Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior 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 behavior to guide strategic business decisions.

2. Dataset Summary

- Rows: 3,900
- Columns: 18

Key data points

- Customer demographics (Age, Gender, Location, Subscription Status)
- Customer demographics (Age, Gender, Location, Subscription Status)
- Shopping behavior (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)
- Missing Data: 37 values in Review Rating column

3. Extract Transform Load using Python

I began with cleaning data and preparing it for analysis using Python.

- **Data Loading**: Imported the dataset from CSV file using pandas
- **Initial Exploration :**
    - **df.info()** to check the structure of data

```
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()** for statistics summary

```
df.describe().round(2)
```

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.00	3900.00	3900.00	3863.00	3900.00
mean	1950.50	44.07	59.76	3.75	25.35
std	1125.98	15.21	23.69	0.72	14.45
min	1.00	18.00	20.00	2.50	1.00
25%	975.75	31.00	39.00	3.10	13.00
50%	1950.50	44.00	60.00	3.80	25.00
75%	2925.25	57.00	81.00	4.40	38.00
max	3900.00	70.00	100.00	5.00	50.00

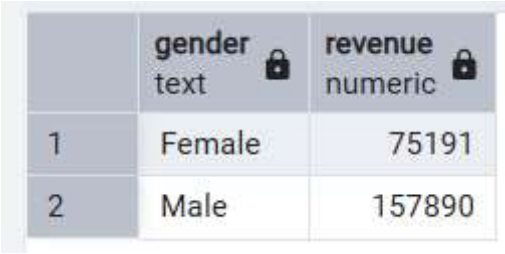
- **Missing Data Handling:** Checked for null values and imputed missing values in the **Review Rating** column using the median rating of each product category.
- Column Standardization: Renamed columns to snake case for better readability and documentation
- **Feature Engineering:**

    - Created **age_group** column by binning customer ages
    - Created **purchase_frequency_days** column from purchase data
- **Database Integration:** Connected Python script to PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

4. Exploratory Data Analysis using SQL (Business Transactions)

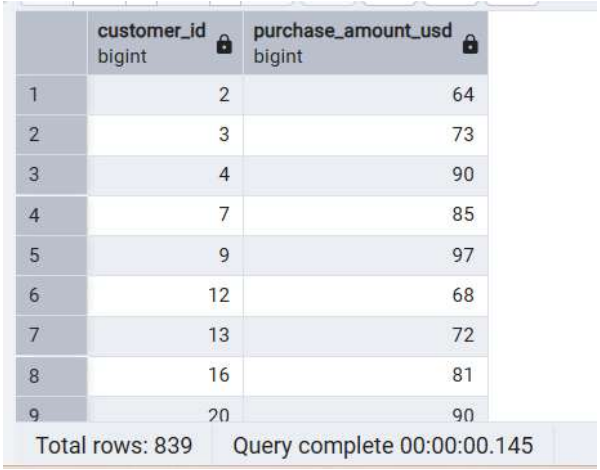
I performed structured analysis in PostgreSQL to answer key business questions:
database: perfectdb
Schema: None
Table: orders
DBMS (Server): PostgreSQL

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



	gender text	revenue numeric
1	Female	75191
2	Male	157890

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



	customer_id bigint	purchase_amount_usd 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

Total rows: 839 Query complete 00:00:00.145

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

	product text	mean_review 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** : Compared average purchase amounts between Standard and Express shipping.

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

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

	subscriptior text	no_of_customers bigint	avg_spent numeric	total_spent numeric
1	No	2847	59.87	170436.00
2	Yes	1053	59.49	62645.00

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

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

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

	customer_category text	no_of_customers bigint
1	Loyal	3116
2	Returning	701
3	New	83

8. **Top 3 Products per Category** : Listed the most purchased products within each category. (I used DENSE_RANK() instead of ROW_NUMBER() because business needs insights of top 3 products per category and cannot ignore product that is performing well in sales)

	item_purchased text	category text	no_of_orders bigint	product_rnk bigint
1	Jewelry	Accessori...	171	1
2	Sunglasses	Accessori...	161	2
3	Belt	Accessori...	161	2
4	Scarf	Accessori...	157	3
5	Pants	Clothing	171	1
6	Blouse	Clothing	171	1
7	Shirt	Clothing	169	2

Total rows: 13 Query complete 00:00:00.100 CRLF Ln

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

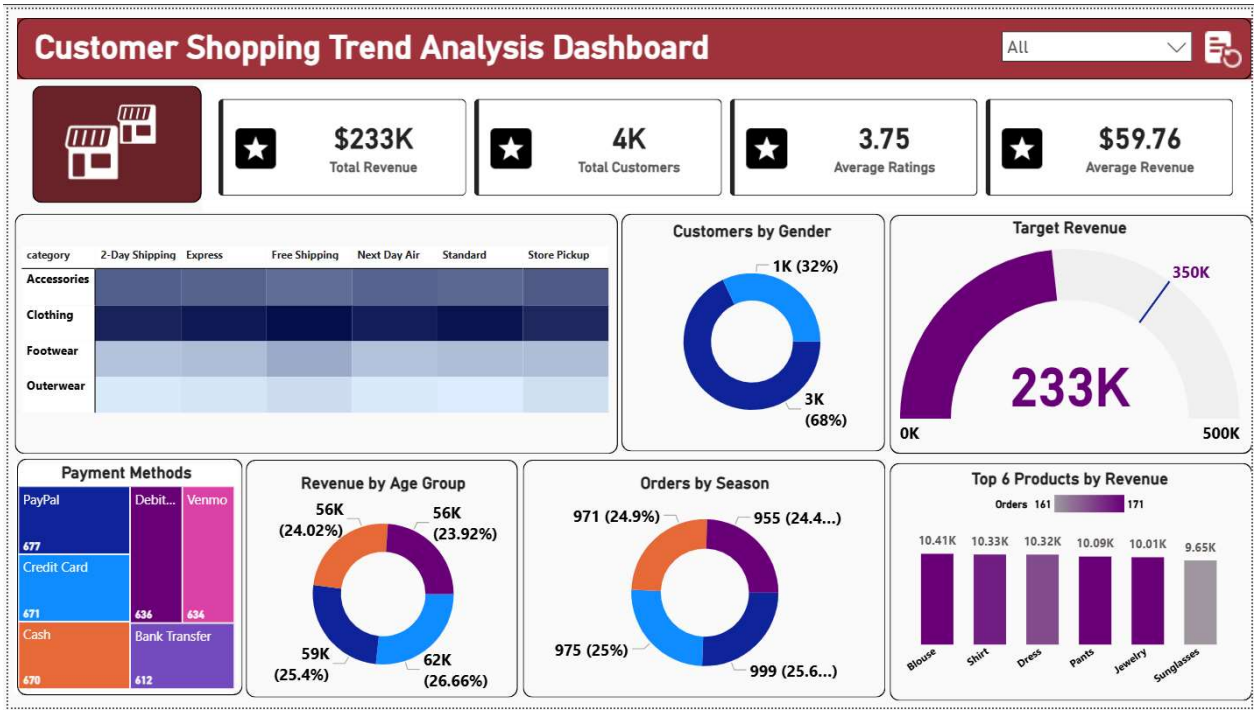
	subscriptior text	customers bigint
1	No	2518
2	Yes	958

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. Data Visualisation in Power BI

I created an interactive dashboard in Power BI to present insights visually to the business stakeholders.



6. Post Analysis Business Recommendations

- **Boost Subscriptions :** Promote exclusive benefits for subscribers
- **Customer Loyalty Programs:** Reward repeat buyers to move them into the 'Loyal' segment based on seasons
- **Product Positioning and Catalogue Optimisation:** Highlight top-rated and best selling products in campaigns and Optimise catalogue based on review ratings and feedback
- **Review Discount Policy**: Reward 'Loyal' customers with attractive discounts to boost sales along with margin control
- **Targeted Marketing:** Focus efforts on high-revenue age groups and express-shipping users

7. Summary

This report analyzes customer shopping behavior across demographics, products, and purchasing patterns to identify key trends and provide data-driven recommendations for improving revenue, engagement, and business strategy.