Q1) Identify the Data type for the Following:

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

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	Continuous
Weight	Continuous
Hair Color	Nominal
Socioeconomic Status	Nominal
Fahrenheit Temperature	Continuous
Height	Continuous
Type of living accommodation	Ordinal
Level of Agreement	Ordinal
IQ(Intelligence Scale)	Ordinal
Sales Figures	Discrete
Blood Group	Nominal
Time Of Day	Continuous
Time on a Clock with Hands	Continuous
Number of Children	Discrete

Religious Preference	Nominal
Barometer Pressure	Continuous
SAT Scores	Continuous
Years of Education	Continuous

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

Ans:

Total Possible Outcomes: 8

HHH, **HHT, HTH, THH**, TTH, THT, HTT, TTT

Possibility of Two Heads and One Tail: 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:

a) The minimum possible outcome for Two Dice is 1+1 = 2
there is no Possibility of Getting "1".
So the probability of Getting 1 when two dice is rolled is 0

b) When we roll two dice

Total Possible Outomes: 6 × 6 = 36

When we roll two dice, the possibility of getting number less than or equal to 4 is: (1,1),(1,2),(1, 3), (2, 2), (3, 1),(2,1).

Favourable outcomes: 6

Therefore, total Probability Getting less than or equal to 4 is: 6/36 => 1/6

c) When we roll two dice,

minimum possible outcome: 2

maximum Possible Outcome: 12

The numbers which are divisible by 2 and 3 in between 2 and 12 are: 6 and 12

(i) Outcomes for 6 are: (1,5),(2,4),(3,3),(4,2),(5,1) =>5

(ii) Outcomes for 12 are: (6,6) =>1

Therefore the probability of Getting numbers which are divisible by 2 and 3 are: 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:

Total Balls is 2 Red + 3 Green + 2 Blue = 7 balls

Total outcomes for picking two balls at a time = $7C_2$

None of the ball is blue is 7-2 = 5 => $5C_2$

Total Probability of Getting none of the balls is Blue is: $\frac{5C_2}{7C_2}$

$$\frac{\frac{5!}{3!\times 2!}}{\frac{7!}{5!\times 2!}} \Rightarrow \frac{\frac{5\times 4\times 3\times 2\times 1}{(3\times 2\times 1)(2\times 1)}}{\frac{7\times 6\times 5\times 4\times 3\times 2\times 1}{(5\times 4\times 3\times 2\times 1)(2\times 1)}} \Rightarrow \frac{10}{21}$$

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

Ans:

The Expected number of Candies for a randomly Selected Child is:

$$1 \times 0.015 + 4 \times 0.20 + 3 \times 0.65 + 5 \times 0.005 + 6 \times 0.01 + 2 \times 0.120$$

$$0.015+0.8+1.95+0.025+0.06+0.24$$

The required answer 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:

The Required parameters are drawn by using python code (Q7.py). The final results are:

S. No	Parameter	Points	Score	Weigh
1	Mean	3.5965	3.2172	17.487
2	Median	3.695	3.325	17.71
3	Mode	3.07 & 3.92	3.44	17.02 & 18.90
4	Variance	0.2859	0.9573	3.193
5	Standard Deviation	0.53467	0.9784	1.7869
6	Range	2.17	3.911	8.3999

Python Program:

```
import pandas as pd
df=pd.read_csv("Q7.csv")
df
# Calculating the parameters Mean, Median, Mode, Variance, Standard Deviation, Range
# ********* for Points column **********
```

```
PMean=df["Points"].mean()
PMedian=df["Points"].median()
PMode=df["Points"].mode()
PVariance=df["Points"].var()
PStd=df["Points"].std()
PRange=df["Points"].max()-df["Points"].min()
print("The Required Parameters of Points Column are:")
print("Mean of the points is: ",PMean)
print("Median of the points is: ",PMedian)
print("Mode of the points is: ",PMode)
print("Variance of the points is: ",PVariance)
print("Standard Deviation of the points is: ",PStd)
print("Range of the points is: ",PRange)
# ********* for Score column **********
SMean=df["Score"].mean()
SMedian=df["Score"].median()
SMode=df["Score"].mode()
SVariance=df["Score"].var()
SStd=df["Score"].std()
SRange=df["Score"].max()-df["Score"].min()
print("The Required Parameters of Score Column are:")
print("Mean of the Score is: ",SMean)
print("Median of the Score is: ",SMedian)
print("Mode of the Score is: ",SMode)
print("Variance of the Score is: ",SVariance)
print("Standard Deviation of the Score is: ",SStd)
print("Range of the Score is: ",SRange)
# ********* for Weigh column **********
WMean=df["Weigh"].mean()
WMedian=df["Weigh"].median()
WMode=df["Weigh"].mode()
WVariance=df["Weigh"].var()
WStd=df["Weigh"].std()
WRange=df["Weigh"].max()-df["Weigh"].min()
print("The Required Parameters of Weigh Column are: ")
```

```
print("Mean of the Weigh is: ",WMean)
print("Median of the Weigh is: ",WMedian)
print("Mode of the Weigh is: ",WMode)
print("Variance of the Weigh is: ",WVariance)
print("Standard Deviation of the Weigh is: ",WStd)
print("Range of the Weigh is: ",WRange)
```

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 = ∑ (probability × Value)

∑ P(x).E(x)

Total Number of Patients are : 9

Probability of selecting each patient = $\frac{1}{9}$

Expected Value is:

=> $\frac{1}{9}\times 108 + \frac{1}{9}\times 110 + \frac{1}{9}\times 110 + \frac{1}{9}\times 123 + \frac{1}{9}\times 134 + \frac{1}{9}\times 135 + \frac{1}{9}\times 145 + \frac{1}{9}\times 167 + \frac{1}{9}\times 187 + \frac{1}{9}\times 199$

=> $\frac{1}{9}(108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)$

=> $\frac{1}{9}$ (1308)

=> 145.33

Expected Value of the Weight of that patient = 145.33

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

Cars speed and distance

Use Q9_a.csv

Ans:

The Required parameters are drawn by using python code (Q9_a.py). The final results are:

Python Program:

```
import pandas as pd
df=pd.read_csv("Q9_a.csv")
df.head()
#### Skewness and Kurtosis of Speed Column ####
x=df["speed"].skew()
y=df["speed"].kurt()
print("The Skewness of Speed is: ",x.round(3))
print("The Kurtosis of Speed is: ",y.round(3))
# ================================================================
#### Skewness and Kurtosis of Distance Column ####
a=df["dist"].skew()
b=df["dist"].kurt()
print("The Skewness of distance is: ",a.round(3))
print("The Kurtosis of distance is: ",b.round(3))
```

1. For Speed:
   a. Skewness = -0.118
   b. Kurtosis = -0.509
2. For Distance:
   a. Skewness = 0.807
   b. Kurtosis = 0.405

SP and Weight(WT)

Use Q9_b.csv

Ans:

The Required parameters are drawn by using python code (Q9_b.py). The final results are:

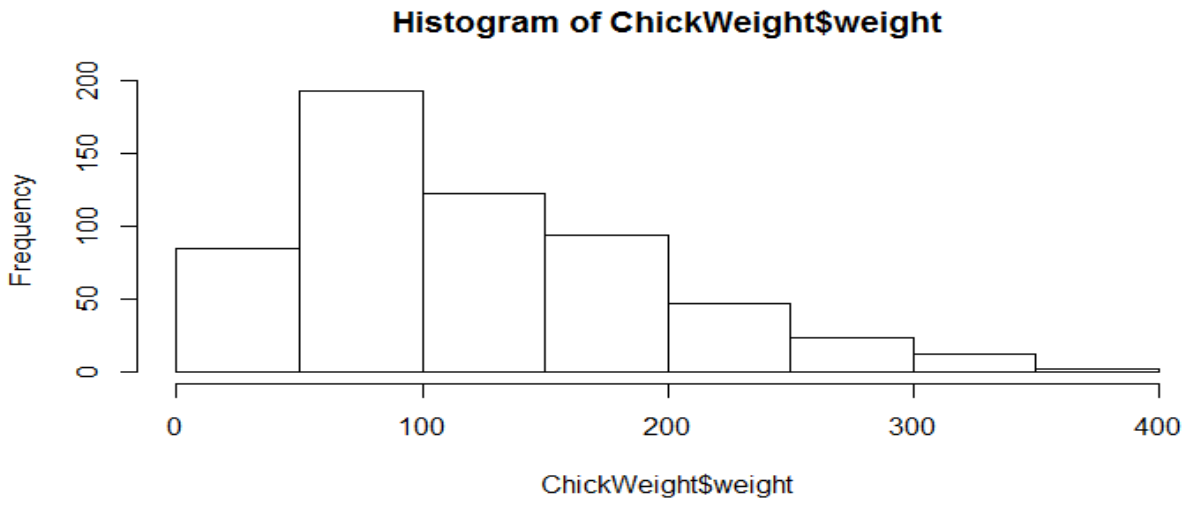
1. For Speed:
   a. Skewness = 1.611
   b. Kurtosis = 2.977

2. For Distance:
    a. Skewness = -0.615
    b. Kurtosis = 0.95

Python Program:

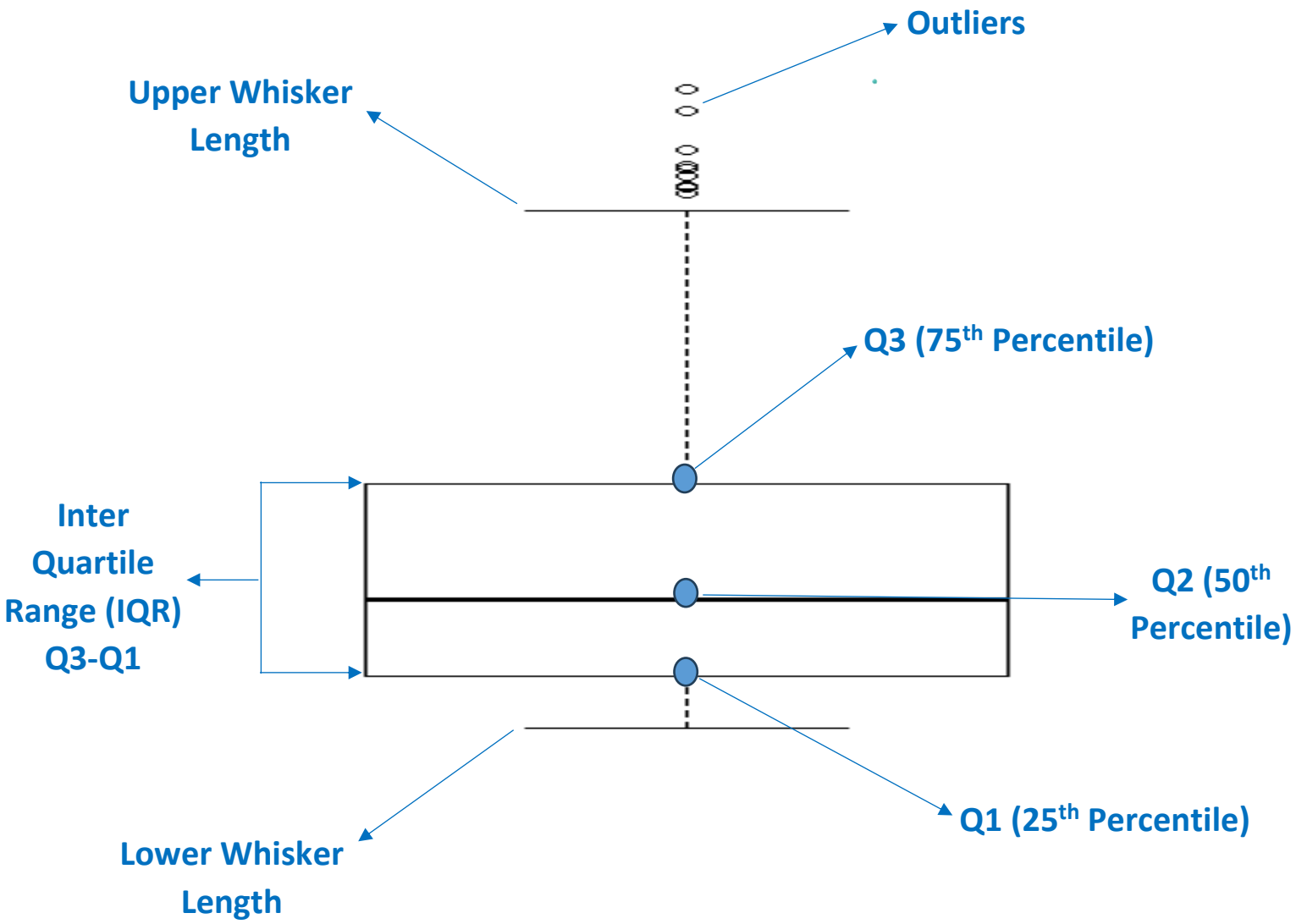
```
import pandas as pd
df=pd.read_csv("Q9_a.csv")
df.head()
#### Skewness and Kurtosis of Speed Column ####
x=df["speed"].skew()
y=df["speed"].kurt()
print("The Skewness of Speed is: ",x.round(3))
print("The Kurtosis of Speed is: ",y.round(3))
# ============================================================
#### Skewness and Kurtosis of Distance Column ####
a=df["dist"].skew()
b=df["dist"].kurt()
print("The Skewness of distance is: ",a.round(3))
print("The Kurtosis of distance is: ",b.round(3))
```

Q10) Draw inferences about the following boxplot & histogram



Ans:

1. The above Histogram is Positively Skewed.
2. In Positively Skewed, Mode <Median<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?

Ans:

Confidence Interval(CI) = $x \pm z \times \frac{\sigma}{\sqrt{n}}$

X= mean => 200
σ = Standard Deviation => 30
n = Sample Size = 2000
Z values from the Z table:
94 = 1.882
98 = 2.326
96 = 2.053

$\frac{\sigma}{\sqrt{n}} = \frac{30}{\sqrt{2000}} = 0.6708$

(i) $200 \pm 1.882 \times 0.6708 = (201.2624, 198.737)$
At 94% Confidence Interval Lies between 198.737 & 201.2624

(ii) $200 \pm 2.326 \times 0.6708 = (201.5602, 198.4398)$
At 94% Confidence Interval Lies between 198.4398 & 201.5602

(iii) $200 \pm 2.053 \times 0.6708 = (201.3771, 198.6229)$
At 94% Confidence Interval Lies between 198.6229 & 201.3771

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

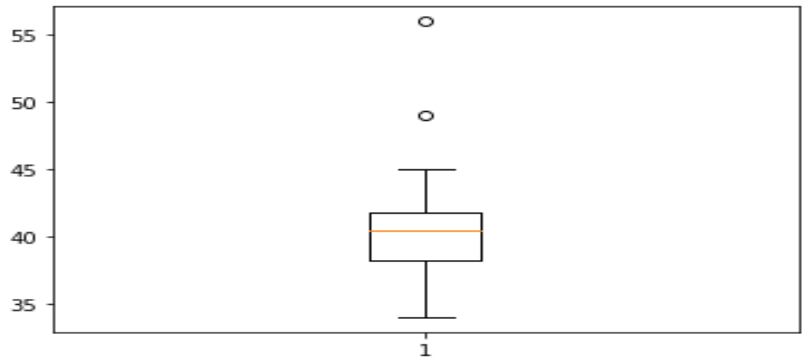
1) What can we say about the student marks?
2) Find mean, median, variance, standard deviation.

Ans:

The Required parameters are drawn by using python code (Q12.py).

1. The final results are:
   Mean = 41.0
   Median = 40.5
   Variance = 25.53
   Standard Deviation = 5.05

2. A Box plot is drawn for the above data:



From the above box plot we can that, it has two outliers 49 & 56.

Python Program:

```
import numpy as np
x=np.array([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])
import pandas as pd
y = pd.DataFrame(x)
y
# Calculatig the parameters
me=y.mean()
md=y.median()
va=y.var()
st=y.std()
print("The Required Parameters are:")
print("Mean = ",me)
print("Median = : ",md)
print("Variance = ",va.round(2))
print("Standard Deviation = ",st.round(2))
#================= Box Plot ==============================
import matplotlib.pyplot as plt
plt.boxplot(y)
```

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

Ans:

When Mean = Median, we can say that our curve has Symmetrical Shape.

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

Ans:

When Mean>Median, we can say that our curve has Positively Skewed.

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

Ans:

When Median>Mean, we can say that our curve has Negatively Skewed.

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

Ans:

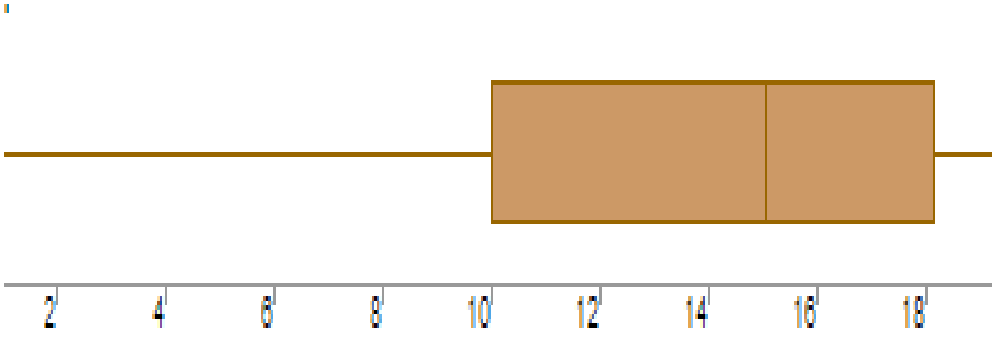
A Positive Kurtosis Value Indicates that the Distribution is Peaked and has Thick Tails.

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

Ans:

A Negative Kurtosis Value Indicates that the Distribution has Lighter Tails than the Normal Distribution.

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



What can we say about the distribution of the data?

Ans:

In the Above Box plot data is not distributed across the plot. Some outliers are influencing the data.

What is nature of skewness of the data?

Ans:

The data seems to be Negatively Skewed. Hence we say that Median > Median.

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

Ans:

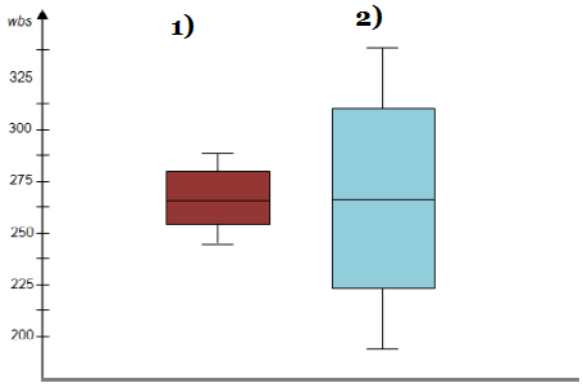
From the above plot we can say that

Q1=10

Q3=18

Inter Quartile Range Q3-Q1 = 8(Approximately)

Q19) Comment on the below Boxplot visualizations?



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

Ans:

From the above Boxplots, Box Plot1 has Less distributed data and Box Plot2 has Highly Distributed Data When compared to Box Plot1.

In Box Plot1, data may spread in between 250-280 (Approximately)

In Box Plot2, data may Spread in between 225-310 (Approximately)

Both of the Box plots will have 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:

The Required Probabilities are drawn by using python code (Q20.py).

Probability for MPG>38 is: 35.0
Probability for MPG<40 is: 73.0
Probability for 2<MPG<50 is: 96.0

Python Program:

```
import pandas as pd
df=pd.read_csv("Cars.csv")
df
m=df["MPG"].mean()
s=df["MPG"].std()
from scipy.stats import norm
nd  = norm(m,s)  # mean, sd
# p(X > 38)
p1=1 - nd.cdf(38)
# p(X < 40)
p2=nd.cdf(40)
# p(2 < X < 50)
p3=nd.cdf(50) - nd.cdf(2)
#Required Probabilities
```

```
print("Probability for MPG>38 is: ",p1.round(2)*100)
print("Probability for MPG<40 is: ",p2.round(2)*100)
print("Probability for 2<MPG<50 is: ",p3.round(2)*100)
```

Q 21) Check whether the data follows normal distribution

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

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

Ans:

(a). The inference was drawn by using the python program Q21 Cars.py

In the Given Data, MPG of Cars Follows Normal Distribution

Python Program:

```
# Normality Test for Cars
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv("Cars.csv")
df.head()
from statsmodels.stats import weightstats as ztests
zcal,pval = ztests.ztest(x1=df["MPG"],value=8,alternative='smaller')
alpha=0.05
if pval > alpha:
   print("In the Given Data, MPG of Cars Follows Normal Distribution")
else:
   print("In the Given Data, MPG of Cars Doesnot Follow Normal Distribution")
```

(b). The inference was drawn by using the python program Q21 wc-at.py

In the Given Data, Adipose Tissue (AT) Follows Normal Distribution

Python Program:

```
# Normality Test for Cars
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv("wc-at.csv")
df.head()
from statsmodels.stats import weightstats as ztests
zcal,pval = ztests.ztest(x1=df["AT"],value=8,alternative='smaller')
alpha=0.05
if pval > alpha:
   print("In the Given Data, Adipose Tissue Follows Normal Distribution")
else:
   print("In the Given Data, Adipose Tissue Doesnot Follow Normal Distribution")
```

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

Ans:

The Required Z scores were calculated by using the python program Q22.py

The Z score value for the Confidence Interval at 90% is: 1.282

The Z score value for the Confidence Interval at 94% is: 1.555

The Z score value for the Confidence Interval at 60% is: 0.253

Python Program:

```
from scipy import stats
from scipy.stats import norm
z1=stats.norm.ppf(0.90)
z2=stats.norm.ppf(0.94)
z3=stats.norm.ppf(0.60)
print("The Z score value for the Confidence Interval at 90% is: ",z1.round(3))
print("The Z score value for the Confidence Interval at 94% is: ",z2.round(3))
print("The Z score value for the Confidence Interval at 60% is: ",z3.round(3))
```

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

Ans:

The Required t scores were calculated by using the python program Q23.py

The t scores of 95% confidence interval for sample size of 25: 2.064

The t scores of 96% confidence interval for sample size of 25: 2.172

The t scores of 99% confidence interval for sample size of 25: 2.797

Python Program:

```
# Calculating t score values for 95%, 96% and 99%
from scipy import stats
from scipy.stats import norm
# df=n-1 =>24
t1=stats.t.ppf(0.975,24)
t2=stats.t.ppf(0.98,24)
t3=stats.t.ppf(0.995,24)
print("t scores of 95% confidence interval for sample size of 25 : ",t1.round(3))
print("t scores of 96% confidence interval for sample size of 25 : ",t2.round(3))
print("t scores of 99% confidence interval for sample size of 25 : ",t3.round(3))
```

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:

The Required Probability was calculated by using the python program Q24.py

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.3216.

Python Program:

```
from scipy import stats
from scipy.stats import norm
# Calculating t value
t=(260-270)/(90/18**0.5)
p=1-stats.t.cdf(abs(-0.4714),df=17)
print("The Required probability is: ",p)
```