CUSTOMER SHOPPING BEHVIOR ANALYSIS

1. Project Overview

This project examines customer shopping behavior using a transactional dataset comprising **3,900 purchase records** across multiple product categories. The objective is to identify meaningful insights related to spending patterns, customer segmentation, product preferences, and subscription behavior in order to support data-driven strategic decision-making.

2. Dataset Summary

- Total Records:3900
- Total Features:18

➢ KEY VARIABLES

- Customer Demographics
  - Age
  - Gender
  - Location
  - Subscription Status
- Purchase Details
  - Item Purchased
  - Product Category
  - Purchase Amount
  - Season
  - Size
  - Color
- Shopping Behavior Metrics
  - Discount Applied
  - Promo Code Usage
  - Previous Purchases
  - Purchase Frequency
  - Review Rating
  - Shipping Type
- Data Quality
  - The dataset contains **37 missing values** in the Review Rating column.
  - No other significant missing data issue were identified

3. Exploratory Data Analysis (EDA) Using Python

The exploratory data analysis phase was conducted using Python to understand the dataset structure, assess data quality, and generate initial statistical insights.

➢ **Data Preparation and Cleaning**

Data Loading:

The dataset was imported into the Python environment using the Pandas library for efficient data manipulation and analysis.

Initial Exploration:

- df.info() was used to examine the dataset structure, including data types, non-null counts, and overall schema validation.
- df.describe(include='all') was applied to generate descriptive statistics for numerical variables, providing insights into central tendency, dispersion, and distribution patterns.

This initial exploration helped identify data inconsistencies, missing values, and potential preprocessing requirements before proceeding to deeper analysis.

Descriptive Statistics of Numerical Variables

```
[9]: df.describe(include='all')
```

[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	Previous Purchases	Payment Method	Frequency of Purchases
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3900.000000	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN	6	7
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	NaN	677	584
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	25.351538	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	14.447125	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	1.000000	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	50.000000	NaN	NaN

[ ]:

[ ]:

Missing Data Handling

A null value assessment was performed across all features. The dataset contained 37 missing values in the **Review Rating** column. These missing values were imputed using the **median review rating within each product category**, ensuring category-level consistency while minimizing the impact of outliers.

Column Standardization

All column names were converted to **snake_case** format to improve readability, maintain consistency with Python conventions, and support cleaner documentation and SQL integration.

Feature Engineering

- **Age Group Creation:**
  An age_group feature was generated by binning customer ages into predefined intervals to enable demographic segmentation and comparative analysis.
- **Purchase Frequency (Days):**
  A new feature, purchase_frequency_days**,** was derived from purchase-related data to quantify customer purchasing intervals in days. This metric supports behavioral segmentation and retention analysis.

Data Consistency Check

A redundancy check was performed between discount_applied and promo_code_used. Since both variables conveyed overlapping promotional information, promo_code_used was removed to eliminate duplication and maintain dataset efficiency.

Database Integration

The cleaned and transformed dataset was integrated with **PostgreSQL** by establishing a connection through Python. The processed DataFrame was then exported to the database, enabling structured SQL-based analysis and advanced querying.

4. Data Analysis Using SQL (Business Transactions)

Following data preprocessing, structured analysis was conducted in **PostgreSQL** to address key business questions and extract actionable insights from transactional data.

1) Revenue by Gender

Following data preprocessing, structured analysis was conducted in **PostgreSQL** to address key business questions and extract actionable insights from transactional data.

	gender text	revenue numeric
1	Female	75191
2	Male	157890

2) High-Spending Discount Users

This analysis aimed to identify customers who used discounts yet recorded a purchase amount above the overall average transaction value**.**

	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:

3) Top 5 Products by Rating

This analysis identified the top five products with the highest average review ratings, based on customer feedback.

	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

This analysis compared the average purchase amount between customers selecting Standard Shipping and Express Shipping.

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5) Subscribers vs. Non-Subscribers

This analysis evaluated differences in purchasing behavior between subscribers and non-subscribers by comparing both average transaction value and total revenue contribution.

	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

This analysis identified the five products with the highest percentage of discounted purchases, highlighting items that rely heavily on promotions to drive sales.

	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

Customers were classified into New, Returning, and Loyal segments based on their purchase history to better understand engagement and retention patterns.

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

8) Customer Segmentation

Customers were classified into New, Returning, and Loyal segments based on their purchase history to better understand engagement and retention patterns.

	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessories	Jewelry	171
2	2	Accessories	Sunglasses	161
3	3	Accessories	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
11	2	Outerwear	Coat	161

9) **Repeat Buyers & Subscriptions**

This analysis examined whether customers with more than five purchases are more likely to be subscribers, exploring the relationship between purchasing frequency and subscription adoption.

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

10) **Revenue by Age Group**

This analysis evaluated the total revenue contribution of each age group, providing insight into which demographic segments generate the highest financial impact.

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

5. Dash Board in Power Bi

Finally we built in interactive Power Bi to present insight visually

6. Business Recommendation

1. Strengthen Subscription Value Proposition

Enhance the attractiveness of the subscription program by offering exclusive, clearly differentiated benefits (e.g., priority shipping, early product access, member-only discounts).

If subscribers do not demonstrate significantly higher average spend or retention, the program must be restructured rather than simply promoted.

2. Implement Tiered Loyalty Programs

Introduce a structured loyalty framework that incentivizes repeat purchases and transitions customers from Returning to Loyal segments.

Reward mechanisms may include milestone-based discounts, points systems, or personalized offers tied to purchase frequency.

3. Optimize Discount Strategy

Reevaluate discount-heavy products to determine whether promotions are driving sustainable growth or eroding margins.

Focus on:

- Reducing excessive discount dependency
- Testing controlled promotional experiments
- Protecting profitability while maintaining competitiveness
- Revenue growth without margin control is financially unstable.

4. Strategic Product Positioning

Prioritize top-rated and best-selling products in marketing campaigns and homepage placements. Products that combine high purchase volume and strong customer ratings should receive premium visibility and inventory priority.

5. Targeted Marketing Campaigns

Allocate marketing resources toward:

- High-revenue age groups
- Express-shipping users (if correlated with higher average spend)
- High-frequency buyers likely to convert to subscribers
- Segmentation-driven marketing reduces acquisition costs and increases ROI compared to generic campaigns.

7. Conclusion

This project demonstrates a complete end-to-end analytics workflow:

- Data cleaning and preprocessing in Python
- Structured business analysis using PostgreSQL
- Insight visualization through Power BI
- Strategic recommendations derived from data

More importantly, the analysis moved beyond descriptive statistics and focused on actionable business decisions.