Customer Shopping Analysis

1. Project Overview

The following project analyzes the customer shopping behavior using transactional data from 3.900 users across various categories of products. The goal of this project is to uncover insights regarding spending patterns, consumer age groups, product preferences and subscription behavior of consumers to guide strategic business decisions.

2. Dataset Summary

- Rows: 3,900
- Columns: 18
- Key Features:
  - Customer Demographic (Age, Gender, Location, Subscription Status)
  - Purchase details (Items Purchased, Category, Purchase Amount, Season, Size, Color)
  - Shopping Behavior (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)
  - Missing Data: 37 Review Ratings

3. Exploratory Data Analysis using Python

We began the project by preparing and cleaning the data in Python as such:

- **Loading Data:** Database was imported to the spyder IDE using pandas.
- **Initial Exploration:** df.info() was used to check the structure and df.describe() was used to check for summary statistics.

	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	N
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	N
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	N
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	N
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	N
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	N
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	N

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

- **Missing Data:** After checking for null values, it was observed that there were 37 different entries within the review rating column that were vacant. To ensure acceptable data without any biases were used for reliable extrapolation of data, the median was taken for each category separately to evaluate the missing entries.
- **Column Header Standardization:** All column headers were converted to snake casing to ensure consistency for analyzing and documentation.
- **Feature Engineering:**
    - A new column **age_group** was created by partitioning and filtering customer ages into sub-categories.
    - Using the pre-existing purchase frequency, a new column **purchase_frequency_days** was created for ease of analysis.

- **Data consistency check:** Verified if **discount_applied** and **promo_code_used** columns were dependent; Since a discount was only applied when a promo code was used, the **promo_code_used** column was dropped.
- **Database Integration:** Connected the python script to PostgreSQL to load the cleaned DataFrame into the database for SQL analysis.

4. Data analysis in SQL

The following key business questions were answered by doing a structured data analysis in PostgreSQL:

1. What is the total revenue generated by male vs. female customers?

	gender text	revenue numeric
1	Female	75191
2	Male	157890

2. Which customers used a discount but still spent more than 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

3. Which are the top 5 products with the highest average 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. Compare the average Purchase Amounts between Standard and Express Shipping.

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and 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. Which 5 products have the highest percentage of purchases with discounts applied?

	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. Segment customers into New, Returning, and Loyal based on their total number of previous purchases, and show the count of each segment.

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

8. What are the top 3 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. Are customers who are repeat buyers (more than 5 previous purchases) also likely to subscribe?

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

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

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

5. PowerBI Dashboard

Lastly, we built an interactive dashboard in **Power BI** to present our evaluations visually.

6. Business Recommendations

- **Boost Subscriptions:** Increasing exclusive benefits catered solely towards subscribers,
- **Review Discount Policy:** Discount margin controls can be balanced with sales boosts caused by them.
- **Targeted Marketing:** Prioritize efforts on high revenue age groups and express shipping,
- **Relevant positioning:** Highlight top-rated and best-selling products in marketing campaigns.