INTRODUCTION

This project analyses customer shopping data to understand purchasing patterns, spending behaviour, product performance, and the impact of discounts, subscriptions, and shipping types.

The goal is to generate **actionable business insights** that can help improve:

- Revenue
- Customer retention
- Product strategy
- Marketing campaigns

OBJECTIVES

- Compare revenue across customer demographics.
- Identify high-value customers.
- Analyse the impact of discounts and subscriptions.
- Determine top-rated products.
- Compare spending across shipping types.
- Provide business recommendations.

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, Colour)

    - Shopping behaviour (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)

    - Missing Data: 37 values in Review Rating column

Tools & Technologies Used

- **Python** (Pandas, Matplotlib/Seaborn) – Data Cleaning & EDA
- **SQL** – Business Queries
- **Power BI** – Dashboard & Visualisations

Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

● **Data Loading**: Imported the dataset using pandas

● **Initial Exploration**: using df.info()

```
<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
```

now using df.describe(include = 'all')

Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3900	3900	3900.000000	3900	3900
2	2	NaN	6	7
No	No	NaN	PayPal	Every 3 Months
2223	2223	NaN	677	584
NaN	NaN	25.351538	NaN	NaN
NaN	NaN	14.447125	NaN	NaN
NaN	NaN	1.000000	NaN	NaN
NaN	NaN	13.000000	NaN	NaN
NaN	NaN	25.000000	NaN	NaN
NaN	NaN	38.000000	NaN	NaN
NaN	NaN	50.000000	NaN	NaN

Now, Handling the Missing Values

Checked for the null values and assign the median rating of each product to them.

Standardization of the Columns

Renamed columns to snake case for better readability and documentation.

Perform the Feature Engineering

1) Created age_group column by binning customer ages.
2) Created purchase_frequency_days column from purchase data.

Now, Checking the Consistency of the Data

Checked if discount_applied and promo_code_used were redundant and then dropped the promo_code_used.

Now after performing Exploratory Data Analysis, we exported the cleaned CSV file for the further analysis using mysql.

Data Analysis using SQL (Business Transactions)

1) **Total revenue generated by male vs female customers**

	gender	sum(purchase_amount)
▶	Male	157890
	Female	75191

2) **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
	22	62

Total customers who used discounts but still bought the items with a price that is above the average purchase price = 839

3) **The top 5 products with the highest average review rating**

	item_purchased	Average_Review_Rating
▶	Gloves	3.86
	Sandals	3.84
	Boots	3.82
	Hat	3.8
	Skirt	3.78

4) The average purchase amount between Standard and Express Shipping

	shipping_type	Average_purchase_amount
▶	Express	60.48
	Standard	58.46

5) Compared average spend and total revenue across subscription status.

	subscription_status	Total_Customers	Average_Spend	total_revenue
▶	Yes	1053	59.49	62645
	No	2847	59.87	170436

6) The 5 products that have the highest percentage of purchases with disounts applied

item_purchased	Discounted_Purchases	Discount_Rate
Hat	77	50.0000
Sneakers	72	49.6552
Coat	79	49.0683
Sweater	79	48.1707
Pants	81	47.3684

7) Classified customers into New, Returning, and Loyal segments based on purchase history.

	customer_segment	number of customers
▶	Loyal	3116
	Returning	701
	New	83

8) The top 3 most purchased products within 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) Checked whether the customers with more than 5 purchases are more likely to subscribe.

	subscription_status	repeat_buyers
▶	Yes	958
	No	2518

10) The revenue contribution of each age group .

	age_group	total_revenue
▶	Young Adult	62143
	Middle-Aged	59197
	Adult	55978
	Senior	55763

Dashboard in Power BI

Business Recommendations

1) **Strengthen Subscription Adoption** – Introduce and promote exclusive incentives for subscribed customers to increase retention and long-term value.

2) **Implement Customer Loyalty Initiatives** – Design reward programs that encourage repeat purchases and gradually transition customers into the “Loyal” segment.
3) **Optimise Discount Strategies** – Carefully evaluate discount policies to ensure they stimulate sales while maintaining healthy profit margins.
4) **Enhance Product Positioning** – Prioritise high-rated and top-selling products in marketing campaigns to maximise visibility and conversion rates.
5) **Adopt Targeted Marketing Approaches** – Direct promotional efforts toward high-revenue age groups and customers who prefer express shipping to improve campaign efficiency.