Customer Shopping Behaviour Analysis

Business Problem Statement

"How can the company leverage consumer shopping data to identify trends, improve customer engagement, and optimize marketing and product strategies?"

. Project Overview

This project analyses customer shopping behaviour using transactional data from 3,900 purchases across various product categories. The goal is to uncover insights into spending patterns, customer segments, product preferences, and subscription behaviour to guide strategic business decisions.

. 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

. Exploratory Data Analysis using Python

Data Loading: Imported the dataset using pandas.

Initial Exploration: Used df.info() to check structure and .describe() 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

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 Data Frame into the database for SQL analysis.

Data Analysis using SQL (Business Transactions)

We performed structured analysis in MySQL 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

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

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

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

Subscribers vs. Non-Subscribers – Compared average spend and total revenue across subscription status.

	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

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	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

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

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

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	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

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

Revenue by Age Group – Calculated total 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

Dashboard in Power BI

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

. Business Recommendations

Strategic Suggestions for Improvement

- **Subscription Growth:** Develop a "Subscriber-Only" discount model targeting the 73% non-subscriber base to increase recurring revenue.
- **Average Order Value (AOV) Boost:** Since the average purchase is ~$60, implement a "Free Shipping" threshold at $75 to encourage larger cart sizes.
- **Category Diversification:** Investigate why **Outerwear** and **Footwear** are lagging behind **Clothing**; consider seasonal marketing campaigns to balance revenue across categories.

- **Retention for Young Adults:** Create loyalty programs specifically for the "Young Adult" segment, as they are currently your most active buyers.
- **Quality Control:** Analyse reviews specifically for products with ratings below 3.0 to identify if the issue lies in product quality or shipping delays.