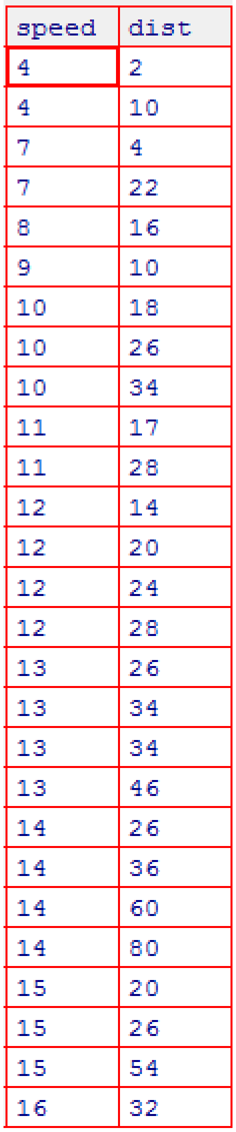
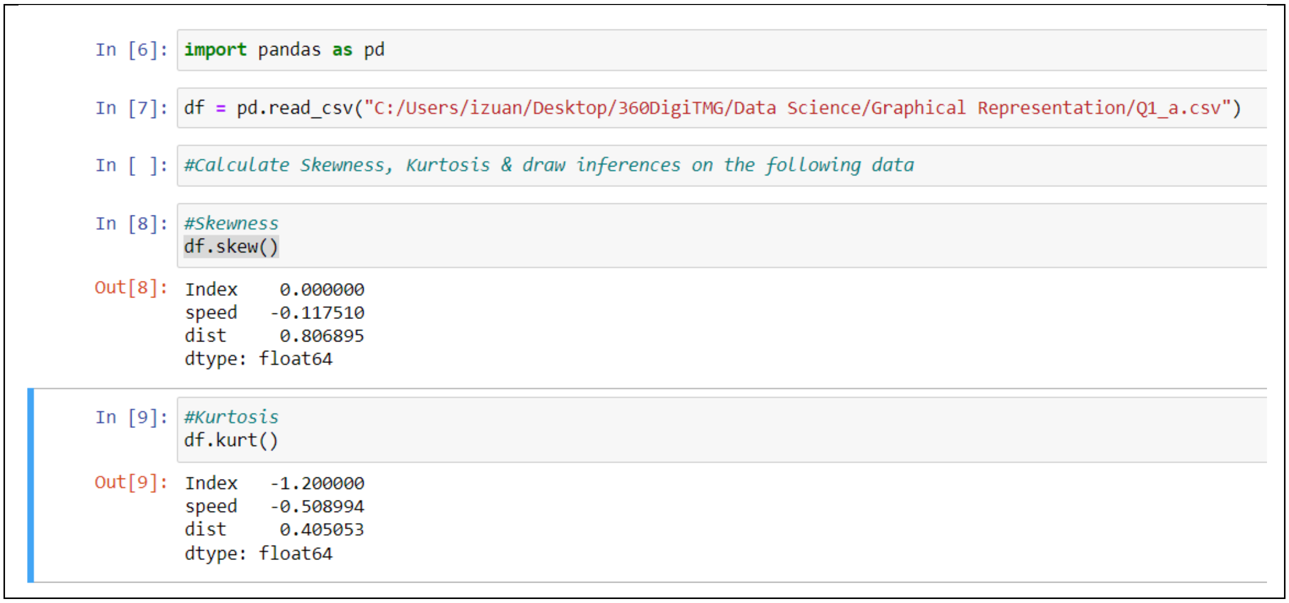
Q1) Calculate Skewness, Kurtosis & draw inferences on the following data

a. Cars speed and distance

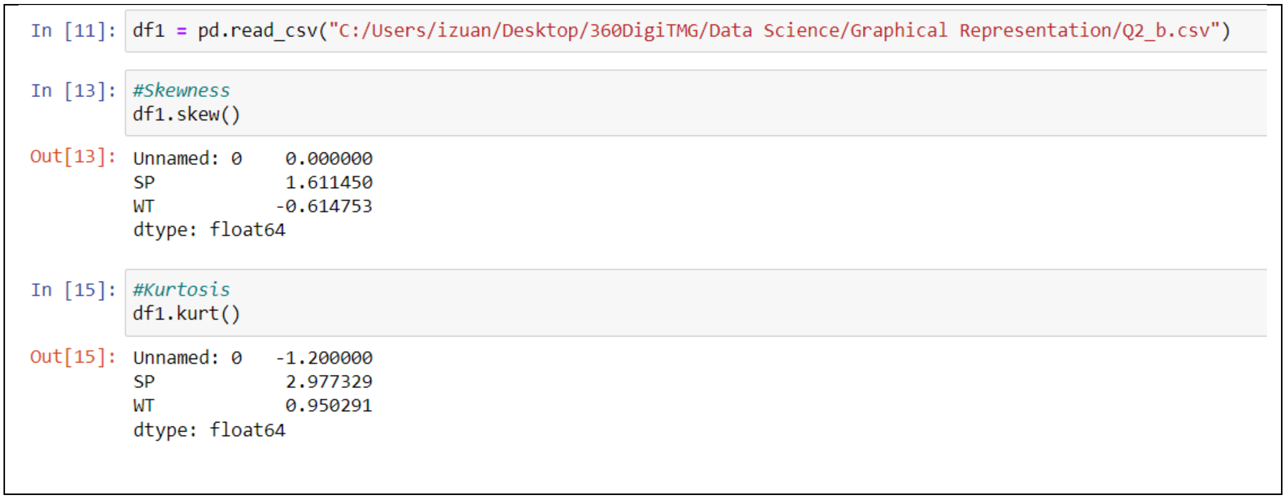


speed	dist
4	2
4	10
7	4
7	22
8	16
9	10
10	18
10	26
10	34
11	17
11	28
12	14
12	20
12	24
12	28
13	26
13	34
13	34
13	46
14	26
14	36
14	60
14	80
15	20
15	26
15	54
16	32

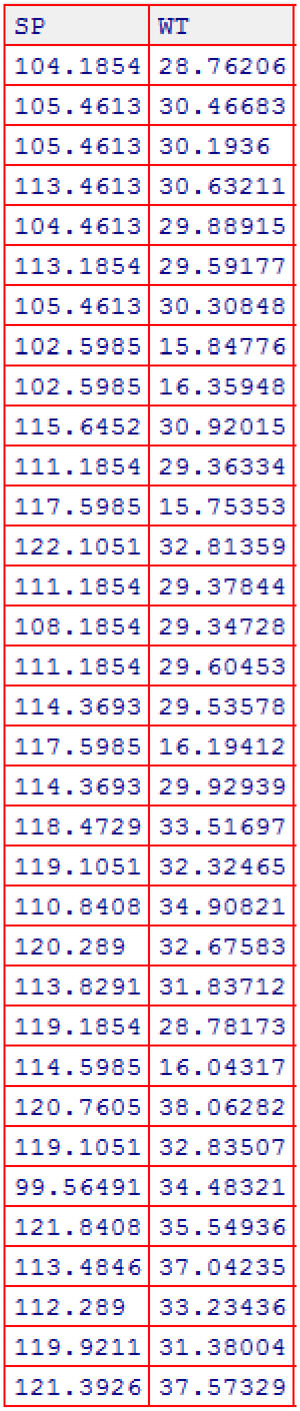


```
In [6]: import pandas as pd

In [7]: df = pd.read_csv("C:/Users/izuan/Desktop/360DigiTMG/Data Science/Graphical Representation/Q1_a.csv")

In [ ]: #Calculate Skewness, Kurtosis & draw inferences on the following data

In [8]: #Skewness
        df.skew()

Out[8]: Index     0.000000
        speed    -0.117510
        dist      0.806895
        dtype: float64

In [9]: #Kurtosis
        df.kurt()

Out[9]: Index    -1.200000
        speed    -0.508994
        dist      0.405053
        dtype: float64
```

b. Top Speed (SP) and Weight (WT)

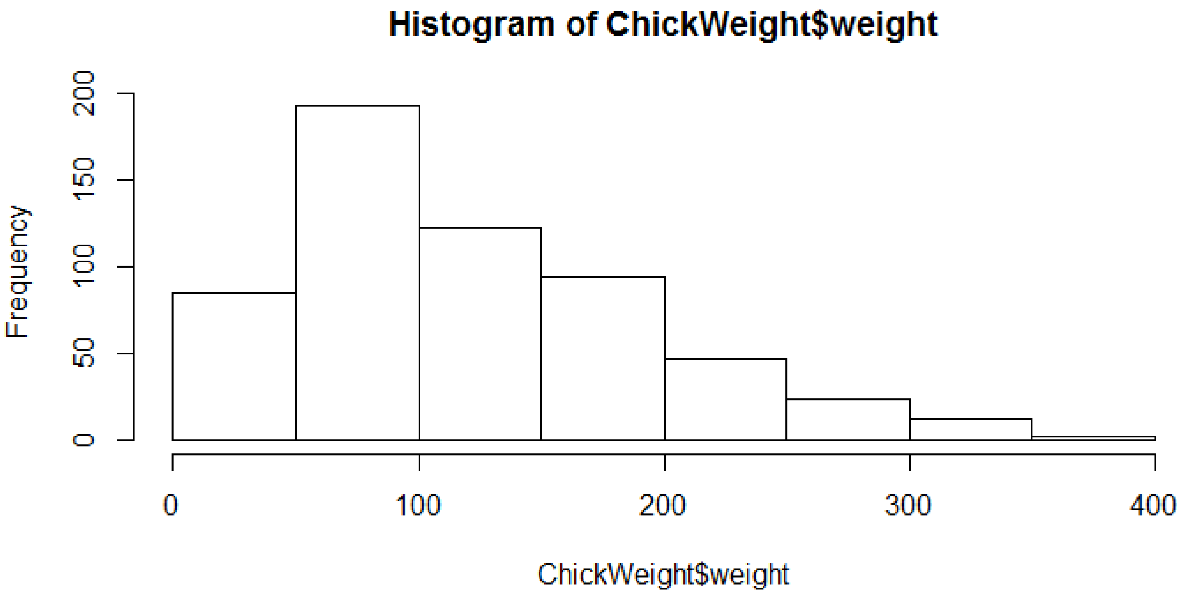


```
In [11]: df1 = pd.read_csv("C:/Users/izuan/Desktop/360DigiTMG/Data Science/Graphical Representation/Q2_b.csv")

In [13]: #Skewness
         df1.skew()

Out[13]: Unnamed: 0     0.000000
         SP             1.611450
         WT            -0.614753
         dtype: float64

In [15]: #Kurtosis
         df1.kurt()

Out[15]: Unnamed: 0    -1.200000
         SP             2.977329
         WT             0.950291
         dtype: float64
```

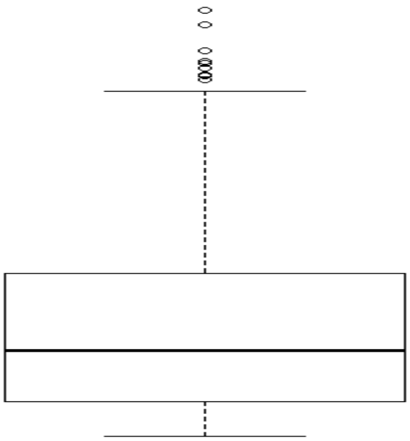


SP	WT
104.1854	28.76206
105.4613	30.46683
105.4613	30.1936
113.4613	30.63211
104.4613	29.88915
113.1854	29.59177
105.4613	30.30848
102.5985	15.84776
102.5985	16.35948
115.6452	30.92015
111.1854	29.36334
117.5985	15.75353
122.1051	32.81359
111.1854	29.37844
108.1854	29.34728
111.1854	29.60453
114.3693	29.53578
117.5985	16.19412
114.3693	29.92939
118.4729	33.51697
119.1051	32.32465
110.8408	34.90821
120.289	32.67583
113.8291	31.83712
119.1854	28.78173
114.5985	16.04317
120.7605	38.06282
119.1051	32.83507
99.56491	34.48321
121.8408	35.54936
113.4846	37.04235
112.289	33.23436
119.9211	31.38004
121.3926	37.57329

Q2) Draw inferences about the following boxplot & histogram

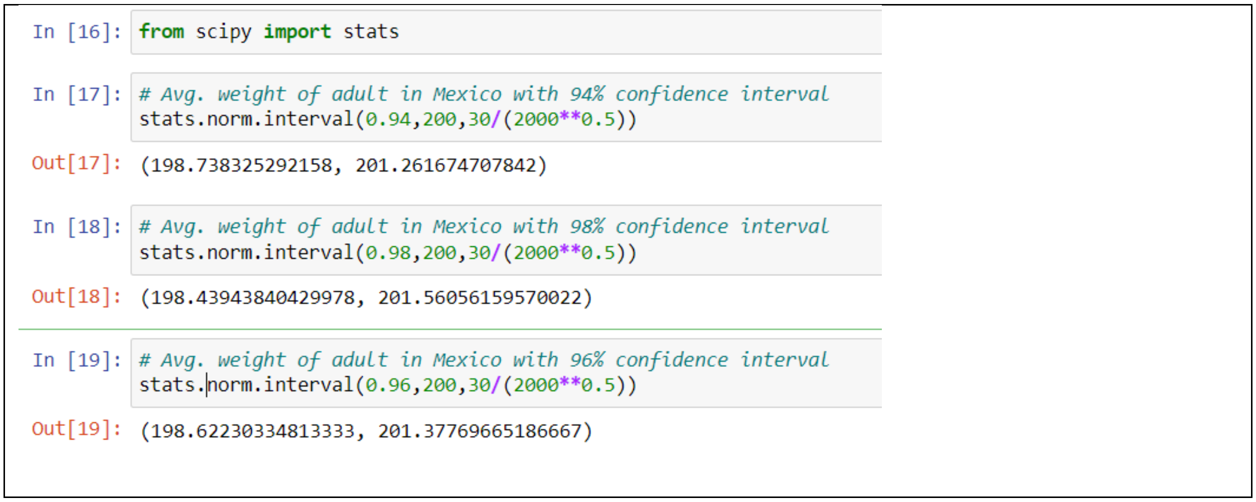


- Chick weight data is right skewed or positively skewed.
- More than 50% Chick Weight is between 50 to 150.
- Most of the chick weight is between 50 to 100.



- The data is right skewed.
- There are outliers at upper side.

Q3) Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

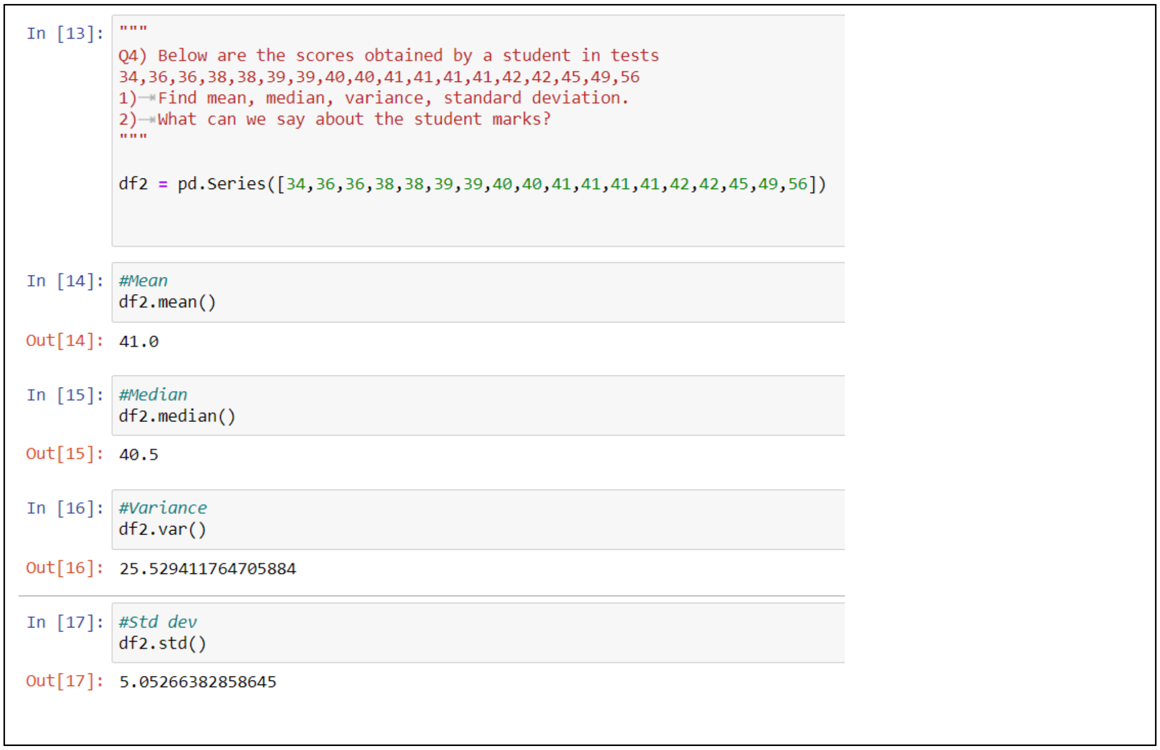


```
In [16]: from scipy import stats

In [17]: # Avg. weight of adult in Mexico with 94% confidence interval
         stats.norm.interval(0.94,200,30/(2000**0.5))

Out[17]: (198.738325292158, 201.261674707842)

In [18]: # Avg. weight of adult in Mexico with 98% confidence interval
         stats.norm.interval(0.98,200,30/(2000**0.5))

Out[18]: (198.43943840429978, 201.56056159570022)

In [19]: # Avg. weight of adult in Mexico with 96% confidence interval
         stats.norm.interval(0.96,200,30/(2000**0.5))

Out[19]: (198.62230334813333, 201.37769665186667)
```

Q4) Below are the scores obtained by a student in tests

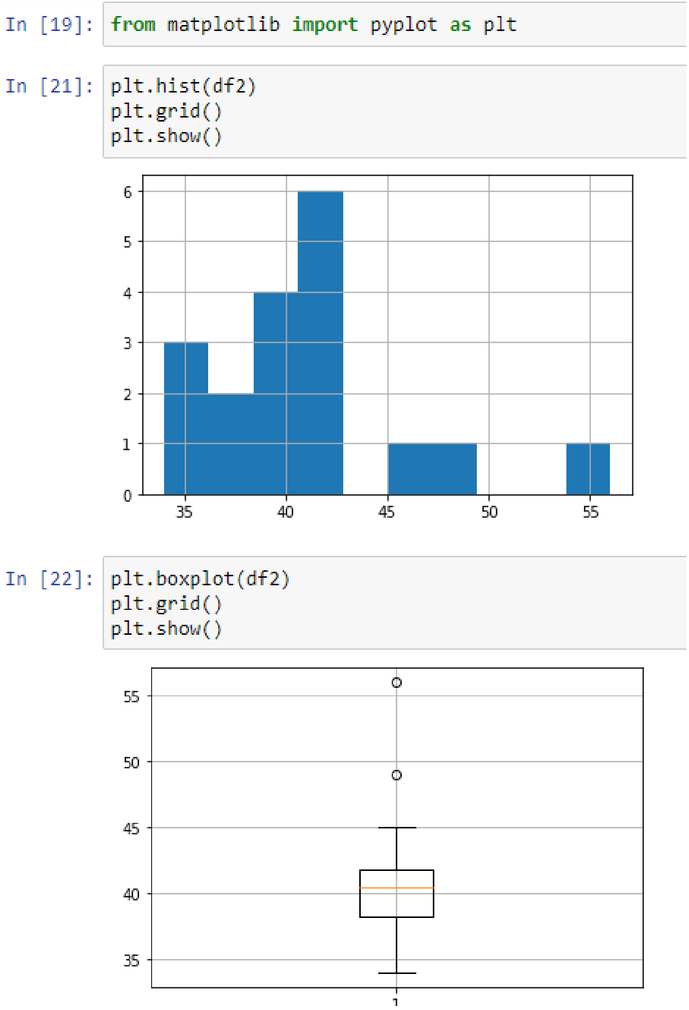
34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

1) Find mean, median, variance, standard deviation.



```
In [13]: """
         Q4) Below are the scores obtained by a student in tests
         34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56
         1)  Find mean, median, variance, standard deviation.
         2)  What can we say about the student marks?
         """

         df2 = pd.Series([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

In [14]: #Mean
         df2.mean()

Out[14]: 41.0

In [15]: #Median
         df2.median()

Out[15]: 40.5

In [16]: #Variance
         df2.var()

Out[16]: 25.529411764705884

In [17]: #Std dev
         df2.std()

Out[17]: 5.05266382858645
```

2) What can we say about the student marks?



In [19]:
```
from matplotlib import pyplot as plt
```

In [21]:
```
plt.hist(df2)
plt.grid()
plt.show()
```

In [22]:
```
plt.boxplot(df2)
plt.grid()
plt.show()
```

- From above plot we can say that the mean of students' marks is 41 which is slightly greater than median. Most of the students got marks in the range 41-42 and there are two outliers of 49,56.

Q5) What is the nature of skewness when mean, median of data are equal?

Normalized skewness

Q6) What is the nature of skewness when mean > median?

Right skewed

Q7) What is the nature of skewness when median > mean?

Left skewed

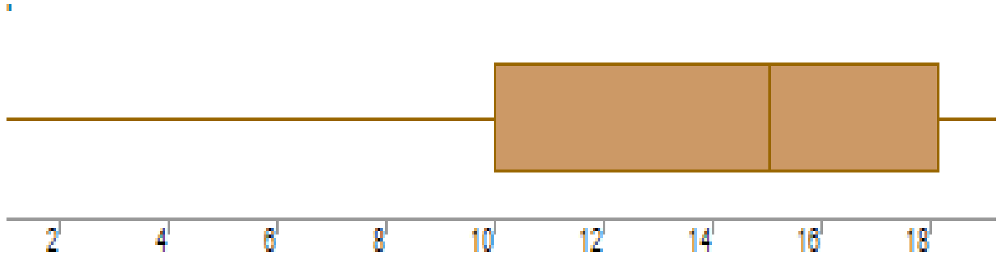
Q8) What does positive kurtosis value indicates for a data?

Indicates a "heavy-tailed" distribution

Q9) What does negative kurtosis value indicates for a data?

Indicates a "light-tailed" distribution

Q10) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

Median = Between 14 – 16
Q3 = 18
Q1 = 10
Min = Less than 2
Max = More than 18

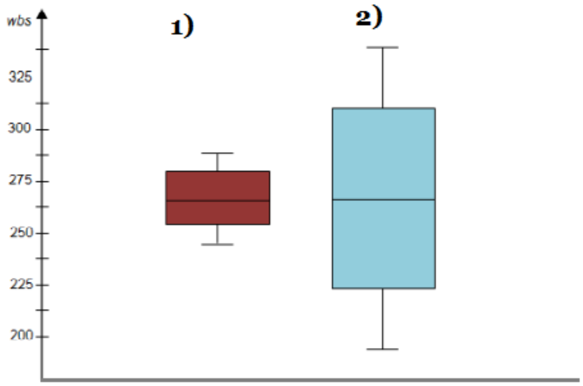
What is nature of skewness of the data?

Skewed to the left.

What will be the IQR of the data (approximately)?

18 – 10 = 8

Q11) Comment on the below Boxplot visualizations?

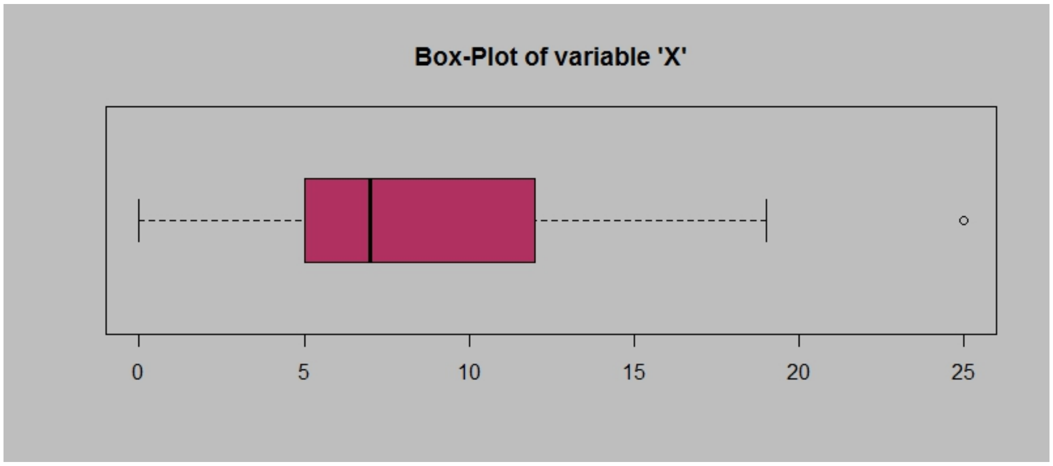


Median appears to be overlapped.
IQR blue box > IQR red box.
Min-Max blue box > Min-Max red box.

Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

Median = ~ 262.5
IQR red = ~275 - ~250 = ~25
IQR blue = ~312.5 - ~225 = ~87.5
ΔIQR = ~87.5 - ~25 = ~62.5

Q12)



Answer the following three questions based on the boxplot above.

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

IQR = ~12 - ~5 = ~7

Implies the middle 50% values in the dataset have a spread of 7.

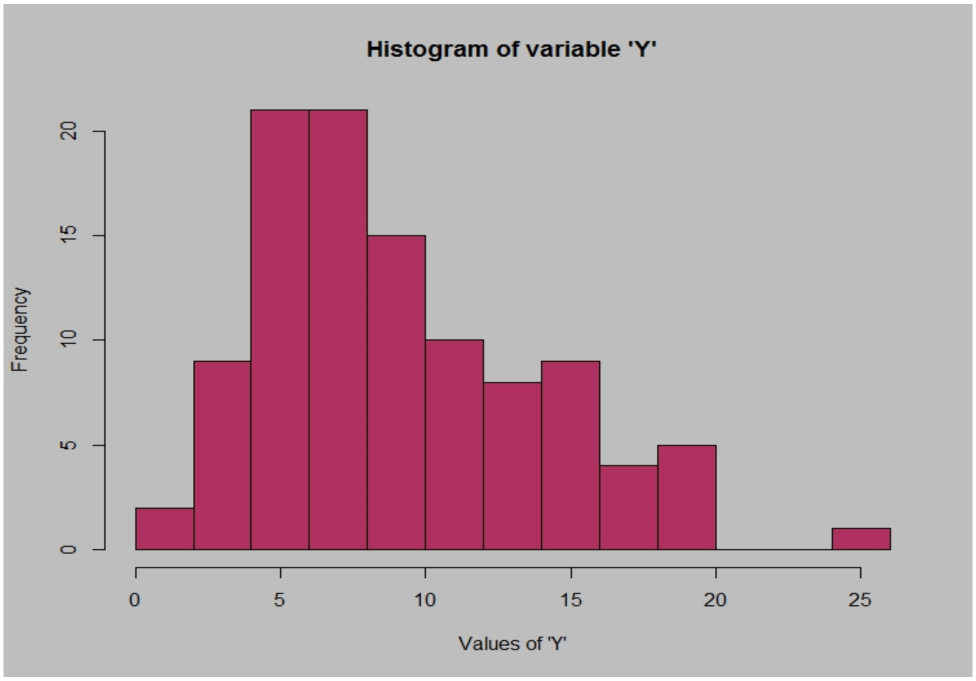
(ii) What can we say about the skewness of this dataset?

Skewed to the right

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

No more outlier

Q13)



Answer the following three questions based on the histogram above.

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

5 - 8

(ii) Comment on the skewness of the dataset.

Skewed to the right

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

Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets.