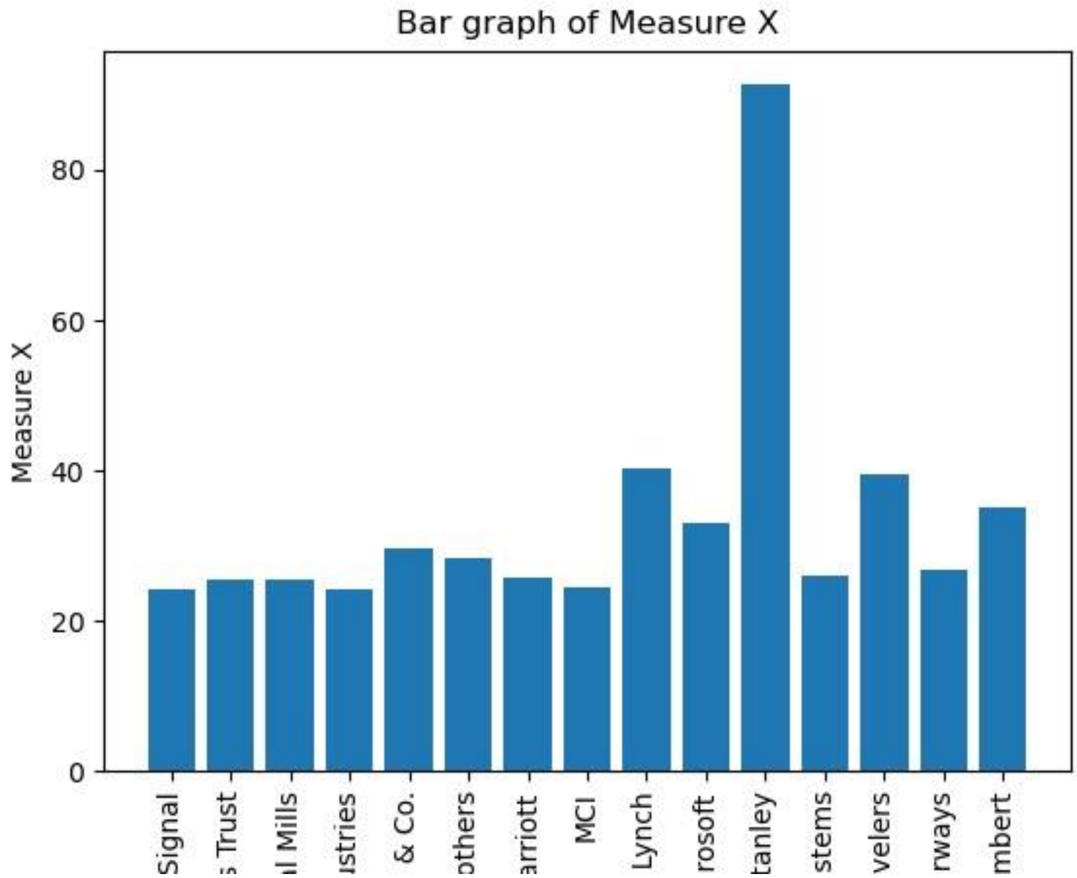
Topics: Descriptive Statistics and Probability

1. Look at the data given below. Plot the data, find the outliers and find out μ, σ, σ^2

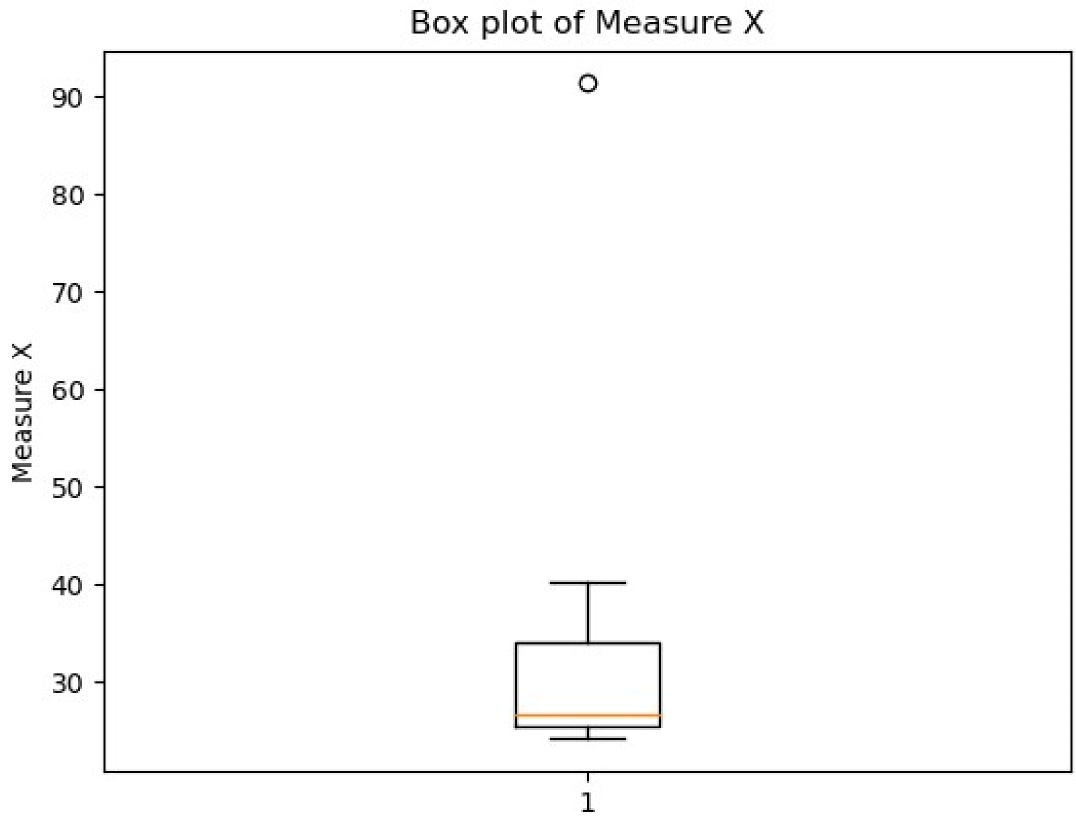
Name of company	Measure X
Allied Signal	24.23%
Bankers Trust	25.53%
General Mills	25.41%
ITT Industries	24.14%
J.P.Morgan & Co.	29.62%
Lehman Brothers	28.25%
Marriott	25.81%
MCI	24.39%
Merrill Lynch	40.26%
Microsoft	32.95%
Morgan Stanley	91.36%
Sun Microsystems	25.99%
Travelers	39.42%
US Airways	26.71%
Warner-Lambert	35.00%

```
#q1
import pandas as pd
import matplotlib.pyplot as plt

# create a dataframe from the given data
data = {'Name of company': ['Allied Signal', 'Bankers Trust', 'General Mills', 'ITT Industries', 'J.P.Morgan & Co.', 'Lehman Brothers', 'Marriott', 'MCI', 'Merrill Lynch', 'Microsoft', 'Morgan Stanley', 'Sun Microsystems', 'Travelers', 'US Airways', 'Warner-Lambert'],
    'Measure X': [24.23, 25.53, 25.41, 24.14, 29.62, 28.25, 25.81, 24.39, 40.26, 32.95, 91.36, 25.99, 39.42, 26.71, 35.00]}
df = pd.DataFrame(data)

# plot a bar graph
plt.bar(df['Name of company'], df['Measure X'])
plt.xticks(rotation=90)
plt.title('Bar graph of Measure X')
plt.xlabel('Name of company')
plt.ylabel('Measure X')
plt.show()

# plot a boxplot
plt.boxplot(df['Measure X'])
```

```
plt.title('Box plot of Measure X')
plt.ylabel('Measure X')
plt.show()

# find the outliers
q1 = df['Measure X'].quantile(0.25)
q3 = df['Measure X'].quantile(0.75)
iqr = q3 - q1
lower_bound = q1 - 1.5 * iqr
upper_bound = q3 + 1.5 * iqr
outliers = df[(df['Measure X'] < lower_bound) | (df['Measure X'] > upper_bound)]
print('Outliers:')
print(outliers)

# find the mean, variance, and standard deviation
mean = df['Measure X'].mean()
variance = df['Measure X'].var()
std_dev = df['Measure X'].std()

print('Mean:', mean)
print('Variance:', variance)
print('Standard deviation:', std_dev)
```

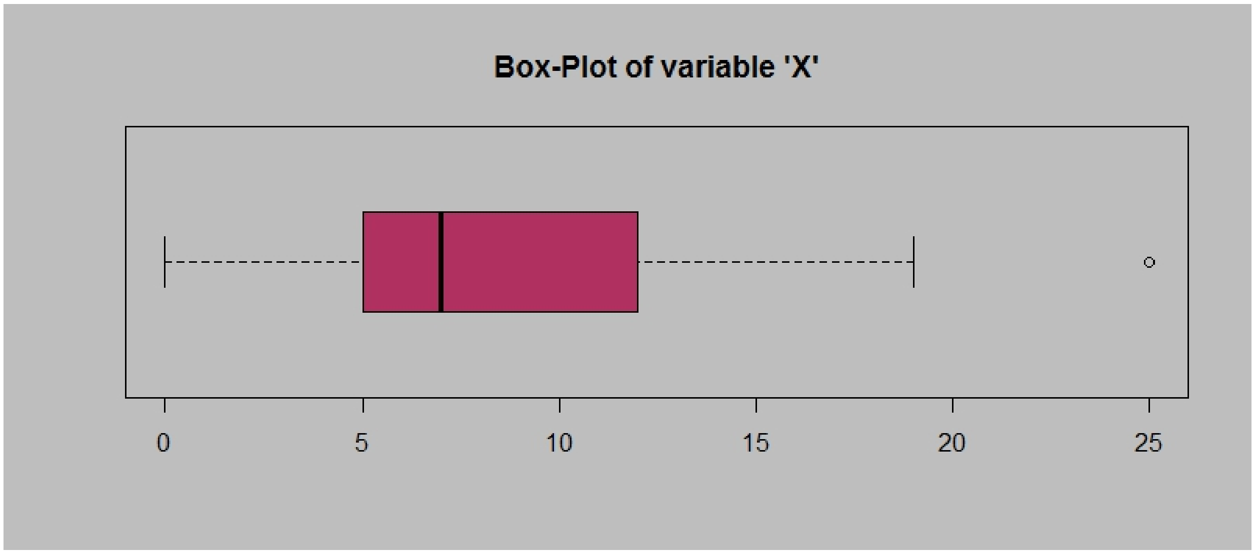


Questions referred to from *Aczel A., Sounderpandian J., Complete Business Statistics (7ed.)*



Outliers:
  Name of company Measure X
10 Morgan Stanley   91.36
Mean: 33.27133333333333
Variance: 287.1466123809524
Standard deviation: 16.945400921222028

2.



Answer the following three questions based on the box-plot above.

(i) What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.

(ii) What can we say about the skewness of this dataset?

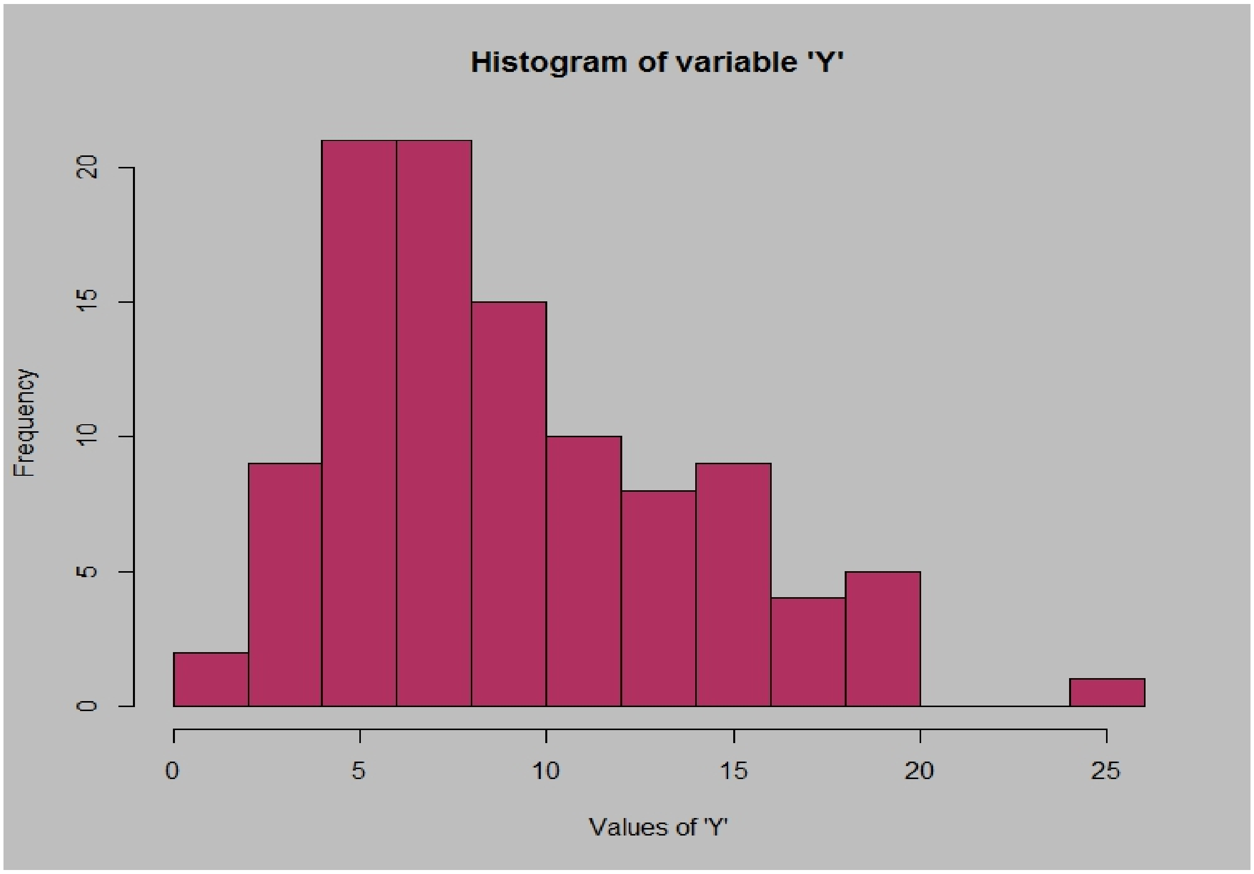
(iii) If it was found that the data point with the value 25 is actually 2.5, how would the new box-plot be affected?

(i) The inter-quartile range of this dataset is approximately **7.5** (12.5 - 5). This value implies that 50% of the data points lie between 5 and 12.5.

(ii) The dataset appears to be positively skewed as the median (7.5) is closer to the lower quartile (5) than the upper quartile (12.5).

(iii) If the data point with the value 25 is actually 2.5, the new box-plot would be affected in the following ways:

- The range of the dataset would decrease from 25 to 12.5.
- The upper quartile would decrease from 12.5 to approximately 10.
- The median would decrease from 7.5 to approximately 6.25.
- The dataset would become more negatively skewed.



Answer the following three questions based on the histogram above.

(i) Where would the mode of this dataset lie?

(ii) Comment on the skewness of the dataset.

(iii) Suppose that the above histogram and the box-plot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset.

(i) The mode of this dataset would lie around the value of 5 to **10**, as it is the highest frequency.

(ii) The dataset is positively skewed, as the tail of the histogram is longer on the right side.

(iii) The histogram and box-plot complement each other in providing information about the dataset. The histogram shows the frequency distribution of the data, while the box-plot shows the median, quartiles, and outliers. Together, they provide a more complete picture of the data.

3. AT&T was running commercials in 1990 aimed at luring back customers who had switched to one of the other long-distance phone service providers. One such commercial shows a businessman trying to reach Phoenix and mistakenly getting Fiji, where a half-naked native on a beach responds incomprehensibly in Polynesian. When asked about this advertisement, AT&T admitted that the portrayed incident did not actually take place but added that this was an enactment of something that "could happen." Suppose that one in 200 long-distance telephone calls is misdirected. What is the probability that at least one in five attempted telephone calls reaches the wrong number? (Assume independence of attempts.)

```
#Q3
import math

# probability of call misdirecting
p = 1/200

# probability of call not misdirecting
q = 1 - p

# number of calls
n = 5

# probability that at least one in five attempted telephone calls reaches the wrong number
P = 1 - math.pow(q, n)
probabablity_percentage=(1 - math.pow(q, n))*100

print(f"The probablity is found to be {P}and percentage is{probabablity_percentage}")
```

```
The probablity is found to be 0.024751246878125O2and percentage is2.475124687812502
```

4. Returns on a certain business venture, to the nearest $1,000, are known to follow the following probability distribution

x	P(x)
-2,000	0.1
-1,000	0.1
0	0.2
1000	0.2
2000	0.3
3000	0.1

(i) What is the most likely monetary outcome of the business venture?
(ii) Is the venture likely to be successful? Explain
(iii) What is the long-term average earning of business ventures of this kind? Explain
(iv) What is the good measure of the risk involved in a venture of this kind? Compute this measure

```
#q4

import math


# probability distribution

p = [-2000, -1000, 0, 1000, 2000, 3000]

q = [0.1, 0.1, 0.2, 0.2, 0.3, 0.1]


# (i) most likely monetary outcome
```

```
most_likely_outcome = p[q.index(max(q))]

print("(i) The most likely monetary outcome of the business venture is ${:,}.".format(most_likely_outcome))


# (ii) is the venture likely to be successful?

p_success = sum(q[2:])

if p_success > 0.5:

    print("(ii) Yes, the venture is likely to be successful. There is a {:.2f}% chance for this venture to make a profit.".format(p_success*100))

else:

    print("(ii) No, the venture is not likely to be successful. There is a {:.2f}% chance for this venture to make a profit.".format((1-p_success)*100))


# (iii) long-term average earning

long_term_average = sum([p[i]*q[i] for i in range(len(p))])

print("(iii) The long-term average earning of business ventures of this kind is ${:,}.".format(int(long_term_average)))


# (iv) good measure of risk

variance = sum([q[i]*(p[i]-long_term_average)**2 for i in range(len(p))])

risk = math.sqrt(variance)

print("(iv) The good measure of the risk involved in a venture of this kind is ${:,}.".format(int(risk)))
```

output

(i) The most likely monetary outcome of the business venture is $2,000. (ii) Yes, the venture is likely to be successful. There is a 80.00% chance for this venture to make a profit.

(iii) The long-term average earning of business ventures of this kind is $800.

(iv) The good measure of the risk involved in a venture of this kind is $1,469.