Customer Shopping Behavior Analysis

1. Project Overview

This project studies how customers shop by analyzing data from 3,900 purchase transactions across different product categories. The main goal is to understand how customers spend money, how they differ from each other, what products they prefer, and how subscriptions affect buying behavior. These insights help businesses make better decisions.

---

2. Dataset Summary

- **Total Records:** 3,900
- **Total Columns:** 18
- **Main Information Includes:**
    - Customer details such as age, gender, location, and subscription status
    - Purchase information like product name, category, price, season, size, and color
    - Shopping behavior details including discounts, promo codes, past purchases, buying frequency, product reviews, and shipping method
- **Missing Data:**
    - 37 missing values in the review rating column

---

3. Exploratory Data Analysis (Using Python)

The analysis started by cleaning and preparing the data in Python to make it ready for deeper analysis.

- Data Loading: Imported the dataset using pandas.

```
# Importing the Required Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Connecting Google Drive with colab.
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
# Loading the dataset using pandas
df = pd.read_csv("/content/drive/MyDrive/Data_Analyst/Customer_Shopping_Behavior_Analysis_Project/customer_shopping_behavior.csv")
```

```
df.head()
```

	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
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes

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

```
# Summary statistics using .describe()
df.describe()
```

	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

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 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 by Gender – Compared total revenue generated by male vs. female customers.

	gender character varying (10)	revenue bigint
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 [PK] integer	purchase_amount integer
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

3. Top 5 Products by Rating – Found products with the highest average review ratings.

	item_purchased character varying (50)	average_product_rating numeric
1	Gloves	3.9
2	Hat	3.8
3	Jacket	3.8
4	Sneakers	3.8
5	Socks	3.8

4. Shipping Type Comparison – Compared average purchase amounts between Standard and Express shipping.

	shipping_type character varying (20)	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 character varying (10)	total_customers bigint	average_spend numeric	total_revenue numeric
1	Yes	1053	59.49	62645.00
2	No	2847	59.87	170436.00

6. Discount-Dependent Products – Identified 5 products with the highest percentage of discounted purchases.

	age_group text	total_sale bigint	contribution_percentage text
1	Mid-Age Adul...	68066	29.20%
2	Adults	65216	27.98%
3	Young Adults	52905	22.70%
4	Seniors	46894	20.12%

7. Customer Segmentation – Classified customers into New, Returning, and Loyal segments based on purchase history.

	customer_segment text	total bigint
1	New Customer	83
2	Loyal Customer	3116
3	Returning Customer	701

8. Top 3 Products per Category – Listed the most purchased products within each category.

	item_rank bigint	category character varying (50)	item_purchased character varying (50)	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 – Checked whether customers with >5 purchases are more likely to subscribe.

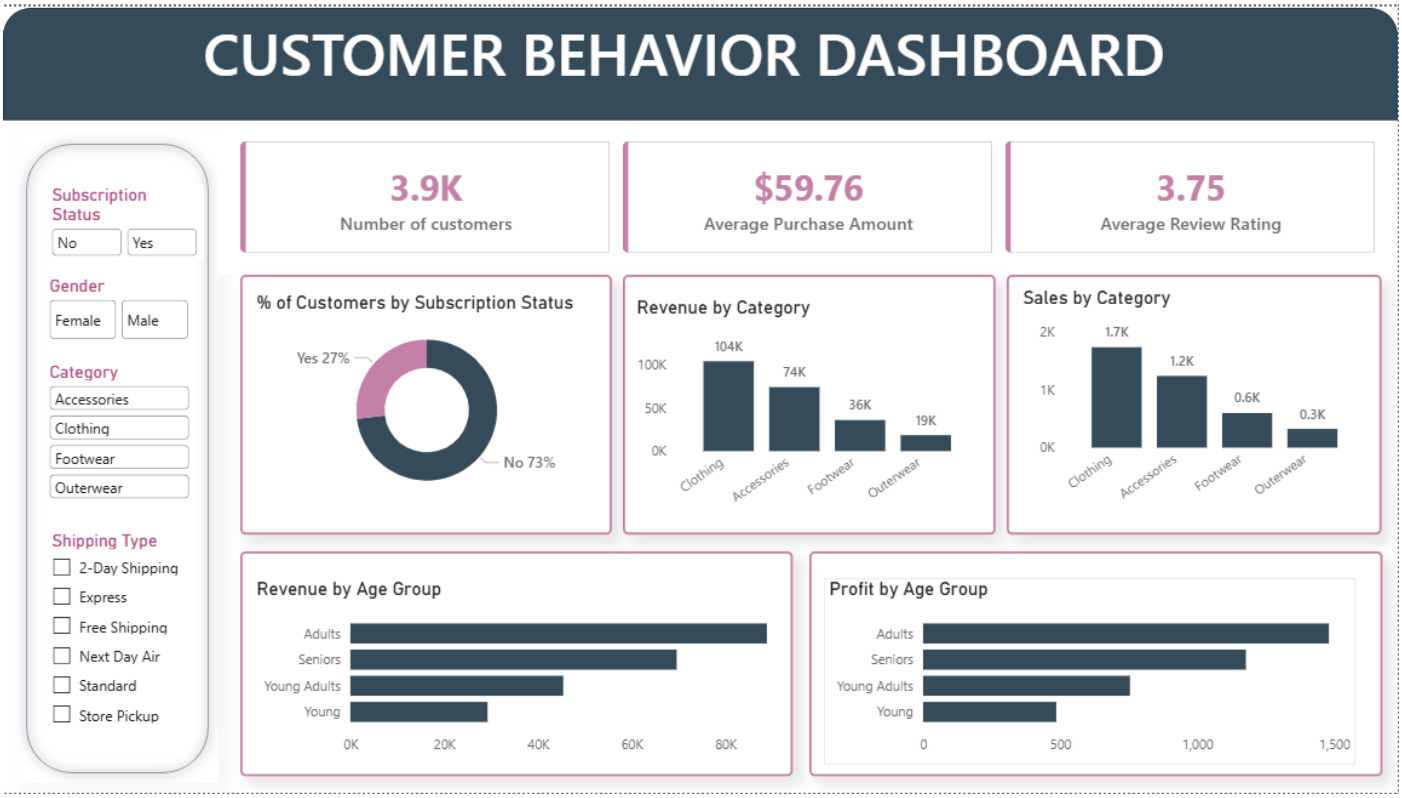
	subscription_status character varying (10)	total_customer bigint
1	No	2518
2	Yes	958

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

	age_group text	revenue bigint	contribution_percentage text
1	Mid-Age Adults	68066	29.20%
2	Adults	65216	27.98%
3	Young Adults	52905	22.70%
4	Seniors	46894	20.12%

5. **Dashboard in Power BI**

Finally, we built an interactive dashboard in Power BI to present insights visually.



6. Business Recommendations

- Offer special benefits and rewards to encourage more customers to subscribe.
- Give rewards to repeat customers so they become loyal buyers.
- Use discounts to increase sales, but make sure profits are not reduced too much.
- Focus marketing on best-selling and highly rated products.
- Spend more marketing effort on age groups that bring in the most revenue and customers who prefer fast shipping.
- Offer discounts to senior customers, as they are more likely to buy when prices are lower.
- To grow steadily, focus on adult customers, improve online sales, offer smart discounts, encourage subscriptions, and collect better product reviews.