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)	Interval
Sales Figures	Ratio
Blood Group	Nominal
Time Of Day	Ordinal
Time on a Clock with Hands	Interval
Number of Children	Ratio
Religious Preference	Nominal
Barometer Pressure	Ratio
SAT Scores	Interval
Years of Education	Interval

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

ANS: Total conditions with 3 coins= 8

Condition where 2 heads and 1 tail occur=3

PROBABILITY= 3/8

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 outcome=36

A) sum =1: probability =0

B) <=4: outcome=6 (11,12,13,21,22,31); probability= 6/36= 1/6

C) **Sum divisible by 2&3: outcome=5 (15,24,33,42,51); probability=5/36**

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?

ANS: Total number of balls=7
Number of ways of drawing 2 balls out of 7= 7C2 = 21
Number of ways of drawing 2 balls out of 5= 5C2 = 10
Required Probability= 21/10

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- Generalised 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

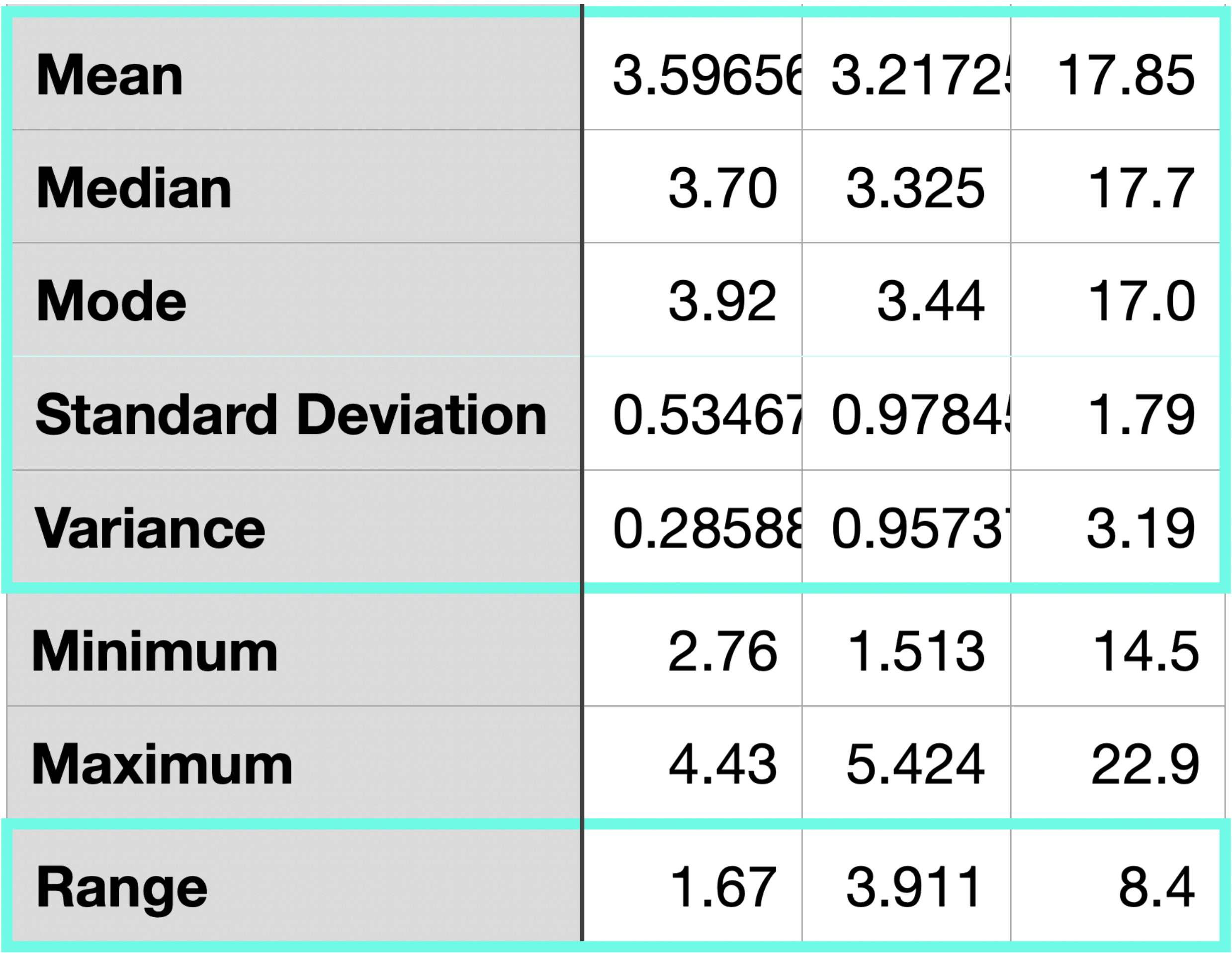
= 1*0.015+4*0.2+3*0.65+5*0,005+6*0.01+2*0.12

= 3.09

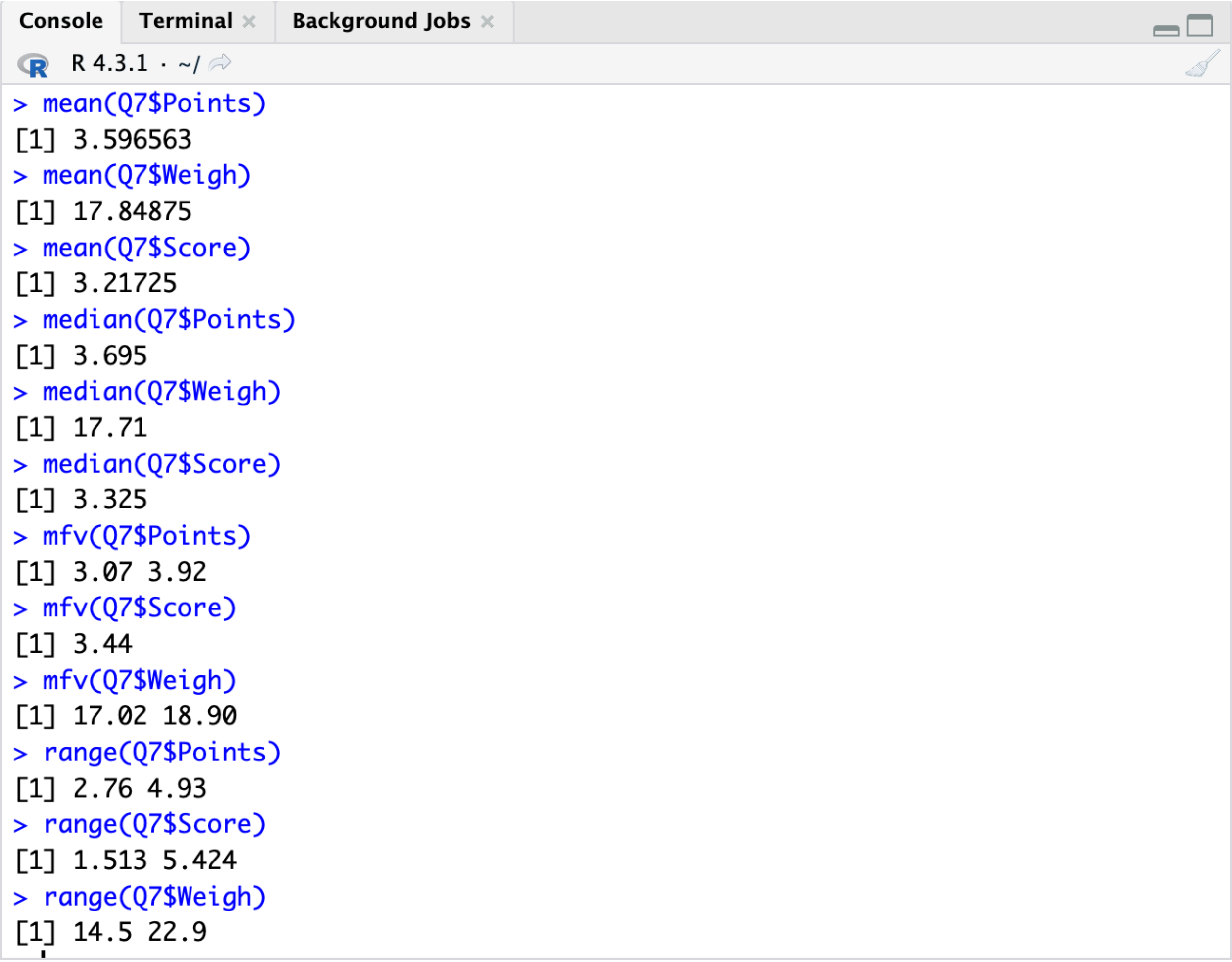
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

- 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



Mean	3.59656	3.21725	17.85
Median	3.70	3.325	17.7
Mode	3.92	3.44	17.0
Standard Deviation	0.53467	0.97845	1.79
Variance	0.28588	0.95737	3.19
Minimum	2.76	1.513	14.5
Maximum	4.43	5.424	22.9
Range	1.67	3.911	8.4



```
> mean(Q7$Points)
[1] 3.596563
> mean(Q7$Weigh)
[1] 17.84875
> mean(Q7$Score)
[1] 3.21725
> median(Q7$Points)
[1] 3.695
> median(Q7$Weigh)
[1] 17.71
> median(Q7$Score)
[1] 3.325
> mfv(Q7$Points)
[1] 3.07 3.92
> mfv(Q7$Score)
[1] 3.44
> mfv(Q7$Weigh)
[1] 17.02 18.90
> range(Q7$Points)
[1] 2.76 4.93
> range(Q7$Score)
[1] 1.513 5.424
> range(Q7$Weigh)
[1] 14.5 22.9
```

ANS: The mode is actually calculated using the formulas in the excel, but there is no mode for the table values.

Likewise the mean, median, mode of the dataset revolves around similar values as seen in the above table.

The standard deviation and variance is also appropriately calculated using the predefined formulas.

The range calculation is done by first finding the maximum and minimum values and then subtracting them.

The same can be done using RStudio, the needed commands are executed and the results are attached above as screenshots.

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: So, the expected weight is:

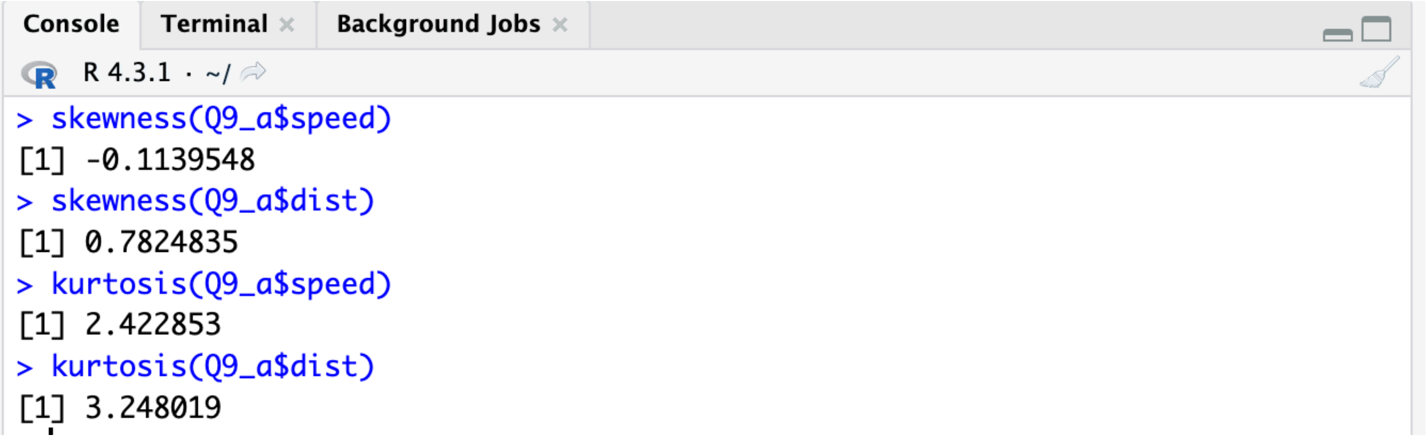
E(X) = (108 * 1/9) + (110 * 1/9) + (123 * 1/9) + (134 * 1/9) + (135 * 1/9) + (145 * 1/9) + (167 * 1/9) + (187 * 1/9) + (199 * 1/9) = 145.33

Therefore, the expected weight of a patient chosen at random is 145.33 pounds.

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

Cars speed and distance

Use Q9_a.csv



```
> skewness(Q9_a$speed)
[1] -0.1139548
> skewness(Q9_a$dist)
[1] 0.7824835
> kurtosis(Q9_a$speed)
[1] 2.422853
> kurtosis(Q9_a$dist)
[1] 3.248019
```

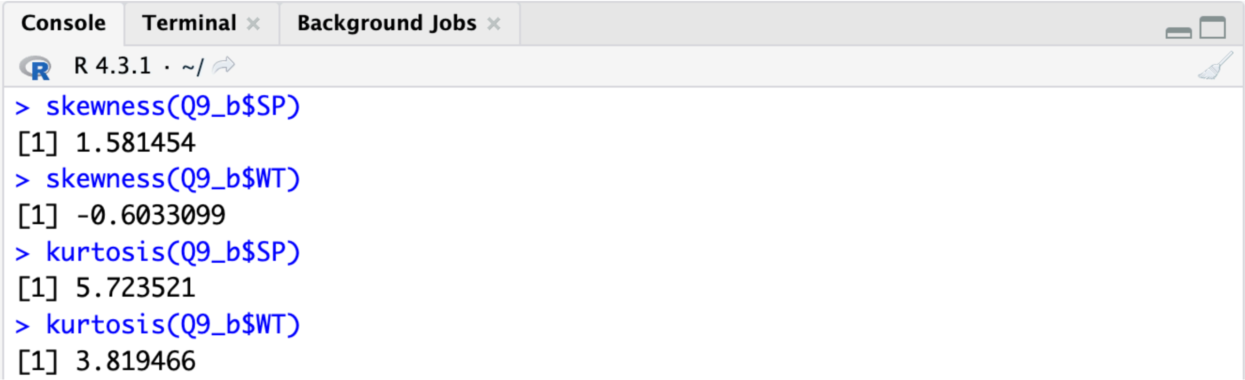
ANS: The skewness of speed is -0.1139548 which indicates a slightly negatively skewed dataset wherein most data points are located right side and has tail pointing towards left.

The skewness of Dist on the other hand is 0.7824835 which indicates positively skewed dataset wherein most data points are located on the left side and has a tail towards the right.

The values of kurtosis for both datas is greater than 2 which indicates a significantly peaked dataset.

SP and Weight(WT)

Use Q9_b.csv



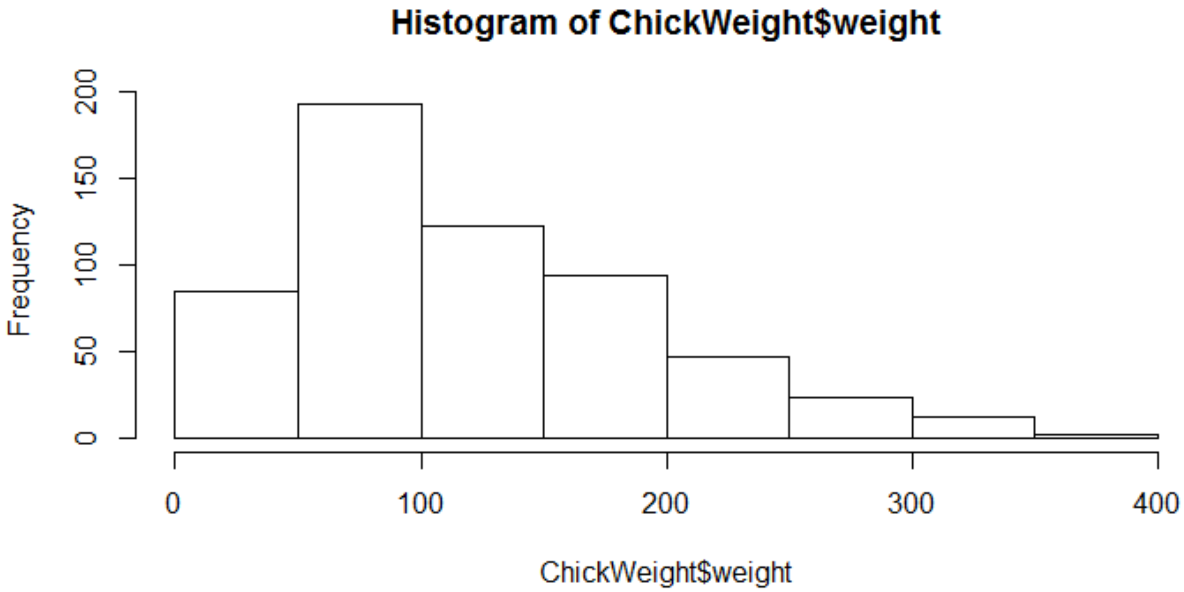
```
> skewness(Q9_b$SP)
[1] 1.581454
> skewness(Q9_b$WT)
[1] -0.6033099
> kurtosis(Q9_b$SP)
[1] 5.723521
> kurtosis(Q9_b$WT)
[1] 3.819466
```

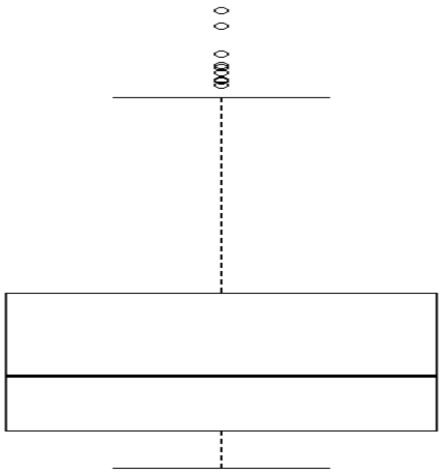
ANS: The value of skewness in SP is 1.581454, this indicates a significant positively skewed dataset, wherein most data points are located on the left side and has a tail towards the right.

The skewness value of WT is -0.6033099, this indicates a slight negatively skewed dataset, wherein most data points are located right side and has a small tail towards left.

The values of kurtosis for both the datasets are more than 2 that indicates the data is highly peaked.

Q10) Draw inferences about the following boxplot & histogram



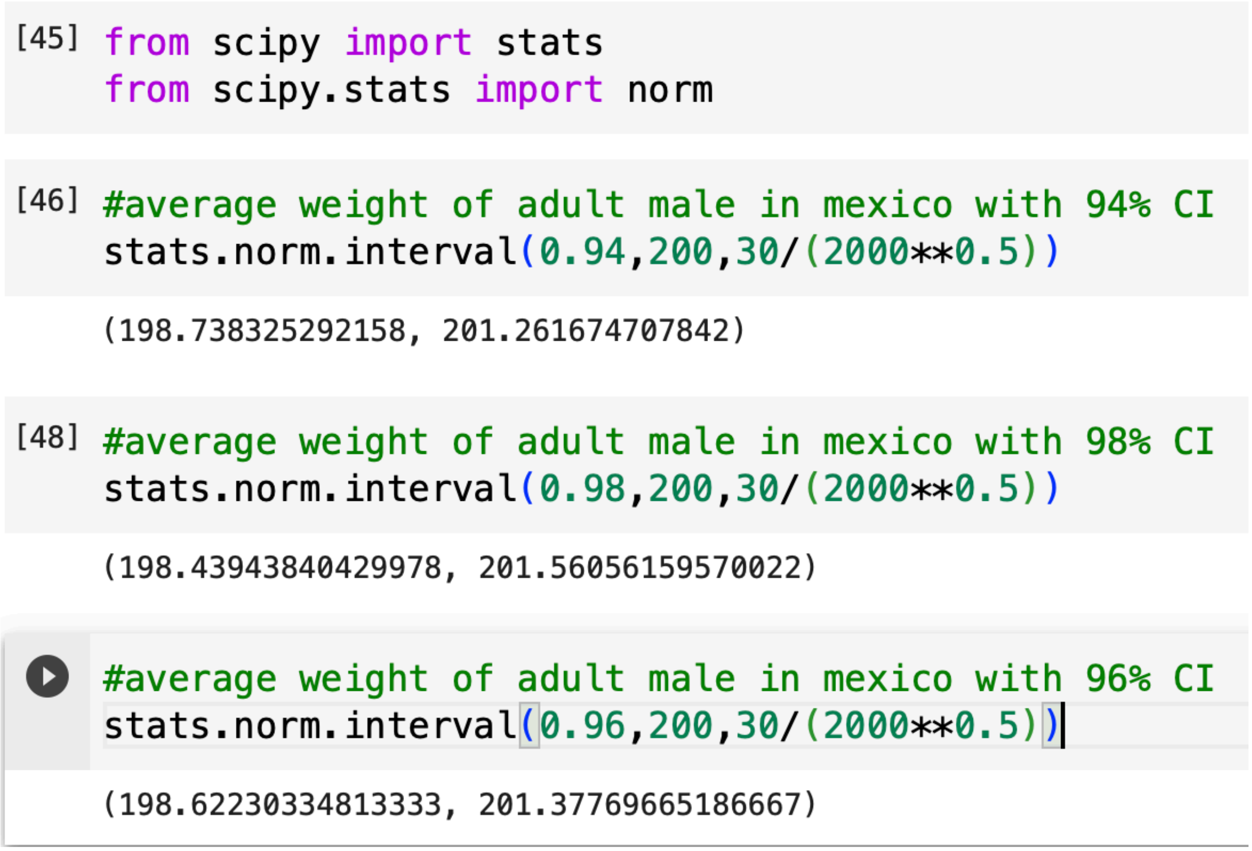


ANS: From the histogram we can conclude that the data is right skewed, that means majority of the data points lie on the left side of the median. The most data points lie in the range of 50 to 100 with 200 frequency and least points lie in 400 around 10-20.

The box plot looks slightly right skewed as the upper part is bigger than the lower part, the middle or the dividing line is called Median line. We can also see that there are outliers on the upper side of the box plot.

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: The following is calculated using python3. The same is calculated using stats.norm.interval function. The screenshot is attached below for reference.



```
[45] from scipy import stats
     from scipy.stats import norm
```

```
[46] #average weight of adult male in mexico with 94% CI
     stats.norm.interval(0.94,200,30/(2000**0.5))

     (198.738325292158, 201.261674707842)
```

```
[48] #average weight of adult male in mexico with 98% CI
     stats.norm.interval(0.98,200,30/(2000**0.5))

     (198.43943840429978, 201.56056159570022)
```

```
     #average weight of adult male in mexico with 96% CI
     stats.norm.interval(0.96,200,30/(2000**0.5))

     (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.

ANS:

Mean= Sum of all scores/no. of scores

= 738/18

= 41

Median= 40+41/2

= 40.5

Mode= 41

Variance=25.53

Standard Deviation=5.05

Variance & Standard Deviation is calculated using pre-defined formulas in python3.

2) What can we say about the student marks?

ANS: We can conclude that the students have scored marks in the range of 34 to 56. Most students have scored 41 marks. The lowest marks are 34 whereas the highest marks are 56. The average marks obtained by students is 41. Variance simply tells how far are the data points are from the mean & standard deviation is simply square root of variance.

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

Zero Skewness.

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

Positively skewed (right)

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

Negatively skewed (left)

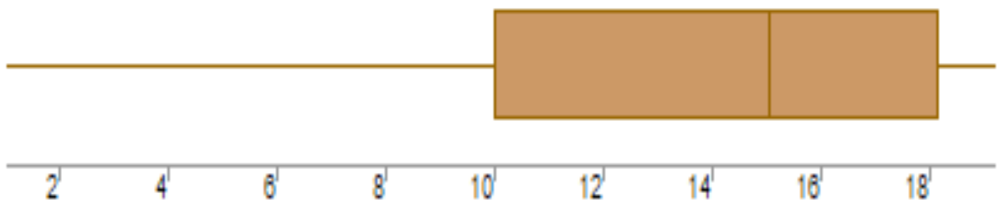
Q16) What does positive kurtosis value indicates for a data ?

Positive value of kurtosis indicates that distribution is peaked and possesses thick tails. It indicates that more data points are located in the tail than around the mean.

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

A negative value of kurtosis means that the distribution has thinner tails and a flatter peak than the normal distribution.

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



What can we say about the distribution of the data?

The data is negatively skewed, the maximum data points lie on the right side of the median. The median is greater than the mean.

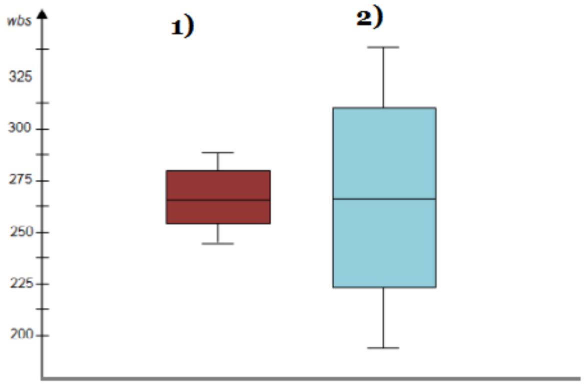
What is nature of skewness of the data?

The data is negatively or left skewed.

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

The IQR here is 8 (18-10).

Q19) Comment on the below Boxplot visualizations?



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

ANS: Let us consider that the upper box plot visualisations are in regard with people's opinion on a subject.

The box plot 1 and 2 have the same median point as the line dividing the box is at the same level.

The box plot 1 states that the people do not have much differences in their opinions, that is they agree with each other on an extent.

But the box plot 2 suggests that the people have a varied opinion and do not agree with each other on a greater extent.

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)

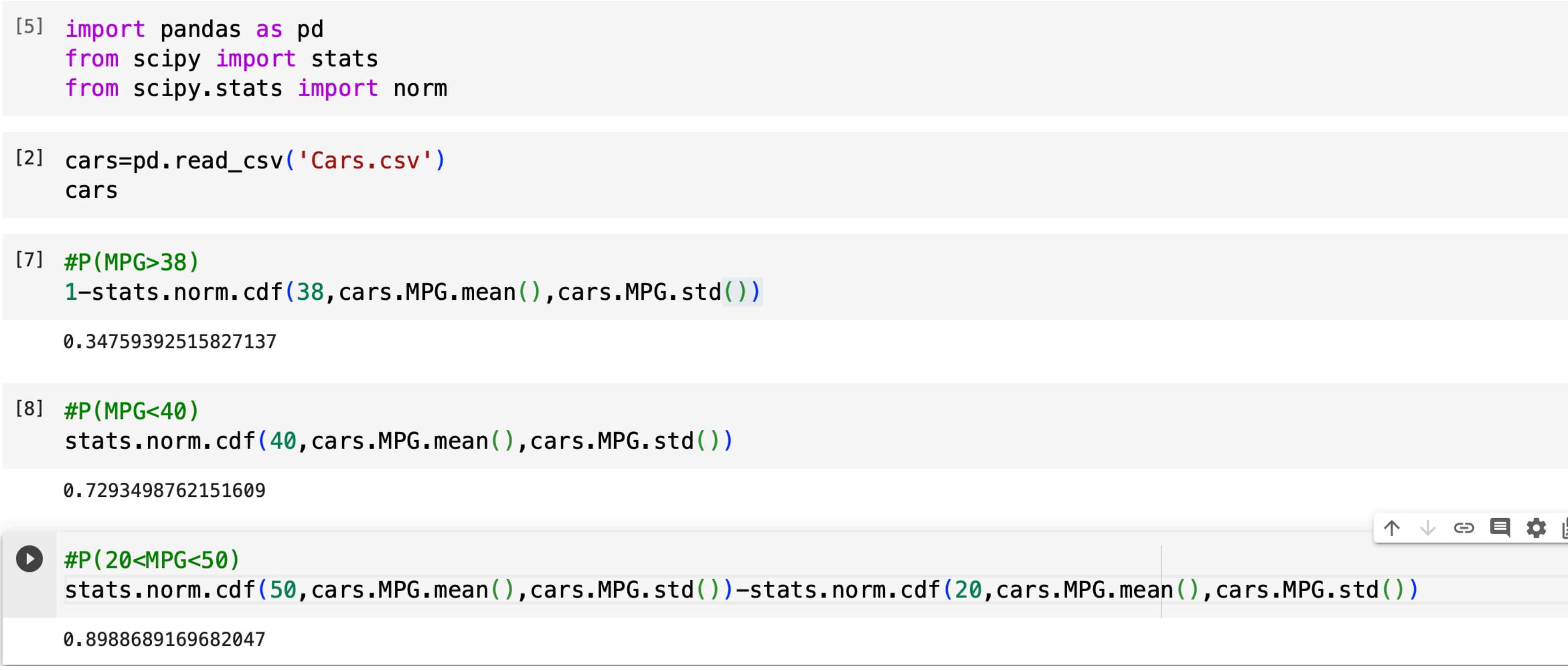
ANS:

A) 0.348 or 34.8%

B) 0.723 or 72.3%

C) 0.899 or 89.9%

Attaching the screenshot below for the same.



```
[5] import pandas as pd
    from scipy import stats
    from scipy.stats import norm
```

```
[2] cars=pd.read_csv('Cars.csv')
    cars
```

```
[7] #P(MPG>38)
    1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())

    0.34759392515827137
```

```
[8] #P(MPG<40)
    stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())

    0.7293498762151609
```

```
#P(20<MPG<50)
stats.norm.cdf(50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(20,cars.MPG.mean(),cars.MPG.std())

0.8988689169682047
```

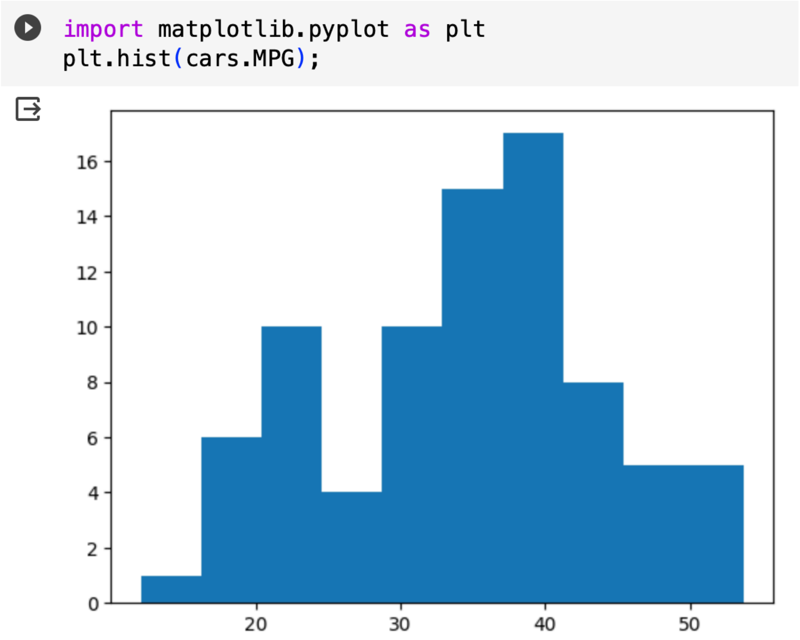
Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution

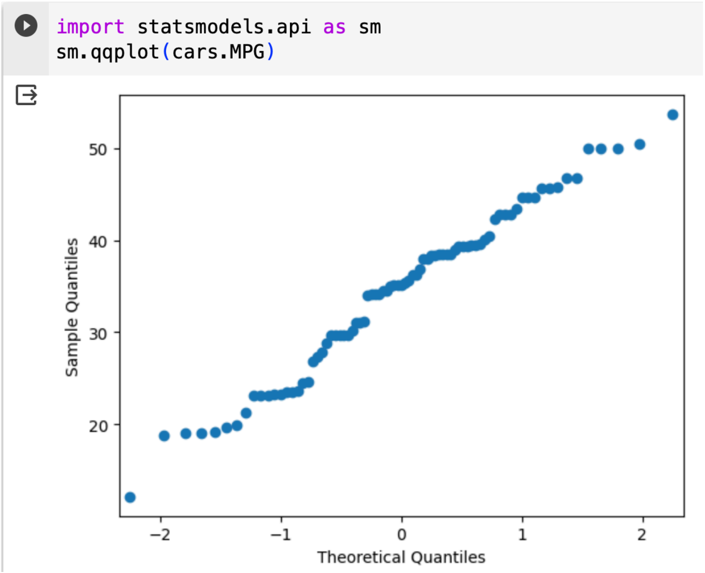
Dataset: Cars.csv

ANS: On plotting the histogram and Q-Q plot for MPG data of cars.csv, it is concluded that the data does not sufficiently follow a pattern of normal distribution.

We can conclude this as neither the histogram has a bell-shaped curved nor the Q-Q plot marks the data points in approximately straight line inclining 45 degrees. Attaching the screenshots for the same.



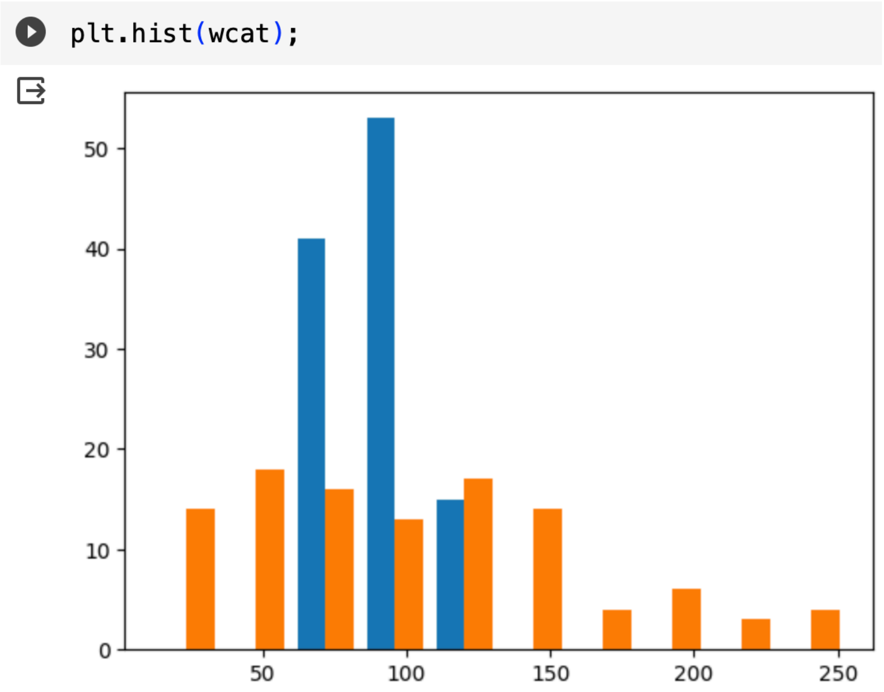
```
import matplotlib.pyplot as plt
plt.hist(cars.MPG);
```



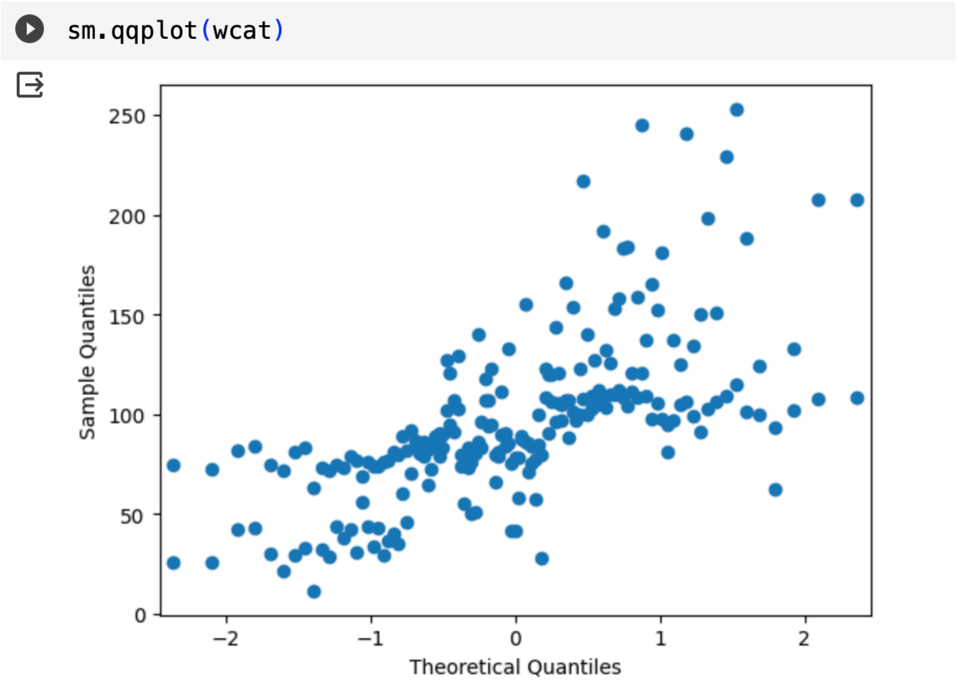
```
import statsmodels.api as sm
sm.qqplot(cars.MPG)
```

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

ANS: It is clearly visible from histogram as well as Q-Q plot that the dataset mentioned above is not normally distributed. We can conclude this as neither the histogram has a bell-shaped curved nor the Q-Q plot marks the data points in approximately straight line inclining 45 degrees. Attaching the screenshots for the same.



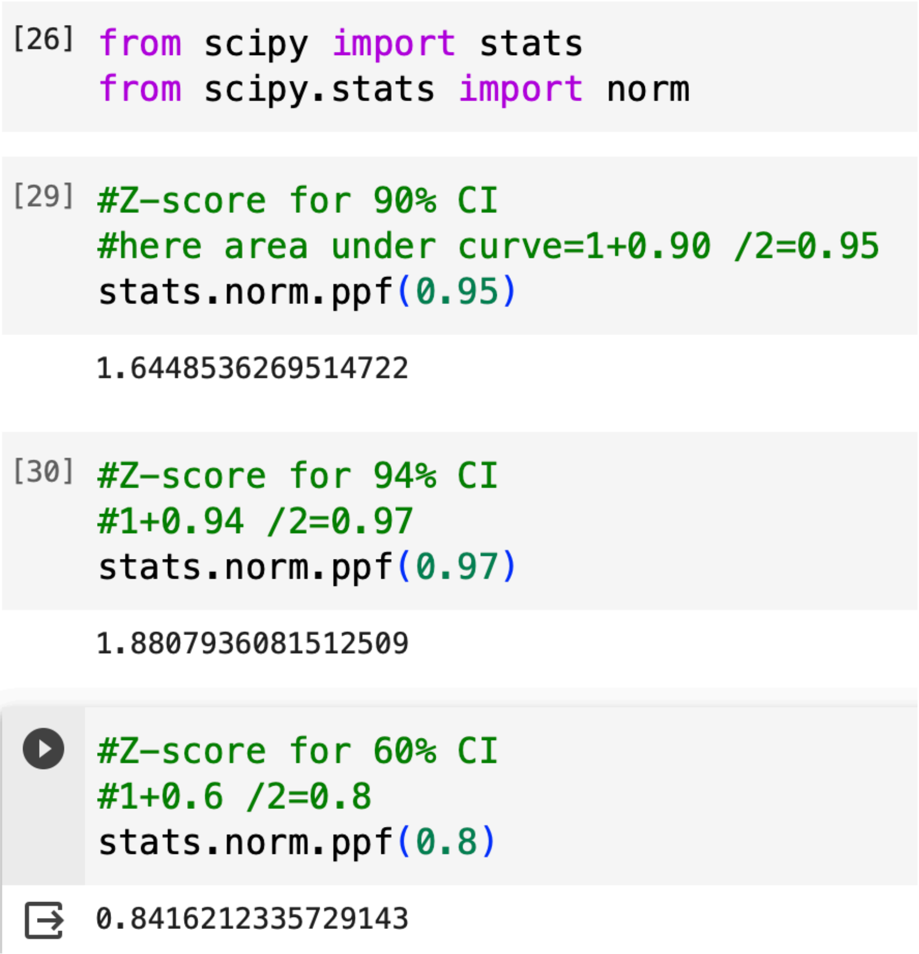
```
plt.hist(wcat);
```



```
sm.qqplot(wcat)
```

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

ANS: The z-scores can be calculated using z-score table but here we will use python3 for the calculation.



```
[26] from scipy import stats
     from scipy.stats import norm
```

```
[29] #Z-score for 90% CI
     #here area under curve=1+0.90 /2=0.95
     stats.norm.ppf(0.95)
```

```
1.6448536269514722
```

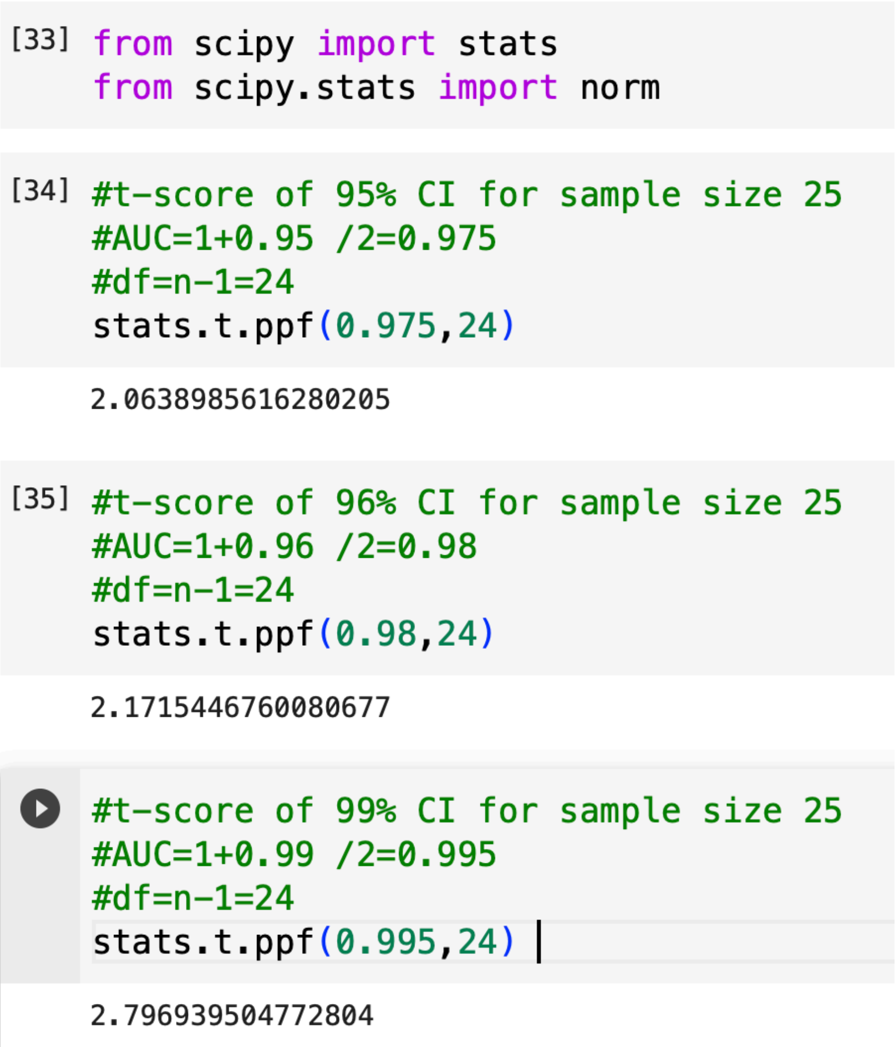
```
[30] #Z-score for 94% CI
     #1+0.94 /2=0.97
     stats.norm.ppf(0.97)
```

```
1.8807936081512509
```

```
#Z-score for 60% CI
#1+0.6 /2=0.8
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



```
[33] from scipy import stats
     from scipy.stats import norm
```

```
[34] #t-score of 95% CI for sample size 25
     #AUC=1+0.95 /2=0.975
     #df=n-1=24
     stats.t.ppf(0.975,24)
```

```
2.0638985616280205
```

```
[35] #t-score of 96% CI for sample size 25
     #AUC=1+0.96 /2=0.98
     #df=n-1=24
     stats.t.ppf(0.98,24)
```

```
2.1715446760080677
```

```
#t-score of 99% CI for sample size 25
#AUC=1+0.99 /2=0.995
#df=n-1=24
stats.t.ppf(0.995,24)
```

```
2.796939504772804
```

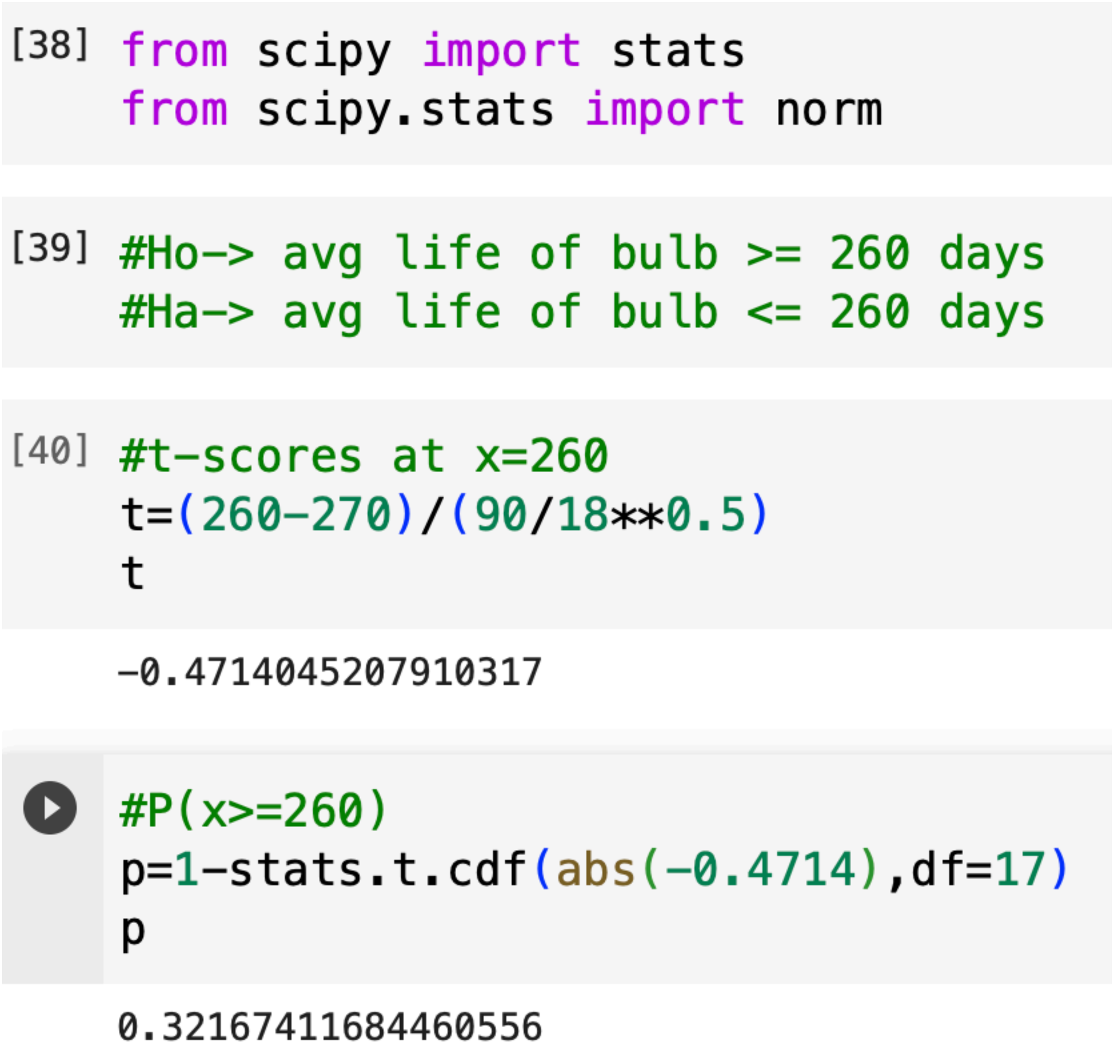
Q 24) A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode → pt(tscore,df)

df → degrees of freedom

ANS: The required probability is 0.322 or 32.2% . The same is calculated using python3. The screenshot is attached for reference.



```
[38] from scipy import stats
     from scipy.stats import norm
```

```
[39] #Ho-> avg life of bulb >= 260 days
     #Ha-> avg life of bulb <= 260 days
```

```
[40] #t-scores at x=260
     t=(260-270)/(90/18**0.5)
     t

     -0.4714045207910317
```

```
     #P(x>=260)
     p=1-stats.t.cdf(abs(-0.4714),df=17)
     p

     0.32167411684460556
```