Customer Shopping Behavior Analysis

1.Introduction

This project investigates customer shopping behaviours across various product categories by examining 3,900 purchase records. The objective is to gain insights into spending habits, customer demographics, product performance, discount utilization, and subscription patterns. These findings will assist businesses in pinpointing opportunities to enhance marketing strategies, pricing, and customer retention.

2. Dataset Summary

The dataset contains 3,900 rows and 18 attributes, covering customer details, product information, and purchase behaviour. Key fields include:

- **Demographics:** age, gender, location, subscription status
- **Purchase details**: item purchased, category, amount, season
- **Shopping behaviour**: discount usage, promo codes, previous purchases, shipping type, and review rating

A small portion of entries had missing values in the review rating column.

3. Data Preparation & Exploration (Python)

Python was used for cleaning, formatting, and preparing the data before loading it into PostgreSQL. The main steps included:

• Data loading and structure checks

Used pandas to load the data and reviewed the structure with .info() and .describe().

[9]:

	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

Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
Yes	14	Venmo	Fortnightly
Yes	2	Cash	Fortnightly
Yes	23	Credit Card	Weekly
Yes	49	PayPal	Weekly
Yes	31	PayPal	Annually

• Handling missing values
Missing review ratings were imputed using the median rating within each product category.

• Renaming columns
Column names were standardized into snake_case to improve readability and maintain consistency.

• Feature engineering
- Created age_ group by binning customer ages into meaningful categories.
- Added purchase_ frequency_ days by analyzing gaps between purchases.

• Removing redundant information
promo_code_used was dropped after confirming that discount_applied already captured the same behaviour.

• Database integration
Connected Python script to PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

4. SQL Analysis in PostgreSQL

Structured queries were used to analyze trends across demographic groups, product categories, and purchasing behaviours.

1. Revenue by Gender : Compared total purchase revenue generated by male and female shoppers.

	gender text	revenue numeric
1	Female	75191
2	Male	157890

2. High-Spending Discount Users : Identified customers who used discounts but still spent above the overall average purchase amount. This group may be valuable for targeted upselling.

	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 00:00:00.271

3. Top 5 Products by Rating : Ranked products by average review rating. Items like Gloves, Sandals, Boots, Hats, and Skirts scored the highest.

	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.
Express shipping customers spent slightly more on average

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. Subscribers vs Non-Subscribers : Evaluated how subscription status influenced spending and revenue. Non-subscribers accounted for a higher total revenue due to larger customer count

	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. Discount-Dependent Products : Found products most commonly purchased with discounts.

	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

7. Customer Segmentation : Classified customers as New, Returning, or Loyal based on the number of purchases.

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

8. Top 3 Products in Each Category : Listed the most frequently purchased items inside each product category.

	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

Total rows: 11 Query complete 00:00:00.193

9. Repeat Buyers and Subscriptions : Checked whether customers who made more than five purchases were more likely to subscribe. Loyal customers showed a higher subscription tendency.

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

10. Revenue Contribution by Age Group : Calculated which age groups contributed most to overall revenue. Young adults and middle-aged customers led in revenue.

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

5. Power BI Dashboard

A Power BI dashboard was created to visualize key insights interactively. The dashboard highlights:

- Total customers: 3.9K
- Average purchase amount: $59.76
- Average review rating: 3.75
- Subscription breakdown
- Revenue by category
- Revenue and sales by age group
- Top products
- Interactive slicers (Gender, Category, Shipping Type, etc.)

6. Business Recommendations

1. Increase Subscription Adoption

Subscribers show higher average spending. Offering extra benefits can boost conversion.

2. Strengthen Loyalty Programs

Returning and loyal customers can be encouraged further with discounts and point systems.

3. Optimize Discount Strategy

Some items rely heavily on discounts; reviewing discount frequency preserves margins.

4. Promote High-Rated Products
Products like Gloves, Sandals, and Boots consistently perform well and should be featured in promotions.
5. Target High-Revenue Age Segments
Young adults and middle-aged shoppers generate strong revenue and should be prioritized in marketing.

7. Conclusion

This project combined Python, PostgreSQL and Power BI to build a complete customer behaviour analysis pipeline.
Data cleaning improved quality, SQL queries delivered business insights, and Power BI visualized the results clearly for decision-makers.