Customer Shopping Behavior Analysis Report

1. Project Overview

This project focuses on analyzing customer shopping behavior using transactional data from 3900 purchases to understand purchasing patterns, spending trends, customer preferences, and engagement behavior. The objective is to extract actionable business insights using a combination of **Python for data preparation**, **SQL for structured analysis**, and **Power BI for visualization**.

By integrating data across tools, this project delivers a complete end-to-end analytics workflow — from raw dataset to business-ready insights.

2. Dataset Summary

- Rows: 3,900
- Columns: 18
- Customer demographics (age, gender, location, subscription status)
- Purchase behavior (items bought, categories, purchase amount, season, size, color)
- Shopping Behavior (discount applied, promo code used, purchase frequency, previous purchases, Review rating, shipping type)
- Missing data: 37 values in Review rating column

3. Exploratory Data Analysis Using Python

Python was used as the primary tool for cleaning, transforming, and preparing the data before sending it to the database.

Data Loading

The dataset was imported into Python using Pandas to enable structured exploration and transformation.

Initial Exploration

Used df.info() to check structure and df.describe() for summary statistics. This step helped identify inconsistencies and missing values early in the process.

```
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)
memory usage: 548.6+ KB
```

```
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
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

Missing values in the **review rating** column were handled using category-level median imputation. This ensured that product-level rating patterns were preserved rather than introducing bias using a global average.

Column Standardization

All column names were standardized using **snake_case formatting**. This improved readability and ensured compatibility across SQL, Power BI, and Python workflows.

Feature Engineering

New features were created to enhance analytical depth:

- **Age group segmentation** to analyze revenue across lifecycle stages
- **Purchase frequency indicators** to measure customer engagement
  These engineered features allowed more meaningful customer segmentation and trend analysis.

Data Consistency Check

Redundant columns related to discount usage were identified and validated for consistency. Non-essential duplicates like promo_code_used were removed to improve data quality.

Database Integration

The cleaned and transformed dataset was successfully connected to **PostgreSQL** using Python, enabling further analysis using SQL queries and linking the dataset with Power BI.

4. Data Analysis Using SQL

Structured business questions were answered using SQL inside PostgreSQL.

Revenue by Gender

Total revenue was aggregated across gender groups to understand spending contribution. This analysis highlighted which gender segment contributed more to overall business revenue.

Male Customers: ₹157,890
Female Customers: ₹75,191

	gender text	revenue numeric
1	Female	75191
2	Male	157890

High-Spending Discount Users

Customers who used discounts and still spent above the average purchase amount were identified. This group represents high-value customers who respond positively to promotional offers.

	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	88
12	29	94
13	32	79

Top 5 Products by Rating

Products were ranked based on average review ratings, revealing the most highly rated items in the catalog.

	item_purchased text	avg_review_rating double precision
1	Gloves	3.8614285714285725
2	Sandals	3.8443750000000003
3	Boots	3.8187500000000005
4	Hat	3.8012987012987005
5	Skirt	3.784810126582278

Shipping Type Comparison

Average purchase value was compared between Standard and Express shipping users to determine whether faster delivery influences higher spending.

	shipping_type text	avg_purchase_amount numeric
1	Express	60.4752321981424149
2	Next Day Air	58.6311728395061728
3	Store Pickup	59.8938461538461538
4	2-Day Shipping	60.7336523125996810
5	Free Shipping	60.4103703703703704
6	Standard	58.4602446483180428

Subscribers and Non-Subscribers

Both average spend and total revenue were compared across subscription groups, revealing differences in long-term value.

	subscription_status text	avg_spend numeric	total_revenue numeric
1	No	59.8651211801896733	170436
2	Yes	59.4919278252611586	62645

Discount-Dependent Products

Products with the highest proportion of discounted purchases were identified. These items show strong dependence on promotional pricing.

	item_purchased text	discount_percentage numeric
1	Hat	50.0000000000000000
2	Sneakers	49.6551724137931034
3	Coat	49.0683229813664596
4	Sweater	48.1707317073170732
5	Pants	47.3684210526315789

Customer Segmentation

Customers were classified into **New**, **Returning**, and **Loyal** segments based on previous purchase behavior. This segmentation enables targeted engagement strategies.

	customer_segment text	customer_count bigint
1	Loyal	3476
2	New	83
3	Returning	341

Top 3 Products per Category

For each category, the three most purchased products were extracted to identify top performers across product lines.

	category text	item_purchased text	total_purchases bigint
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
4	Clothing	Blouse	171
5	Clothing	Pants	171
6	Clothing	Shirt	169
7	Footwear	Sandals	160
8	Footwear	Shoes	150
9	Footwear	Sneakers	145
10	Outerwear	Jacket	163
11	Outerwear	Coat	161

Repeat Buyers & Subscriptions

The relationship between repeat purchase behavior and subscription status was analyzed to check whether loyal buyers are more likely to subscribe.

	subscription_status text	customer_count bigint
1	No	2518
2	Yes	958

Revenue by Age Group

Revenue contribution was analyzed across age groups to identify the most profitable demographic segments.

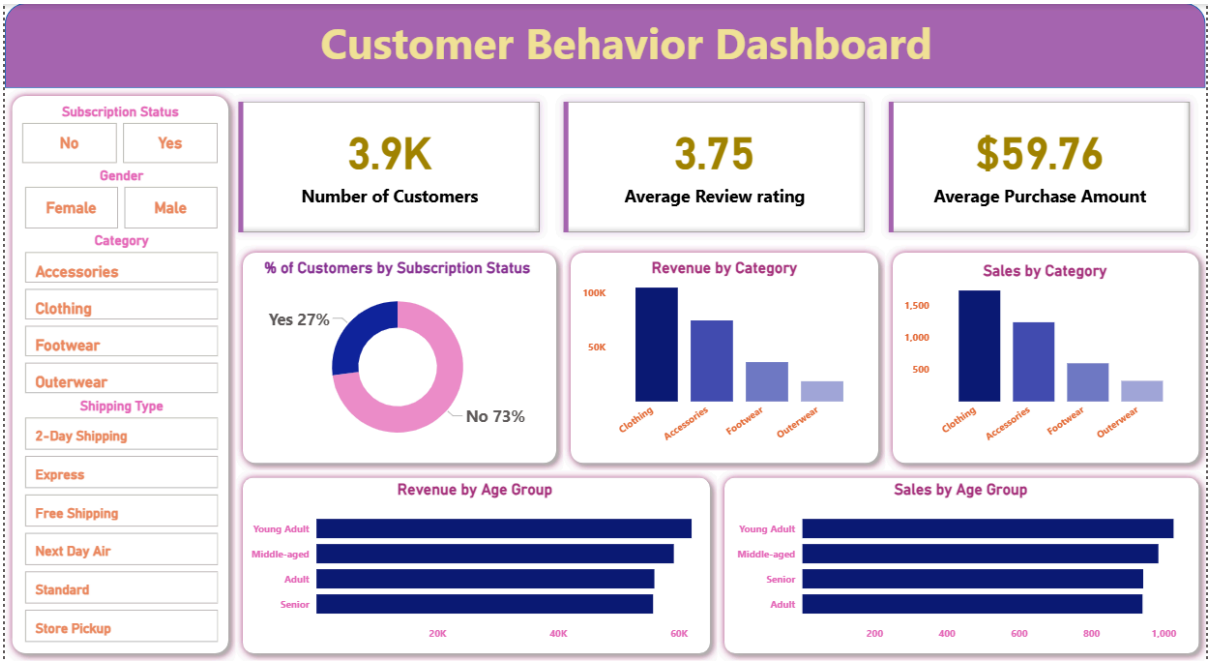
	age_group text	total_revenue numeric
1	55+	69590
2	25-34	45400
3	45-54	45370
4	35-44	43463
5	Under 25	29258

5. Dashboard in Power BI

A dynamic **Power BI dashboard** was created to visualize key metrics and trends. The dashboard includes:

- Total customers, average rating, and average purchase value
- Revenue and sales distribution by category
- Subscription status breakdown
- Interactive filters for slicing by gender, location, shipping type, and age group

This dashboard enables both strategic and operational decision-making through real-time data exploration.



6. Business Recommendations

Boost Subscriptions

Introduce exclusive benefits such as priority shipping, member-only discounts, and early product access to encourage more users to convert into subscribers.

Customer Loyalty Programs

Reward repeat buyers with loyalty points, cashback offers, and milestone rewards to increase retention and long-term revenue.

Review Discount Policy

Use discounts strategically on products that respond well to promotions while protecting margins on high-demand items.

Product Positioning

Promote both top-rated and best-selling products prominently to maximize conversions and brand trust.

Targeted Marketing

Focus marketing campaigns on **high-revenue age groups**, loyal customers, and express-shipping users who show higher spending potential.