Customer Shopping Behavior Analysis

Project Overview

This initiative examines customer purchasing behavior by analyzing transactional data encompassing 3,900 transactions spread across multiple product categories. The objective is to extract meaningful insights regarding purchasing trends, distinct customer groups, item preferences, and subscription engagement to inform decision making strategies.

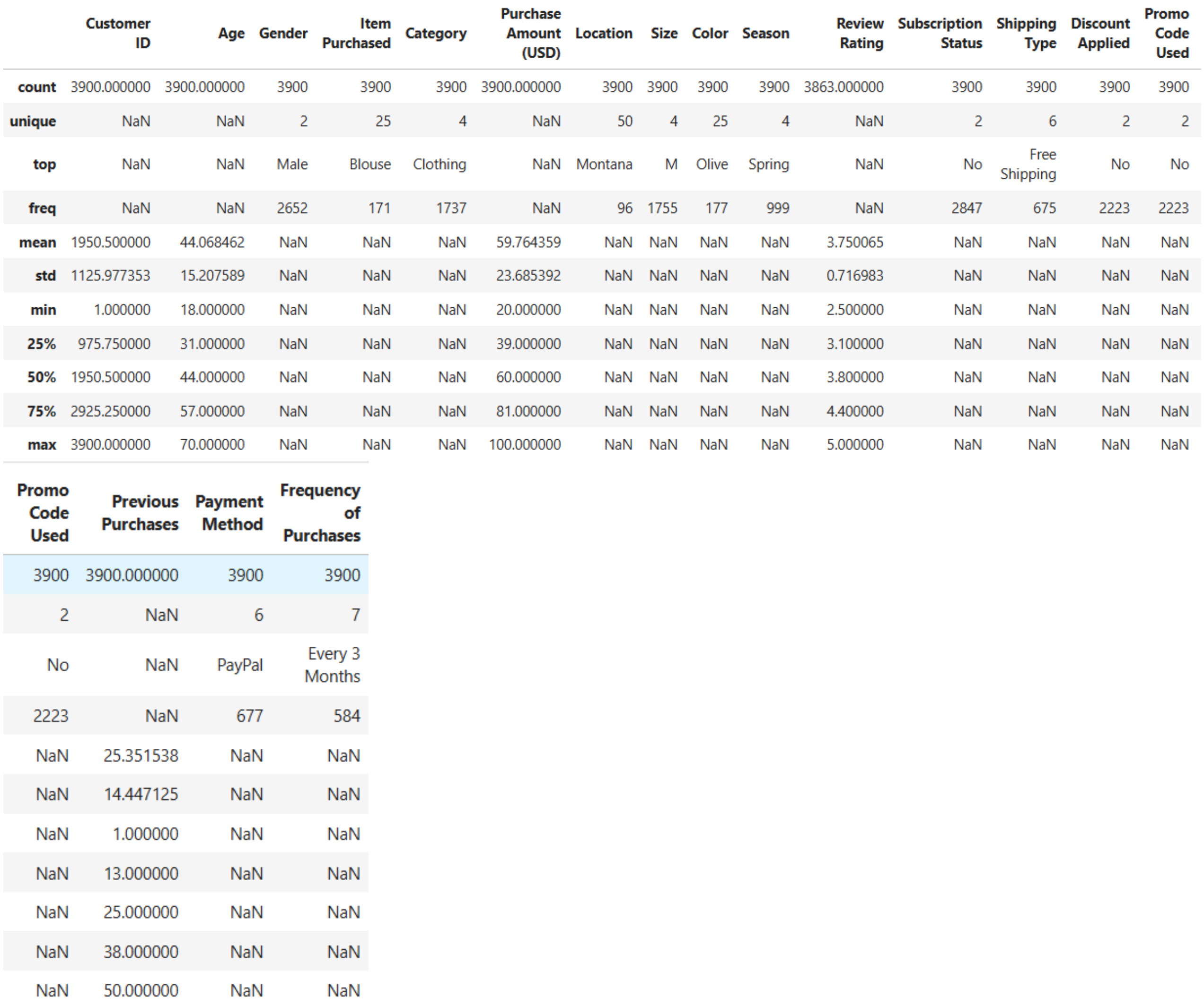
Dataset Summary

- **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

Exploratory Data Analysis using Python

The foundation involved data organization and refinement in Python:

- **Data Loading:** The dataset was brought into the environment utilizing pandas.
- **Initial Exploration:** Structural assessment was performed via df.info() and statistical summaries were generated using df.describe().



	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

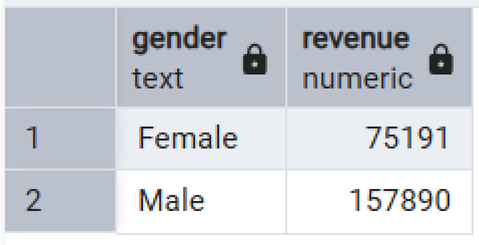
- **Missing Data Handling:** Null values were examined throughout the dataset with missing entries in the Review Rating column being filled using the median rating corresponding to each product category.
- **Column Standardization:** All column names were reformatted into snake case to enhance clarity and consistency in documentation.
- **Feature Engineering:**
  - An age_group column was constructed by segmenting customer ages into groups.
  - A purchase_frequency_days column was generated from transaction timing information.
- **Data Consistency Check:** An examination was conducted to determine whether discount_applied and promo_code_used contained overlapping information; promo_code_used was subsequently removed.

- **Database Integration:** The processed DataFrame was transferred from Python into a PostgreSQL database environment to enable SQL based examination.

Data Analysis using SQL (Business Transactions)

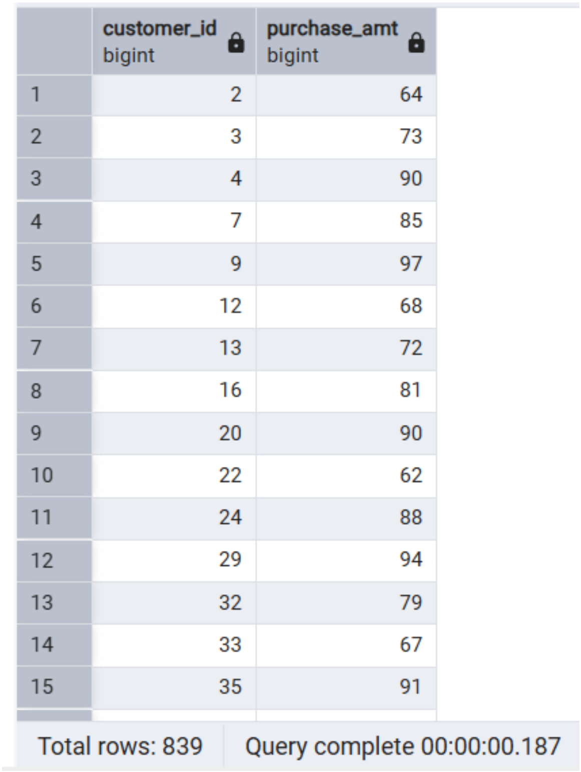
Systematic inquiries were executed in PostgreSQL to address key operational questions:

1. **Revenue by Gender -** Examination of total revenue generation across male and female customer segments.



	gender text	revenue numeric
1	Female	75191
2	Male	157890

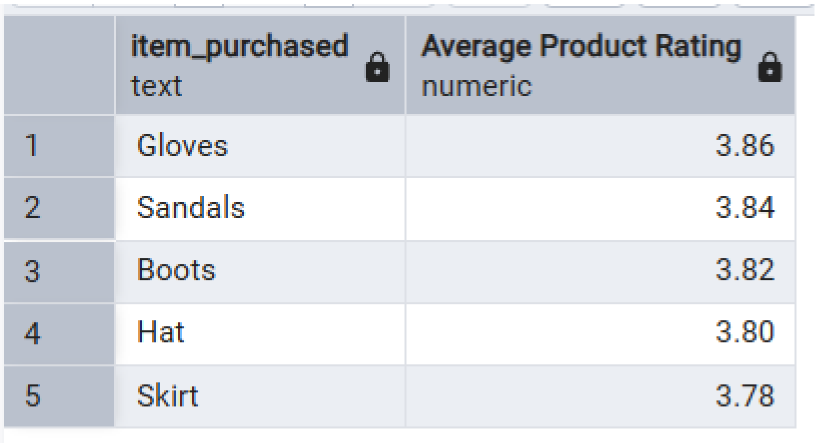
2. **High-Spending Discount Users -** Identification of discount-using customers whose expenditure exceeded the average transaction value.



	customer_id bigint	purchase_amt 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
10	22	62
11	24	88
12	29	94
13	32	79
14	33	67
15	35	91

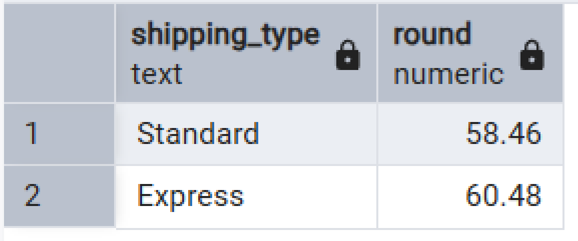
Total rows: 839 Query complete 00:00:00.187

3. **Top 5 Products by Rating -** Discovery of items with the most favorable average review scores.



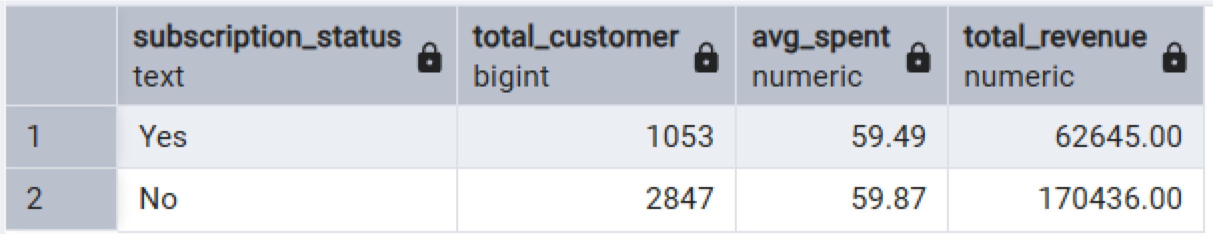
	item_purchased text	Average Product Rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

4. **Shipping Type Comparison -** Assessment of average purchase values between Standard and Express delivery methods.



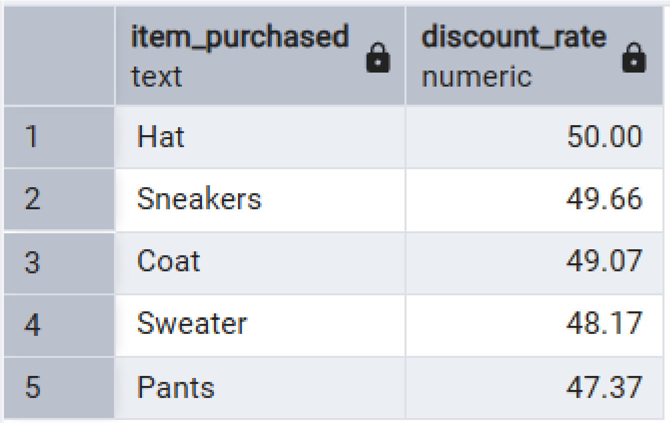
	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. **Subscribers vs. Non-Subscribers -** Analysis of average spending and aggregate revenue variation based on subscription enrollment.



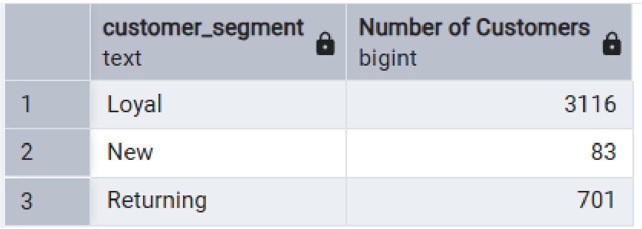
	subscription_status text	total_customer bigint	avg_spent numeric	total_revenue numeric
1	Yes	1053	59.49	62645.00
2	No	2847	59.87	170436.00

6. **Discount-Dependent Products -** Detection of 5 items demonstrating the greatest reliance on discounted transactions.



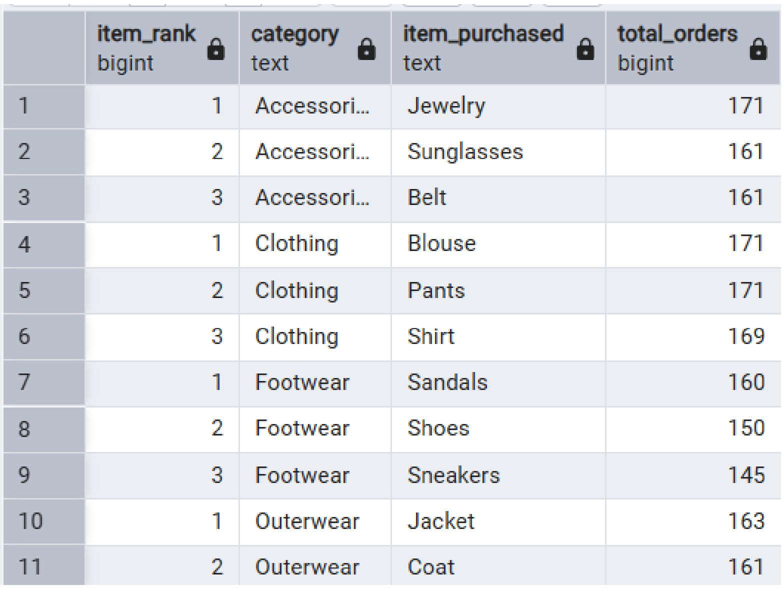
	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

7. **Customer Segmentation -** Categorization of customers into New, Returning, and Loyal classifications determined by purchase frequency.



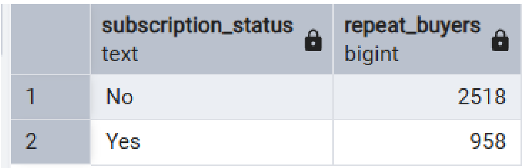
	customer_segment text	Number of Customers bigint
1	Loyal	3116
2	New	83
3	Returning	701

8. **Top 3 Products per Category -** Compilation of best-performing items within each merchandise category.



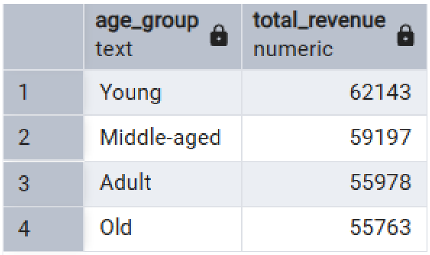
	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	Belt	161
4	1	Clothing	Blouse	171
5	2	Clothing	Pants	171
6	3	Clothing	Shirt	169
7	1	Footwear	Sandals	160
8	2	Footwear	Shoes	150
9	3	Footwear	Sneakers	145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

9. **Repeat Buyers & Subscriptions -** Investigation of whether customers exceeding 5 purchases demonstrate higher subscription propensity.



	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

10. **Revenue by Age Group -** Calculation of revenue contribution from each age segment.



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