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/Categorical

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

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)	Ordinal
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	Ratio
SAT Scores	Interval
Years of Education	Interval

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

Ans:

The probability of two heads and one tail is 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

Ans:

a) The probability of sum is Equal to 1 is 0

b) The probability of sum is Less than or equal to 4 is 6/36 or 1/6

c) The probability of sum is divisible by 2 and 3 is 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?

Ans:

The probability that none of the balls drawn blue is 10/21 or 5/7

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

<u>ANS</u>:

The Expected number of candies for a randomly selected child is **3.09**

<u>Explanation</u>:

= 1*0.015 + 4*0.20 + 3*0.65 + 5*0.005 + 6*0.01 + 2*0.12

= 3.09

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

Use Q7.csv file

<u>ANS</u>:

In [1]:

```
import pandas as pd
```

In [2]:

```
q7 = pd.read_csv('Q7.csv')
```

Mean

In [3]:

```
q7.mean()
```

Out[3]:

```
Points     3.596563
Score      3.217250
Weigh     17.848750
dtype: float64
```

Median

In [4]:

```
q7.median()
```

Out[4]:

```
Points     3.695
Score      3.325
Weigh     17.710
dtype: float64
```

Mode

In [5]:

```
q7[['Points','Score','Weigh']].mode()
```

Out[5]:

	Points	Score	Weigh
0	3.07	3.44	17.02
1	3.92	NaN	18.90

- Mean:

  Points = 3.596563
  Score = 3.217250
  Weigh = 17.848750

- Median:

  Points = 3.695
  Score = 3.325
  Weigh = 17.710

- Mode:

  Points = 3.07
  Score = 3.44
  Weigh = 17.02

Variance

In [6]:

```
q7.var()
```

Out[6]:

```
Points    0.285881
Score     0.957379
Weigh     3.193166
dtype: float64
```

Variance:

Points = 0.28
Score = 0.95
Weigh = 3.19

Standard Deviation

In [7]:

```
q7.std()
```

Out[7]:

```
Points    0.534679
Score     0.978457
Weigh     1.786943
dtype: float64
```

Standard Deviation:

Points = 0.53
Score = 0.97
Weigh = 1.78

Range

In [8]:

```
Points_range = q7['Points'].max() - q7['Points'].min()
Score_range = q7['Score'].max() - q7['Score'].min()
Weigh_range = q7['Weigh'].max() - q7['Weigh'].min()
print(Points_range)
print(Score_range)
print(Weigh_range)
```

```
2.17
3.9110000000000005
8.399999999999999
```

Range:

Points = 2.17
Score = 3.91
Weigh = 8.39

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?

ANS:

The Weight of the patient when one of the patients is chosen at random is 145.33(in pounds).

Explanation:

No. of data given here is 9, then probability of choosing one patient is 1/9, so

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

=145.33

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

Cars speed and distance

In [9]:
```
from scipy.stats import skew,kurtosis
```

In [10]:
```
q9_a = pd.read_csv('Q9_a.csv')
q9_a.head()
```

Out[10]:

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

In [11]:
```
print('The skewness of speed is ',q9_a['speed'].skew())
```
```
The skewness of speed is  -0.11750986144663393
```

In [12]:
```
print('The kurtosis of speed is ',q9_a['speed'].kurtosis())
```
```
The kurtosis of speed is  -0.5089944204057617
```

In [13]:
```
print('The skewness of distance is ',q9_a['dist'].skew())
```
```
The skewness of distance is  0.8068949601674215
```

In [14]:
```
print('The kurtosis of dist is ',q9_a['dist'].kurtosis())
```
```
The kurtosis of dist is  0.4050525816795765
```

Speed:

Skewness = -0.117
Kurtosis = -0.508

Distance:

Skewness = 0.806
Kurtosis = 0.405

Inference:

1. The skewness of speed is negative, it signifies that the data are negatively sk ewed or left skew
2. Mean<median<mode
3. The kurtosis of speed is Negative, it signifies that the curve is flat with thin t ails and data is platykurtic data.
4. The skewness of distance is positive, it signifies that the data are positively s kewed or right skew
5. Mean>median>mode
6. The kurtosis of distance is positive, it signifies that the curve is peaked with t hick tails and data is called as leptokurtic data.

SP and Weight (WT)

Use Q9_b.csv

```
In [15]: q9_b = pd.read_csv('Q9_b.csv')
         q9_b.head()
```

Out[15]:

	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 [16]: print('The skewness of speed is ',q9_b['SP'].skew())

The skewness of speed is  1.6114501961773586

In [17]: print('The kurtosis of speed is ',q9_b['SP'].kurtosis())

The kurtosis of speed is  2.9773289437871835

In [18]: print('The skewness of weight is ',q9_b['WT'].skew())

The skewness of weight is  -0.6147533255357768

In [19]: print('The kurtosis of weight is ',q9_b['WT'].kurtosis())

The kurtosis of weight is  0.9502914910300326
```

Speed:

Skewness	=	1.611
Kurtosis	=	2.977

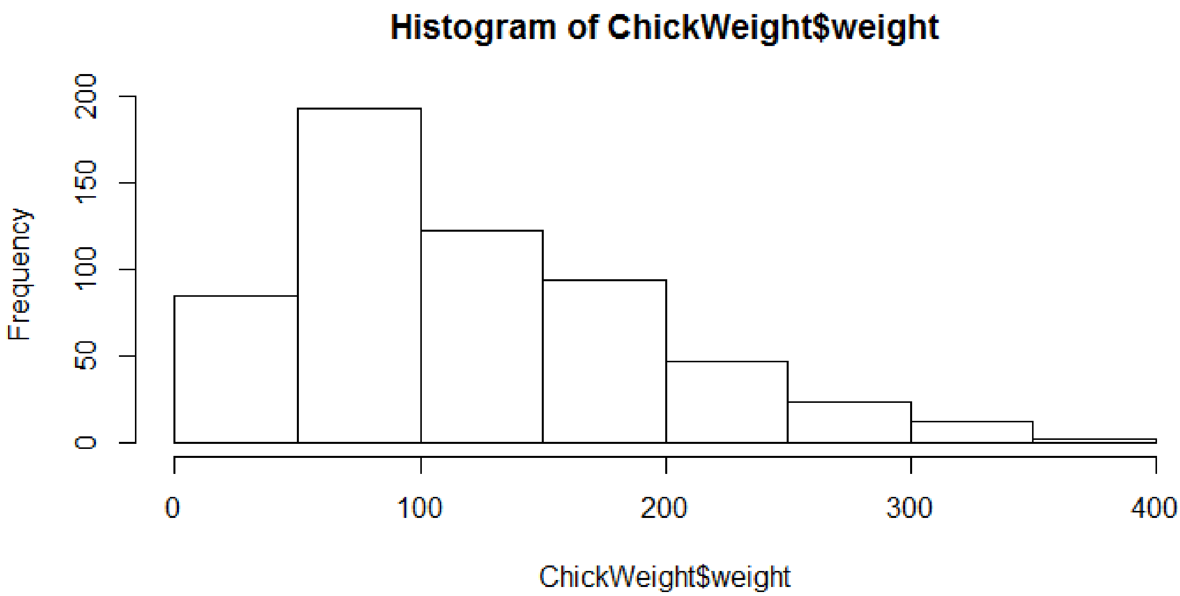
Weight:

Skewness	=	-0.614
Kurtosis	=	0.950

Inference:

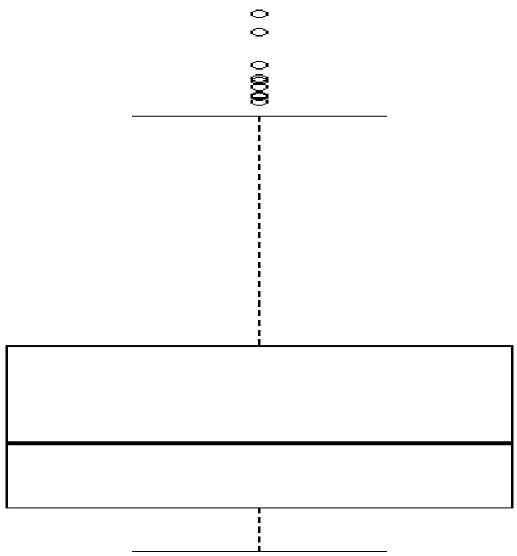
1. The skewness of SP is Positive, it signifies that the data are positively skewe d or right skewed
2. Mean>median>mode
3. The kurtosis of SP is Positive, it signifies that the curve is peaked with thick tails and data is Leptokurtic data.
4. The skewness of Weight is negative, it signifies that the data are negatively s kewed or left skewed
5. Mean<median<mode
6. The kurtosis of Weight is positive, it signifies that the curve is peaked with t hick tails and data is called as leptokurtic data.

Q10) Draw inferences about the following boxplot & histogram



Inference:

- The above histogram shows that the values are positively skewed or right skewed since more no of data present in left side
- The skewness values is greater than zero
- Thin tail present towards the right side
- Mean>median>mode



Inference:

- The boxplot shows that greater density of data present towards left side
- The data are positively skewed or right skewed
- The data contains outliers in positive direction, the dots represents the presence of outliers

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?

<u>ANS</u>:

```
In [46]: # t = 94%
         stats.t.ppf(.97,df=1999)

Out[46]: 1.8818614764780113

In [47]: # t = 98%
         stats.t.ppf(.99,df=1999)

Out[47]: 2.328214776106972

In [48]: #t = 96%
         stats.t.ppf(.98,df=1999)

Out[48]: 2.055089962825778
```

t94 = 1.88

t98 = 2.33

t96 = 2.05

Confidence Interval = $\overline{X} \pm t \frac{s}{\sqrt{n}}$

94% confidence interval:

= 200-1.882*(30/√2000) to 200+1.882*(30/√2000)

=198.73 to 201.26

98% confidence interval:

= 200-2.328*(30/√2000) to 200+2.328*(30/√2000)

=198.43 to 201.56

96% confidence interval:

= 200-2.055*(30/√2000) to 200+2.055*(30/√2000)

=198.62 to 201.378

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.

ANS:

```
In [20]: students_marks = pd.DataFrame([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

In [21]: students_marks.mean()
Out[21]: 0    41.0
         dtype: float64

In [22]: students_marks.median()
Out[22]: 0    40.5
         dtype: float64

In [23]: mode = students_marks.value_counts().head(1)
         mode
Out[23]: 41    4
         dtype: int64

In [24]: students_marks.var()
Out[24]: 0    25.529412
         dtype: float64

In [25]: students_marks.std()
Out[25]: 0    5.052664
         dtype: float64
```

Mean	=	41
Median	=	40.5
Mode	=	41
Variance	=	25.52
Std deviation	=	5.05

2) What can we say about the student marks?

ANS:

- The average mark of students in test is 41
- The maximum and minimum marks are 56 and 34
- Most of the students mark between 35 to 45

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

ANS:

No Skewness

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

ANS:

Positive Skewness

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

ANS:

Negative Skewness

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

ANS:

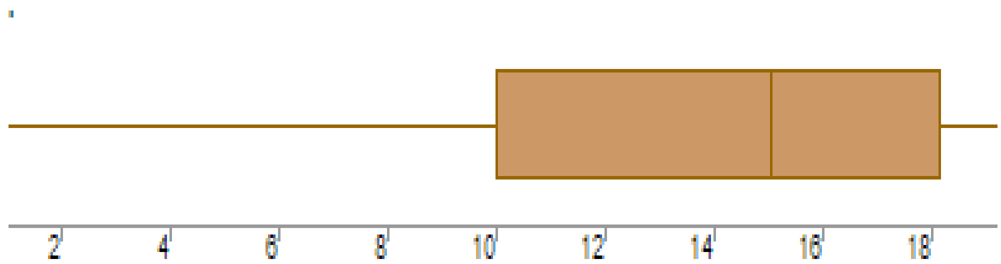
In Positive Kurtosis (>3), the distribution is peaked and has thick tails (i.e. most of the values in distribution located in tails rather than around the mean).

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

ANS:

In Negative Kurtosis (<3), the distribution is flat and has thin tails(i.e. it has fewer values in its shorter tails)

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



What can we say about the distribution of the data?

<u>ANS</u>:

- The distribution of this data is not Normally distributed
- Most of the values in left so it is left skewed
- The median value of the distribution is 15
- The Lower and Upper quartile are 10 and 18

What is nature of skewness of the data?

<u>ANS</u>:

Most of the values are skewed towards left side so the nature of the skewness is left skewed

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

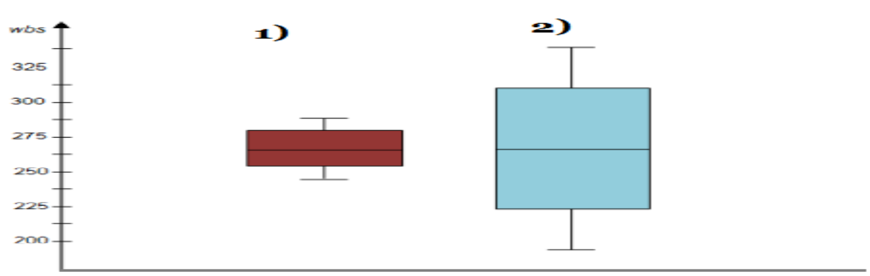
<u>ANS</u>:

IQR = QR3-QR1

=18-10

IQR=10

Q19) Comment on the below Boxplot visualizations?



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

ANS:

The given box plots are Normally distributed, both have median around 262.5 and they have no outlier.

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)

ANS:

```
In [29]: cars['MPG'].mean()
Out[29]: 34.422075728024666

In [30]: cars['MPG'].std()
Out[30]: 9.131444731795982

In [49]: #P(MPG>38)
         1-stats.norm.cdf(x=38,loc=34.422,scale=9.131)
Out[49]: 0.3475838101730251

In [50]: #P(MPG<40)
         stats.norm.cdf(x=40,loc=34.422,scale=9.131)
Out[50]: 0.729362470706113

In [51]: #P(20<MPG<50)
         stats.norm.cdf(x=50,loc=34.422,scale=9.131)-stats.norm.cdf(x=20,loc=34.422,scale=9.131)
Out[51]: 0.8988852898457339
```

- P(MPG>38) = 0.34
- P(MPG<40) = 0.72
- P (20<MPG<50) = 0.89

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution

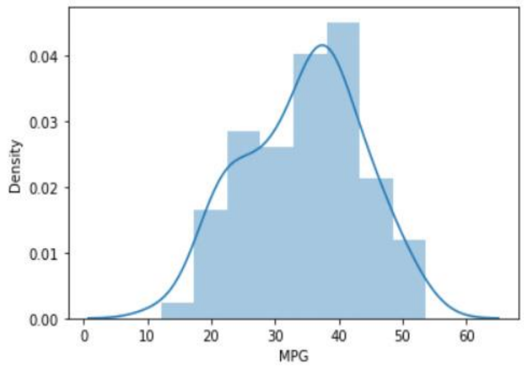
Dataset: Cars.csv

ANS:

```
In [34]: import warnings
         warnings.filterwarnings('ignore')
```

```
In [35]: sns.distplot(a=cars['MPG'])
         skew(cars['MPG'])
```

```
Out[35]: -0.17463433818755686
```



Skewness around -0.5 to 0.5 is acceptable, here skewness of MPG is -0.17 hence it follows normal distribution.

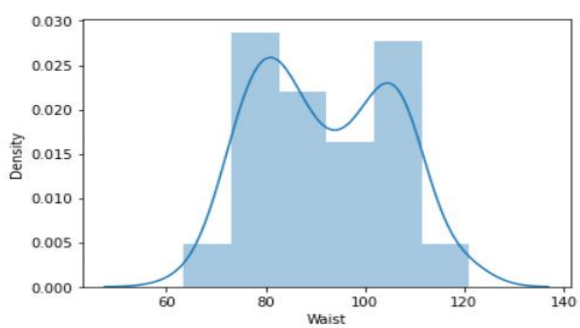
b) Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

ANS:

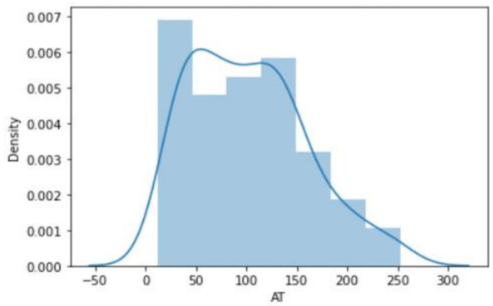
```
In [53]: sns.distplot(a=wc_at['Waist'])
         skew(wc_at['Waist'])
```

```
Out[53]: 0.1322041763592883
```



Skewness around -0.5 to 0.5 is acceptable, here skewness of Waist is 0.13 hence it follows normal distribution

```
In [54]: sns.distplot(a=wc_at['AT'])
         skew(wc_at['AT'])
Out[54]: 0.5767896975987847
```



Skewness around -0.5 to 0.5 is acceptable, here skewness of AT is 0.57 hence it follows normal distribution

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

<u>ANS:</u>

```
In [39]: stats.norm.ppf(1.90/2)
Out[39]: 1.6448536269514722

In [40]: stats.norm.ppf(1.94/2)
Out[40]: 1.8807936081512509

In [41]: stats.norm.ppf(1.60/2)
Out[41]: 0.8416212335729143
```

Z90 = 1.645
Z94 = 1.88
Z60 = 0.842

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

<u>ANS</u>:

```
In [42]: stats.t.ppf(.975,df=24)
Out[42]: 2.0638985616280205

In [43]: stats.t.ppf(.98,df=24)
Out[43]: 2.1715446760080677

In [44]: stats.t.ppf(.995,df=24)
Out[44]: 2.796939504772804
```

t95= 2.064

t96 = 2.172

t99 = 2.797

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

ANS:

```
In [52]: 1-stats.norm.cdf(x=260,loc=270,scale=90)

Out[52]: 0.5442358810453114
```

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 54.4%