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	Categorical
Number of kids	Discrete
Number of tickets in Indian railways	Discrete
Number of times married	Discrete
Gender (Male or Female)	Categorical

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

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

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

ANS:

Total number of Possibilities=2*2*2=8

Probability of two heads= No. Of Possibilities / Total No.Of Possibilities

= (HHT,HTH,THH)/8

=3/8=0.375

Probability of one tail = No. Of Possibilities / Total No.Of Possibilities

=(HTH,THH,HHT) / 8

=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:

Total No.Of Possibilities=6*6=36

a) There is no values is equal to one

b) Probability of (<or =4) = No. Of Possibilities /Total No.Of Possibilities

= 6/36=1/6 = 0.16

c) divisible by 2 = No. Of Possibilities / Total No.Of Possibilities
=18/36=9/18=1/2
= 0.50
divisible by 3 = No. Of Possibilities / Total No.Of Possibilities
=12/36=4/12=1/3
=0.33

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=2+3+2=7

Let S be the sample

n(S)=Number of ways of drawing 2 balls out of 7

=7C2=7*6/2*1=42/2=21

Let E be the ways of drawing 2 balls,none of which is blue

n(E)=Number of ways of drawing 2 balls out of (2+3)balls

= 5C2=5*4/2*1=20/2=10

Therefore,

P(E)=n(E)/n(S)=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:

Expected number of candies for a randomly selected child

=1*0.015+4*0.20+3*0.65+5*0.005+6*0.01+2*0.120

=0.015+0.8+1.95+0.025+0.06+0.24=3.09

Expected number of candies for a randomly selected child =3.09

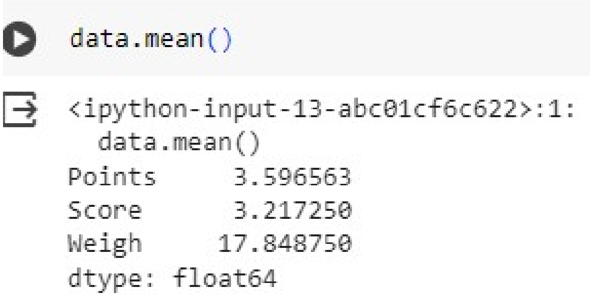
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

- 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:

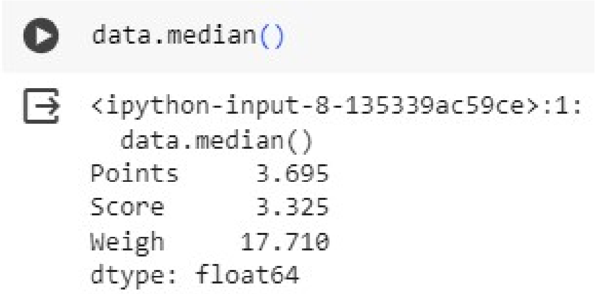
<u>MEAN:</u>



```
data.mean()
```

```
<ipython-input-13-abc01cf6c622>:1:
  data.mean()
Points     3.596563
Score      3.217250
Weigh     17.848750
dtype: float64
```

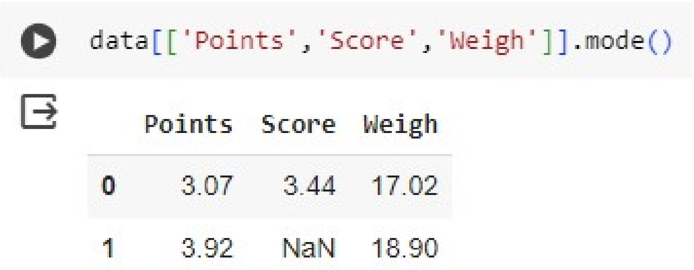
<u>MEDIAN:</u>



```
data.median()
```

```
<ipython-input-8-135339ac59ce>:1:
  data.median()
Points     3.695
Score      3.325
Weigh     17.710
dtype: float64
```

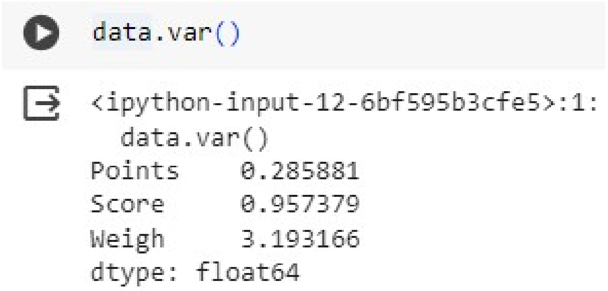
MODE:



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

	Points	Score	Weigh
0	3.07	3.44	17.02
1	3.92	NaN	18.90

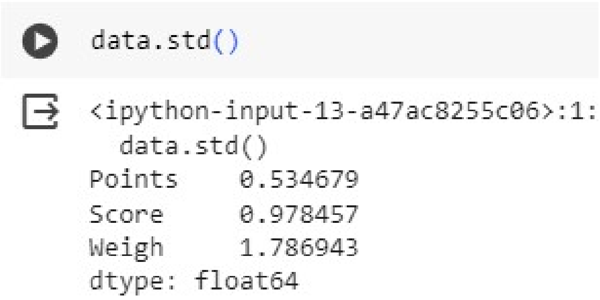
VARIANCE:



```
data.var()
```

```
<ipython-input-12-6bf595b3cfe5>:1:
  data.var()
Points    0.285881
Score     0.957379
Weigh     3.193166
dtype: float64
```

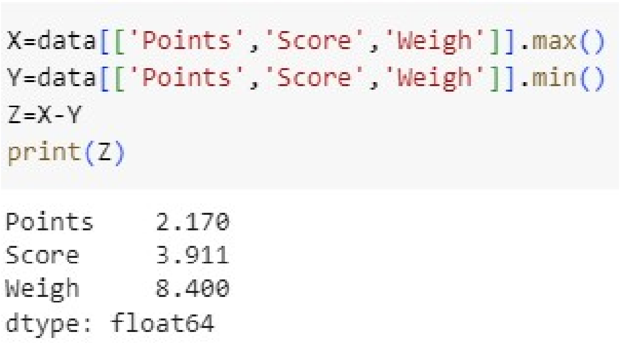
STANDARD DEVIATION:



```
data.std()
```

```
<ipython-input-13-a47ac8255c06>:1:
  data.std()
Points    0.534679
Score     0.978457
Weigh     1.786943
dtype: float64
```

RANGE:



```
X=data[['Points','Score','Weigh']].max()
Y=data[['Points','Score','Weigh']].min()
Z=X-Y
print(Z)
```

```
Points    2.170
Score     3.911
Weigh     8.400
dtype: float64
```

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=sum of all the values/Total no.of values

=108+110+123+134+135+145+167+187+199/9

=1308/9

=145.333

Expected Value of the Weight of that patient is 145.33

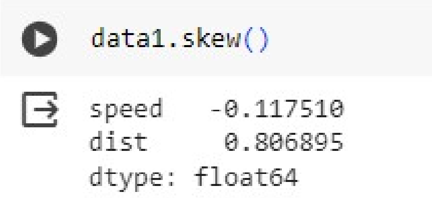
Q9) Calculate Skewness, Kurtosis & draw inferences on the following data

Cars speed and distance

Use Q9_a.csv

ANS:

<u>Skewness</u>



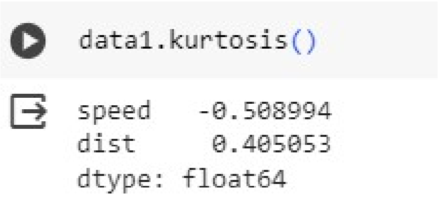
```
data1.skew()
```

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

* Speed is Negative skewness means left side,that is data distributed in right side

* Distance is Positive skewness means right side,that is data distributed in left side

Kurtosis



```
data1.kurtosis()
```

```
speed   -0.508994
dist     0.405053
dtype: float64
```

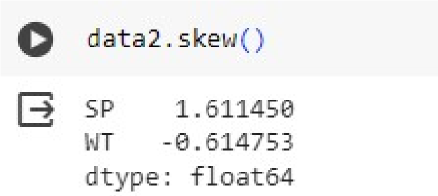
* Speed is negative ,so data is wide not in peak

* Distance is positive ,so data is high in peak

SP and Weight(WT)

Use Q9_b.csv

Skewness



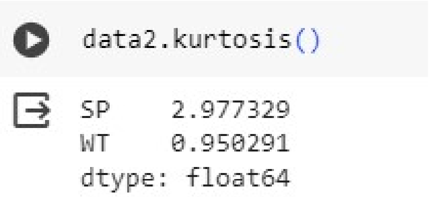
```
data2.skew()
```

```
SP    1.611450
WT   -0.614753
dtype: float64
```

* sp is positive skewness,so data distributed in right side

* weight is negative skewness,so data distributed in left side

Kurtosis



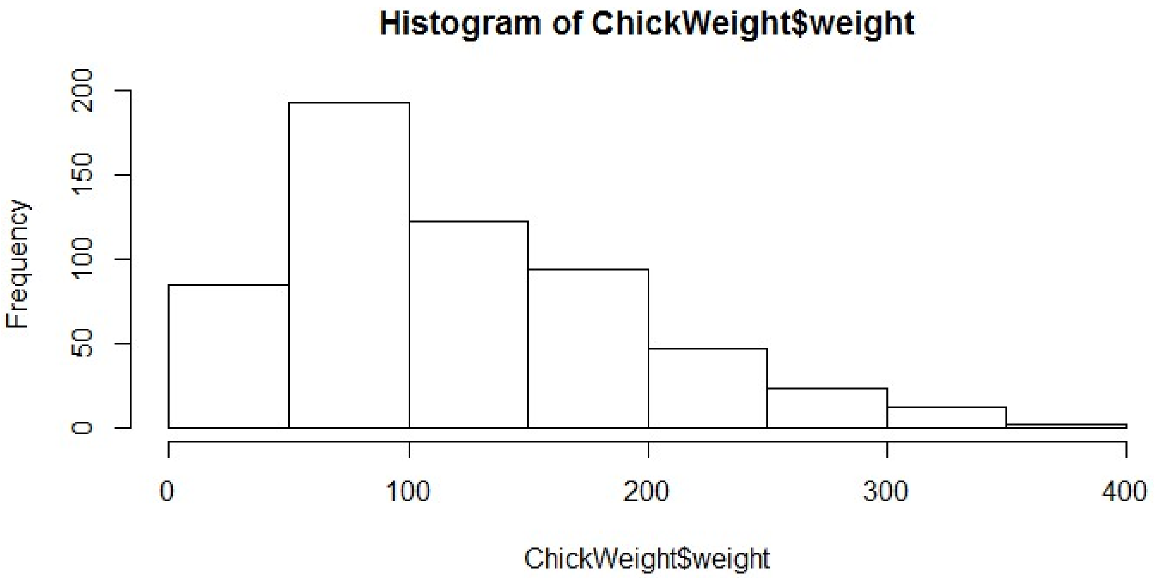
```
data2.kurtosis()
```

```
SP    2.977329
WT    0.950291
dtype: float64
```

* SP is positive ,so data is high in peak

* weight is also positive,so data is high in peak

Q10) Draw inferences about the following boxplot & histogram



ANS:

* 50 to 100 weight having more frequency 180

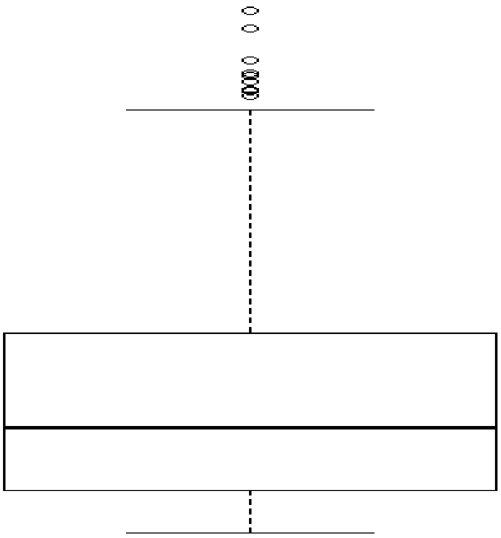
* 350 to 400 weight having very less frequency 5

* In this histogram have positive skewness,so data is right skewed

* 0 to 50 weight having 80 frequency

* 100 to 150 weight having 120 frequency

* Data is not a normal distribution



ANS:

* 7 outliers are present in above box plot

* Positive skewness,so data is right skewed

* Data is not normally distributed

* Q1 is smaller than the Q3

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:

To calculate confidence interval,

sample mean($\overline{X}$)=200

standard deviation(S)=30

sample size(n)=2000

Confidence Interval=$\overline{X}$± $Z_{(1-\alpha)}$ [σ /√ n]

For 94%,96%,98% confidence Interval, the Z-score approximately

Z=1.88,2.05,2.33

Let calculate the confidence Interval,

For 94% confidence Interval

Confidence Interval= 200 ± 1.22[30/√2000]

= 200 ± 1.22*0.6738

= 200±0.822

= (200.822,199.178)

For 96% Confidence Interval

Confidence Interval= 200± 2.05[30/√ 2000

= 200± 2.05*0.6738

= 200± 1.381

= (201.381,198.619)

For 98% Confidence Interval

Confidence Interval= 200±2.33[30/√2000]

= 200± 2.33*0.6738

= 200± 1.569

= (201.56,198.43)

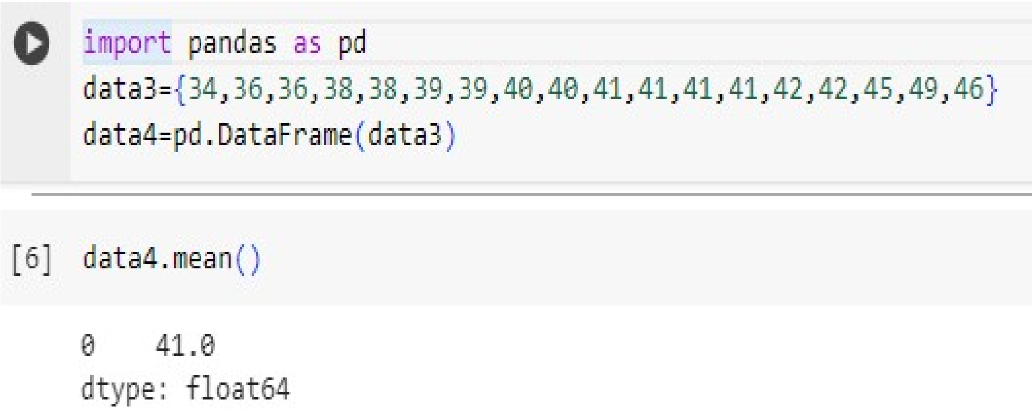
Q12) Below are the scores obtained by a student in tests

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:

<u>MEAN:</u>

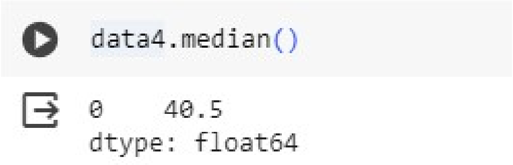


```
import pandas as pd
data3={34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,46}
data4=pd.DataFrame(data3)
```

```
[6] data4.mean()
```

```
0    41.0
dtype: float64
```

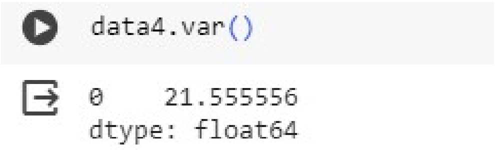
<u>MEDIAN:</u>



```
data4.median()
```

```
0    40.5
dtype: float64
```

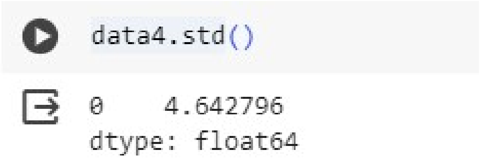
<u>VARIANCE:</u>



```
data4.var()
```

```
0    21.555556
dtype: float64
```

<u>STANDARD DEVIATION:</u>



```
data4.std()
```

```
0    4.642796
dtype: float64
```

MEAN=41.0

MEDIAN=40.5
VARIANCE=21.55
STANDARD DEVIATION=4.642

2)What can we say about the student marks?
ANS:

* Average of student marks is 41
* The students marks range from 34 to 56
* Most of the students score between 35 to 42
* Mode of the student mark is 41

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

ANS:

When the values of the mean,median of data are equal,there is no skewness,so data is normally distributed.

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

ANS:

If the mean is greater than median ,the distribution is postive skewness

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

ANS:

If the mean is lesser than median ,the distribution is negative skewness

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

ANS:

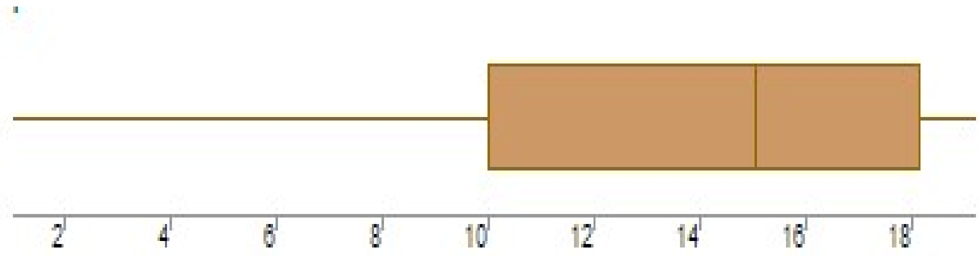
A distribution with a positive kurtosis value indicated that the distribution has heavier tails than normal distribution

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

ANS:

A distribution with a negative kurtosis value indicated that the distribution has flatter than normal distribution

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



What can we say about the distribution of the data?

ANS:

*In this boxplot have no outliers

*Median between 15 to 16

*Data present in the range from 10 to 18

*This boxplot not following Normal Distribution

*The nature of the skewness is left skewness of the data

What is nature of skewness of the data?

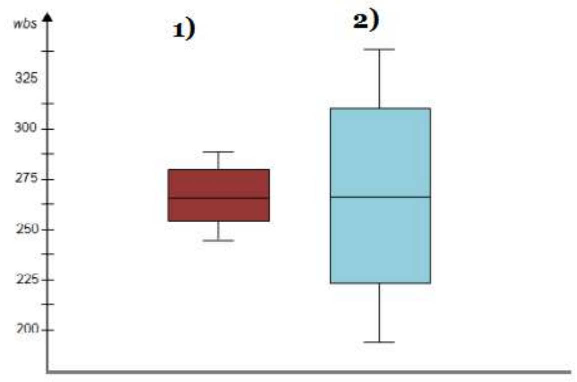
ANS: Left Skewness

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

ANS:

IQR=Q3-Q1=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:

***** Both the plots are normally distributed

***** In this box plot there is no outliers

* We can say box plot 1 is sample and box plot 2 is population

* Q1 is 25%,Q3 is 75%. IQR is 50% on both plots. So we can say both the distributions following normal distribution, that is mean,median, mode are equal

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

Data _set: Cars.csv

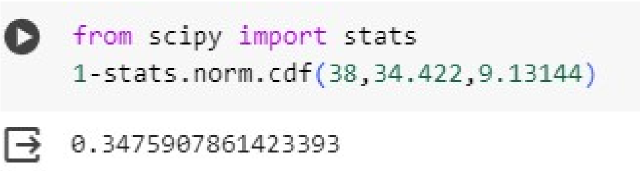
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

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

ANS:

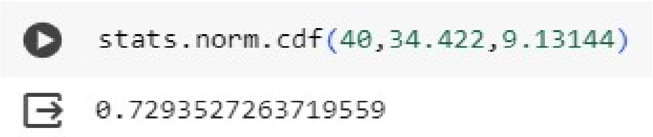
a) P(MPG>38)



```
from scipy import stats
1-stats.norm.cdf(38,34.422,9.13144)
```

```
0.3475907861423393
```

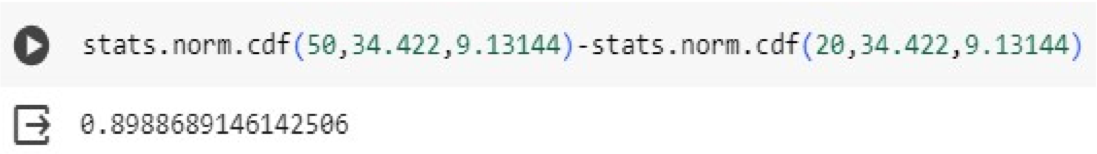
b) P(MPG<40)



```
stats.norm.cdf(40,34.422,9.13144)
```

```
0.7293527263719559
```

c) P(20<MPG<50)



```
stats.norm.cdf(50,34.422,9.13144)-stats.norm.cdf(20,34.422,9.13144)
```

```
0.8988689146142506
```

P(MPG<38)= 0.735

P(MPG>40)= 0.729

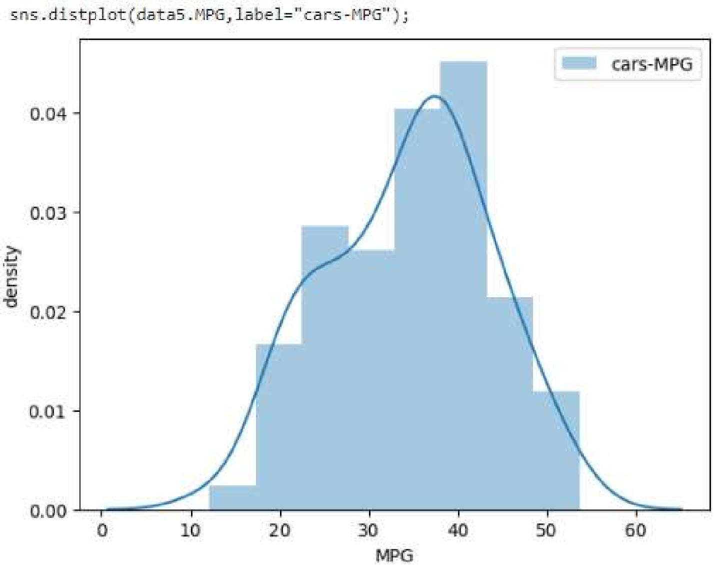
P(20<MPG<50)= 0.898

21) Check whether the data follows normal distribution

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

ANS:

(MPG OF CARS)



```
sns.distplot(data5.MPG,label="cars-MPG");
```

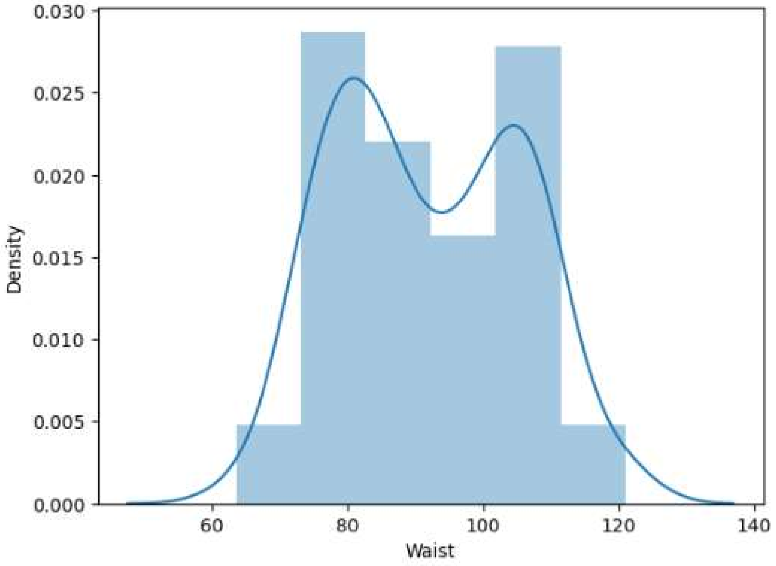
MPG of cars following the normal distribution

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

Dataset: wc-at.csv

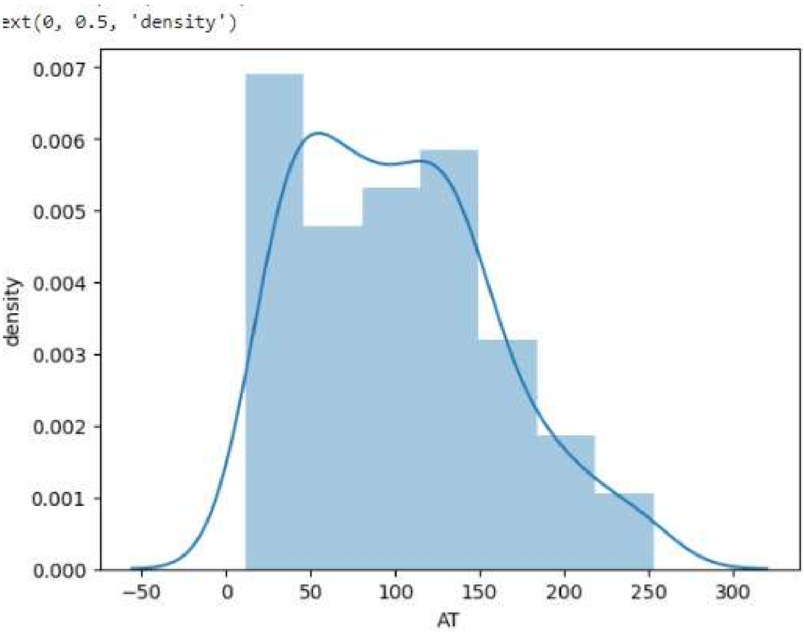
ANS:

(Waist Circumstance)



Waist circumstance does not follow normal distribution

(Adipose Tissue)



```
ext(0, 0.5, 'density')
```

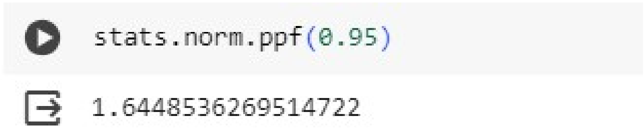
Adipose tissue does not follow normal distribution

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

ANS:

For 90% confidence interval(Two tail)

The critical z-value for a 90% confidence interval is approximately
Z=1.645



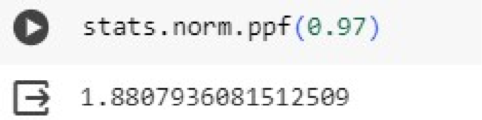
```
stats.norm.ppf(0.95)
```

```
1.6448536269514722
```

For 94% confidence interval(Two tail)

The critical Z-value for 94% confidence interval is approximately

Z=1.88



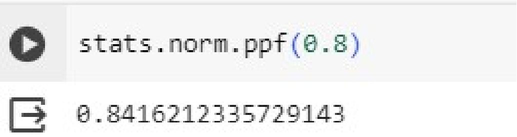
```
stats.norm.ppf(0.97)
```

```
1.8807936081512509
```

For 60% confidence interval(Two tail)

The critical Z-value for 60% confidence interval is approximately

Z=0.84



```
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

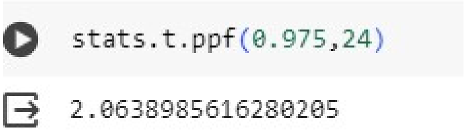
ANS:

The degrees of freedom (df) sample of size 25 is n-1=24

95% confidence interval

For 95% confidence interval with df=24,the critical t-value approximately

t=2.064



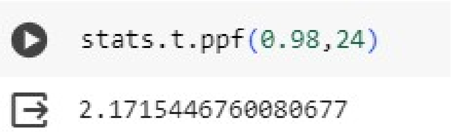
```
stats.t.ppf(0.975,24)
```

```
2.0638985616280205
```

96% confidence interval

For 96% confidence interval with df=24,the critical t-value approximately

t=2.171



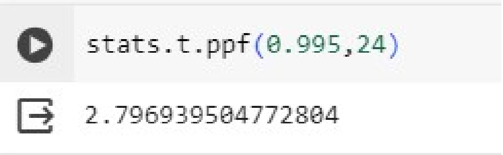
```
stats.t.ppf(0.98,24)
```

```
2.1715446760080677
```

99% confidence interval

For 99% confidence interval with df=24,the critical t-value approximately

t=2.797



```
stats.t.ppf(0.995,24)
```

```
2.796939504772804
```

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:

Sample mean=260

Population mean=270

Standard deviation=90

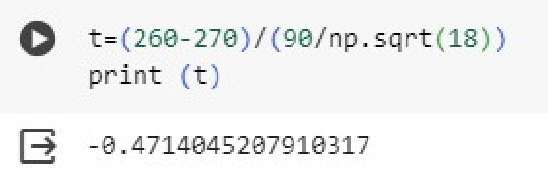
Sample size=18

T-score=$\bar{x}$- μ /[s/√n]

=260-270/[90/√18]

=-10/[90/4.24]=-0.4712

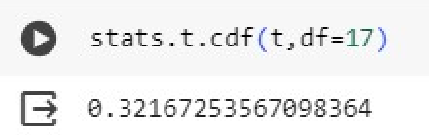
T-score value



```
t=(260-270)/(90/np.sqrt(18))
print (t)
```

```
-0.4714045207910317
```

After put degree of freedom



```
stats.t.cdf(t,df=17)
```

```
0.32167253567098364
```