LAB-DAY-4
DATE:19/1/26

16. Question: Develop a Python program to calculate the frequency distribution of words in the customer reviews dataset?

CODE:

```
from collections import Counter
import re
reviews = ["Good product", "Very good quality", "Product is good"]
cleaned = " ".join(reviews).lower()
cleaned = re.sub(r'[^\w\s]', '', cleaned)
words = cleaned.split()
freq = Counter(words)
print(freq)
```

OUTPUT:

```
••• Counter({'good': 3, 'product': 2, 'very': 1, 'quality': 1, 'is': 1})
```

Start coding or generate with AI

17. Create a Python program that fulfills these requirements and helps your team gain insights from the customer feedback data.

CODE:

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

```
df = pd.read_csv("data.csv")
# Stop words
stop_words = {"the","and","is","in","to","of"}
# Preprocessing
text = " ".join(df['feedback']).lower()
text = re.sub(r'[^\w\s]', '', text)
words = [w for w in text.split() if w not in stop_words]
freq = Counter(words)
N = int(input("Enter N: "))
top_words = freq.most_common(N)
for word, count in top_words:
    print(word, count)
# Bar plot
words, counts = zip(*top_words)
plt.bar(words, counts)
plt.xlabel("Words")
plt.ylabel("Frequency")
plt.title("Top Words in Customer Feedback")
plt.show()
```

OUTPUT:

```
Enter N: 5
product 5
good 4
service 4
quality 3
excellent 3
```

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

Question:

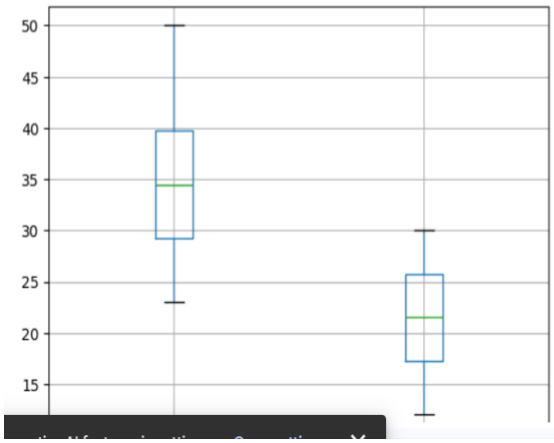
1. Calculate the mean, median and standard deviation of age and %fat using Pandas.
2. Draw the boxplots for age and %fat.
3. Draw a scatter plot and a q-q plot based on these two variables

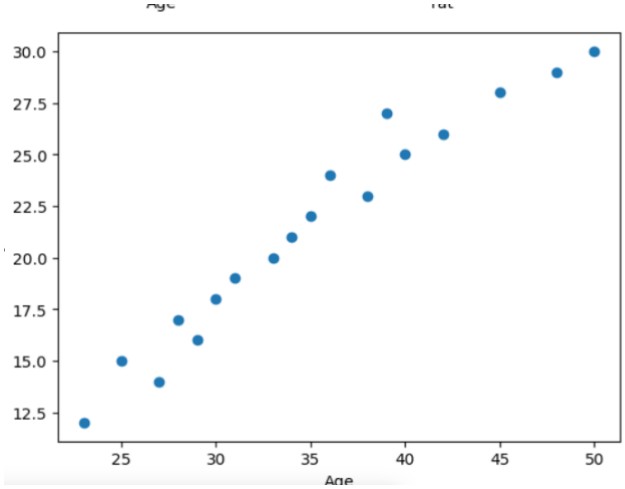
CODE:

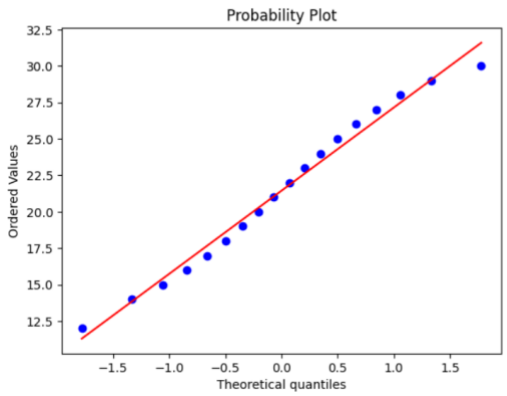
```
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats
data = {
   'Age': [23,25,30,35,40,28,33,38,45,50,29,31,34,36,42,48,27,39],
   'Fat': [12,15,18,22,25,17,20,23,28,30,16,19,21,24,26,29,14,27]
}
df = pd.DataFrame(data)
# Statistics
print(df.mean())
print(df.median())
print(df.std())
# Boxplots
df.boxplot()
plt.show()
# Scatter plot
plt.scatter(df['Age'], df['Fat'])
plt.xlabel("Age")
plt.ylabel("Body Fat %")
plt.show()
# Q-Q plot
stats.probplot(df['Fat'], plot=plt)
plt.show()
```

OUTPUT:

```
Age    35.166667
Fat    21.444444
dtype: float64
Age    34.5
Fat    21.5
dtype: float64
Age    7.793285
Fat    5.436502
dtype: float64
```







19. "What is the 95% confidence interval for the mean reduction in blood pressure for patients who received the new drug? Also, what is the 95% confidence interval for the mean reduction in blood pressure for patients who received the placebo?

CODE:

```
import numpy as np
import scipy.stats as stats
drug = np.random.normal(15, 5, 50)
placebo = np.random.normal(5, 4, 50)
def confidence_interval(data):
    mean = np.mean(data)
    std = np.std(data, ddof=1)
    n = len(data)
    ci = stats.t.interval(0.95, n-1, mean, std/np.sqrt(n))
    return ci
print("Drug CI:", confidence_interval(drug))
print("Placebo CI:", confidence_interval(placebo))
```

OUTPUT:

```
Drug CI: (np.float64(13.950982764706948),
Placebo CI: (np.float64(3.9769725041731254
```

20. "Based on the data collected from the A/B test, is there a statistically significant difference in the mean conversion rates between website design A and website design B?"

CODE:

```
import scipy.stats as stats
design_A = [0.12,0.13,0.14,0.11,0.15]
design_B = [0.16,0.18,0.17,0.19,0.20]
```

```
t_stat, p_value = stats.ttest_ind(design_A, design_B)
print("P-value:", p_value)
if p_value < 0.05:
    print("Statistically significant difference")
else:
    print("No statistically significant difference")
```

OUTPUT:

```
P-value: 0.0010528257933665399
Statistically significant difference
```