ASSIGNMENT 1 PART 1 BASIC STATS

Q1) Identify the Data type for the Following:

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

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	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: Sample space for 1 coin tossed (S)={H,T, n(S)=2

P(coin tossed)=1/2

3 Coins are tossed : Sample space for 3 coin tossed
n(S)={HHH,HHT,HTH,THH,TTT,TTH,THT,HTT}

P(2 heads and 1 tail)=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: Sample space for 2 dice rolled S={(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)}

n(S)=36

a) P(Sum is equal to 1): Total Favorable cases (Having sum =1) = 0. As minimum sum is 2 for outcome (1,1). Hence, probability is 0.
b) P(Less than or equal to 4): {(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}
   n(P)=6/36=1/6
c) P(Sum is divisible by 2 and 3): {(1 , 5) , (3 , 3) , (4 , 2) , (5 , 1) , (6 , 6)}
   n(P)=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 = (2 + 3 + 2) = 7.

n(S)= Number of ways of drawing 2 balls out of 7 = 7C2=21

Let E = Event of drawing 2 balls, none of which is blue.

n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls. = 5C2=10

so, probability=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 * 0.12

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

= 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.

Ans: Weigh Mean is less as compared to Points and Score. But the median of Weigh is more. This can be due to some outliers in the data. Also, the calculated SD and Range is more as compared with Points and Score.

CALCULATION MADE	**POINTS**	**SCORE**	**WEIGH**
MEAN	**3.5965625**	**3.21725**	**17.84875**

MEDIAN	**3.695**	**3.325**	**17.71**
MODE	**3.92**	**3.44**	**17.02**
STANDARD DEVIATION	**0.534678736**	**0.978457443**	**1.786943236**
VARIANCE	**0.285881351**	**0.957378968**	**3.193166129**
RANGE	**2.17**	**3.911**	**8.4**

Use Q7.csv file is also attached with calculated necessary details.

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 (X * Probability of X)

= (1/9) (108) + (1/9) (110) + (1/9) (123) + (1/9) (134) + (1/9) (145) + (1/9) (167) + (1/9) (187) + (1/9) (199)

= 145.33

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

Cars speed and distance

Use Q9_a.csv

Ans:

	Speed	Distance
SKEW	**-0.117509861**	**0.80689496**
KURT	**-0.50899442**	**0.405052582**

A negative skew indicates that the tail is on the left side of the distribution, which extends towards more negative values. Here Speed has a negative skew so the tail will fall on the left side of the distribution and Distance has a positive skew so the tail will fall on the right side of the distribution.

SP and Weight(WT)

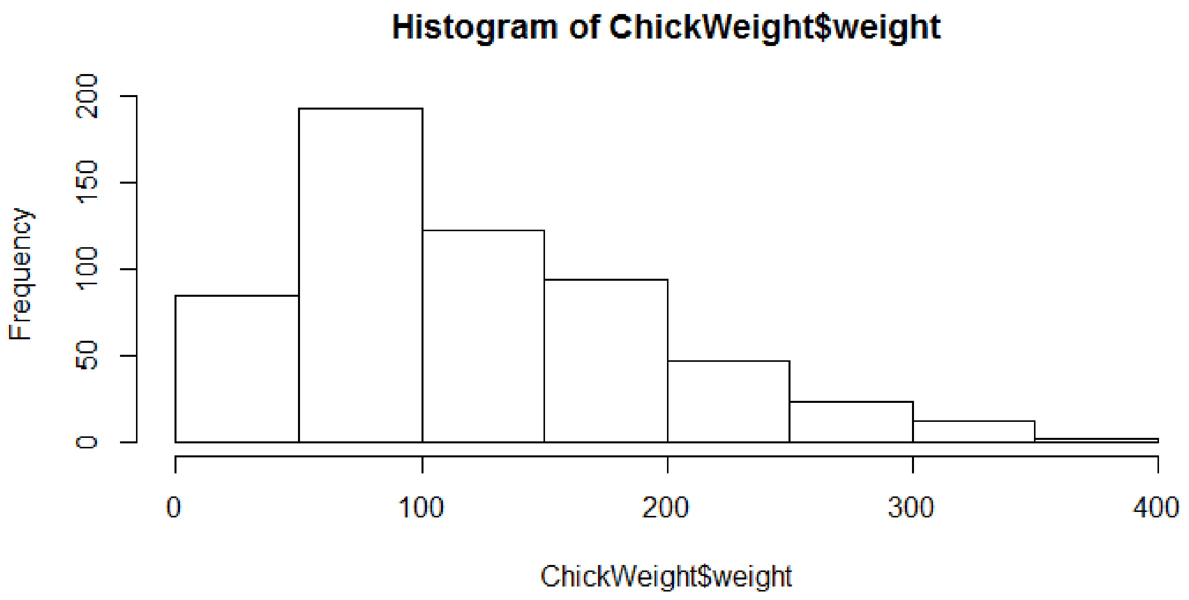
Use Q9_b.csv

Ans:

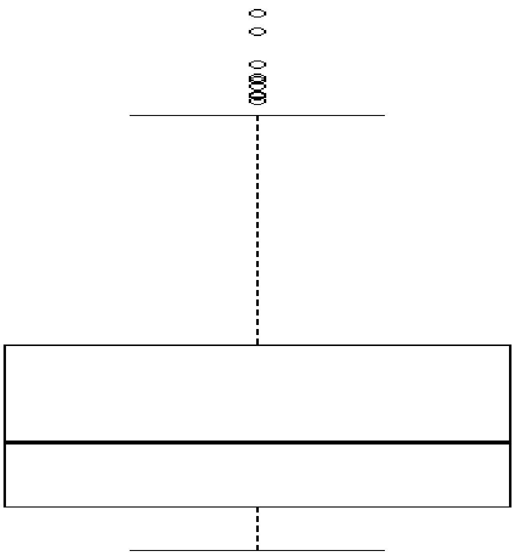
	SP	WT
SKEW	**1.611450196**	**-0.614753326**
KURT	**2.977328944**	**0.950291491**

Here the WT has negative skew so tail will fall on left side of the distribution. Also, the KURT value for SP is quite greater than WT. This infers that some outlier is present in the data which will lead to flat distribution.

Q10) Draw inferences about the following boxplot & histogram



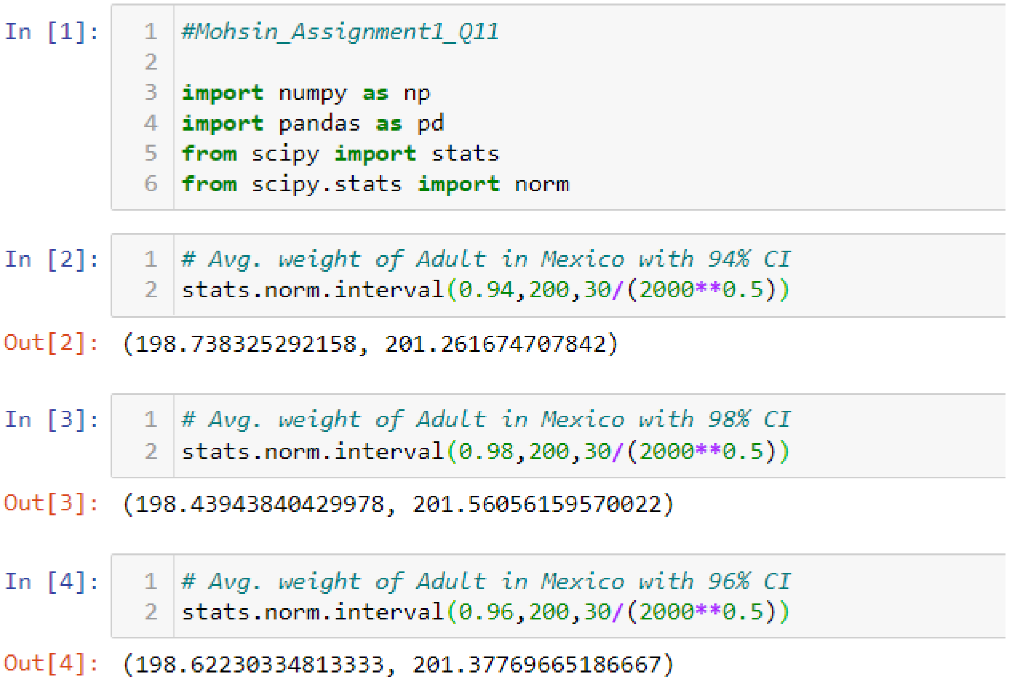
Ans: The histograms peak has right skew and tail is on right. Mean > Median. We have outliers on the higher side.



Ans: The boxplot has outliers on the maximum 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:



```
In [1]:
#Mohsin_Assignment1_Q11

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

```
In [2]:
# Avg. weight of Adult in Mexico with 94% CI
stats.norm.interval(0.94,200,30/(2000**0.5))

Out[2]: (198.738325292158, 201.261674707842)
```

```
In [3]:
# Avg. weight of Adult in Mexico with 98% CI
stats.norm.interval(0.98,200,30/(2000**0.5))

Out[3]: (198.43943840429978, 201.56056159570022)
```

```
In [4]:
# Avg. weight of Adult in Mexico with 96% CI
stats.norm.interval(0.96,200,30/(2000**0.5))

Out[4]: (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) Mean =41, Median =40.5, Variance =25.52 and Standard Deviation =5.05

2) we don't have outliers and the data is slightly skewed towards right because mean is greater than median.

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

Ans: No skewness is present we have a perfect symmetrical distribution

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

Ans: Skewness and tail is towards Right

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

Ans: Skewness and tail is towards left

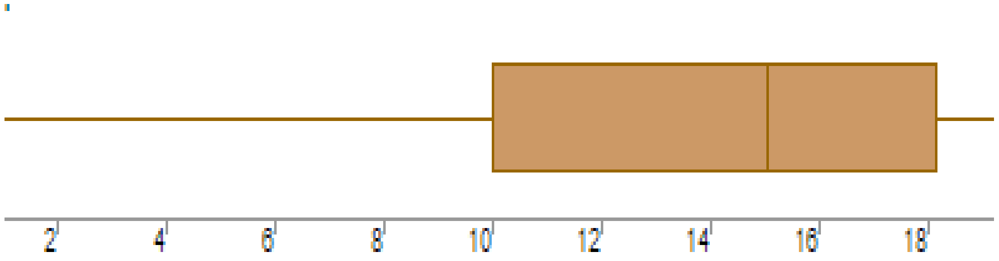
Q16) What does positive kurtosis value indicates for a data ?

Ans: Positive kurtosis means the curve is more peaked and it is Leptokurtic

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

Ans: Negative Kurtosis means the curve will be flatter and broader

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



What can we say about the distribution of the data?

Ans: The above Boxplot is not normally distributed the median is towards the higher value

What is nature of skewness of the data?

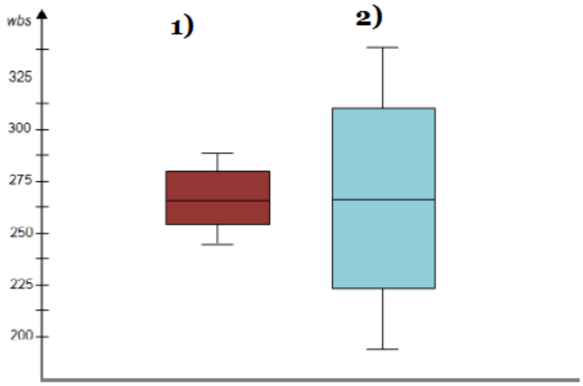
Ans: The data is a skewed towards left. The whisker range of minimum value is greater than maximum

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

Ans: The Inter Quantile Range = Q3 Upper quartile – Q1 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.



Ans: First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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

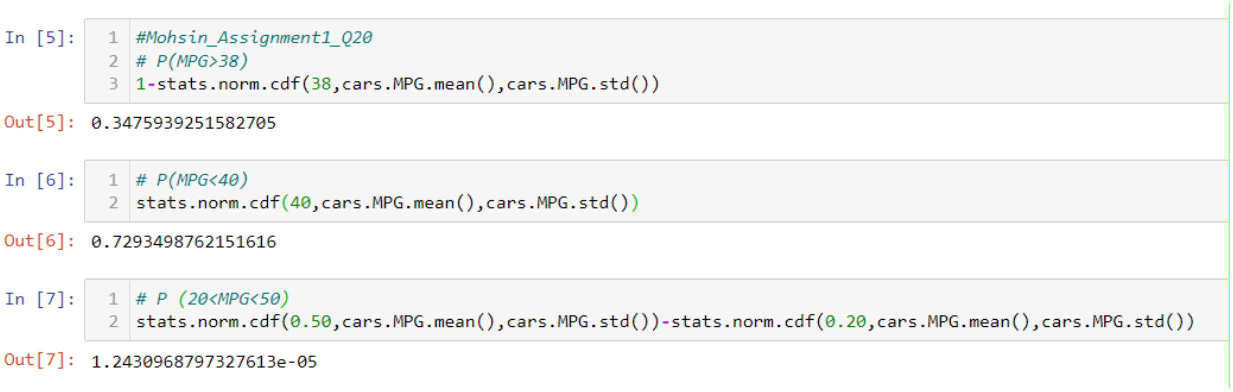
Data _set: Cars.csv

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:

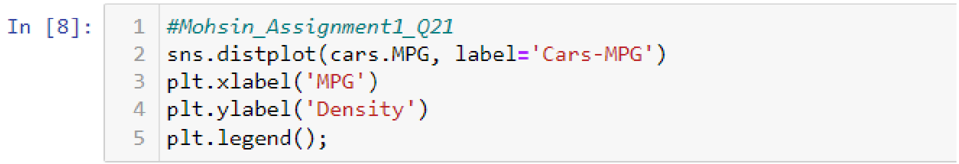


```
In [5]: #Mohsin_Assignment1_Q20
        # P(MPG>38)
        1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())

Out[5]: 0.3475939251582705

In [6]: # P(MPG<40)
        stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())

Out[6]: 0.7293498762151616

In [7]: # P (20<MPG<50)
        stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std())

Out[7]: 1.2430968797327613e-05
```

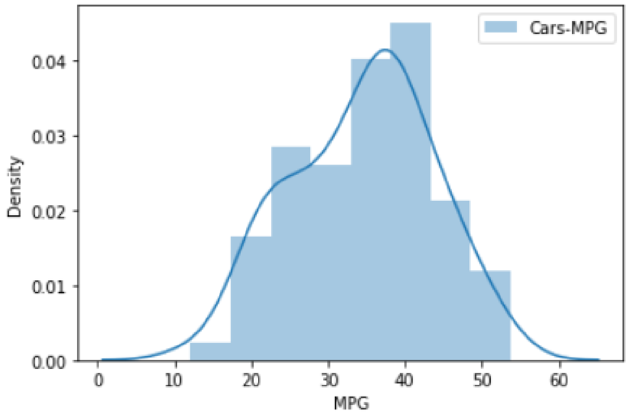
Q 21) Check whether the data follows normal distribution

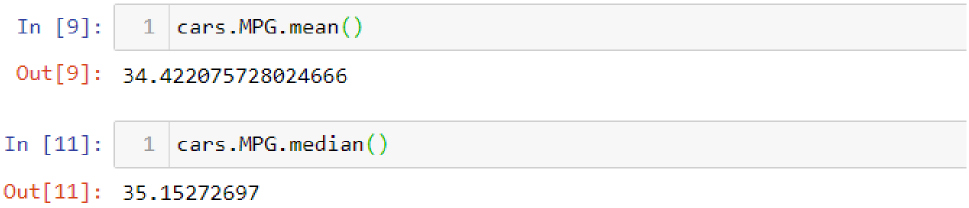
a) Check whether the MPG of Cars follows Normal Distribution
   Dataset: Cars.csv

Ans:



```
In [8]: #Mohsin_Assignment1_Q21
        sns.distplot(cars.MPG, label='Cars-MPG')
        plt.xlabel('MPG')
        plt.ylabel('Density')
        plt.legend();
```





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

Out[9]: 34.422075728024666

In [11]: cars.MPG.median()

Out[11]: 35.15272697
```

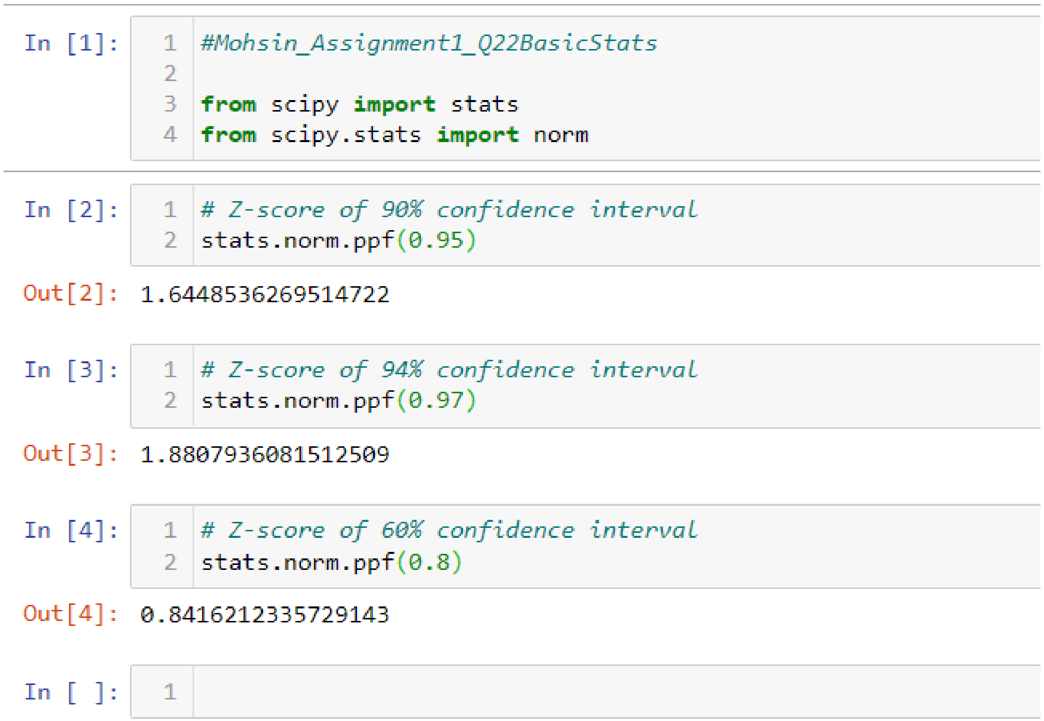
Inference: MPG of cars follows normal distributionMPG of cars follows normal distribution

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

Ans: Adipose Tissue (AT) and Waist does not follow Normal Distribution
Note: Code is attached.

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

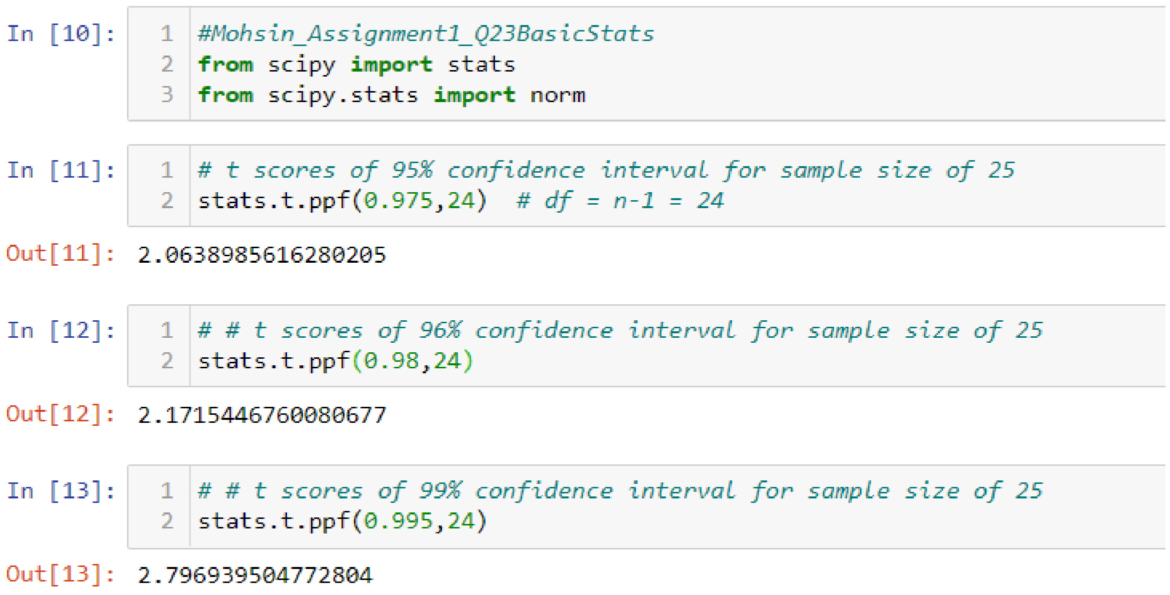
Ans:



```
In [1]:  #Mohsin_Assignment1_Q22BasicStats

         from scipy import stats
         from scipy.stats import norm

In [2]:  # Z-score of 90% confidence interval
         stats.norm.ppf(0.95)

Out[2]:  1.6448536269514722

In [3]:  # Z-score of 94% confidence interval
         stats.norm.ppf(0.97)

Out[3]:  1.8807936081512509

In [4]:  # Z-score of 60% confidence interval
         stats.norm.ppf(0.8)

Out[4]:  0.8416212335729143

In [ ]:
```

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

Ans:



```
In [10]: #Mohsin_Assignment1_Q23BasicStats
         from scipy import stats
         from scipy.stats import norm

In [11]: # t scores of 95% confidence interval for sample size of 25
         stats.t.ppf(0.975,24)  # df = n-1 = 24

Out[11]: 2.0638985616280205

In [12]: # # t scores of 96% confidence interval for sample size of 25
         stats.t.ppf(0.98,24)

Out[12]: 2.1715446760080677

In [13]: # # t scores of 99% confidence interval for sample size of 25
         stats.t.ppf(0.995,24)

Out[13]: 2.796939504772804
```

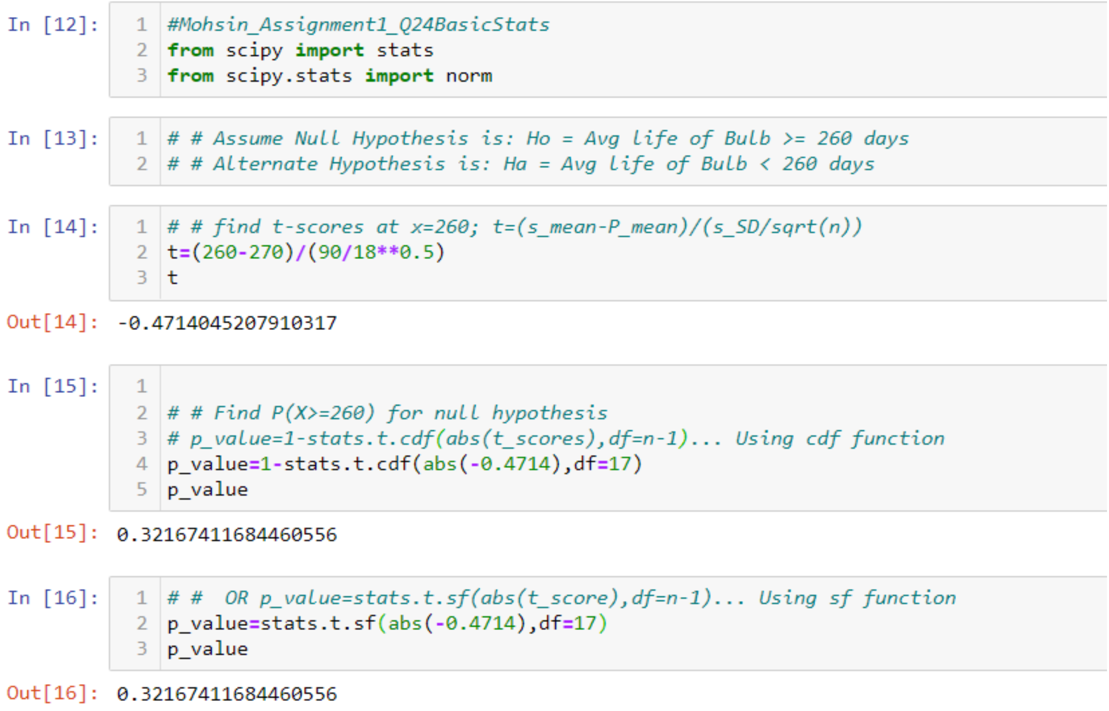
Q 24)   A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs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

Hint:

rcode → pt(tscore,df)

df → degrees of freedom

Ans:



```
In [12]: #Mohsin_Assignment1_Q24BasicStats
         from scipy import stats
         from scipy.stats import norm

In [13]: # # Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days
         # # Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

In [14]: # # find t-scores at x=260; t=(s_mean-P_mean)/(s_SD/sqrt(n))
         t=(260-270)/(90/18**0.5)
         t

Out[14]: -0.4714045207910317

In [15]: 
         # # Find P(X>=260) for null hypothesis
         # p_value=1-stats.t.cdf(abs(t_scores),df=n-1)... Using cdf function
         p_value=1-stats.t.cdf(abs(-0.4714),df=17)
         p_value

Out[15]: 0.32167411684460556

In [16]: # #  OR p_value=stats.t.sf(abs(t_score),df=n-1)... Using sf function
         p_value=stats.t.sf(abs(-0.4714),df=17)
         p_value

Out[16]: 0.32167411684460556
```