CUSTOMER SHOPPING BEHAVIOR ANALYSIS

Python + SQL + Power BI

By Riya Bhatt

1. PROJECT OVERVIEW

This project analyzes customer shopping behavior 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 behavior to guide strategic business decisions.

BUSINESS PROBLEM STATEMENT

A leading retail company wants to better understand its customers' shopping behavior in order to improve sales, customer satisfaction, and long-term loyalty. The management team has noticed changes in purchasing patterns across demographics, product categories, and sales channels (online vs. offline). They are particularly interested in uncovering which factors, such as discounts, reviews, seasons, or payment preferences, drive consumer decisions and repeat purchases.

You are tasked with analyzing the company's consumer behavior dataset to answer the following overarching business question: "How can the company leverage consumer shopping data to identify trends, improve customer engagement, and optimize marketing and product strategies?"

DELIVERABLES

1. Data Preparation & Modeling (Python): Clean and transform the raw dataset for analysis.

2. Data Analysis (SQL): Organize the data into a structured format, simulate business transactions, and run queries to extract insights on customer segments, loyalty, and purchase drivers.

3. Visualization & Insights (Power BI): Build an interactive dashboard that highlights key patterns and trends, enabling stakeholders to make data-driven decisions.

4. Report and Presentation: Write a clear project report summarizing your key findings and business recommendations. Prepare a presentation that visually communicates insights and actionable recommendations to stakeholders.

5. GitHub Repository: Include all Python scripts, SQL queries, and dashboard files in a well-structured repository.

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, Color)
- Shopping behavior (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

We began with data preparation and cleaning in Python:

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

In [3]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Customer ID             3900 non-null   int64  
 1   Age                     3900 non-null   int64  
 2   Gender                  3900 non-null   object 
 3   Item Purchased          3900 non-null   object 
 4   Category                3900 non-null   object 
 5   Purchase Amount (USD)   3900 non-null   int64  
 6   Location                3900 non-null   object 
 7   Size                    3900 non-null   object 
 8   Color                   3900 non-null   object 
 9   Season                  3900 non-null   object 
 10  Review Rating           3863 non-null   float64
 11  Subscription Status     3900 non-null   object 
 12  Shipping Type           3900 non-null   object 
 13  Discount Applied        3900 non-null   object 
 14  Promo Code Used         3900 non-null   object 
 15  Previous Purchases      3900 non-null   int64  
 16  Payment Method          3900 non-null   object 
 17  Frequency of Purchases  3900 non-null   object 
dtypes: float64(1), int64(4), object(13)
```

```
# Summary statistics using .describe()
df.describe(include='all')
```

	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	3
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	

- **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.

DATA ANALYSIS USING SQL (BUSINESS TRANSACTIONS)

We performed structured analysis in PostgreSQL to answer key business

Q1. Revenue by Gender – Compared total revenue generated by male vs. female customers.

	gender	revenue
▶	Male	157890
	Female	75191

Q2. High-Spending Discount Users – Identified 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	99

Total rows: 839 | Query complete 00:00

Q3. Top 5 Products by Rating – Found products with the highest average review ratings.

item_purchased	Average_Product_Rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.80
Skirt	3.78

Q4. Compare the average Purchase Amounts between Standard and Express Shipping.

shipping_type	avg(purchase_amount)
Express	60.4752
Standard	58.4602

Q5. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers.

	subscription_status	total_customer	avg_spend	total_spend
▶	Yes	1053	59.4919	62645
	No	2847	59.8651	170436

Q6. Which 5 products have the highest percentage of purchases with discounts applied?

item_purchased	discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

Q7. Segment customers into New, Returning, and Loyal based on their total number of previous purchases, and show the count of each segment.

customer_segment	Number of Customers
Loyal	3116
Returning	701
New	83

Q8. What are the top 3 most purchased products within each category?

item_rank	category	item_purchased	total_orders
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
1	Clothing	Blouse	171
2	Clothing	Pants	171
3	Clothing	Shirt	169
1	Footwear	Sandals	160
2	Footwear	Shoes	150
3	Footwear	Sneakers	145
1	Outerwear	Jacket	163
2	Outerwear	Coat	161

Q9. Are customers who are repeat buyers (more than 5 previous purchases) also likely to subscribe?

	subscription_status	repeat_buyers
▶	Yes	958
	No	2518

Q10. What is the revenue contribution of each age group?

age_group	total_revenue
Young Adult	62143
Middle-Age	59197
Adult	55978
Senior	55763

DASHBOARD IN POWER BI

Customer Behaviour Analysis

Subscription Status

No | Yes

Subscription Status

Female | Male

Category

Accessories

Clothing

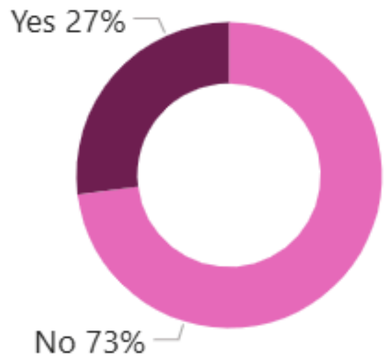
Footwear

3,900
No of customers

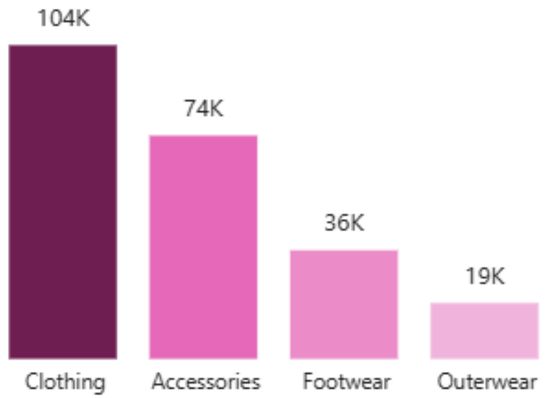
$59.76
Avg purchase amount

3.75
Avg review rating

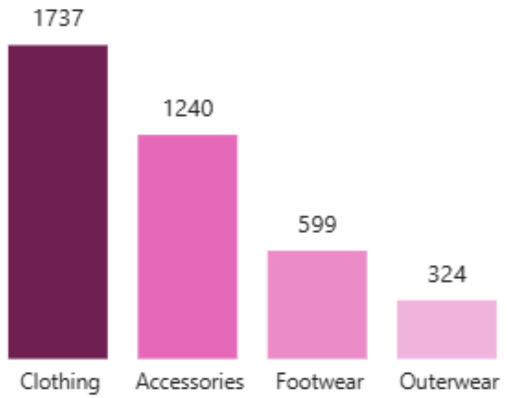
Customers with Subscription



Revenue by Category



Sales by Category



Revenue by Age Group

Sales by Age Group

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.

THANK YOU