# **Experiment 1: Data Loading, Cleaning, and Preprocessing**

## **Aim:**

The goal of this experiment is to understand how to load a dataset, clean it, and perform basic preprocessing steps to prepare it for analysis

### **1. Importing Required Libraries**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

**Explanation:**We import numpy for numerical computations, pandas for handling data in tabular format, and matplotlib.pyplot for data visualization.

### **2. Loading the Dataset**

****df = pd.read\_csv("D:/tsa\_107/data/Microsoft\_Stock.csv")

**Explanation:**We use pd.read\_csv() to load a CSV file containing Microsoft stock data.

### **3. Understanding the Data**

****print(df.describe())

**Explanation:**df.describe() provides a statistical summary of the dataset, including mean, median, min, and max values for each numerical column.

### **4. Displaying the First Few Rows**

****df.head()

**Explanation:**df.head() shows the first five rows of the dataset, giving us an overview of the available columns and their values.

### **5. Visualizing Outliers with Box Plot**

****plt.figure(figsize=(10, 8))

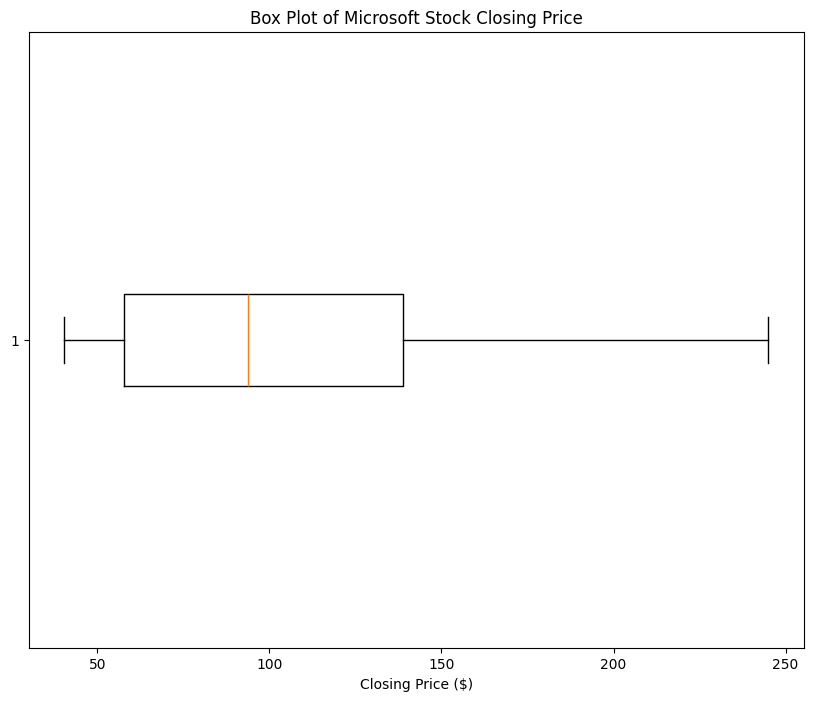
plt.boxplot(df['Close'], vert=False)

plt.title('Box Plot of Microsoft Stock Closing Price')

plt.xlabel('Closing Price ($)')

plt.show()

**Explanation:**A box plot helps visualize the distribution of closing stock prices, detecting outliers and understanding the spread of the data.



### **7. Checking for Missing Values**

****print(df.isnull().sum())

**Explanation:**df.isnull().sum() displays the number of missing values in each column, helping us identify if any data cleaning is required.

### **8. Visualizing the Distribution of Closing Prices with Histogram**

****plt.figure(figsize=(8, 6))

plt.hist(df['Close'], bins=20, color='purple', edgecolor='black')

plt.title('Distribution of Microsoft Stock Closing Prices')

plt.xlabel('Closing Price ($)')

plt.ylabel('Frequency')

plt.grid(True)

plt.show()

**Explanation:**A histogram shows the frequency distribution of closing prices, helping us understand how often specific price ranges occur in the dataset.