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.

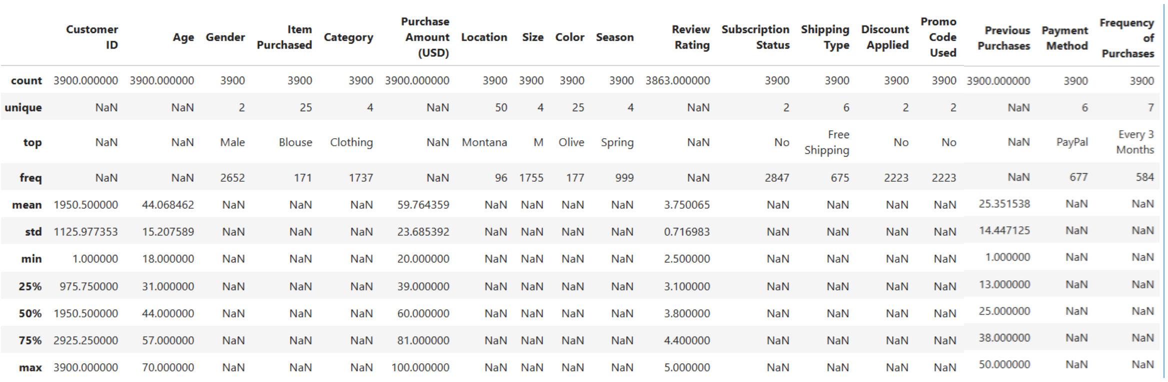
2. Dataset Summary

- Rows: 3,900 - Columns: 18 - Key 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 .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	Payment Method	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	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN	6	7
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	NaN	677	584
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	25.351538	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	14.447125	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	1.000000	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	50.000000	NaN	NaN

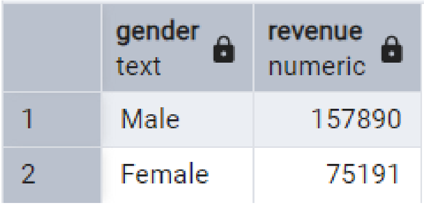
- **Missing Data Handling:** Checked for null values and imputed missing values in the Review Rating column using the median rating of each product category.
- **Column Standardization:** Renamed columns to **snake case** for better readability and documentation.

- **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 mostly the same; dropped promo_code_used to simplify dataset.
- **Database Integration:** Connected Python script to PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

4. Data Analysis using SQL (Business Questions)

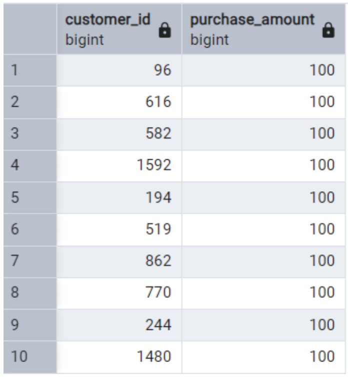
We used PostgreSQL queries to answer 10 practical business questions about revenue, segments, discounts, subscriptions, and product performance.

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



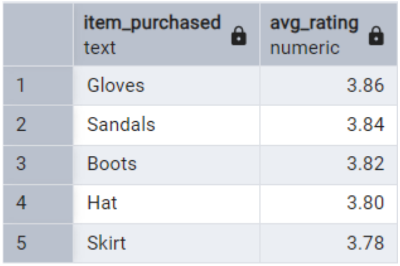
	gender text	revenue numeric
1	Male	157890
2	Female	75191

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



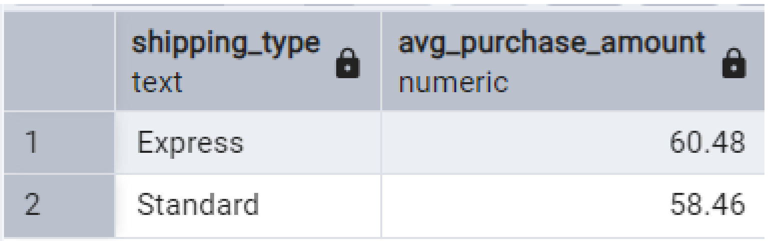
	customer_id bigint	purchase_amount bigint
1	96	100
2	616	100
3	582	100
4	1592	100
5	194	100
6	519	100
7	862	100
8	770	100
9	244	100
10	1480	100

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



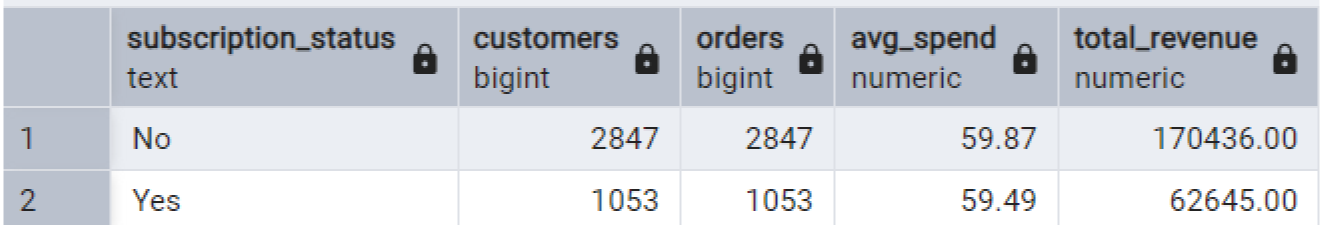
	item_purchased text	avg_rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

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



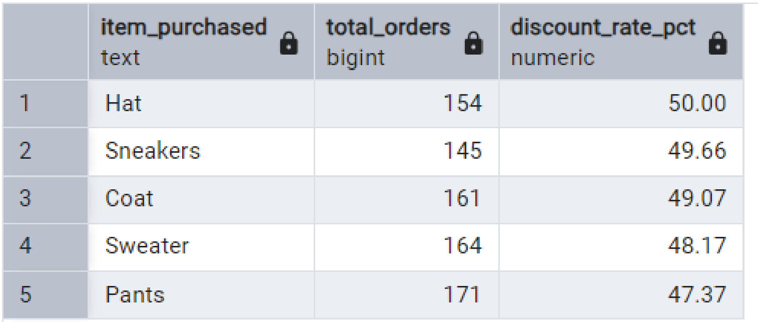
	shipping_type text	avg_purchase_amount numeric
1	Express	60.48
2	Standard	58.46

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



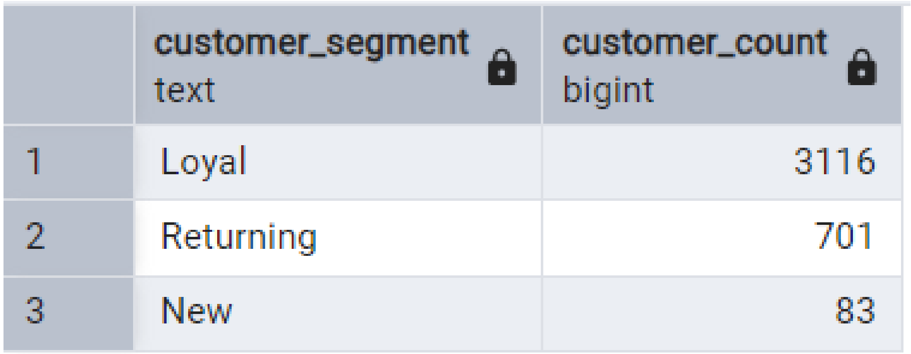
	subscription_status text	customers bigint	orders bigint	avg_spend numeric	total_revenue numeric
1	No	2847	2847	59.87	170436.00
2	Yes	1053	1053	59.49	62645.00

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



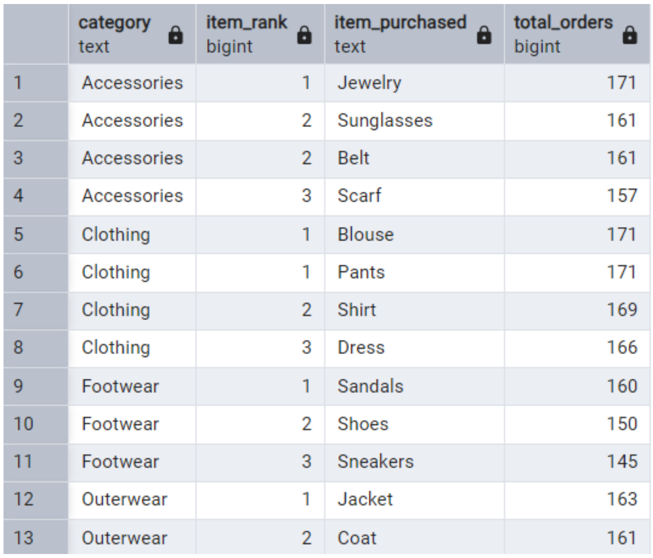
	item_purchased text	total_orders bigint	discount_rate_pct numeric
1	Hat	154	50.00
2	Sneakers	145	49.66
3	Coat	161	49.07
4	Sweater	164	48.17
5	Pants	171	47.37

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



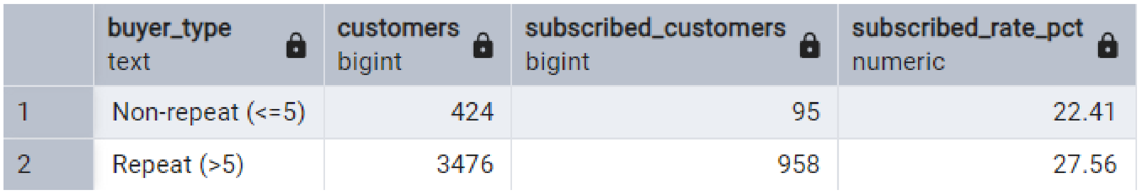
	customer_segment text	customer_count bigint
1	Loyal	3116
2	Returning	701
3	New	83

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



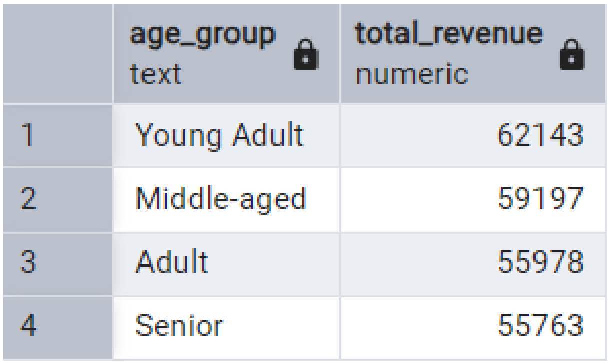
	category text	item_rank bigint	item_purchased text	total_orders bigint
1	Accessories	1	Jewelry	171
2	Accessories	2	Sunglasses	161
3	Accessories	2	Belt	161
4	Accessories	3	Scarf	157
5	Clothing	1	Blouse	171
6	Clothing	1	Pants	171
7	Clothing	2	Shirt	169
8	Clothing	3	Dress	166
9	Footwear	1	Sandals	160
10	Footwear	2	Shoes	150
11	Footwear	3	Sneakers	145
12	Outerwear	1	Jacket	163
13	Outerwear	2	Coat	161

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



	buyer_type text	customers bigint	subscribed_customers bigint	subscribed_rate_pct numeric
1	Non-repeat (<=5)	424	95	22.41
2	Repeat (>5)	3476	958	27.56

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



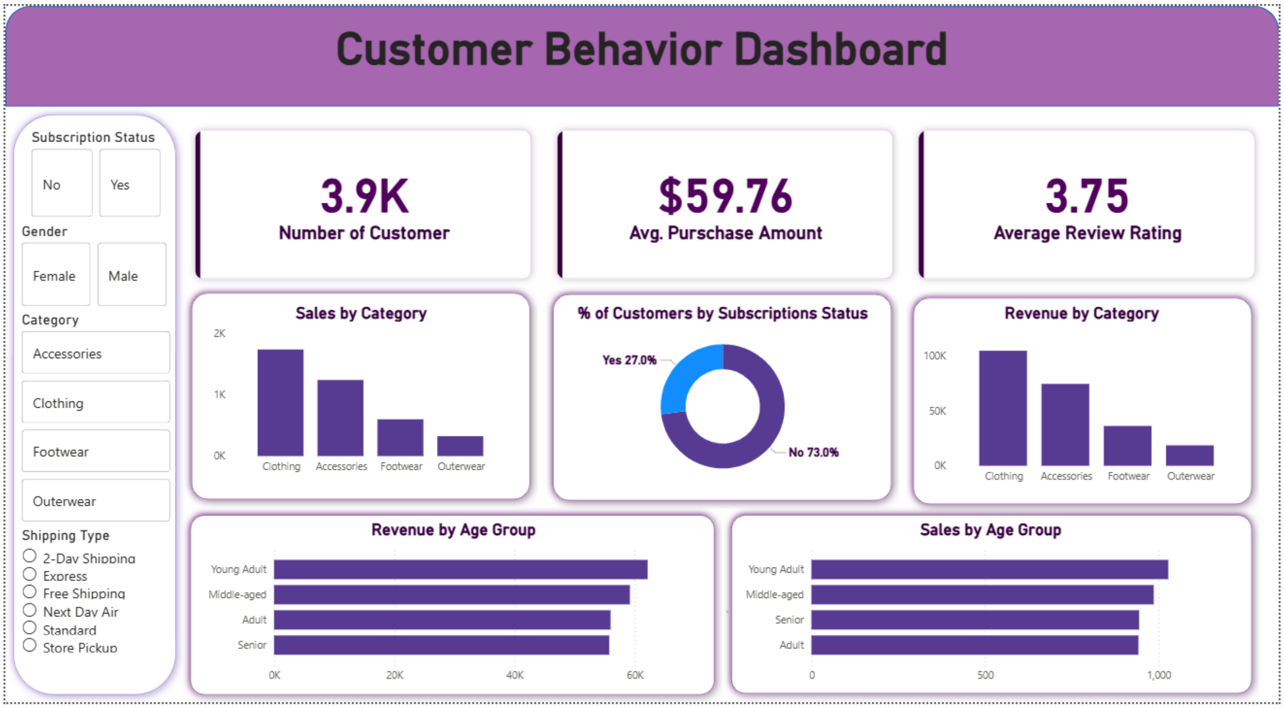
	age_group text	total_revenue numeric
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

SQL Techniques Used

- Filtering and sorting (WHERE, ORDER BY)
- Aggregations (SUM, AVG, COUNT)
- GROUP BY and HAVING
- CASE for customer segmentation
- Simple subqueries

5. Dashboard in Power BI

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



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.