Customer Shopping Behavior Analysis

1. Project Overview

In this project, I analyzed customer purchasing behavior using transactional data from approximately 3,900 orders across multiple product categories. My goal was to identify spending patterns, customer segments, product preferences, and subscription-related trends that can support informed business decisions.

2. Dataset Description

The dataset contains 3,900 rows and 18 columns and includes the following types of information:
Customer data: age, gender, location, subscription status
Purchase details: item purchased, product category, price, season, size, and color
Behavioral metrics: discount usage, promo codes, number of previous purchases, purchase frequency, review ratings, and shipping type
During data inspection, I identified 37 missing values in the review rating column, which required preprocessing.

3. Data Cleaning and Exploratory Analysis (Python)

I performed data preparation and exploratory analysis using Python:

- Loaded and explored the dataset using pandas
- Analyzed dataset structure and summary statistics with .info() and .describe()
- Handled missing review ratings by imputing the median rating within each product category
- Standardized column names to snake_case to improve readability and consistency
- Engineered new features:
    - Created age groups to simplify demographic analysis
    - Derived purchase frequency metrics from historical purchase data
- Checked discount-related fields for redundancy and removed the promo code column as it duplicated discount logic
- Exported the cleaned dataset to PostgreSQL for further analysis using SQL

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discou Appli
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	39
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	22
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	Na
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	Na
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	Na
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	Na
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	Na
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	Na
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	Na

Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3900	3900	3900.000000	3900	3900
2	2	NaN	6	7
No	No	NaN	PayPal	Every 3 Months
2223	2223	NaN	677	584
NaN	NaN	25.351538	NaN	NaN
NaN	NaN	14.447125	NaN	NaN
NaN	NaN	1.000000	NaN	NaN
NaN	NaN	13.000000	NaN	NaN
NaN	NaN	25.000000	NaN	NaN
NaN	NaN	38.000000	NaN	NaN
NaN	NaN	50.000000	NaN	NaN

4. Business Analysis Using SQL

Using PostgreSQL, I wrote SQL queries to answer key business questions, including:

1. **Comparing total revenue by gender**.

	gender text	revenue numeric
1	Female	75191
2	Male	157890

2. **Identifying customers who used discounts but still spent above the average purchase amount**.

	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
11	24	89

Total rows: 839 Query complete 00:00:

3. **Finding products with the highest average customer 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. **Comparing average order values between standard and express shipping.**

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. **Analyzing spending behavior of subscribers versus non-subscribers.**

	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. **Detecting products most dependent on discounts.**

	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. **Segmenting customers into new, returning, and loyal groups based on purchase history.**

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

8. **Identifying the top three best-selling products within each category.**

	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. **Evaluating whether frequent buyers are more likely to subscribe**.

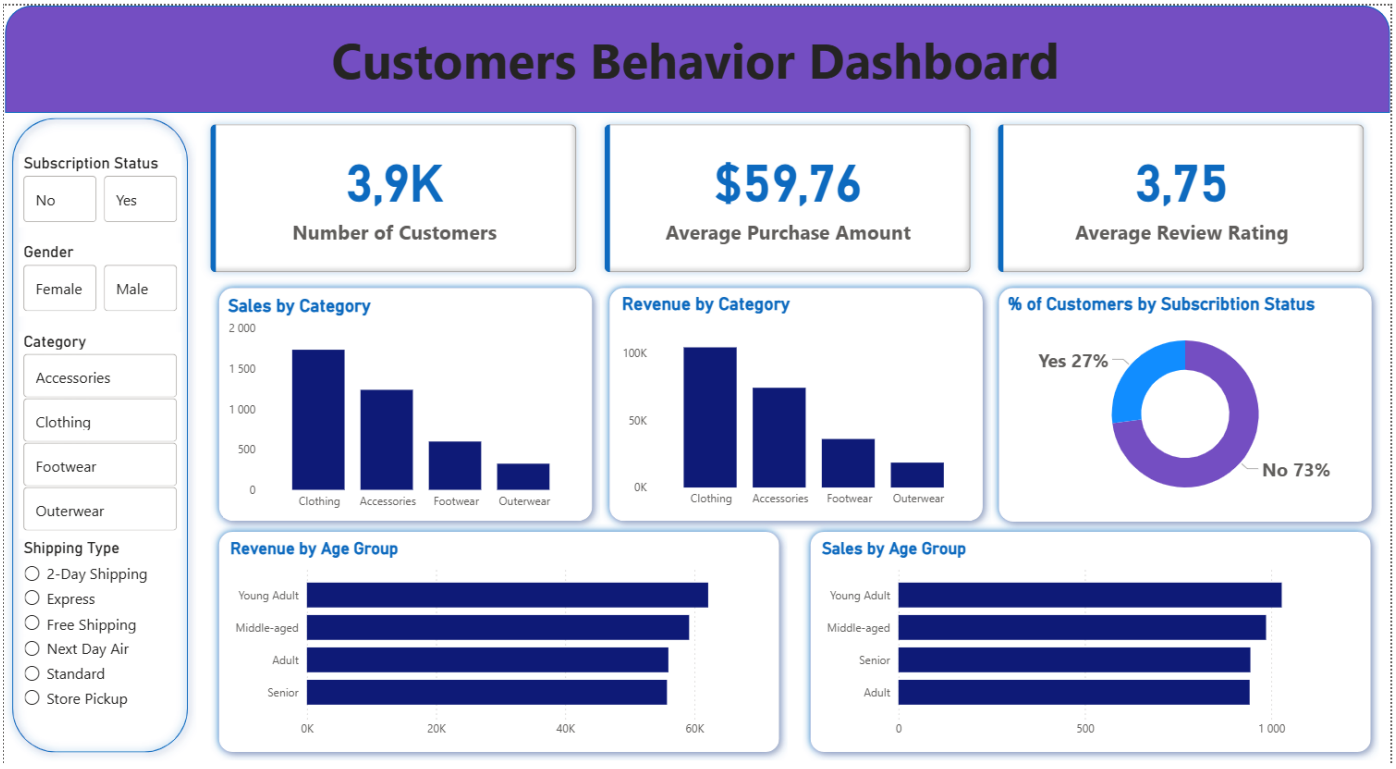
	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

10. **Calculating revenue contribution by age group.**

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

5. Data Visualization (Power BI)

I designed an interactive **Power BI dashboard** to visualize key insights from the analysis.
The dashboard highlights customer distribution, subscription impact, revenue by category and age group, and overall purchasing trends, making the results easy to interpret for non-technical stakeholders.



6. Key Insights and Recommendations

- **Based on the analysis, I formulated several business recommendations:**
- **Increase subscription adoption by promoting exclusive subscriber benefits**

- **Strengthen loyalty programs to convert repeat buyers into loyal customers**

- **Optimize discount usage to drive sales without reducing margins**
- **Emphasize top-rated and best-selling products in marketing campaigns**

- **Focus targeted marketing on high-revenue age groups and customers who prefer express shipping**