

DAY 4 LAB EXPERIMENTS

NAME : GUTTI CHAITANYA

Reg No : 192424185

DATE : 19/12/2025

EXP_16 To Develop a Python Program to Calculate the frequency distribution of words in the customer reviews Dataset

The screenshot shows a Jupyter Notebook interface with a dark theme. On the left, there's a file tree with 'sample_data' containing 'customer_reviews_500.csv' and 'word_frequency_distribution.csv'. The main area contains the following Python code:

```
import pandas as pd
import re
from collections import Counter

# Load the customer reviews dataset
reviews_df = pd.read_csv("customer_reviews_500.csv")

# Combine all reviews into a single text
all_reviews = " ".join(reviews_df["review_text"].astype(str))

# Convert text to lowercase and remove punctuation
clean_text = re.sub(r'[!-zA-Z]', '', all_reviews.lower())

# Split text into individual words
words = clean_text.split()

# Calculate frequency distribution of words
word_frequency = Counter(words)

# Convert the result into a DataFrame
frequency_df = pd.DataFrame(word_frequency.items(), columns=["Word", "Frequency"])

# Sort by frequency in descending order
frequency_df = frequency_df.sort_values(by="Frequency", ascending=False)

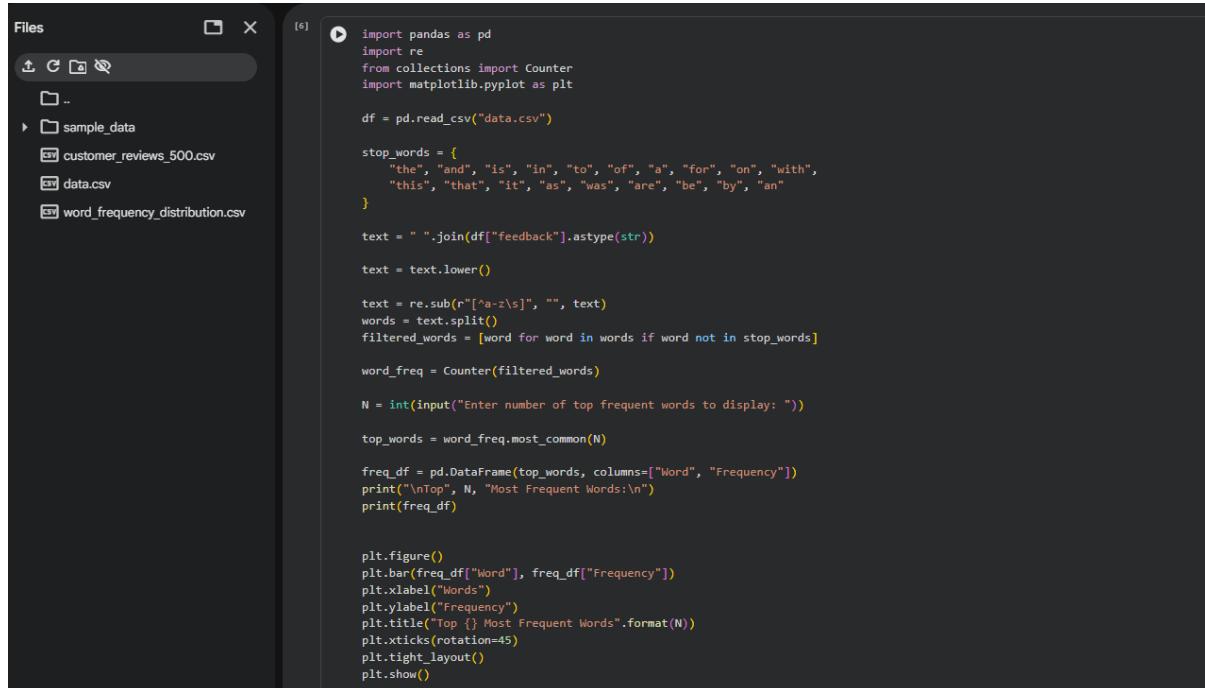
# Display top 10 most frequent words
print(frequency_df.head(10))

# Save the frequency distribution to a CSV file
frequency_df.to_csv("word_frequency_distribution.csv", index=False)
```

Below the code, the output shows the top 10 most frequent words from the dataset:

Rank	Word	Frequency
7	and	249
19	product	239
6	quality	150
0	not	136
27	good	135
14	money	113
2	the	111
23	very	99
16	bad	92
3	price	58

EXP_17 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.



The screenshot shows a Jupyter Notebook interface with a sidebar containing files: sample_data, customer_reviews_500.csv, data.csv, and word_frequency_distribution.csv. The main area contains Python code for reading a CSV file, removing stop words, and creating a bar chart of the top frequent words.

```
import pandas as pd
import re
from collections import Counter
import matplotlib.pyplot as plt

df = pd.read_csv("data.csv")

stop_words = {
    "the", "and", "is", "in", "to", "of", "a", "for", "on", "with",
    "this", "that", "it", "as", "was", "are", "be", "by", "an"
}

text = ".join(df["feedback"].astype(str))

text = text.lower()

text = re.sub(r"[^a-z\s]", "", text)
words = text.split()
filtered_words = [word for word in words if word not in stop_words]

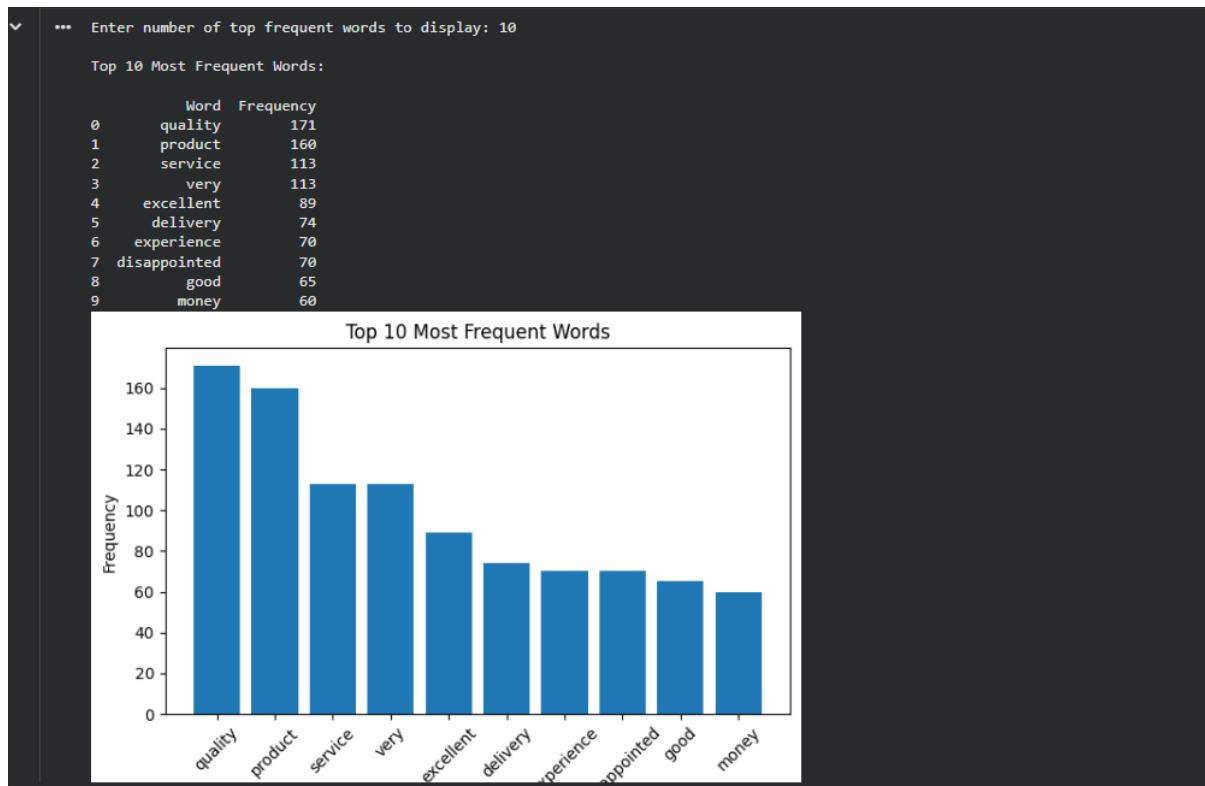
word_freq = Counter(filtered_words)

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

top_words = word_freq.most_common(N)

freq_df = pd.DataFrame(top_words, columns=["Word", "Frequency"])
print("\nTop ", N, "Most Frequent Words:\n")
print(freq_df)

plt.figure()
plt.bar(freq_df["Word"], freq_df["Frequency"])
plt.xlabel("Words")
plt.ylabel("Frequency")
plt.title("Top {} Most Frequent Words".format(N))
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



EXP_18 Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.

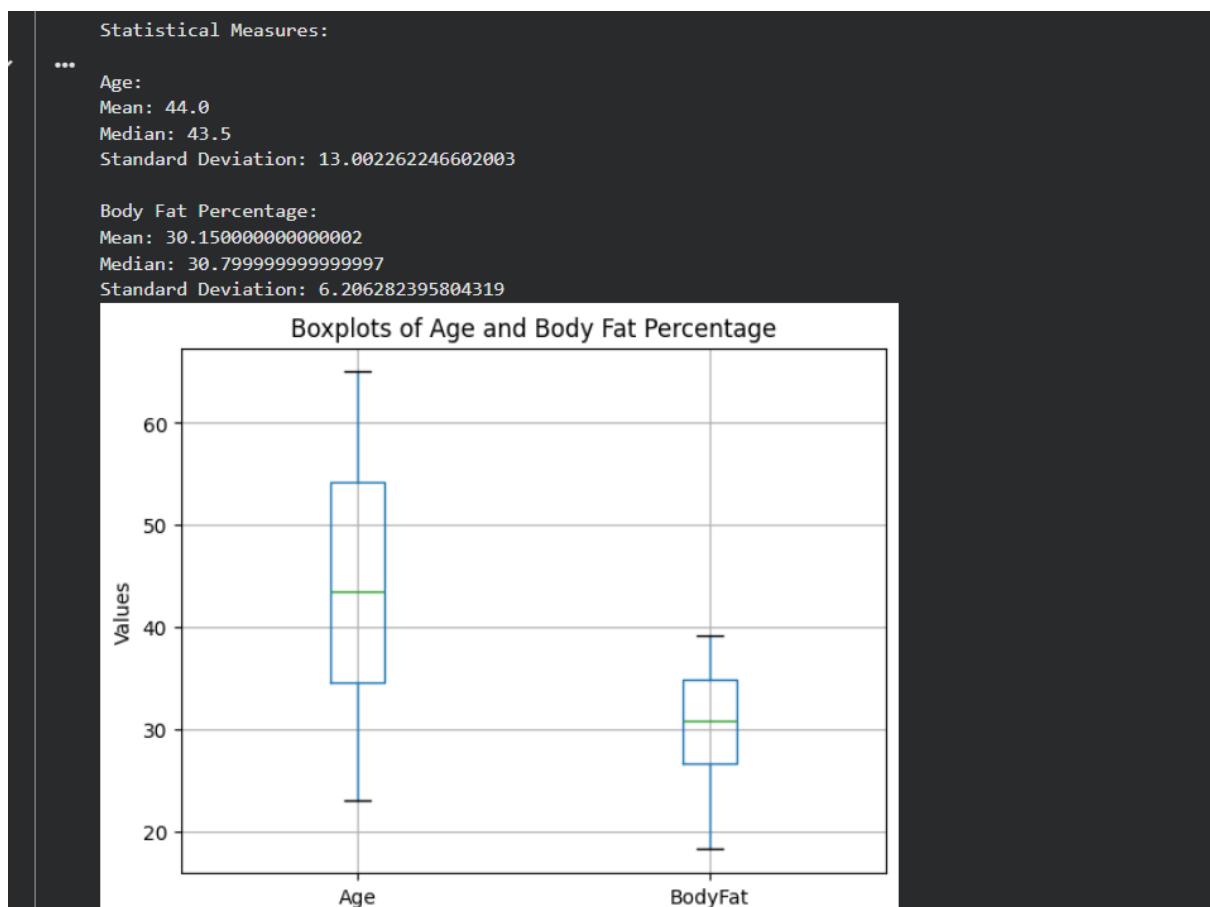
The screenshot shows a Jupyter Notebook interface. On the left is a file browser titled 'Files' with several CSV files listed. The main area contains Python code for statistical analysis and plotting. The code imports 'scipy.stats' and reads a CSV file 'age_bodyfat_18.csv'. It prints 'Statistical Measures' and calculates and prints the mean, median, and standard deviation for both 'Age' and 'BodyFat'. It then creates two boxplots: one for 'Age' and one for 'BodyFat', with the title 'Boxplots of Age and Body Fat Percentage'. Finally, it generates a Q-Q plot for 'BodyFat'.

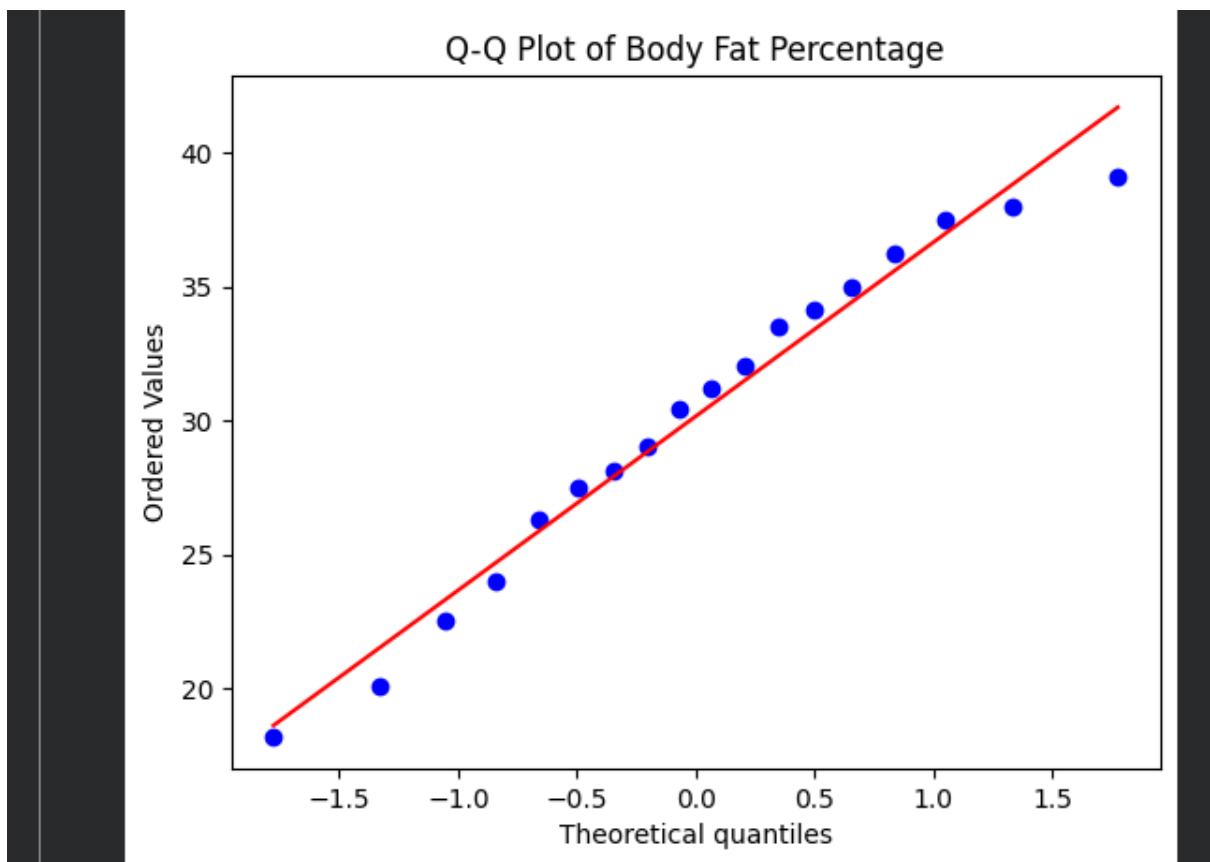
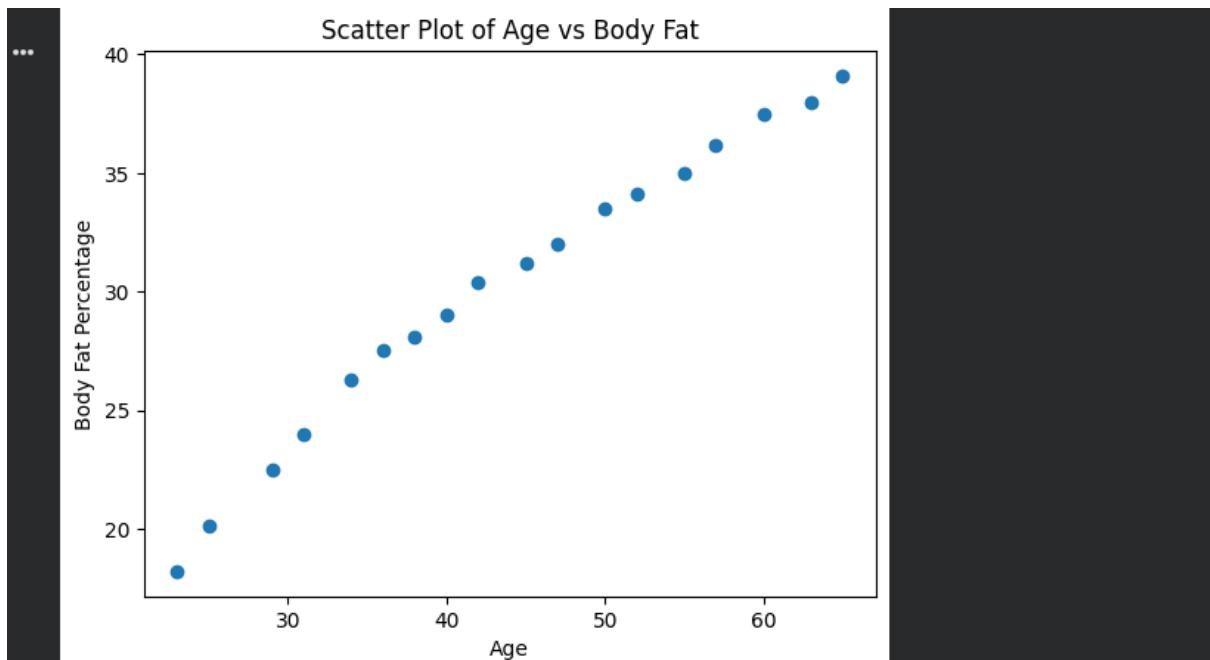
```
from scipy import stats
df = pd.read_csv("age_bodyfat_18.csv")
print("Statistical Measures:\n")
print("Age:")
print("Mean:", df["Age"].mean())
print("Median:", df["Age"].median())
print("Standard Deviation:", df["Age"].std(), "\n")
print("Body Fat Percentage:")
print("Mean:", df["BodyFat"].mean())
print("Median:", df["BodyFat"].median())
print("Standard Deviation:", df["BodyFat"].std())
plt.figure()
df.boxplot(column=["Age", "BodyFat"])
plt.title("Boxplots of Age and Body Fat Percentage")
plt.ylabel("Values")
plt.show()

plt.figure()
plt.scatter(df["Age"], df["BodyFat"])
plt.xlabel("Age")
plt.ylabel("Body Fat Percentage")
plt.title("Scatter Plot of Age vs Body Fat")
plt.show()

stats.probplot(df["BodyFat"], dist="norm", plot=plt)
plt.title("Q-Q Plot of Body Fat Percentage")
plt.show()
```

Statistical Measures:
Age:
Mean: 44.0





EXP_19 Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.

```
[11] 0s
import pandas as pd
import numpy as np
from scipy import stats

df = pd.read_csv("blood_pressure_reduction_50.csv")

drug_group = df[df["Group"] == "Drug"]["BP_Reduction"]
placebo_group = df[df["Group"] == "Placebo"]["BP_Reduction"]

def confidence_interval(data, confidence=0.95):
    mean = np.mean(data)
    std_err = stats.sem(data)
    margin = std_err * stats.t.ppf((1 + confidence) / 2, len(data) - 1)
    return mean - margin, mean + margin

drug_ci = confidence_interval(drug_group)
placebo_ci = confidence_interval(placebo_group)

print("95% Confidence Interval for Drug Group:", drug_ci)
print("95% Confidence Interval for Placebo Group:", placebo_ci)

...
95% Confidence Interval for Drug Group: (np.float64(9.76658498706352), np.float64(12.925350543671751))
95% Confidence Interval for Placebo Group: (np.float64(1.991428176754361), np.float64(4.283933312120505))
```

EXP_20 The marketing team has conducted an A/B test to evaluate the effectiveness of two different website designs (A and B) in terms of conversion rate. They randomly divided the website visitors into two groups, with one group experiencing design A and the other experiencing design B. After a week of data collection, you now have the conversion rate data for both groups.

```
[12] 0s
import pandas as pd
from scipy import stats

df = pd.read_csv("ab_test_conversion_rates.csv")
design_A = df[df["Design"] == "A"]["Conversion_Rate"]
design_B = df[df["Design"] == "B"]["Conversion_Rate"]

t_stat, p_value = stats.ttest_ind(design_A, design_B)

print("T-statistic:", t_stat)
print("P-value:", p_value)

alpha = 0.05

if p_value < alpha:
    print("Result: There IS a statistically significant difference between Design A and Design B.")
else:
    print("Result: There is NO statistically significant difference between Design A and Design B.")

...
T-statistic: -8.551459943248732
P-value: 3.292961428703664e-15
Result: There IS a statistically significant difference between Design A and Design B.
```