Name: *Sharan Kumar Punjal* **Batch**: *Data Science 19/05 (Hyderabad)*

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	Nominal
Fahrenheit Temperature	Interval
Height	Ratio
Type of living accommodation	Nominal
Level of Agreement	Ratio
IQ(Intelligence Scale)	Nominal
Sales Figures	Ratio
Blood Group	Nominal
Time Of Day	Ratio

Time on a Clock with Hands	Ratio
Number of Children	Countable
Religious Preference	Nominal
Barometer Pressure	Interval
SAT Scores	Nominal
Years of Education	Countable

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

Soln. Probability of success (heads), $p = 0.5$

Probability of failure (tails), $q = 0.5$

Number of coins = 3

Required number of successes = 2

Probability of getting two heads and one tail

$= 3C2 * p^2 * q^1$

= **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

Soln. a) Probability of the sum being equal to 1 when two dice are rolled is **0** as there is no chance of getting a sum of 1 in this scenario.

b) Probability of getting a sum of 2, P(2) = 1/36

Probability of getting a sum of 3, P(3) = 2/36

Probability of getting a sum of 4, P(4) = 3/36

Probability of getting a sum less than or equal to 4, P(<=4)

= P(2) +P(3) + P(4)

= 1/36 + 2/36 + 3/36

= 6/36 = 1/6 = **0.16**

c) Probability of getting sum divisible by 2, a = 18/36
Probability of getting sum divisible by 3, b = 12/36
Probability of getting sum divisible by both 2 and 3 = a*b
= 18/36 * 12/36
= 1/6 = **0.167**

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?

Soln. Probability of drawing a blue ball (p) = 2/7

Probability of not drawing a blue ball (q) = 5/7

Probability of none of the balls are blue when two balls are drawn

= 2C0 * p^0 * q^2

= **0.51**

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

Soln. Expected value of the distribution = np,

Where n - no. of trials

p – Probability of success

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 + 0.80 + 1.95 + 0.025 + 0.01 + 0.240

= 3.04

The expected number of candies for a randomly selected child is **3.04**

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

Soln. The given dataset consists of cars along with theirs Points, Score and Weight

For Points: Mean of the dataset = **3.60**

Median of the dataset = **3.69**

Mode of the dataset = **(3.07, 3.92)**

Standard Deviation of the dataset = **0.53**

Variance of the dataset = **0.285**

Range of the dataset = Max-Min = 4.93-2.76 = **2.17**

The values calculated gives us an idea of the average value being 3.60 with a deviation in values of 0.53 over a range of 2.17. The central value of Points turns out to be 3.69. The recurring value/mode of the Points are 3.07 and 3.92.

For Score: Mean of the dataset = **3.217**

Median of the dataset = **3.325**

Mode of the dataset = **3.44**

Standard Deviation of the dataset = **0.978**

Variance of the dataset = **0.956**

Range of the dataset = Max-Min = 5.424-1.513 = **3.911**

The values calculated gives us an idea of the average value being 3.217 with a deviation in values of 0.978 over a range of 3.911. The central value of Score turns out to be 3.325. The recurring value/mode of the Score is 3.44.

For Weigh: Mean of the dataset = **17.85**

Median of the dataset = **17.71**

Mode of the dataset = **(17.02, 18.90)**

Standard Deviation of the dataset = **1.78**

Variance of the dataset = **3.17**

Range of the dataset = Max-Min = 22.9-14.5 = **8.40**

The values calculated gives us an idea of the average value being 17.85 with a deviation in values of 1.78 over a range of 8.40. The central value of Weigh turns out to be 17.71. The recurring value/mode of the Weigh are 17.02 and 18.90.

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?

Soln. The expected value for weight for a patient chosen at random

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

= **145.33**

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

Cars speed and distance

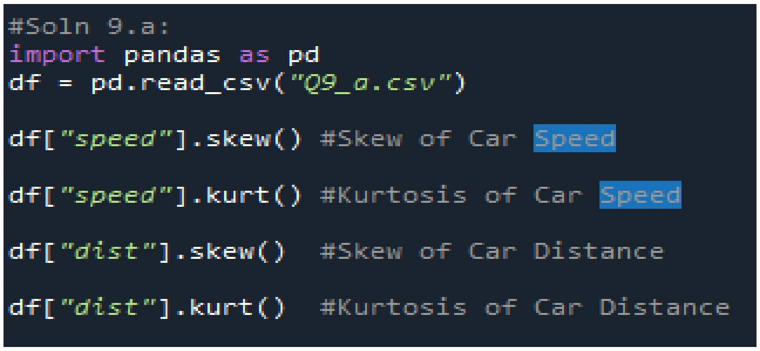
Use Q9_a.csv

Soln. Cars speed: The skewness of the Speed data is **-0.117** (negatively skewed), values are more distributed on the right side.

The Kurtosis of the Speed data is **-0.509** which denotes a platykurtic distribution where the values have a flat distribution and moderately spread out.

Distance: The skewness of the Distance data is **0.807** (positively skewed), values are more distributed on the right side.

The Kurtosis of the Distance data is **0.405** which denotes a leptokurtic distribution where the values are distributed tall and thin.



```
#Soln 9.a:
import pandas as pd
df = pd.read_csv("Q9_a.csv")

df["speed"].skew() #Skew of Car Speed

df["speed"].kurt() #Kurtosis of Car Speed

df["dist"].skew()  #Skew of Car Distance

df["dist"].kurt()  #Kurtosis of Car Distance
```

SP and Weight(WT)

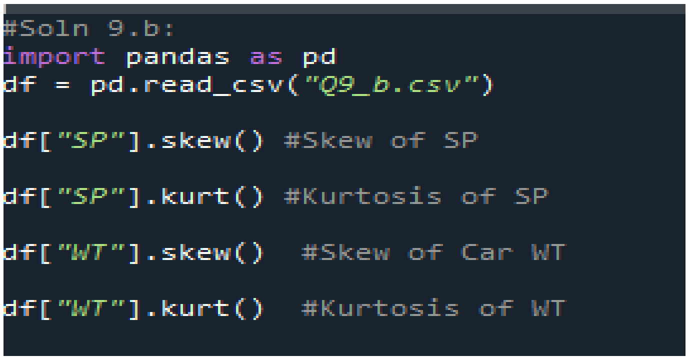
Use Q9_b.csv

Soln. SP: The skewness of the SP data is **1.611** (positively skewed), values are
. more distributed on the left side.

The Kurtosis of the SP data is **2.977** which denotes a leptokurtic distribution where the values are distributed and tall and thin.

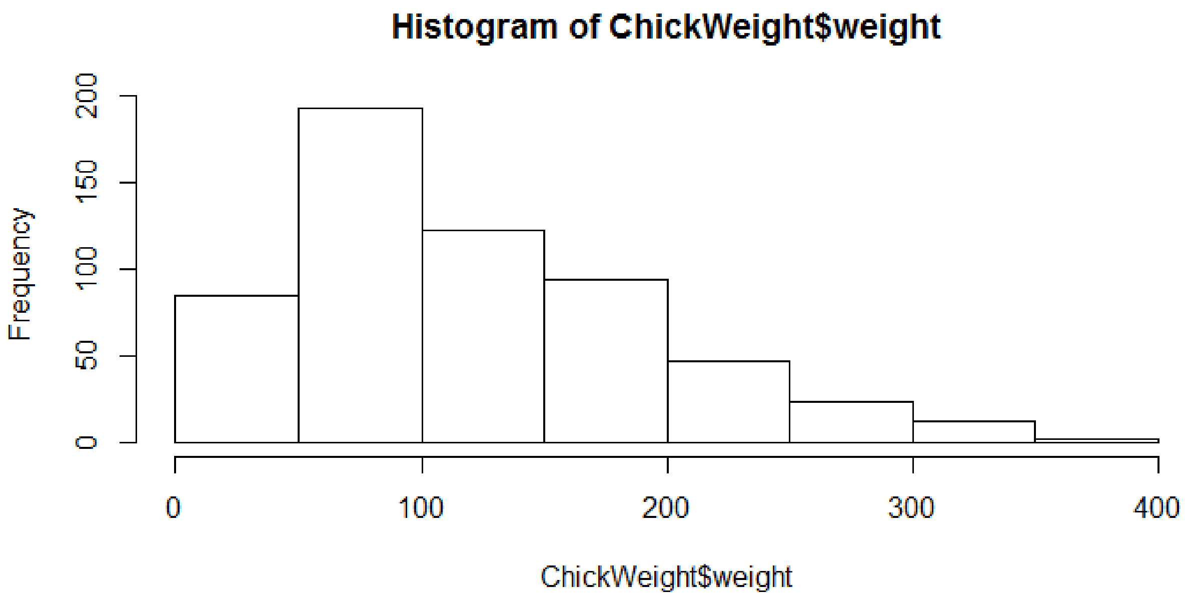
WT: The skewness of the WT data is **-0.614** (negatively skewed), values are
. more distributed on the right side.

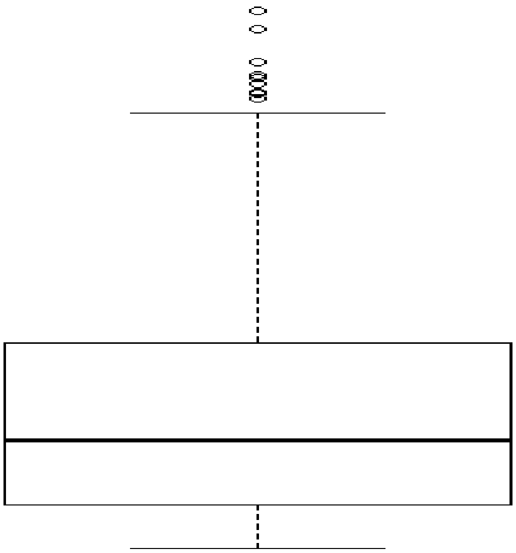
The Kurtosis of the WT data is **0.950** which denotes a leptokurtic distribution where the values are distributed and tall and thin.



```
#Soln 9.b:
import pandas as pd
df = pd.read_csv("Q9_b.csv")

df["SP"].skew() #Skew of SP

df["SP"].kurt() #Kurtosis of SP

df["WT"].skew()  #Skew of Car WT

df["WT"].kurt()  #Kurtosis of WT
```

Q10) Draw inferences about the following boxplot & histogram





Soln. Inference of Histogram:

The range of the values of ChickWeight$weight on the Histogram starts from 0 and ends at 400 with an interval of 50. The normal distribution curve of the data is positively skewed. Majority of the data are in the range of 50-100 ChickWeight$weight with 200 occurences. Least number of the ChickWeight values occur between the 350-400 range.

Inference of boxplot:

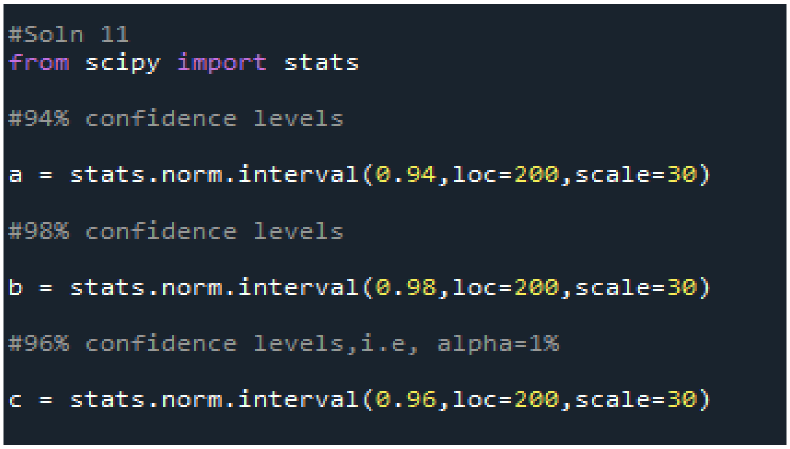
On observing the box-plot, we can conclude that the bulk of the data is on the left side/lower side of values, hence the data is postively skewed. There are 7 outliers and areover the upper whisker length of the box plot.

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?

Soln. The average weight of an adult male in Mexico with a 94% confidence interval: **(143.58, 256.42)**

The average weight of an adult male in Mexico with a 98% confidence interval: **(130.21, 269.79)**

The average weight of an adult male in Mexico with a 96% confidence interval: **(138.39, 261.61)**



```
#Soln 11
from scipy import stats

#94% confidence levels

a = stats.norm.interval(0.94,loc=200,scale=30)

#98% confidence levels

b = stats.norm.interval(0.98,loc=200,scale=30)

#96% confidence levels,i.e, alpha=1%

c = stats.norm.interval(0.96,loc=200,scale=30)
```

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?

Soln. 1) Mean of the scores obtained by the student = **41**

Standard Deviation of scores obtained by the student = **4.91**

Variance of scores obtained by the student = **24.11**

2) We can conclude that the average scores obtained by the student is 41 with deviation in scores being 4.91. The highest score obtained by the student is 56 and the least score obtained is 34.

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

Soln. The nature of skewness when mean and median of the data are equal is **symmetrical** in shape.

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

Soln. The nature of skewness when mean>median is **positively skewed**.

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

Soln. The nature of skewness when median>mean is **negatively skewed**.

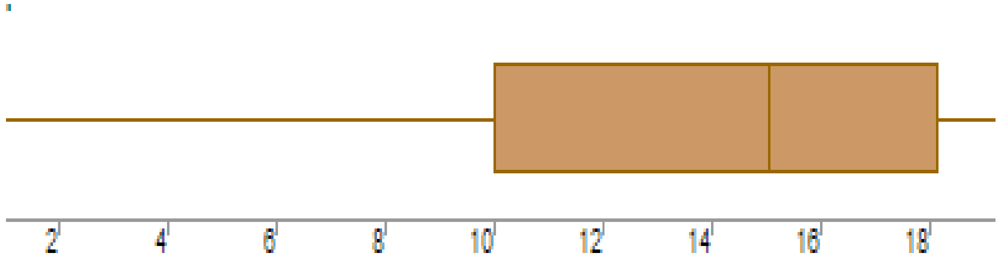
Q16) What does positive kurtosis value indicates for a data?

Soln. Positive kurtosis value indicates narrow spread of peak values and thicker spread of data in the tail values

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

Soln. Negative kurtosis value indicates wider spread of peak values and thinner spread of data in the tail values.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

Soln. Upon inspection of the box plot we can say that the spread of data ranges from about 0 to about 20. 50% of the data lies between 10 to a little over 18. The central value of the data upon inspection from the box plot can be said to be 15.

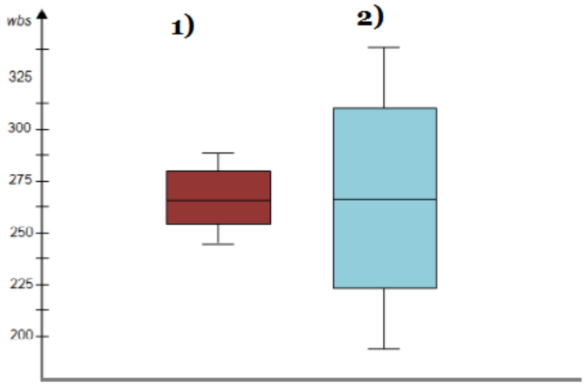
The nature of skewness of the given data distribution is negatively skewed owing to the central value of the entire data.

The IQR or Inter-Quartile Range (approximate) = 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.

Soln. Both the box plots have the same median value. Boxplot 1 has a smaller range of values stating a bulk of the data has a smaller deviation compared to Boxplot 2 which has a wider range of values where data has a larger deviation. On observing the boxplots we can conclude that the data is normally distributed.

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.

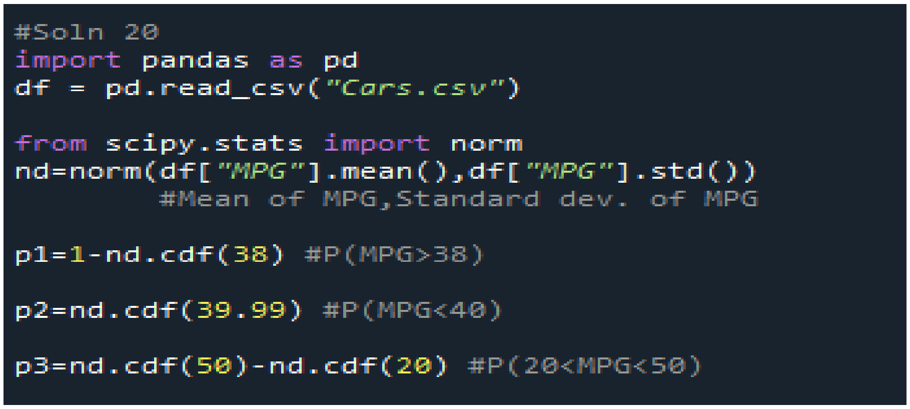
MPG <- Cars$MPG

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

Soln. a. P(MPG>38) = **0.347**

b. P(MPG<40) = **0.729**

c. P(20<MPG<50) = **0.899**

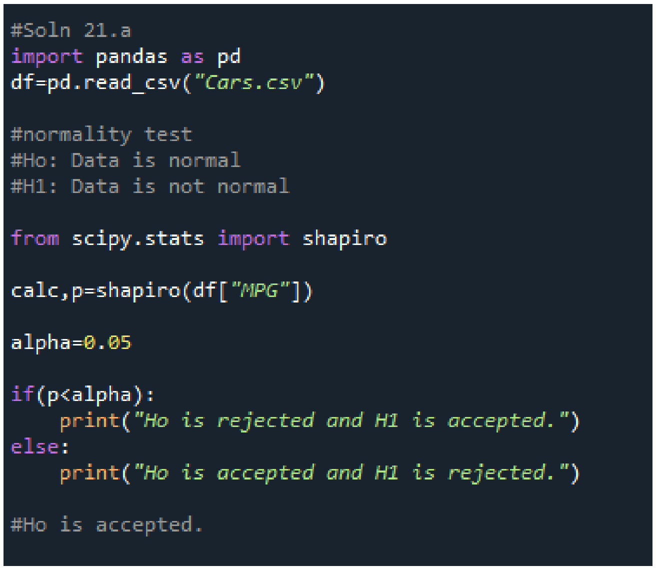


```
#Soln 20
import pandas as pd
df = pd.read_csv("Cars.csv")

from scipy.stats import norm
nd=norm(df["MPG"].mean(),df["MPG"].std())
        #Mean of MPG,Standard dev. of MPG

p1=1-nd.cdf(38) #P(MPG>38)

p2=nd.cdf(39.99) #P(MPG<40)

p3=nd.cdf(50)-nd.cdf(20) #P(20<MPG<50)
```

Q 21) Check whether the data follows normal distribution

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

Soln.

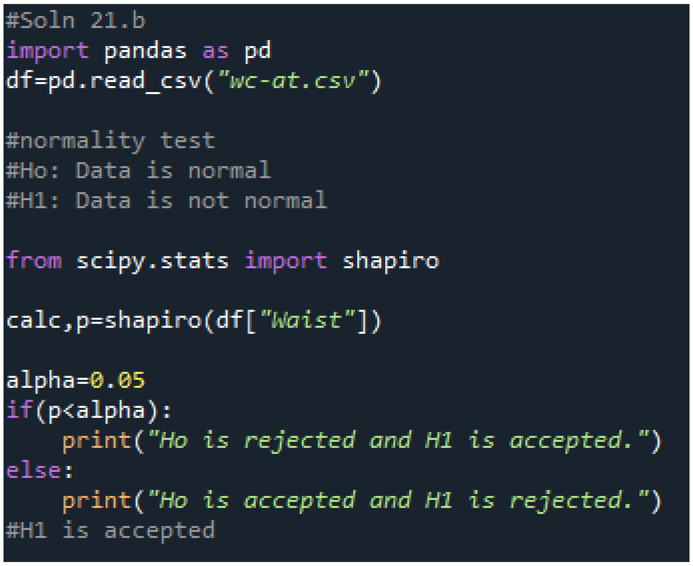


```
#Soln 21.a
import pandas as pd
df=pd.read_csv("Cars.csv")

#normality test
#Ho: Data is normal
#H1: Data is not normal

from scipy.stats import shapiro

calc,p=shapiro(df["MPG"])

alpha=0.05

if(p<alpha):
    print("Ho is rejected and H1 is accepted.")
else:
    print("Ho is accepted and H1 is rejected.")

#Ho is accepted.
```

Using the Shapiro test we are able to conclude with enough statistical significance that the MPG of Cars **follows Normal Distribution.**

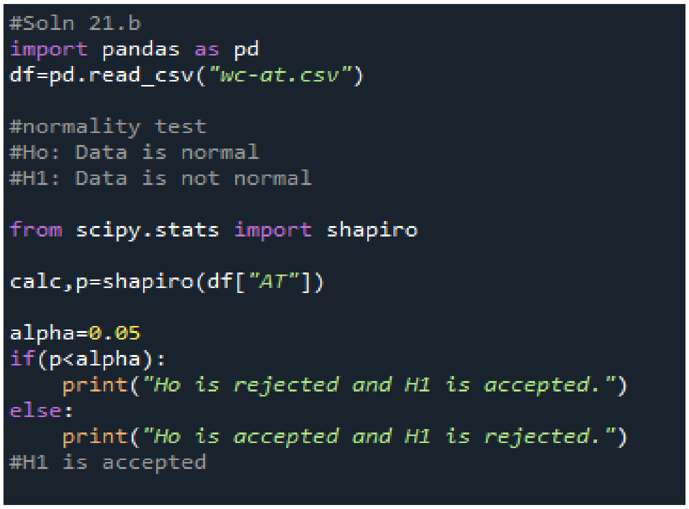
b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution
   Dataset: wc-at.csv

Soln.



```
#Soln 21.b
import pandas as pd
df=pd.read_csv("wc-at.csv")

#normality test
#Ho: Data is normal
#H1: Data is not normal

from scipy.stats import shapiro

calc,p=shapiro(df["Waist"])

alpha=0.05
if(p<alpha):
    print("Ho is rejected and H1 is accepted.")
else:
    print("Ho is accepted and H1 is rejected.")
#H1 is accepted
```

Using the Shapiro test we are able to conclude with enough statistical significance that the Waist Circumference **does not follow Normal Distribution.**



```
#Soln 21.b
import pandas as pd
df=pd.read_csv("wc-at.csv")

#normality test
#Ho: Data is normal
#H1: Data is not normal

from scipy.stats import shapiro

calc,p=shapiro(df["AT"])

alpha=0.05
if(p<alpha):
    print("Ho is rejected and H1 is accepted.")
else:
    print("Ho is accepted and H1 is rejected.")
#H1 is accepted
```

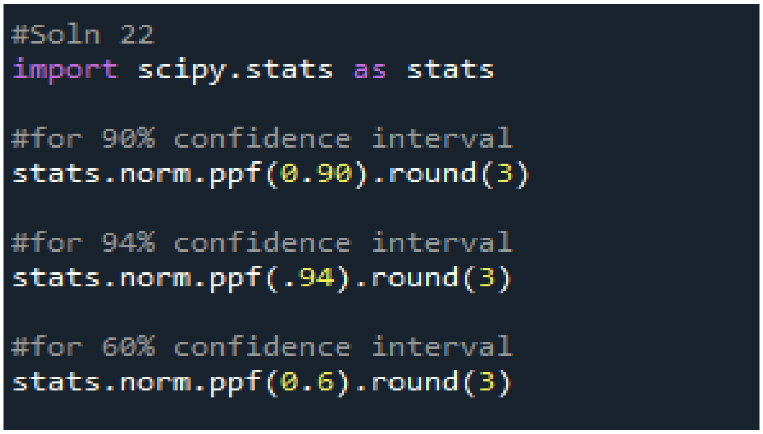
Using the Shapiro test we are able to conclude with enough statistical significance that the Adipose Tissue **does not follow Normal Distribution.**

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

Soln. Z-score for 90% confidence interval = **1.282**

Z-score for 94% confidence interval = **1.555**

Z-score for 60% confidence interval = **0.253**



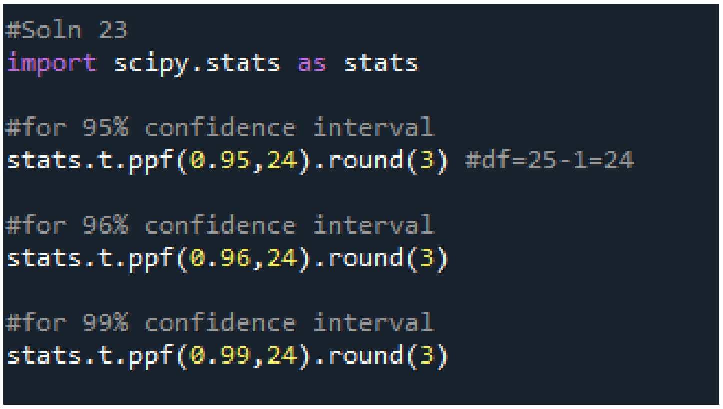
```
#Soln 22
import scipy.stats as stats

#for 90% confidence interval
stats.norm.ppf(0.90).round(3)

#for 94% confidence interval
stats.norm.ppf(.94).round(3)

#for 60% confidence interval
stats.norm.ppf(0.6).round(3)
```

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

Soln. t-score for 95% confidence interval = **1.711**

t-score for 96% confidence interval = **1.828**

t-score for 99% confidence interval = **2.492**



```
#Soln 23
import scipy.stats as stats

#for 95% confidence interval
stats.t.ppf(0.95,24).round(3) #df=25-1=24

#for 96% confidence interval
stats.t.ppf(0.96,24).round(3)

#for 99% confidence interval
stats.t.ppf(0.99,24).round(3)
```

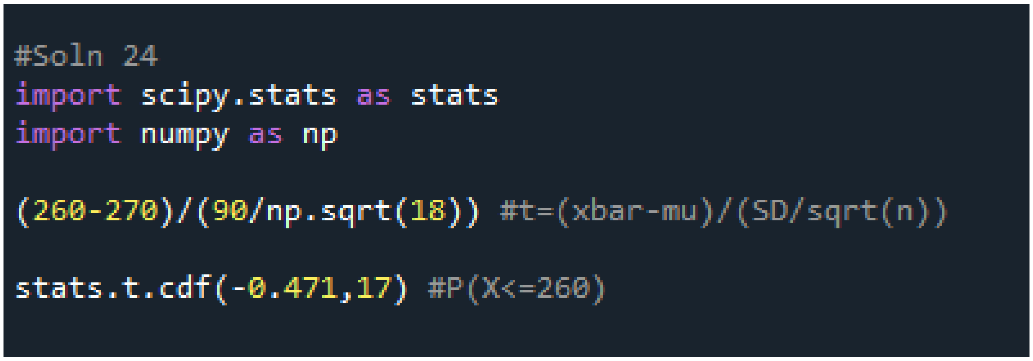
Q 24) A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would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

Hint:

rcode → pt(tscore,df)

df → degrees of freedom

Soln. P(X<=260) = **0.322**



```
#Soln 24
import scipy.stats as stats
import numpy as np

(260-270)/(90/np.sqrt(18)) #t=(xbar-mu)/(SD/sqrt(n))

stats.t.cdf(-0.471,17) #P(X<=260)
```