Q1) Identify the Data type for the Following:

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

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

Nominal, Ordinal, Interval, Ratio.

ANS:

Nominal: the data can only be categorized.

Ordinal: the data can be categorized and ranked.

Interval: the data can be categorized and ranked, and evenly spaced.

Ratio: the data can be categorized, ranked, evenly spaced and has a natural zero.

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	Ratio
Time on a Clock with Hands	Ratio
Number of Children	Ordinal
Religious Preference	nominal
Barometer Pressure	Ratio
SAT Scores	Ratio
Years of Education	Interval

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

ANS:

Outcomes when three coins are tossed: {HHH,HHT,HTH,HTT,TTT,TTH,THT,THH}

Possibilities ={HHT,THH,HTH}

=3

Probability = number of possibilities of 2 heads and 1 tail/ total number of outcomes

=3/8

=37.5%

Therefore, the probability of getting 2 heads and 1 tail is 37.5%

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:

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)}

Chances of getting sum equal to 1=0

Probability (Chances of getting sum equal to 1)=0/36

=0

Chances of getting sum less than or equal to 4=6

Probability (Chances of getting sum less than or equal to 4)=6/36

=1/6

=16.6%

Chances of getting a number which sum is divisible by 2&3=6

Probability (Chances of getting a number which sum is divisible by 2&3) =6/36

=1/6

=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?

ANS:

Total number of balls(red, green, blue)=2+3+2

=7

Probability of one ball which can't be blue=5/7

Now,

Total no.of balls =6

Balls which can't be blue =4

So, the probability of second ball not being blue = 4/6

So probability of two balls not being blue = 5/7* 4/6

=20/42

=0.4761904761904

=47.6%

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

ANS:

Expected value=Sum of product of probability and X

$$EV = \sum P(X_i) \times X_i$$

CHILD	CANDIES COUNT(X)	PROBABILITY(P)	P*X		
A	1	0.015	0.015		
B	4	0.2	0.8		
C	3	0.65	1.95		
D	5	0.005	0.025		
E	6	0.01	0.06		
F	2	0.12	0.24		
			3.09	#--->expected value	

Therefore, 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.

Use Q7.csv file

ANS:

	POINTS	SCORE	WEIGHT
MEAN	3.596563	3.21725	17.84875
MEDIAN	3.695	3.325	17.71
MODE	3.92	3.44	17.02
VARIANCE	0.276948	0.927461	3.093379688
SD	0.534679	0.978457	1.786943236
RANGE:			
HIGHEST VALUE	4.93	5.424	22.9
LOWEST VALUE	2.76	1.513	14.5
RANGE:	2.17	3.911	8.4

INFERENCE**:**

The most occurring value in points is 3.92,the most occurring value in score is 3.44, the most occurring value in weights is 17.02

If the weight of the Merc 280 is reduced to 17.02, there is a possibility that the car will have maximum sales.

	Points	Score	Weigh
Mazda RX4	3.9	2.62	16.46
Mazda RX4 Wag	3.9	2.875	17.02
Datsun 710	3.85	2.32	18.61
Hornet 4 Drive	3.08	3.215	19.44
Hornet Sportabout	3.15	3.44	17.02
Valiant	2.76	3.46	20.22
Duster 360	3.21	3.57	15.84
Merc 240D	3.69	3.19	20
Merc 230	3.92	3.15	22.9
Merc 280	3.92	3.44	18.3
Merc 280C	3.92	3.44	18.9

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:

N	X	probabilty(P)	PX			
1	108	0.11	12			
2	110	0.11	12.22222			
3	123	0.11	13.66667			
4	134	0.11	14.88889			
5	135	0.11	15			
6	145	0.11	16.11111			
7	167	0.11	18.55556			
8	187	0.11	20.77778			
9	199	0.11	22.11111			
			145.3333	#sum	#--->expected value	

Therefore , the expected value of the patient is 145.3333.

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

Cars speed and distance

Use Q9_a.csv

SP and Weight(WT)

Use Q9_b.csv

ANS:

Skewness:

Skewness is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images. A distribution can have right (or positive), left (or negative), or zero skewness.

Kurtosis :

Kurtosis is a measure of the tailedness of a distribution. Tailedness is how often outliers occur.

9(a):

Index	speed	dist			
1	4	2		-0.11751	#skewness
2	4	10		0.405053	#kurtosis
3	7	4			
4	7	22			
5	8	16			
6	9	10			
7	10	18			
8	10	26			
9	10	34			

INFERENCE:

1. For the given data, the skewness is negative . So, it is left skewed.
2. Since skewness is -0.11751 and kurtosis is 0.405053 , the graph is asymmetrical.

9(b):

	SP	WT			
1	104.1854	28.76206			
2	105.4613	30.46683		0.120049	#skewness
3	105.4613	30.1936		-1.67197	#kurtosis
4	113.4613	30.63211			
5	104.4613	29.88915			
6	113.1854	29.59177			
7	105.4613	30.30848			
8	102.5985	15.84776			
9	102.5985	16.35948			
10	115.6452	30.92015			
11	111.1854	29.36334			
12	117.5985	15.75353			
13	122.1051	32.81359			
14	111.1854	29.37844			
15	108.1854	29.24728			

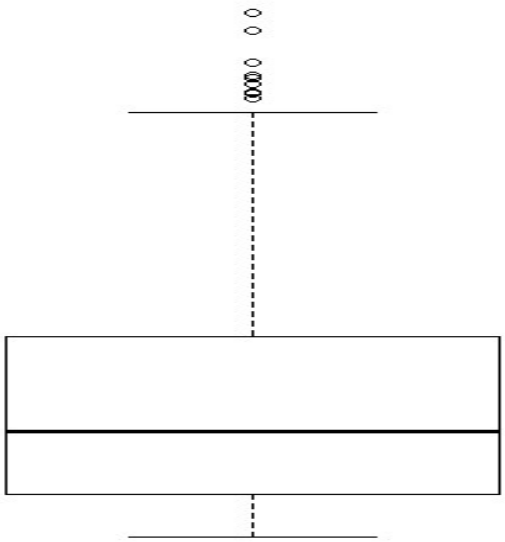
INFERENCE:

1. Kurtosis of SP and WT is -1.67197 , that indicates that the distribution has lighter tails than the normal distribution.
2. The skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a distribution. Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).

Q10) Draw inferences about the following boxplot & histogram

ANS:

1. The graph shows, right skewedness because the frequency of the chickweight is decreasing as the weight increases.
2. ChicktheWeight 100 has the highest frequency.



ANS: The outliers exist in the upper part of the boxplot, which is the negative side.

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:

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

```
     import scipy.stats as stats
     import math
```

```
[4]  # Given data
     sample_mean = 200  # Sample mean weight
     sample_std = 30     # Sample standard deviation
     sample_size = 2000  # Sample size
     population_size = 3000000  # Population size
```

```
[5]  # Calculate the standard error
     standard_error = sample_std / math.sqrt(sample_size)

     # Calculate the critical values for 94%, 98%, and 96% confidence levels
     confidence_levels = [0.94, 0.98, 0.96]
     critical_values = [stats.norm.ppf(1 - (1 - cl) / 2) for cl in confidence_levels]

     # Calculate the confidence intervals
     confidence_intervals = [
         (sample_mean - critical_value * standard_error,
          sample_mean + critical_value * standard_error)
         for critical_value in critical_values
     ]
```

```
[6]  # Print the confidence intervals
     for cl, interval in zip(confidence_levels, confidence_intervals):
         print(f"{int(cl * 100)}% Confidence Interval: {interval}")

     94% Confidence Interval: (198.738325292158, 201.261674707842)
     98% Confidence Interval: (198.43943840429978, 201.56056159570022)
     96% Confidence Interval: (198.62230334813333, 201.37769665186667)
```

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. From the data, we can say that the student scores 40-41 marks every time (average).

34	40.5	#median
36	41	#mean
36	41	#mode
38	24.11111	#variance
38	4.910307	#sd
39		
39		
40		
40		
41		
41		
41		
41		
42		
42		
45		
49		
56		

3. Mean :41
4. Median :40.5
5. Variance:24.11111
6. Standard deviation :4.910307

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

ANS: When the mean and median of a dataset are equal, it implies that the data is symmetrically distributed around the central value. In such a case, the skewness of the data is zero.

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

ANS: When mean > median, the data has a positive skewness, indicating a longer and fatter right tail in the distribution.

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

ANS: When median > mean, the data has a negative skewness, indicating a longer and fatter left tail in the distribution.

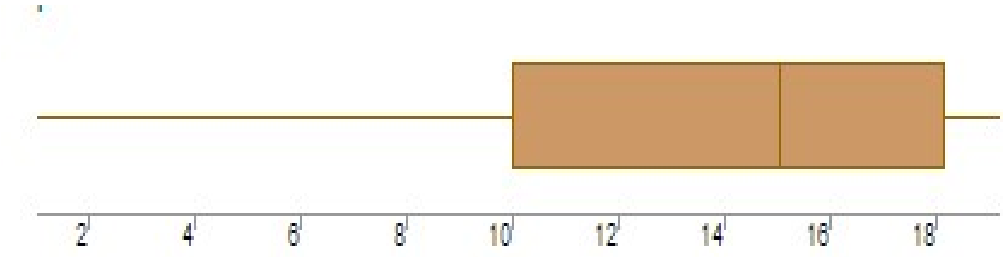
Q16) What does positive kurtosis value indicates for a data ?

ANS: A positive kurtosis value indicates that the data has heavy tails and a relatively peaked compared to a normal distribution.

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

ANS: A negative kurtosis value in a dataset indicates that the distribution has lighter or less extreme tails than a normal distribution.

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



What can we say about the distribution of the data?

ANS: the data is not equally distributed across the plane. There might be outliers influencing the data. The median of the data is 14.7(approx)

25% of the data lies between 0-10

50% of the data lies between 10-18

25% of the data lies after 18-20(approx.)

What is nature of skewness of the data?

ANS: it is left skewed data, as the whisker lengt on the upper quadrant is higher than the data on the lower quadrant. Median will be greater than the mean as it is left skewed.

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

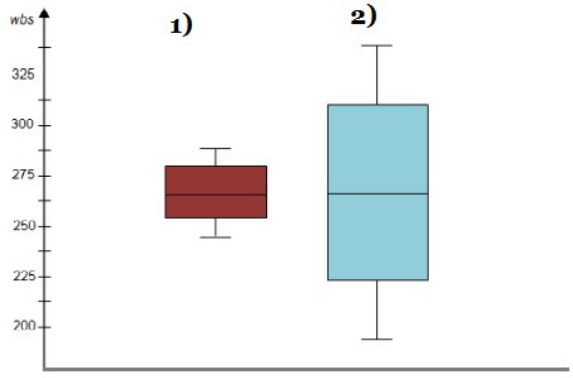
Q1=10

Q2=14.7

Q3=18

IQR=Q3-Q1=8(Approx)

Q19) Comment on the below Boxplot visualizations?



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

ANS:

1. There are no outliers.
2. The boxplot 1 and boxplot 2 has the same median but different sizes.
3. Q1,Q3 of boxplot1 is 275, 250 respectively.
4. Q1,Q3 of boxplot2 is 300, 225 respectively.

5. IQR for boxplot1 and boxplot2 is 262.5
6. The data spread in boxplot 1 and 2 are both symmetrical.

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:

HP	MPG	VOL	SP	WT			
49	53.70068	89	104.1854	28.76206		53.70068	#maximum
55	50.0134	92	105.4613	30.46683		12.10126	#minimum
55	50.0134	92	105.4613	30.1936		41.59942	#range
70	45.69632	92	113.4613	30.63211		34.42208	#mean
53	50.50423	92	104.4613	29.88915		9.074903	#sd
70	45.69632	89	113.1854	29.59177		2788.188	#sum
55	50.0134	92	105.4613	30.30848			
62	46.71655	50	102.5985	15.84776		probabilities	
62	46.71655	50	102.5985	16.35948	a.	0.346692	#P(MPG>38)
80	42.29908	94	115.6452	30.92015	b.	0.730608	#P(MPG<40)
73	44.65283	89	111.1854	29.36334		0.956973	#50
92	39.35409	50	117.5985	15.75353		0.056005	#20
92	39.35409	99	122.1051	32.81359	c.	0.900969	#P (20<MPG<50)
73	44.65283	89	111.1854	29.37844			
66	45.73489	89	108.1854	29.34728			

a. P(MPG>38)

```
=1-NORM.DIST(38,G5,G6,G7)
```

0.346692	#P(MPG>38)

=34%

b. P(MPG<40)

```
=NORM.DIST(40,G5,G6,G7)
0.730608 #P(MPG<40)
```

=73%

c. P (20<MPG<50)

=(=NORM.DIST(50,G5,G6,G7))-(=NORM.DIST(20,G5,G6,G7))

= =G12-G13

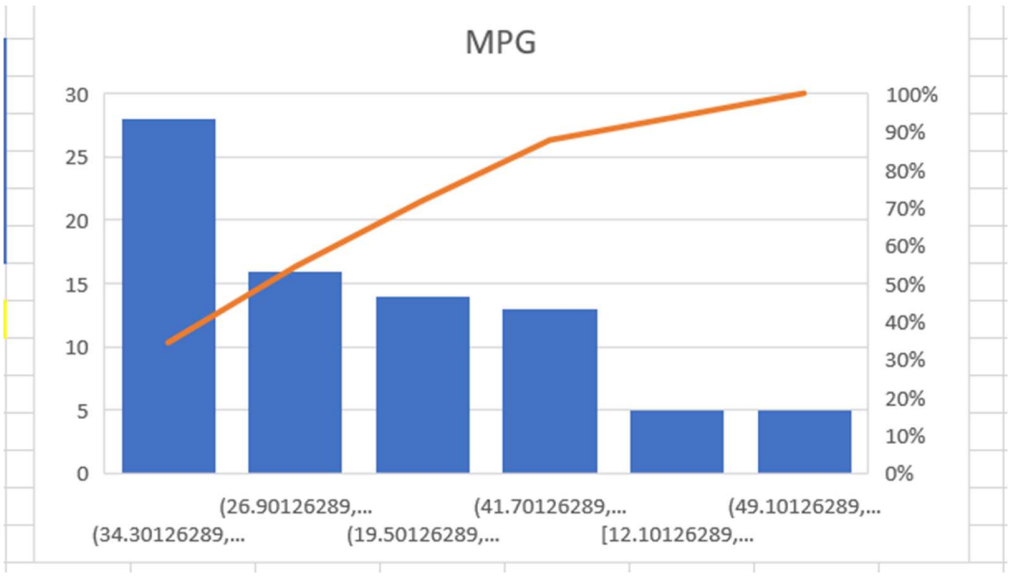
= 0.900969 #P (20<MPG<50)

=90%

Q 21) Check whether the data follows normal distribution

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

ANS:

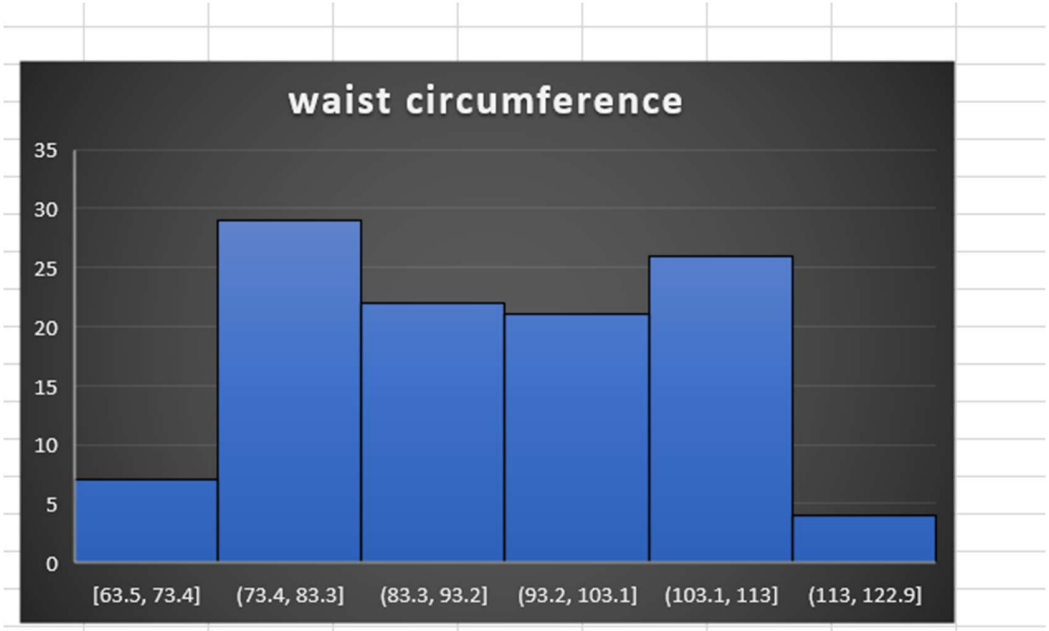


The data is not normally distributed. The graph line is not bell-shaped.

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

Dataset: wc-at.csv

ANS:



The waist circumference is normally distributed.

91.90183	#mean
90.8	#median

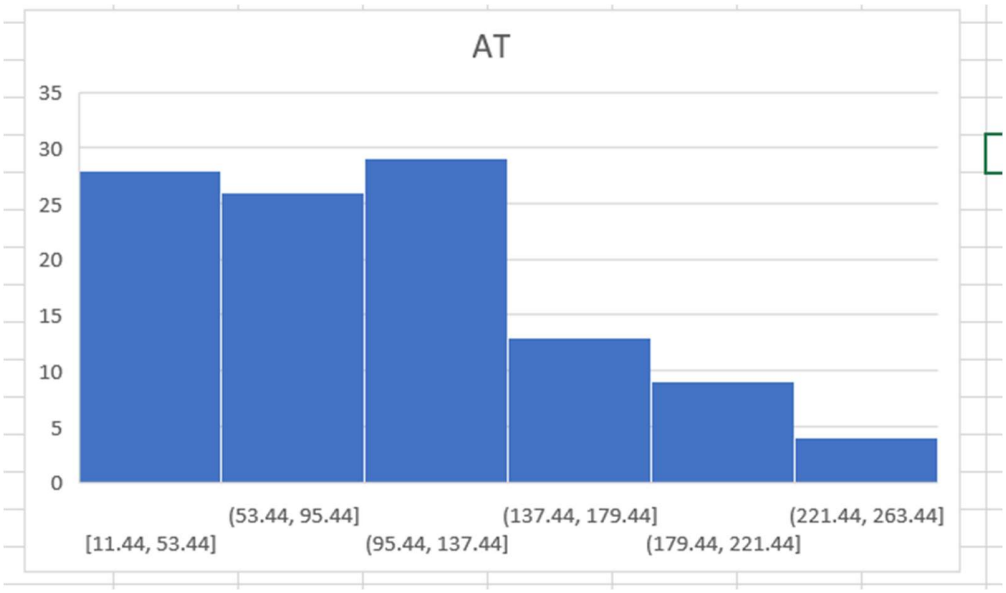
The mean and median are almost the same, so we can consider this as normal distribution.

0.134056	#skewness

The skewness is also almost zero.

-1.10267	#kurtosis

The peak is wide and has thinner tail, so it has negative kurtosis.



From the graph, we can say that the data is not normally distributed.

```
101.894 #mean
 96.54  #median
```

The mean is greater than median, so it is righted skewed (positive skewness).

```
0.584869 #skewness
```

The peak is wide and thin, so it has negative kurtosis

```
-0.28558 #kurtosis
```

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

ANS:

```
import numpy as np
import pandas as pd
```

```
[3] import scipy.stats as stats
    import math
```

```
# Define the confidence levels
confidence_levels = [0.90, 0.94, 0.60]

# Calculate the Z-scores for each confidence level
z_scores = [stats.norm.ppf(1 - (1 - cl) / 2) for cl in confidence_levels]

# Print the Z-scores
for cl, z_score in zip(confidence_levels, z_scores):
    print(f"{int(cl * 100)}% Confidence Interval Z-score: {z_score:.2f}")
```

```
90% Confidence Interval Z-score: 1.64
94% Confidence Interval Z-score: 1.88
60% Confidence Interval Z-score: 0.84
```

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

ANS:

```
import numpy as np
import pandas as pd
```

```
[3] import scipy.stats as stats
    import math
```

```
[11] # Sample size
     sample_size = 25

     # Degrees of freedom for a t-distribution (n - 1)
     degrees_of_freedom = sample_size - 1

     # Define the confidence levels
     confidence_levels = [0.95, 0.96, 0.99]

     # Calculate the t-scores for each confidence level
     t_scores = [stats.t.ppf(1 - (1 - cl) / 2, df=degrees_of_freedom) for cl in confidence_levels]

     # Print the t-scores
     for cl, t_score in zip(confidence_levels, t_scores):
         print(f"{int(cl * 100)}% Confidence Interval t-score: {t_score:.2f}")
```

```
95% Confidence Interval t-score: 2.06
96% Confidence Interval t-score: 2.17
99% Confidence Interval t-score: 2.80
```

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:

```
import numpy as np
import pandas as pd
```

```
[3] import scipy.stats as stats
    import math
```

```
[12] # Given data
     sample_mean = 260
     population_mean = 270
     sample_std = 90
     sample_size = 18

     # Calculate the t-score
     t_score = (sample_mean - population_mean) / (sample_std / math.sqrt(sample_size))

     # Calculate the degrees of freedom
     degrees_of_freedom = sample_size - 1

     # Calculate the probability using the cumulative distribution function (CDF)
     probability = stats.t.cdf(t_score, df=degrees_of_freedom)

     print(f"The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is: {probability:.4f}")

     The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is: 0.3217
```