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. Dataset Summary

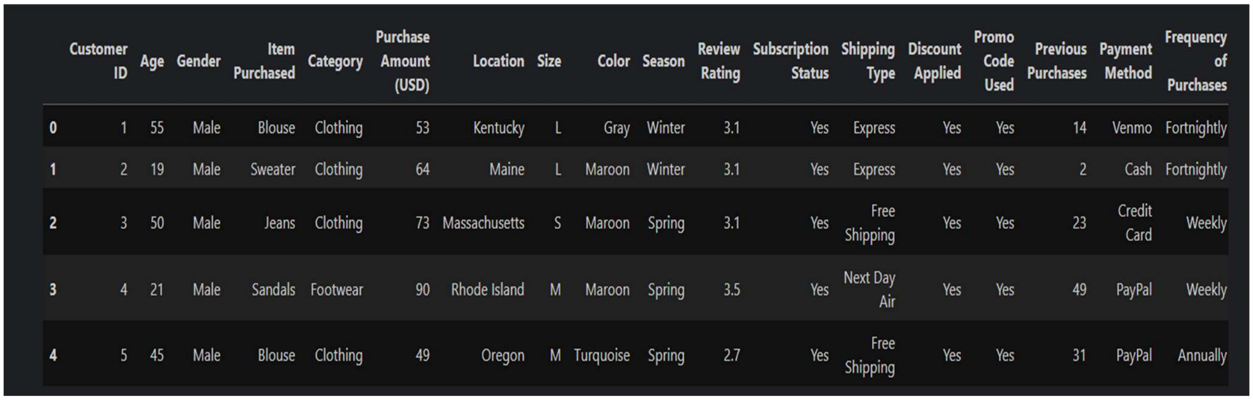
- Rows: 3,900
- Columns: 18
- Key Features:
    - Customer demographics (Age, Gender, Location, Subscription Status)
    - Purchase details (Item Purchased, Category, Purchase Amount, Season, Size, Color)
    - Shopping behaviour (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)
- Missing Data: 37 values in the Review Rating column

2. Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

● **Data Loading**: Imported the dataset using pandas.

● **Initial Exploration**: Used df.info() to check structure.



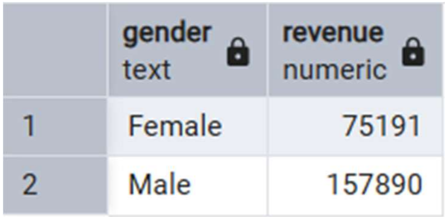
	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually

- ● **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 the age_group column by binning customer ages.
  - ○ Created frequency_purchase_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)

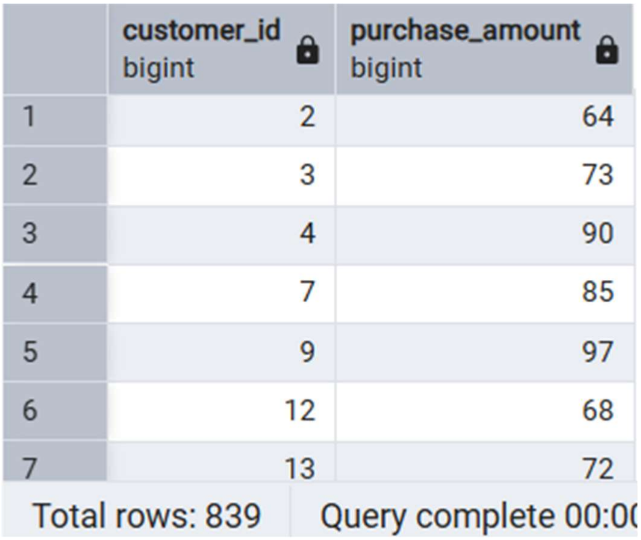
We performed structured analysis in PostgreSQL to answer key business questions:

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



	gender text	revenue numeric
1	Female	75191
2	Male	157890

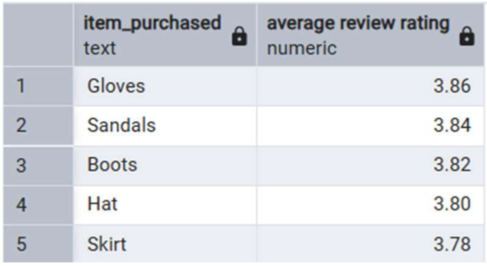
2. High-Spending Discount Users – Identified customers who used discounts but still spent above the average purchase amount.



	customer_id bigint	purchase_amount bigint
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68
7	13	72

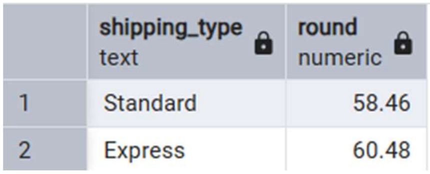
Total rows: 839 Query complete 00:0

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



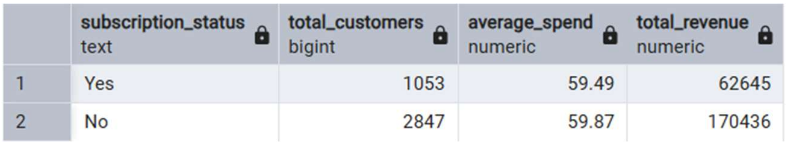
	item_purchased text	average review rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

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



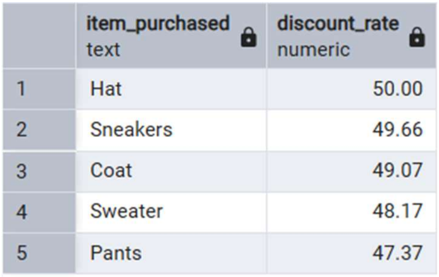
	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

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



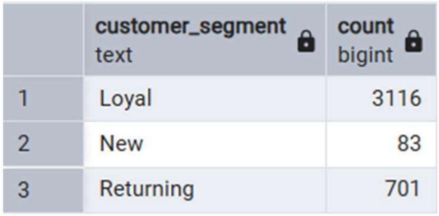
	subscription_status text	total_customers bigint	average_spend numeric	total_revenue numeric
1	Yes	1053	59.49	62645
2	No	2847	59.87	170436

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



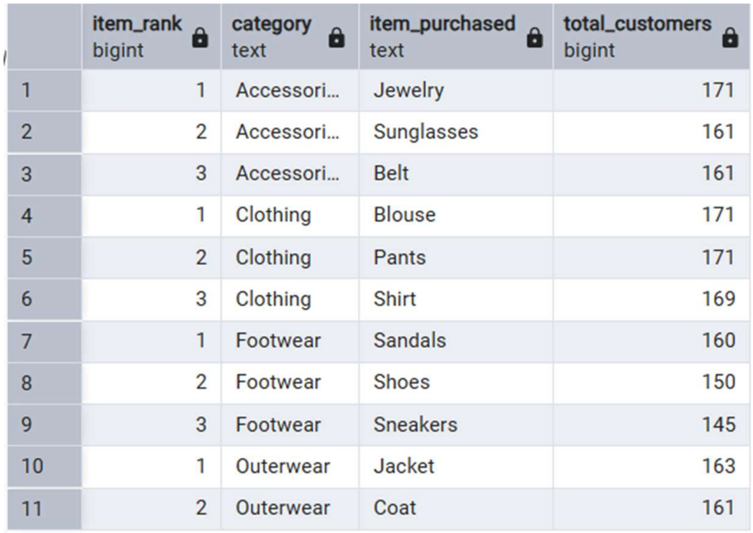
	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

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



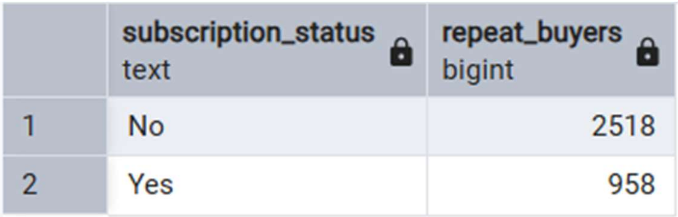
	customer_segment text	count bigint
1	Loyal	3116
2	New	83
3	Returning	701

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



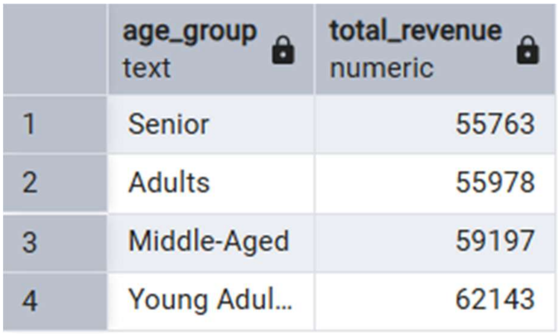
	item_rank bigint	category text	item_purchased text	total_customers bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	Belt	161
4	1	Clothing	Blouse	171
5	2	Clothing	Pants	171
6	3	Clothing	Shirt	169
7	1	Footwear	Sandals	160
8	2	Footwear	Shoes	150
9	3	Footwear	Sneakers	145
10	1	Outerwear	Jacket	163
11	2	Outerwear	Coat	161

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



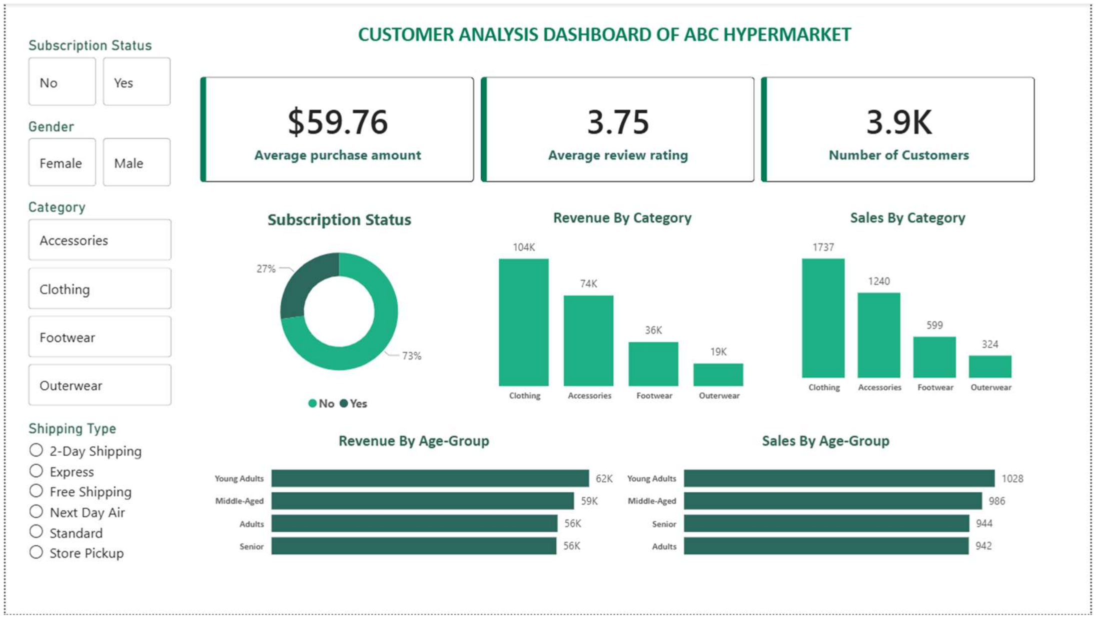
	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

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



	age_group text	total_revenue numeric
1	Senior	55763
2	Adults	55978
3	Middle-Aged	59197
4	Young Adul...	62143

5. Dashboard in Power BI



6. Business Recommendations

- **Boost Subscriptions** – Promote exclusive benefits for subscribers, especially for females
- **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, like placing clothes visible to everyone.
- **Targeted Marketing** – Focus efforts on high-revenue age groups and express-shipping users.