

# Customer Shopping Behavior Analysis Report

## 1. Project Overview

This project examines customer purchasing behavior by utilizing transactional data derived from 3,900 purchases spanning multiple product categories. The objective is to reveal insights regarding spending trends, customer demographics, product choices, and subscription habits to inform strategic business decisions.

## 2. Dataset Summary

- Total Rows: 3,900
- Total 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 entries in the Review Rating column

## 3. Exploratory Data Analysis using Python

We commenced with the preparation and cleaning of data 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	Discount Applied
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	39
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	1
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	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN

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 Handling:** Verified the presence of null values and filled in the missing values in the **Review Rating** column by utilizing the median rating for 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.

## 4. Data Analysis using SQL (Business Transactions)

We performed structured analysis in PostgreSQL to answer key business questions:

1. **Revenue by Gender** – Analyzed the total revenue produced by male customers in comparison to female customer

	gender text	revenue numeric
1	Female	75191
2	Male	157890

2. **High-Spending Discount Users** – Identified customers who utilized discounts yet still exceeded 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. **Top 5 Products by Rating** – Identified products that possess 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

4. **Shipping Type Comparison** – Analyzed the average purchase amounts for Standard and Express shipping.

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. **Subscribers vs. Non-Subscribers** – Analyzed the average expenditure and overall revenue based on 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

6. **Discount-Dependent Products** – Five products with the highest percentage of discounted purchases have been identified.

	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. **Customer Segmentation** – Categorized customers into New, Returning, and Loyal segments according to their purchase history..

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

8. **Top 3 Products per Category** – Identified the top-selling items in 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. **Repeat Buyers & Subscriptions** – Verified if customers with more than five purchases are more inclined to subscribe.

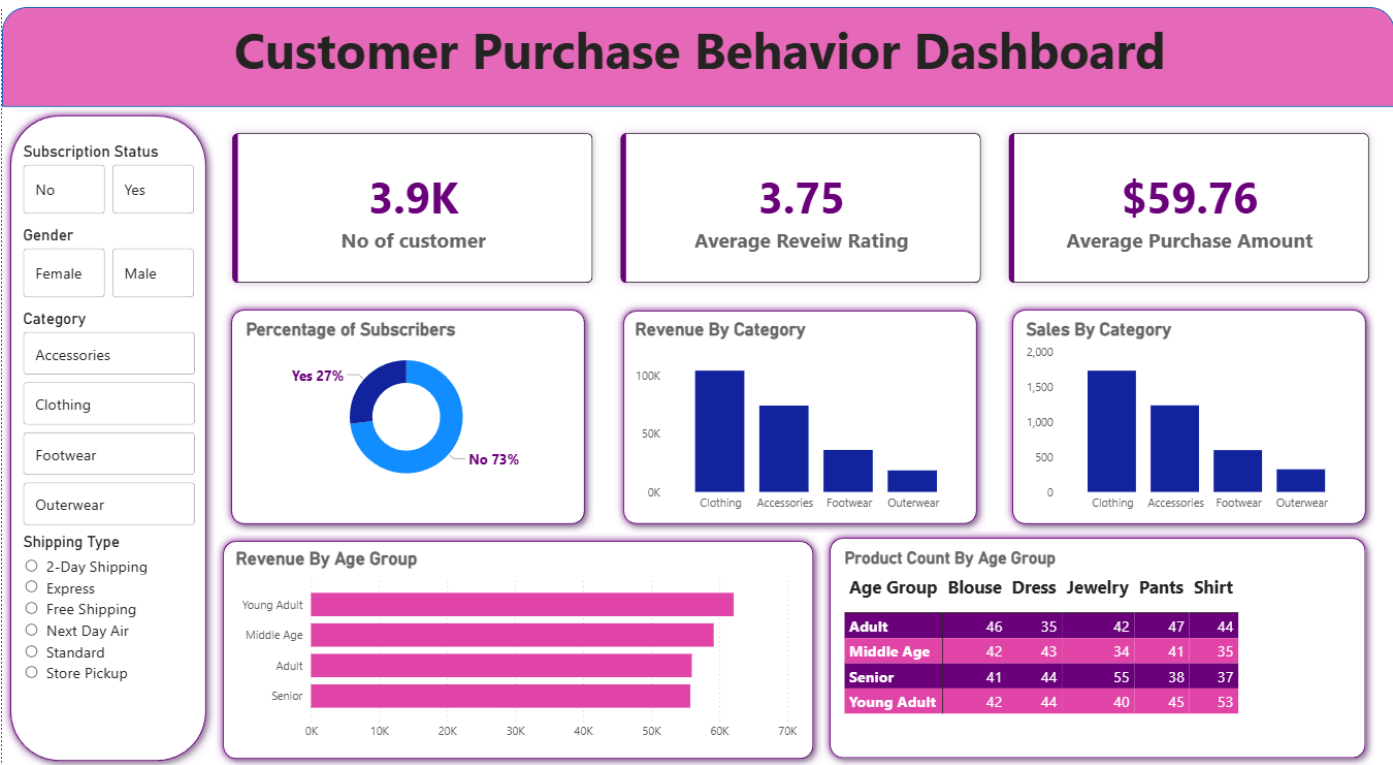
	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

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

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

# 5. Dashboard in Power BI

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



## 6. Business Recommendations

- **Boost Subscriptions** – Advertise unique advantages for subscribers.
- **Customer Loyalty Programs** – Incentivize repeat customers to transition them into the "Loyal" category.
- **Review Discount Policy** – Achieve a balance between sales increases and margin management.
- **Product Positioning** – Emphasize the highest-rated and most popular products in marketing campaigns.
- **Targeted Marketing** – Concentrate efforts on age demographics that generate significant revenue and the utilization of express shipping.