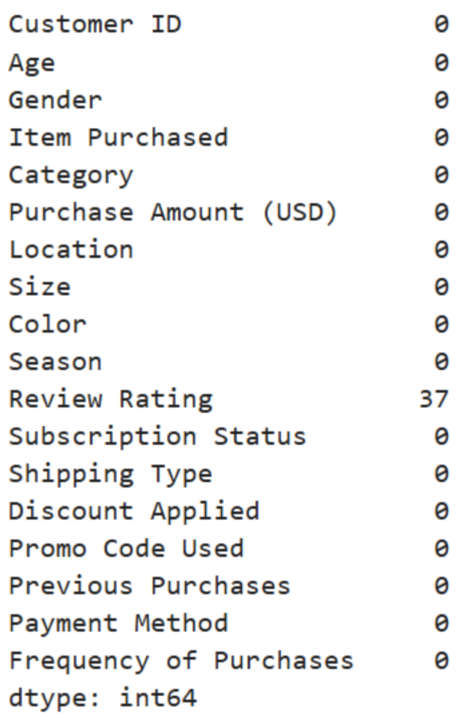
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Customer ID             3900 non-null   int64  
 1   Age                     3900 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 
```

- df.describe() for summary statistics

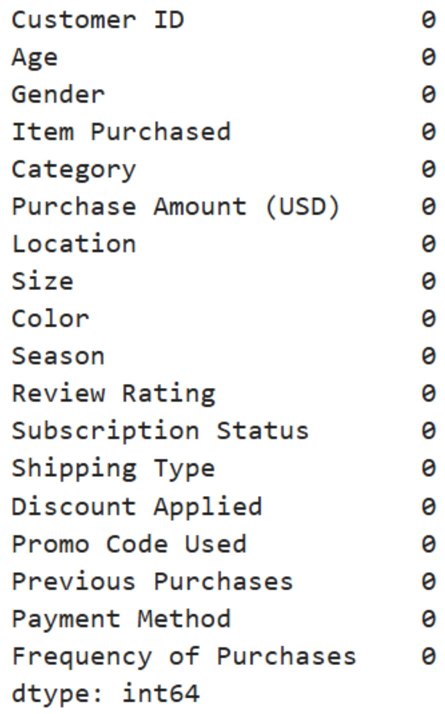


	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring
freq	NaN	NaN	2652	171	1737	NaN	96	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

- **Missing Values Handling:**
  Review ratings with missing values were analyzed and handled appropriately



```
Customer ID                0
Age                        0
Gender                     0
Item Purchased             0
Category                   0
Purchase Amount (USD)      0
Location                   0
Size                       0
Color                      0
Season                     0
Review Rating             37
Subscription Status        0
Shipping Type              0
Discount Applied           0
Promo Code Used            0
Previous Purchases         0
Payment Method             0
Frequency of Purchases     0
dtype: int64
```



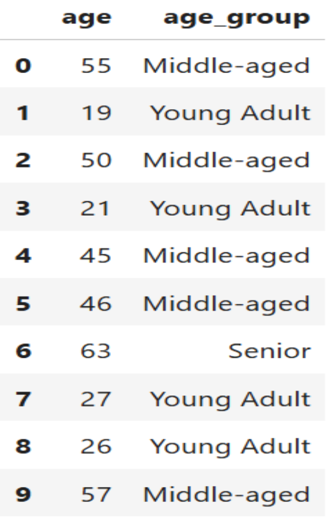
```
Customer ID               0
Age                       0
Gender                    0
Item Purchased            0
Category                  0
Purchase Amount (USD)     0
Location                  0
Size                      0
Color                     0
Season                    0
Review Rating             0
Subscription Status       0
Shipping Type             0
Discount Applied          0
Promo Code Used           0
Previous Purchases        0
Payment Method            0
Frequency of Purchases    0
dtype: int64
```

- **Basic Explorations:**
  - Distribution of purchase amounts
  - Customer demographics overview
  - Discount usage frequency

Python helped in understanding the dataset structure before moving to deeper analysis.

- **Column Standardization:**
  Standardized column names by converting them to *snake_case* to improve readability, consistency, and ease of use across Python, SQL, and Power BI workflows.
- **Feature Engineering:**

  Created an **age_group** variable by categorizing customer ages into meaningful segments to support demographic analysis.



	age	age_group
0	55	Middle-aged
1	19	Young Adult
2	50	Middle-aged
3	21	Young Adult
4	45	Middle-aged
5	46	Middle-aged
6	63	Senior
7	27	Young Adult
8	26	Young Adult
9	57	Middle-aged

Derived a **purchase_frequency_days** feature to better represent customer purchasing behavior over time.

- **Data Consistency Validation:**
  Assessed the relationship between **discount_applied** and **promo_code_used** variables. As both fields conveyed overlapping information, **promo_code_used** was removed to reduce redundancy and improve data clarity.
- **Database Integration:**
  Connected the Python environment to a **PostgreSQL** database and loaded the cleaned and transformed dataset to enable efficient SQL-based analysis.

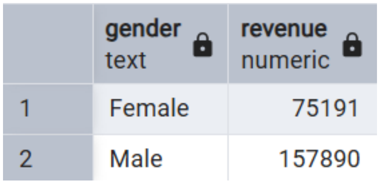
IV. SQL Analysis

Initial Data Validation

Before conducting advanced SQL analysis, a preliminary query was executed to inspect the structure and contents of the dataset. This step ensured that the data was correctly loaded into the PostgreSQL database and that all columns were accessible for querying.

1. Total Revenue by Gender

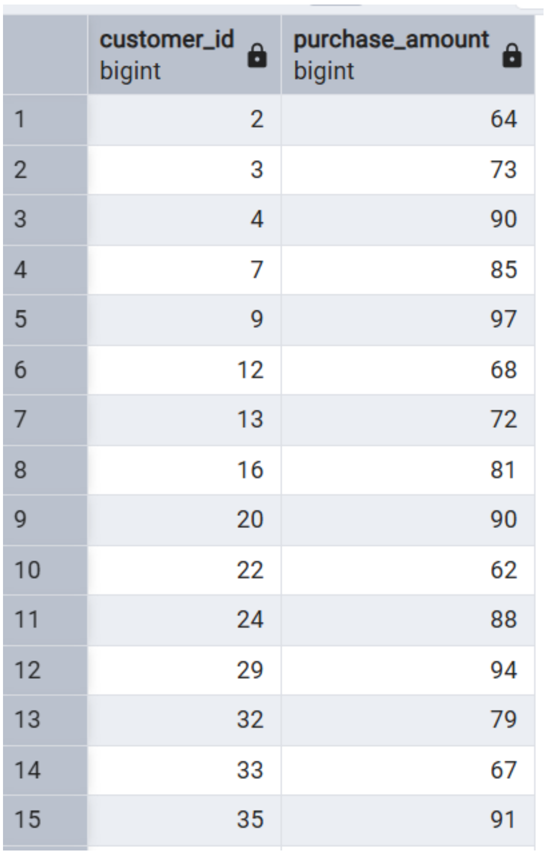
This query analyzes the total revenue generated by customers based on gender.
The objective is to identify whether purchasing power differs between male and female customers, which can help businesses tailor marketing strategies and product positioning accordingly.



	gender text	revenue numeric
1	Female	75191
2	Male	157890

2. High-Value Customers Using Discounts

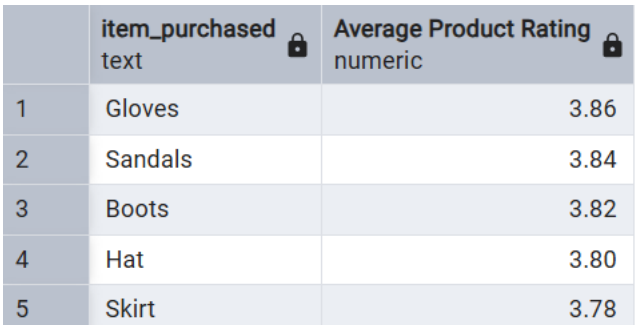
This analysis identifies customers who applied a discount but still spent more than the average purchase amount.
It helps uncover high-value customers who remain strong contributors to revenue despite promotional incentives, indicating effective discount strategies.



	customer_id bigint	purchase_amount bigint
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68
7	13	72
8	16	81
9	20	90
10	22	62
11	24	88
12	29	94
13	32	79
14	33	67
15	35	91

3. Top 5 Products by Average Review Rating

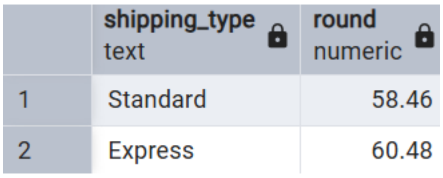
This query calculates the average review rating for each product and ranks them to identify the top 5 highest-rated items.
The results highlight products with strong customer satisfaction, which can be prioritized for promotion or inventory expansion.



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

4. Average Purchase Amount by Shipping Type

This analysis compares the average purchase amount between customers who selected Standard shipping versus Express shipping.
The goal is to understand whether faster shipping options are associated with higher spending behavior.



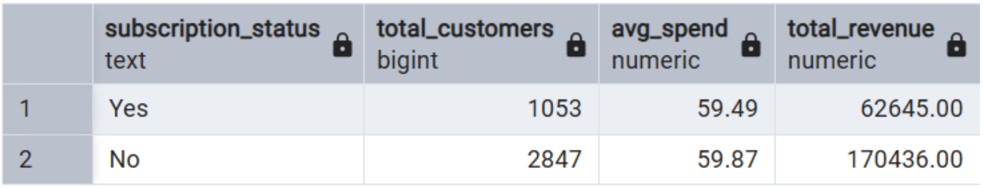
	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. Spending Behavior of Subscribed vs. Non-Subscribed Customers

This query compares subscriber and non-subscriber customers in terms of:

- Number of customers
- Average spending per customer
- Total revenue contribution

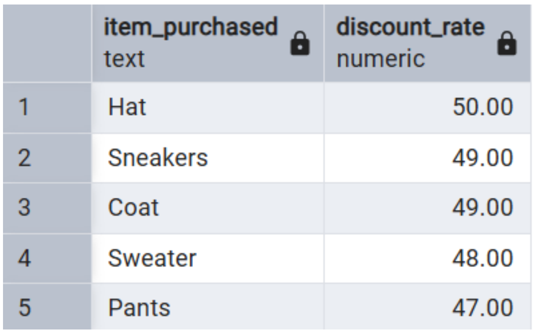
The analysis helps assess the financial impact of the subscription program and its effectiveness in increasing customer value.



	subscription_status 🔒 text	total_customers 🔒 bigint	avg_spend 🔒 numeric	total_revenue 🔒 numeric
1	Yes	1053	59.49	62645.00
2	No	2847	59.87	170436.00

6. Products with the Highest Discount Utilization

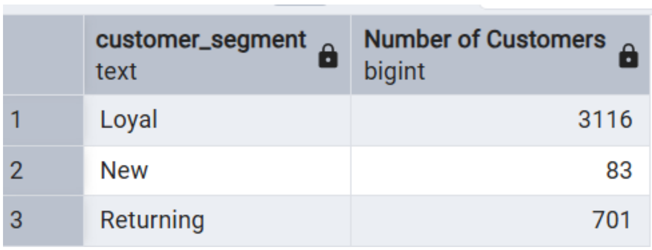
This analysis identifies the top 5 products with the highest percentage of purchases made using discounts.
It provides insight into which products are most dependent on promotional pricing and can help evaluate discount strategy effectiveness.



	item_purchased 🔒 text	discount_rate 🔒 numeric
1	Hat	50.00
2	Sneakers	49.00
3	Coat	49.00
4	Sweater	48.00
5	Pants	47.00

7. Customer Segmentation Based on Purchase History

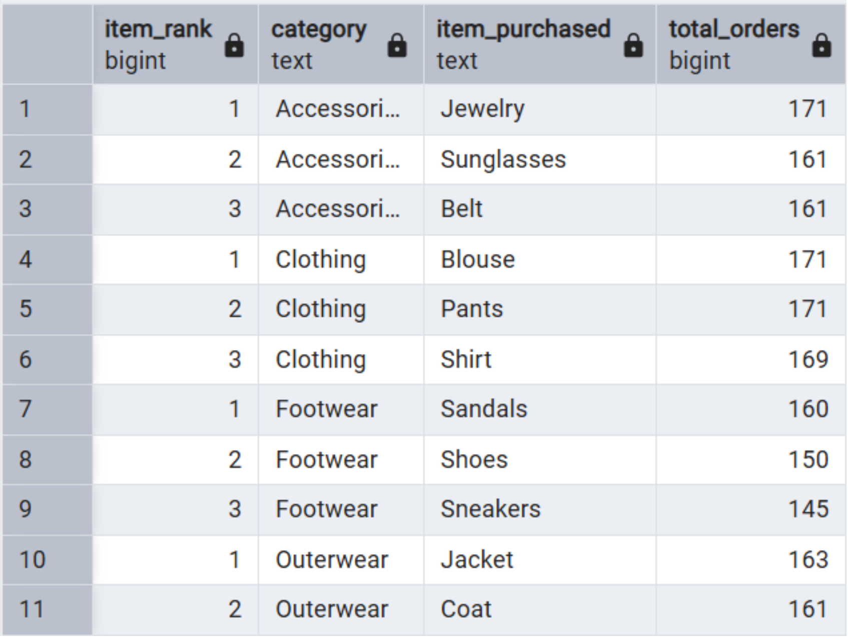
Customers are segmented into **New**, **Returning**, and **Loyal** categories based on their number of previous purchases.
This segmentation helps understand customer distribution across lifecycle stages and supports targeted retention strategies.



	customer_segment 🔒 text	Number of Customers 🔒 bigint
1	Loyal	3116
2	New	83
3	Returning	701

8. Top 3 Most Purchased Products per Category

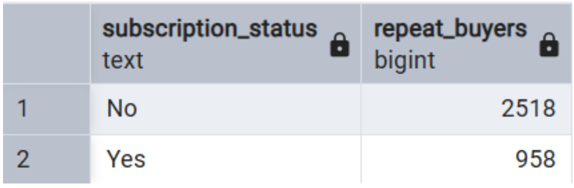
This query ranks products within each category based on total number of orders and extracts the top 3 per category.
It highlights category-level bestsellers and supports inventory and merchandising decisions.



	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	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. Subscription Behavior of Repeat Buyers

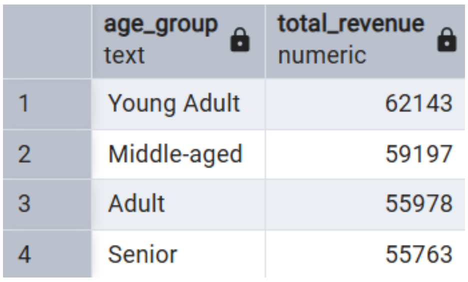
This analysis examines whether repeat buyers (customers with more than 5 previous purchases) are more likely to be subscribed.
It helps evaluate the relationship between customer loyalty and subscription adoption.



	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

10. Revenue Contribution by Age Group

This query calculates total revenue generated by each age group.
The objective is to identify the most valuable age segments, enabling more effective demographic targeting and strategic planning.



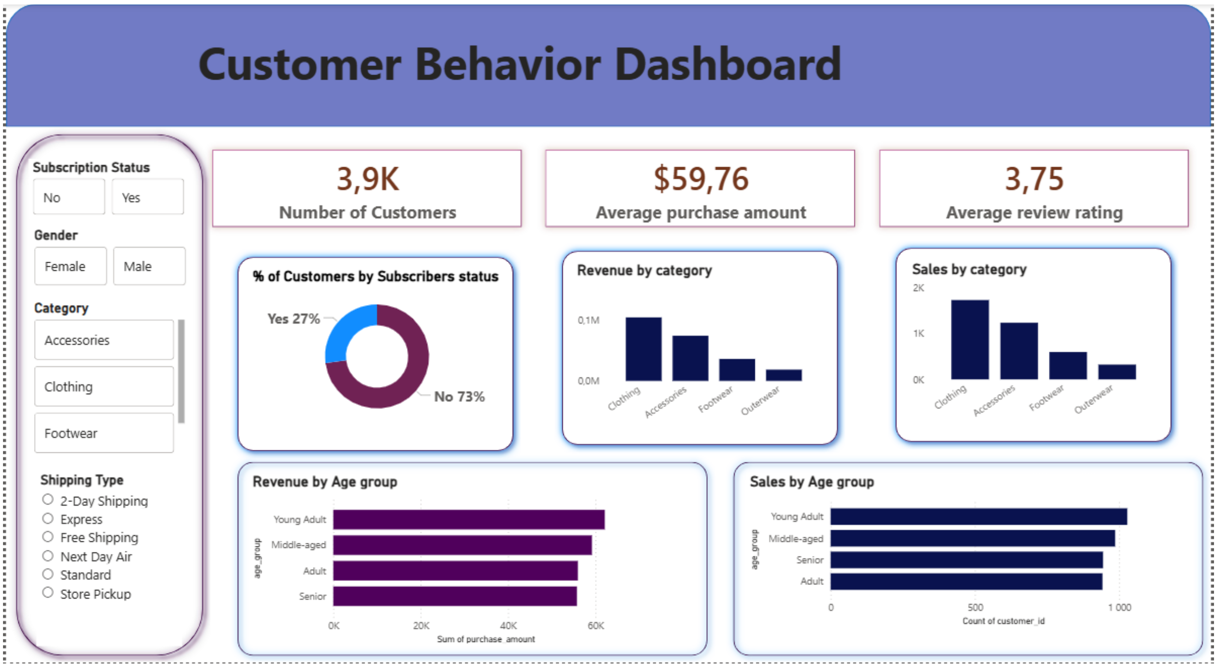
	age_group text	total_revenue numeric
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

Conclusion of SQL Analysis

The SQL analysis provided actionable insights into customer behavior, spending patterns, product performance, and subscription impact.
These insights were later visualized using Power BI to support data-driven decision-making.

V. Dashboard in Power BI

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



Customer Behavior Dashboard

3,9K
Number of Customers

$59,76
Average purchase amount

3,75
Average review rating

VI. Business Recommendations

1. **Target High-Value Customer Segments**
   Focus marketing efforts on customer segments (gender and age groups) that generate the highest revenue to maximize return on investment.

2. **Use Discounts Strategically**
   Apply discounts selectively to avoid margin loss, especially for customers who already spend above the average purchase amount.

3. **Promote Top-Rated Products**
   Leverage highly rated products in marketing campaigns and ensure sufficient stock to capitalize on strong customer satisfaction.

4. **Strengthen Subscription Programs**
   Enhance subscription benefits to increase adoption and retention, particularly among repeat and loyal customers.

5. **Optimize Shipping Incentives**
   Encourage higher spending by offering Express shipping incentives for high-value orders.

6. **Adjust Product Pricing and Promotions**
   Review products with high discount dependency and test alternative pricing or promotional strategies.

7. **Develop Segment-Specific Customer Strategies**
   Design tailored actions for New, Returning, and Loyal customers to improve retention and lifetime value.