Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior using transactional data from 3,900 purchases across various product categories. The goal is to uncover insights into spending patterns, customer segments, product preferences, 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 Rating, Shipping Type)

- 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: Imported the dataset using pandas.

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

```
[5 rows x 18 columns]
       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
<class 'pandas.core.frame.DataFrame'>
```

```
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 the 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 Check: Verified if discount_applied and promo_code_used were redundant; dropped promo_code_used

.● Database Integration: Connected Python script to MYSQL and loaded the cleaned DataFrame into the database for SQL analysis.

4 Data Analysis using SQL (Business Transactions)

We performed structured analysis in MYSQL to answer key business questions:

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

revenue	gender
157890	Male
75191	Female

2. High-Spending Discount Users – Identified customers who used discounts but still spent above the average purchase amount.

Result Grid | Filter Rows:

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

customer 16

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

Result Grid | Filter Rows:

AVG(review_rating)	item_purchased
3.8614285714285725	Gloves
3.8443750000000003	Sandals
3.8187500000000005	Boots
3.8012987012987005	Hat
3.784810126582278	Skirt

Result 1

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

Result Grid | Filter Rows:

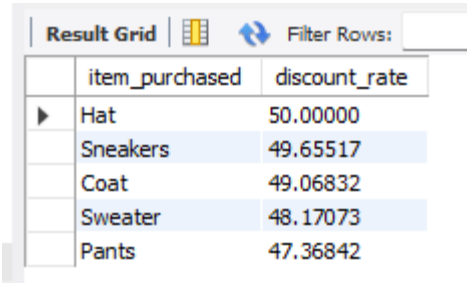
avg(purchase_amount)	shipping_type
60.4752	Express
58.4602	Standard

5. Subscribers vs. Non-Subscribers – Compared average spend and total revenue across subscription status.

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

count(customer_id)	avg(purchase_amount)	sum(purchase_amount)	subscription_status
1053	59.4919	62645	Yes
2847	59.8651	170436	No

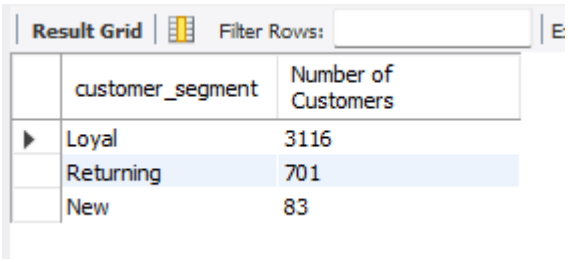
6. Discount-Dependent Products – Identified 5 products with the highest percentage of discounted purchases.



Result Grid | Filter Rows:

item_purchased	discount_rate
Hat	50.00000
Sneakers	49.65517
Coat	49.06832
Sweater	48.17073
Pants	47.36842

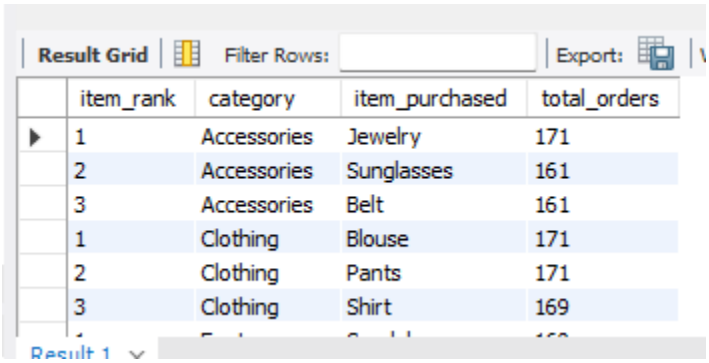
7. Customer Segmentation – Classified customers into New, Returning, and Loyal segments based on purchase history.



Result Grid | Filter Rows:

customer_segment	Number of Customers
Loyal	3116
Returning	701
New	83

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

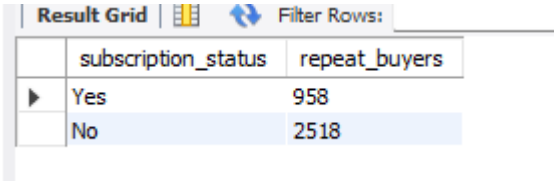


Result Grid | Filter Rows: | Export:

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

Result 1

9. Repeat Buyers & Subscriptions – Checked whether customers with >5 purchases are more likely to subscribe.



Result Grid | Filter Rows:

subscription_status	repeat_buyers
Yes	958
No	2518

10. Revenue by Age Group – Calculated total revenue contribution of each age group.

Result Grid | Filter Rows:

age_group	total_revenue
Young Adult	74658
Middle Age	67916
Adult	65013
Senior	25494

65013

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 Programs – Reward repeat buyers to move them into the “Loyal” segment.
- Review Discount Policy – Balance sales boosts with margin control.
- Product Positioning – Highlight top-rated and best-selling products in campaigns.
- Targeted Marketing – Focus efforts on high-revenue age groups and express-shipping users.