Customer Shopping Behavior Analysis

1. Project Overview

This project explores customer shopping behavior based on transactional data from 3,900 records covering various product categories. The purpose of this study is to reveal patterns in customer spending, segmentation, product preferences, and subscription activity, helping businesses make informed decisions.

2. Dataset Summary

The dataset comprises 3,900 rows and 18 columns. Major attributes include:

• Customer demographics: Age, Gender, Location, and Subscription Status
• Transactional information: Item Purchased, Category, Purchase Amount, Season, Size, Color
• Behavioral attributes: Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, and Shipping Type
A total of 37 values were missing in the Review Rating field.

3. Exploratory Data Analysis using Python

Data preprocessing and cleaning were performed using Python. Steps included:

• Loading the dataset with pandas
• Inspecting its structure with df.info() and descriptive statistics using df.describe()

```
df.describe(include='all')
```
✓ 0.0s

	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

• Handling missing values by imputing the Review Rating column with category-wise median values

• Renaming columns into snake_case format for consistency
• Creating new features such as 'age_group' (by binning age values) and 'purchase_frequency_days' (based on the frequency of transactions)

```
df[['age','age_group']].head(10)
```
✓ 0.0s

	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

• Checking redundancy between 'discount_applied' and 'promo_code_used'—and dropping the latter
• Integrating the cleaned dataset into PostgreSQL for SQL-based analysis

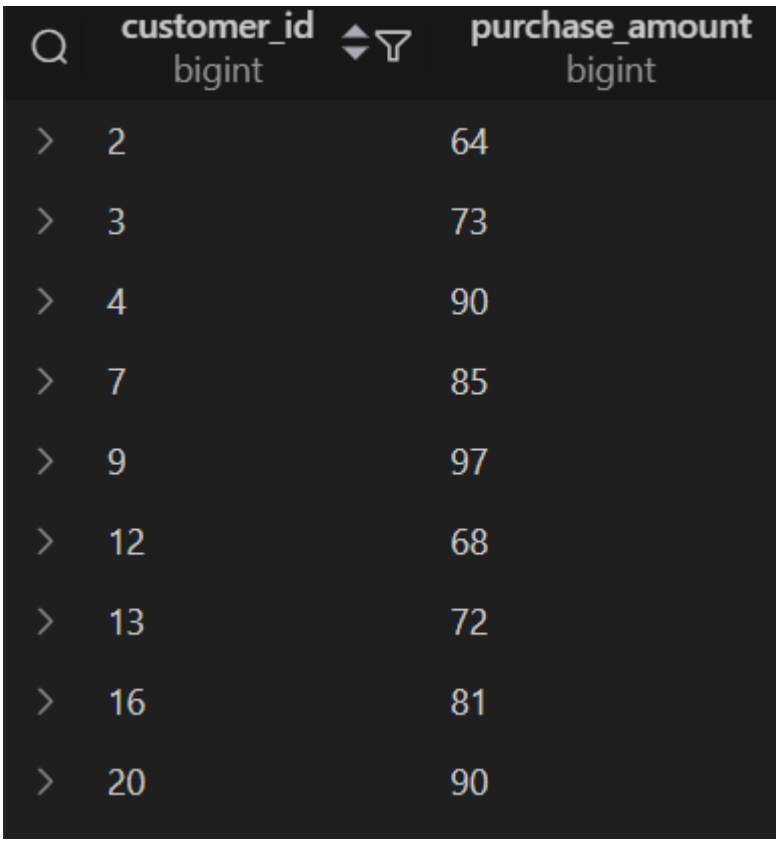
4. Data Analysis using SQL

Structured SQL queries were developed to answer key business questions:

1. Compared revenue generated by male and female customers.

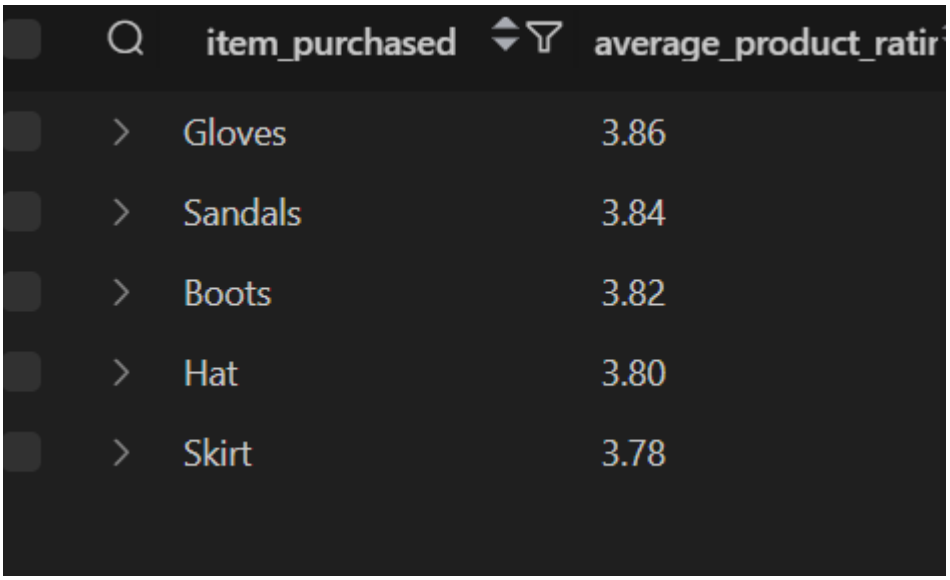
gender	total_revenue
Female	75191
Male	157890

2. Identified customers using discounts who still spent above the average purchase amount.



customer_id bigint	purchase_amount bigint
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90

3. Listed the top 5 products based on average review ratings.



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

4. Compared average purchase values between Standard and Express shipping methods.

shipping_type	urchase_amour
Standard	58.46
Express	60.48

5. Analyzed spending and total revenue for subscribed versus non-subscribed customers.

subscription_status	total_customers bigint	avg_spend	total_revenue
Yes	1053	59.49	62645.00
No	2847	59.87	170436.00

6. Found products with the highest proportion of discounted purchases.

item_purchased	discount_rate
Hat	50.00
Sneakers	49.00
Coat	49.00
Sweater	48.00
Pants	47.00

7. Segmented customers as New, Returning, or Loyal using previous purchase counts.

customer_segment	bigint
Loyal	3116
New	83
Returning	701

8. Determined the top 3 most purchased products within each category using window functions.

item_rank bigint	category	iten	total_orders bigint
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

9. Measured the relationship between repeat buyers (more than 5 previous purchases) and subscription likelihood.

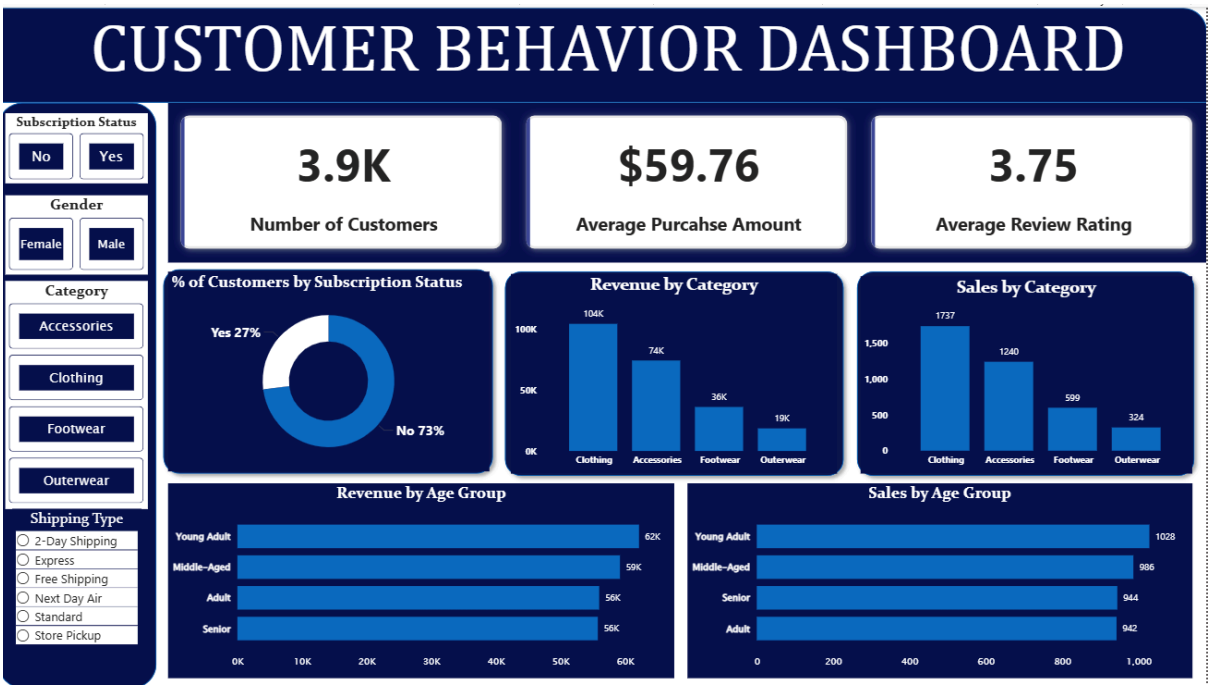
subscription_status	Repeat_Buyers bigint
No	2518
Yes	958

10. Calculated total revenue contribution by age group.

age_group	Total_revenue
Young Adult	62143
Middle-Aged	59197
Adult	55978
Senior	55763

5. Dashboard in Power BI

An interactive Power BI dashboard was designed to visually present the SQL findings. The dashboard displays metrics such as revenue by gender, purchase behavior by age group, product ratings, and discount utilization. Filters for season, category, and location allow dynamic exploration of customer behavior.



6. Business Recommendations

Based on the insights derived, the following strategies were recommended:

• Strengthen Subscription Offers: Create targeted campaigns promoting subscriber-only benefits.
• Encourage Customer Loyalty: Launch point-based or tiered reward programs for frequent shoppers.
• Optimize Discount Policies: Balance discounts to retain profitability while sustaining sales volume.
• Promote Top-rated Products: Showcase high-rated and frequently purchased products in marketing materials.
• Focus Marketing on High-performing Segments: Prioritize express-shipping customers and high-spending age groups.