Customer Shopping Behaviour Analysis

Project Overview:

This project analyzes 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.

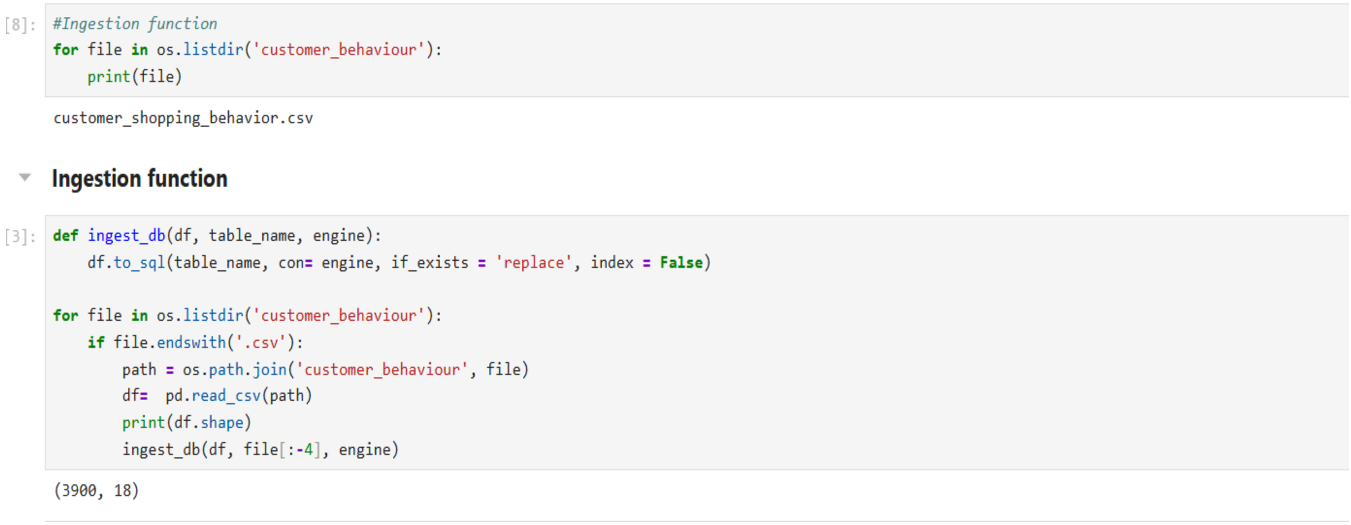
Dataset Summary

- Rows: 3,900
- Columns: 18
- Key Features:
- Customer demographics (Age, Gender, Location, Subscription Status)
- Purchase details (Item Purchased, Category, Purchase Amount, Season, Size, Colour)
- Shopping behaviour (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

● Data Loading: Imported the dataset using pandas.

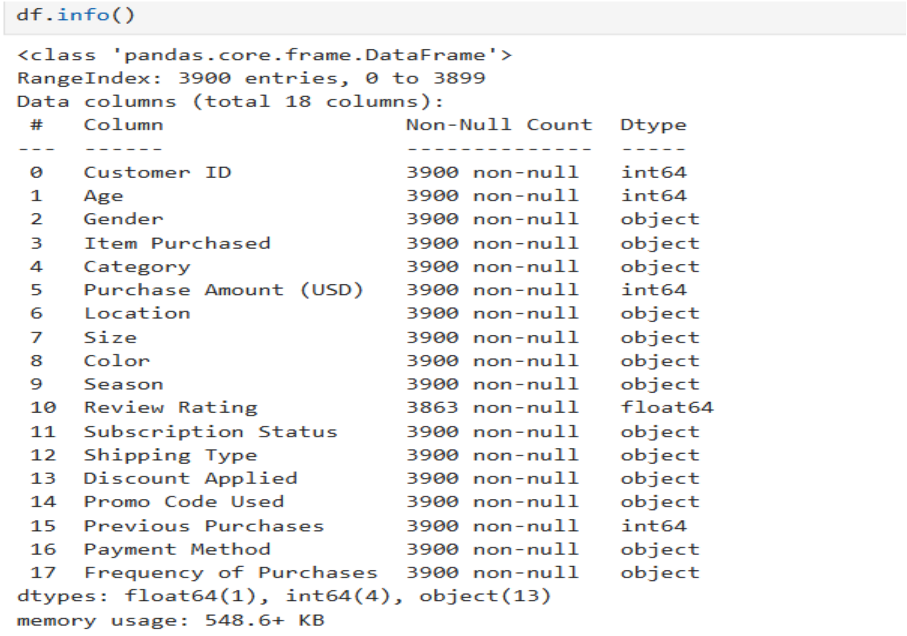
● Initial Exploration: Used df.info() to check structure and df.describe() for summary statistics.



```
[8]: #Ingestion function
     for file in os.listdir('customer_behaviour'):
         print(file)

     customer_shopping_behavior.csv
```

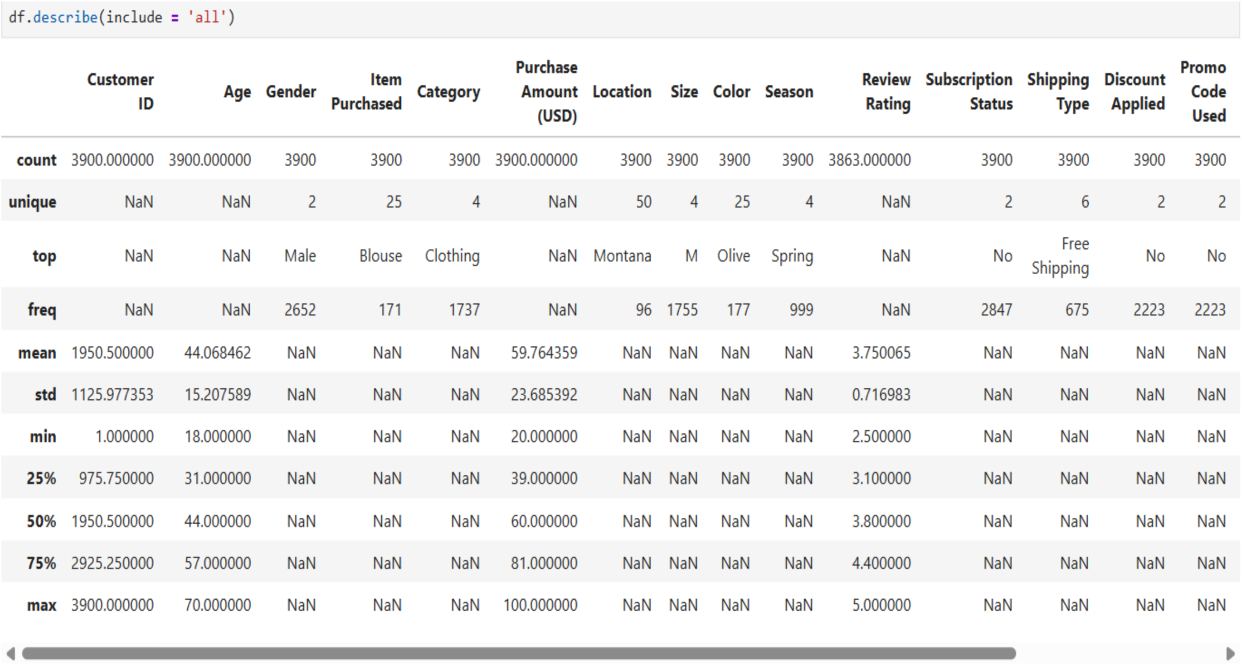
Ingestion function

```
[3]: def ingest_db(df, table_name, engine):
         df.to_sql(table_name, con= engine, if_exists = 'replace', index = False)

     for file in os.listdir('customer_behaviour'):
         if file.endswith('.csv'):
             path = os.path.join('customer_behaviour', file)
             df=  pd.read_csv(path)
             print(df.shape)
             ingest_db(df, file[:-4], engine)

     (3900, 18)
```



```
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

● **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.

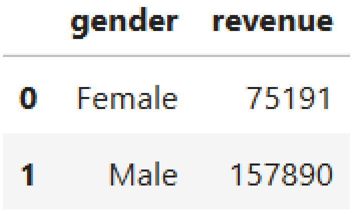
● **Data Consistency Check:** Verified if discount_applied and promo_code_used were redundant; dropped promo_code_used.

● **Database Integration:** Connected Python script to Sqlite and loaded the cleaned DataFrame into the database for SQL analysis.

Data Analysis using SQL (Business Transactions)

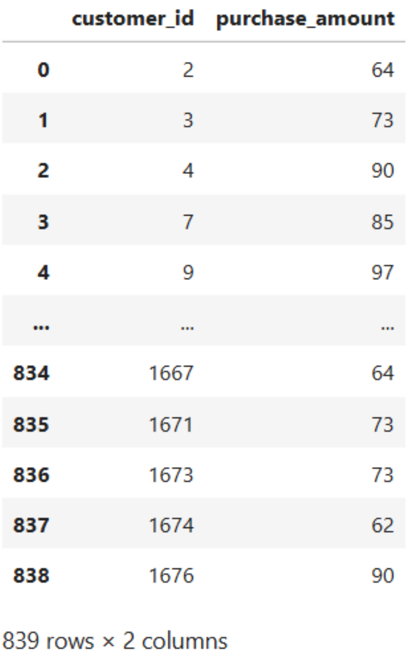
Performed structured analysis in PostgreSQL to answer key business questions:

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



	gender	revenue
0	Female	75191
1	Male	157890

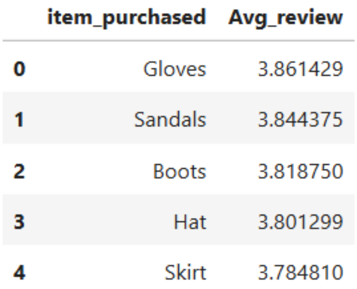
2. High-Spending Discount Users – Identified customers who used discounts but still spent above the average purchase amount.



	customer_id	purchase_amount
0	2	64
1	3	73
2	4	90
3	7	85
4	9	97
...	...	...
834	1667	64
835	1671	73
836	1673	73
837	1674	62
838	1676	90

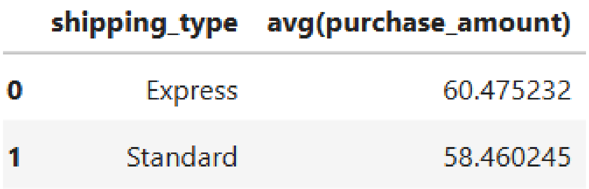
839 rows × 2 columns

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



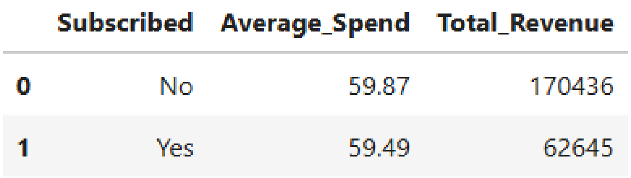
	item_purchased	Avg_review
0	Gloves	3.861429
1	Sandals	3.844375
2	Boots	3.818750
3	Hat	3.801299
4	Skirt	3.784810

4. Shipping Type Comparison – Compared average purchase amounts between Standard and Express shipping.



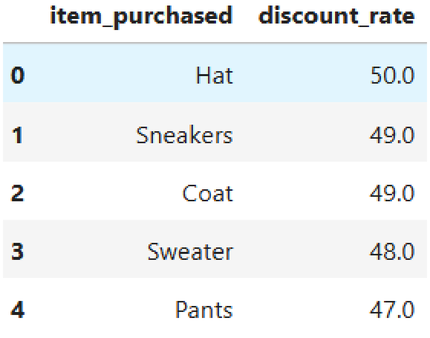
	shipping_type	avg(purchase_amount)
0	Express	60.475232
1	Standard	58.460245

5. Subscribers vs. Non-Subscribers – Compared average spend and total revenue across subscription status.



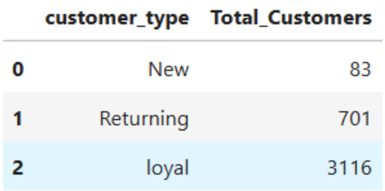
	Subscribed	Average_Spend	Total_Revenue
0	No	59.87	170436
1	Yes	59.49	62645

6. Discount-Dependent Products – Identified 5 products with the highest percentage of discounted purchases.



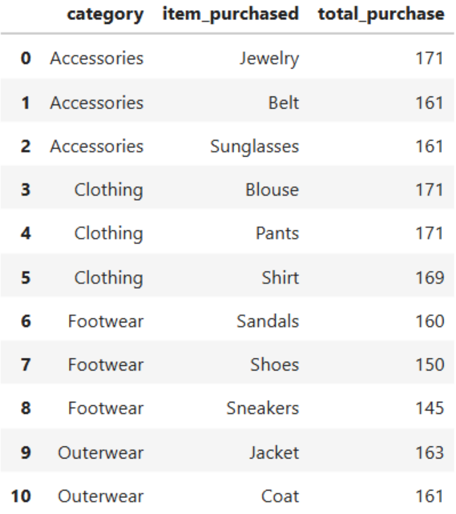
	item_purchased	discount_rate
0	Hat	50.0
1	Sneakers	49.0
2	Coat	49.0
3	Sweater	48.0
4	Pants	47.0

7. Customer Segmentation – Classified customers into New, Returning, and Loyal segments based on purchase history.



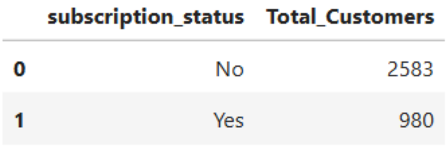
	customer_type	Total_Customers
0	New	83
1	Returning	701
2	loyal	3116

8. Top 3 Products per Category – Listed the most purchased products within each category.



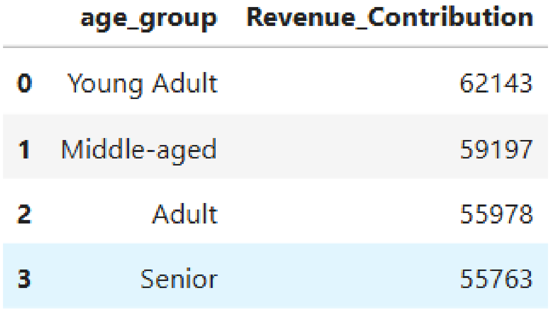
	category	item_purchased	total_purchase
0	Accessories	Jewelry	171
1	Accessories	Belt	161
2	Accessories	Sunglasses	161
3	Clothing	Blouse	171
4	Clothing	Pants	171
5	Clothing	Shirt	169
6	Footwear	Sandals	160
7	Footwear	Shoes	150
8	Footwear	Sneakers	145
9	Outerwear	Jacket	163
10	Outerwear	Coat	161

9. Repeat Buyers & Subscriptions – Checked whether customers with >5 purchases are more likely to subscribe.



	subscription_status	Total_Customers
0	No	2583
1	Yes	980

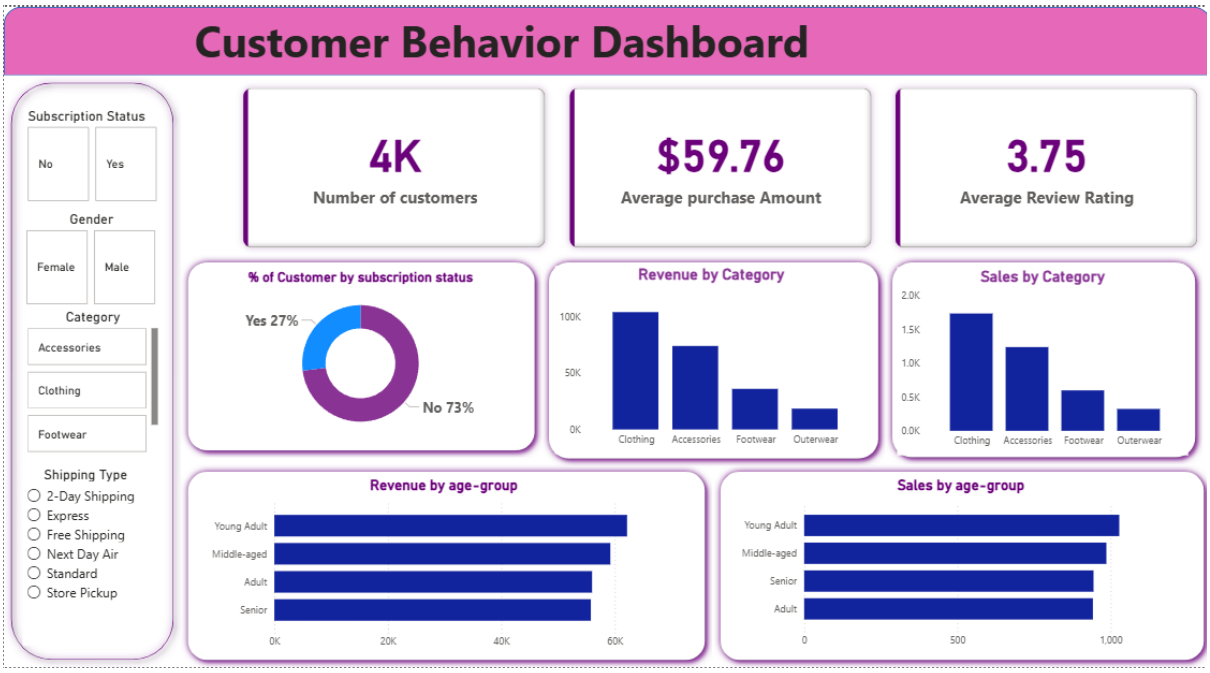
10. Revenue by Age Group – Calculated total revenue contribution of each age group.



	age_group	Revenue_Contribution
0	Young Adult	62143
1	Middle-aged	59197
2	Adult	55978
3	Senior	55763

Dashboard in Power BI

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



Customer Behavior Dashboard

4K
Number of customers

$59.76
Average purchase Amount

3.75
Average Review Rating

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.