

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 subscriptions behavior to guide strategic business decisions.

2. Dataset Summary

- Rows: 3900
- Columns: 18
- Key Features:
 - Customers demographics (Age, Gender, Location, Subscription type)
 - Purchase details (Item purchased, Category, Purchase amount, Season, Size, Color)
 - Shopping behaviors (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 using python

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

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3863.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.750065	25.351538
std	1125.977353	15.207589	23.685392	0.716983	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.800000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

```

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)
memory usage: 548.6+ KB

```

- **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:** Rename columns to `snake case` for better readability and documentation.
- **Feature Engineering:** Created an `age_group` feature to categorize customers into `Children (10–20)`, `Young (21–30)`, `Mid-senior (31–50)`, and `Senior (51+)`, helping analyze spending, subscriptions, and ratings by age segment.
- **Age Distribution Insight:** Using the Shapiro-Wilk test, we found that the `age` distribution is not normal, meaning certain age groups are more common than others. This indicates that our customer base is concentrated within specific age ranges rather than evenly spread out.
- **Purchase Amount Distribution Insight:** The Shapiro-Wilk test shows that the `purchase amount` is not normally distributed. This means customer spending patterns are uneven. a small group of customers contributes a much higher amount compared to others.

- **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 and Average purchase amount by Gender:** Compared average and total revenue generated by male v/s female customers.

	gender text	total_spending numeric	avg_spending numeric
1	Male	157890	59.54
2	Female	75191	60.25

- 2. Customer Count by Age group:** Analyzed which age group has the highest number of customers to identify the most represented demographic segment.

	age_group text	total_customers bigint
1	Senior	1476
2	Min-senior	1475
3	Young	737
4	Children	212

- 3. Total Spending by Age group:** Analyzed which age group contributes the highest total spending to understand which segment drives the most revenue.

	age_group text	total_amount numeric
1	Senior	88480
2	Min-senior	87322
3	Young	44775
4	Children	12504

- 4. High-Spending Customers Using Discounts:** Identified the number of customers who used a discount but still spent more than the average purchase amount, highlighting value-driven yet high-spending customers.

	customers bigint
1	839

- 5. Category Performance Analysis:** Determined which product category has the highest total sales, best average rating, and the most customers to identify top-performing categories driving business growth.

	category text	total_amount numeric	average_rating numeric	total_customers bigint
1	Clothing	104264	3.72	1737
2	Accessories	74200	3.77	1240
3	Footwear	36093	3.79	599
4	Outerwear	18524	3.75	324

- 6. Top 5 Products by Average Rating:** Identified the five products with the highest average customer review ratings to highlight the best-performing items in terms of customer satisfaction.

	item_purchased text	average_rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.79

- 7. Top 5 Locations by Sales:** Identified the five locations with the highest total sales to understand key markets contributing the most to overall revenue.

	location text	total_sales numeric
1	Montana	5784
2	Illinois	5617
3	California	5605
4	Idaho	5587
5	Nevada	5514

8. Average Purchase Amount by Shipping: Compared the average purchase amount across different shipping types to evaluate which delivery option is associated with higher customer spending.

	shipping_type text	average_amount numeric
1	Standard	3.82
2	Express	3.77
3	2-Day Shipping	3.77
4	Next Day Air	3.72
5	Free Shipping	3.72
6	Store Pickup	3.71

9. Subscribers vs. Non-Subscribers: Compared average spend and total revenue across subscription status.

	subscription_status text	average_spend numeric	total_revenue numeric
1	No	59.87	170436
2	Yes	59.49	62645

10. Top 5 Products by Discount Usage: Identified the five products with the highest percentage of purchases made using discounts to understand which items attract customers through promotional offers.

	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

11. Customer Segmentation by Purchase History: Categorized customers into *New*, *Returning*, and *Loyal* segments based on their total number of previous purchases to analyze customer retention and loyalty patterns.

	customer_segment text	total_customers bigint
1	loyal	3116
2	returning	701
3	new	83

12. Top 3 Products per Category: Identified the three most purchased products within each category to determine the best-selling items driving sales in their respective segments.

	rn 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

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 buys to move them to the Loyal segment.
- **Review Discount Policy:** Balance sales boosts 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.