Customer Shopping Behavior Analysis – Project Report

1. Introduction

This project focuses on understanding how customers shop, what influences their purchasing decisions, and how different demographic and behavioral factors impact overall sales. Using a dataset of **3,900 purchase records**, this study identifies trends in spending habits, product choices, subscription patterns, and customer loyalty.The insights derived support better decision-making for marketing, product strategy, and customer retention.

2. Dataset Overview

The dataset includes **18 attributes** representing customer profiles, transaction details, and shopping behavior.
Key data fields include:

- Customer demographics: *Age, Gender, Location, Subscription Status*
- Purchase information: *Product, Category, Purchase Amount, Season, Size, Color*
- Behavior indicators: *Discount Applied, Promo Code Used, Shipping Type, Review Rating*
- Missing data: *37 missing ratings*, later treated during preprocessing

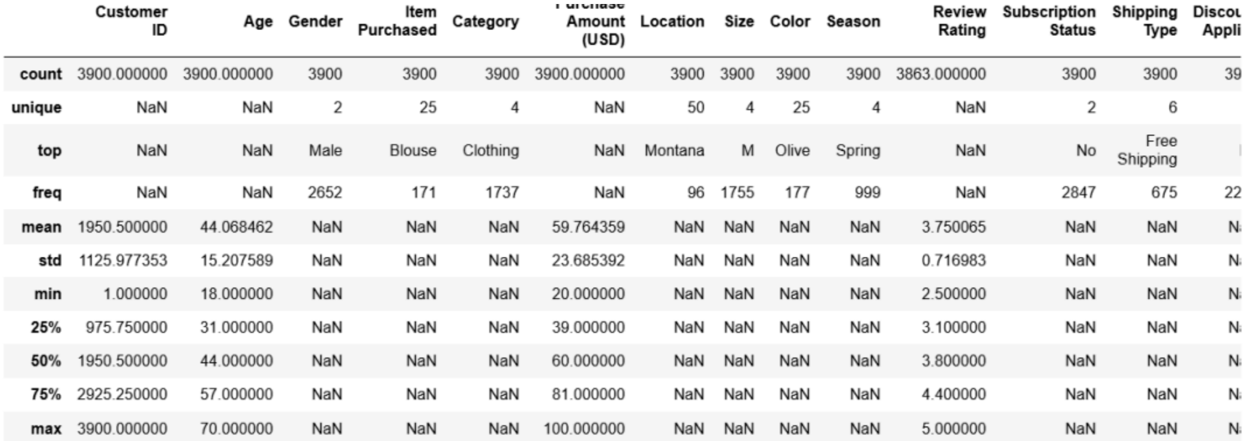
The dataset size and variety allow analysis across demographics, product categories, and transaction types.

3. Data Preprocessing & Python-Based Exploration

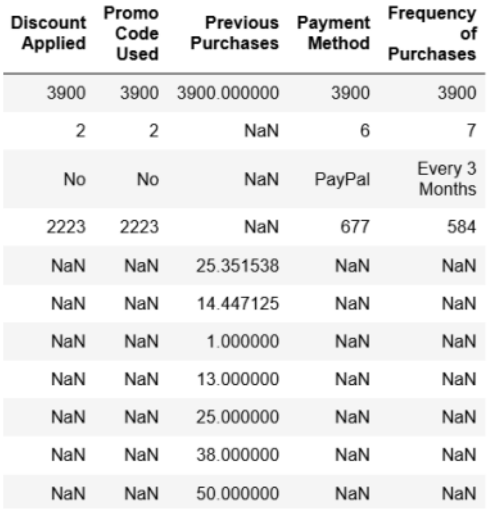
Data cleaning and preparation were done using Python.

3.1 Initial Checks

- Loaded dataset using **pandas**
- Used .info() and .describe() to understand structure and statistical summary



	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



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

3.2 Handling Missing Values

- Identified missing values in *Review Rating*
- Imputed missing ratings using **median rating of each product category**, ensuring consistency (Reference: Missing value description on page 1–2 of the uploaded file Customer Shopping Behavior Anal…)

3.3 Column Standardization

- Converted column names to *snake_case* for easier handling in code

3.4 Feature Engineering

- Created **age_groups** (e.g., Young Adult, Middle-aged)
- Generated **purchase_frequency_days** to measure shopping intervals

3.5 Data Cleaning Decisions

- Removed **promo_code_used** after identifying overlap with discount usage

3.6 PostgreSQL Integration

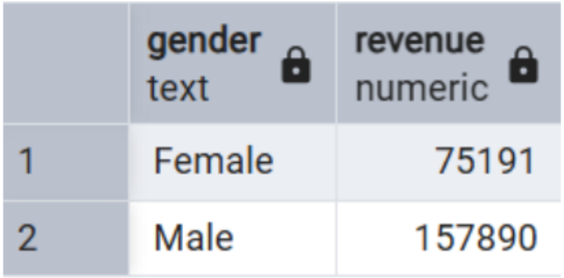
- Final cleaned DataFrame imported into PostgreSQL for SQL-based analysis

4. SQL Analysis & Business Queries

Using PostgreSQL, several analytical queries were executed to extract insights.

4.1 Revenue by Gender

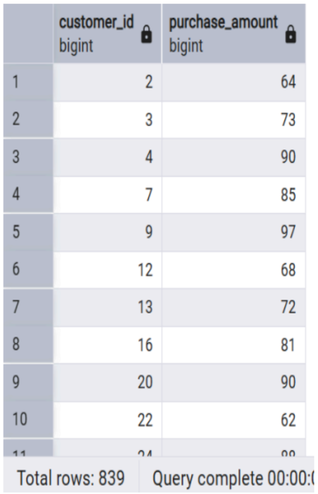
- Compared revenue contribution between male and female shoppers (Values reflected in page 3 table Customer Shopping Behavior Anal…)



	gender text	revenue numeric
1	Female	75191
2	Male	157890

4.2 High-Spending Discount Users

- Identified customers using discounts but still spending **above average**

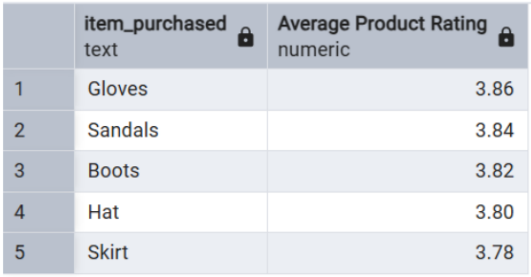


	customer_id bigint	purchase_amount bigint
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68
7	13	72
8	16	81
9	20	90
10	22	62
11	24	88

Total rows: 839 Query complete 00:00:

4.3 Top 5 Highest-Rated Products

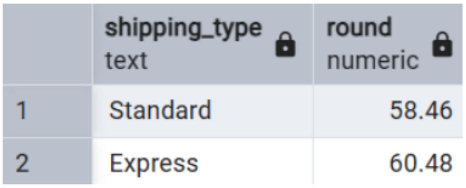
Products such as *Gloves, Sandals, Boots, Hats,* and *Skirts* showed the highest average ratings (Table on page 4 shows exact values Customer Shopping Behavior Anal…)



	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

4.4 Shipping Type Analysis

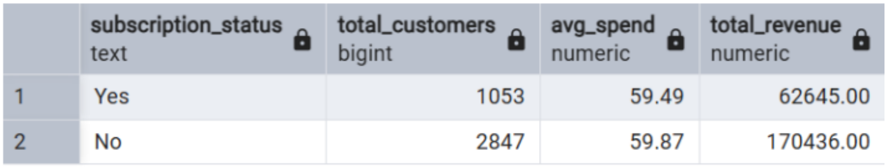
- Compared average order values for **Standard vs Express** shipping
  Page 4 shows that **Express** purchases had a slightly higher average amount



	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

4.5 Subscribers vs Non-Subscribers

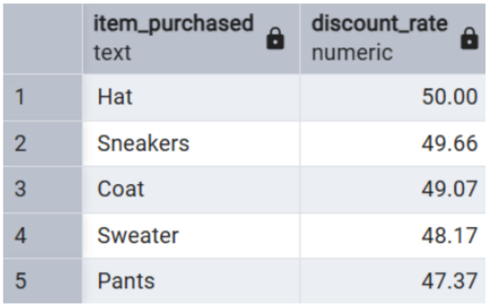
- Checked average spend and revenue from both groups
- Non-subscribers were more in number, but subscriber groups showed close average spend



	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

4.6 Discount-Driven Products

- Identified products with the highest percentage of discounted purchases such as *Hat* and *Sneakers*

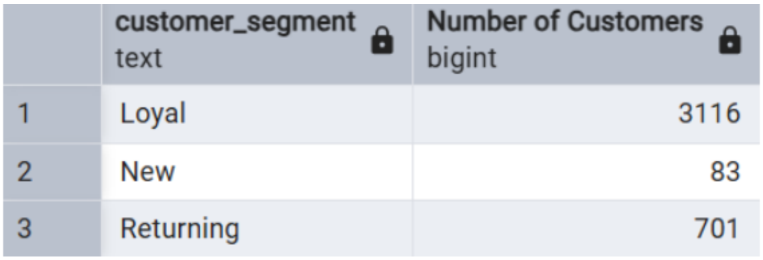


	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

4.7 Customer Segmentation

Segmented users into:

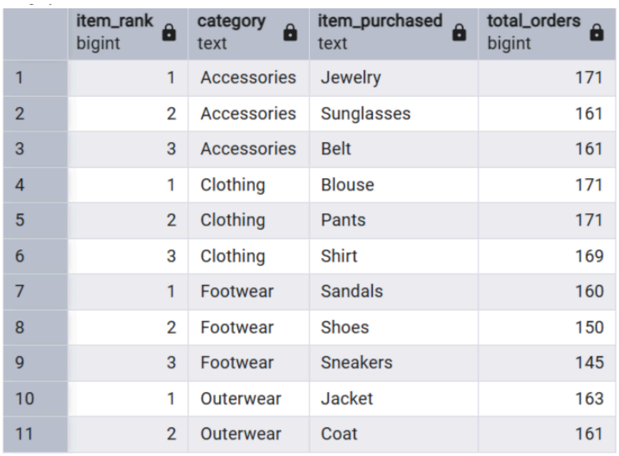
- **New**
- **Returning**
- **Loyal**
  As shown on page 5, Loyal customers dominate the dataset.



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

4.8 Top 3 Products per Category

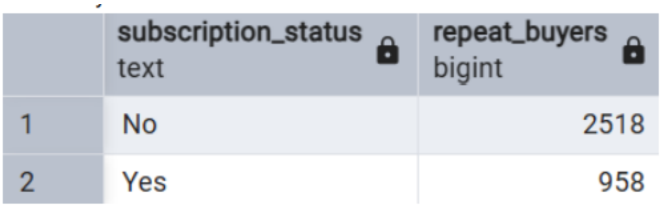
Used ranking logic to list the top-purchased items in each category



	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

4.9 Repeat Purchases vs Subscription

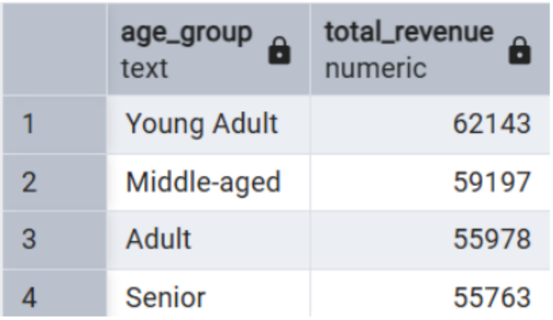
Checked if customers with >5 purchases tend to subscribe
(Page 6 table used—Yes vs No subscription counts)



	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

4.10 Revenue by Age Group

- Identified age groups contributing the most revenue
  Pages 6–7 indicate **Young Adults** created the highest revenue share

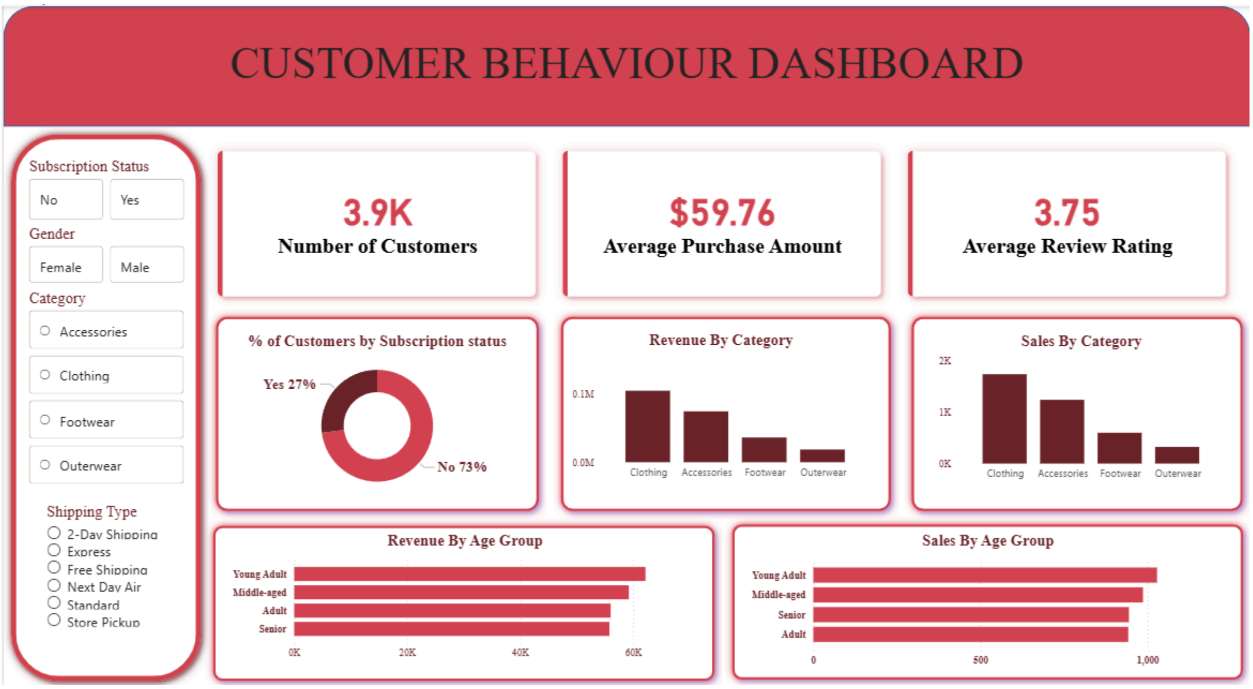


	age_group text	total_revenue numeric
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

5. Interactive Dashboard in Power BI

A Power BI dashboard was designed to visualize:

- Customer demographics
- Subscription distribution
- Revenue by category
- Sales by age group
- Average purchase amounts
- Review ratings



This helps in quick interpretation of trends and customer patterns.

6. Business Insights & Recommendations

Based on the overall analysis, the following recommendations were developed:

6.1 Strengthen Subscription Plans

Subscribers contribute significantly to revenue.
Offer:

- Exclusive discounts
- Early access to new items
- Personalized shopping suggestions

6.2 Enhance Customer Loyalty Program

Loyal customers form the majority.
Introduce:

- Tiered reward programs
- Cashback on repeat purchases

6.3 Adjust Discount Strategy

Products with extremely high discount usage should be reviewed to maintain profits.

6.4 Promote High-Rated & Best-Selling Products

Highlight top-rated and most purchased items in marketing campaigns to drive conversions.

6.5 Age-Based Targeted Marketing

Young Adults and Middle-aged customers show high spending power—target them with category-specific ads.

6.6 Promote Express Shipping Offers

Express shipping customers spend more on average; offering bundled express shipping may boost conversions.

7. Conclusion

This project successfully analyzed customer shopping behavior using Python, SQL, and Power BI.
Key achievements include:

- Cleaning and preparing a multi-feature dataset
- Executing complex SQL queries for behavioral insights
- Building a structured dashboard for visualization
- Delivering actionable business recommendations

The study demonstrates how data analytics can significantly improve strategic decisions in retail, particularly around customer retention, product promotion, and revenue optimization.