Problem Statement:

A big retail company wants to understand how customers shop so they can increase sales, make customers happier, and keep them coming back.

Recently, they have noticed that buying patterns are changing based on things like age, product types, and whether customers shop online or in-store. They also want to know what really influences customers to buy — for example, discounts, product reviews, seasons, or preferred payment methods.

The task is to study the customer shopping data and have to answer for this question

How can the company use customer shopping data to find trends, improve customer engagement, and make better marketing and product decisions?

Process of Project:

Data Cleaning & Preparation:
By using Python to clean the raw data, fix errors, handle missing values, and prepare it properly so it can be analyze easily.

Data Analysis:
And by using SQL, we have to Organize the cleaned data into tables, create structured business transaction data, and write SQL queries to find insights about customer groups, loyalty patterns, and factors that influence purchases.

Dashboard & Insights:
Creating an interactive Power BI dashboard that clearly shows important trends and patterns. This will help managers understand the data and make better decisions.

Project Report & Presentation:
Preparing a clear and simple report explaining with findings and business suggestions. Also, creating a presentation that visually explains the key insights and practical recommendations.

Customer Shopping Behavior Analysis

Project Overview:

This project focuses on analyzing customer shopping behavior using transaction data from 3,900 purchases across different product categories.

Objective is to understand:

- How customers spend
- Which products they prefer
- How different customer groups behave
- Whether subscriptions influence buying patterns

These insights will help the company make better business decisions related to marketing, product strategy, and customer engagement.

2. Dataset Summary

Total Records (Rows): 3,900

Total Features (Columns): 18

Key Information in the Dataset:

1. Customer Details: Age, Gender, Location, Subscription Status

2. Purchase Information: Item Purchased, Product Category, Purchase Amount, Season, Size, Color

3. Shopping Behavior Details: Discount Applied, Promo Code Used, Previous Purchases, Purchases Frequency, Review Rating, Shipping Type

Exploratory Data Analysis (EDA) Using Python:

I started the project by cleaning and preparing the data using Python. Below is a simple explanation of the steps I performed:

1. **Data Loading:** Imported the dataset using the pandas library.

   ```
   dataset = pd.read_csv("customer_shopping_behavior.csv")
   ```

2. **Initial Exploration:**

   - Used df.info() to understand the structure of the dataset (columns, data types, null values).
   - Used describe() to view summary statistics like mean, minimum, maximum, etc.

```
In [10]: dataset.describe(include = "all")
```

Out[10]:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN

3. Handling Missing Data:

- Checked for missing (null) values.
- Found 37 missing values in the Review Rating column.
- Filled the missing values using the median review rating of each product category to maintain accuracy.

4. Column Standardization:

- Renamed column names into snake_case format (for example: Purchase Amount → purchase_amount) to improve readability and maintain consistency.

5. Feature Engineering:

- Created a new column called age_group by grouping customer ages into categories (e.g., 18–25, 26–35, etc.).
- Created a new column called purchase_frequency_days to better represent customer purchase behavior in numeric form.

```
In [15]: # creating a column age_group
         groups = ["Young Adult","Adult","Middle_age","Senior"]
         dataset["age_group"] = pd.qcut(dataset["age"],q=4,labels = groups)

In [16]: dataset[["age","age_group"]].head(9)
```

Out[16]:

	age	age_group
0	55	Middle_age
1	19	Young Adult
2	50	Middle_age
3	21	Young Adult
4	45	Middle_age
5	46	Middle_age
6	63	Senior
7	27	Young Adult
8	26	Young Adult

6. **Data Consistency Check:**

   - Checked whether discount_applied and promo_code_used were providing the same information.
   - Since both were similar, I removed the promo_code_used column to avoid duplication.

7. **Database Integration:**

   - Connected the Python script to MySQL Database.
   - Uploaded the cleaned and prepared dataset into the database for further analysis using SQL.

Data Analysis Using MySQL:

After cleaning the data, I used MySQL Database to perform structured analysis and answer important business questions. Below is a simple explanation of what we analyzed:

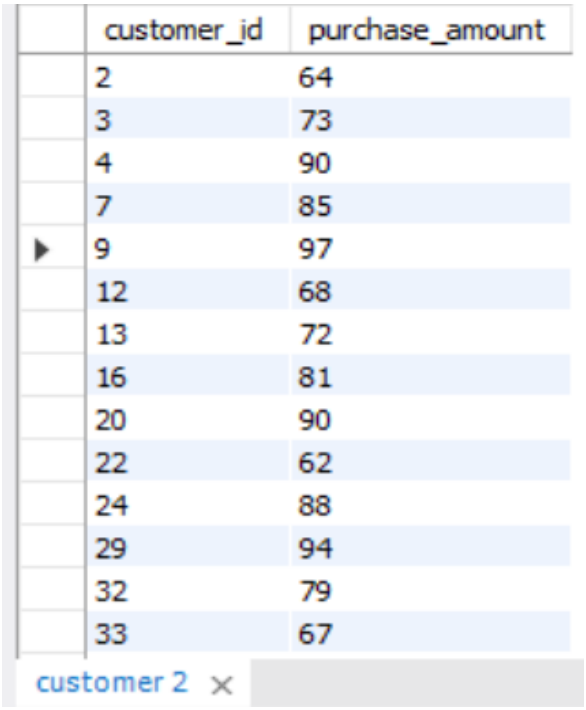
1. **Revenue by Gender**

   Compared the total revenue generated by male and female customers to understand which group contributes more to sales.

	gender	total_revenue
▶	Male	157890
	Female	75191

2. High-Spending Discount Users

Identified customers who used discounts but still spent more than the average purchase amount. This helps find valuable customers who respond well to offers.

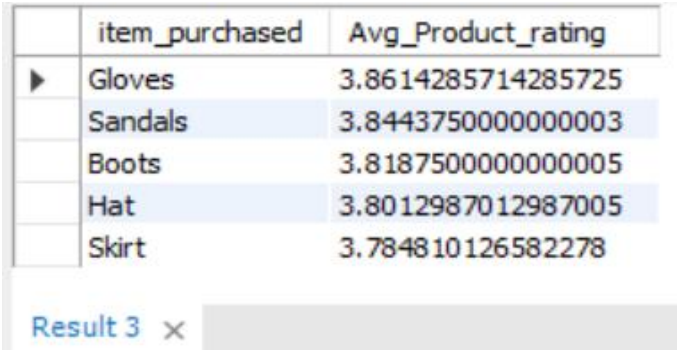


customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90
22	62
24	88
29	94
32	79
33	67

customer 2 ×

3. Top 5 Products by Rating

Found the five products with the highest average review ratings, helping the company understand which items customers love the most.

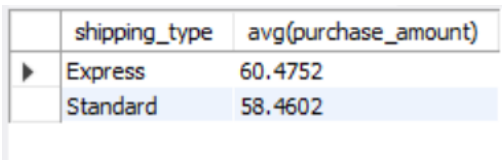


item_purchased	Avg_Product_rating
Gloves	3.8614285714285725
Sandals	3.8443750000000003
Boots	3.8187500000000005
Hat	3.8012987012987005
Skirt	3.784810126582278

Result 3 ×

4. Shipping Type Comparison

Compared the average purchase amount between customers who chose Standard shipping and those who chose Express shipping.



shipping_type	avg(purchase_amount)
Express	60.4752
Standard	58.4602

5. Subscribers vs. Non-Subscribers

Compared average spending and total revenue between subscribed and non-subscribed customers to see if subscriptions increase customer value.

subscription_status	total_customers	avg_spend	total_revenue
Yes	1053	59.4919	62645
No	2847	59.8651	170436

6. Discount-Dependent Products

Identified the top 5 products that were most frequently purchased using discounts. This shows which products depend heavily on promotions.

item_purchased	discount_rate
Hat	50.0000
Sneakers	49.6552
Coat	49.0683
Sweater	48.1707
Pants	47.3684

Result 6

7. Customer Segmentation

Grouped customers into three categories based on purchase history:

- New Customers – Few purchases
- Returning Customers – Moderate purchases
- Loyal Customers – Frequent purchases

This helps in targeted marketing strategies.

customer_segment	number of customers
loyal	3116
returning	701
New	83

8. Top 3 Products per Category

Listed the top three most purchased products within each category to understand popular items.

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

Result 8

9. Repeat Buyers & Subscriptions

Analyzed whether customers with more than 5 purchases are more likely to subscribe, helping measure loyalty impact.

subscription_status	repeat_buyers
Yes	958
No	2518

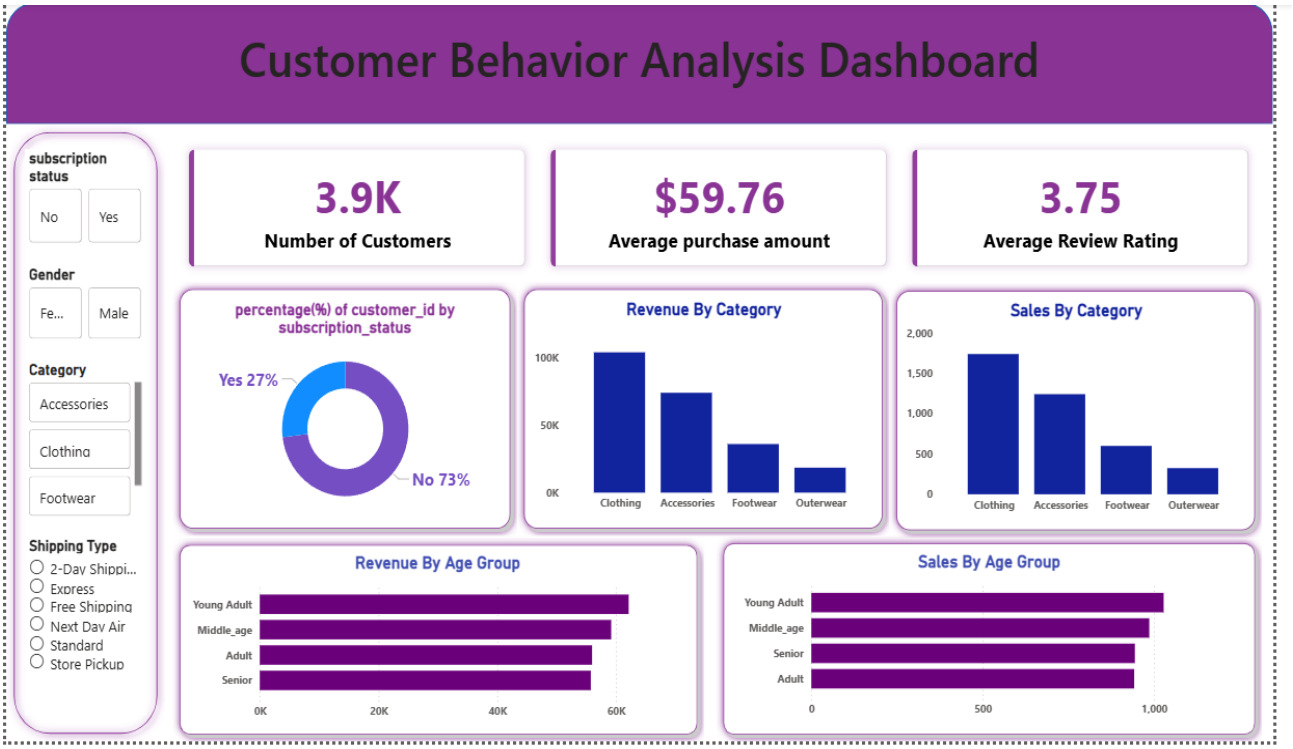
10. Revenue by Age Group

Calculated the total revenue contributed by each age group, helping identify which age segment drives the most sales.

age_group	total_revenue
Young Adult	62143
Middle_age	59197
Adult	55978
Senior	55763

Dashboard in Power BI:

Finally, I built an interactive dashboard in Power BI to present insight visually.



Business Recommendations:

Based on the analysis, here are the key business suggestions:

1. Boost Subscriptions

Encourage more customers to subscribe by offering exclusive benefits such as special discounts, early access to products, or free shipping. This can increase repeat purchases and long-term customer value.

2. Strengthen Customer Loyalty Programs

Create or improve loyalty programs that reward repeat customers. Offer points, cashback, or special rewards to move customers from “New” and “Returning” into the “Loyal” segment.

3. Review Discount Strategy

Analyze the discount policy carefully. While discounts increase sales, they may reduce profit margins. The company should balance promotional offers with profitability.

4. Improve Product Positioning

Promote top-rated and best-selling products more actively in marketing campaigns. Highlighting popular products can increase customer trust and drive more sales.

5. Use Targeted Marketing

Focus marketing efforts on high-revenue age groups and customers who prefer express shipping, as they tend to spend more. Personalized campaigns can improve engagement and conversion rates.

Conclusion:

- By analyzing customer shopping data, we understood how different customers buy and what influences their decisions.
- We identified important factors like age group, discounts, subscriptions, and product ratings that impact sales.
- The analysis helped find high-value customers and popular products.
- Using these insights, the company can improve marketing strategies, increase customer loyalty, and grow overall revenue.