**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 |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| 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 |
| 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?**

Sol:

No. of possibilities when a coin is tossed = 2

when a coin is tossed 3 times = 2^3=8

probabilities obtained when a coin is tossed 3 times = { (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 = 0.375 (or) 37.5%

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

Sol:

Probability of two dice rolled are 6*6 = 36

a) There sum is equal to 1 = 0

b) There sum is less than or equal to 4 = 6
   The probability = 6/36 = 1/6 = 0.166 (or) 1.6 %

c) Sum is divisible by 2 and 3
   The no. of outcomes divisible by 2 and 3 = 24
   The probability of outcome/total outcome = 6/36 = 1/6
   The probability of 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?**

Sol:

Total no of balls = (2+3+2) =7

Probability of the first ball is not blue = 5/7

Probability of the second ball is not blue = 4/6

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**

Sol:

Expected value = ∑ probability*candies count

= (1*0.015) + (4*0.20) + (3*0.65) + (5*0.005) + (6*0.01) + (2*0.120) = 3.09

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.**

Use Q7.csv file

Sol:

| 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 |

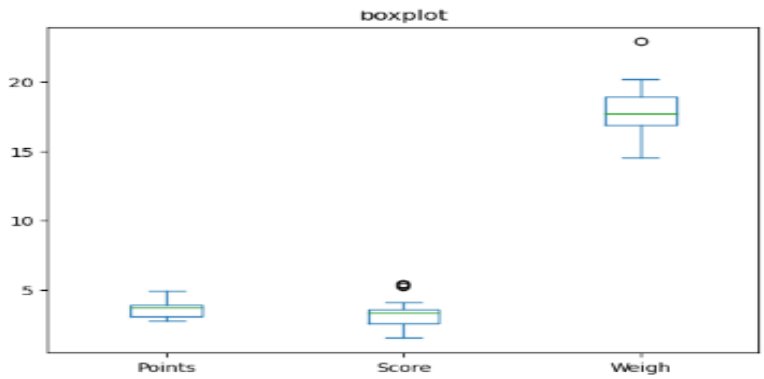
Inferences:

1. Distribution is right skewed in Points and Score, because Median>Mean

2. Distribution is left skewed in Weigh, because Mean>Median

3. The columns Points and Weigh is multimodal

4. 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?**

Sol:

Expected value = ∑probability*weights

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

=1308/9

=145.33

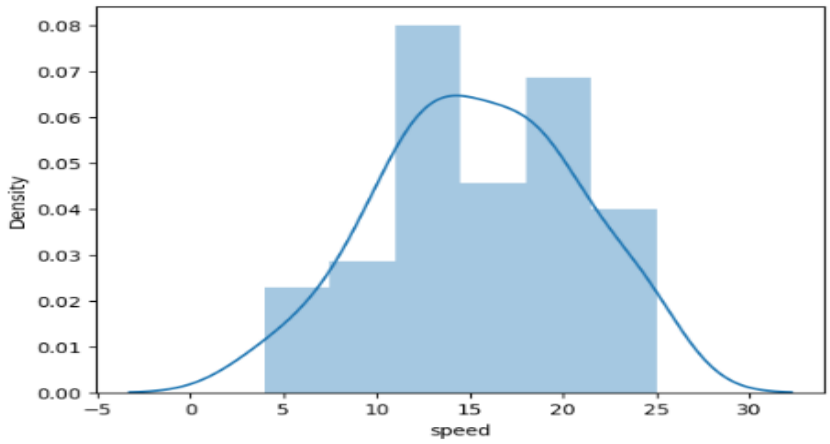
**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9_a.csv**

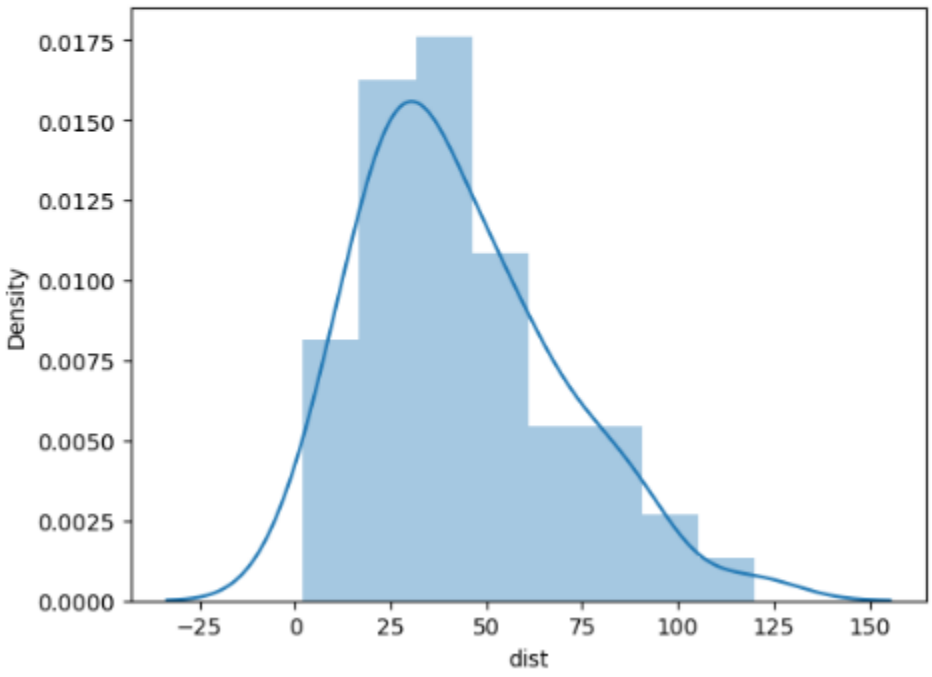
Sol:

```
In [19]: #for speed
         print("Skewness of speed :",round(scp.stats.skew(qn.speed),4))
         print("kurstosis of speed :",round(scp.stats.kurtosis(qn.speed),4))

         Skewness of speed : -0.114
         kurstosis of speed : -0.5771

In [20]: sns.distplot(qn['speed'])
Out[20]: <AxesSubplot:xlabel='speed', ylabel='Density'>
```



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.

```
In [21]: #for distance
         print("Skewness of speed :",round(scp.stats.skew(qn.dist),4))
         print("kurtosis of speed :",round(scp.stats.kurtosis(qn.dist),4))

         Skewness of speed : 0.7825
         kurtosis of speed : 0.248

In [22]: sns.distplot(qn['dist'])

Out[22]: <AxesSubplot:xlabel='dist', ylabel='Density'>
```




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.

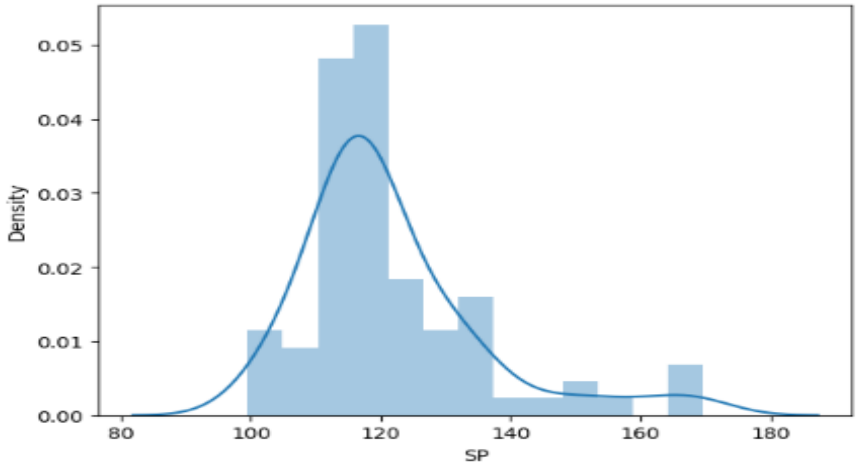
**SP and Weight(WT)**

**Use Q9_b.csv**

## SP

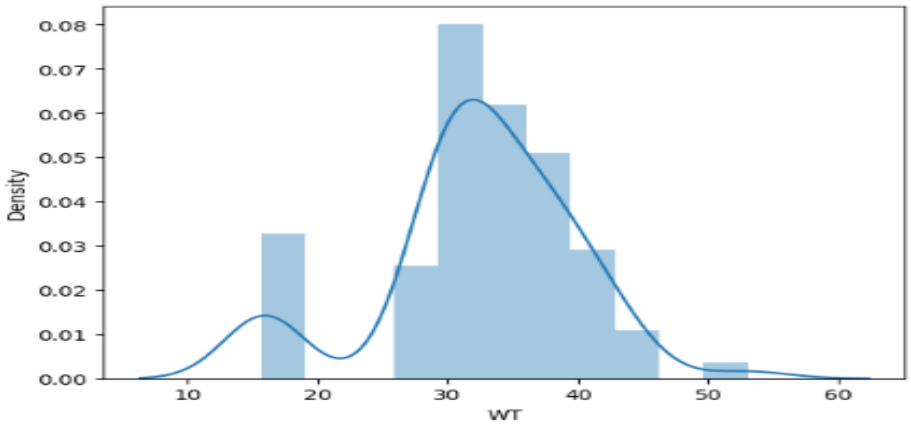
From the below graph we can infer that the SP data is right skewed(+ve ) and it is leptokurtic kurtosis (+ve kurtosis) i.e. more peakdness distribution. There are outlier values in the upper extreme zone.

```
In [25]: #for SP
         print('skewness of SP:',round(scp.stats.skew(qnn.SP),4))
         print('kurstosis of SP:',round(scp.stats.kurtosis(qnn.SP),4))

         skewness of SP: 1.5815
         kurstosis of SP: 2.7235

In [26]: sns.distplot(qnn['SP'])
Out[26]: <AxesSubplot:xlabel='SP', ylabel='Density'>
```




## Weight (WT)

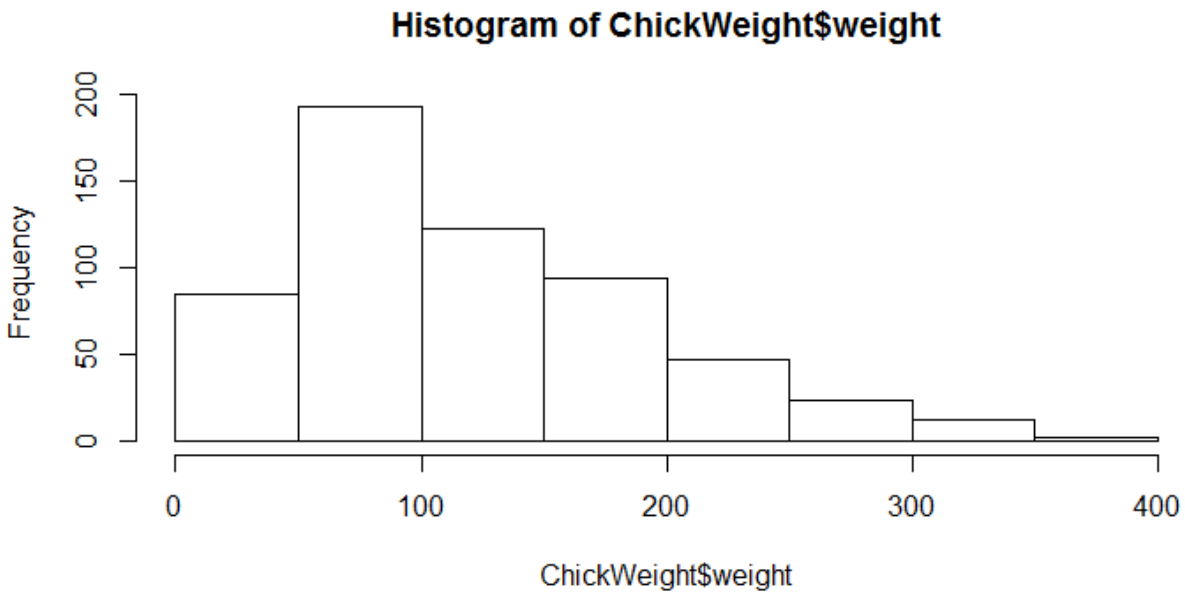
```
In [27]: print('skewness of WT:',round(scp.stats.skew(qnn.WT),4))
         print('kurstosis of WT:',round(scp.stats.kurtosis(qnn.WT),4))

         skewness of WT: -0.6033
         kurstosis of WT: 0.8195

In [28]: sns.distplot(qnn['WT'])
Out[28]: <AxesSubplot:xlabel='WT', ylabel='Density'>
```



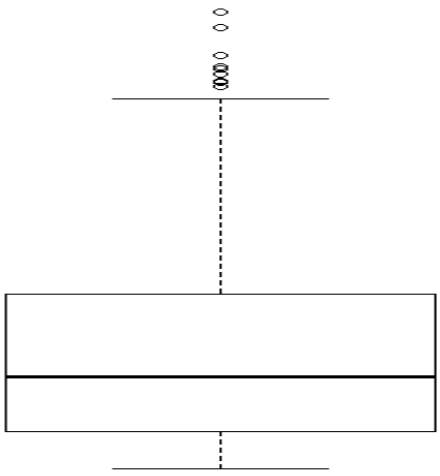
From the above graph we can infer that the WT data is left skewed and it is leptokurtic kurtosis (+ve kurtosis) i.e. more peakdness in distribution.This data have outliers in the upper and lower extreme zones.

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

---




In this histogram distribution is positively skewed hence it is right skewed distribution in this case Median>mean



1. From this box plot we could infer that more no of outlier values are present above the upper extreme zone.
2. This graph shows that it is right skewed
3. The range of whisker is very wide in the upper quartile region
4. 1.5IQR gives the limit if the upper extreme point and the values of data beyond this upper extreme point are termed as Outliers.
5. 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?**

Sol:

```
In [29]: from scipy import stats

In [30]: cdf_98 = stats.norm.interval(0.98,200,30/np.sqrt(2000))
         print('Confidence Interval at 98% is :',np.round(cdf_98,2))

         Confidence Interval at 98% is : [198.44 201.56]

In [31]: cdf_96 = stats.norm.interval(0.96,200,30/np.sqrt(2000))
         print('Confidence Interval at 96% is : ',np.round(cdf_96,2))

         Confidence Interval at 96% is :  [198.62 201.38]

In [32]: cdf_94 = stats.norm.interval(0.94,200,30/np.sqrt(2000))
         print('Confidence Interval at 94% is :',np.round(cdf_94,2))

         Confidence Interval at 94% is : [198.74 201.26]
```

**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**

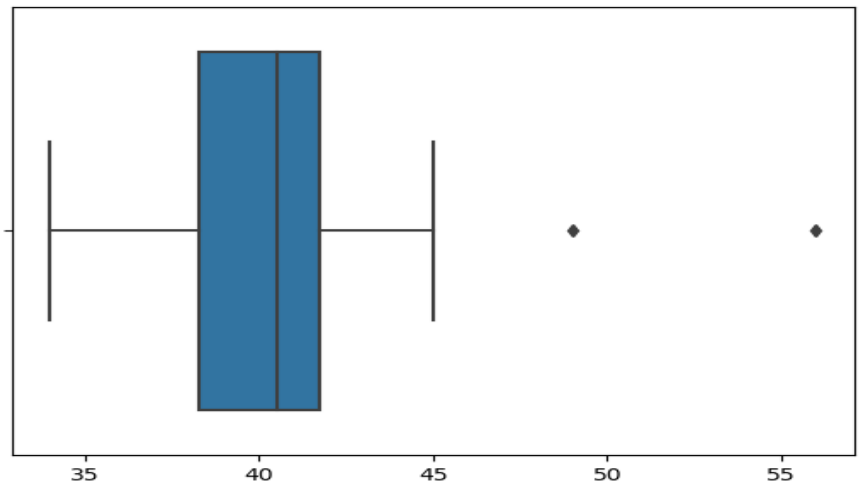
**1) Find mean, median, variance, standard deviation.**

## Q.12

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

In [41]: print("The Mean is :",marks.mean())
         print("The Median is :",marks.median())
         print("the Variance is :",round(marks.var(),4))
         print("The Standard deviation is :",round(marks.std(),4))

         The Mean is : 41.0
         The Median is : 40.5
         the Variance is : 25.5294
         The Standard deviation is : 5.0527
```

**2) What can we say about the student marks?**

```
In [42]: sns.boxplot(marks)
Out[42]: <AxesSubplot:>
```




1. From this boxplot we can infer that the mean and median are close and its symmetrically distributed we can also call it as normal distribution.
2. And also, we can see two outliers (49,56) above the upper extreme region.

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

Sol:

The nature of skewness is normally distributed when mean, median of data is equal

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

Sol:

The nature of skewness is negatively skewed i.e., left skewed when mean>median

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

Sol:

The nature of skewness is positively skewed i.e., right skewed when median>mean

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

Sol:

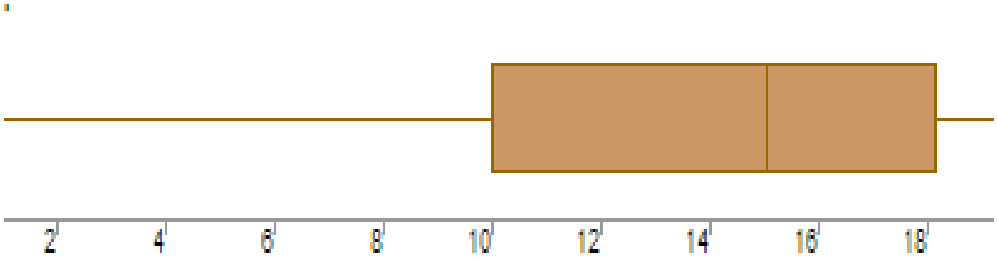
The positive kurtosis value indicates that the peakedness in distribution is very high known as leptokurtic

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

Sol:

The negative kurtosis value indicates that the peakedness in distribution is flattened known as platykurtic

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




**What can we say about the distribution of the data?**

Sol:

The distribution of the data is skewed so mean will be pulled away from the center.

**What is nature of skewness of the data?**

Sol:

Nature of skewness of the data is negatively skewed i.e., left skewed distribution,
In this case median>mean

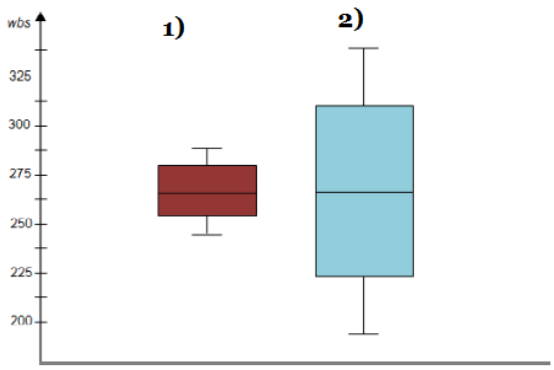
**What will be the IQR of the data (approximately)?**

Sol:

IQR = upper quartile-lower quartile

IQR = 18 - 10 = 8

**Q19) Comment on the below Boxplot visualizations?**




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

Sol:

1. In both the boxplots the distribution is symmetrical
2. Boxplot 2 has a wider number of data than box plot 1
3. Boxplot 2 is flatter in distribution as compared to box plot 1 which is more peaked.

**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)**

Sol:

```
In [56]: print('The probability the MPG >38 :',round(1-stats.norm.cdf(38,df['MPG'].mean(),df['MPG'].std()),4))

         The probability the MPG >38 : 0.3476

In [58]: print('The probability the MPG <40 :',round(stats.norm.cdf(40,df['MPG'].mean(),df['MPG'].std()),4))

         The probability the MPG <40 : 0.7293

In [62]: print('The probability the MPG between 20 and 50 :',round((stats.norm.cdf(50,df['MPG'].mean(),df['MPG'].std())
                                                                  -stats.norm.cdf(20,df['MPG'].mean(),df['MPG'].std())),4))

         The probability the MPG between 20 and 50 : 0.8989
```

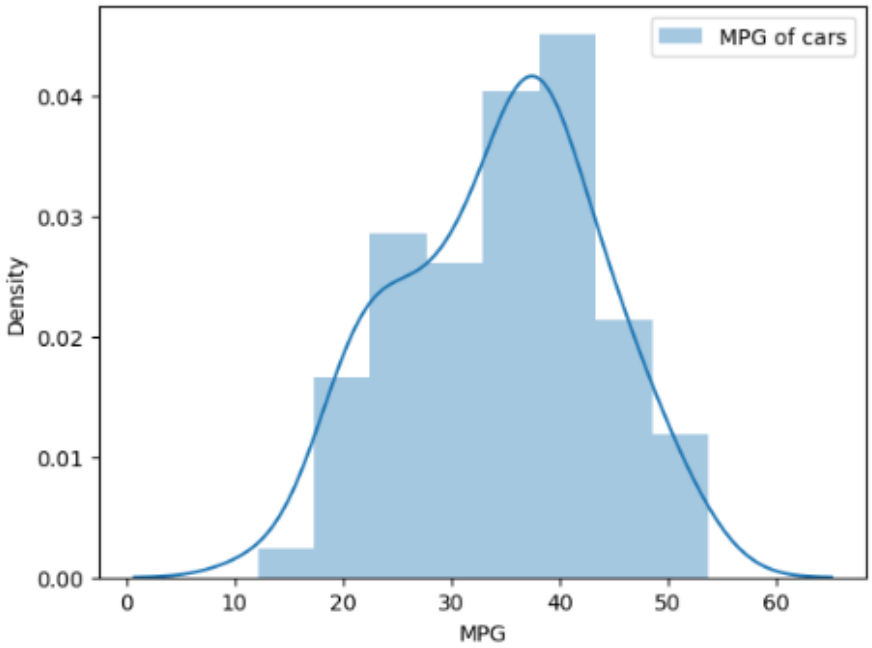
**Q 21) Check whether the data follows normal distribution**

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

Sol:

Median>Mean, It is left skewed distribution

```
In [66]: sns.distplot(df.MPG ,label='MPG of cars')
         plt.xlabel('MPG')
         plt.ylabel('Density')
         plt.legend();
         print("mean",round(df.MPG.mean(),4))
         print("median",round(df.MPG.median(),4))

         mean 34.4221
         median 35.1527
```



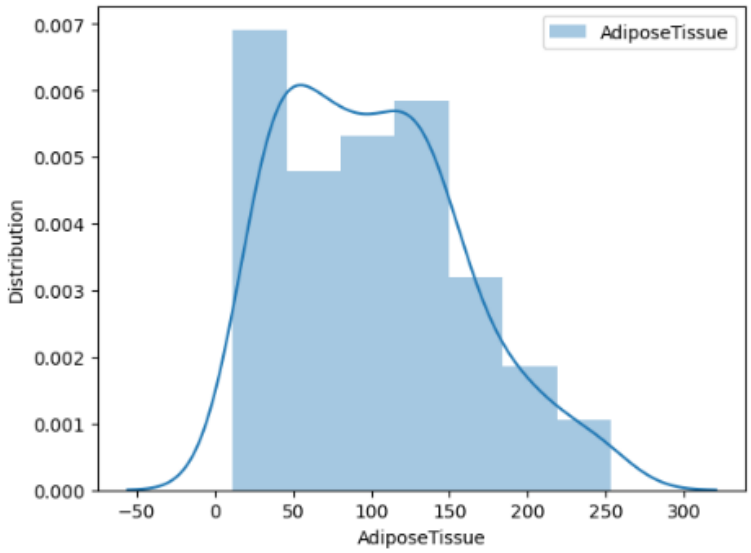

**b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution**
**Dataset: wc-at.csv**

Sol:

Mean > Median, It is Right Skewed distribution.

```
In [48]: sns.distplot(wt.AT, label='AdiposeTissue')
         plt.xlabel('AdiposeTissue')
         plt.ylabel('Distribution')
         plt.legend();
         print('mean',round(wt.AT.mean(),4))
         print('median',round(wt.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**

Sol:

For 90% confidence interval A = (1+0.90)/2 = 0.95

For 94% confidence interval A = (1+0.94)/2 = 0.97

For 60% confidence interval A = (1+0.60)/2 = 0.80

## Q.22

```
In [63]: print('The Z score for the confidence interval of 90% is :',round(stats.norm.ppf(.95),4))
         print('The Z score for the confidence interval of 94% is :',round(stats.norm.ppf(.97),4))
         print('The Z score for the confidence interval of 60% is :',round(stats.norm.ppf(.8),4))
```

```
The Z score for the confidence interval of 90% is : 1.6449
The Z score for the confidence interval of 94% is : 1.8808
The Z score for the confidence interval of 60% is : 0.8416
```

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

Sol:

Sample(n) : 25 ; df = n-1 = 24

For 95% confidence interval B = (1+0.95)/2 = 0.975

For 96% confidence interval B = (1+0.96)/2 = 0.98

For 99% confidence interval B = (1+0.99)/2 = 0.995

**Q.23**

```
In [64]: print("T-scores of 95% confidence interval",round(stats.t.ppf(0.975,24),4))
         print("T-scores of 96% confidence interval",round(stats.t.ppf(0.98,24),4))
         print("T-scores of 99% confidence interval",round(stats.t.ppf(0.995,24),4))

         T-scores of 95% confidence interval 2.0639
         T-scores of 96% confidence interval 2.1715
         T-scores 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**

Sol:

Sample (n)=18; df = n-1 = 17; Mean of sampled bulbs(x)=260 ,

Standard deviation(σ)=90, Mean of population(μ)=270

t=(x-μ)/(s/√n) = (260-270) / (90/√18) = -0.471

```
In [65]: print('T-Scores:',round(1-stats.t.cdf(.471,17),4))
         T-Scores: 0.3218
```

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