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.

	customer_id	age	gender	item_purchased	category	purchase_amount_(usd)	location	size	color	season	review_rating	subscription_status	shipping_type	disco
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3900.000000	3900	3900	
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750051	NaN	NaN	
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.713590	NaN	NaN	
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	

discount_applied	promo_code_used	previous_purchases	payment_method	frequency_of_purchases
3900	3900	3900.000000	3900	3900
2	2	NaN	6	7
No	No	NaN	PayPal	Every 3 Months
2223	2223	NaN	677	584
NaN	NaN	25.351538	NaN	NaN
NaN	NaN	14.447125	NaN	NaN
NaN	NaN	1.000000	NaN	NaN
NaN	NaN	13.000000	NaN	NaN
NaN	NaN	25.000000	NaN	NaN
NaN	NaN	38.000000	NaN	NaN
NaN	NaN	50.000000	NaN	NaN

● **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 PostgreSQL and loaded the cleaned DataFrame into the 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** – Compared total revenue generated by male vs. female customers.

	gender text	total_revenue numeric
1	Female	75191
2	Male	157890

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

	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

Total rows: 839 Query complete 00:00:00.138

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

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

	shipping_type text	avg_amount numeric
1	Standard	58.46
2	Express	60.48

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

	subscription_status text	total_customer bigint	total_revenue numeric	avg_spend numeric
1	No	2847	170436	59.87
2	Yes	1053	62645	59.49

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

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

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

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

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

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

Total rows: 11 | Query complete 00:00:00.127

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

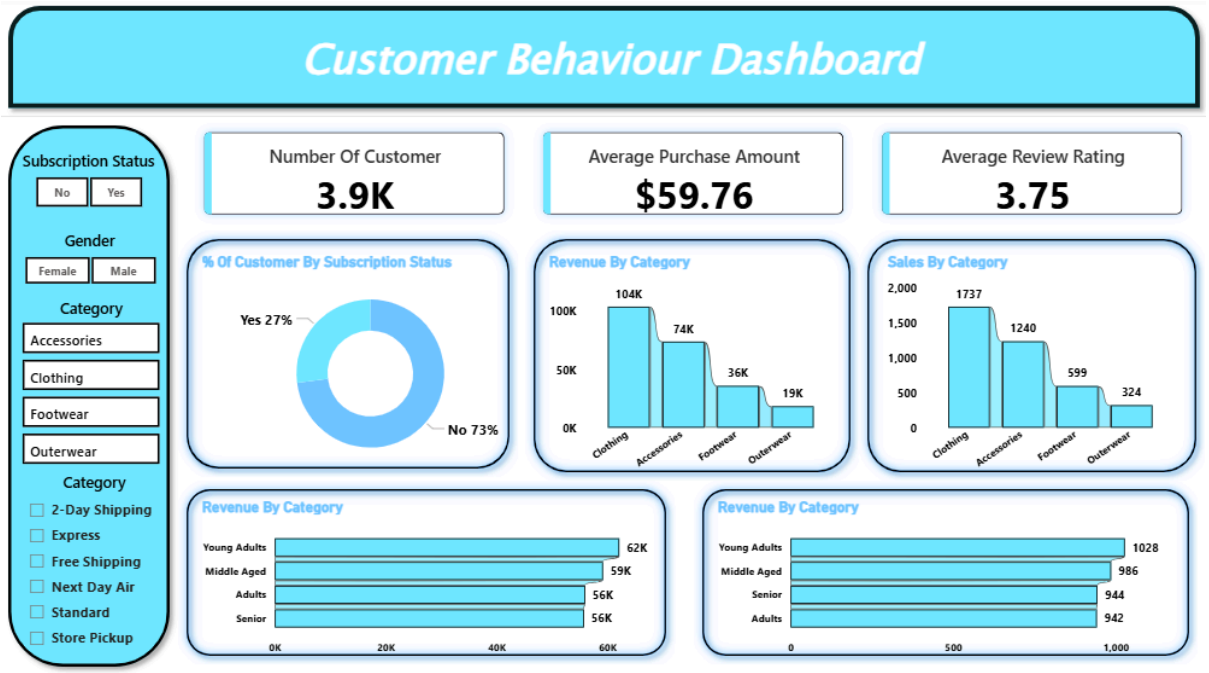
	subscription_status text	repeat_cus bigint
1	No	2583
2	Yes	980

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

	age_group text	sum numeric
1	Young Adul...	62143
2	Middle Aged	59197
3	Adults	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 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.