**Name: S KUSHAVARADHA**

**Reg. No.: 192124209**

**11. Scenario : You are a data scientist working for a company that sells products online.**

**You have been tasked with creating a simple plot to show the sales of a product over time.**

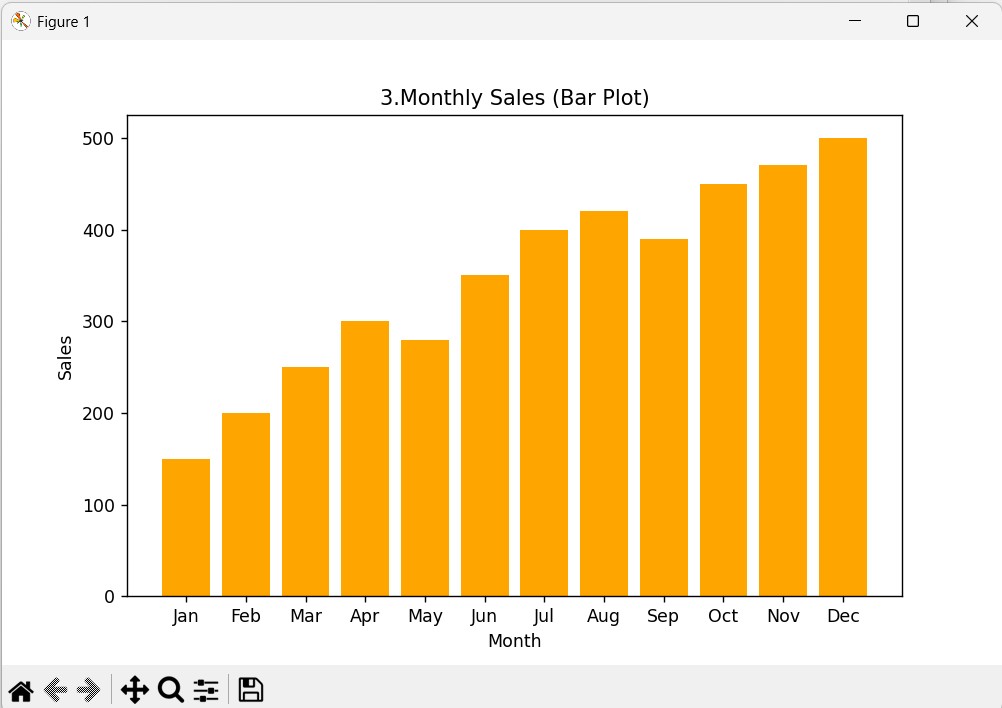
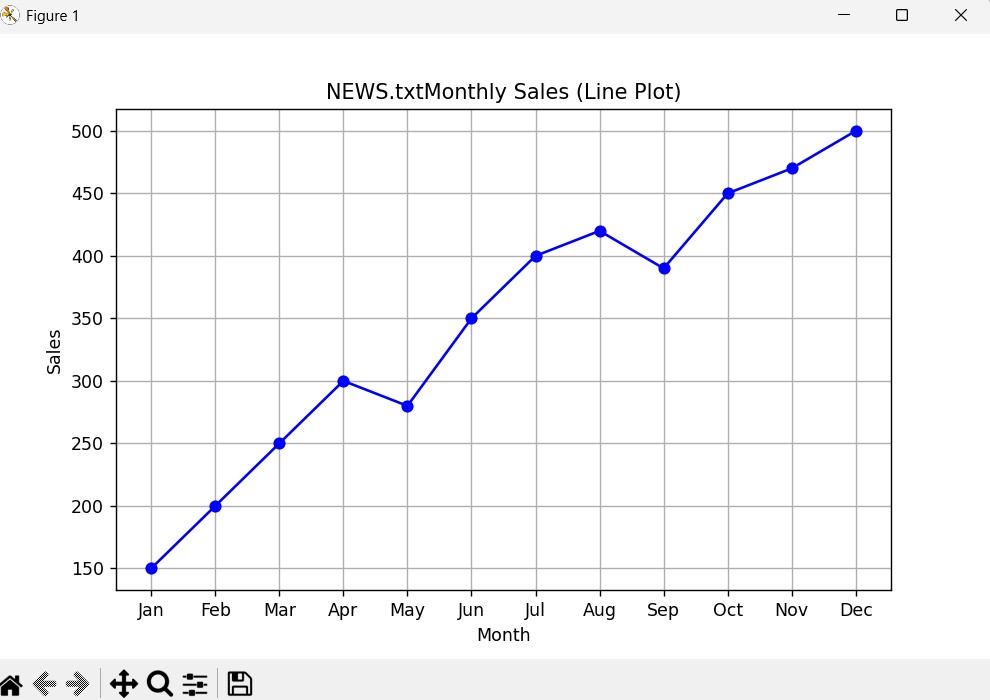
**Question:**

1. **Write code to create a simple line plot in Python using Matplotlib to predict sales happened in a month?**
2. **Write code to create a scatter plot in Python using Matplotlib to predict sales happened in a month?**
3. **Develop a Python program to create a bar plot of the monthly sales data.** import matplotlib.pyplot as plt months = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun',

'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec']

sales = [150, 200, 250, 300, 280, 350, 400, 420, 390, 450, 470, 500] plt.figure(figsize=(8, 5)) plt.plot(months, sales, marker='o', linestyle='-', color='blue') plt.title('NEWS.txtMonthly Sales (Line Plot)') plt.xlabel('Month') plt.ylabel('Sales') plt.grid(True) plt.show() plt.figure(figsize=(8, 5)) plt.scatter(months, sales, color='green') plt.title('2.Monthly Sales (Scatter Plot)') plt.xlabel('Month') plt.ylabel('Sales') plt.grid(True) plt.show() plt.figure(figsize=(8, 5)) plt.bar(months, sales, color='orange')

plt.title('3.Monthly Sales (Bar Plot)') plt.xlabel('Month') plt.ylabel('Sales') plt.show()



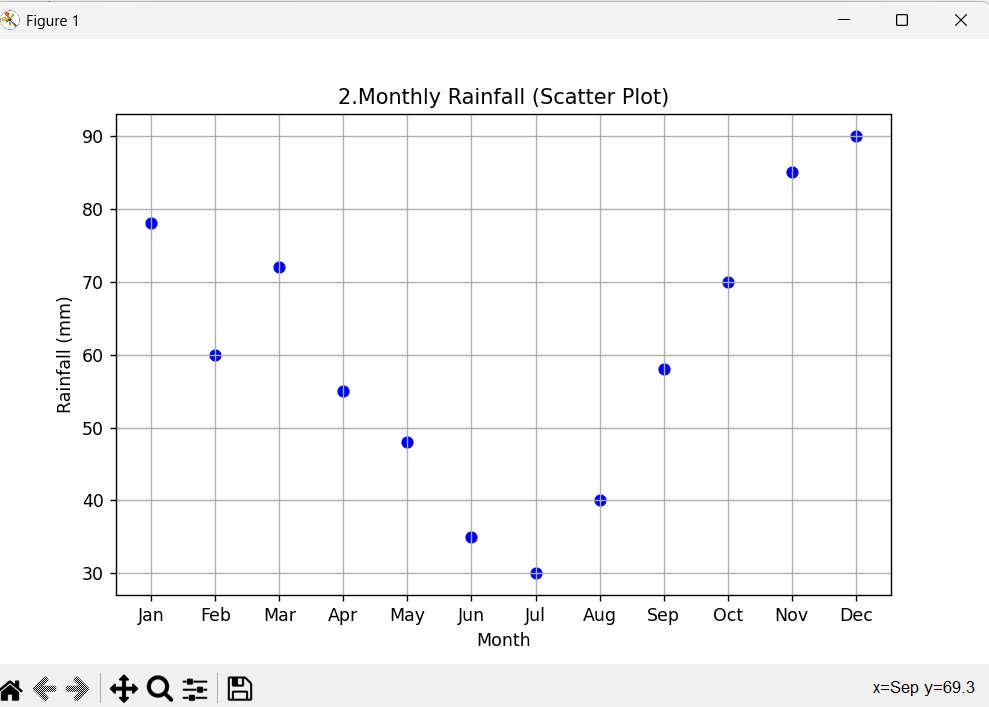
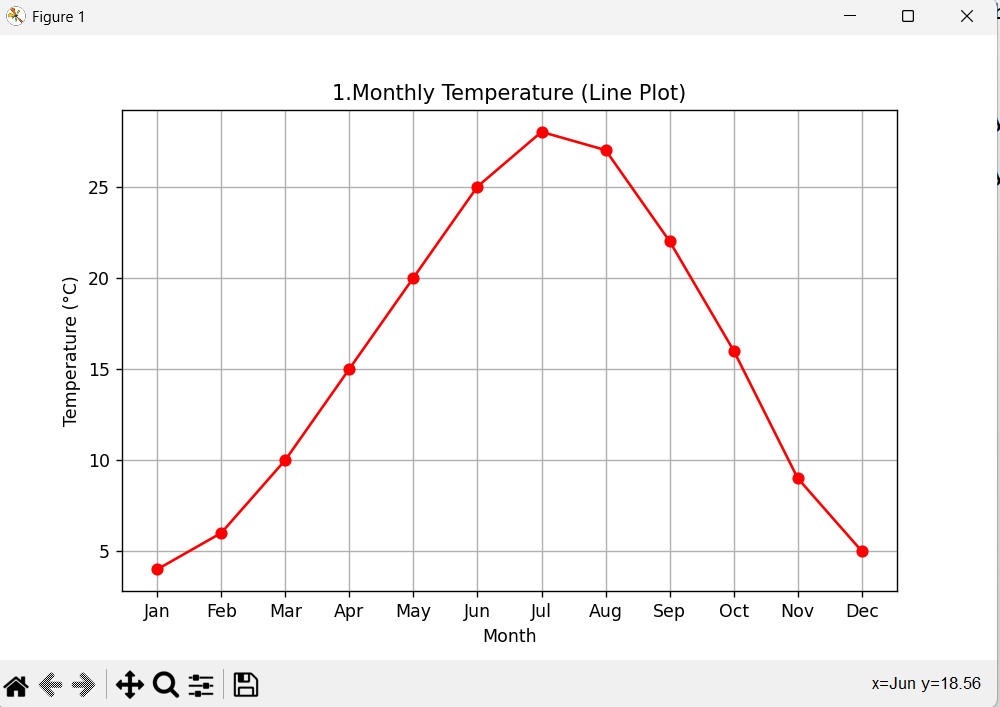
**12. Scenario: You are working on a data analysis project that involves analyzing the monthly temperature and rainfall data for a city. You have a dataset containing the monthly temperature and rainfall values for each month of a year. Your task is to develop a Python program that generates line plots and scatter plots to visualize the temperature and rainfall data.**

**Question:**

**1. Develop a Python program to create a line plot of the monthly temperature data.**

**2: Develop a Python program to create a scatter plot of the monthly rainfall data.** import matplotlib.pyplot as plt months = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec']

temperature = [4, 6, 10, 15, 20, 25, 28, 27, 22, 16, 9, 5] rainfall = [78, 60, 72, 55, 48, 35, 30, 40, 58, 70, 85, 90] plt.figure(figsize=(8, 5)) plt.plot(months, temperature, marker='o', color='red') plt.title('1.Monthly Temperature (Line Plot)') plt.xlabel('Month') plt.ylabel('Temperature (°C)') plt.grid(True) plt.show() plt.figure(figsize=(8, 5)) plt.scatter(months, rainfall, color='blue') plt.title('2.Monthly Rainfall (Scatter Plot)') plt.xlabel('Month') plt.ylabel('Rainfall (mm)') plt.grid(True) plt.show()



**13. Scenario: You are working on a text analysis project and need to determine the frequency distribution of words in a given text document. You have a text document named “sample\_text.txt” containing a paragraph of text. Your task is to develop a Python program that reads the text document, processes the text, and generates a frequency distribution of the words.**

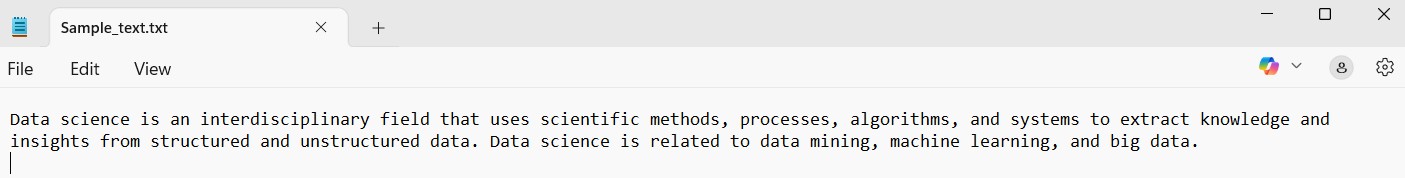
**Question: How would you develop a Python program to calculate the frequency distribution of words in a text document?**

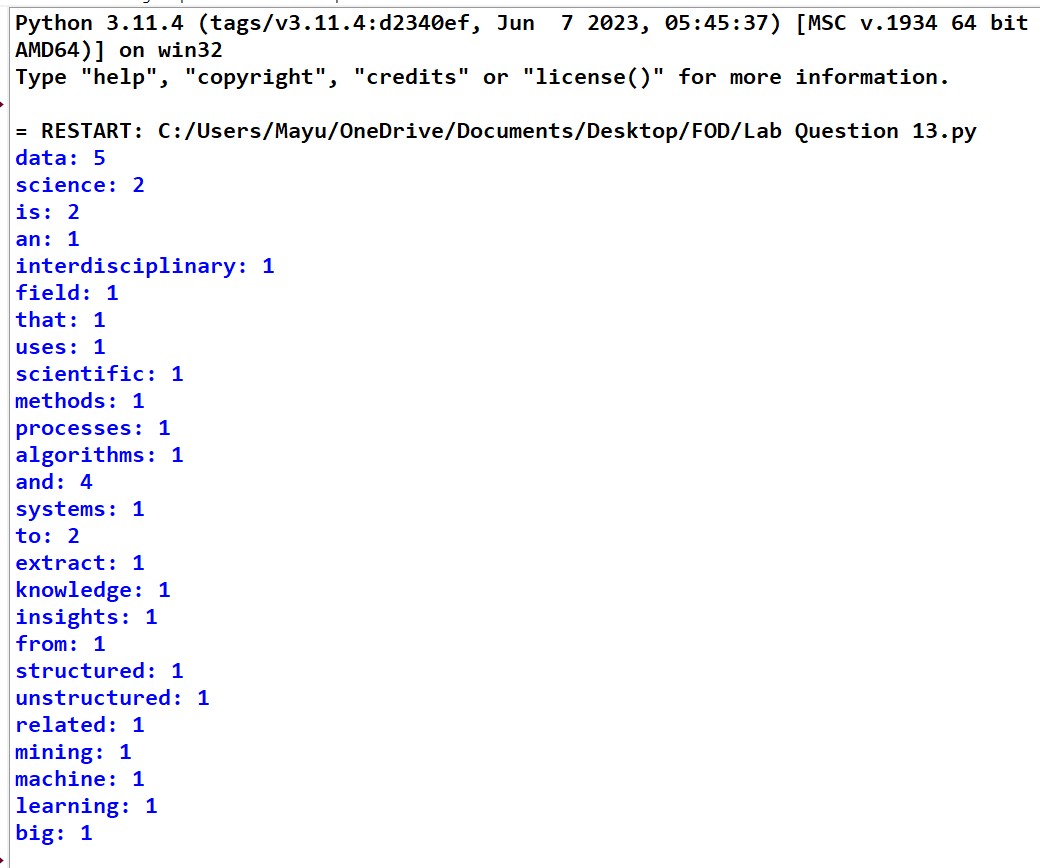
import string from collections import Counter

with open('sample\_text.txt', 'r') as file:

text = file.read() text = text.lower() text = text.translate(str.maketrans('', '', string.punctuation)) words = text.split()

word\_freq = Counter(words) for word, freq in word\_freq.items(): print(f"{word}: {freq}")





**14. Scenario: You are a data analyst working for a company that sells products online. You have been tasked with analyzing the sales data for the past month. The data is stored in a Pandas data frame.**

**Question: Develop a code in python to find the frequency distribution of the ages of the customers who have made a purchase in the past month.**

import pandas as pd

data = {

'CustomerID': [101, 102, 103, 104, 105, 106, 107, 108],

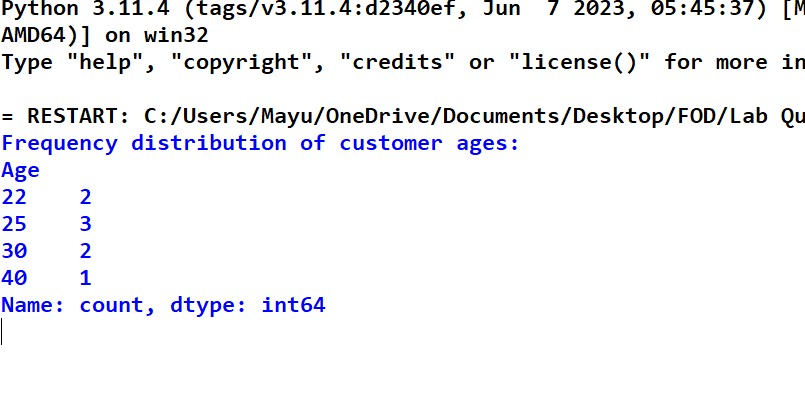
'Age': [25, 30, 22, 25, 30, 40, 22, 25],

'PurchaseAmount': [200, 150, 180, 210, 160, 300, 190, 220]

}

df = pd.DataFrame(data)

age\_frequency = df['Age'].value\_counts().sort\_index() print("Frequency distribution of customer ages:") print(age\_frequency)



**15. Scenario: You are a data analyst working for a social media platform. As part of your analysis, you have a dataset containing user interaction data, including the number of likes received by each post. Your task is to develop a Python program that calculates the frequency distribution of likes among the posts.**

**Question: Develop a Python program to calculate the frequency distribution of likes among the posts?** import pandas as pd

data = {

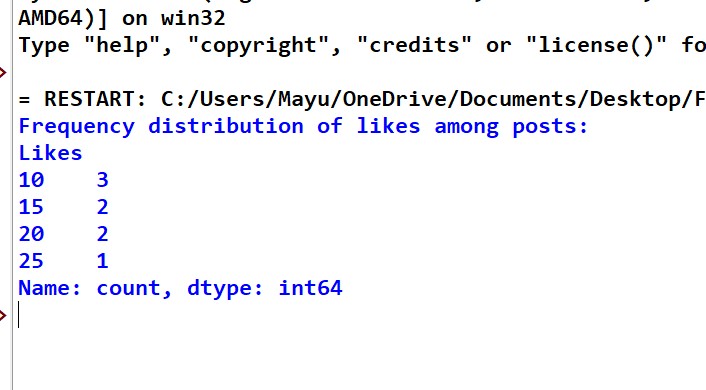
'PostID': [201, 202, 203, 204, 205, 206, 207, 208],

'Likes': [10, 15, 10, 20, 15, 10, 25, 20]

}

df = pd.DataFrame(data) like\_frequency = df['Likes'].value\_counts().sort\_index() print("Frequency distribution of likes among posts:")

print(like\_frequency)



**16. Scenario: You are working on a project that involves analyzing customer reviews for a product. You have a dataset containing customer reviews, and your task is to develop a Python program that calculates the frequency distribution of words in the reviews.**

**Question: Develop a Python program to calculate the frequency distribution of words in the customer reviews dataset?** import pandas as pd from collections import Counter import string data = {

'ReviewID': [1, 2, 3, 4],

'ReviewText': [

"Great product, really loved it!",

"Good quality, but too expensive.",

"Amazing product, worth the price.",

"Not bad, but expected better quality."

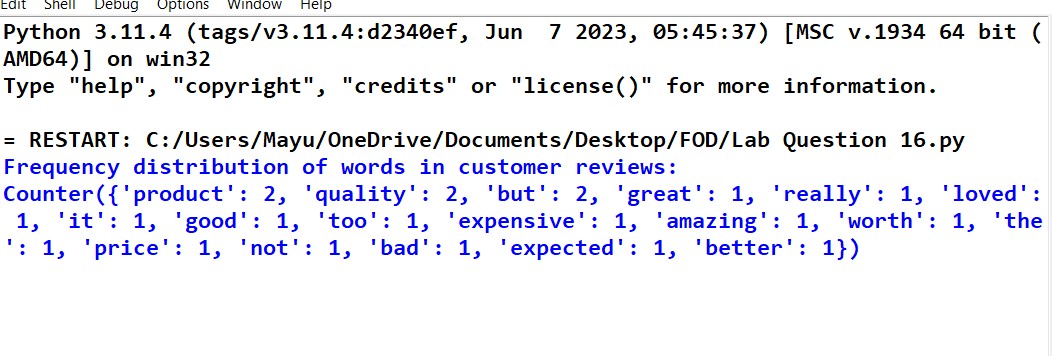
]

}

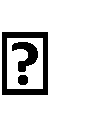
df = pd.DataFrame(data)

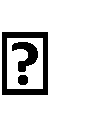
all\_reviews = ' '.join(df['ReviewText'].str.lower()) all\_reviews = all\_reviews.translate(str.maketrans('', '', string.punctuation)) words = all\_reviews.split() word\_freq = Counter(words) print("Frequency distribution of words in customer reviews:")

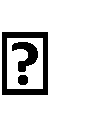
print(word\_freq)

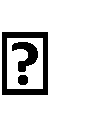


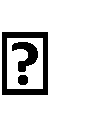
**17. Scenario: You are a data analyst working for a marketing research company. Your team has collected a large dataset containing customer feedback from various social media platforms. The dataset consists of thousands of text entries, and your task is to develop a Python program to analyze the frequency distribution of words in this dataset. Your program should be able to perform the following tasks:**

 **Load the dataset from a CSV file (data.csv) containing a single column named “feedback” with each row representing a customer comment.**

 **Preprocess the text data by removing punctuation, converting all text to lowercase, and eliminating any stop words (common words like “the,” “and,” “is” etc. that don’t carry significant meaning).**

 **Calculate the frequency distribution of words in the preprocessed dataset.**

 **Display the top N most frequent words and their corresponding frequencies, where N is provided as user input.**

 **Plot a bar graph to visualize the top N most frequent words and their frequencies.**

**Question: Create a Python program that fulfills these requirements and gain insights from the customer feedback data.**

import pandas as pd import string from collections import Counter import matplotlib.pyplot as plt from nltk.corpus import stopwords import nltk nltk.download('stopwords') df = pd.read\_csv('data.csv') stop\_words = set(stopwords.words('english')) df['feedback'] = df['feedback'].str.lower()

df['feedback'] = df['feedback'].apply(lambda x: x.translate(str.maketrans('', '', string.punctuation))) def preprocess\_text(text):

words = text.split()

return [word for word in words if word not in stop\_words]

df['processed\_feedback'] = df['feedback'].apply(preprocess\_text) all\_words = [word for feedback in df['processed\_feedback'] for word in feedback] word\_freq = Counter(all\_words)

N = int(input("Enter the number of top frequent words to display: "))

top\_n\_words = word\_freq.most\_common(N)

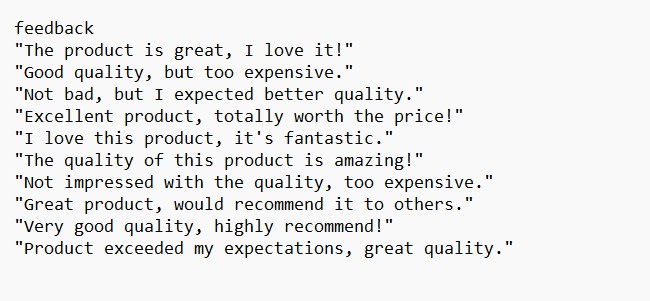
print(f"\nTop {N} most frequent words:") for word, freq in top\_n\_words:

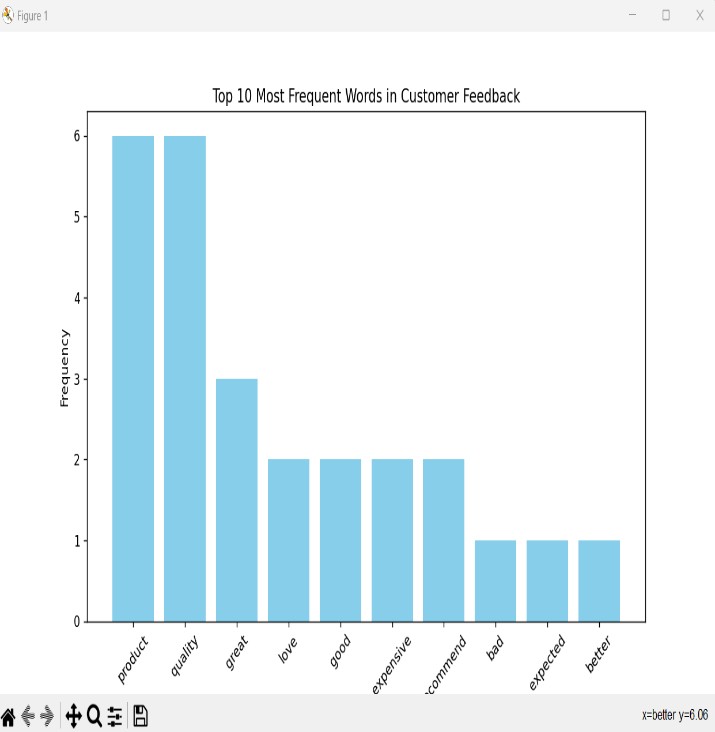
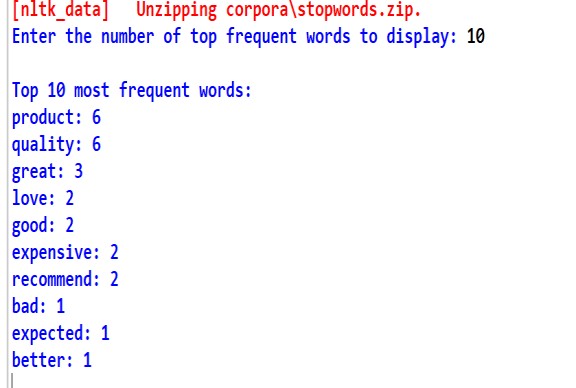
print(f"{word}: {freq}")

words, frequencies = zip(\*top\_n\_words)

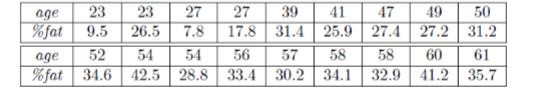
plt.figure(figsize=(10, 6)) plt.bar(words, frequencies, color='skyblue') plt.title(f"Top {N} Most Frequent Words in Customer Feedback") plt.xlabel('Words') plt.ylabel('Frequency') plt.xticks(rotation=45)

plt.show()

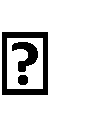


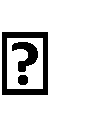


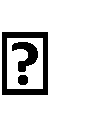
**18. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.**



**Question:**

 **Calculate the mean, median and standard deviation of age and %fat using Pandas.**

 **Draw the boxplots for age and %fat.**

 **Draw a scatter plot and a q-q plot based on these two variables.**

**CODE:**

import pandas as pd import matplotlib.pyplot as plt import statsmodels.api as sm

data = {

'age': [23, 23, 27, 27, 39, 41, 47, 49, 50,

52, 54, 54, 56, 57, 58, 58, 60, 61],

'%fat': [9.5, 26.5, 7.8, 17.8, 31.4, 25.9, 27.4, 27.2, 31.2,

34.6, 42.5, 28.8, 33.4, 30.2, 34.1, 32.9, 41.2, 35.7]

}

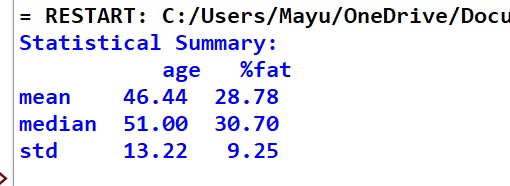
df = pd.DataFrame(data)

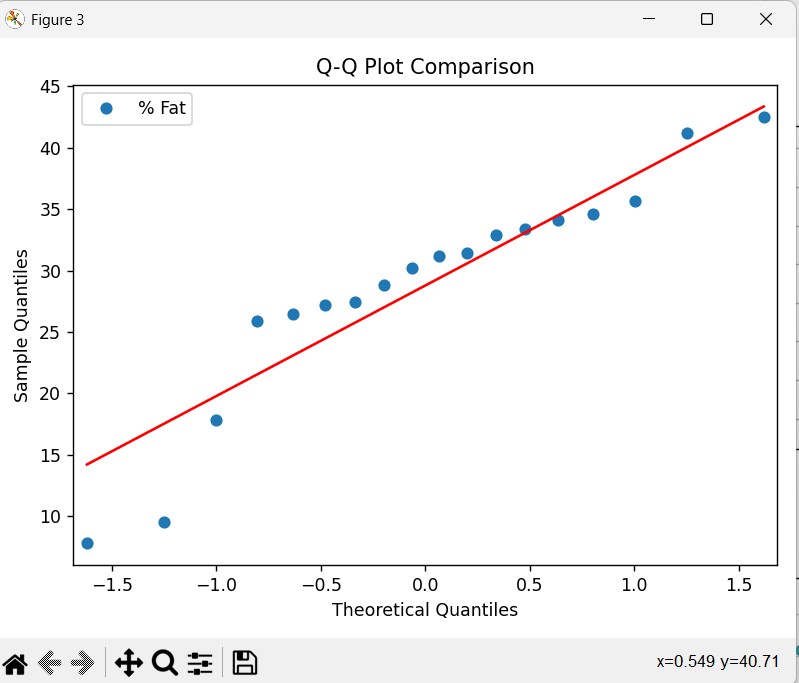
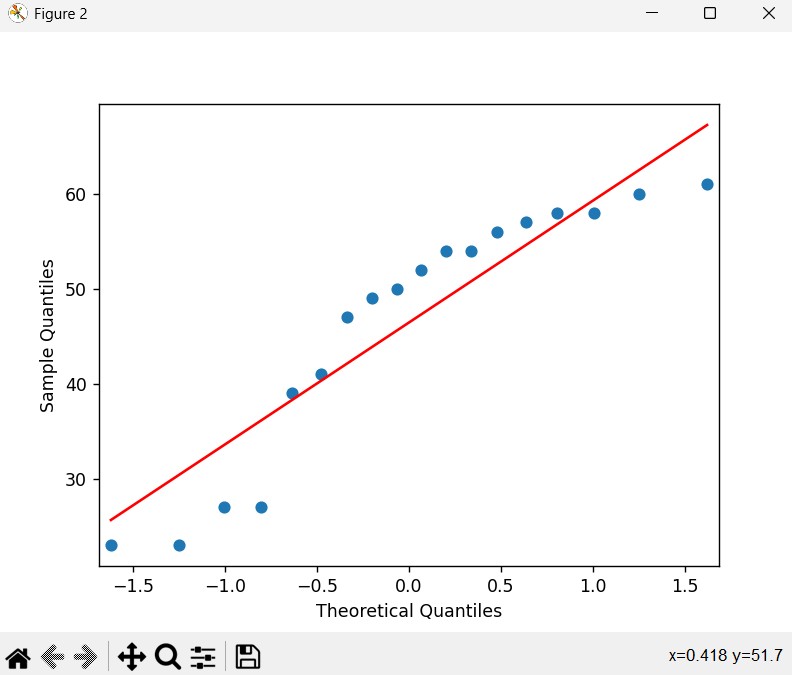
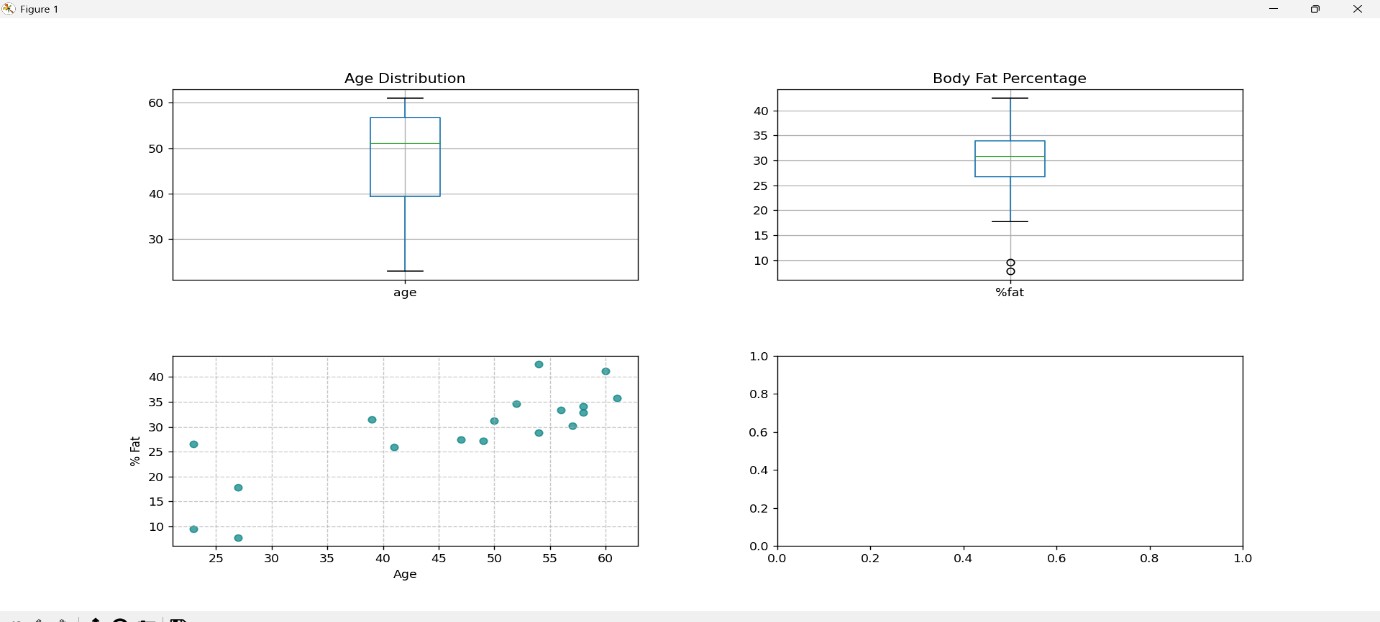
stats = df.agg(['mean', 'median', 'std']) plt.figure(figsize=(15, 10))

plt.subplots\_adjust(hspace=0.4, wspace=0.3)

plt.subplot(2, 2, 1) df.boxplot(column='age')

plt.title('Age Distribution') plt.subplot(2, 2, 2) df.boxplot(column='%fat') plt.title('Body Fat Percentage') plt.subplot(2, 2, 3) plt.scatter(df['age'], df['%fat'], c='teal', alpha=0.7) plt.xlabel('Age') plt.ylabel('% Fat') plt.grid(True, linestyle='--', alpha=0.7) plt.subplot(2, 2, 4) sm.qqplot(df['age'], line='s', label='Age') sm.qqplot(df['%fat'], line='s', label='% Fat') plt.legend() plt.title('Q-Q Plot Comparison') plt.tight\_layout() plt.show() print("Statistical Summary:\n", stats.round(2)) **OUTPUT:**



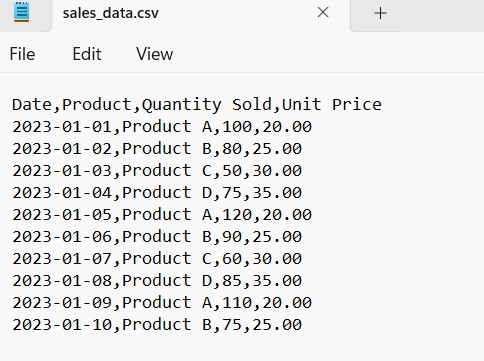


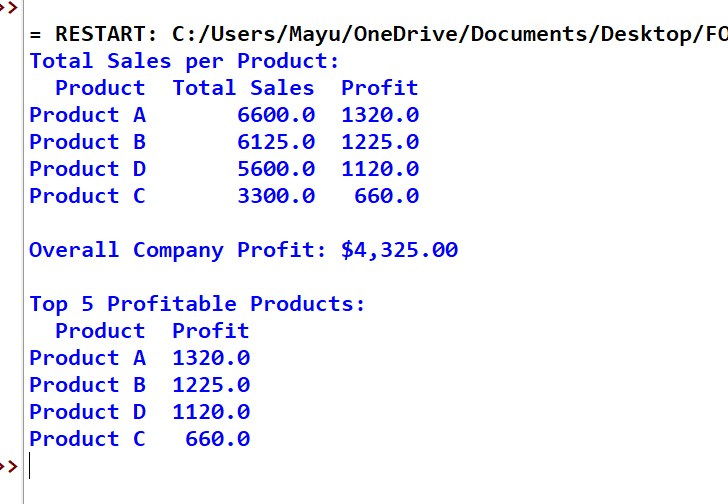
**19. Sales and Profit Analysis: a) Load the “sales\_data.csv” file into a Pandas data frame, which contains columns “Date,” “Product,” “Quantity Sold,” and “Unit Price”**

1. **Create a new column named “Total Sales” that calculates the total sales for each transaction (Quantity Sold \* Unit Price).**
2. **Calculate the total sales for each product and the overall profit, considering a 20% profit margin on each product. Display the top 5 most profitable products.**

**CODE:**

import pandas as pd df = pd.read\_csv('sales\_data.csv') df['Total Sales'] = df['Quantity Sold'] \* df['Unit Price'] product\_sales = df.groupby('Product').agg({'Total Sales': 'sum'}).reset\_index() product\_sales['Profit'] = product\_sales['Total Sales'] \* 0.20 overall\_profit = product\_sales['Profit'].sum() top\_products = product\_sales.nlargest(5, 'Profit') print("Total Sales per Product:") print(product\_sales.sort\_values('Total Sales', ascending=False).to\_string(index=False)) print(f"\nOverall Company Profit: ${overall\_profit:,.2f}") print("\nTop 5 Profitable Products:") print(top\_products[['Product', 'Profit']].to\_string(index=False)) **OUTPUT:**





**20. Customer Segmentation: a) Load “customer\_data.” file into a Pandas data frame, which contains “Customer ID,” ”Age,” “Gender,” and “Total Spending.”**

1. **Segment customers into three groups based on their total spending: “High Spenders,” ”Medium Spenders,” and “Low Spenders.” Assign these segments to a new column in the data frame.**
2. **Calculate the average age of customers in each spending segment.**

**CODE:**

import pandas as pd df = pd.read\_csv('customer\_data.csv') quantiles = df['Total Spending'].quantile([0.33, 0.67])

df['Spending Segment'] = pd.cut(df['Total Spending'], bins=[-1, quantiles[0.33], quantiles[0.67], float('inf')], labels=['Low Spenders', 'Medium Spenders', 'High Spenders'])

avg\_age = df.groupby('Spending Segment')['Age'].mean() print("Customer Segmentation:") print(df[['Customer ID', 'Spending Segment']]) print("\nAverage Age per Spending Segment:") print(avg\_age) print("\nData Quality Checks:")

df.info()

print("\nMissing Values:") print(df.isna().sum()) print("\nGender Distribution:") print(df['Gender'].value\_counts()) print("\nSpending Segment Statistics:")

print(df.groupby('Spending Segment', observed=True)['Total Spending'].agg(['mean', 'median', 'std']))

**OUTPUT:**

