Customer Shopping Behavior Analysis

1. Project Overview

This project examines customer shopping behavior using transactional data from 3,900 purchases across multiple product categories. It aims to identify spending trends, customer segments, product preferences, and subscription patterns to support data-driven business decisions.

2. Dataset Summary

- Rows: 3,900

- Columns: 17

- Key Features:

- Customer demographics (Age, Gender, Location, Subscription Status)
- Purchase details (Item Purchased, Category, Purchase Amount, Season, Size, Color)
- Shopping behavior (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)

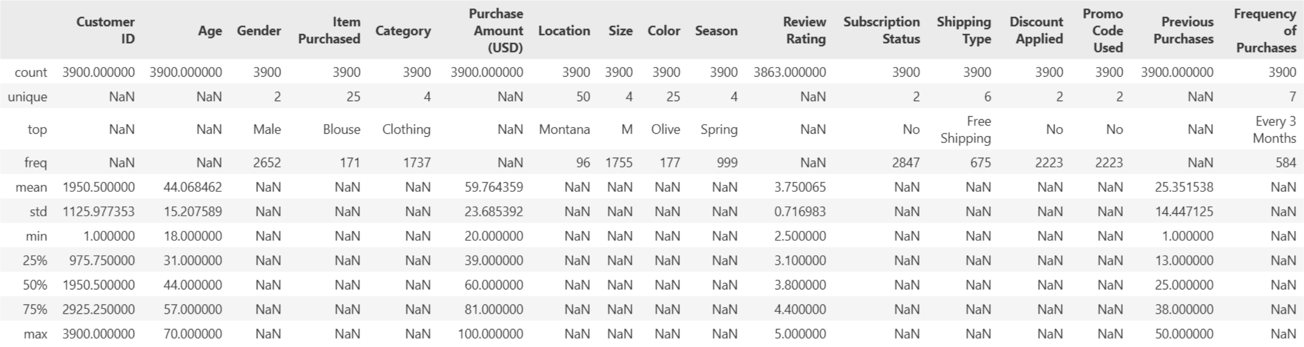
- Missing Data: 37 values in Review Rating column

3. 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 and df.describe() for summary statistics.



	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	Frequency of Purchases
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3900.000000	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN	7
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	NaN	Every 3 Months
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	NaN	584
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	25.351538	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	14.447125	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	1.000000	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	50.000000	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.

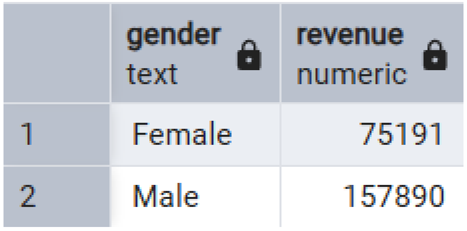
- **Data Consistency Check:** Verified if discount_applied and promo_code_used were redundant; dropped promo_code_used.

- **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)

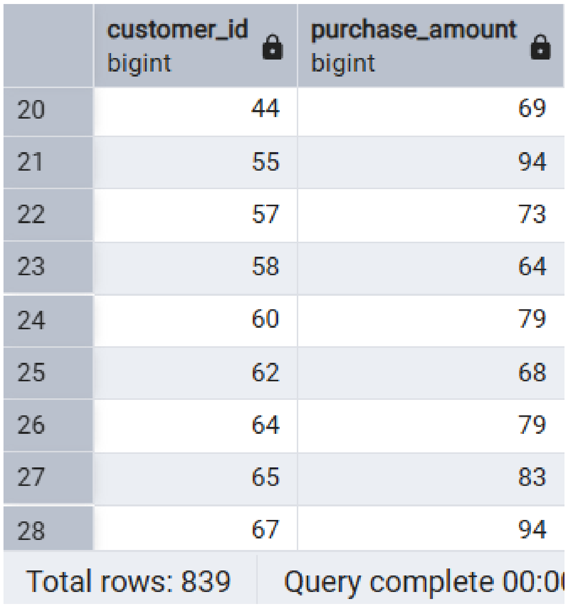
We performed structured analysis in PostgreSQL to answer key business questions:

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



	gender text	revenue numeric
1	Female	75191
2	Male	157890

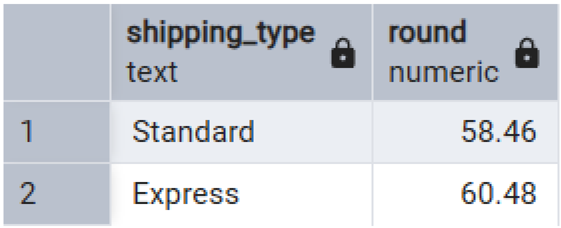
2. **High-Spending Discount Users** – Identified customers who used discounts but still spent above the average purchase amount.



	customer_id bigint	purchase_amount bigint
20	44	69
21	55	94
22	57	73
23	58	64
24	60	79
25	62	68
26	64	79
27	65	83
28	67	94

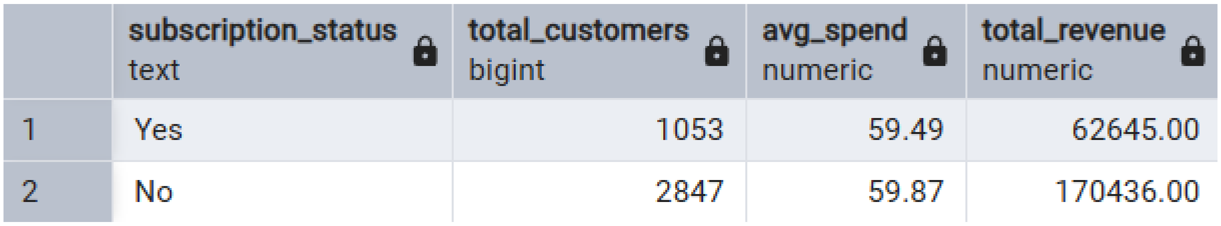
Total rows: 839 Query complete 00:0

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



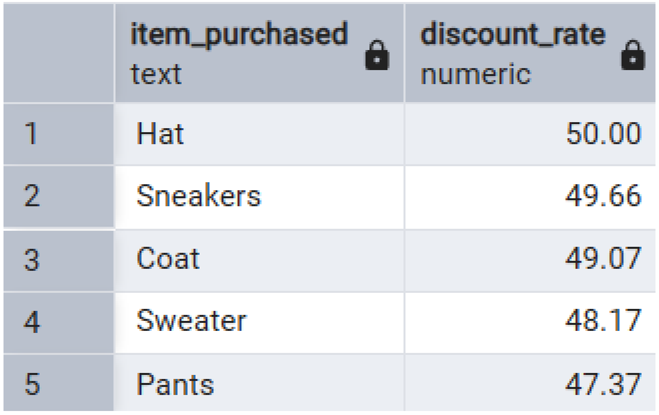
	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

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



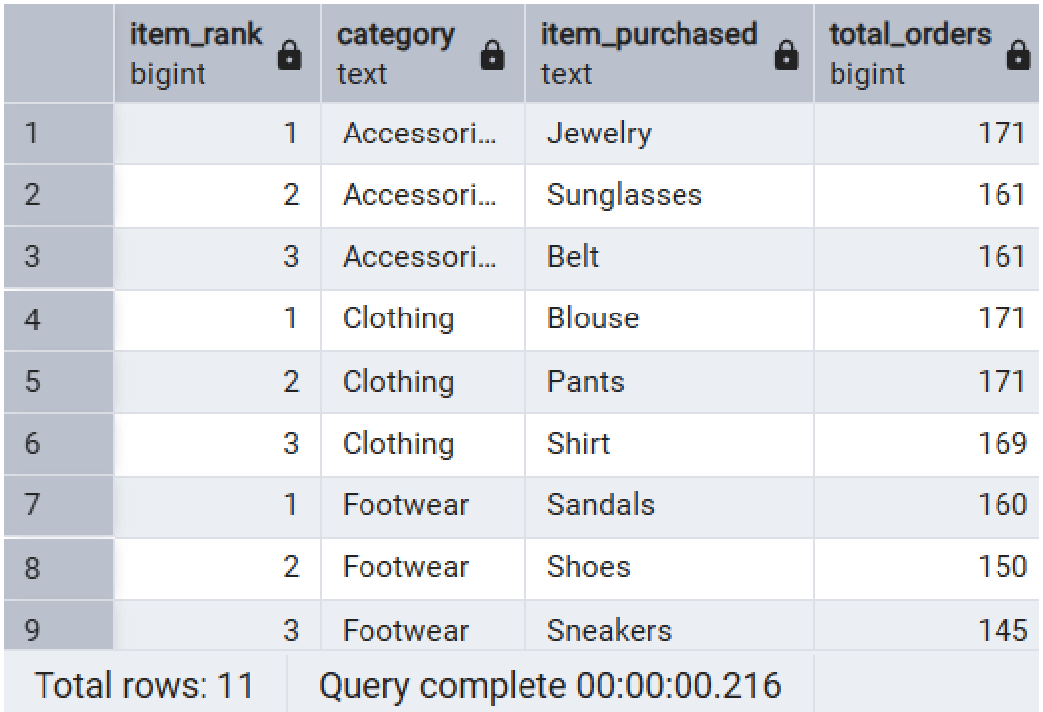
	subscription_status	total_customers	avg_spend	total_revenue
	text	bigint	numeric	numeric
1	Yes	1053	59.49	62645.00
2	No	2847	59.87	170436.00

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



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

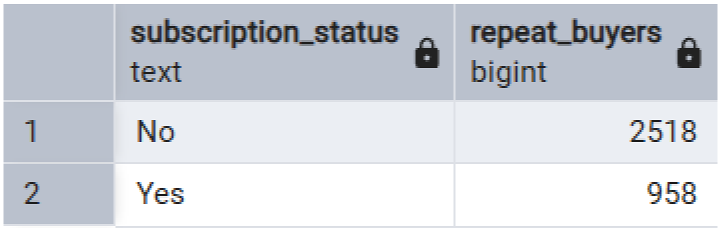
6. **Top 3 Products per Category** – Listed the most purchased products within each category.



	item_rank	category	item_purchased	total_orders
	bigint	text	text	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

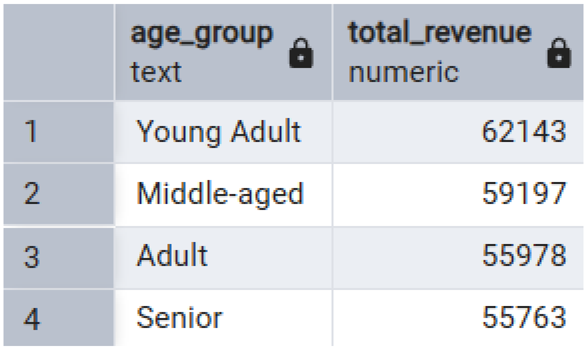
Total rows: 11 Query complete 00:00:00.216

7. **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

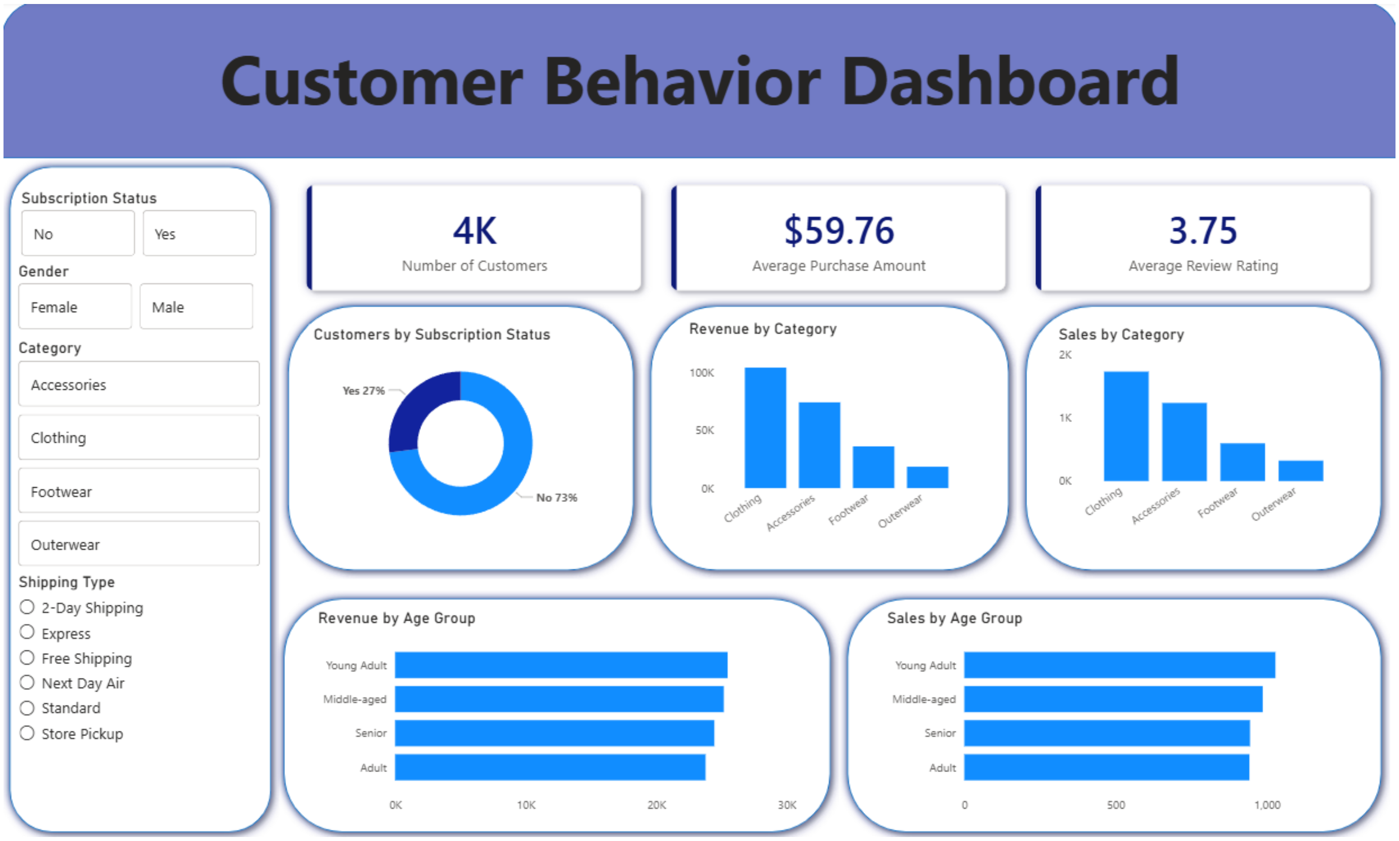
8. **Revenue by Age Group** – Calculated total revenue contribution of each age group.



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

5. Dashboard in Power BI

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



Customer Behavior Dashboard

Subscription Status: No | Yes

Gender: Female | Male

Category: Accessories | Clothing | Footwear | Outerwear

Shipping Type
○ 2-Day Shipping
○ Express
○ Free Shipping
○ Next Day Air
○ Standard
○ Store Pickup

4K Number of Customers

$59.76 Average Purchase Amount

3.75 Average Review Rating

6. Business Recommendations

- **Target High-Spending Discount Users:**
  Customers who use discounts while still spending above the average purchase amount represent a valuable segment. Personalized premium offers and exclusive product bundles should be designed for this group to increase retention and overall revenue.
- **Promote Top Products Within Each Category:**
  the top three most purchased products in each category should be prominently featured in marketing campaigns, homepage placements, and recommendation systems to maximize visibility and conversion rates.
- **Encourage Express Shipping Upsell:**
  Since customers choosing express shipping tend to have higher average purchase values, offering free or discounted express shipping above a specific spending threshold can effectively increase basket size.
- **Use Gender-Based Revenue Insights for Personalization:**
  Differences in revenue contribution by gender indicate opportunities for personalization. Gender-informed product recommendations, email marketing, and targeted promotions can improve customer engagement and sales performance.
- **Focus Marketing on High-Revenue Age Groups:**
  Marketing resources should be prioritized toward age groups that generate the highest total revenue. Age-specific messaging and tailored product recommendations can further enhance campaign effectiveness.
- **Improve the Review Collection Process:**
  To address missing review ratings, customers should be encouraged to leave feedback through post-purchase reminders. More complete review data enhances customer trust and supports more accurate product recommendations.