Customer Shopping Behavior Analysis

1. Project Overview

This project examines 3,900 customer purchase transactions to uncover key revenue drivers, analyze customer segmentation and subscription behavior, and evaluate how product categories and discount strategies influence overall sales performance and customer value.

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, Purchase Frequency, Review Rating, Shipping Type
- **Data Quality:**
  - The dataset contained **37 missing values in the Review Rating column**, accounting for **less than 1% of the total data**, which were handled to ensure unbiased product rating analysis.

3. Exploratory Data Analysis using Python

Python was used for data preparation, validation, and feature engineering before downstream SQL analysis.

- **Data Loading:** Imported the dataset using pandas and verified data types and structure.
- **Initial Exploration:** Used df.info() and df.describe() to understand column distributions, ranges, and potential data quality issues.

	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'>
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:** Identified missing values in the *Review Rating* column and imputed them using the median rating within each product category to preserve category-level rating behavior.
- **Column Standardization:** Renamed all columns to **snake_case** to improve consistency and maintainability across Python, SQL, and Power BI.
- **Feature Engineering:**
  - Created an age_group column by binning customer ages to enable demographic-level analysis.
  - Created a purchase_frequency_days feature to support customer behavior and segmentation analysis.

- **Data Consistency Checks:** Evaluated overlap between *discount_applied* and *promo_code_used* and removed *promo_code_used* as a redundant feature.
- **Database Integration:** Loaded the cleaned and transformed dataset into **MYSQL** to perform structured business analysis using SQL.

4. Data Analysis using SQL (Business Transactions)

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

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

	gender	revenue
▶	Female	81585
	Male	151496

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

	customer_id	purchase_amount
▶	2	64
	3	73
	4	90
	7	85
	9	97
	12	68
	13	72
	16	81
	20	90
	22	62
	24	88

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

	item_purchased	Average Product Rating
▶	Gloves	3.86
	Sandals	3.84
	Boots	3.82
	Hat	3.8
	Skirt	3.78

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

	shipping_type	avg_purchase_amount
▶	Express	60.48
	Standard	58.46

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

subscription_status	total_customers	avg_spend	total_revenue
Yes	1053	59.49	62645
No	2847	59.87	170436

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

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

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

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.

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
1	Footwear	Sandals	160
2	Footwear	Shoes	150
3	Footwear	Sneakers	145
1	Outerwear	Jacket	163
2	Outerwear	Coat	161

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

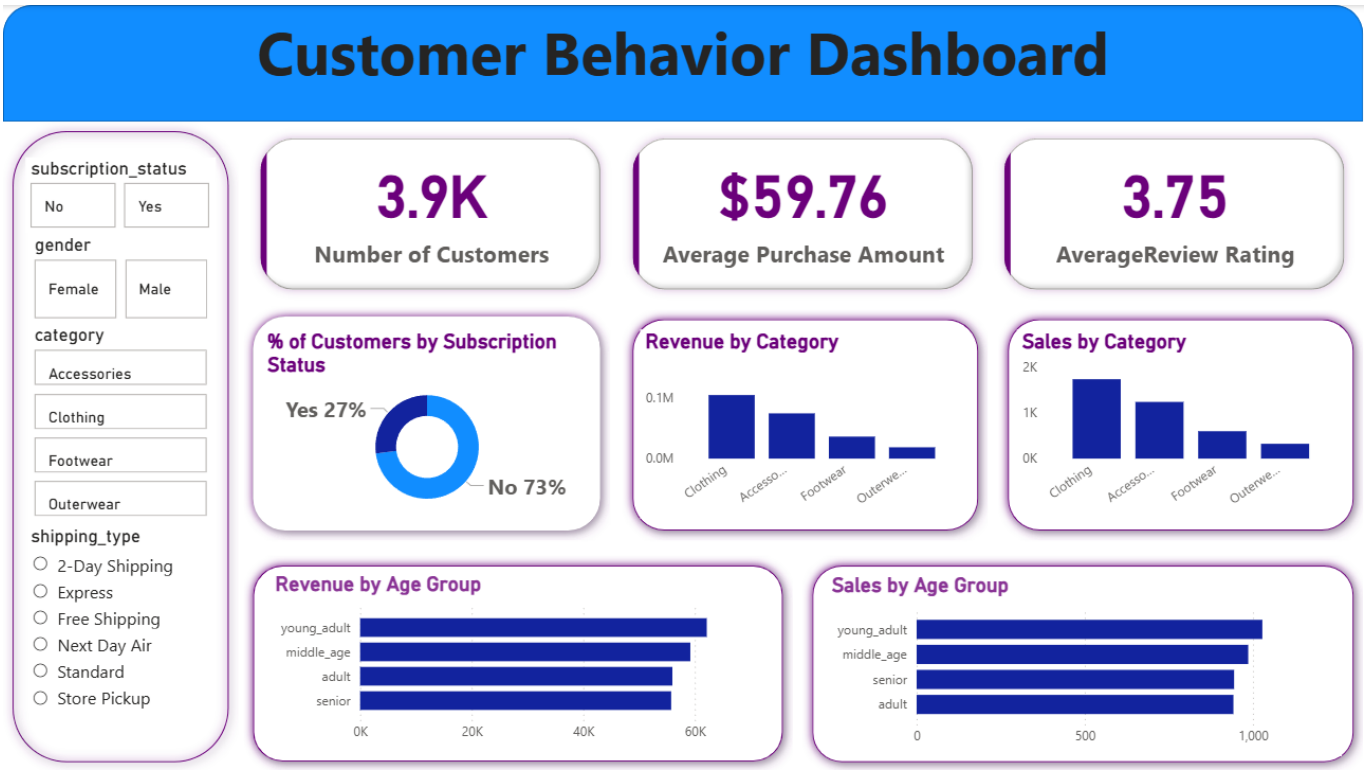
subscription_status	repeat_buyers
Yes	958
No	2518

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

age_group	total_revenue
young_adult	62143
middle_age	59197
adult	55978
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 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.