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	Discou Appli
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	39
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	22
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	N
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	N
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	N
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	N
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	N
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	N
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	N

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. **What is the total revenue generated by male vs. female customers?**

   Objective: Compare total spending between male and female customers to understand gender-based contribution to overall sales.

Gender	Revenue
Male	157890
Female	75191

2. **How many customers belong to each gender, and what's their average and total spending?**

   Objective: Analyze the customer distribution by gender and understand how much each gender spends on average.

Gender	Total_Customers	Revenue	Avg_Purchase
Male	2652	157890	59.54
Female	1248	75191	60.25

3. **Which customers used discounts but still spent more than the overall average purchase amount?**

   Objective: Identify high-value customers who utilized discounts yet made above-average purchases

customer_id	purchase_amount
43	100
96	100
194	100
205	100
244	100
249	100
456	100

customers 7 ×

4. **Which gender has the highest number of high-spending discount users?**

   Objective**:** Check whether male or female customers are more likely to spend above average even with discounts applied

gender	high_spenders_with_discount
Male	839

5. **What are the top 5 products with the highest average review ratings?**

   Objective: Identify the best-rated products to understand customer satisfaction trends

item_purchased	average_rating_product	total_reviews
Gloves	3.86	140
Sandals	3.84	160
Boots	3.82	144
Hat	3.8	154
Handbag	3.78	153

6. **What is the average purchase amount for Standard vs. Express shipping types?**

   Objective**:** Compare how much customers spend depending on their shipping preference.

shipping_type	avg_purchase_amount
Express	60.48
Standard	58.46

7. **Which shipping type contributes the most to overall revenue?**

   Objective**:** Analyze order volume, average spend, and total revenue by shipping method.

shipping_type	total_orders	total_revenue	avg_purchase_amount
Express	646	39067	60.48
Standard	654	38233	58.46

8. **Do subscribed customers spend more than non-subscribers?**

   Objective**:** Compare subscribers vs. non-subscribers in terms of total customers, average spend, and revenue contribution.

subscription_status	total_customer	avg_spend	total_spend	revenue_percentage
No	2847	59.87	170436	73.12
Yes	1053	59.49	62645	26.88

9. **What is the discount usage rate per product?**

   Objective**:** Identify which products are most frequently purchased using discounts.

Result Grid Filter Rows:

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

10. How do purchase behaviors differ between discounted and non-discounted orders for each product?

Objective**:** Compare how discounts influence spending patterns across products.

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

item_purchased	total_orders	discounted_orders	discount_rate	avg_discounted_spend	avg_non_discounted_spend
Hat	154	77	50.00	59.69	62.06
Sneakers	145	72	49.66	60.03	59.08
Coat	161	79	49.07	56.58	58.60
Sweater	164	79	48.17	58.35	57.08
Pants	171	81	47.37	59.72	58.37
Boots	144	67	46.53	62.72	62.55
Jeans	124	57	45.97	56.12	64.91

Result 20

11. How can customers be segmented based on previous purchases (New, Returning, Loyal)?

Objective**:** Classify customers based on loyalty and analyze their contribution to total sales.

customer_segment	customer_count	percentage_share
Loyal	3116	79.90
Returning	701	17.97
New	83	2.13

12. What is the average spending of each customer segment?

Objective**:** Understand which customer group contributes the most to overall revenue.

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

customer_segment	total_customers	avg_purchase	total_revenue
Loyal	3116	59.54	185517
Returning	701	60.93	42711
New	83	58.47	4853

13. What are the top 3 most purchased products within each category?

Objective**:** Determine product popularity across different categories.

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

category	item_rank	item_purchased	total_order
Accessories	1	Jewelry	171
Accessories	2	Sunglasses	161
Accessories	3	Belt	161
Clothing	1	Blouse	171
Clothing	2	Pants	171
Clothing	3	Shirt	169
Footwear	1	Sandals	160

Result 23

14. Are repeat buyers (more than 5 previous purchases) more likely to subscribe?

Objective**:** Evaluate the relationship between repeat buying and subscription behavior.

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

subscription_status	repeat_buyers	total_customers	repeat_buyer_percentage
Yes	958	1053	90.98
No	2518	2847	88.44

15. What is the revenue contribution of each age group?

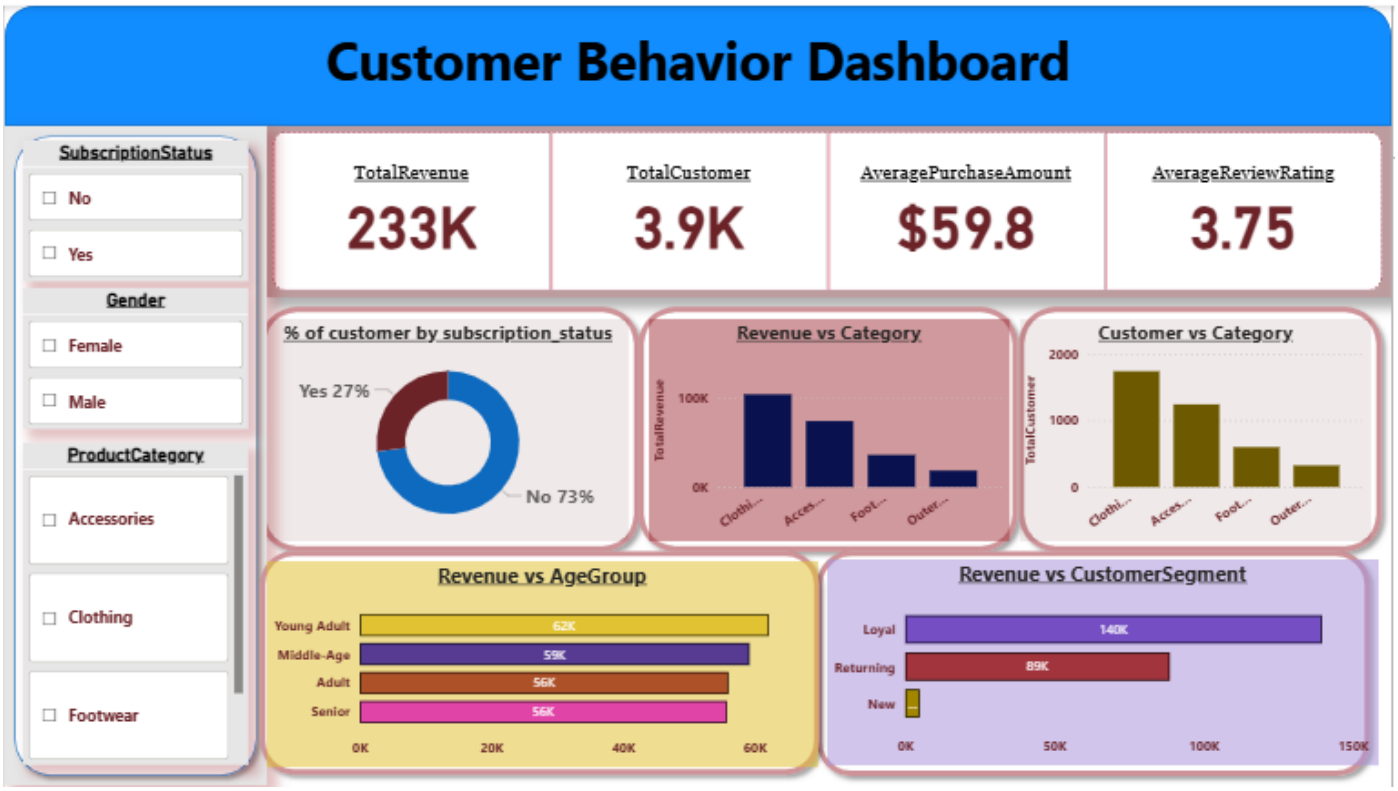
Objective**:** Analyze how much each age group contributes to total revenue.

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

age_group	total_customers	total_revenue	revenue_percentage
Young Adult	1028	62143	26.66
Middle-Age	986	59197	25.40
Adult	942	55978	24.02
Senior	944	55763	23.92

5. Dashboard in Power BI

Finally, we built an interactive dashboard in **Power BI** to present insights visually.



6. Business Summary Insights & Recommendations

Overall Performance:

$233K revenue from 3.9K customers (Avg. Purchase: $59.8 | Rating: 3.75)
-Improve customer experience and aim for >4.0 rating through post-purchase feedback.

Subscription Program:

Only 27% customers subscribed; subscribers spend more.
-Run awareness campaigns, offer 10% sign-up discounts, and add exclusive member perks.

Product Category Insights:

Clothing leads revenue; footwear lags in value per buyer.
-Bundle footwear with accessories and promote top-rated, high-performing products.

Customer Segmentation:

Loyal customers generate the highest revenue share.
-Enhance loyalty programs and send personalized offers to boost repeat purchases.

Age Group Analysis:

Middle-age & adults drive major revenue (~$50K–60K).
-Focus marketing on these groups and introduce trendy, affordable options for youth.

Shipping Insights:

Express shipping users show higher spending behavior.
-Offer limited free express upgrades and target them with premium product deals.

Gender-Based Observations:

Males contribute more revenue; females rate higher satisfaction.
-Promote female-favorite products and design premium bundles for male customers.