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	Ordinal
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	Ratio
Years of Education	Ratio

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

Ans.

If three coins are tossed.

Probability = {HHH,HTH,HHT,THH,TTH,THT,HTT,TTT}=8

If two heads and one tail are obtained then possible outcomes

={HHT,HTH,THH}=3

Probability = 3/8 = 0.375

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.

Two dice are rolled then possible outcomes =

{(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)}

= 36

a) Possible outcomes of sum equal to 1 = 0

Probability = 0/36 =0

b) Possible outcomes of sum less than or equal to 4 = 6

Probability = 6/36=1/6

c) Sum is divisible by 2 and 3

Probability = 6/36 = 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.

Possible outcomes = 21

Two balls are draws random then possible outcomes = 10

Probablity of none of the ball is blue = 10/21=0.4761

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

Expected number of candies

= E(x)

= (1*0.015)+(4*0.2)+(3*0.65)+(5*0.05)+(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

- 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

	POINTS	SCORE	WEIGHT
MEAN	115.09	102.952	571.16
MEDIAN	3.695	3.325	17.71
VARIANCE	0.285881	0.957379	3.193166
STD.DEV	0.534679	0.978457	1.786943
MAX	4.93	5.424	22.9
MIN	2.76	1.513	14.5
RANGE	2.17	3.911	8.4

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.

Expected value = ∑(x)/n

= 1308/9

= 145.33

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

Cars speed and distance

Use Q9_a.csv

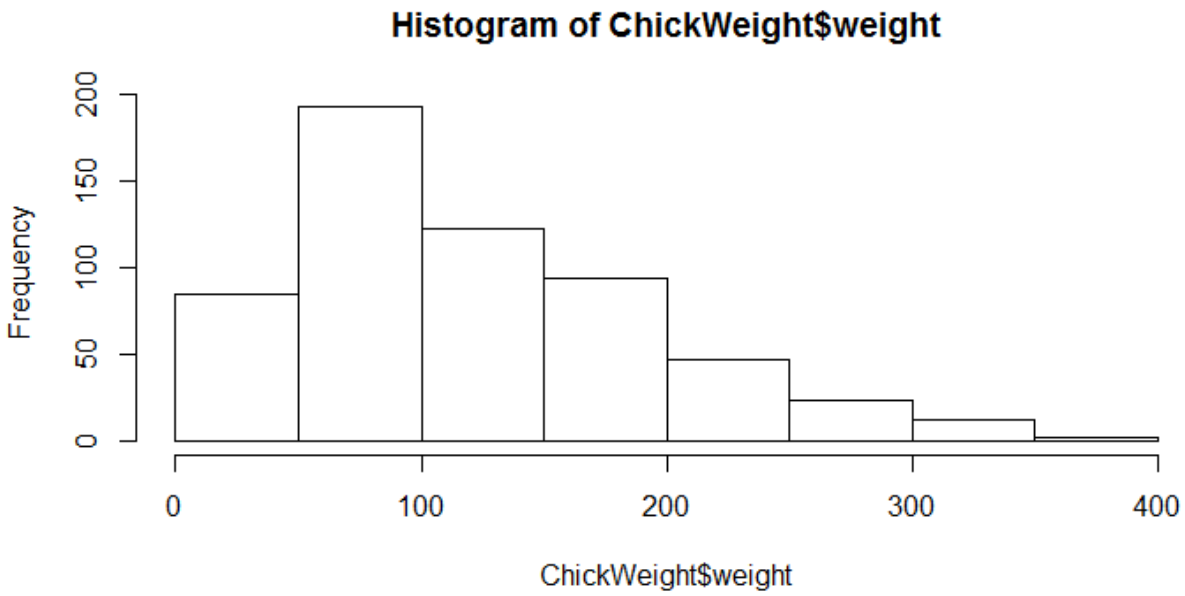
```
In [1]: from scipy import stats
        import scipy as sp
        from scipy.stats import kurtosis
        from scipy.stats import skew
        import pandas as pd

In [4]: df=pd.read_csv("Q9_a.csv")

In [5]: print(skew(df,axis=0,bias=True))
        [ 0.         -0.11395477  0.78248352]

In [6]: print(kurtosis(df,axis=0,bias=True))
        [-1.20096038 -0.57714742  0.24801866]
```

SP and Weight(WT)

Use Q9_b.csv

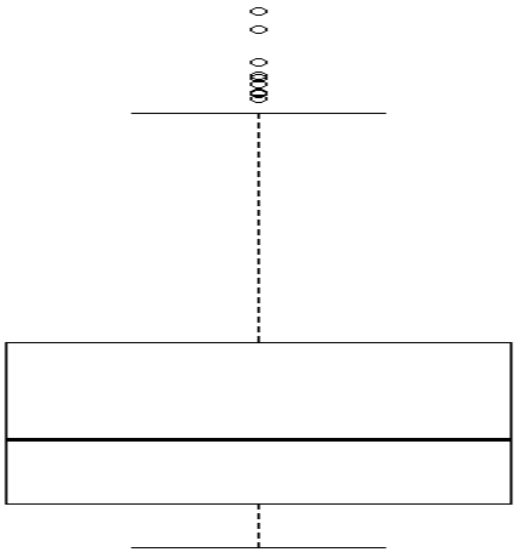
```
In [1]: from scipy import stats
        import scipy as sp
        from scipy.stats import kurtosis
        from scipy.stats import skew
        import pandas as pd

In [2]: df=pd.read_csv("Q9_b.csv")

In [3]: print(print(skew(df,axis=0,bias=True)))
        [ 0.          1.58145368 -0.60330993]
        None

In [4]: print(kurtosis(df,axis=0,bias=True))
        [-1.20036585  2.72352149  0.81946588]
```

Q10) Draw inferences about the following boxplot & histogram



Ans.

In the above Histogram we may conclude that the Distribution is right skewed.



In the above box plot we conclude that the outliers present in the datasets and the distribution is right skewed.

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?

Ans.

Population (N) = 3,000,000, n=2000, X = 200 , SD = 30

Confidence interval Estimate = $X \pm Z \frac{SD}{\sqrt{n}}$

$= 200 \pm Z \frac{30}{\sqrt{2000}}$

94% of confidence interval

$Z=\frac{(1-0.94)}{2} + (0.94) = 0.97 = 1.89$

Therefore $200 \pm 1.89 \frac{30}{\sqrt{2000}} = 200 \pm 2.56$

Lower limit = 197.44

Upper limit = 202.56

98 % of confidence interval

$Z = \frac{(1-0.98)}{2} + (0.98) = 0.99 = 2.33$

Therefore $200 \pm 2.33 \frac{30}{\sqrt{2000}} = 200 \pm 3$

Lower limit = 197

Upper limit = 202

96 % of confidence interval

Z = $\frac{(1-0.96)}{2}$+(0.96) = 0.99 =2.06

Therefore 200 ± 2.03 $\frac{30}{\sqrt{2000}}$ = 200 ± 3

Lower limit = 197.27

Upper limit = 202.73

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?**

Ans.

1) **Mean** = 41

Median = 40.5

Variance = 25.52

SD = 5.052

2) From the students test score we obtained that mean is greater than Median. Distribution is skewed towards toward right and no outlier are present.

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

Ans .

When mean median of data are equal then there is no skewness and distribution is symmetric.

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

Ans.

Nature of skewness when mean>median then it is right skewed distribution.

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

Ans.

Nature of skewness when median>mean then it is left skewed Distribution.

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

Ans.

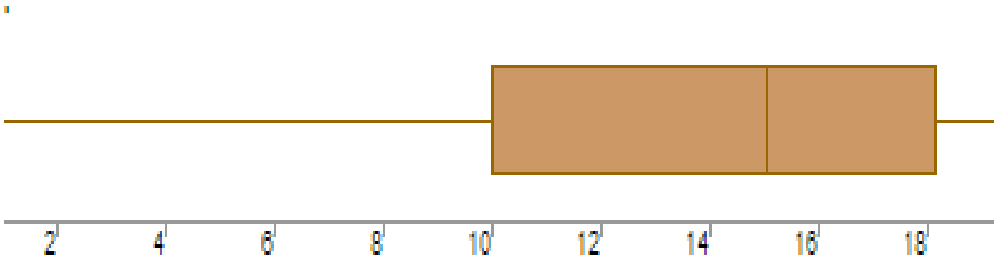
Positive kurtosis value indicates that the distribution is more peaked and Possesses thick tails.

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

Ans.

Negative Kurtosis value indicates that the distribution is not proper peaked Thin tails.

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



What can we say about the distribution of the data?

Ans.

Negative Skewed distribution.

What is nature of skewness of the data?

Ans.

Left skewed distribution.

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

Ans.

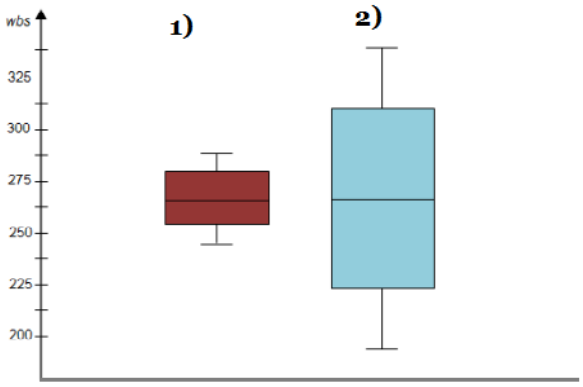
= IQR

= Upper Quartile – Lower Quartile

= 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.

Ans.

The above boxplot follow the same median and follow the normal Distribution, Outlier does not exist in above box-plot and the range of 1st box-plot is less than 2nd.

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 [1]:
```
import pandas as pd
```

In [2]:
```
df=pd.read_csv("Cars.csv")
```

In [3]:
```
df.mean()
```

Out[3]:
```
HP     117.469136
MPG     34.422076
VOL     98.765432
SP     121.540272
WT      32.412577
dtype: float64
```

In [4]:
```
df.std()
```

Out[4]:
```
HP     57.113502
MPG     9.131445
VOL    22.301497
SP     14.181432
WT      7.492813
dtype: float64
```

In [5]:
```
# p(MGP>38)
from scipy import stats
print(1-stats.norm.cdf(38,34.422076,9.131445))
```

```
0.34759394041453007
```

In [6]:
```
# p(MPG<40)
print(stats.norm.cdf(40,34.422076,9.131445))
```

```
0.7293498604157946
```

In [7]:
```
# p(20<MPG<50)
print(stats.norm.cdf(50,34.422076,9.131445)-stats.norm.cdf(20,34.422076,9.13144))
```

```
0.898869006747862
```

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

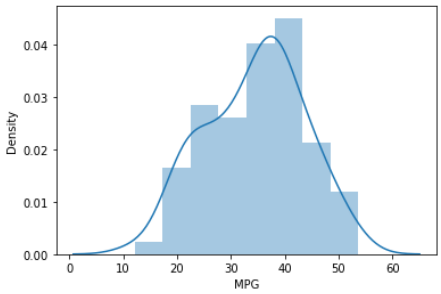
In [1]:
```
import pandas as pd
import seaborn as sns
```

In [2]:
```
df=pd.read_csv("Cars.csv")
```

In [3]:
```
sns.distplot(df["MPG"])
```

```
C:\Users\chand\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function an
d will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar fle
xibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[3]:
```
<AxesSubplot:xlabel='MPG', ylabel='Density'>
```



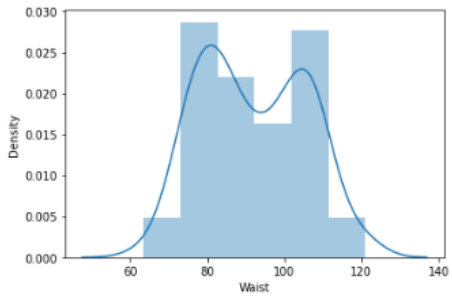
b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

```
In [1]: import pandas as pd
        import seaborn as sns

In [2]: df=pd.read_csv("wc-at.csv")

In [3]: sns.distplot(df["Waist"])
```

```
C:\Users\chand\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function an
d will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar fle
xibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

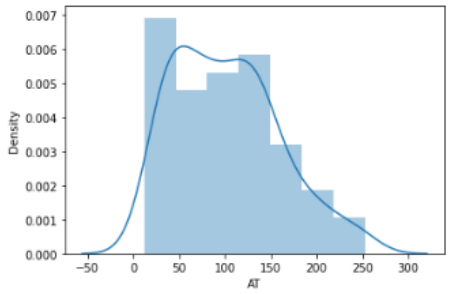
```
Out[3]: <AxesSubplot:xlabel='Waist', ylabel='Density'>
```




```
In [4]: sns.distplot(df["AT"])
```

```
C:\Users\chand\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function an
d will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar fle
xibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
Out[4]: <AxesSubplot:xlabel='AT', ylabel='Density'>
```



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

1) Z scores of 90% confidence interval

Z = 1-0.90/2 =0.10/2=0.05

Z = 1.645

2) Z scores of 94% confidence interval

Z = 1.0.94/2 = 0.06/2 = 0.03

Z = 1.881

3) Z scores of 60% confidence interval

Z =1-0.60/2= 0.040/2 = 0.2

Z= 0.253

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

Ans.

n = 25 , n-1 = 25 -1 = 24

```
In [5]: from scipy import stats
        from scipy.stats import norm

In [6]:  t scores of 95% confidence interval for sample size of 25
        stats.t.cdf(0.975,24)

Out[6]: 0.8303570471638759

In [7]:  t scores of 95% confidence interval for sample size of 25
        stats.t.cdf(0.98,24)

Out[7]: 0.8315688116127068

In [8]:  t scores of 95% confidence interval for sample size of 25
        stats.t.cdf(0.995,24)

Out[8]: 0.8351685156761681
```

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

Standred deviation = 90

Last average = 270, last average = 260

n = 18

df = n-1 = 18-1 = 17

tscore = -10/21.23

= -0.47

Recode → pt(tscore,df)

= pt (-0.47,17)

[1] 0.3221639

Probability of 18 randmaly seleted bulbs would have an average life of no more than 260 days

= 32%