

# Customer Shopping Behavior Analysis

## 1. Project Overview

This project explores customer shopping patterns using transactional data from 3,900 purchases across multiple product categories. The main objective is to uncover insights into spending behavior, product preferences, seasonal trends, and subscription patterns that can help optimize marketing strategies and improve customer retention.

## 2. Dataset Summary

- Rows: 3,900
- Columns: 18

Key Features:

- Customer Demographics: Age, Gender, Location, Subscription Status
- Purchase Details: Item Purchased, Category, Purchase Amount (USD), Season, Size, Color
- Shopping Behavior: Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type

## 3. Exploratory Data Analysis using Python

Data preparation and cleaning were performed using Python (Pandas).

- Data Loading: Imported the dataset and verified data types and null values.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...
895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No
896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No
897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No
898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No
899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No

- Initial Exploration: Used `df.info()` to understand data structure and `df.describe()` for summary statistics.

```
<class 'pandas.core.frame.DataFrame'>
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                        3900 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
```

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.749949	25.351538
std	1125.977353	15.207589	23.685392	0.716223	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	3.700000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

- Feature Engineering :Created **age\_group** column by binning customer ages ,  
**purchase\_frequency\_days** column from purchase data.
- Database Integration :Connected python script to MS-SQL and loaded the DataFrame for SQL analysis.

#### 4. Data Analysis using SQL

Structured queries were executed in PostgreSQL to answer key business questions:

1. Revenue by Gender: Compared total revenue generated by male vs. female customers.

	gender	total_revenue
1	Male	157890
2	Female	75191

2. Top Product Categories: Identified categories contributing the most to overall sales.

	category	total_revenue	total_orders
1	Clothing	104264	1737
2	Accessories	74200	1240
3	Footwear	36093	599
4	Outerwear	18524	324

3. Which are the top 5 products with the highest average review rating?

	product	Average_Review_Rating
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.81
4	Hat	3.81
5	T-shirt	3.78

4. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers.

	subscription_status	total_customers	average_spend	total_revenue
1	Yes	1053	59	62645
2	No	2847	59	170436

5. Compare the average purchase amounts between standard and express shipping?

	shipping_type	avg_amount
1	Standard	58
2	Express	60

6. Which 5 products have the highest percentage of purchase with discounts applied?

	products	discount_rate
1	Hat	50
2	Coat	49
3	Sneakers	49
4	Sweater	48
5	Pants	47

7. What are the top 3 purchase products within each category?

	item_rnk	category	item_purchased	total_orders
1	1	Accessories	Jewelry	171
2	2	Accessories	Belt	161
3	3	Accessories	Sunglasses	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

8. Which segment customers into new, returning ,and loyal based on their total Numbers of previous purchase and show the count of each segment?

	customer_segments	total_cutomers
1	Returning	701
2	loyal	3116
3	New	83

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

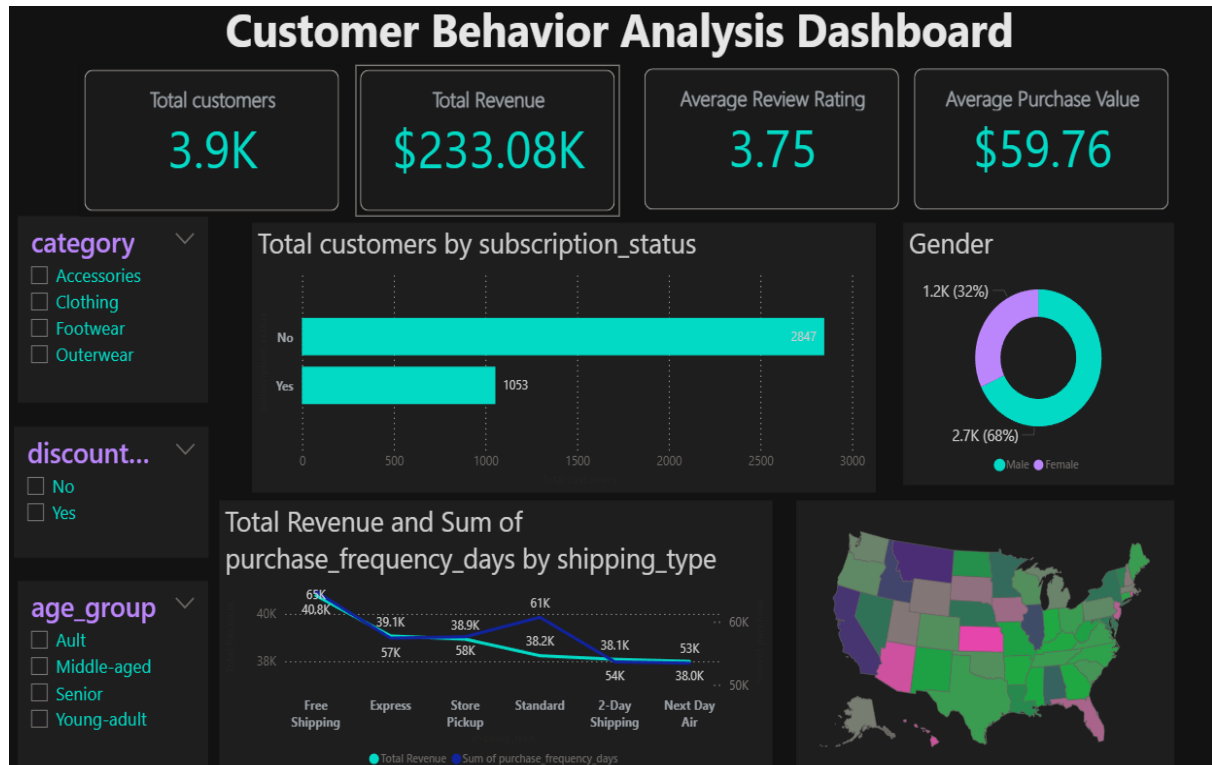
	age_group	revenue
1	Young-adult	62143
2	Middle-aged	59197
3	Ault	55978
4	Senior	55763

10. Seasonal Trends: Analyzed sales distribution across different seasons.

	season	total_orders	total_revenue	avg_purchase_value
1	Fall	975	60018	61
2	Spring	999	58679	58
3	Winter	971	58607	60
4	Summer	955	55777	58

## 5. Dashboard Visualization using Power BI

An interactive Power BI dashboard was designed to visualize:



## 6. Business Recommendations

Based on the analysis, the following strategies were proposed:

- **Boost Subscriptions:** Offer exclusive benefits and discounts to subscribers to increase retention.
- **Customer Loyalty Programs:** Reward repeat customers to convert them into loyal, long-term buyers.
- **Review Discount Policy:** Balance discount frequency to maintain both customer satisfaction and profit margins.
- **Product Positioning:** Highlight top-rated and high-revenue products in promotional campaigns.
- **Targeted Marketing:** Focus marketing on high-spending age groups and customers preferring express shipping.

## 7. Conclusion

The project successfully integrated Python, SQL, and Power BI to derive actionable insights from customer transaction data. These findings can guide data-driven marketing strategies, enhance personalization, and improve overall customer engagement.