Customer Shopping Behavior Analysis

Prepared by: Muddassir Mushtaq Mulla

---

1. Project Overview

This project analyzes customer shopping behavior using a dataset of **3,900 purchases**. The goal is to understand how customers shop, what products they prefer, which segments spend more, and how discounts or subscriptions influence buying decisions. The analysis was completed using **Python**, **PostgreSQL (SQL)**, and **Power BI** for visualization.

---

2. Dataset Summary

The dataset contains **18 columns** covering:

Customer Information

- Age, Gender, Location
- Subscription Status

Purchase Details

- Item Purchased, Category, Size, Color, Season
- Purchase Amount
- Shipping Type

Shopping Behavior

- Discount Applied, Promo Code Used
- Previous Purchases, Frequency of Purchases
- Review Rating

Data Quality

- **37 missing values** in Review Rating
- There was slight overlap between the discount and promo-related features.

---

3. Exploratory Data Analysis (Python)

Work in Python focused on cleaning, understanding, and preparing the data.

3.1 Data Loading and Inspection

- Loaded data using Pandas
- Checked structure using df.info()
- Generated summary statistics using df.describe()

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

[4]:

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

Previous Purchases	Payment Method	Frequency of Purchases
3900.000000	3900	3900
NaN	6	7
NaN	PayPal	Every 3 Months
NaN	677	584
25.351538	NaN	NaN
14.447125	NaN	NaN
1.000000	NaN	NaN
13.000000	NaN	NaN
25.000000	NaN	NaN
38.000000	NaN	NaN
50.000000	NaN	NaN

3.2 Handling Missing Values

- Filled 37 missing review ratings
- Used **median rating per product category** to impute values

3.3 Data Standardization

- Converted column names to **snake_case** for clarity
- Formatted categorical text for consistency

3.4 Feature Engineering

New features were added to make the analysis more meaningful:

- **Age Groups** (Teen, Young Adult, Adult, Senior)
- **Purchase Frequency** (days between purchases)
- **Removed Redundant Columns** by dropping promo_code_used

3.5 Database Loading

- Cleaned data was uploaded to **PostgreSQL** for SQL-based business analysis

---

4. Business Analysis with SQL

SQL queries were used to answer important business questions.

Key Insights Extracted via SQL

1. **Revenue by Gender** – Compared spending between male and female customers

	gender text	revenue numeric
1	Female	75191
2	Male	157890

2. **High-Spending Discount Users** – Identified customers who use discounts but still spend above average

	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

3. **Top 5 Products by Review Rating**

	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** – Standard vs. Express spending

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. **Subscribers vs. Non-Subscribers** – Revenue and average spend difference

	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** – Products frequently bought on discount

	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** – Classified customers as New, Returning, or Loyal

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

8. **Top 3 Products in Each Category** – Using SQL window functions

	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 and Subscription Likelihood**

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

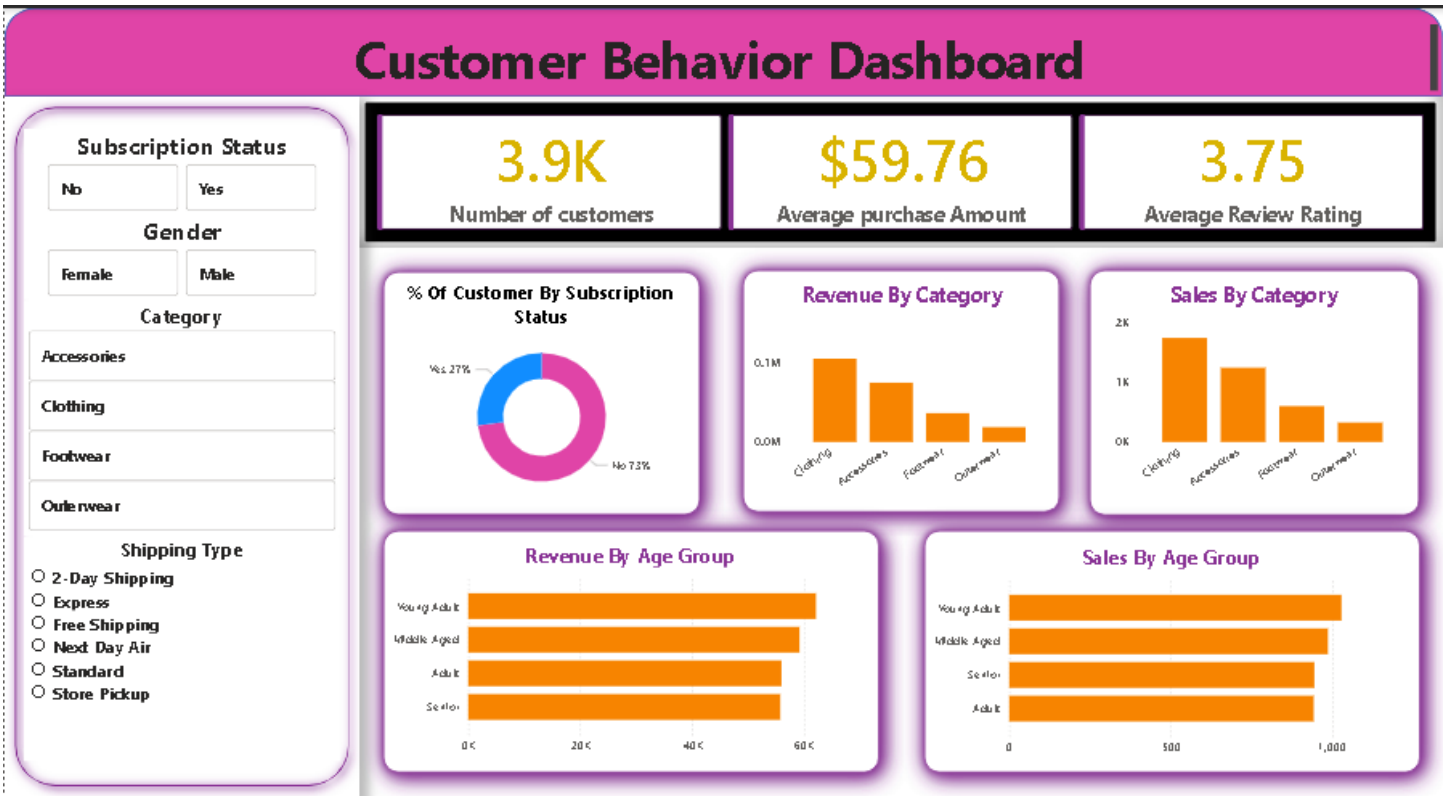
10. **Revenue by Age Group**

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

---

5. Power BI Dashboard

A multi-page Power BI dashboard was created to visually present the insights.



Dashboard Sections

- **Sales Overview:** Revenue, units sold, and top categories
- **Customer Demographics:** Age and gender patterns
- **Subscription Analysis:** Revenue uplift from subscribers
- **Shipping Behavior:** Spend comparison by shipping type
- **Discount Impact:** Influence of promotions on purchases

---

6. Business Recommendations

Based on the combined Python, SQL, and Power BI analysis, the following actions are recommended:

1. Increase Subscriptions

- Offer exclusive deals and free express shipping to encourage customers to subscribe

2. Build a Loyalty Program

- Provide points, cashback, or rewards to repeat buyers
- Target customers with more than 5 purchases

3. Optimize Discount Strategy

- Review products that rely heavily on discounts
- Consider value-based pricing or product bundles

4. Promote Top-Rated and Best-Selling Items

- Highlight these products on the homepage, ads, and emails

5. Target High-Revenue Age Groups

- Focus marketing efforts on the age groups contributing most to sales

6. Improve Shipping Experience

- Offer express shipping benefits for premium customers

---

7. Conclusion

This project provides a clear understanding of customer behavior, spending patterns, and product performance. By applying the recommended strategies, the business can increase revenue, improve customer satisfaction, and strengthen customer loyalty.

Prepared by: Muddassir Mushtaq Mulla