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

Q1) Identify the Data type for the Following:

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

Number of Children	Nominal
Religious Preference	Nominal
Barometer Pressure	Interval
SAT Scores	Interval
Years of Education	Ordinal

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

Sample space = 8

P(HHT)= 3/8

NOTE: p(x) = no.of possible events/sample space

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 2and 3

sample space = 36

a) P(x =1) = 0/36 =0
b) P(x<=4) = 6/36
c) P(x=6,12) =6/36

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?

Sample space = 7c2 = 21

Possible event(none of the balls drawn blue) = 5c2 = 10.

P(x = none of the balls drawn 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

ANS)

P(no.of candies for randomly selected child) = (1*0.015)+(4*0.20)+(3*0.65)+(5*0.005)+(6*0.01)+(2*0.120) =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

```
mean for points,score,weight = 3.59,3.217,17.848
median for points,score, weigh = 3.695, 3.325, 17.71
mode for points = 3.07,3.92
mode for score  = 3.44
mode for weigh  = 17.02,18.90
variance for points, score, weigh = 0.285, 0.957, 3.193
std for points, score, weigh = 0.5346, 0.978, 1.786
Range for points, score, weigh = 2.17, 3.911, 8.4
```

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
df = pd.read_csv("/content/Q7.csv")
df.head(2)
```

```
df.describe()
```

```
df['Points'].mode()
df['Score'].mode()
df.Weigh.mode()
df.Points.var()
df.Score.var()
df.Weigh.var()
```

```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7f1cc34d0880>
```

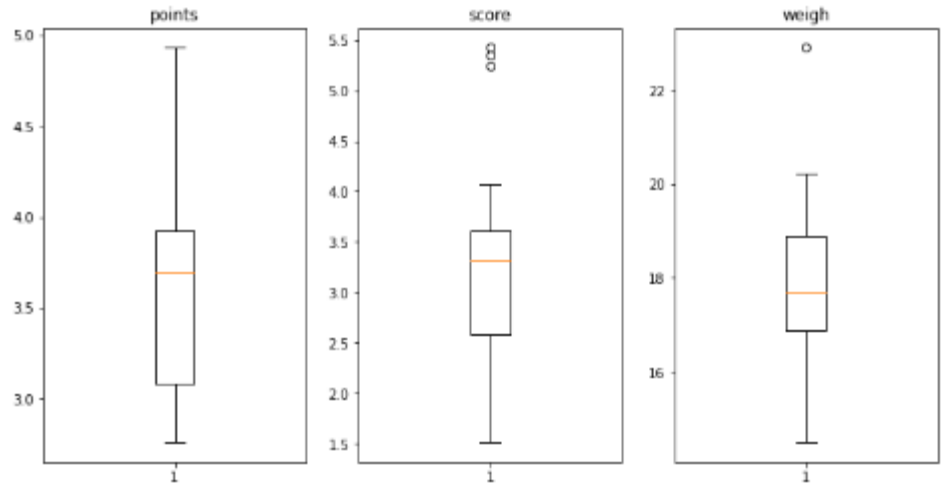
Inferences:

```
inferences based on pairplots :
if points are increasing scores are decreasing which means points and score are negati
vely correlated
```

inferences based on boxplots:
Points:
 1.there are no outliers in this variable
 2.right/postive skewed distribution

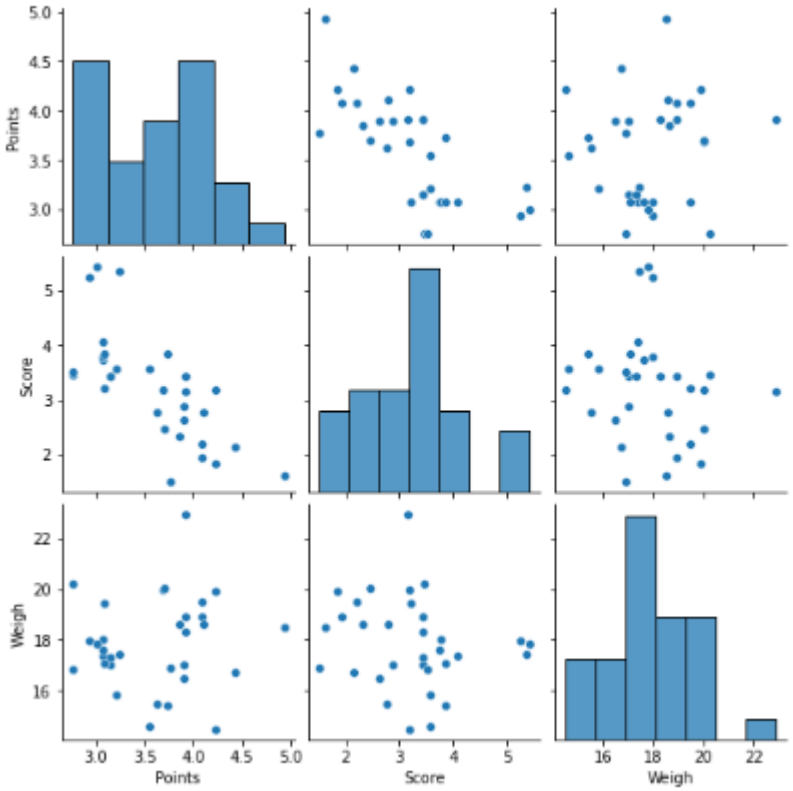
Score:
  1.found three outliers
Weigh:
  1.found one outlier

```
plt.subplot(1,3,1)
plt.boxplot(df['Points'])
plt.title('points')
plt.subplot(1,3,2)
plt.boxplot(df['Score'])
plt.title("score")
plt.subplot(1,3,3)
plt.boxplot(df['Weigh'])
plt.title("weigh")
plt.show()
```



```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7f1cc34d0880>
```



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)

P(x=weight of the randomly chosen patient) = 1/9(108+110+123+134+135+145+167+187+100) = 145.33

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

Cars speed and distance

Use Q9_a.csv

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
[ ] df1 = pd.read_csv("/content/Q9_a.csv")
    df2 = pd.read_csv("/content/Q9_b.csv")
```

```
[ ] cars = df1
```

```
[ ] cars.skew()
```

```
Index    0.000000
speed   -0.117510
dist     0.806895
dtype: float64
```

```
[ ] from scipy.stats import skew, kurtosis
```

```
[ ] cars_dist_skewness = skew(cars['dist'])
    cars_dist_skewness
```

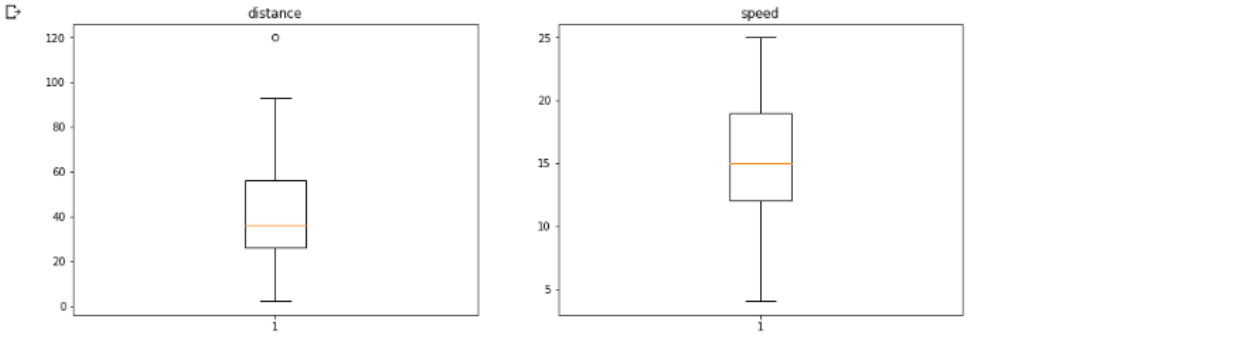
```
0.7824835173114966
```

```
[ ] cars_speed_skewness = skew(cars['speed'])
    cars_speed_skewness
```

```
-0.11395477012828319
```

```
[ ] fig,ax = plt.subplots(figsize=(15,5))
    plt.subplot(1,2,1)
    plt.boxplot(cars['dist'])
    plt.title('distance')
    plt.subplot(1,2,2)
```

```
fig,ax = plt.subplots(figsize=(15,5))
plt.subplot(1,2,1)
plt.boxplot(cars['dist'])
plt.title('distance')
plt.subplot(1,2,2)
plt.boxplot(cars['speed'])
plt.title('speed')
plt.show()
```



nferences:
:ars distance: found one outlier in distance and its a right skewed.
:ars speed: no outliers and slightly left skewed.

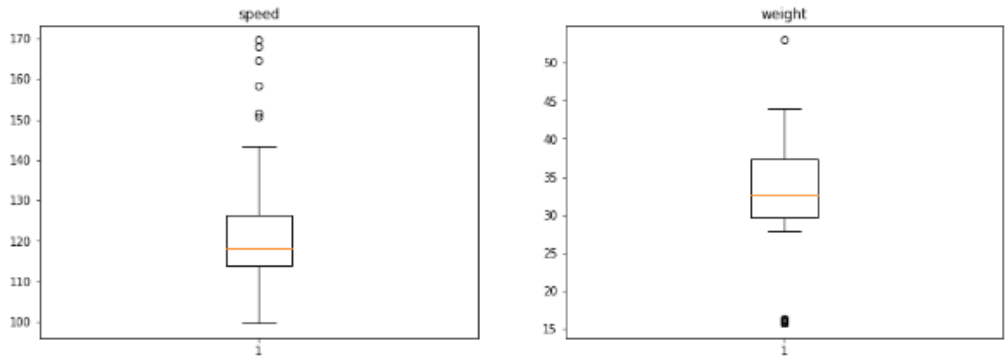
SP and Weight(WT)

Use Q9_b.csv

```
df2.skew()
```

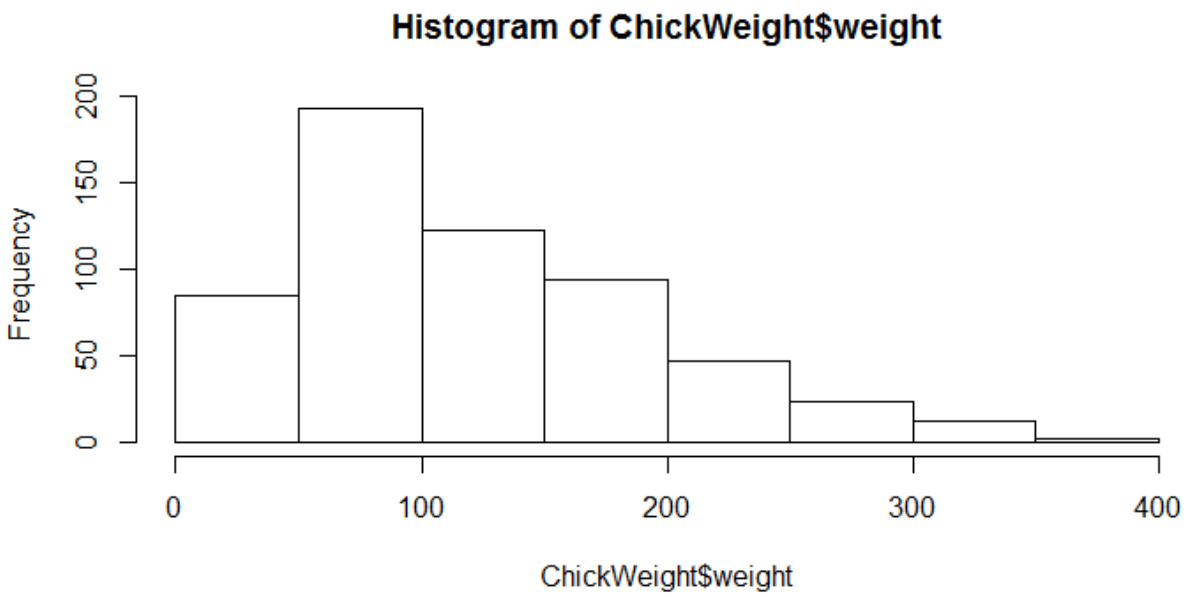
```
Unnamed: 0    0.000000
SP            1.611450
WT           -0.614753
dtype: float64
```

```
fig,ax = plt.subplots(figsize=(15,5))
plt.subplot(1,2,1)
plt.boxplot(df2['SP'])
plt.title('speed')
plt.subplot(1,2,2)
plt.boxplot(df2['WT'])
plt.title('weight')
plt.show()
```

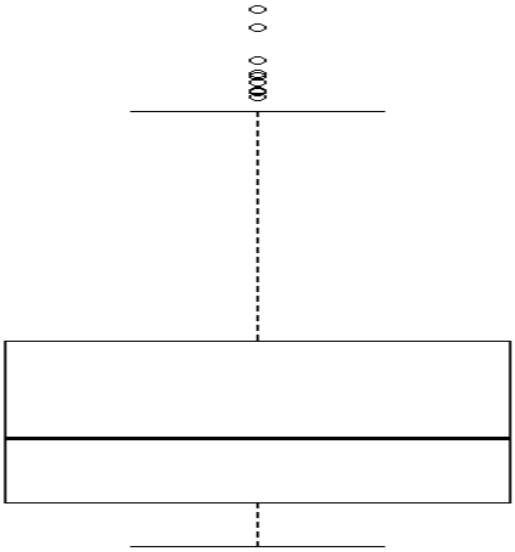


cars speed and weight inferences:
speed:found outliers and right skewed
weight: found outliers and left skewed

Q10) Draw inferences about the following boxplot & histogram



Ans) Histogram inferences: right skewed and most of the points lies between 50 to 150.



Ans)

Boxplot inferences: outliers observered on maximum side and 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)sample size(n) = 2000 **population size(N) = 3000000**

Sample avg(x) = 200pounds

Sample std(s) = 30

Formula for confidence interval:

Stats.t.interval(1-alpha,df=n-1,loc=samplemean,scale=samplestd/sqrt(n))

```
[6] #94% C.I
    stats.t.interval(0.94,df=1999,loc=200,scale=30/(np.sqrt(2000)))

    (198.7376089443071, 201.2623910556929)

[7] #98% C.I
    stats.t.interval(0.98,df=1999,loc=200,scale=30/(np.sqrt(2000)))

    (198.4381860483216, 201.5618139516784)

[8] #96% C.I
    stats.t.interval(0.96,df=1999,loc=200,scale=30/(np.sqrt(2000)))

    (198.6214037429732, 201.3785962570268)
```

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: Mean =41, Median =40.5, Variance =25.52 and Standard Deviation =5.05

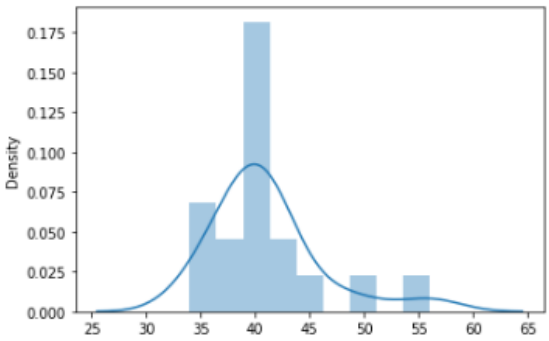
```
[18] data = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]
```

```
[20] students = pd.DataFrame(data)
     students.info()
```

```
students.describe()
```

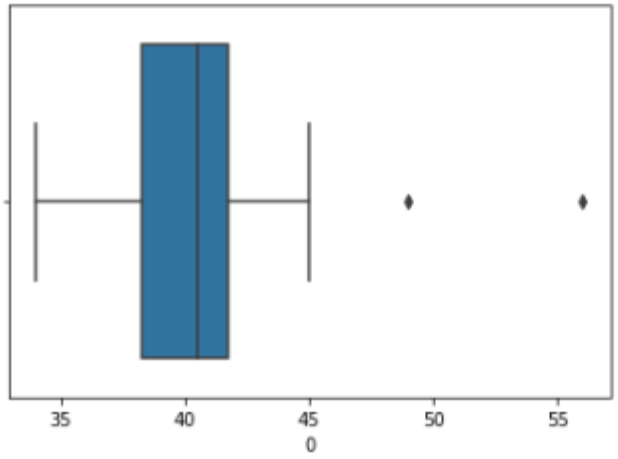
```
sns.distplot(students)
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecat
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7ff11bb288b0>
```




```
sns.boxplot(students[0])
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: Future
  warnings.warn(
<matplotlib.axes._subplots.AxesSubplot at 0x7ff117f6d760>
```



we can we say about the student marks is right skewed and found two outliers.

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) right skewed/postive skewness, tail towards right

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

Ans) left skewed/negative skewness, tail towards left

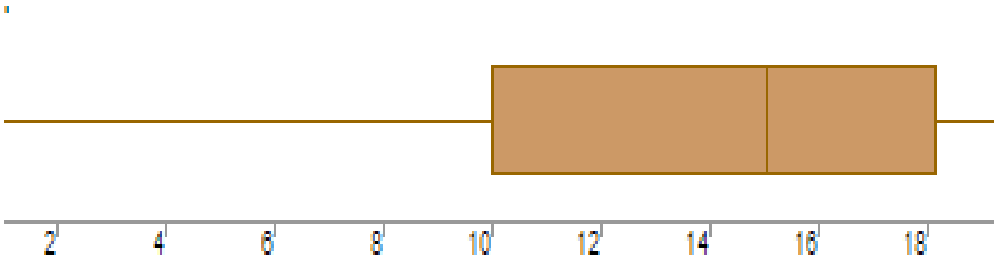
Q16) What does positive kurtosis value indicates for adata ?

Ans) postive kurtosis means high peaked value and variance will be low

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

Ans) negative kurtosis means low peaked value and vairance will be low and the curve will look like flat.

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



What can we say about the distribution of the data?

Ans) it is not symmetric and mean is not equal to median also so we can say distribution is not normally distributed.

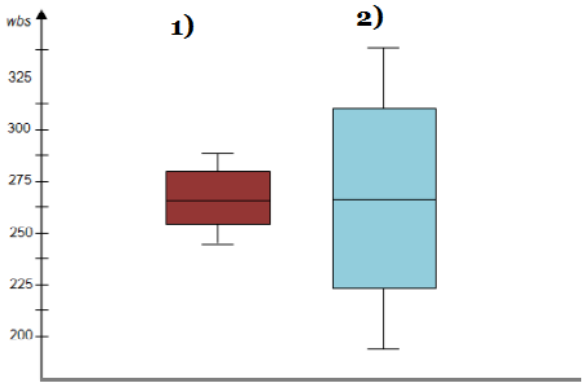
What is nature of skewness of the data?

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.

Inferences:

1. The range of boxplot 1 is small with respect to box plot 2
2. 50% quartile(median) is almost same for both box plots
3. No skewness observed in both plots

Q 20) Calculate probability from the given dataset for the below cases

Data _set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

a. P(MPG>38)
b. P(MPG<40)
c. P (20<MPG<50)

```
[8] cars = pd.read_csv("/content/Cars.csv")
    cars.head(4)
```

```
[9] cars.MPG.mean()

    34.42207572802469
```

```
[10] cars.MPG.std()

    9.131444731795982
```

```
[11] # computing p(x>38)
    p = 1-stats.norm.cdf(38,loc=34.422,scale=9.13)
    p

    0.34756795338871926
```

```
[12] #computing p(x<40)
    stats.norm.cdf(40,loc=34.422,scale=9.13)

    0.7293846197612225
```

```
#computing p(20<X<50)
#p(20<mpg<50) = p(50)-p(20)
stats.norm.cdf(50,loc=34.422,scale=9.13) - stats.norm.cdf(20,loc=34.422,scale=9.13)

0.8989225042585001
```

a. P(MPG>38) = 0.3475
b. P(MPG<40) = 0.729
c. P (20<MPG<50) = 0.898

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

```
from scipy import stats
import numpy as np
import pandas as pd
```

```
cars = pd.read_csv("/content/Cars.csv")
cars.head()
```

	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

```
stats.shapiro(cars['MPG'])
```

```
ShapiroResult(statistic=0.9779686331748962, pvalue=0.17639249563217163)
```

Pvalue > 0.05 which means reject null hypothesis.

Alternative hypothesis is cars['MPG'] 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

```
from scipy import stats
import numpy as np
import pandas as pd
```

```
data = pd.read_csv("/content/wc-at.csv")
data.head()
```

```
stats.shapiro(data['Waist'])
```

```
ShapiroResult(statistic=0.9558576345443726, pvalue=0.0011704121716320515)
```

```
stats.shapiro(data['AT'])
```

```
ShapiroResult(statistic=0.9523370862007141, pvalue=0.0006539996829815209)
```

Both Waist and AT p values are less than 0.05
So fail to reject null hypothesis
Null hypothesis = Waist and AT are not follwing normal distribution.

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

```
from scipy import stats
import numpy as np
```

```
#Z scoresof  90% confidence interval
#(1+confidenceInterval)/2 = z statValue
stats.norm.ppf(0.95)
```

```
1.6448536269514722
```

```
#Z scoresof  94% confidence interval
stats.norm.ppf(0.97)
```

```
1.8807936081512509
```

```
#Z scoresof  60% confidence interval
stats.norm.ppf(0.8)
```

```
0.8416212335729143
```

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

```
# 95% confidence interval t score
stats.t.ppf(0.95,df=24)
```

```
1.7108820799094275
```

```
# 96% confidence interval t score
stats.t.ppf(0.96,df=24)
```

```
1.8280511719596342
```

```
# 99% confidence interval t score
stats.t.ppf(0.99,df=24)
```

```
2.4921594731575762
```

Q 24)A Government companyclaims 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) sample size (n) = 18

avg = 260

std = 90

computing critical value(t) why because we don't know population std.

formula for computing t value:

t =(samplemean – population mean)/(sample std/sqrt(n))

computing p values for critical values, as below

from scipy import stats

p_value = stats.t.cdf(t_value, df = n-1)

p_value = 0.32 =32%

```
import pandas as pd
import numpy as np
from scipy import stats
```

```
[5] #computing t value
    t = (260-270)/(90/np.sqrt(18))
    t
```

```
-0.4714045207910317
```

```
#computing p value
p = stats.t.cdf(-0.4714, 17)
p
```

```
0.32167411684460556
```