Customer Shopping Behavior Analysis

1. Project Overview

This project analyses 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

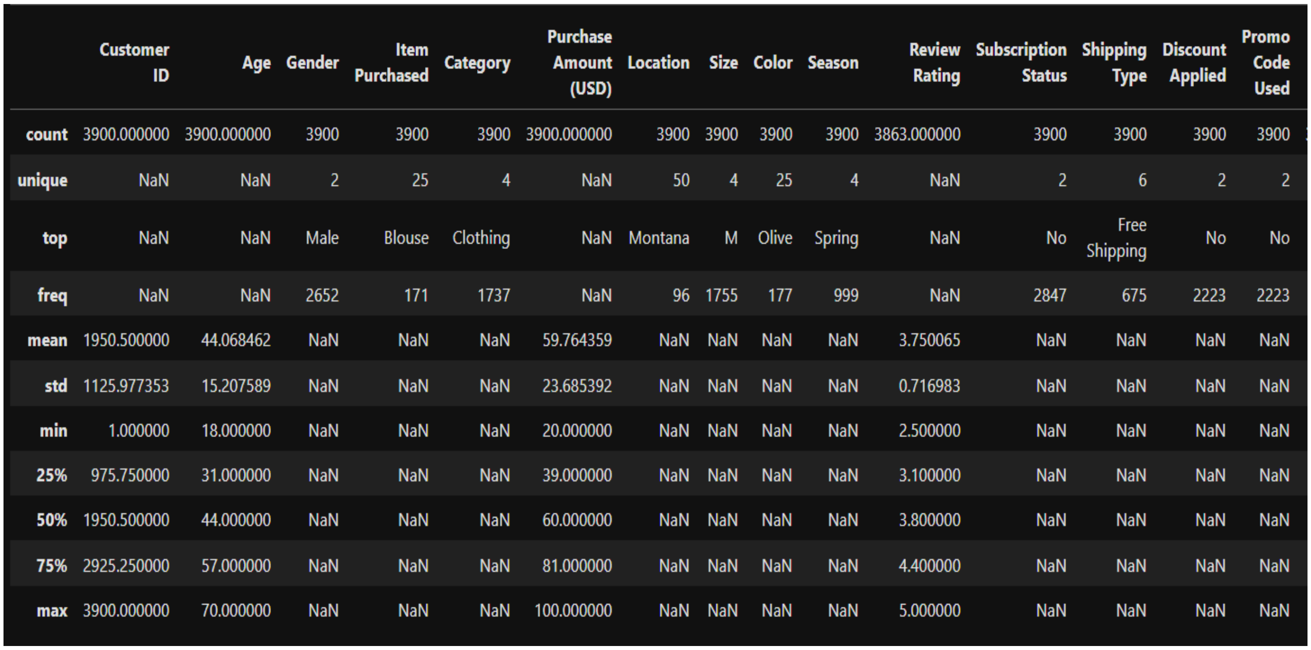
Rows: 3,900 - Columns: 18 – Key Features:

- ○ 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

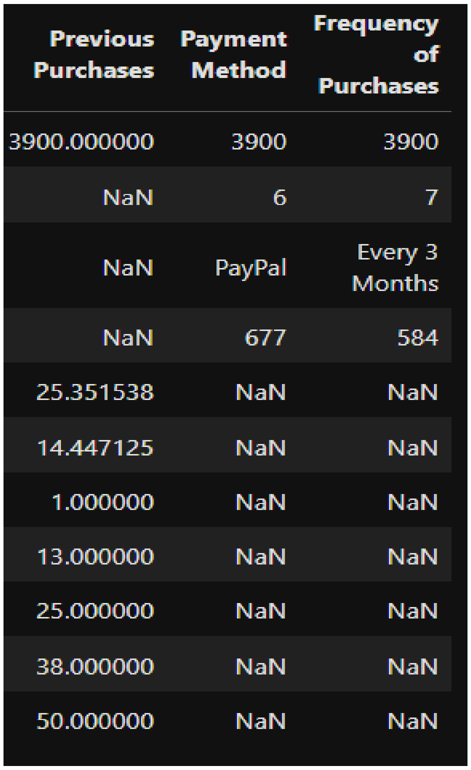
3. Exploratory Data Analysis using Python.

We began with data preparation and cleaning in Python:

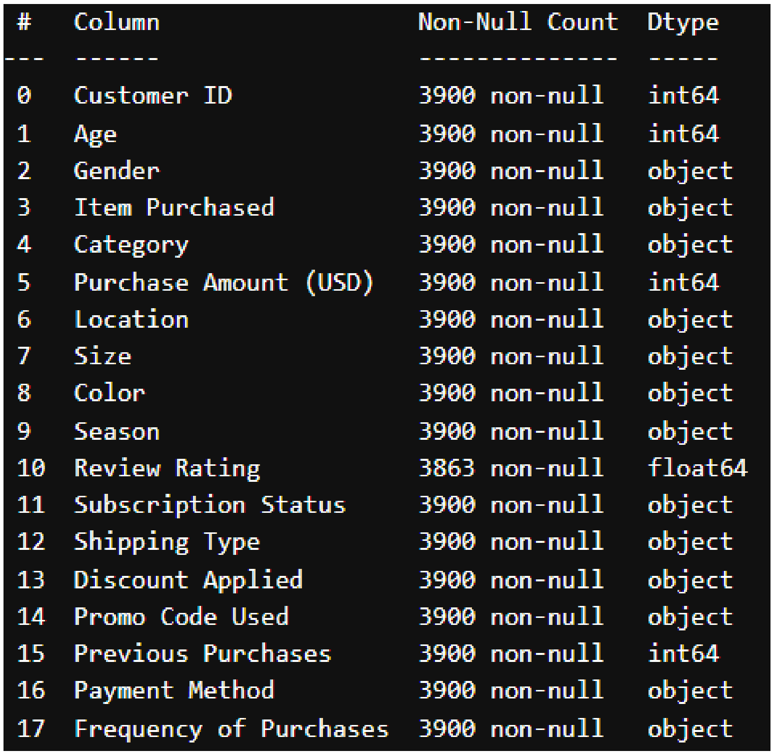
- **Data Loading:** Imported dataset using **Pandas**.
- **Initial Exploration:** Used **df.info()** to check structure and **.describe()** for summary statistic.



	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



Previous Purchases	Payment Method	Frequency of Purchases
3900.000000	3900	3900
NaN	6	7
NaN	PayPal	Every 3 Months
NaN	677	584
25.351538	NaN	NaN
14.447125	NaN	NaN
1.000000	NaN	NaN
13.000000	NaN	NaN
25.000000	NaN	NaN
38.000000	NaN	NaN
50.000000	NaN	NaN



```
#   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
```

- **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 **frequency_purchased_days** column from purchase data.

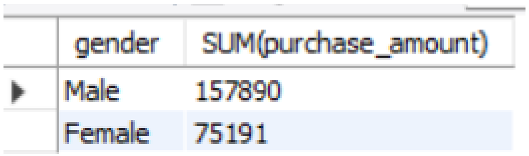
- **Data Consistency Check:** Verified if **discount_applied** and **promo_code_used** were redundant; dropped **promo_code_used.**

- **Database Integration:** Connected Python script to MySQL and loaded the cleaned Data Frame into the database for SQL analysis.

4. Data Analysis using SQL.

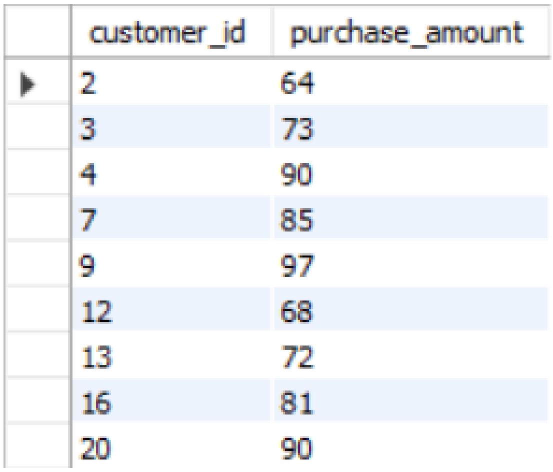
We performed structured analysis in MySQL to answer key business questions:

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



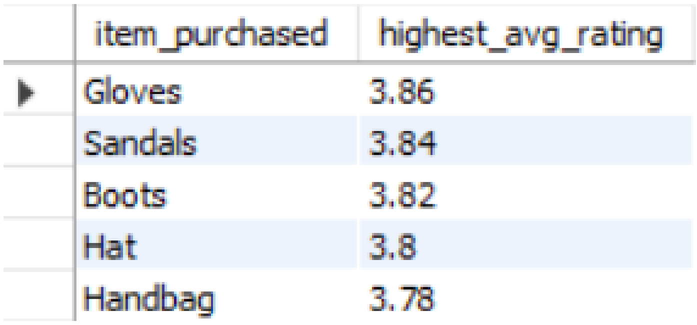
	gender	SUM(purchase_amount)
▶	Male	157890
	Female	75191

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



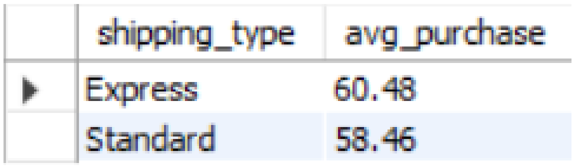
	customer_id	purchase_amount
▶	2	64
	3	73
	4	90
	7	85
	9	97
	12	68
	13	72
	16	81
	20	90

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



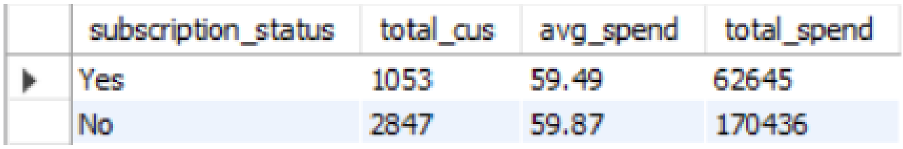
	item_purchased	highest_avg_rating
▶	Gloves	3.86
	Sandals	3.84
	Boots	3.82
	Hat	3.8
	Handbag	3.78

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



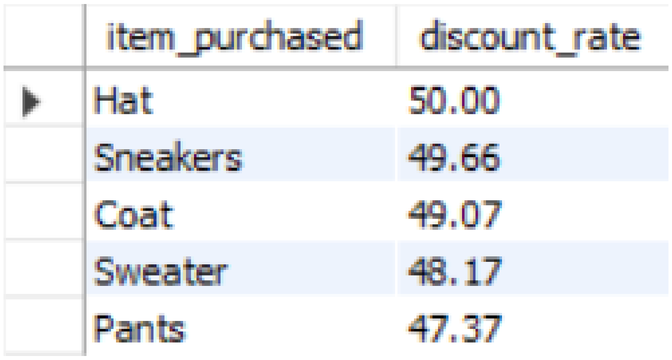
	shipping_type	avg_purchase
▶	Express	60.48
	Standard	58.46

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



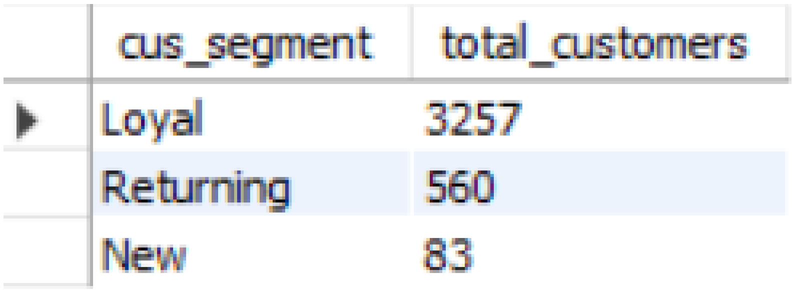
	subscription_status	total_cus	avg_spend	total_spend
▶	Yes	1053	59.49	62645
	No	2847	59.87	170436

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



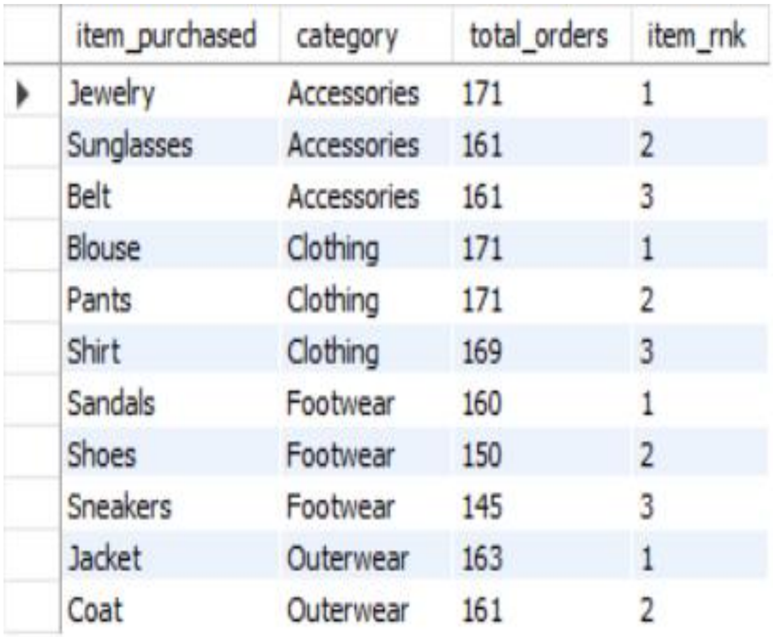
	item_purchased	discount_rate
▶	Hat	50.00
	Sneakers	49.66
	Coat	49.07
	Sweater	48.17
	Pants	47.37

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



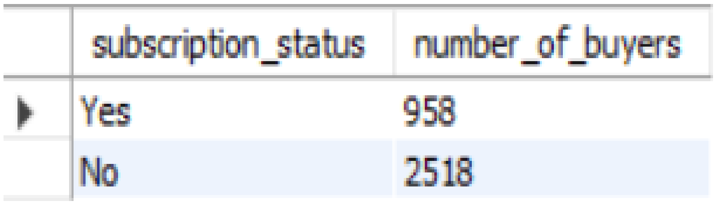
	cus_segment	total_customers
▶	Loyal	3257
	Returning	560
	New	83

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



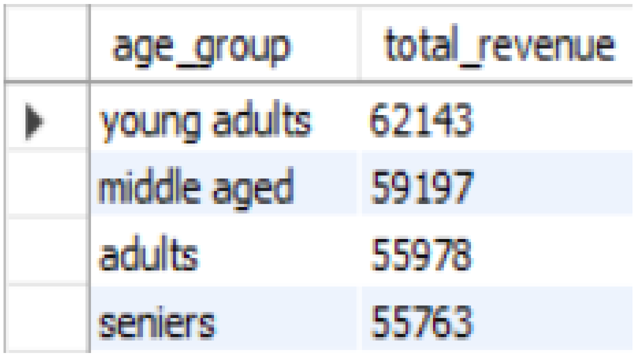
item_purchased	category	total_orders	item_rnk
Jewelry	Accessories	171	1
Sunglasses	Accessories	161	2
Belt	Accessories	161	3
Blouse	Clothing	171	1
Pants	Clothing	171	2
Shirt	Clothing	169	3
Sandals	Footwear	160	1
Shoes	Footwear	150	2
Sneakers	Footwear	145	3
Jacket	Outerwear	163	1
Coat	Outerwear	161	2

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



subscription_status	number_of_buyers
Yes	958
No	2518

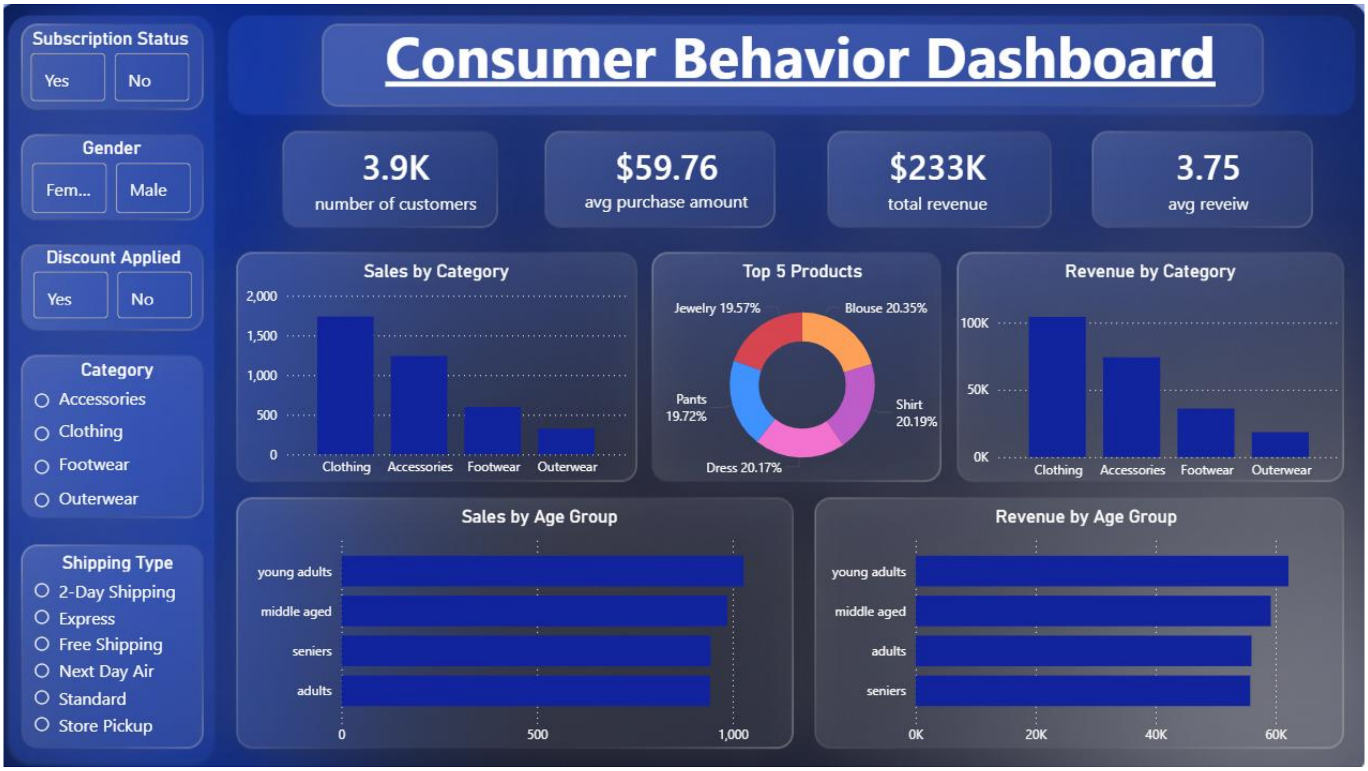
10. **Revenue by Age Group** – Calculated total revenue contribution of each age group.



age_group	total_revenue
young adults	62143
middle aged	59197
adults	55978
seniers	55763

5. Dashboard in Power BI

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



6. Business Recommendations

- **Subscriptions Optimization** – Introduce exclusive benefits and savings for subscribers.
- **Customer Loyalty Integration** – Reward repeat buyers to move them into the “Loyal” segment.
- **Review Discount Policy** – Audit promotional discounting to ensure customer acquisition remains profitable without eroding gross margins.
- **Product Placement** – Highlight top-rated and best-selling products in campaigns.
- **Targeted Marketing** – Focus efforts on high-revenue age groups and express-shipping users.