CUSTOMER SHOPPING ANALYSIS

PROJECT REPORT

Business Problem

A retail company wants to understand customer buying behavior to boost sales and loyalty. The task is to analyze shopping data to identify key trends, purchase drivers, and ways to improve customer engagement and marketing strategies.

Project Overview

This project examines customer shopping behavior using 3,900 transaction records. The objective is to identify spending trends, customer segments, product choices, and subscription patterns to support strategic business decisions.

Dataset Summary

- **Size:** 3,900 rows, 18 columns
- **Key Features:**
    - Demographics: Age, Gender, Location, Subscription Status
    - Purchase info: Item, Category, Amount, Season, Size, Color
    - Behavior: Discounts, Promo Codes, Previous Purchases, Frequency, Review Rating, Shipping Type
- **Missing Values:** 37 missing entries in the Review Rating column

Exploratory Data Analysis (Python)

Prepared and cleaned the dataset using Python:

- **Data Loading:** Imported the dataset with pandas.
- **Initial Exploration:** Used df.info() and df.describe() to review structure and summary stats.

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

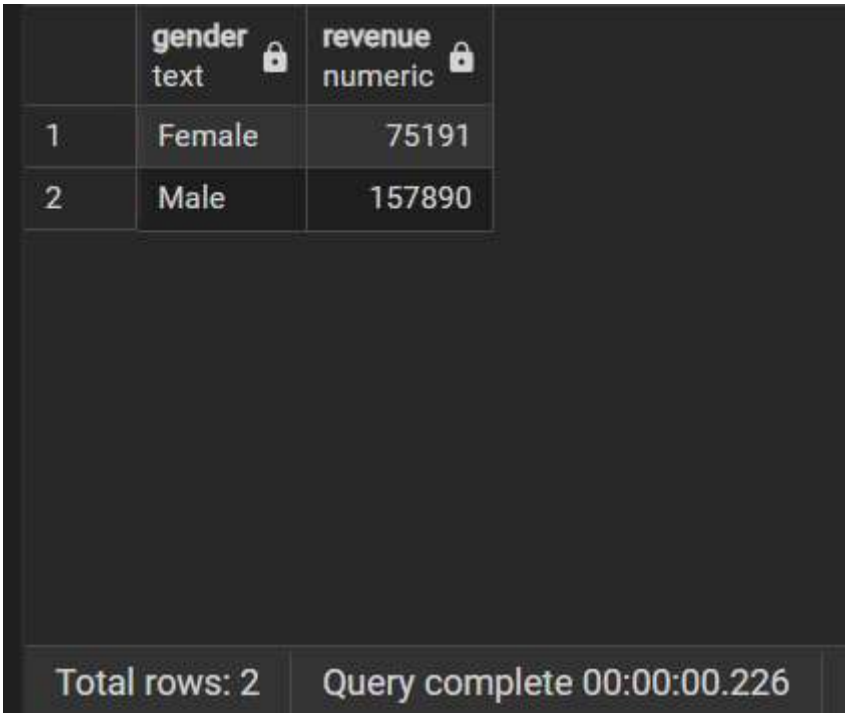
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             3900 non-null   int64
 1   Age                     3900 non-null   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
```

- **Missing Values:** Identified nulls and filled missing Review Rating values using the median rating per product category.
- **Column Standardization:** Converted column names to snake_case for clarity.
- **Feature Engineering:**
    - Added an age_group variable by binning ages.
    - Created purchase_frequency_days from purchase timestamps.
- **Data Consistency:** Checked overlap between discount_applied and promo_code_used, then removed the redundant promo column.
- **Database Integration:** Connected to PostgreSQL and loaded the cleaned DataFrame for SQL-based analysis.

Data Analysis using SQL

Using PostgreSQL,I ran structured queries to answer major business questions, including:

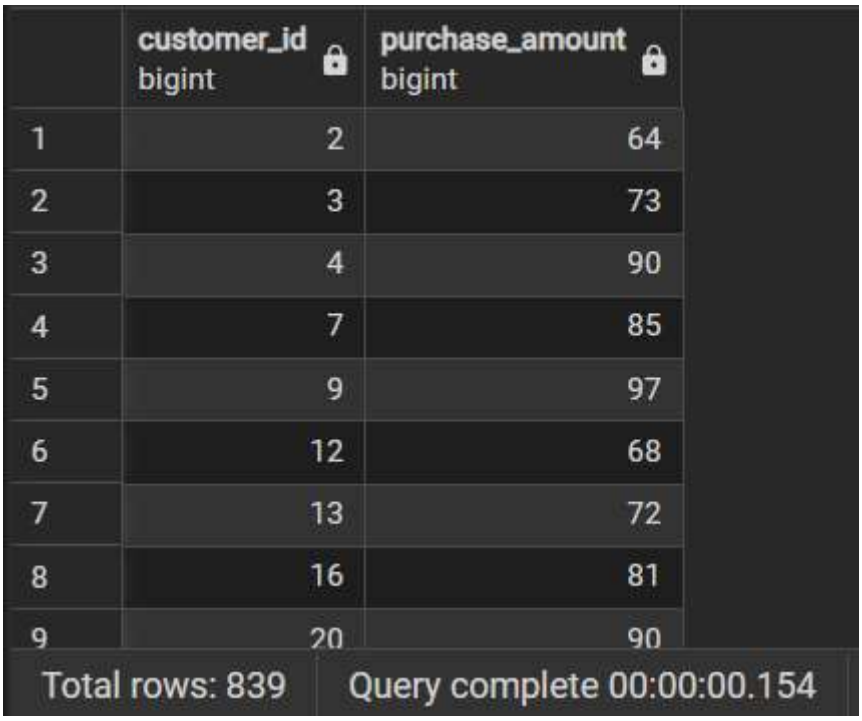
I. **Revenue by Gender** – Calculated and compared total revenue generated by male vs. female customers.



	gender text	revenue numeric
1	Female	75191
2	Male	157890

Total rows: 2 | Query complete 00:00:00.226

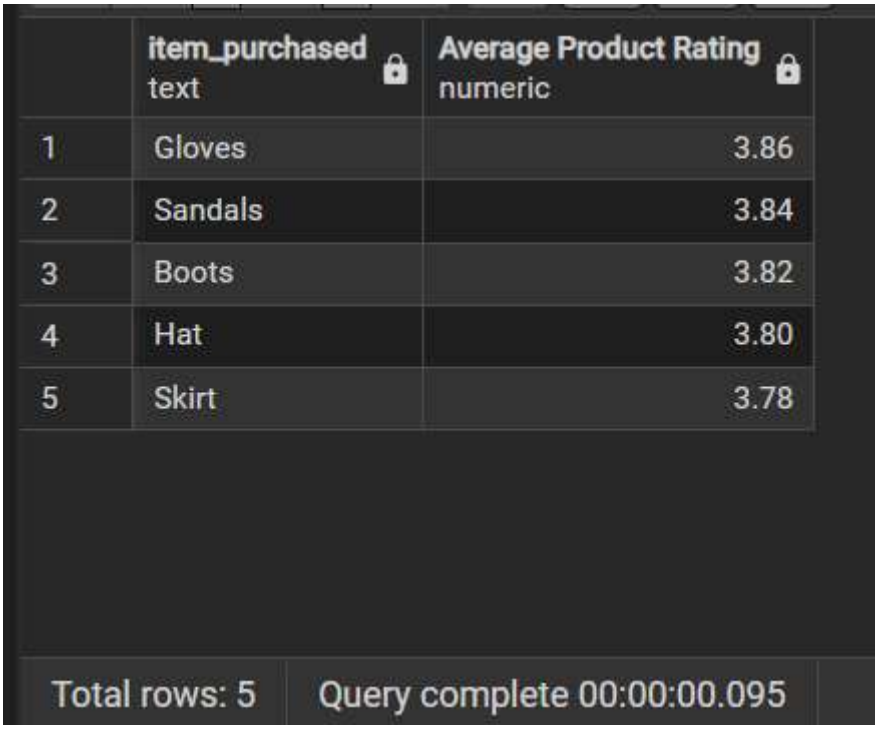
II. **High-Spend Discount Users** – Identified customers who used a discount but still spent more than the overall average purchase amount.



	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
8	16	81
9	20	90

Total rows: 839 Query complete 00:00:00.154

III. **Top-Rated Products** – Found the top 5 products with the highest average review ratings.



	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

Total rows: 5 Query complete 00:00:00.095

IV. **Shipping Type vs. Spend** – Compared average purchase amounts between Standard and Express shipping orders.

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

Total rows: 2 | Query complete 00:00:00.221

V. **Subscriber vs. Non-Subscriber Spend** – Compared customer count, average spend, and total revenue between subscribed and non-subscribed customers.

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

Total rows: 2 | Query complete 00:00:00.149

VI. **Discount-Intensive Products** – Identified the 5 products with the highest percentage of purchases made with discounts applied.

	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

Total rows: 5 | Query complete 00:00:00.156

VII. **Customer Segmentation** – Segmented customers into New, Returning, and Loyal based on previous purchases and counted how many fall into each group.

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

Total rows: 3 | Query complete 00:00:00.084

VIII. **Top Products per Category** – Retrieved the top 3 most purchased products within each product 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

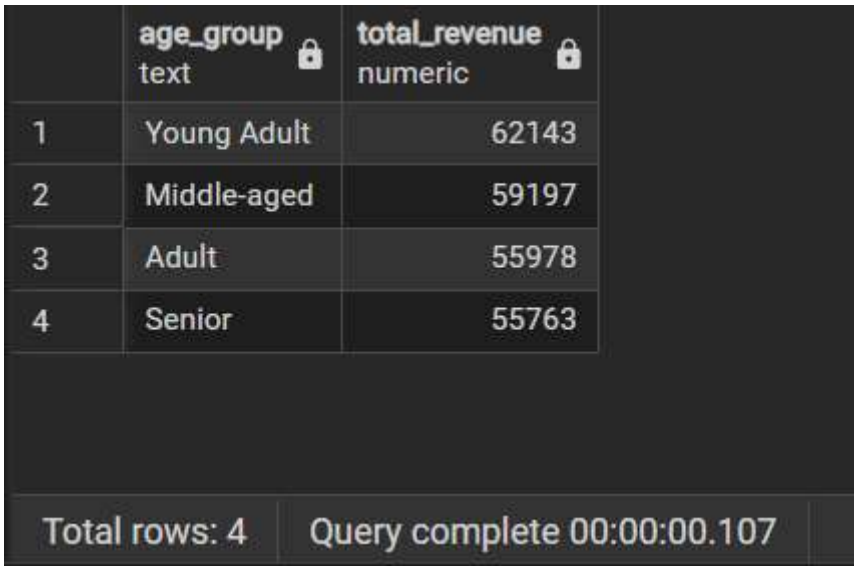
Total rows: 11 Query complete 00:00:00.117

IX. **Repeat Buyers & Subscription** – Checked whether repeat buyers (more than 5 past purchases) are more likely to be subscribers by comparing their subscription status.

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

Total rows: 2 Query complete 00:00:00.094

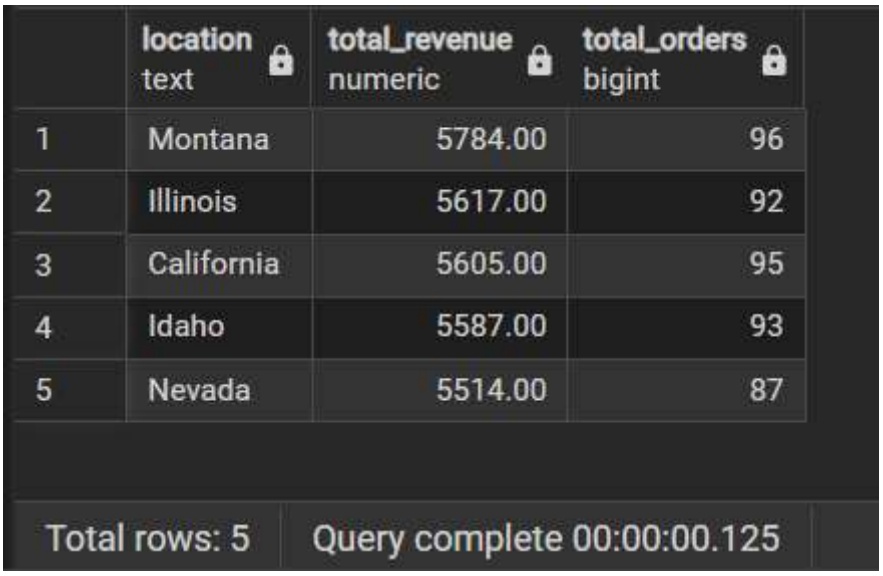
X. **Revenue by Age Group** – Calculated total revenue contributed by each age group and ranked them.



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

Total rows: 4 | Query complete 00:00:00.107

XI. **Top Revenue-Generating Locations** – Identified the 5 locations with the highest total revenue and number of orders.



	location text	total_revenue numeric	total_orders bigint
1	Montana	5784.00	96
2	Illinois	5617.00	92
3	California	5605.00	95
4	Idaho	5587.00	93
5	Nevada	5514.00	87

Total rows: 5 | Query complete 00:00:00.125

XII. **Category-Level Performance** – Measured average purchase amount, total revenue, and average review rating for each product category.

	category text	avg_purchase_amount numeric	total_revenue numeric	avg_review_rating numeric
1	Clothing	60.03	104264.00	3.72
2	Accessori...	59.84	74200.00	3.77
3	Footwear	60.26	36093.00	3.79
4	Outerwear	57.17	18524.00	3.75

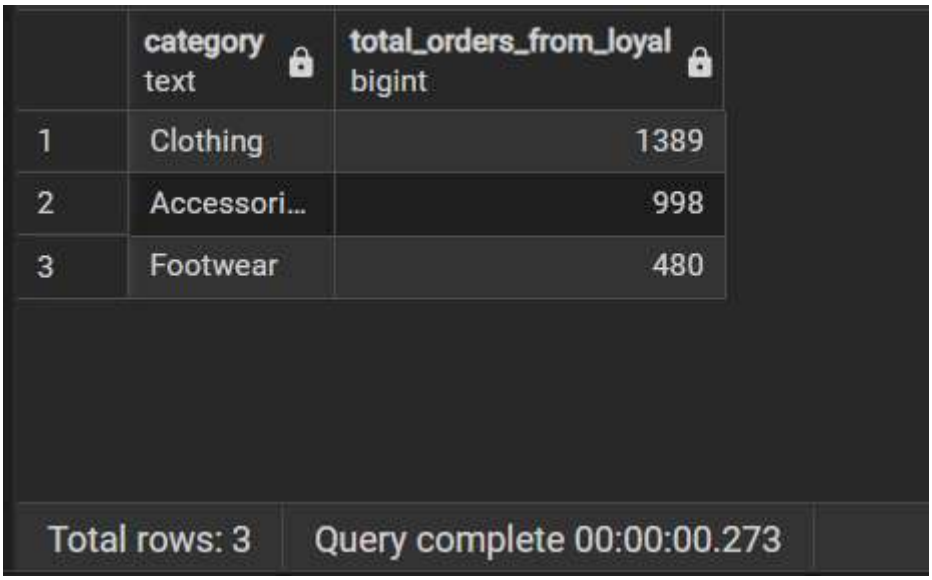
Total rows: 4 Query complete 00:00:00.168

XIII. **Impact of Discounts on Spend & Ratings** – Compared order counts, average purchase amount, and average review rating between discounted and non-discounted orders.

	discount_applied text	total_orders bigint	avg_purchase_amount numeric	avg_review_rating numeric
1	No	2223	60.13	3.76
2	Yes	1677	59.28	3.74

Total rows: 2 Query complete 00:00:00.128

XIV. **Loyal Customers' Favorite Categories** – Among loyal customers (more than 10 previous purchases), found the 3 most frequently purchased categories.



	category text	total_orders_from_loyal bigint
1	Clothing	1389
2	Accessori...	998
3	Footwear	480

Total rows: 3 | Query complete 00:00:00.273

XV. **Payment Methods for High-Value Orders** – Analyzed how payment methods are distributed among high-value orders (above average purchase amount) and their percentage share.

	payment_method text	total_orders bigint	percentage_of_high_value_orders numeric
1	Credit Card	339	17.27
2	Cash	338	17.22
3	Debit Card	334	17.01
4	PayPal	332	16.91
5	Bank Transfer	310	15.79
6	Venmo	310	15.79

Total rows: 6 | Query complete 00:00:00.124

Power BI Dashboard

Created an interactive Power BI dashboard to visually present the insights from the analysis. It highlights key trends in customer spending, product preferences, discounts, shipping behavior, and customer segments, enabling stakeholders to quickly interpret patterns and make data-driven decisions.

Dashboard Insights

- Majority of customers are **Loyal (≈80%)**, while New customers are very few.
- **Average spend is $59.76**, with **43% of orders using discounts** → customers are price-sensitive.
- **Clothing** and **Accessories** are the top revenue-generating categories.
- **Jewelry, Blouse, and Pants** are the most purchased products.
- **Middle-aged and Adult loyal customers** spend the most on average.
- All payment methods perform similarly, with PayPal and Credit Card slightly higher.

Business Recommendations

- **Increase Subscriptions:** Offer exclusive perks and personalized deals to boost sign-ups.
- **Strengthen Loyalty Programs:** Reward repeat buyers to grow the Loyal customer base.
- **Optimize Discounts:** Use targeted promotions to increase sales without hurting margins.
- **Promote Best Products:** Highlight top-rated and best-selling items in marketing campaigns.
- **Target High-Value Segments:** Focus ads on high-spending age groups, key locations, and Express-shipping users.
- **Improve Low-Rated Products:** Track low reviews and prioritize quality fixes.
- **Enable Smart Cross-Selling:** Suggest complementary items based on shopping patterns.