# Basic Statistics assignment-1

Balaji N
nbalaji743@gmail.com

Q1) Identify the Data type for the Following:

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

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

Nominal, Ordinal, Interval, Ratio.

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

# Basic Statistics assignment-1

Balaji N
nbalaji743@gmail.com

| | |
|---|---|
| Barometer Pressure | Interval |
| SAT Scores | Ordinal |
| Years of Education | Interval |

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

Total no of possibilities= $2^3 = 8$

Total no sample { (HHH), (HHT), (HTH), (THH),

(HTT), (TTH), (THT), (TTT)}

Probability of Two heads and one tail={(HTT), (TTH), (THT)}

=3\8

The probability that two heads and one tail is **0.375 (or) 37.5%**

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

The sample space = $6^2 = 36$

a) Equal to 1
The probability that the sum is equal to 1 is **0**
Because there is **no combination** that the dice provides sum 1

b) Less than or equal to 4
The total possibilities of sum being less than or equal to 4 =6
The probability is = 6/36 =**1/6**
The probability that the sum is less than or equal to 4 is **0.166 (or) 16.6%**

c) Sum is divisible by 2 and 3
The total possible outcomes divisible by 2 and 3 = 24
The probability is possible outcome/total outcome= **6/36=1/6**
The probability that the Sum is divisible by 2 and 3= **0.166 (or) 16.6%**

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?

Solution:

The total no of balls=(2+3+2)=7

The probability of the first ball is not blue = 5/7("blue ball \ total no of balls")

The probability of the second ball is not blue = 4/6("blue ball \ total no of balls")

probability that none of the balls drawn is blue **P = (5/7)*(4/6)=20/42**

**= 10/21**

**=0.476**

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

Solution:

Expected value = ∑probability*candies

= (1*0.015)+ (4*0.20)+ (3*0.65)+ (5*0.005)+ (6*0.01)+(2*0.120)

= **3.09**

The Expected number of candies for a randomly selected child is **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.

# Basic Statistics assignment-1

Balaji N
nbalaji743@gmail.com

Solution:

| RESULTS | Points | Score | Weigh |
|---|---|---|---|
| Mean | 3.596563 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| 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 | 2.17 | 3.911 | 8.4 |

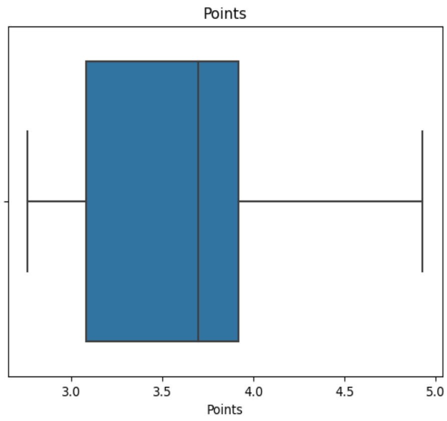
## Inferences:

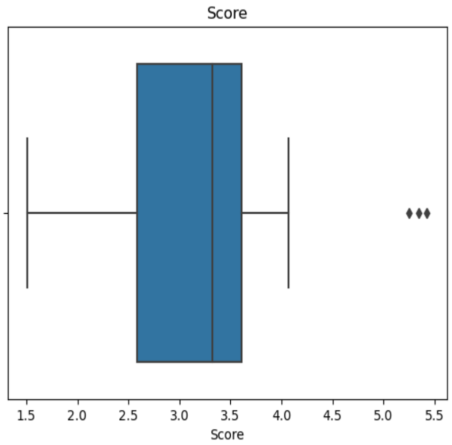
1.) The columns points and Weigh is **multimodal**

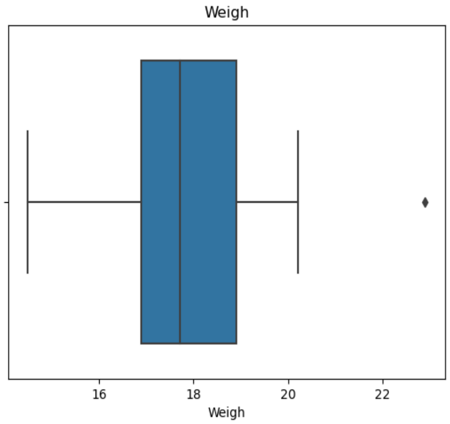
2.) For the Points and score Median> mean , hence the distribution is **Right skewed distribution**

3.) For Weigh  Mean> Median, hence the distribution is **Left skewed distribution**

4.) The boxplot is plotted below by which we can view **Inter-Quartile Region** and **outliers**







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 = ∑probability*weights

= (1/9)*[108+110+123+134+135+145+167+187+199)

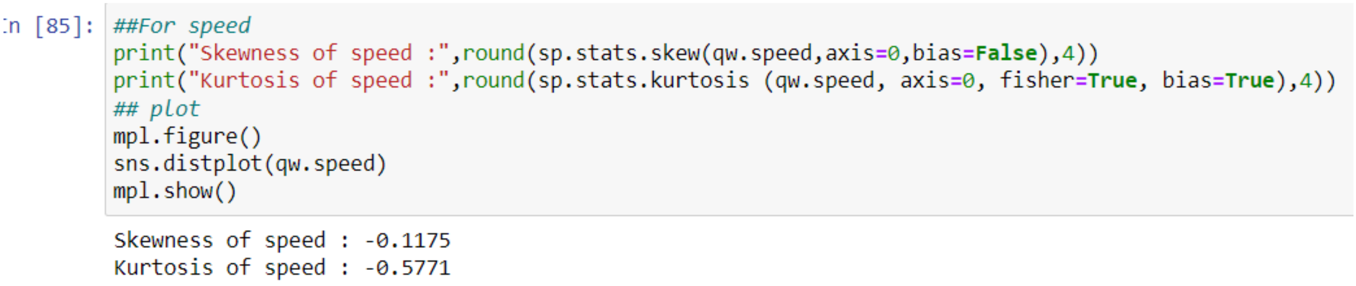
=1308/9

**=145.33**

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

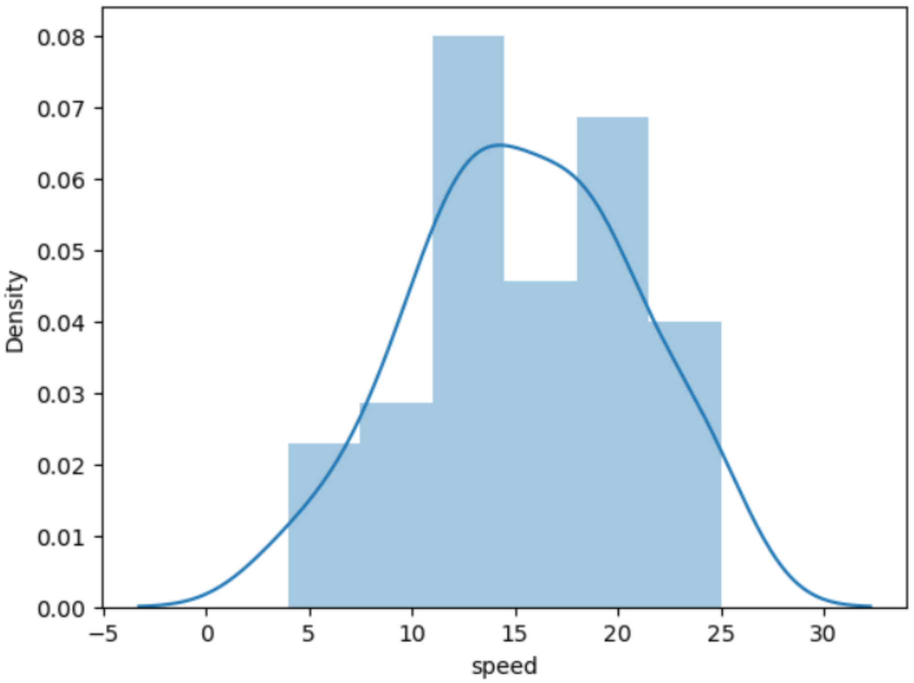
### Cars speed and distance

### Use Q9_a.csv

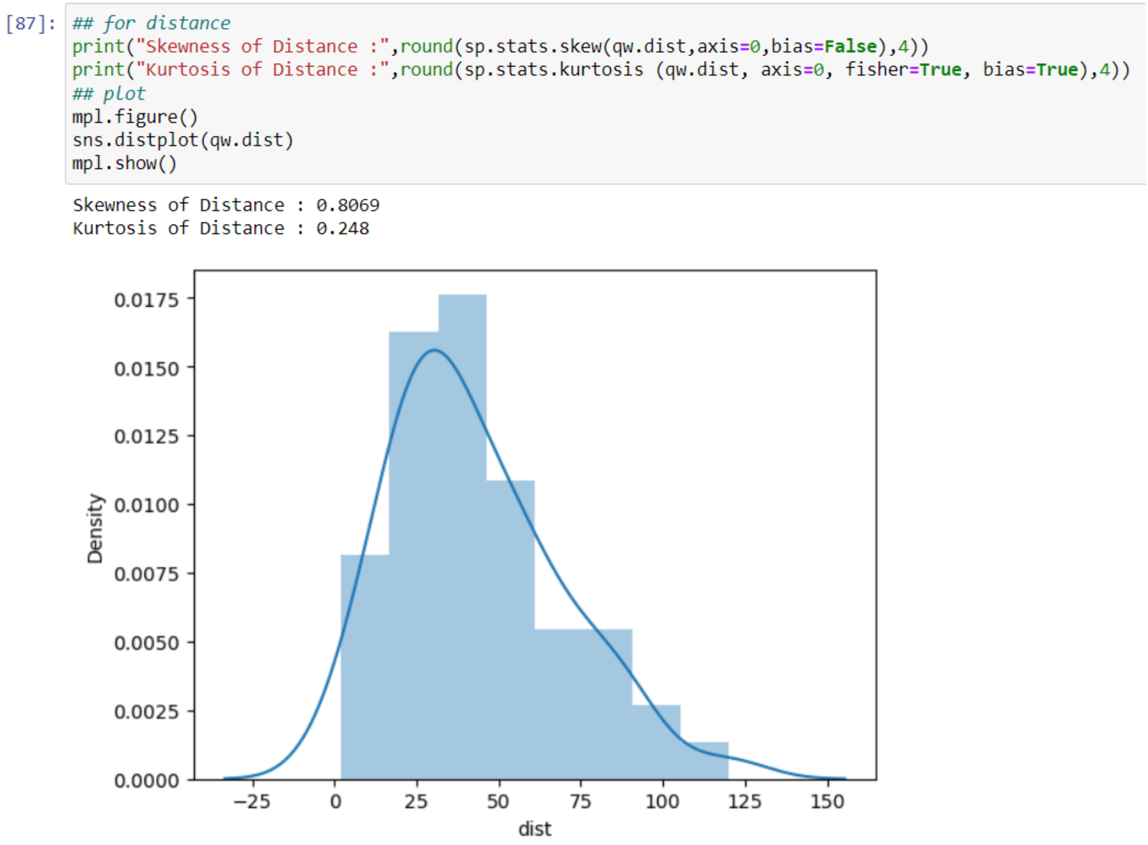


```
n [85]: ##For speed
        print("Skewness of speed :",round(sp.stats.skew(qw.speed,axis=0,bias=False),4))
        print("Kurtosis of speed :",round(sp.stats.kurtosis (qw.speed, axis=0, fisher=True, bias=True),4))
        ## plot
        mpl.figure()
        sns.distplot(qw.speed)
        mpl.show()
```

```
Skewness of speed : -0.1175
Kurtosis of speed : -0.5771
```



- From the above graph we can infer that the speed data is left skewed and it is platykurtic kurtosis (-ve kurtosis) i.e. flatter distribution.



```
[87]: ## for distance
      print("Skewness of Distance :",round(sp.stats.skew(qw.dist,axis=0,bias=False),4))
      print("Kurtosis of Distance :",round(sp.stats.kurtosis (qw.dist, axis=0, fisher=True, bias=True),4))
      ## plot
      mpl.figure()
      sns.distplot(qw.dist)
      mpl.show()

      Skewness of Distance : 0.8069
      Kurtosis of Distance : 0.248
```

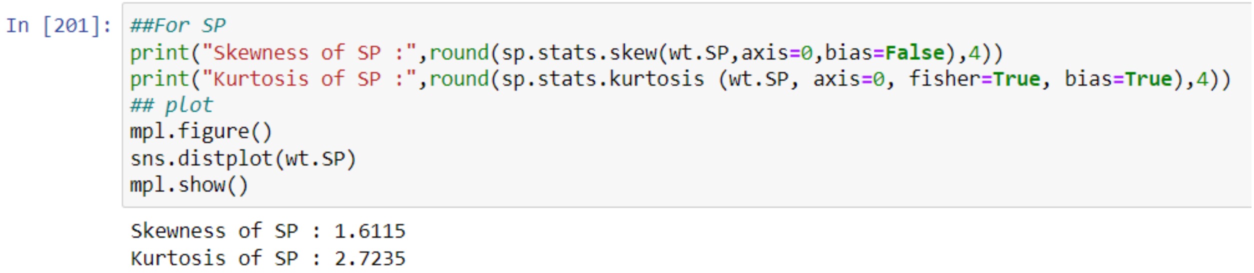
- From the above graph we can infer that the Distance data is right skewed(+ve ) and it is leptokurtic kurtosis (+ve kurtosis) i.e. more peakdness distribution.

# Basic Statistics assignment-1

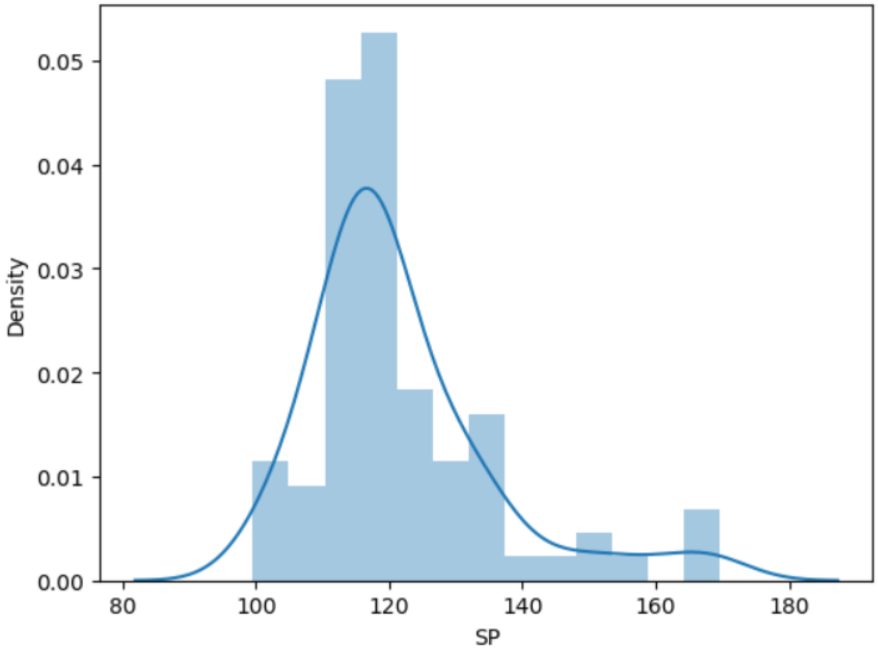
Balaji N
nbalaji743@gmail.com

## SP and Weight(WT)

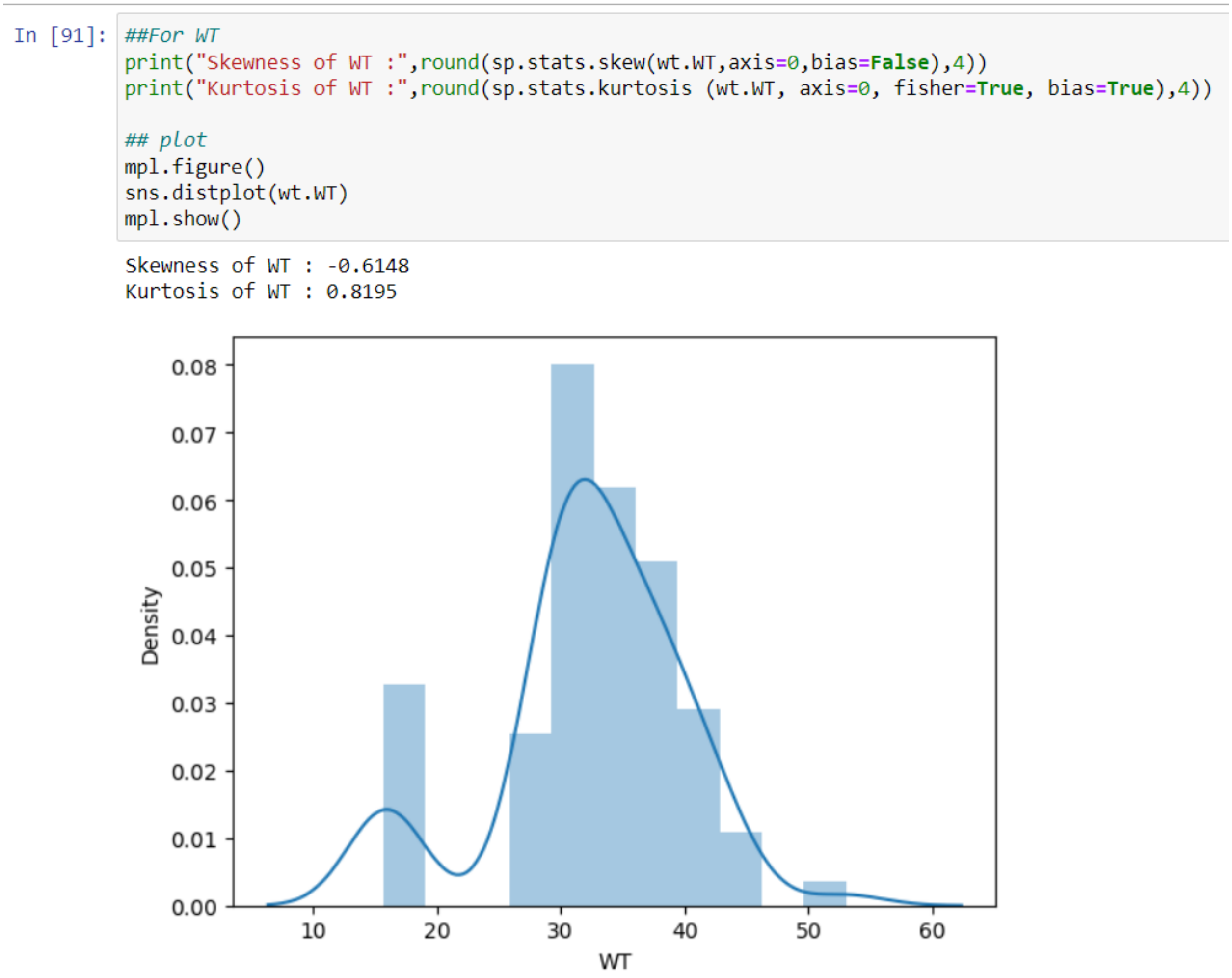
- From the below graph we can infer that the SP data is right skewed(+ve ) and it is leptokurtic kurtosis (+ve kurtosis) i.e. more peakdness distribution.
- There is outlier values in the upper extreme zone.



```
In [201]: ##For SP
          print("Skewness of SP :",round(sp.stats.skew(wt.SP,axis=0,bias=False),4))
          print("Kurtosis of SP :",round(sp.stats.kurtosis (wt.SP, axis=0, fisher=True, bias=True),4))
          ## plot
          mpl.figure()
          sns.distplot(wt.SP)
          mpl.show()

          Skewness of SP : 1.6115
          Kurtosis of SP : 2.7235
```



- From the below  graph we can infer that the WT data is left skewed and it is leptokurtic kurtosis (+ve kurtosis) i.e. more peakdness distribution.
- This data have outliers in the upper and lower extreme zones.



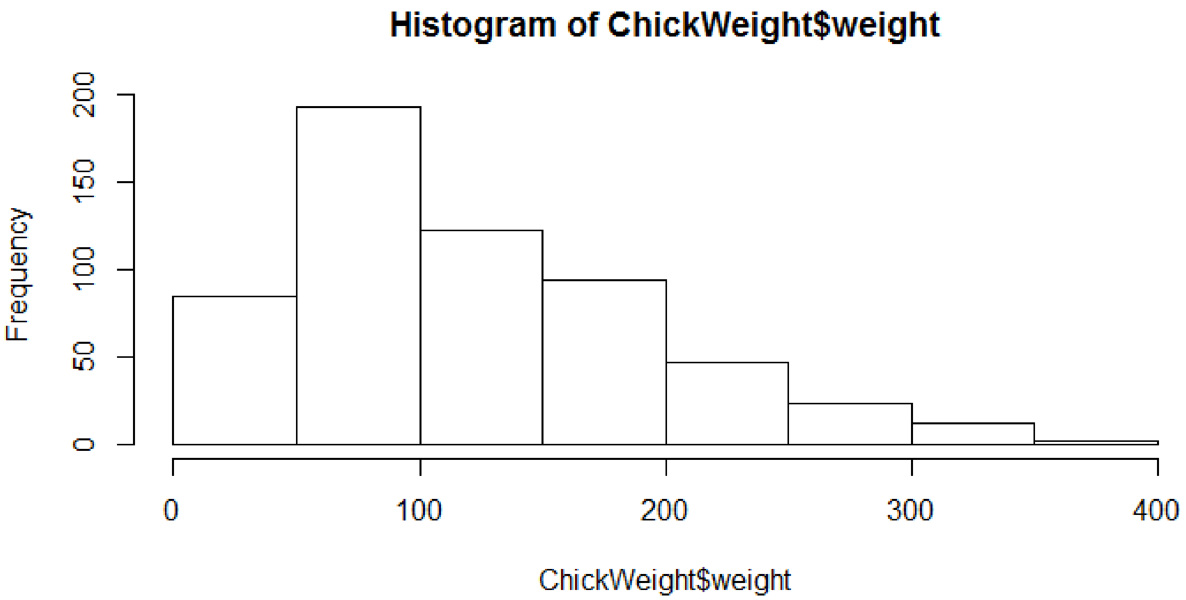
```
In [91]: ##For WT
         print("Skewness of WT :",round(sp.stats.skew(wt.WT,axis=0,bias=False),4))
         print("Kurtosis of WT :",round(sp.stats.kurtosis (wt.WT, axis=0, fisher=True, bias=True),4))

         ## plot
         mpl.figure()
         sns.distplot(wt.WT)
         mpl.show()

         Skewness of WT : -0.6148
         Kurtosis of WT : 0.8195
```

**Use Q9_b.csv**

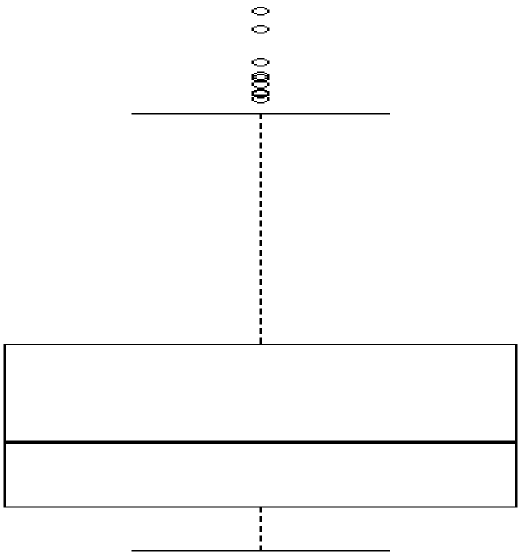
# Basic Statistics assignment-1

Balaji N
nbalaji743@gmail.com

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



- This histogram represents  Right skewed distribution and in this case **Median>mean**



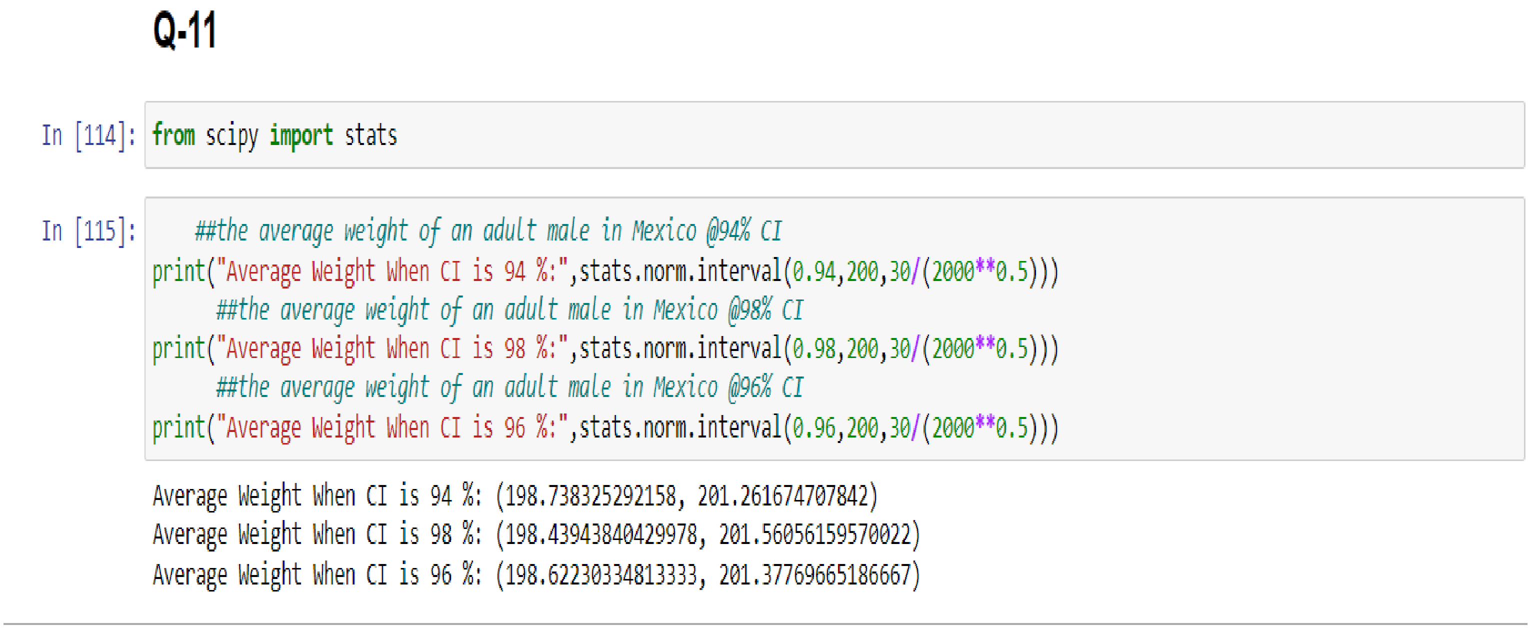
- From this box plot we could infer that more no of outlier values are present in the upper extreme zone.

- This graph shows that it is **right skewed**
- The range of whisker is very wide in the upper quartile region
- 1.5IQR gives the limit if the upper extreme point and the values of data beyond this upper extreme point are termed as Outliers .
- These outliers are to be omitted in the distribution to obtain a normal distribution.

**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?

**Solution:**

Interval= point estimate ± margin of error
Point estimate =200 pounds
Margin of error = (standard deviation)/(sqrt(samples));
Standard deviation is 200 ; The no of sample drawn is 2000



## Q-11

```
In [114]: from scipy import stats
```

```
In [115]:     ##the average weight of an adult male in Mexico @94% CI
          print("Average Weight When CI is 94 %:",stats.norm.interval(0.94,200,30/(2000**0.5)))
               ##the average weight of an adult male in Mexico @98% CI
          print("Average Weight When CI is 98 %:",stats.norm.interval(0.98,200,30/(2000**0.5)))
               ##the average weight of an adult male in Mexico @96% CI
          print("Average Weight When CI is 96 %:",stats.norm.interval(0.96,200,30/(2000**0.5)))
```

```
Average Weight When CI is 94 %: (198.738325292158, 201.261674707842)
Average Weight When CI is 98 %: (198.43943840429978, 201.56056159570022)
Average Weight When CI is 96 %: (198.62230334813333, 201.37769665186667)
```

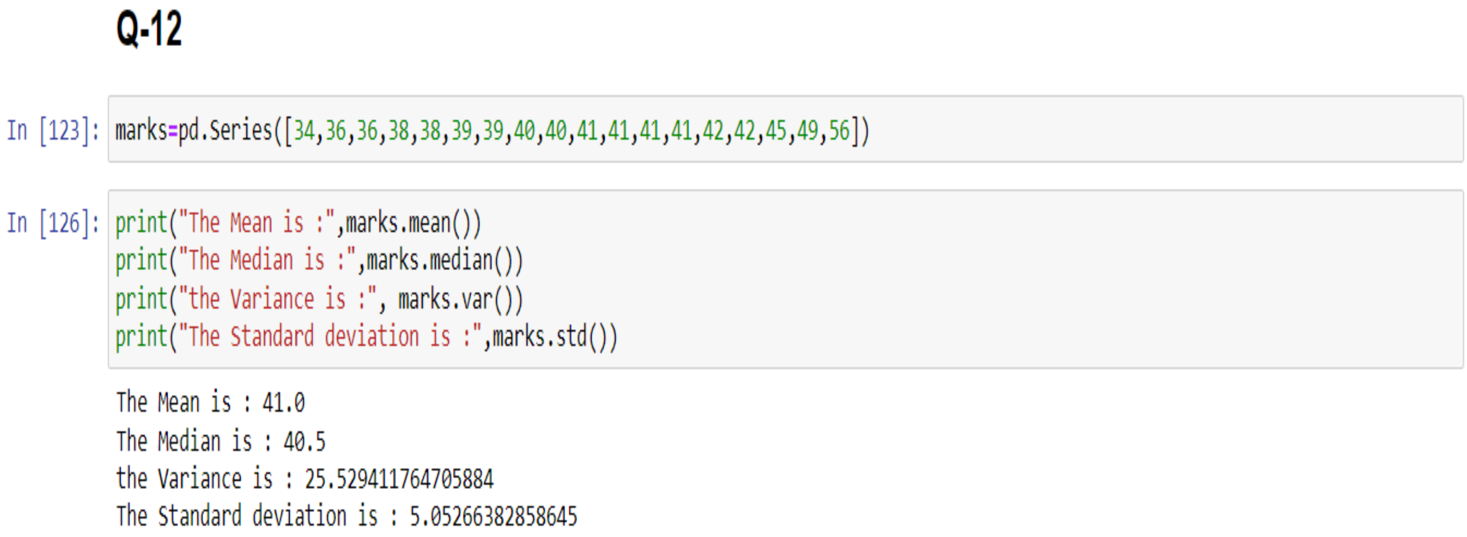
# Basic Statistics assignment-1

Balaji N
nbalaji743@gmail.com

**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.



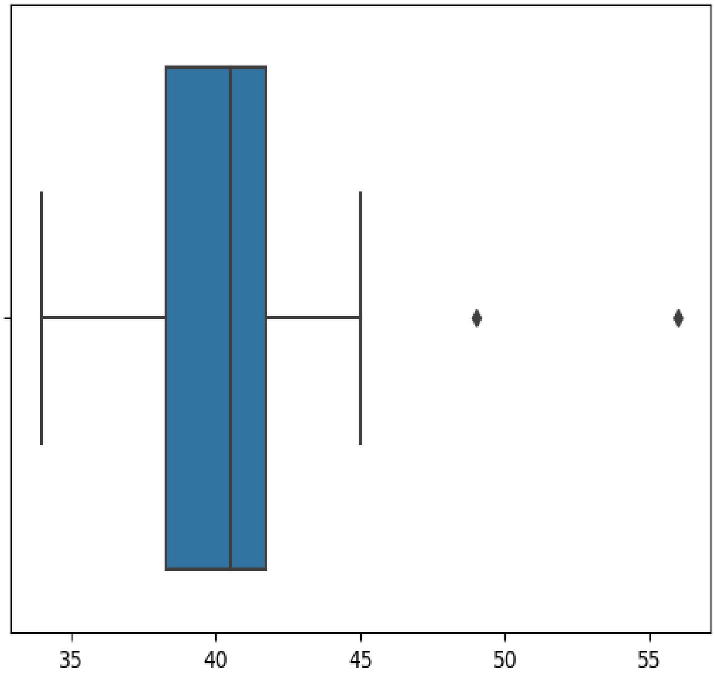
## Q-12

```
In [123]: marks=pd.Series([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])
```

```
In [126]: print("The Mean is :",marks.mean())
          print("The Median is :",marks.median())
          print("the Variance is :", marks.var())
          print("The Standard deviation is :",marks.std())
```

```
The Mean is : 41.0
The Median is : 40.5
the Variance is : 25.529411764705884
The Standard deviation is : 5.05266382858645
```

2) What can we say about the student marks?



- From the students marks it is seen that the mean and median is almost equal and hence the data is **symmetrically distributed**

- This data have two **outlier** values which is beyond the upper **quartile region**(49,56)

# Basic Statistics assignment-1

Balaji N
nbalaji743@gmail.com

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

- In case of Mean=median , the nature of skewness is **symmetrically distributed**

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

- In case of Mean>median , the nature of skewness is **Negative** in nature which is **left skewed distribution**

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

- In case of Mean<median , the nature of skewness is **Positive** in nature which is **Right skewed distribution**

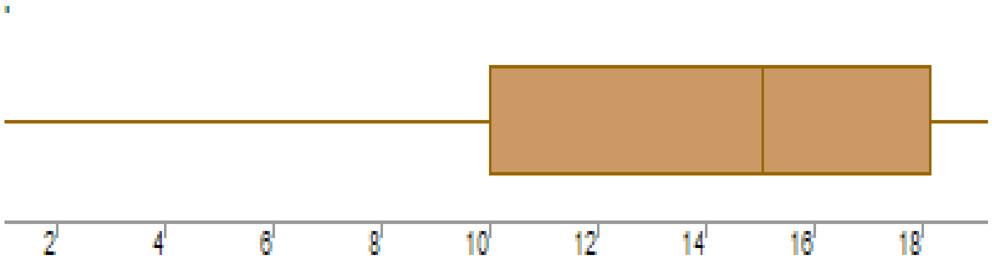
Q16) What does positive kurtosis value indicates for a data ?

- Positive kurtosis for a data Implies that the data is of **more peaked distribution** than the normal distribution

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

- Negative kurtosis for a data Implies that the data is of **flatter distribution** than the normal distribution

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



What can we say about the distribution of the data?

- The given data is not symmetrically distributed ,also more no of values lies below the lower quartile region

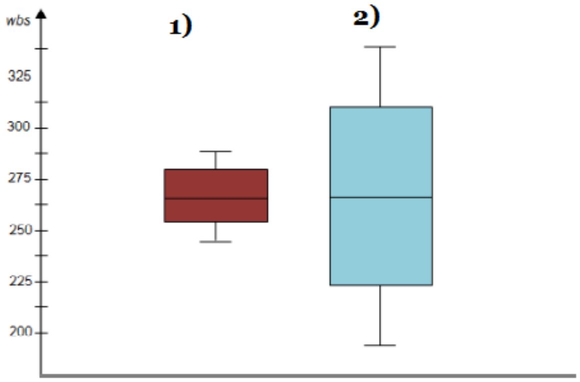
What is nature of skewness of the data?

- This data is left skewed Distribution (-ve skewness)

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

- IQR = upper quartile-lower quartile
  =18-10
  **IQR = 8**

Q19) Comment on the below Boxplot visualizations?



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

**Answer:**

- Box plot 2 has a more wider number of data than box plot 1
- Boxplot 2 if flatter in distribution as compared to box plot 1 which is more peaked.
- Both these data are symmentrically distributed.

# Basic Statistics assignment-1

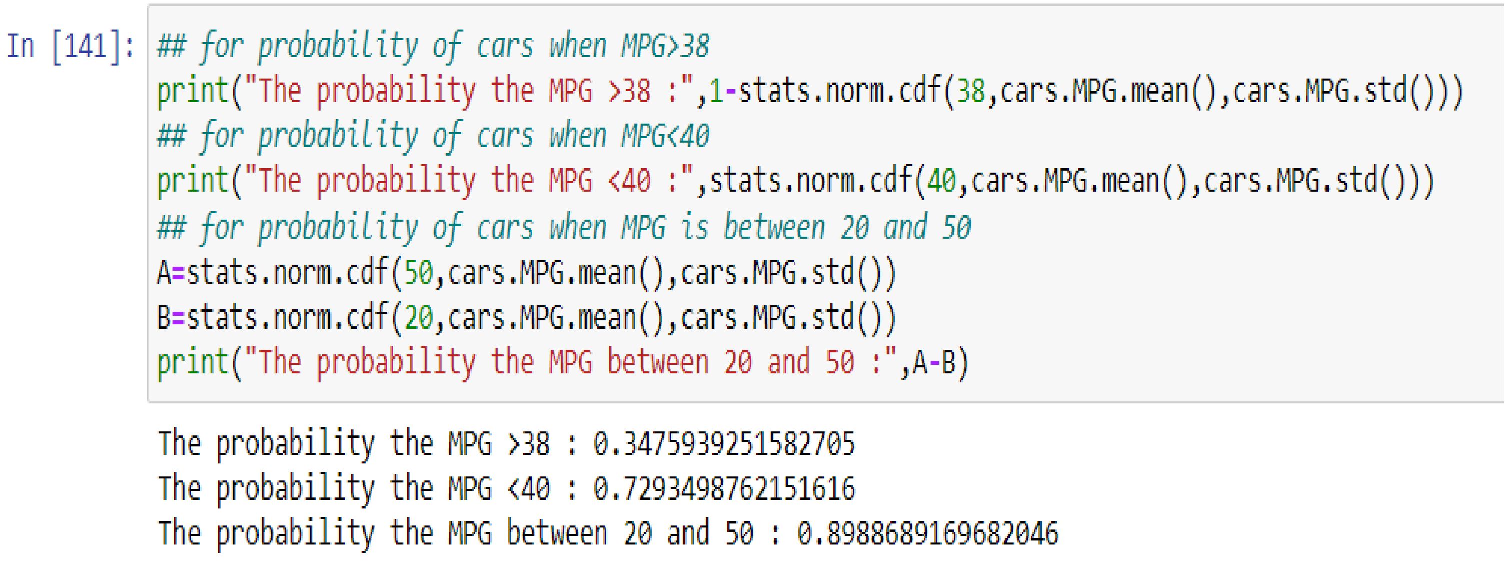
Balaji N
nbalaji743@gmail.com

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(MPG>38)
b. P(MPG<40)
c. P (20<MPG<50)



```
In [141]: ## for probability of cars when MPG>38
          print("The probability the MPG >38 :",1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std()))
          ## for probability of cars when MPG<40
          print("The probability the MPG <40 :",stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std()))
          ## for probability of cars when MPG is between 20 and 50
          A=stats.norm.cdf(50,cars.MPG.mean(),cars.MPG.std())
          B=stats.norm.cdf(20,cars.MPG.mean(),cars.MPG.std())
          print("The probability the MPG between 20 and 50 :",A-B)
```

```
The probability the MPG >38 : 0.3475939251582705
The probability the MPG <40 : 0.7293498762151616
The probability the MPG between 20 and 50 : 0.8988689169682046
```

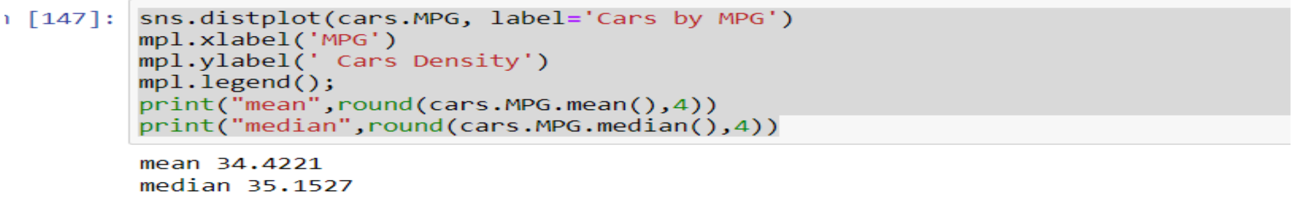
# Basic Statistics assignment-1

Balaji N
nbalaji743@gmail.com

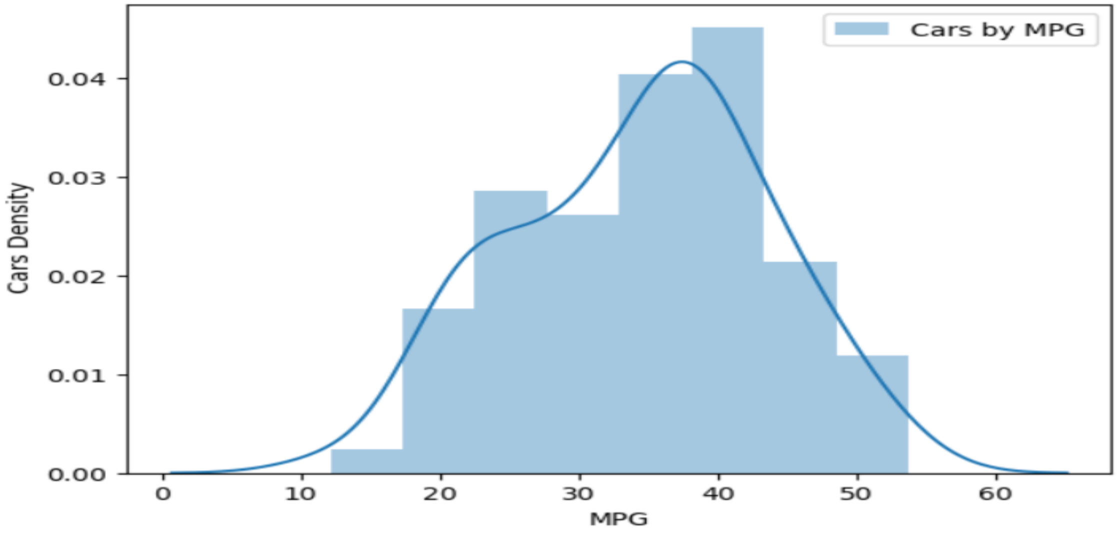
Q 21) Check whether the data follows normal distribution

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

- Median >mean . hence it is **left skewed** distribution



```
ı [147]: sns.distplot(cars.MPG, label='Cars by MPG')
         mpl.xlabel('MPG')
         mpl.ylabel(' Cars Density')
         mpl.legend();
         print("mean",round(cars.MPG.mean(),4))
         print("median",round(cars.MPG.median(),4))

         mean 34.4221
         median 35.1527
```

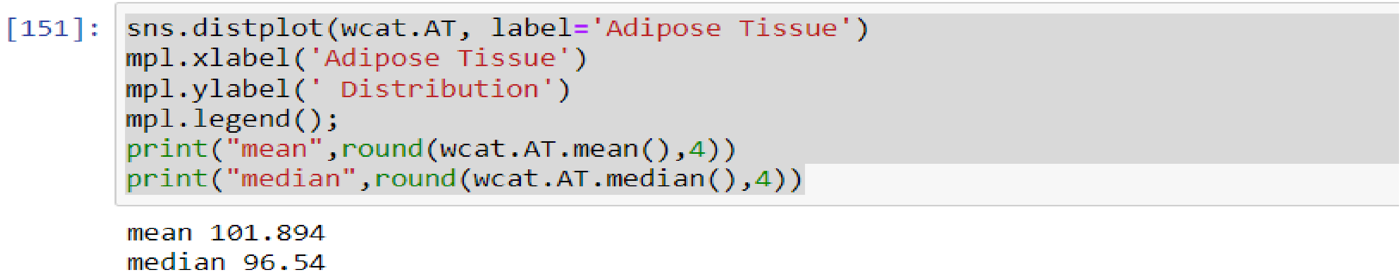


b)Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution
Dataset: wc-at.csv

- Mean > Median , The distribution is **Right Skewed** distribution.

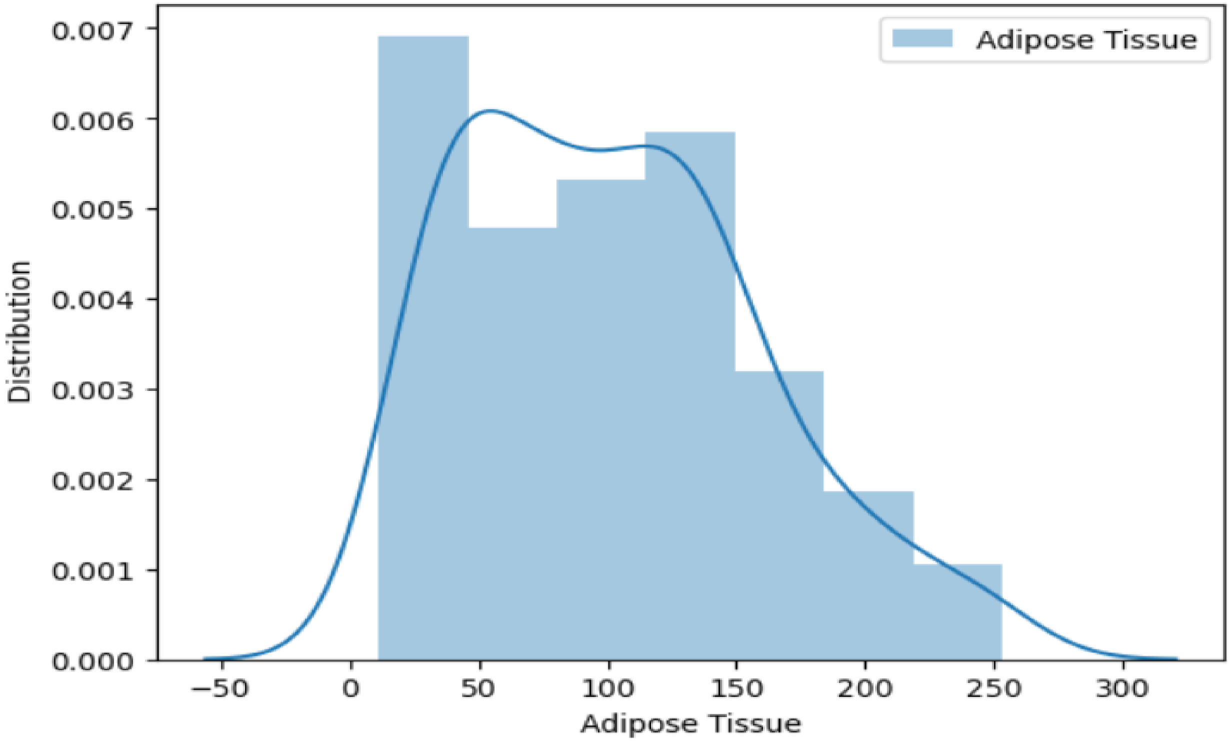
# Basic Statistics assignment-1

Balaji N
nbalaji743@gmail.com



```
[151]: sns.distplot(wcat.AT, label='Adipose Tissue')
       mpl.xlabel('Adipose Tissue')
       mpl.ylabel(' Distribution')
       mpl.legend();
       print("mean",round(wcat.AT.mean(),4))
       print("median",round(wcat.AT.median(),4))
```

```
mean 101.894
median 96.54
```

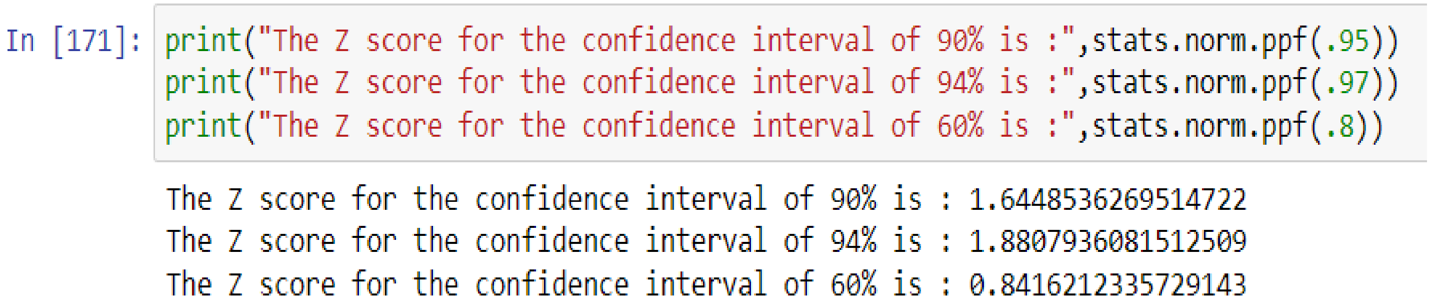


Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

**Solution:**

- For 90% confidence interval → $A=\frac{1+0.90}{2} = 0.95$
- For 94% confidence interval → $A=\frac{1+0.94}{2} = 0.97$
- For 60% confidence interval → $A=\frac{1+0.60}{2} = 0.80$

## Q-22



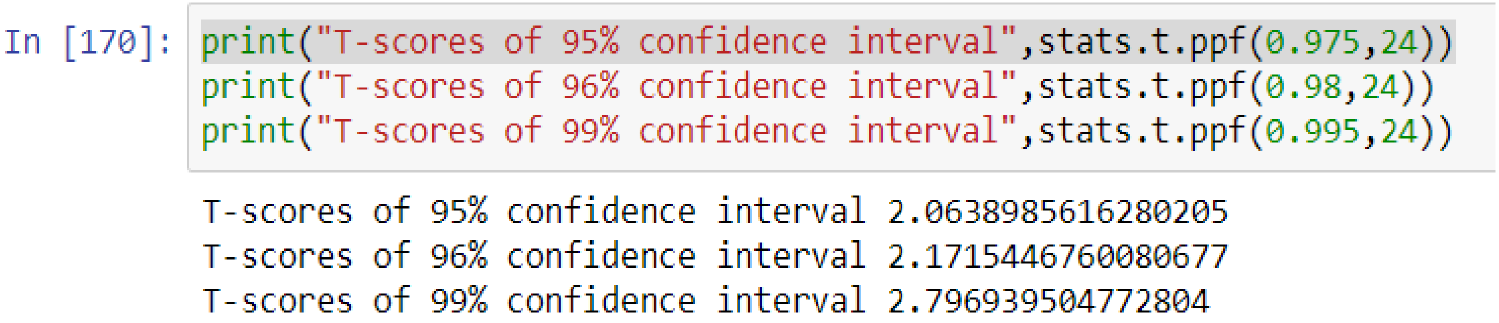
```
In [171]: print("The Z score for the confidence interval of 90% is :",stats.norm.ppf(.95))
          print("The Z score for the confidence interval of 94% is :",stats.norm.ppf(.97))
          print("The Z score for the confidence interval of 60% is :",stats.norm.ppf(.8))

          The Z score for the confidence interval of 90% is : 1.6448536269514722
          The Z score for the confidence interval of 94% is : 1.8807936081512509
          The Z score for the confidence interval of 60% is : 0.8416212335729143
```

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

**Solution:**

- For 95% confidence interval → A=$\frac{1+0.95}{2}$ = 0.975
- For 96% confidence interval → A=$\frac{1+0.96}{2}$ = 0.98
- For 99% confidence interval → A=$\frac{1+0.99}{2}$ = 0.995

## Q-23



```
In [170]: print("T-scores of 95% confidence interval",stats.t.ppf(0.975,24))
          print("T-scores of 96% confidence interval",stats.t.ppf(0.98,24))
          print("T-scores of 99% confidence interval",stats.t.ppf(0.995,24))

          T-scores of 95% confidence interval 2.0638985616280205
          T-scores of 96% confidence interval 2.1715446760080677
          T-scores of 99% confidence interval 2.796939504772804
```

# Basic Statistics assignment-1

Balaji N
nbalaji743@gmail.com

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

Solution:

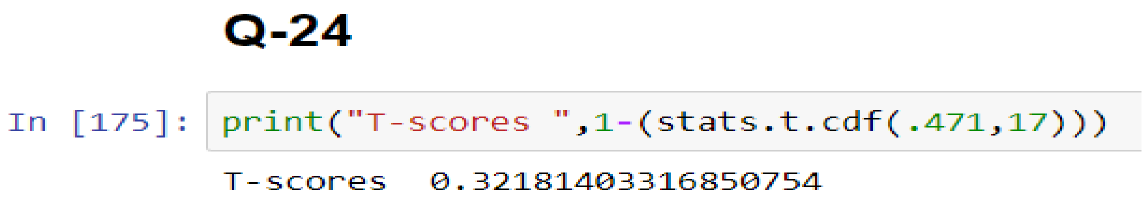
No of items in the sample(n)=18 ;df=n-1=17

Mean of the sampled bulbs(x)=260

Standard deviation(σ)=90

Mean of population(μ)=270

$t = \frac{x-\mu}{s/\sqrt{n}} = \frac{260-270}{90/\sqrt{18}} = -0.471$



## Q-24

```
In [175]: print("T-scores ",1-(stats.t.cdf(.471,17)))
          T-scores  0.32181403316850754
```

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is **0.321**