Customer Shopping Behaviour Analysis

1.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.

2.Data Summary

- Rows : 3900

- Columns: 18

- Key Features:

Customer Demographics (age, Gender, Location, Subscription Status)

Purchase Details (Item purchased, Category, Purchase Amount, Season, Size, Colour)

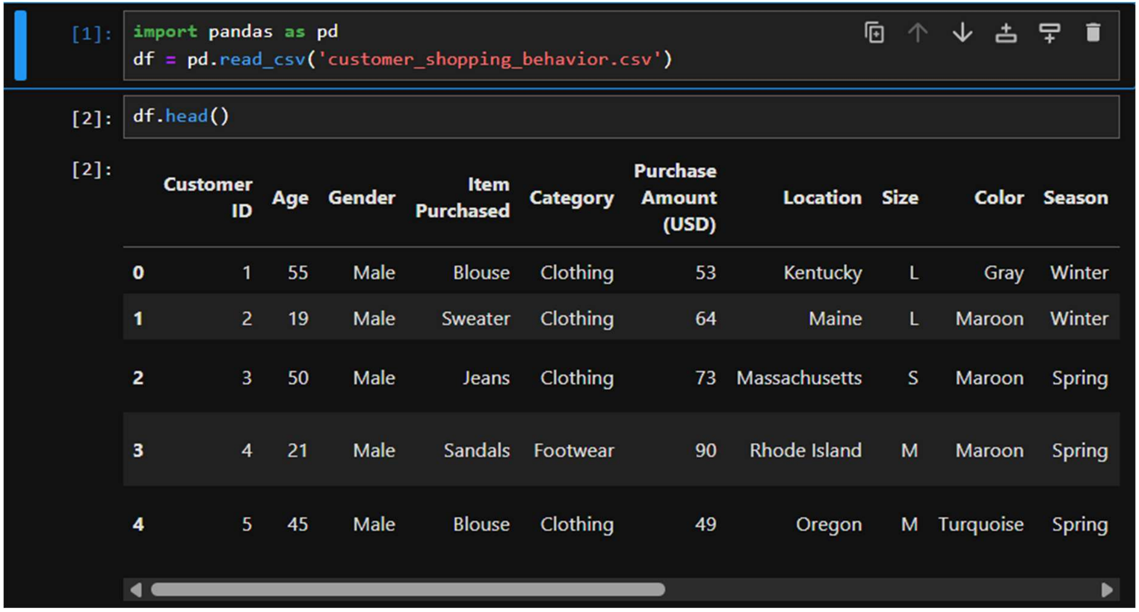
Shopping behaviour (Discount Applies, Promo Code Used, Previous Purchases, Frequency of Purchase, Review Rating, Shipping Type)

- Missing Data: 37 values in Review Rating column

3. Exploratory Data Analysis using Python

Started with data Preparation and cleaning in python:

- **Data Loading:**



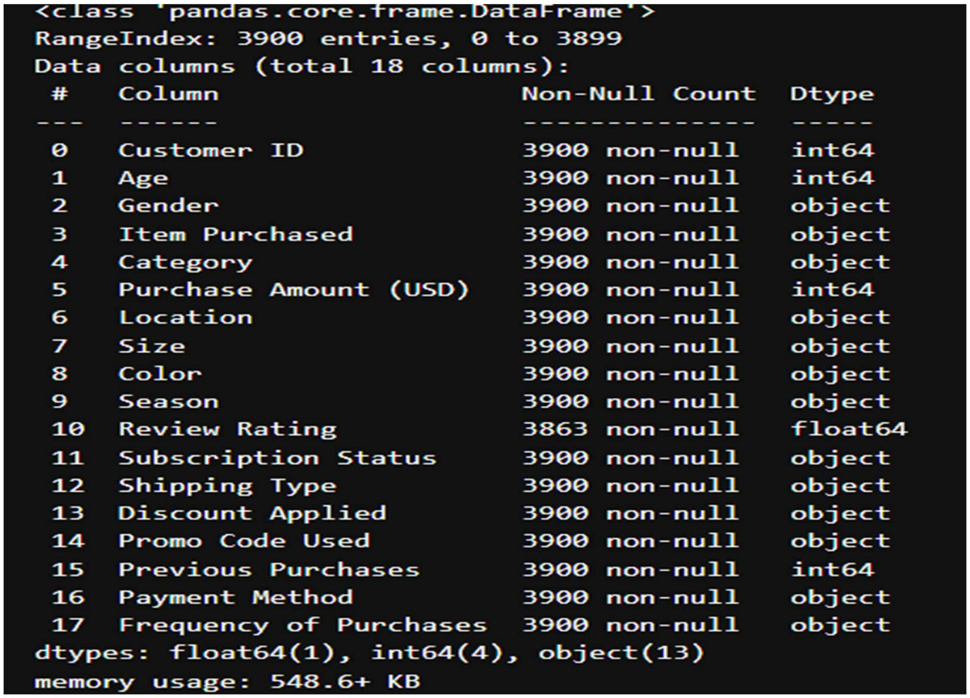
```
[1]: import pandas as pd
     df = pd.read_csv('customer_shopping_behavior.csv')
```

```
[2]: df.head()
```

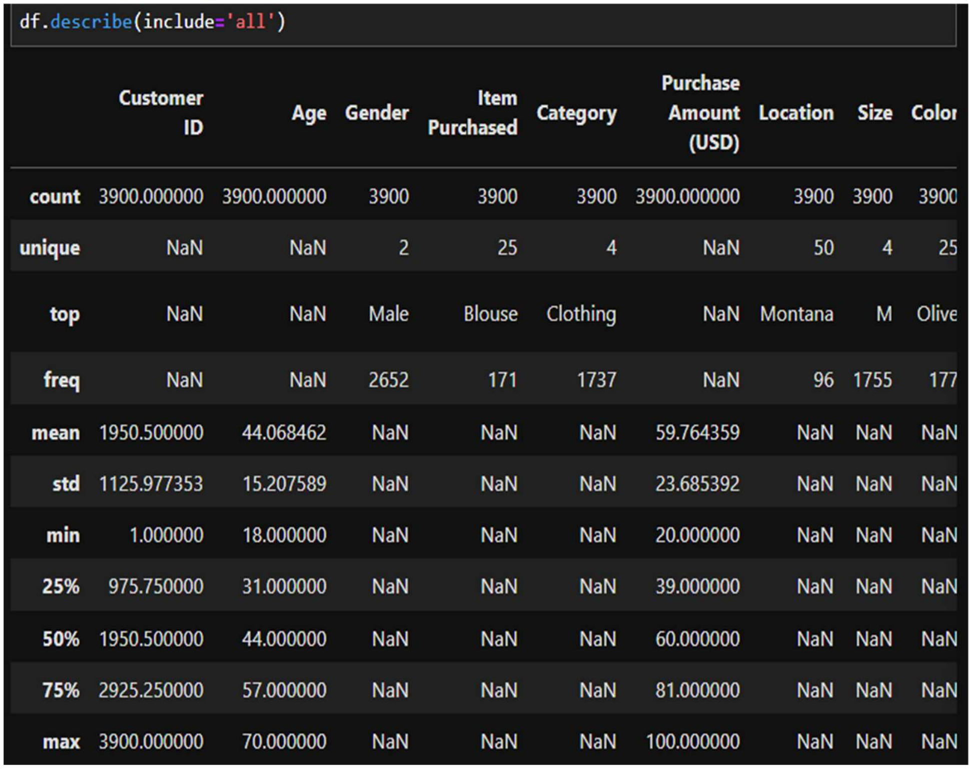
[2]:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring

- **Initial Exploration:**
Used df.info( ) to check structure and .describe( ) for summary statistics.



```
<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
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	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.

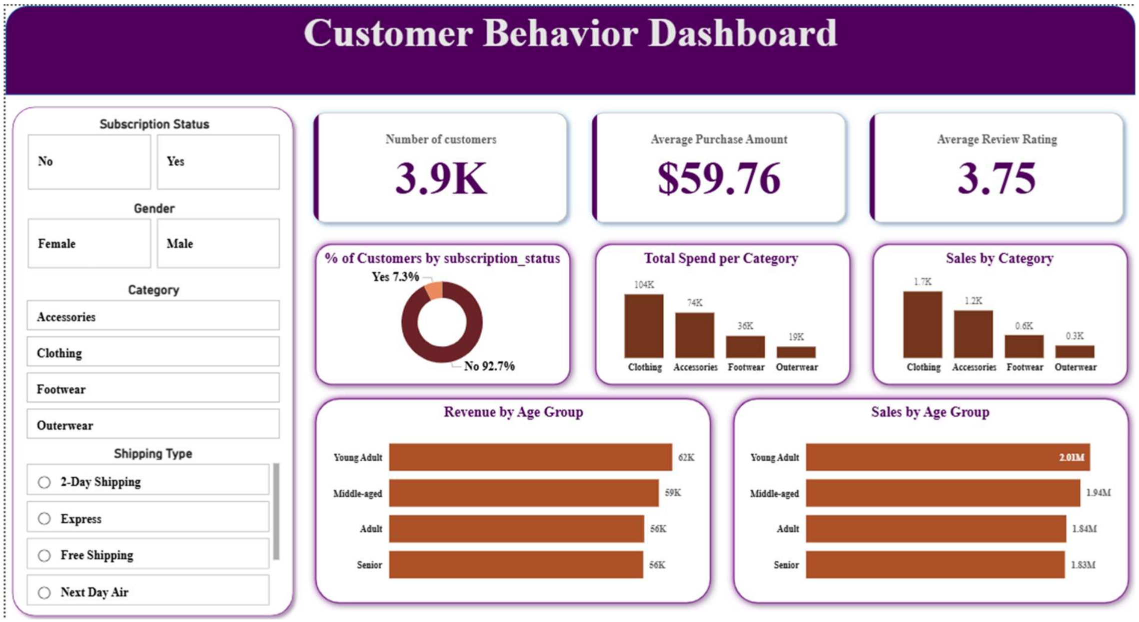
- ○ Created purchase_frequency_days column from purchase data.
- ➢ **Data Consistency Check:** Verified if discount_applied and promo_code_used were redundant; dropped promo_code_used.
- ➢ **Database Integration:** Connected Python script to PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

4. Data Analysis using SQL (Business Transactions)

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

1. **Revenue by Gender –** Compared total revenue generated by male vs. female customers.
2. **High-Spending Discount Users –** Identified customers who used discounts but still spent above the average purchase amount.
3. **Top 5 Products by Rating** – Found products with the highest average review ratings.
4. **Shipping Type Comparison** – Compared average purchase amounts between Standard and Express shipping.
5. **Subscribers vs. Non-Subscribers** – Compared average spend and total revenue across subscription status.
6. **Discount-Dependent Products** – Identified 5 products with the highest percentage of discounted purchases.
7. **Customer Segmentation –** Classified customers into New, Returning, and Loyal segments based on purchase history
8. **Top 3 Products per Category** – Listed the most purchased products within each category.

5. Dashboard in Power BI



Customer Behavior Dashboard

Number of customers: 3.9K

Average Purchase Amount: $59.76

Average Review Rating: 3.75

6. 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