Data Preprocessing and Analysis of Mall Customer Dataset

Understanding Customer Spending Patterns

Presented By : Jobina K V

Introduction

- **Data preprocessing is the process of cleaning and preparing raw data so it can be used for analysis.**
- **It ensures the data is correct, consistent, and ready for exploration.**

Main Purpose is :

- **Understand dataset structure like columns, types and missing values.**
- **Scale numeric data to bring values to a similar range.**
- **Clean and organize data for better analysis.**
- **Identify patterns and trends in the dataset.**

Problem Statement

Customer spending behavior is affected by many factors like :

- **Age**
- **Gender**
- **Annual Income**
- **Spending Score**

The problem is to understand how these factors influence spending and identify patterns in customer behavior.

Proposal Solution

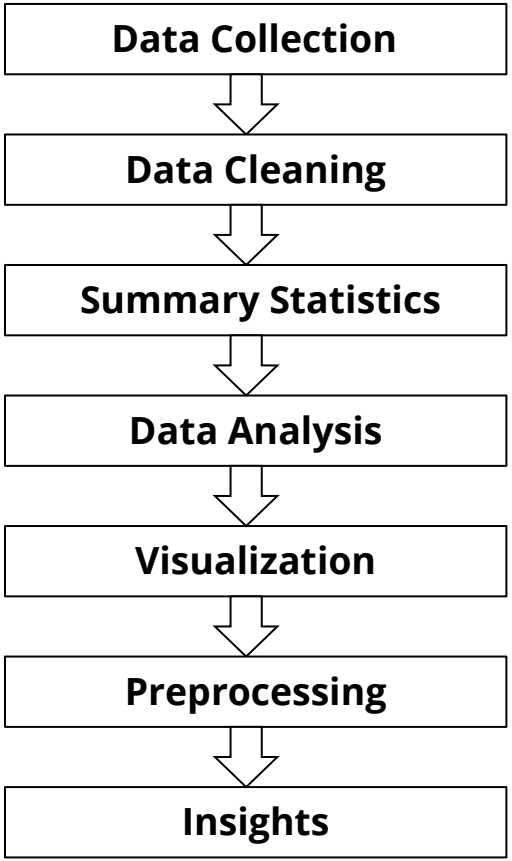
- **The problem is customer spending is influenced by many factors. To solve this, we perform Data Preprocessing on the Mall Customer dataset to clean, scale, and organize the data.**
- **We then explore patterns in spending based on features like Age, Gender, Annual Income, and Spending Score.**
- **This analysis will help identify which factors most affect customer spending behavior.**

Dataset Overview & Structure

- **The Mall Customer dataset includes records of 50 customers with details about Gender, Age, Annual Income, and Spending Score.**
- **It is analyzed to understand customer spending patterns and identify key factors affecting spending behavior.**

```
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              50 non-null     int64
 1   Gender                  50 non-null     object
 2   Age                     50 non-null     int64
 3   Annual Income (k$)      50 non-null     int64
 4   Spending Score (1-100)  50 non-null     int64
dtypes: int64(4), object(1)
memory usage: 2.1+ KB
```

Workflow


Data Collection
Data Cleaning
Summary Statistics
Data Analysis
Visualization
Preprocessing
Insights

Tools Used

- **Software: Python,Google Colab**
- **Libraries: Pandas, Matplotlib, NumPy, Sklearn.Preprocessing**

Implementation

- **Step 1 : Selected the first 50 customers from the dataset for analysis.**
- **Step 2 : Checked data types, missing values, and duplicates to ensure data quality.**
- **Step 3 : Generated summary statistics for numeric features like Age, Annual Income, and Spending Score to understand their mean, median, and range.**
- **Step 4 : Scaled numeric columns (`Age`, `Annual Income`, `Spending Score`) using MinMaxScaler.**
- **Step 5 : Replaced categorical values in Gender column with short codes (“M” for Male, “F” for Female).**
- **Step 6 : Plotted bar charts and histograms to study distributions and patterns :**

➔ **Average Spending Score by Gender**

➔ **Distribution of Age, Annual Income, and Spending Score**

Data info

```
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              50 non-null     int64
 1   Gender                  50 non-null     object
 2   Age                     50 non-null     int64
 3   Annual Income (k$)      50 non-null     int64
 4   Spending Score (1-100)  50 non-null     int64
dtypes: int64(4), object(1)
memory usage: 2.1+ KB
```

Data Cleaning

	0
CustomerID	0
Gender	0
Age	0
Annual Income (k$)	0
Spending Score (1-100)	0

dtype: int64

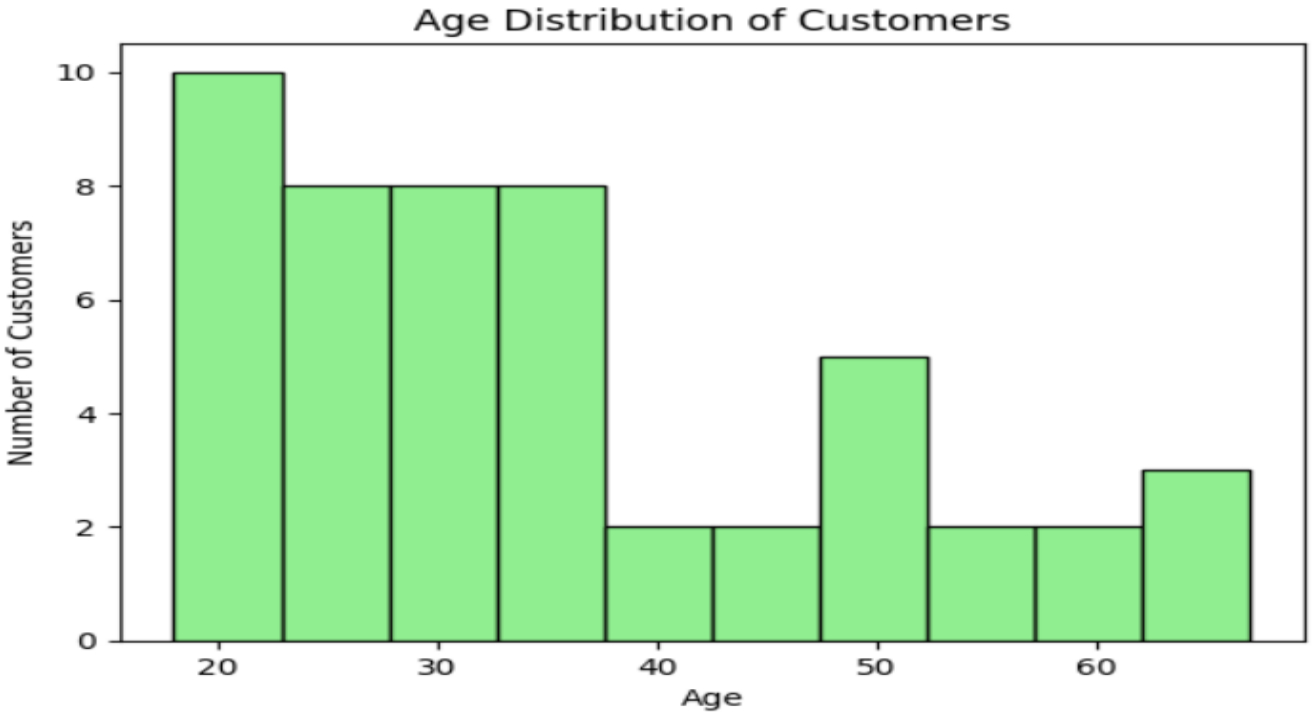
```
df.duplicated().sum()
```

```
np.int64(0)
```

Statistics Summary

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	50.00000	50.000000	50.000000	50.00000
mean	25.50000	35.280000	27.400000	49.48000
std	14.57738	13.751497	8.369039	30.21774
min	1.00000	18.000000	15.000000	3.00000
25%	13.25000	23.250000	20.000000	26.50000
50%	25.50000	31.000000	28.000000	44.50000
75%	37.75000	45.750000	34.000000	75.75000
max	50.00000	67.000000	40.000000	99.00000

Age Distribution

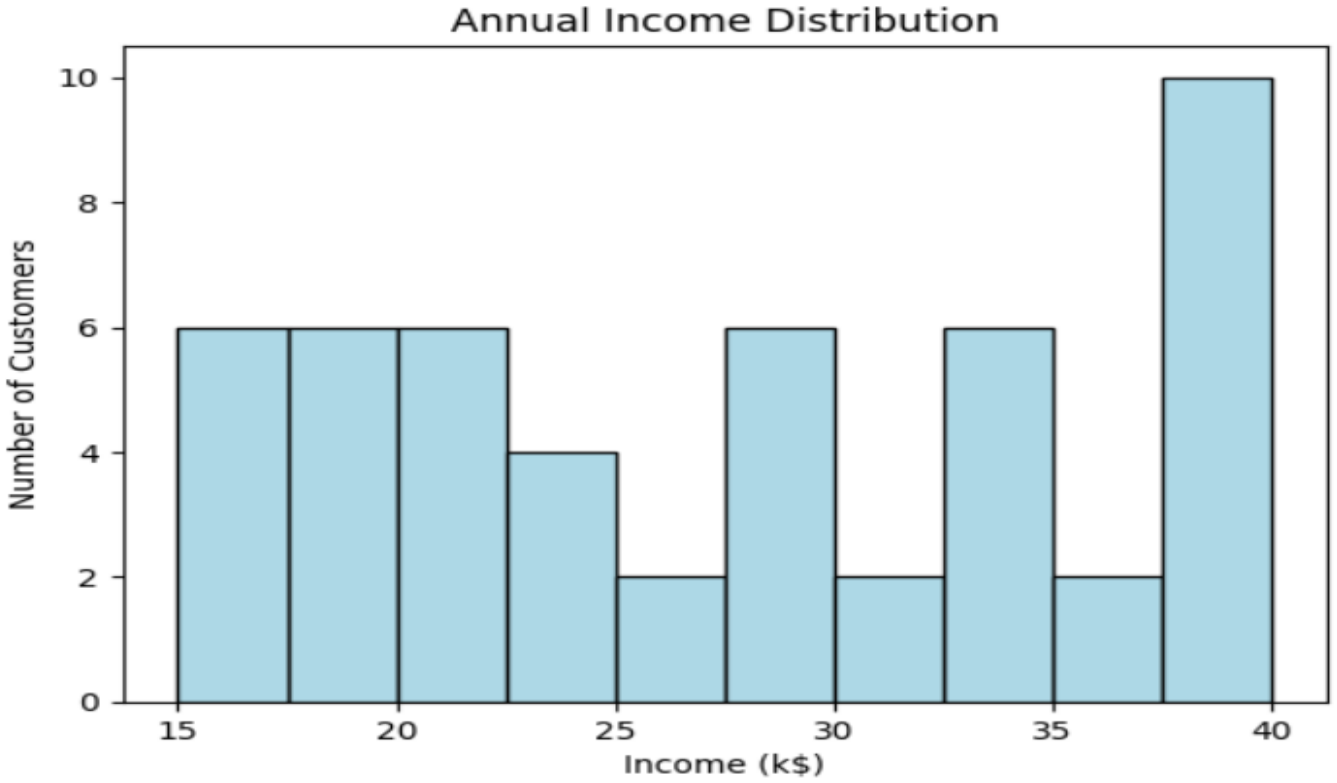


- **The histogram shows the distribution of customers' ages in the dataset before scaling.**
- **Purpose is Understand the main customer age group, which might affect their spending habits.**

Key Insight :

- **Most customers are 20–40 years old.**
- **Few customers younger than 20 or older than 60.**

Annual Income Distribution

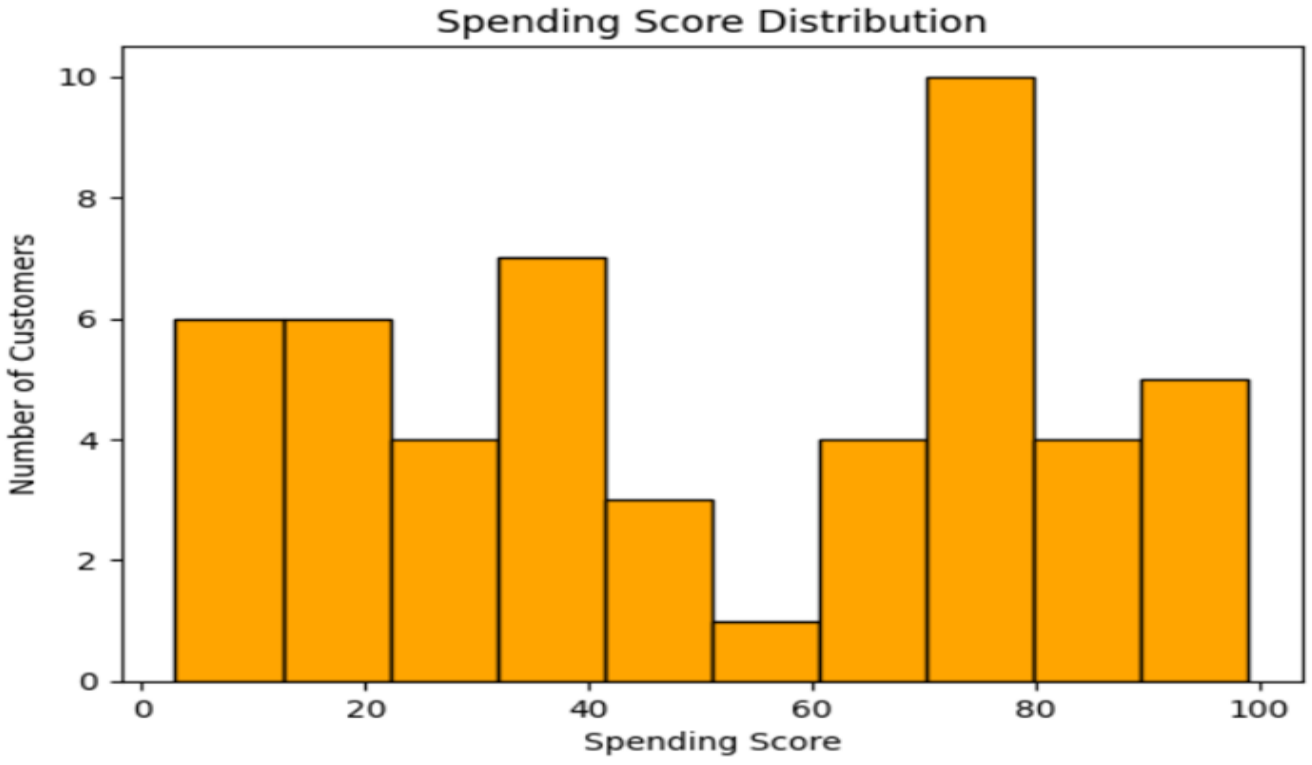


- **The histogram shows the distribution of customers' annual incomes in the dataset before scaling.**
- **Purpose is to find the common income range of customers, which may affect how they spend.**

Key Insight :

- **Most customers earn between 15k–40k annually.**
- **Very few customers have incomes at the extreme low or high ends.**

Spending Score Distribution

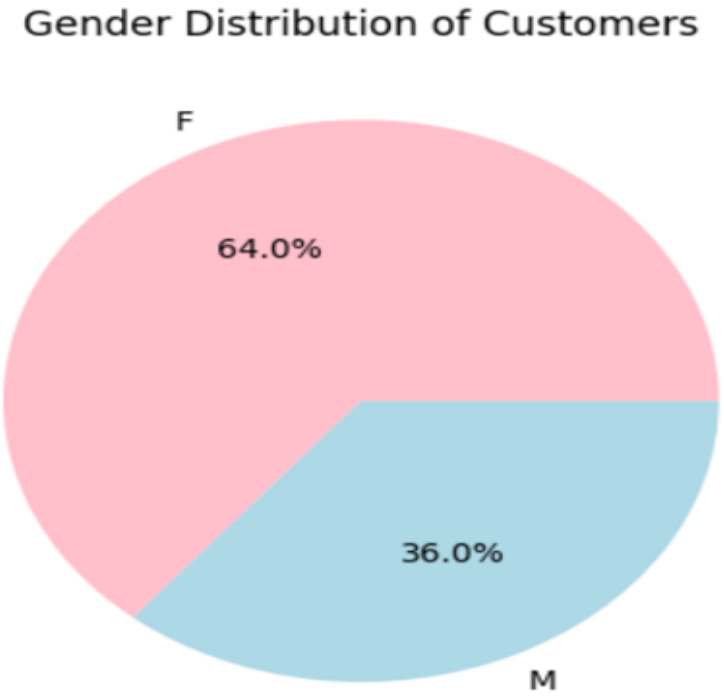


- **The histogram shows the distribution of customers' spending scores in the dataset before scaling.**
- **Purpose Identify how customers are spread across low, medium, and high spending levels.**

Key Insight :

- **Spending scores vary a lot, but many customers fall in the medium to high spending range.**
- **This means customers have different spending habits.**

Gender Distribution

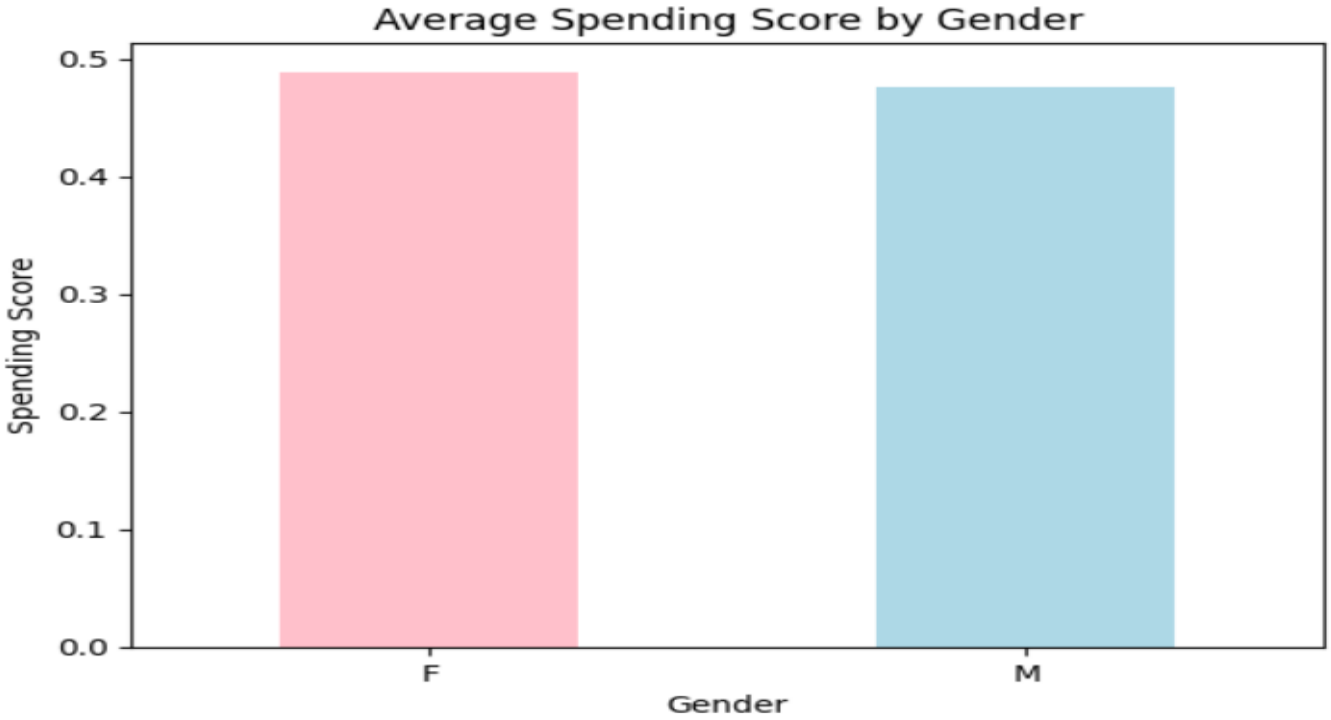


- **The Pie Chart Show the Percentage of male and female customers in the dataset.**
- **Purpose is Understand the customer gender ratio and compare how males and females spend.**

Key Insight

- **There are more female customers, about sixty-four percent, compared to male customers, about thirty-six percent.**

Average Spending Score by Gender



- **This Bar chart showing Average spending score of male and female customers after scaling.**
- **Purpose is Identify which gender contributes more to sales.**

Key Insight :

- **Female customers spent slightly more than males on average.**
- **Bar chart allows quick comparison of spending by gender.**

Result

- **Customers’ age, income, and spending scores show clear patterns after preprocessing.**
- **Female customers spend slightly more on average than male customers.**
- **Most customers fall in the 20–40 age group and have an annual income between 15k–40k.**
- **Spending scores vary widely, showing different customer buying behaviors.**

Conclusion

- **This data preprocessing and analysis helped us understand key factors affecting customer spending behavior.**
- **From this dataset, we found that:**
  - ➔ **Customers' age and income influence their spending patterns.**
  - ➔ **Female customers tend to spend slightly more than male customers.**
  - ➔ **Most customers fall in the 20–40 age range with an income of 15k–40k.**
  - ➔ **Spending scores vary widely, showing different buying behaviors.**
- **These insights can help businesses understand customer preferences and target marketing strategies effectively.**

Reference

- **Mall Customer Dataset – Kaggle**
- **Python & Library References – Pandas, Matplotlib, NumPy, sklearn.preprocessing**

THANK YOU