Customer Shopping Behavior Analysis

1. Project Overview

This project studies **customer shopping behavior** based on **3,900 purchase records** from different product categories.
The aim is to understand **spending habits, product preferences, customer types, and subscription trends** to help businesses make better marketing and sales decisions.

---

2. Dataset Summary

- **Rows:** 3,900
- **Columns:** 18
- **File Type:** CSV

Main Features:

- **Customer Details:** Age, Gender, Location, Subscription Status
- **Purchase Info:** Item Purchased, Category, Amount, Season, Size, Color
- **Shopping Behavior:** Discount Applied, Promo Code Used, Review Rating, Shipping Type, Previous Purchases
- **Missing Data:** 37 missing values in the *Review Rating* column

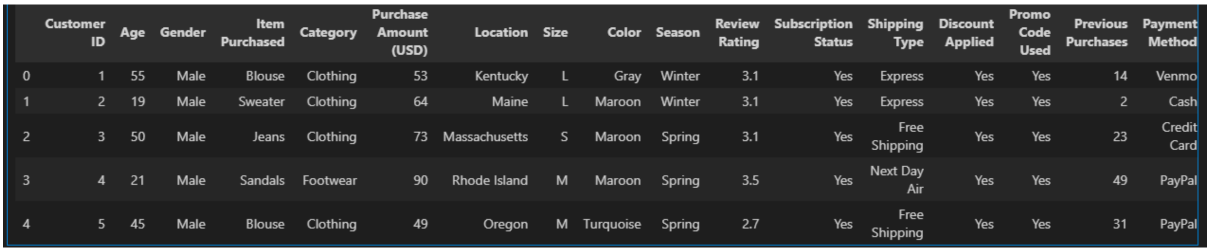
---

3. Data Cleaning & Analysis using Python

All data preparation was done in **Python (VS Code)** using **Pandas and NumPy**.

Steps:

1. **Loaded Data:** Imported the CSV file and checked structure using .info() and .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
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal

2. **Handled Missing Values:** Filled missing ratings with the median value for each product category.

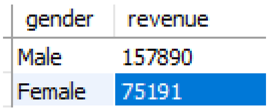
3. **Renamed Columns:** Changed column names to lowercase with underscores for consistency.

4. **Feature Engineering:**
   - Created an **age_group** column from customer ages.
   - Calculated **purchase_frequency_days** based on previous purchases.
5. **Removed Duplicates & Unused Columns:** Dropped the promo_code_used column as it was similar to discount_applied.
6. **Database Connection:** Connected Python to **MySQL** and uploaded the cleaned data for SQL analysis.

4. Data Analysis using SQL

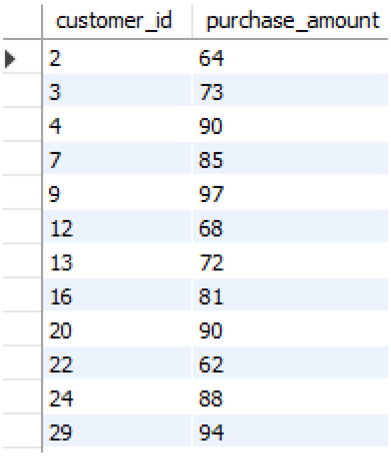
SQL was used to answer key business questions:

1. **Revenue by Gender** – Compared total sales from male vs female customers.



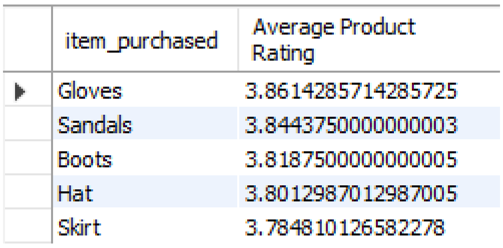
gender	revenue
Male	157890
Female	75191

2. **High-Spending Discount Users** – Found customers who used discounts but still spent more than the average.



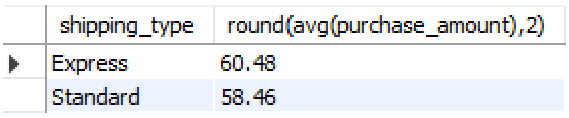
customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90
22	62
24	88
29	94

3. **Top 5 Rated Products** – Identified products with the best average review scores.



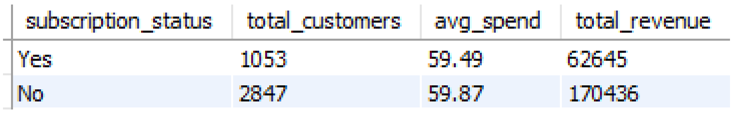
item_purchased	Average Product Rating
Gloves	3.8614285714285725
Sandals	3.8443750000000003
Boots	3.8187500000000005
Hat	3.8012987012987005
Skirt	3.7848101265822278

4. **Shipping Type Comparison** – Checked average order value for Standard vs Express shipping.



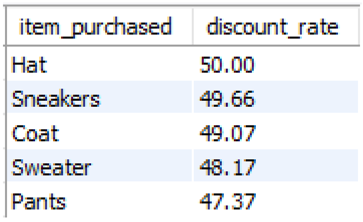
shipping_type	round(avg(purchase_amount),2)
Express	60.48
Standard	58.46

5. **Subscribers vs Non-Subscribers** – Compared average spending and total revenue.



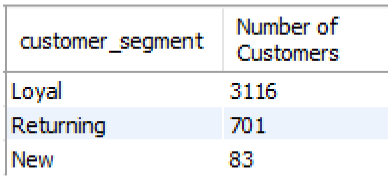
subscription_status	total_customers	avg_spend	total_revenue
Yes	1053	59.49	62645
No	2847	59.87	170436

6. **Discount-Dependent Products** – Found products mostly bought with discounts.



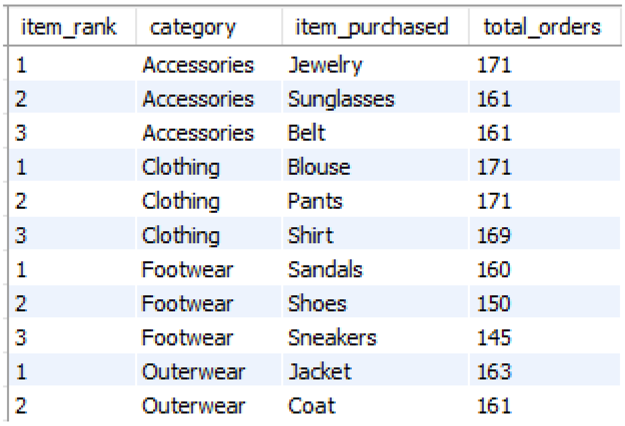
item_purchased	discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

7. **Customer Segments** – Grouped customers as *New, Returning,* and *Loyal* based on purchase frequency.



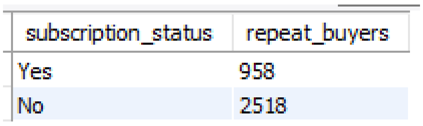
customer_segment	Number of Customers
Loyal	3116
Returning	701
New	83

8. **Top Products by Category** – Listed the most popular items in each category.



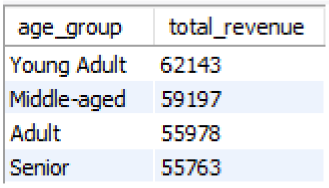
item_rank	category	item_purchased	total_orders
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
1	Clothing	Blouse	171
2	Clothing	Pants	171
3	Clothing	Shirt	169
1	Footwear	Sandals	160
2	Footwear	Shoes	150
3	Footwear	Sneakers	145
1	Outerwear	Jacket	163
2	Outerwear	Coat	161

9. **Repeat Buyers & Subscriptions** – Checked if frequent buyers are more likely to subscribe.



subscription_status	repeat_buyers
Yes	958
No	2518

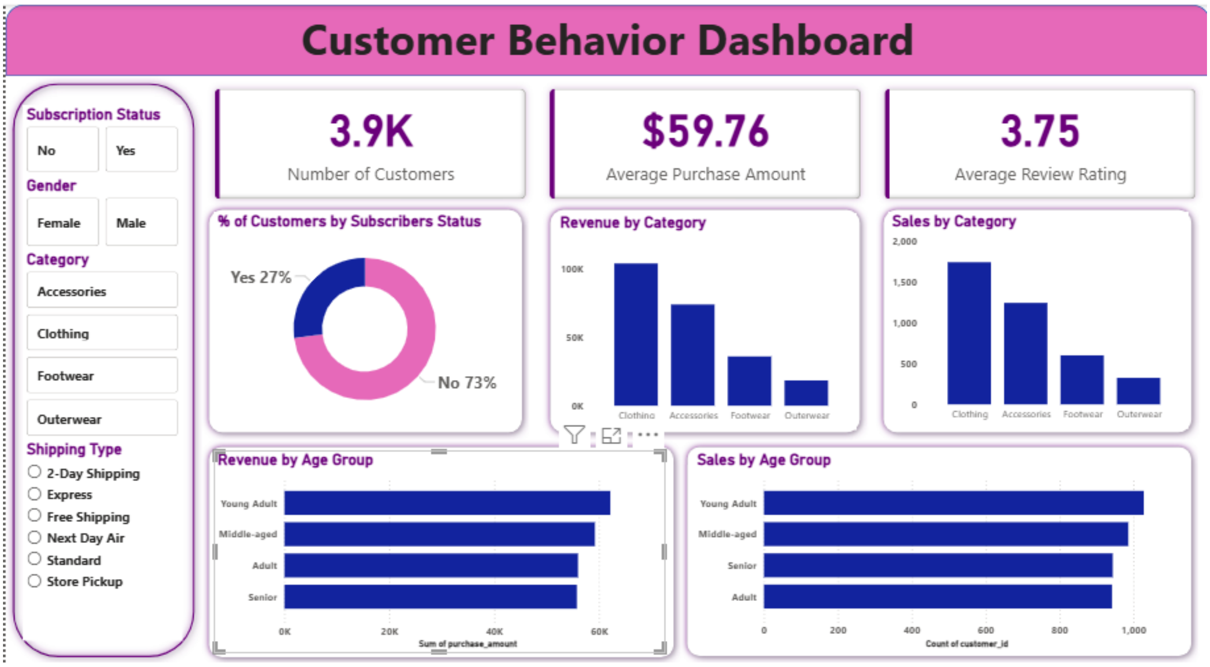
10. **Revenue by Age Group** – Calculated contribution from each age group.



age_group	total_revenue
Young Adult	62143
Middle-aged	59197
Adult	55978
Senior	55763

5. Power BI Dashboard

An **interactive Power BI dashboard** was created to visualize insights clearly.


Customer Behavior Dashboard

It includes:

- Total Sales and Revenue
- Top Customers and Categories
- Revenue by Age, Gender, and Region
- Discount Impact and Subscription Trends

6. Business Insights & Recommendations

1. **Increase Subscriptions** – Offer special discounts or early access to subscribers.
2. **Loyalty Program** – Reward repeat buyers to improve retention.
3. **Balance Discounts** – Use smart discounts to boost sales without reducing profit.
4. **Highlight Best Products** – Promote high-rated and fast-selling products.
5. **Target Marketing** – Focus on top-spending age groups and Express shipping users.

7. Tools Used

- **Python:** Pandas, NumPy, Matplotlib, Seaborn
- **Database:** MySQL
- **Visualization:** Power BI
- **Environment:** VS Code

This project shows end-to-end data analytics skills — from cleaning and analysis in Python, querying with SQL, to presenting insights visually in Power BI.