

Q1) Identify the Data type for the Following:

Activity	Data Type
Number of beatings from Wife	Discrete Data
Results of rolling a dice	Discrete Data
Weight of a person	Continuous Data
Weight of Gold	Continuous Data
Distance between two places	Continuous Data
Length of a leaf	Continuous Data
Dog's weight	Continuous Data
Blue Color	Nominal Data
Number of kids	Discrete Data
Number of tickets in Indian railways	Discrete Data
Number of times married	Discrete Data
Gender (Male or Female)	Nominal Data

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

Data	Data Type
Gender	Nominal Data
High School Class Ranking	Ordinal Data
Celsius Temperature	Interval Data
Weight	Interval Data
Hair Color	Nominal Data
Socioeconomic Status	Ordinal Data
Fahrenheit Temperature	Interval Data
Height	Interval Data
Type of living accommodation	Nominal Data
Level of Agreement	Ordinal Data
IQ(Intelligence Scale)	Interval Data
Sales Figures	Ratio Data
Blood Group	Nominal Data
Time Of Day	Interval Data
Time on a Clock with Hands	Interval Data
Number of Children	Ratio Data
Religious Preference	Nominal Data

Barometer Pressure	Interval Data
SAT Scores	Interval Data
Years of Education	Ratio Data

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

=  $3/8$  or 37.5%

Q4) Two Dice are rolled, find the probability that sum is

- a) Equal to 1
- b) Less than or equal to 4
- c) Sum is divisible by 2 and 3

a) Probability that sum is equal to 1 = 0

b) Probability that sum is less than or equal to 4 =  $6/36 = 16.66\%$

c) Probability that sum is divisible by 2 & 3 =  $5/36 = 13.88\%$

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

=  $10/21 = 47.61\%$

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

CHILD	Candies count	Probability
A	1	0.015
B	4	0.20
C	3	0.65
D	5	0.005
E	6	0.01
F	2	0.120

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

Ans:

Expected Value =  $\sum (\text{Value} * \text{Probability})$  or Mean

Expected Number of Candies for randomly selected child =

$[(1*0.014)+(4*0.20)+(3*0.65)+(5*0.005)+(6*0.01)+(2*0.120)]$

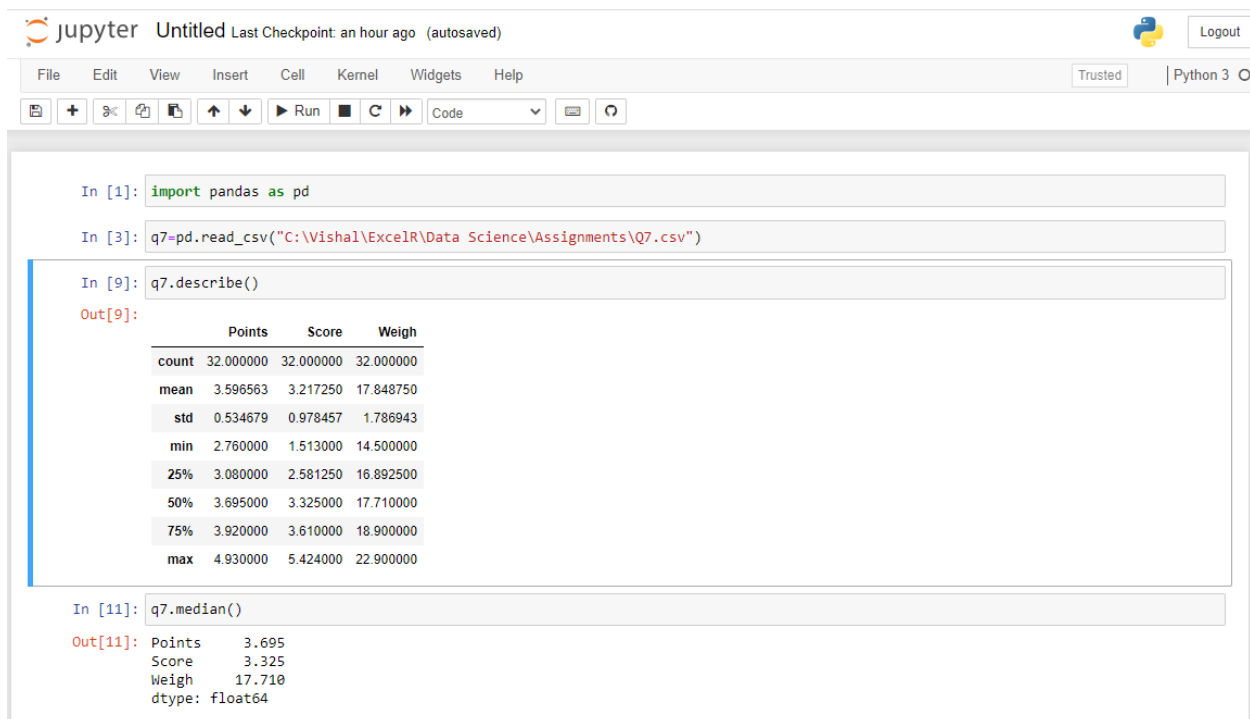
=3.09

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

- For Points, Score, Weigh>  
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

Use Q7.csv file

Using Python,



```

In [1]: import pandas as pd

In [3]: q7=pd.read_csv("C:\\Vishal\\ExcelR\\Data Science\\Assignments\\Q7.csv")

In [9]: q7.describe()
Out[9]:
```

	Points	Score	Weigh
count	32.000000	32.000000	32.000000
mean	3.596563	3.217250	17.848750
std	0.534679	0.978457	1.786943
min	2.760000	1.513000	14.500000
25%	3.080000	2.581250	16.892500
50%	3.695000	3.325000	17.710000
75%	3.920000	3.610000	18.900000
max	4.930000	5.424000	22.900000

```

In [11]: q7.median()
Out[11]: Points    3.695
         Score      3.325
         Weigh     17.710
         dtype: float64

```

```
In [12]: q7.mode()
```

```
Out[12]:
```

	Unnamed: 0	Points	Score	Weigh
0	AMC Javelin	3.07	3.44	17.02
1	Cadillac Fleetwood	3.92	NaN	18.90
2	Camaro Z28	NaN	NaN	NaN

```
In [13]: q7.var()
```

```
Out[13]: Points    0.285881  
Score      0.957379  
Weigh      3.193166  
dtype: float64
```

	Points	Score	Weigh
Mean	3.596563	3.217250	17.848750
Median	3.695	3.325	17.710
Mode	3.92 & 3.07	3.44	17.02 & 18.90
Variance	0.285881	0.957379	3.193166
Standard Deviation	0.534679	0.978457	1.786943
Range	2.76 – 4.93	1.513 – 5.424	14.5 – 22.9

Q8) Calculate Expected Value for the problem below

a) The weights (X) of patients at a clinic (in pounds), are  
108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

Expected Value =  $\sum (\text{Value} * \text{Probability})$  or Mean

Here, probability of each patient getting selected is 1/9

Hence, Expected Value =  $(1308) * 1/9 = 145.33$

Q9) Calculate Skewness, Kurtosis & draw inferences on the following data

**Cars speed and distance**

Use Q9\_a.csv

```
In [15]: q9a=pd.read_csv("C:\Vishal\ExcelR\Data Science\Assignments\Q9_a.csv")
```

```
In [17]: q9a.skew()
```

```
Out[17]: Index      0.000000  
speed    -0.117510  
dist      0.806895  
dtype: float64
```

```
In [18]: q9a.kurtosis()
```

```
Out[18]: Index      -1.200000  
speed    -0.508994  
dist      0.405053  
dtype: float64
```

```
In [ ]:
```

	Skewness	Kurtosis
<b>Car Speed</b>	-0.117510	-0.508994
<b>Distance</b>	0.806895	0.405053

## SP and Weight(WT)

### Use Q9\_b.csv

```
In [19]: q9b=pd.read_csv("C:\Vishal\ExcelR\Data Science\Assignments\Q9_b.csv")
```

```
In [20]: q9b.SP.skew()
```

```
Out[20]: 1.6114501961773586
```

```
In [22]: q9b.WT.skew()
```

```
Out[22]: -0.6147533255357768
```

```
In [23]: q9b.SP.kurtosis()
```

```
Out[23]: 2.9773289437871835
```

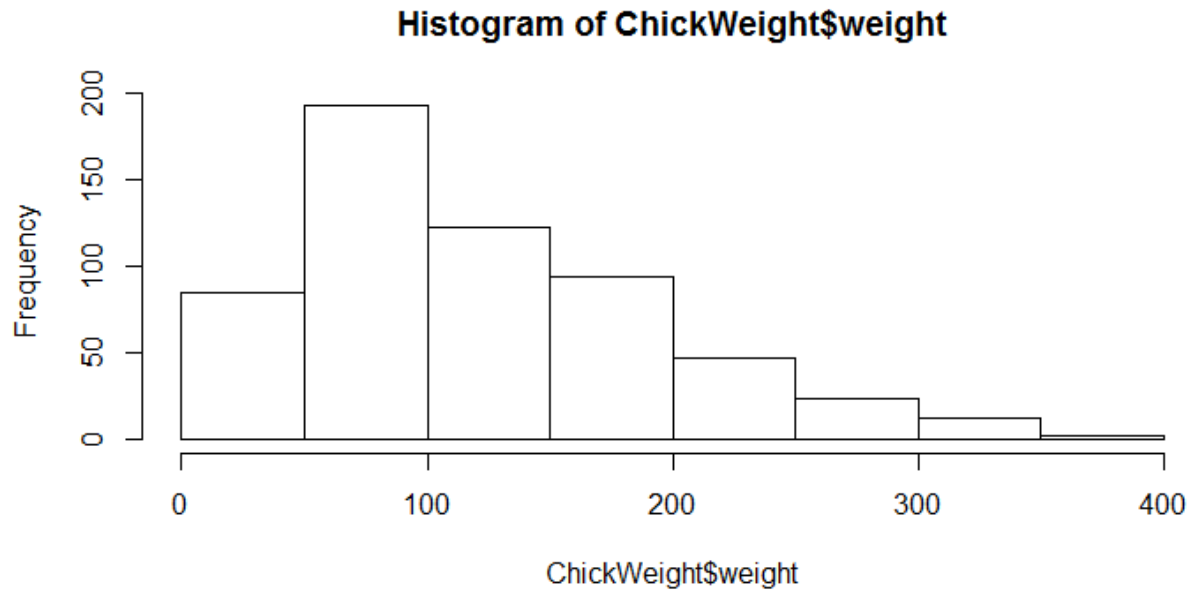
```
In [24]: q9b.WT.kurtosis()
```

```
Out[24]: 0.9502914910300326
```

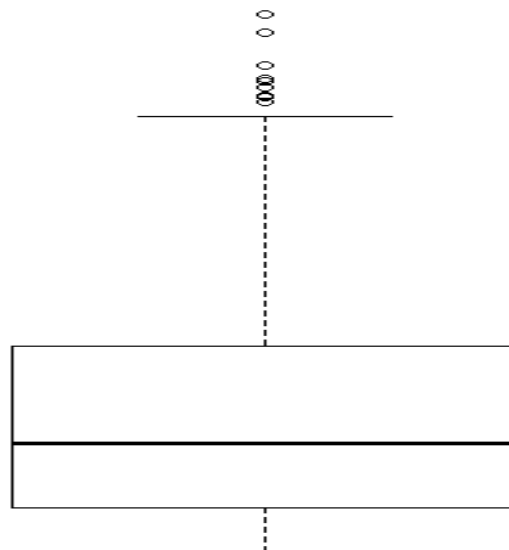
```
In [ ]:
```

	Skewness	Kurtosis
<b>SP</b>	1.611450	2.977328
<b>WT</b>	-0.614753	0.950291

**Q10) Draw inferences about the following boxplot & histogram**



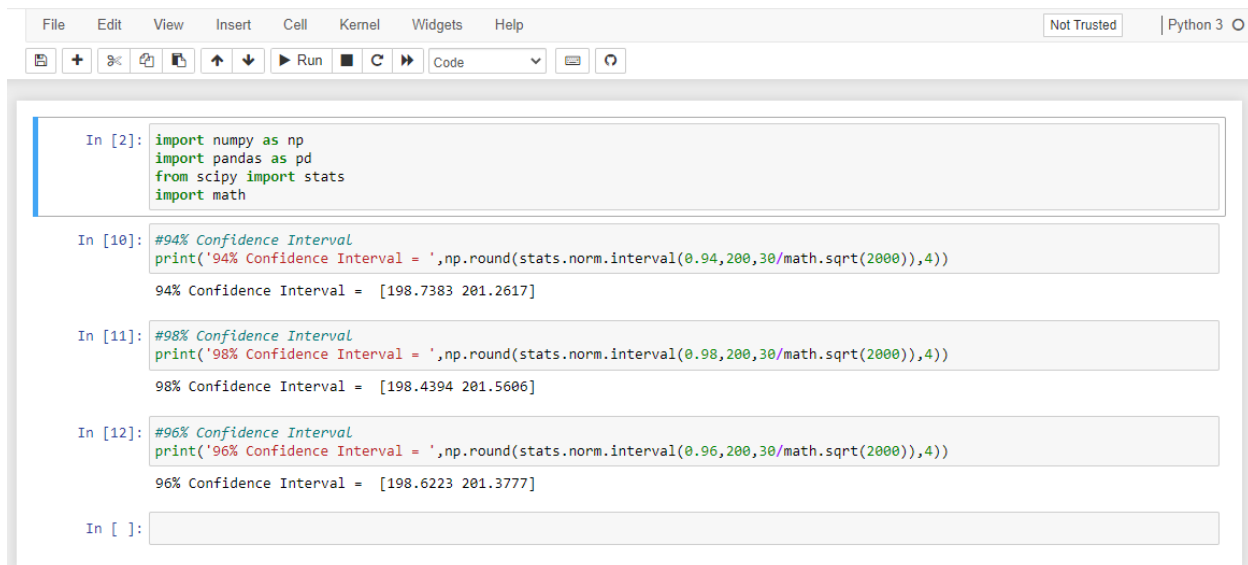
From Histogram, we can conclude that, mode will lie between weights 50-100 and the distribution is positively skewed.



From the boxplot, we can conclude that, outliers are present on upper side.

**Q11)** 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?

Using Python,

A screenshot of a Jupyter Notebook interface. The top bar shows menu items: File, Edit, View, Insert, Cell, Kernel, Widgets, Help. On the right, it says 'Not Trusted' and 'Python 3'. Below the menu is a toolbar with icons for saving, adding cells, undo, redo, and running code. The notebook area contains four input cells. The first cell (In [2]:) imports numpy as np, pandas as pd, scipy stats, and math. The second cell (In [10]:) calculates the 94% confidence interval using stats.norm.interval(0.94, 200, 30/math.sqrt(2000)), 4) and prints the result: '94% Confidence Interval = [198.7383 201.2617]'. The third cell (In [11]:) calculates the 98% confidence interval using stats.norm.interval(0.98, 200, 30/math.sqrt(2000)), 4) and prints the result: '98% Confidence Interval = [198.4394 201.5606]'. The fourth cell (In [12]:) calculates the 96% confidence interval using stats.norm.interval(0.96, 200, 30/math.sqrt(2000)), 4) and prints the result: '96% Confidence Interval = [198.6223 201.3777]'. The last cell (In [ ]:) is empty.

```
In [2]: import numpy as np
import pandas as pd
from scipy import stats
import math

In [10]: #94% Confidence Interval
print('94% Confidence Interval = ', np.round(stats.norm.interval(0.94, 200, 30/math.sqrt(2000)), 4))
94% Confidence Interval = [198.7383 201.2617]

In [11]: #98% Confidence Interval
print('98% Confidence Interval = ', np.round(stats.norm.interval(0.98, 200, 30/math.sqrt(2000)), 4))
98% Confidence Interval = [198.4394 201.5606]

In [12]: #96% Confidence Interval
print('96% Confidence Interval = ', np.round(stats.norm.interval(0.96, 200, 30/math.sqrt(2000)), 4))
96% Confidence Interval = [198.6223 201.3777]

In [ ]:
```

Hence,

94% Confidence Interval = (198.7383, 201.2617)

98% Confidence Interval = (198.4394, 201.5606)

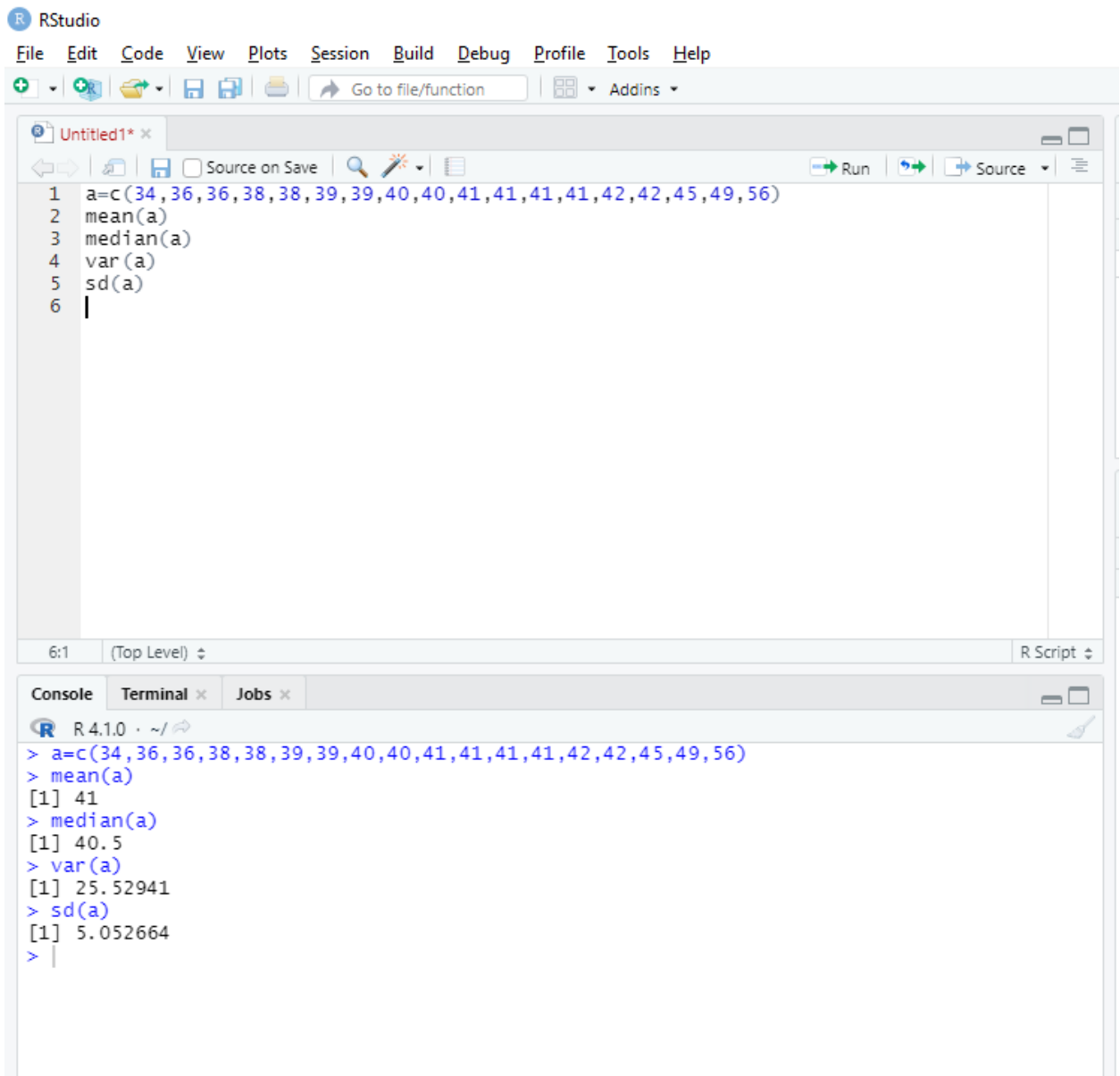
96% Confidence Interval = (198.6223, 201.3777)

**Q12)** 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.

## Using R:



The screenshot shows the RStudio environment. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for saving, running, and other functions. The main editor window, titled 'Untitled1\*', contains the following R code:

```
1 a=c(34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)
2 mean(a)
3 median(a)
4 var(a)
5 sd(a)
6 |
```

The bottom panel shows the Console window with the output of the code:

```
> a=c(34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)
> mean(a)
[1] 41
> median(a)
[1] 40.5
> var(a)
[1] 25.52941
> sd(a)
[1] 5.052664
> |
```

Hence,

Mean: 41

Median: 40.5

Variance: 25.5294

Standard Deviation: 5.052

2) What can we say about the student marks?

We can conclude that, the average marks obtained by students are 41 and maximum students scored 41 marks.



Q13) What is the nature of skewness when mean, median of data are equal?

When, mean & median of data are equal, the value of skewness will be zero. This will be a normal distribution (symmetric).

Q14) What is the nature of skewness when mean > median ?

When, mean of data is greater than the median, then the distribution will be positively skewed.

Q15) What is the nature of skewness when median > mean?

When, mean of data is less than the median, then the distribution will be negatively skewed.

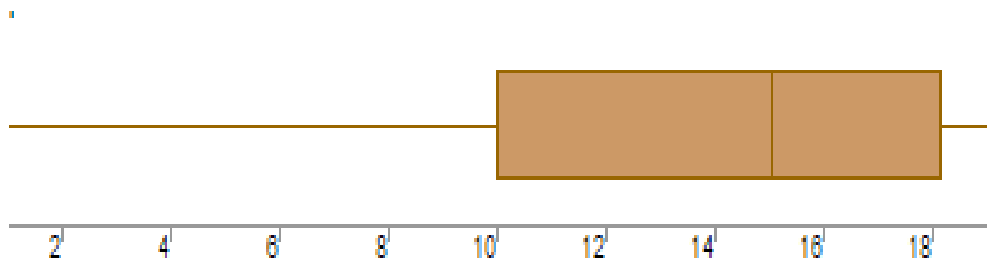
Q16) What does positive kurtosis value indicates for a data ?

Positive kurtosis indicates that, the distribution will have sharp peak and heavier tails. The distribution is known as Leptokurtic.

Q17) What does negative kurtosis value indicates for a data?

Negative kurtosis indicates that, the distribution will have flat peak and lighter tails. The distribution is known as Platykurtic.

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

We can conclude that, the distribution is not normal. Also, for this distribution, Mean < Median < Mode.

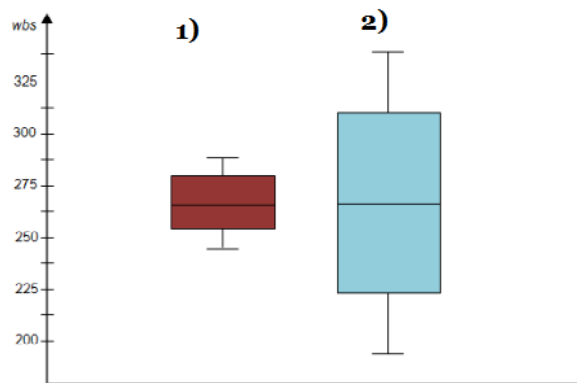
What is nature of skewness of the data?

As, Mean < Median < Mode and also  $Q_3 - Q_2 < Q_2 - Q_1$ , the data/distribution is Left or Negatively Skewed.

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

IQR of data will be between 10 to 18.1

Q19) Comment on the below Boxplot visualizations?



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

For boxplot 1 has minimum value which is almost more than 25% of minimum value of the boxplot 2.

The data in boxplot 1 is more consistent and tends towards upper side as compared to boxplot 2 which is having more variable data.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars\$MPG

- a.  $P(\text{MPG} > 38)$
- b.  $P(\text{MPG} < 40)$
- c.  $P(20 < \text{MPG} < 50)$

Probability = (Interested Events)/(Total Events)

$P(\text{MPG} > 38) = 33/81 = 40.74\%$

$P(\text{MPG} < 40) = 61/81 = 75.30\%$

$P(20 < \text{MPG} < 50) = 69/81 = 85.18\%$

Q 21) Check whether the data follows normal distribution

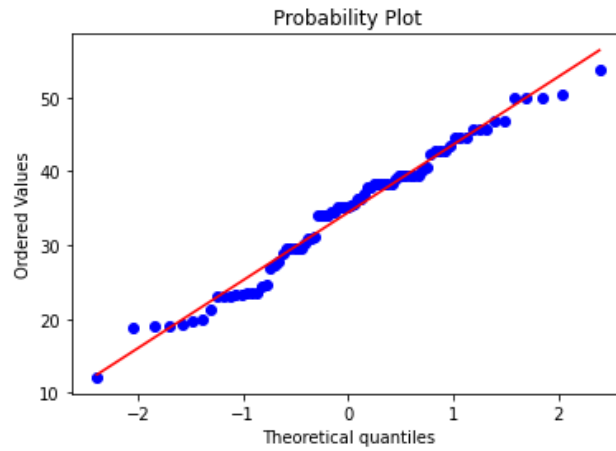
- a) Check whether the MPG of Cars follows Normal Distribution  
Dataset: Cars.csv

MPG of cars follows Normal Distribution

- b) Check Whether the Adaptive Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution  
Dataset: wc-at.csv

Adaptive Tissue does not follow a Normal Distribution

Waist Circumference does not follow a Normal Distribution

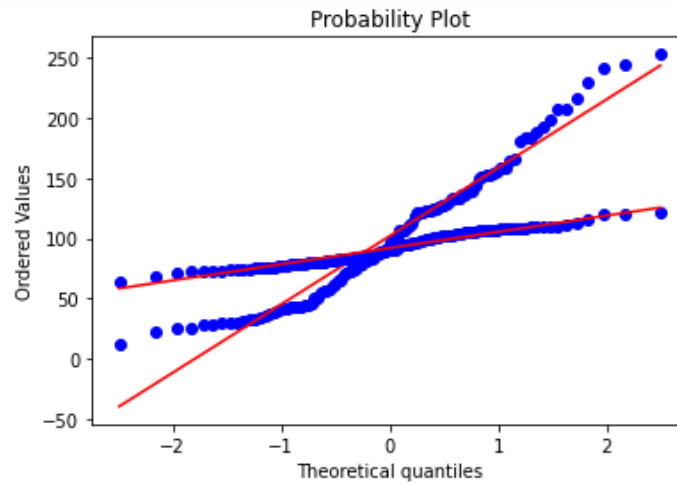


```
In [10]: #Using Shapiro-wilk test
shapiro=stats.shapiro(MPG)
print('Probability = ',shapiro[1])
```

Probability = 0.17639249563217163

```
In [11]: if shapiro[1] > 0.05:
          print('MPG of cars follows Normal Distribution')
        else:
          print('MPG of cars does not follows Normal Distribution')
```

MPG of cars follows Normal Distribution



```
In [13]: #Using Shapiro-wilk test
shapiroAT=stats.shapiro(AT)
shapiroWaist=stats.shapiro(Waist)
print('Probability of Adaptive Tissue = ',shapiroAT[1])
print('Probability Waist Circumference = ',shapiroWaist[1])
```

```
Probability of Adaptive Tissue = 0.0006539996829815209
Probability Waist Circumference = 0.001170447445474565
```

```
In [15]: if shapiroAT[1] > 0.05:
          print('Adaptive Tissue follows Normal Distribution')
        else:
          print('Adaptive Tissue does not follows Normal Distribution')
        if shapiroWaist[1] > 0.05:
          print('Waist Circumference follows Normal Distribution')
        else:
          print('Waist Circumference does not follows Normal Distribution')
```

```
Adaptive Tissue does not follows Normal Distribution
Waist Circumference does not follows Normal Distribution
```

Q 22) Calculate the Z scores of 90% confidence interval, 94% confidence interval, 60% confidence interval

```
In [1]: import scipy  
        from scipy import stats
```

```
In [2]: #z-score of 90% CI  
        print('z-score of 90% CI', stats.norm.ppf(.95))  
  
z-score of 90% CI 1.6448536269514722
```

```
In [4]: #z-score of 94% CI  
        print('z-score of 94% CI', stats.norm.ppf(.97))  
  
z-score of 94% CI 1.8807936081512509
```

```
In [5]: #z-score of 60% CI  
        print('z-score of 60% CI', stats.norm.ppf(.80))  
  
z-score of 60% CI 0.8416212335729143
```

Z score of 90% confidence interval = 1.6448

Z score of 94% confidence interval = 1.8807

Z score of 60% confidence interval = 0.8416

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

```
In [20]: #t-score of 95% CI  
         print('t-score of 95% CI = ', stats.t.ppf(1-0.025, 24))  
  
t-score of 95% CI = 2.0638985616280205
```

```
In [21]: #t-score of 96% CI  
         print('t-score of 96% CI = ', stats.t.ppf(1-0.02, 24))  
  
t-score of 96% CI = 2.1715446760080677
```

```
In [22]: #t-score of 99% CI  
         print('t-score of 99% CI = ', stats.t.ppf(1-0.005, 24))  
  
t-score of 99% CI = 2.796939504772804
```

t score of 95% confidence interval =2.0638

t score of 96% confidence interval =2.1715

t score of 99% confidence interval =2.7969

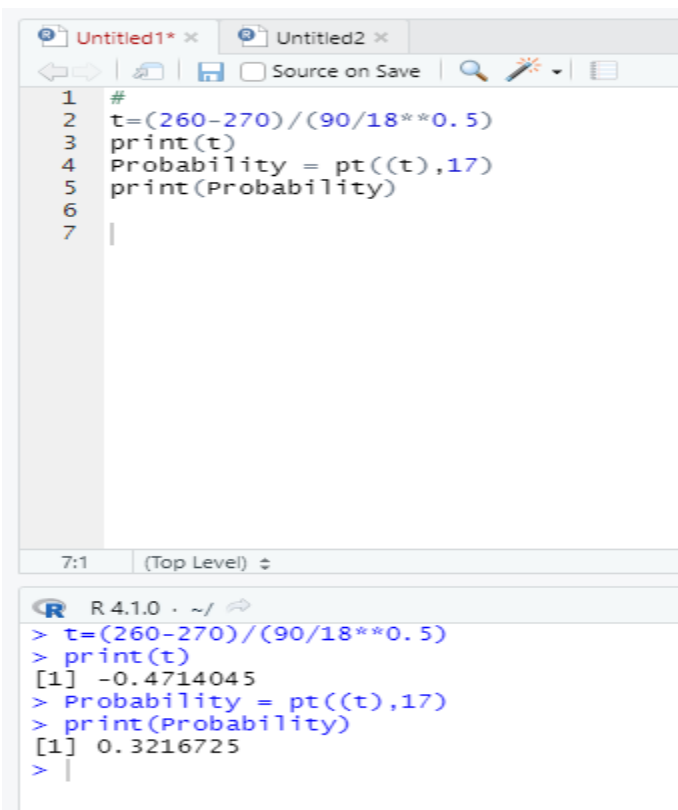
Q 24) A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode → `pt(tscore,df)`

df → degrees of freedom

Probability that the bulb would have an average life of no more than 260 days is 32.16%



```
1 #
2 t=(260-270)/(90/18**0.5)
3 print(t)
4 Probability = pt((t),17)
5 print(Probability)
6
7 |
```

```
R 4.1.0 . ~/
> t=(260-270)/(90/18**0.5)
> print(t)
[1] -0.4714045
> Probability = pt((t),17)
> print(Probability)
[1] 0.3216725
> |
```