

Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior using transactional data to understand patterns in purchases, demographics, customer segments, shopping preferences, and subscription behavior to guide strategic business decisions.

2. Dataset Summary

- Rows : 3.900
- Columns : 18
- Key Features :
 - Customer demographic (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

Began with data preparation and data cleaning in python :

- **Data Loading** : imported the dataset using `pandas`
- **Initial Exploration** : using `df.info()` to check structure and `.describe()` to summary statistics

```
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)
```

	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
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3900.000000	3900
unique	Nan	Nan	2	25	4	Nan	50	4	25	4	Nan	2	6	2	2	Nan	6
top	Nan	Nan	Male	Blouse	Clothing	Nan	Montana	M	Olive	Spring	Nan	No	Free Shipping	No	No	Nan	PayPal
freq	Nan	Nan	2652	171	1737	Nan	96	1755	177	999	Nan	2847	675	2223	2223	Nan	677
mean	1950.500000	44.068462	Nan	Nan	Nan	59.764359	Nan	Nan	Nan	Nan	3.750065	Nan	Nan	Nan	Nan	25.351538	Nan
std	1125.977353	15.207589	Nan	Nan	Nan	23.685392	Nan	Nan	Nan	Nan	0.716983	Nan	Nan	Nan	Nan	14.447125	Nan
min	1.000000	18.000000	Nan	Nan	Nan	20.000000	Nan	Nan	Nan	Nan	2.500000	Nan	Nan	Nan	Nan	1.000000	Nan
25%	975.750000	31.000000	Nan	Nan	Nan	39.000000	Nan	Nan	Nan	Nan	3.100000	Nan	Nan	Nan	Nan	13.000000	Nan
50%	1950.500000	44.000000	Nan	Nan	Nan	60.000000	Nan	Nan	Nan	Nan	3.800000	Nan	Nan	Nan	Nan	25.000000	Nan
75%	2925.250000	57.000000	Nan	Nan	Nan	81.000000	Nan	Nan	Nan	Nan	4.400000	Nan	Nan	Nan	Nan	38.000000	Nan
max	3900.000000	70.000000	Nan	Nan	Nan	100.000000	Nan	Nan	Nan	Nan	5.000000	Nan	Nan	Nan	Nan	50.000000	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 data into the database.

4. Data Analysis using SQL

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
	text	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	purchase_amount
	bigint	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
11	24	88
12	29	94
13	32	79
14	33	67
15	35	91
16	37	69
17	40	60
18	41	76

Total rows: 839 | Query complete 00:00

3. **Top 5 Product by Rating** – Found product 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	round 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_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 Product** – Identified 5 product with highest percentage of discounted purchases.

	item_purchased text	discount_rate 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	category text	item_purchased text	total_order 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
11	2	Outerwear	Coat	161

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

	subscription_status text	repeat_buyer bigint
1	No	2518
2	Yes	958

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

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

5. Dashboard in Power BI

Built an interactive dashboard in **Power BI** to present insight 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 product in campaigns.
- **Targeted Marketing** – Focus efforts on high-revenue age groups and express-shipping users.