Customer Shopping Behaviour Analysis

1. Project Overview

This project analyses customer shopping behaviour using 3900 purchases across various product categories. The goal of this project is to improve sales, customer satisfaction, and long-term loyalty by uncovering factors such as discounts, reviews, seasons, or payment preferences and repeat purchases.

2. Dataset Summary

Rows: 3,900

Columns: 18

Key Features:

- Customer demographics (Age, Gender, Location, Subscription Status)
- Purchase details (Item Purchased, Category, Purchase Amount, Season, Size, Colour)
- Shopping behaviour (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)

Missing Data: 37 values in Review Rating column.

3. Exploratory Data Analysis using Python

Data preparation and cleaning is done using Python (Pandas):

- Data Loading: Imported the dataset using pandas.
- Exploration: Used df.info() to check structure and df.describe(include='all') for summary statistics.

	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

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

- Missing Data Handling: Checked for null values. Review Rating column has 37 null values. Imputed those null values with median rating of respective product category.
- Column Name Standardization: Renamed column names to snake case for better readability and documentation.
- Feature Engineering:
  1. Created age_group column by binning customer ages.
  2. Created purchase_frequency_days column from purchase data.
- Data Consistency Check: Verified if discount_applied and promo_code_used were redundant, they were same so dropped promo_code_used.
- Data Integration: Connected Python script to PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

4. Data Analysis using SQL (Business Transactions)

Performed structured analysis in PostgreSQL to answer key business questions:

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

	gender text	revenue numeric
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 bigint	discount_applied text	purchase_amount bigint
1	2	Yes	64
2	3	Yes	73
3	4	Yes	90
4	7	Yes	85
5	9	Yes	97
6	12	Yes	68
7	13	Yes	72
8	16	Yes	81
9	20	Yes	90
10	22	Yes	62
11	24	Yes	88

Total rows: 839 Query complete 00:00:00.402

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

	item_purchased text	avg_review 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 – Compared average purchase amounts between Standard and Express shipping.

	shipping_type text	avg_purchase_amount 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 text	count 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.

	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.00
3	Coat	49.00
4	Sweater	48.00
5	Pants	47.00

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

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

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

	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	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

Total rows: 11 Query complete 00:00:00.556

9. Repeat Buyers & Subscriptions – Checked whether customers with >5 purchases are more likely to subscribe.

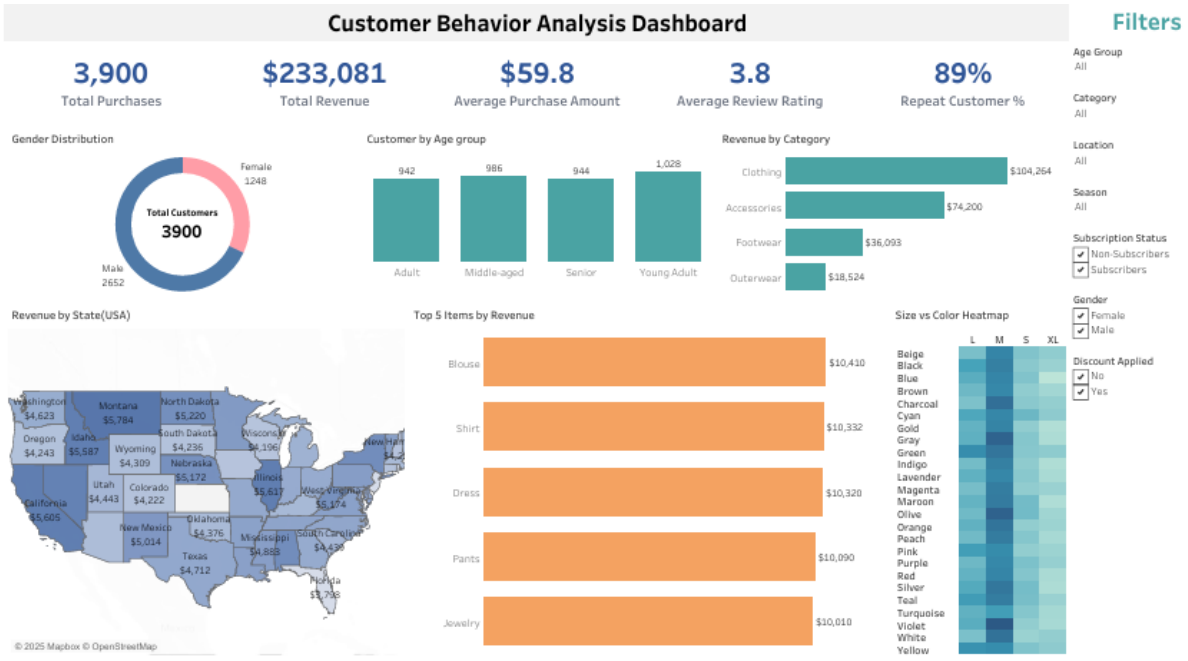
	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

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

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

5. Dashboard in Tableau

At last built an interactive dashboard in Tableau



6. Business Recommendations

- Boost Subscriptions - Promote exclusive benefits for subscribers.
- Customer Loyalty Programs - Reward repeat buyers to move them into the “Loyal” segment.
- Review Discount Policy - Balance sales boosts with margin control.
- Product Positioning - Highlight top-rated and best-selling products in campaigns.
- Targeted Marketing - Focus efforts on high-revenue age groups and express-shipping users.