Q1) Identify the Data type for the Following:

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

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	Ordinal
Level of Agreement	Ordinal
IQ(Intelligence Scale)	Ratio
Sales Figures	Interval
Blood Group	Nominal
Time Of Day	Ordinal
Time on a Clock with Hands	Ordinal
Number of Children	Interval
Religious Preference	Nominal

Barometer Pressure	Ratio
SAT Scores	Interval
Years of Education	Nominal

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

- ➔ S={HHH,HHT,HTH,THH,TTH,THT,HTT,TTT}
- ➔ P(2 heads and 1 tail) = 3/8

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

S= {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),

(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),

(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),

(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),

(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),

(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)}

a) P(sum is = 1) = 0
b) P(sum < = 4) = 6/36 or 1/6
d) P(Sum is divisible by 2 and 3) = 6/36 or 1/6

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?

- ➔ 10/21

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

- ➔ 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

In [1]:
```
import pandas as pd
import numpy as np
```

In [23]:
```
df = pd.read_csv('Q7.csv')
df.head()
```

Out[23]:

	Unnamed: 0	Points	Score	Weigh
0	Mazda RX4	3.90	2.620	16.46
1	Mazda RX4 Wag	3.90	2.875	17.02
2	Datsun 710	3.85	2.320	18.61
3	Hornet 4 Drive	3.08	3.215	19.44
4	Hornet Sportabout	3.15	3.440	17.02

```
In [20]: print('Points:\n Mean:',df.Points.mean(),'\n','Median:',df.Points.median(),'\n',
               'Standard deviation:',df.Points.std(),'\n','Variance:',df.Points.var(),
               '\n','Range:',df.Points.max()-df.Points.min(),'\n',
               'Mode:','\n',df.Points.mode())
```

```
Points:
 Mean: 3.5965625000000006
 Median: 3.6950000000000003
 Standard deviation: 0.5346787360709716
 Variance: 0.285881350806451 66
 Range: 2.17
 Mode:
 0    3.07
1    3.92
dtype: float64
```

```
In [19]: print('Score:\n Mean:',df.Score.mean(),'\n','Median:',df.Score.median(),'\n',
               'Standard deviation:',df.Score.std(),'\n','Variance:',df.Score.var(),
               '\n','Range:',df.Score.max()-df.Score.min(),'\n','Mode:','\n',df.Score.mode(0))
```

```
Score:
 Mean: 3.2172499999999995
 Median: 3.325
 Standard deviation: 0.9784574429896967
 Variance: 0.9573789677419356
 Range: 3.9110000000000005
 Mode:
 0    3.44
dtype: float64
```

```
In [22]: print('Weight:\n Mean:',df.Weigh.mean(),'\n','Median:',df.Weigh.median(),'\n',
               'Standard deviation:',df.Weigh.std(),'\n','Variance:',df.Weigh.var(),
               '\n','Range:',df.Weigh.max()-df.Weigh.min(),'\n','Mode:','\n',df.Weigh.mode(0))
```

```
Weight:
 Mean: 17.848750000000003
 Median: 17.71
 Standard deviation: 1.7869432360968431
 Variance: 3.193166129032258
 Range: 8.399999999999999
 Mode:
 0    17.02
1    18.90
dtype: float64
```

➔ We can say that there is no high variance in the data and the datapoints are normally distributed.

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?

➔ 145

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

Cars speed and distance

Use Q9_a.csv

```
In [24]: import pandas as pd
         import numpy as np
         import scipy.stats as stats
```

```
In [26]: cars = pd.read_csv('Q9_a.csv')
         cars.head()
```

Out[26]:

	Index	speed	dist
0	1	4	2
1	2	4	10
2	3	7	4
3	4	7	22
4	5	8	16

```
In [30]: print('Cars Speed Skewness:',cars.speed.skew(),'\n','Cars Speed Kurtosis:',cars.speed.kurtosis())

Cars Speed Skewness: -0.11750986144663393
 Cars Speed Kurtosis: -0.5089944204057617
```

The skewness and Kurtosis of Speed column, both are negative

```
In [32]: print('Cars Distance Skewness:',cars.dist.skew(),'\n','Cars Distance Kurtosis:',cars.dist.kurtosis())

Cars Distance Skewness: 0.8068949601674215
 Cars Distance Kurtosis: 0.4050525816795765
```

The skewness and Kurtosis of Distance column, both are Positive

SP and Weight(WT)

Use Q9_b.csv

In [35]:
```
data = pd.read_csv('Q9_b.csv')
data.head()
```

Out[35]:

	Unnamed: 0	SP	WT
0	1	104.185353	28.762059
1	2	105.461264	30.466833
2	3	105.461264	30.193597
3	4	113.461264	30.632114
4	5	104.461264	29.889149

In [37]:
```
print('SP Skewness:',data.SP.skew(),'\n','SP Kurtosis:',data.SP.kurtosis())
```

```
SP Skewness: 1.6114501961773586
 SP Kurtosis: 2.9773289437871835
```

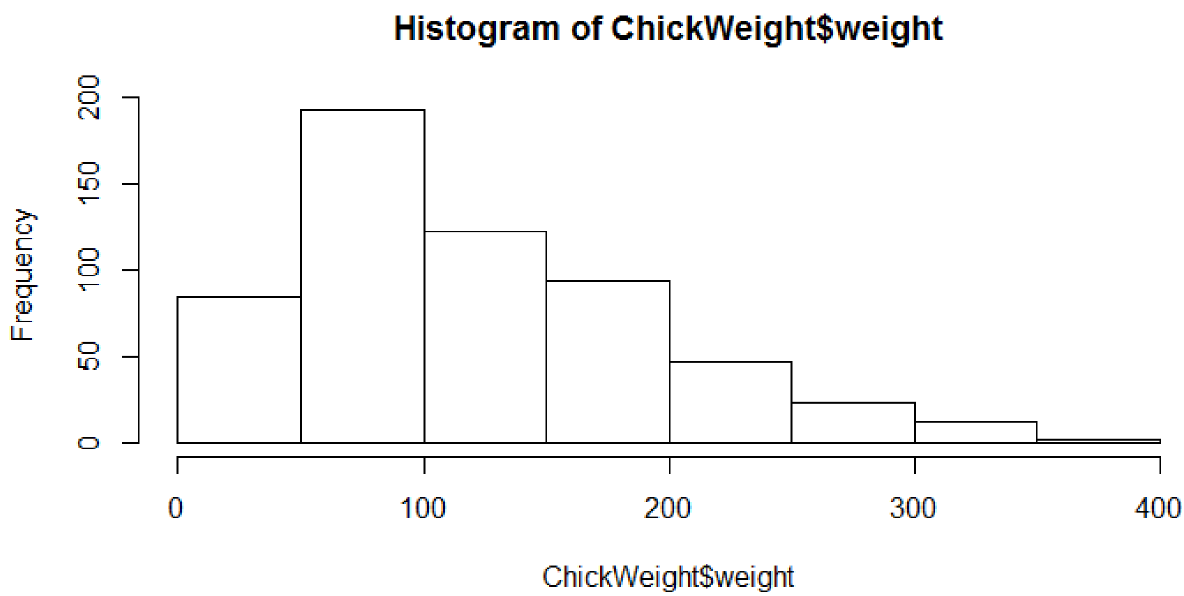
The skewness and Kurtosis of SP column, both are Positive

In [38]:
```
print('WT Skewness:',data.WT.skew(),'\n','WT Kurtosis:',data.WT.kurtosis())
```

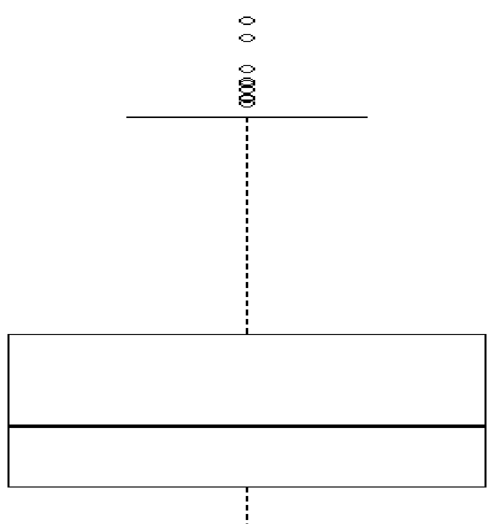
```
WT Skewness: -0.6147533255357768
 WT Kurtosis: 0.9502914910300326
```

There is negative skewness and positive Kurtosis for WT column

Q10) Draw inferences about the following boxplot & histogram



- The above histogram represents the frequency distribution of weights column from chickweight dataset.
- By looking at the histogram, we can say that it represents positive skewness.
- Most of the datapoints lie between 50 – 100 in the x axis.



- The box plot shows there are some points in the data which are outliers.
- This means that they are at greater distance from the mean.

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?

```
In [41]: #Q11
         print('94% confidence interval with mean of 200 and sd of 30 from sample of 2000:')
         stats.t.interval(alpha=0.94, df=2000, loc=200, scale=30)

         94% confidence interval with mean of 200 and sd of 30 from sample of 2000:

Out[41]: (143.54417173267188, 256.4558282673281)
```

```
In [42]: #Q11
         print('96% confidence interval with mean of 200 and sd of 30 from sample of 2000:')
         stats.t.interval(alpha=0.96, df=2000, loc=200, scale=30)

         96% confidence interval with mean of 200 and sd of 30 from sample of 2000:

Out[42]: (138.34732124381935, 261.65267875618065)
```

```
In [43]: #Q11
         print('98% confidence interval with mean of 200 and sd of 30 from sample of 2000:')
         stats.t.interval(alpha=0.98, df=2000, loc=200, scale=30)

         98% confidence interval with mean of 200 and sd of 30 from sample of 2000:

Out[43]: (130.1535847418068, 269.8464152581932)
```

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

```
In [59]: print('Mean:',np.mean(marks),'\n','Median:',np.median(marks),
               '\n','Variance:',np.var(marks),'\n','Standard Deviation:',np.std(marks))

         Mean: 41.0
          Median: 40.5
          Variance: 24.11111111111111
          Standard Deviation: 4.910306620885412
```

There is variance in the marks of students

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

- ➔ The bell curve will be symmetric and there will be 0 skewness

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

- ➔ Positive skewness

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

- ➔ Negative skewness

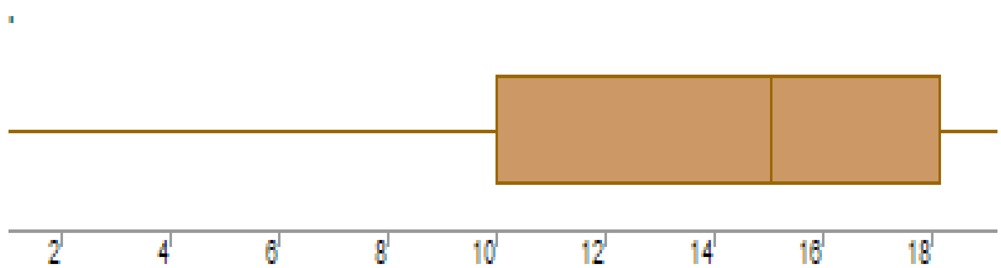
Q16) What does positive kurtosis value indicates for a data ?

- ➔ When we see positive kurtosis curve the peak will be thin and the ends will be thick

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

- ➔ When we see negative kurtosis curve, the curve will be flat and the ends will be thin

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



What can we say about the distribution of the data?

- ➔ We can say that the data is not normally distributed as the boxplot lies towards the right

What is nature of skewness of the data?

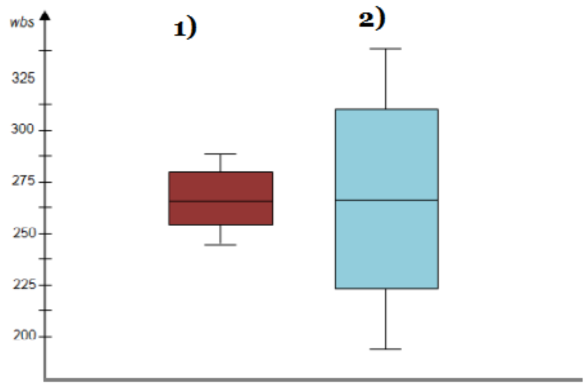
- ➔ The data is negatively skewed

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

- ➔ Q3 = 18 & Q1 = 10
- ➔ Therefor, IQR = Q3 – Q1

➔ = 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.

- ➔ The first boxplot is comparatively short than the second boxplot, this means that there is high variance in the data of boxplot 2 and less variance in the data of boxplot 1.
- ➔ The medians of both these boxplots are all at the same level. However the box plots in these examples show very distributed 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(MPG>38)
b. P(MPG<40)
c. P (20<MPG<50)

In [60]:
```
cars_data = pd.read_csv('Cars.csv')
cars_data
```

Out[60]:

	HP	MPG	VOL	SP	WT
0	49	53.700681	89	104.185353	28.762059
1	55	50.013401	92	105.461264	30.466833
2	55	50.013401	92	105.461264	30.193597
3	70	45.696322	92	113.461264	30.632114
4	53	50.504232	92	104.461264	29.889149
...	...	...	...	...	...
76	322	36.900000	50	169.598513	16.132947
77	238	19.197888	115	150.576579	37.923113
78	263	34.000000	50	151.598513	15.769625
79	295	19.833733	119	167.944460	39.423099
80	236	12.101263	107	139.840817	34.948615

81 rows × 5 columns

In [61]:
```
stats.norm.cdf(38,cars_data.MPG.mean(),cars_data.MPG.std())
```

Out[61]: 0.6524060748417295

In [62]:
```
stats.norm.cdf(60,cars_data.MPG.mean(),cars_data.MPG.std())
```

Out[62]: 0.9974534201888031

In [63]:
```
stats.norm.cdf(50,cars_data.MPG.mean(),cars_data.MPG.std()) - stats.norm.cdf(20,cars_data.MPG.mean(),cars_data.MPG.std())
```

Out[63]: 0.8988689169682046

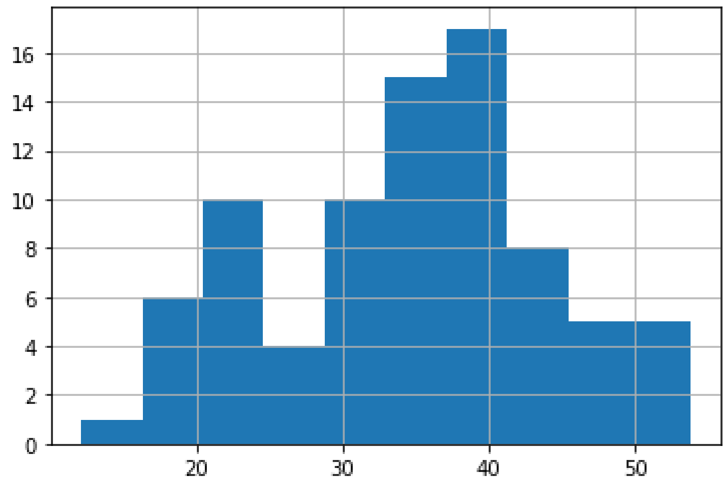
Q 21) Check whether the data follows normal distribution

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

➔ Not normally distributed

```
In [65]: cars_data.MPG.hist()
```

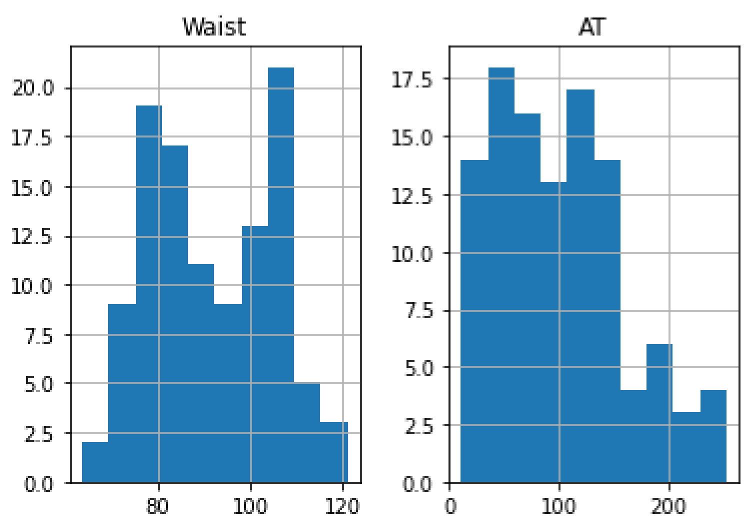
```
Out[65]: <AxesSubplot:>
```



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

```
In [69]: wcat = pd.read_csv('wc-at.csv')
         wcat.hist()
```

```
Out[69]: array([[<AxesSubplot:title={'center':'Waist'}>,
                 <AxesSubplot:title={'center':'AT'}>]], dtype=object)
```



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

```
In [5]: #Q22
        #Calculate the Z scores of  90% confidence interval,94% confidence interval, 60% confidence interval
        print('Z score at 90% CI:', stats.norm.ppf(0.90))
        print('Z score at 94% CI:',stats.norm.ppf(0.94))
        print('Z score at 60% CI:',stats.norm.ppf(0.60))

        Z score at 90% CI: 1.2815515655446004
        Z score at 94% CI: 1.5547735945968535
        Z score at 60% CI: 0.2533471031357997
```

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

```
In [6]: #Q23
        #Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25
        print('t score at 95% CI:',stats.t.ppf(0.95,25))
        print('t score at 96% CI:',stats.t.ppf(0.96,25))
        print('t score at 99% CI:',stats.t.ppf(0.99,25))

        t score at 95% CI: 1.7081407612518986
        t score at 96% CI: 1.8248284689556018
        t score at 99% CI: 2.4851071754106413
```

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

```
In [8]: #Q24
        #x = mean of the sample of bulbs =  260
        #mue = population mean = 270
        #s = standard deviation of the sample = 90
        #n = number of items in the sample = 18
        # t score formula t value = -0.471
        print('Probability:',stats.t.cdf(-0.471,18))

        Probability: 0.3216492583174122
```