|  |  |
| --- | --- |
| 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 | Ordinal |
| 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 | Interval |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Sol. 3/8

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

Sol a) Zero

b) 1/6

c) 1/6

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?

Sol 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

Sol. Expected no. 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

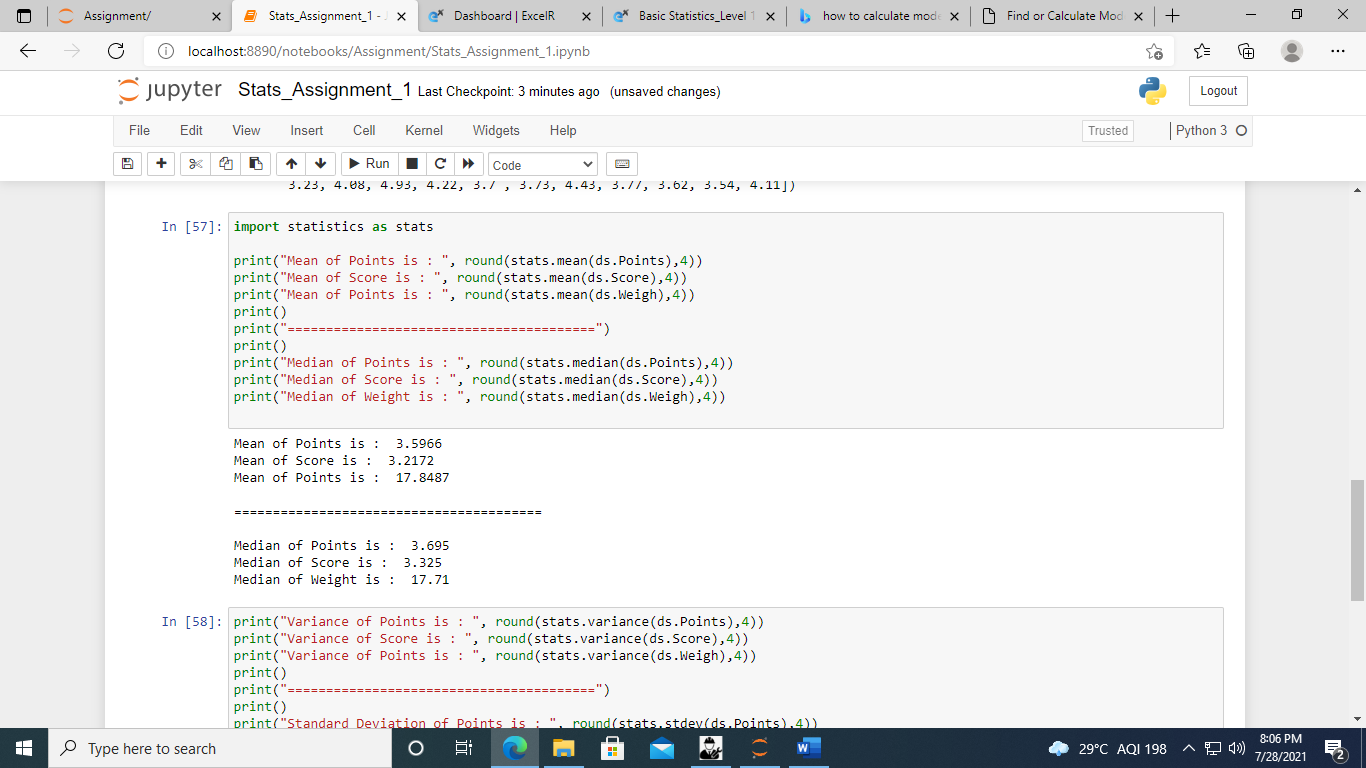
Therefore, expected no. of candies for a randomly selected child is 3 candies.

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

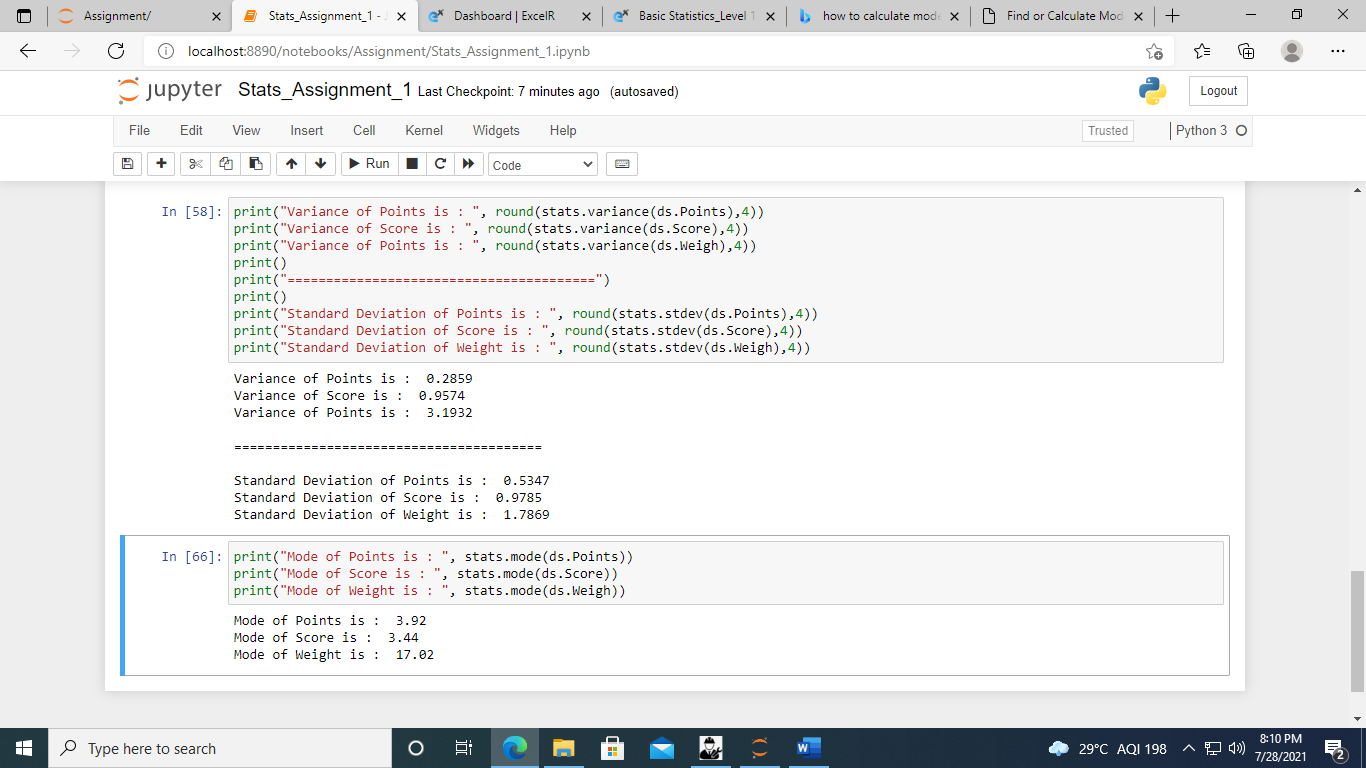
* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

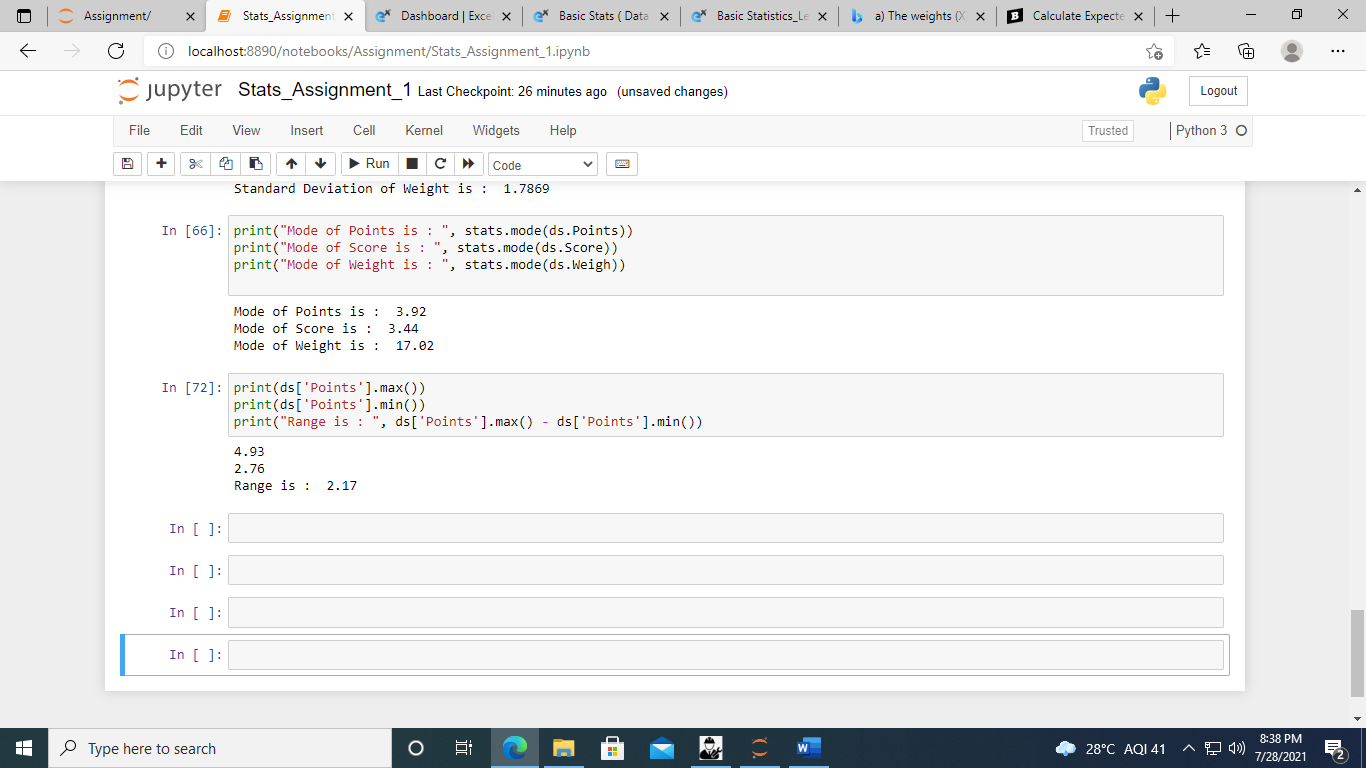


Since the mean, median and mode for each column is approximately similar. This indicates that the data is distributed normally.



Standard Deviation for ‘Points’ and ‘Score’ columns is less than 1 which implies that there is a low variance in data ie more data is gathered around the mean.

SD for ‘Weight’ column is 1.7869 (i.e around 2) meaning 95% of the data lies under the bell curve.



The maximum value is 4.93 and minimum value is 2.76 in the Points column. So, the data in points column ranges from 2.76 to 4.93.

Q8) Calculate Expected Value for the problem below

1. 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?

Sol. If the patient is chosen at random then the expected weight of that patient will be mean of weight.

Mean = 1308/9

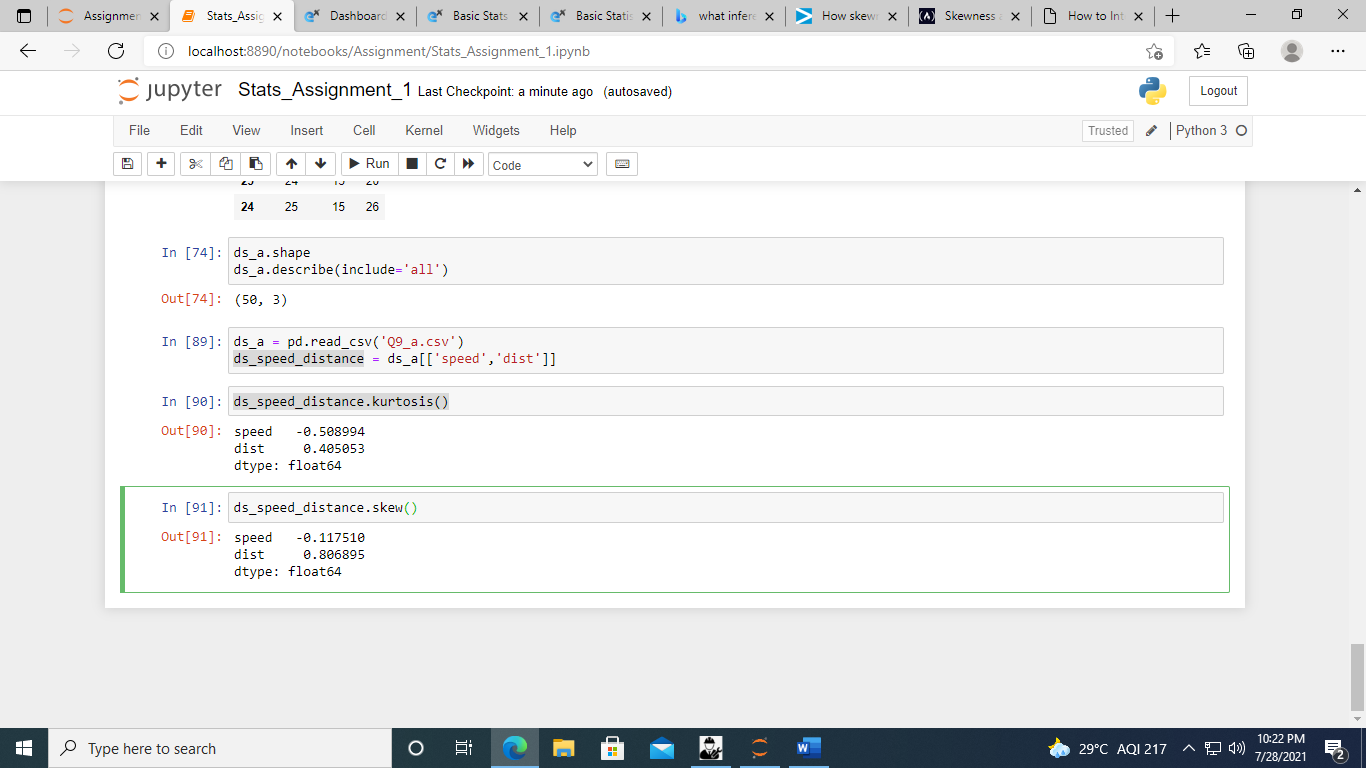
= 145.33

So, if a patient is chosen randomly the expected weight of the patient will be 145.33 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**



**Inferences from skewness**

Negative skewness ( -0.117510) of speed indicates that data is left skewed. Ie median>mean. Ie more values of speed are greater than the average speed and extreme values of speed are lesser than the mean.

Since, the skewness ( -0.117510) is between (-0.5) to (0.5). Therefore data of speed is fairly symmetrical.

Positive skewness (0.806895) of dist indicates that it is right skewed that is mean > median. Ie. More values of dist are lower than the average dist and extreme values of dist are greater than the mean.

Since, the skewness (0.806895) is between 0.5 & 1, therefore data of dist column is moderately skewed.

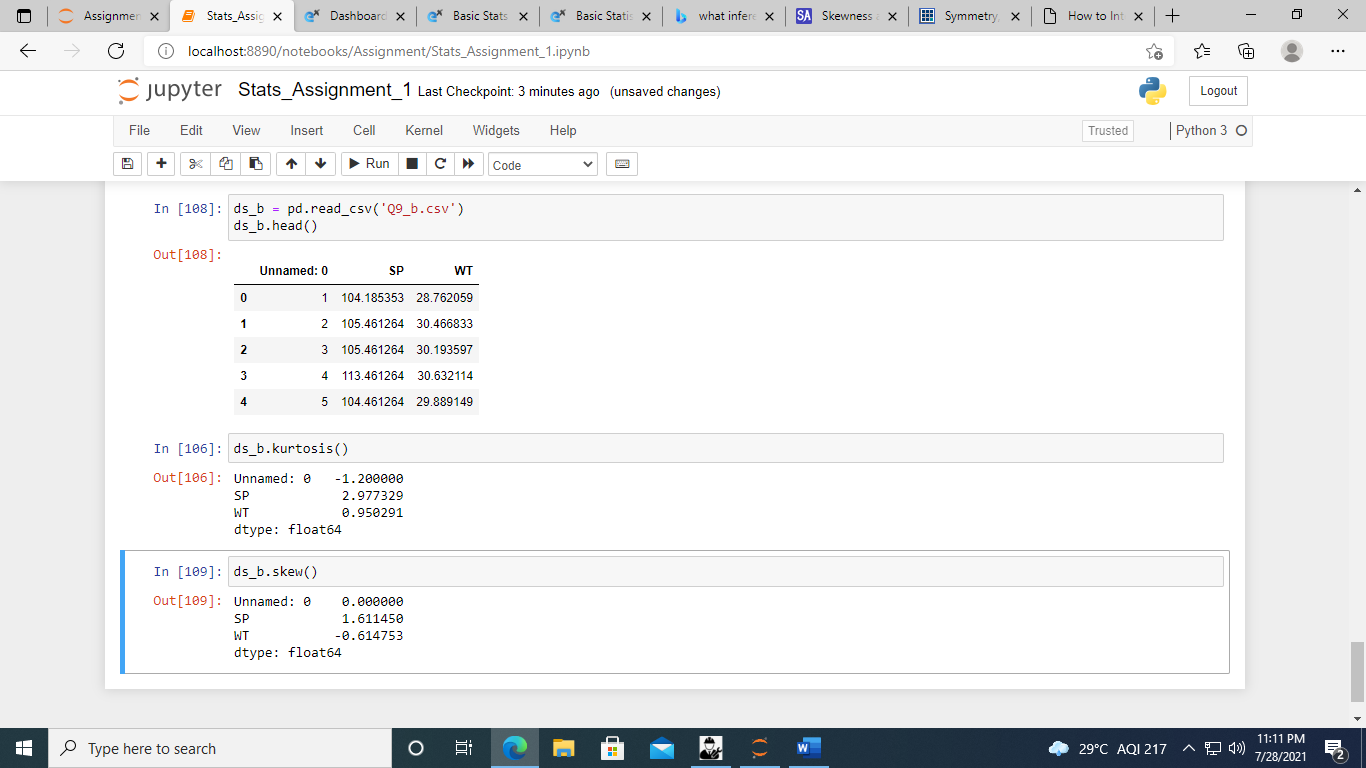
**Inferences from kurtosis**

Negative kurtosis (- 0.508994 ) indicates light-tailed distribution. Ie. The speed values has no outliers.

Positive kurtosis ( 0.405053 ) indicates heavy-tailed distribution .ie. the dist values has large outliers.

**SP and Weight(WT)**

**Use Q9\_b.csv**



**Inferences from Skewness**

SP is (1.611450) is right skewed that is mean > median i.e extreme values of SP are present on the right side of the mean.

Since (1.611450) for SP is above 1, therefore, we can say that the data is highly skewed.

WT is ( -0.614753) is left skewed. Ie. Median > mean. Ie. Extreme values of WT are lower than the mean.

Since ( -0.614753) for WT lies between -0.5 and -1 , therefore, we can say that the data is moderately skewed.

**Inferences from Kurtosis**

SP is ( 2.977329 ) is positive value which means it is heavily tailed on either side ie SP values has large outliers.

WT is (0.950291) is positive means it is heavily tailed and has large outliers.

**Q10) Draw inferences about the following boxplot & histogram**



Sol. Inferences From Histogram

1. In this histogram x – axis represents the different chicken weights whereas the y axis represents the frequencies of that weight.

For eg There are approx. 200 chickens with weight from 50 to 100.

1. We can say that it has somewhat normal distribution curve but it is right skewed. Ie. It has long tail on the right side with few large extreme values that are spread out on the right side.
2. Also the mean is pulled towards the right side. Ie means is greater than median.

Inferences from boxplot

1. The data is skewed to the right. Ie. Mean is greater than the median.

**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?

**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?

Sol.

1. Mean = sum of all/ no. of occurrences = 41
2. Median = 40+41/2 = 81/2 = 40.5
3. Variance = 25.5294
4. Standard Deviation = 5.0527

Since mean > median, distribution of student marks is right skewed. Ie. Most of his scores are less than the mean. A score which is very high compared to mean also exists(ie. large outlier).

Variance is high (25) ie the scores of the student are very spread out from the mean.

Since the SD is 5 it means that the marks of the student are very spread out from the average mark. SD up to 3 is desirable.

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

It has zero skewness . ie. The distribution of data is symmetrical.

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

The data is right skewed. Ie more data is concentrated on the left side with few extreme values on the right side.

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

The data is left skewed. i.e more data is concentrated on the right side with few small values on the left side.

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

Positive kurtosis value indicates that distribution is peaked with heavy tails ie it has outliers on tails.

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

Negative kurtosis value indicates light tails in distribution i.e lack of outliers.

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



What can we say about the distribution of the data?

Ans . The distribution of data is not normal.

What is nature of skewness of the data?

Ans. This shows a negatively skewed data where the median has shifted to the right.

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

IQR = Q3-Q1

= 18.2 – 10

= 7.8  
  
Q19) Comment on the below Boxplot visualizations?



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

Sol. The Boxplot 1 is very slightly right skewed.

The Boxplot 2 is divided into 2 equal halves by the median therefore we can say that it shows normal distribution.

In Boxplot 1 the data is concentrated around the median and is not very much spread out from the median. The range of data is very small. Whereas in Boxplot 2 data is spread out from the median and the range of data is very large.

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

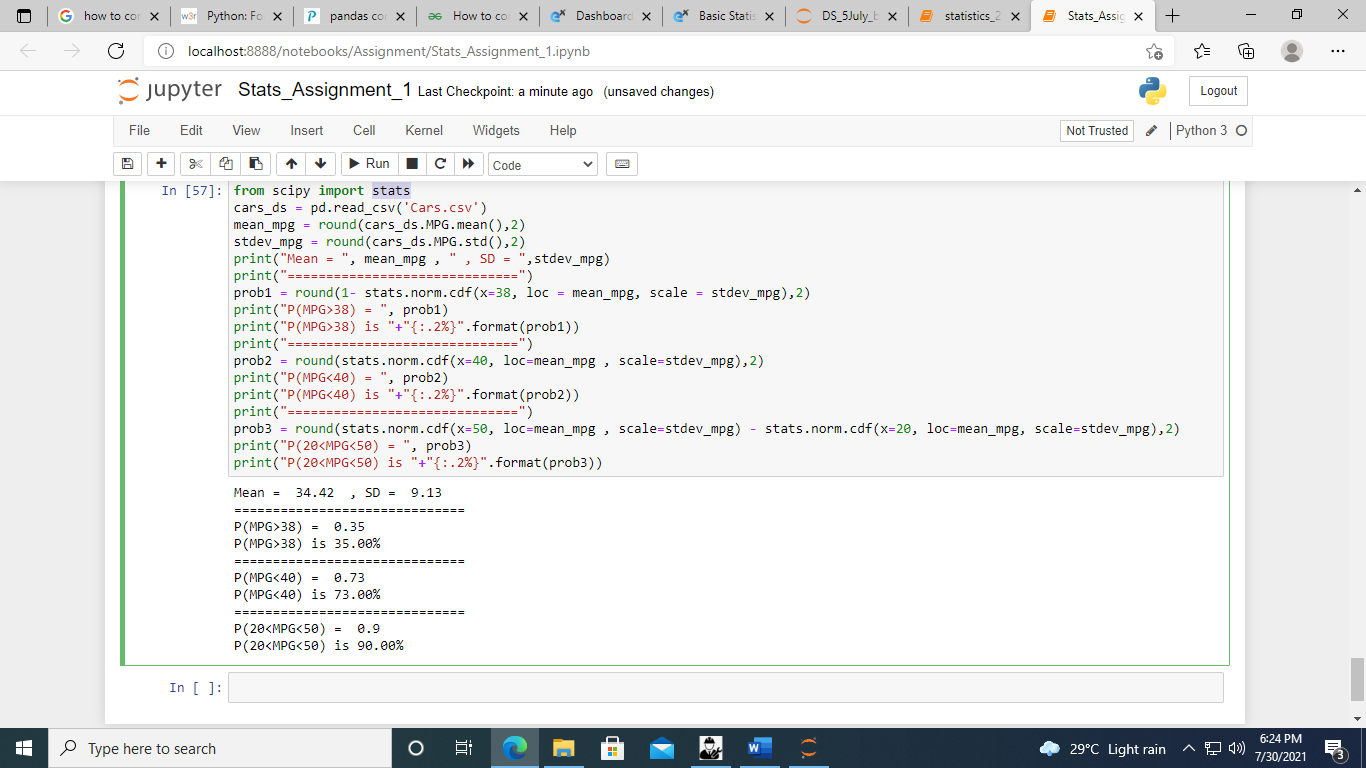
Data \_set: Cars.csv

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

MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)

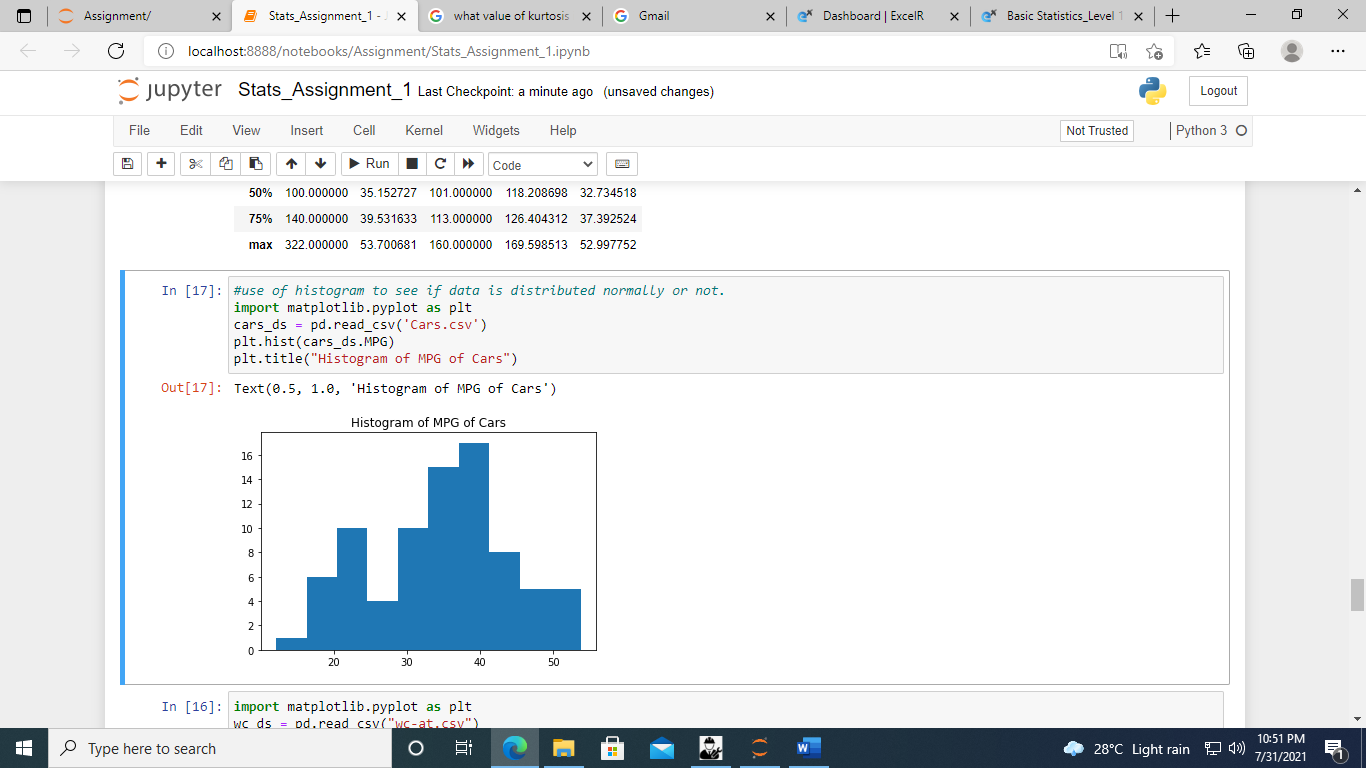
c. P (20<MPG<50)



Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

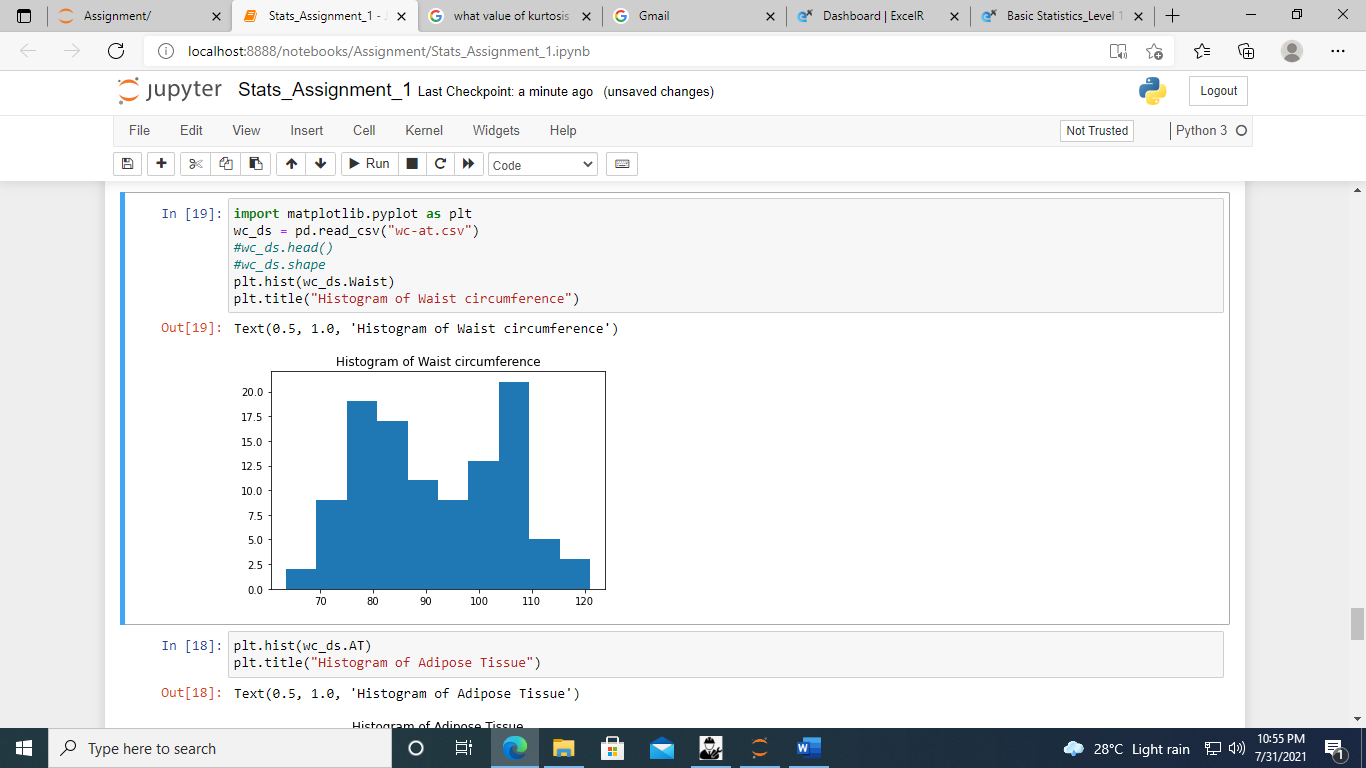
Dataset: Cars.csv



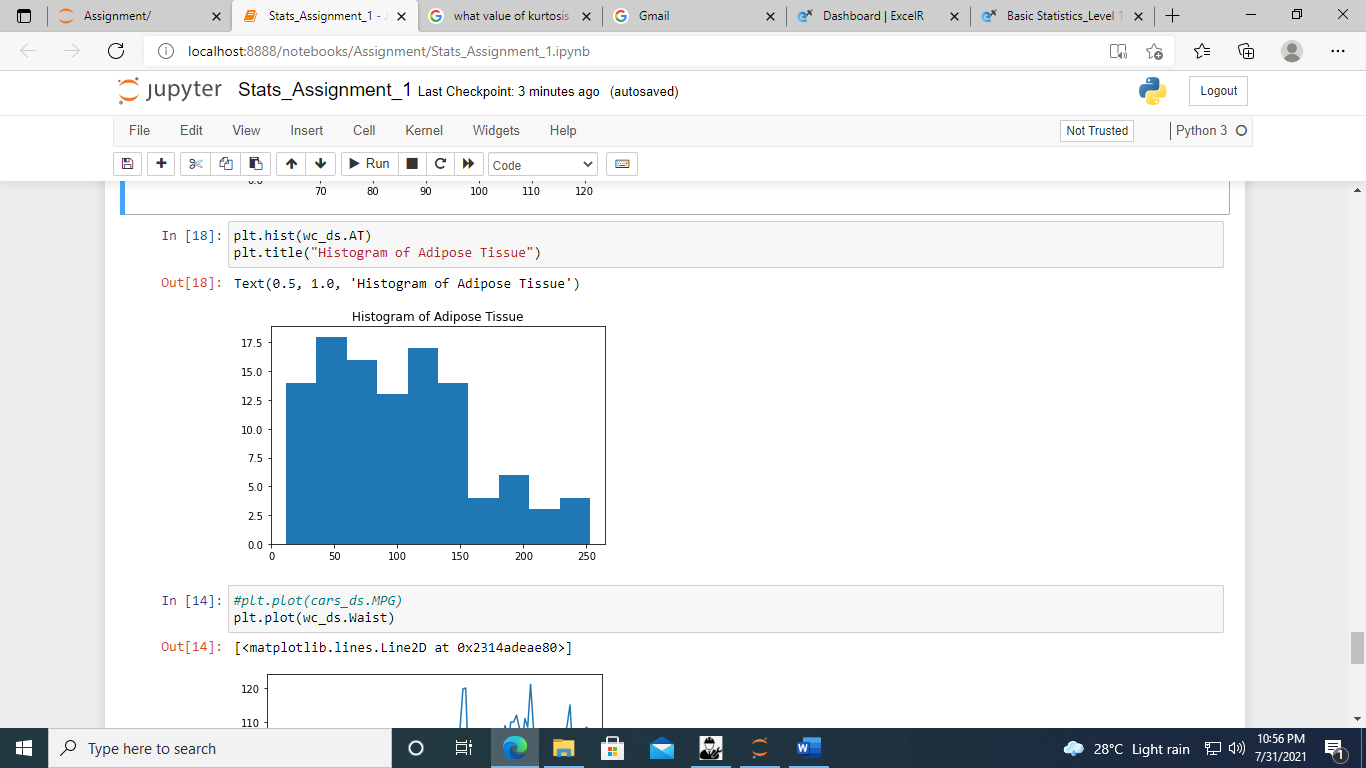
Ans. As we can see from the histogram that MPG of Cars is not similar to the bell curve. Therefore, we can say that the MPG of cars does not follow Normal Distribution.

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

Dataset: wc-at.csv

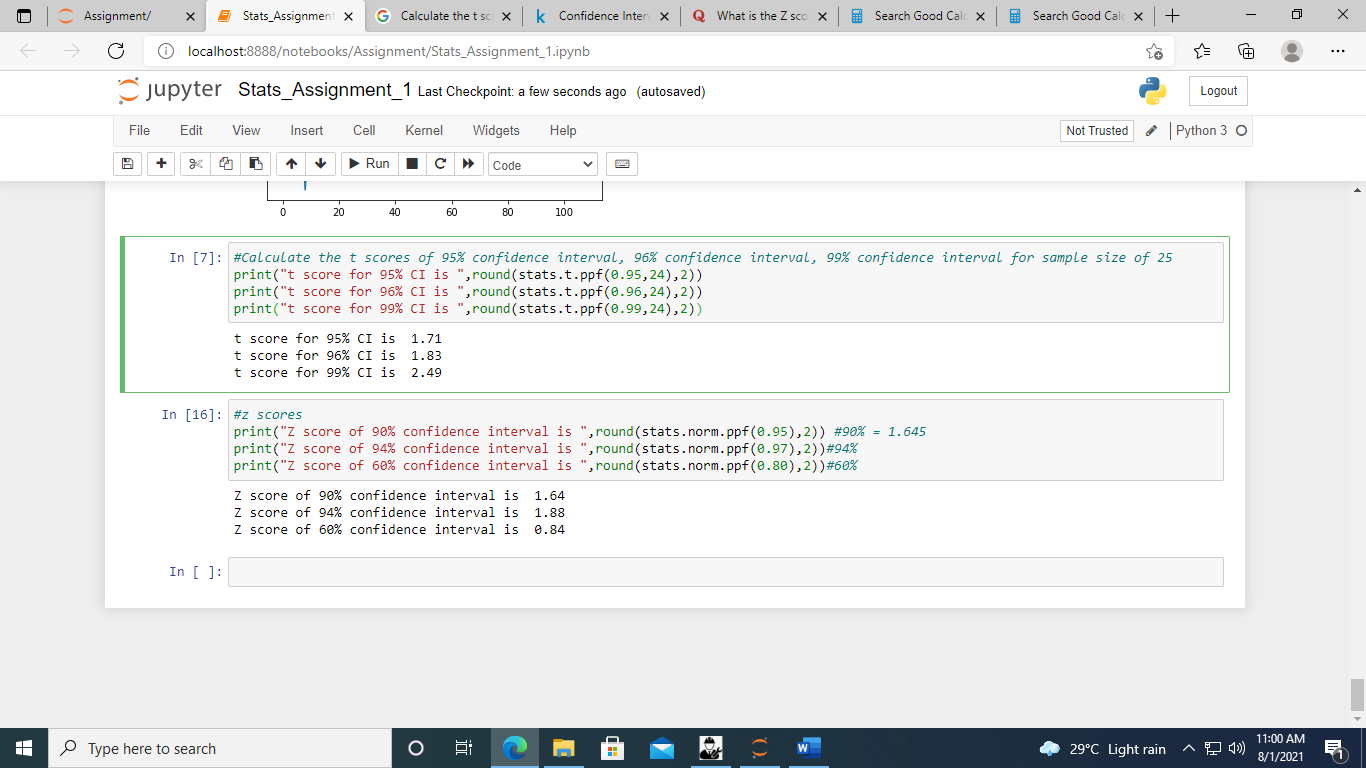


Ans. As we can see from histogram of waist circumference that it is not similar to bell curve for normal Distribution. Therefore, we can say that waist circumference data does not follow ND.



Ans. As we can see from histogram of Adipose tissue that it is not similar to bell curve for normal Distribution. Therefore, we can say that adipose tissue data does not follow ND.

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



A confidence of 90% interval means

it has 2 tails of 10/2 = 5%, so it goes from 5% to 95%

Therefore, Z critical = 1.64

A confidence of 94% interval means

It has 2 tails of 6/2 = 3%, so it goes from 3% to 97%

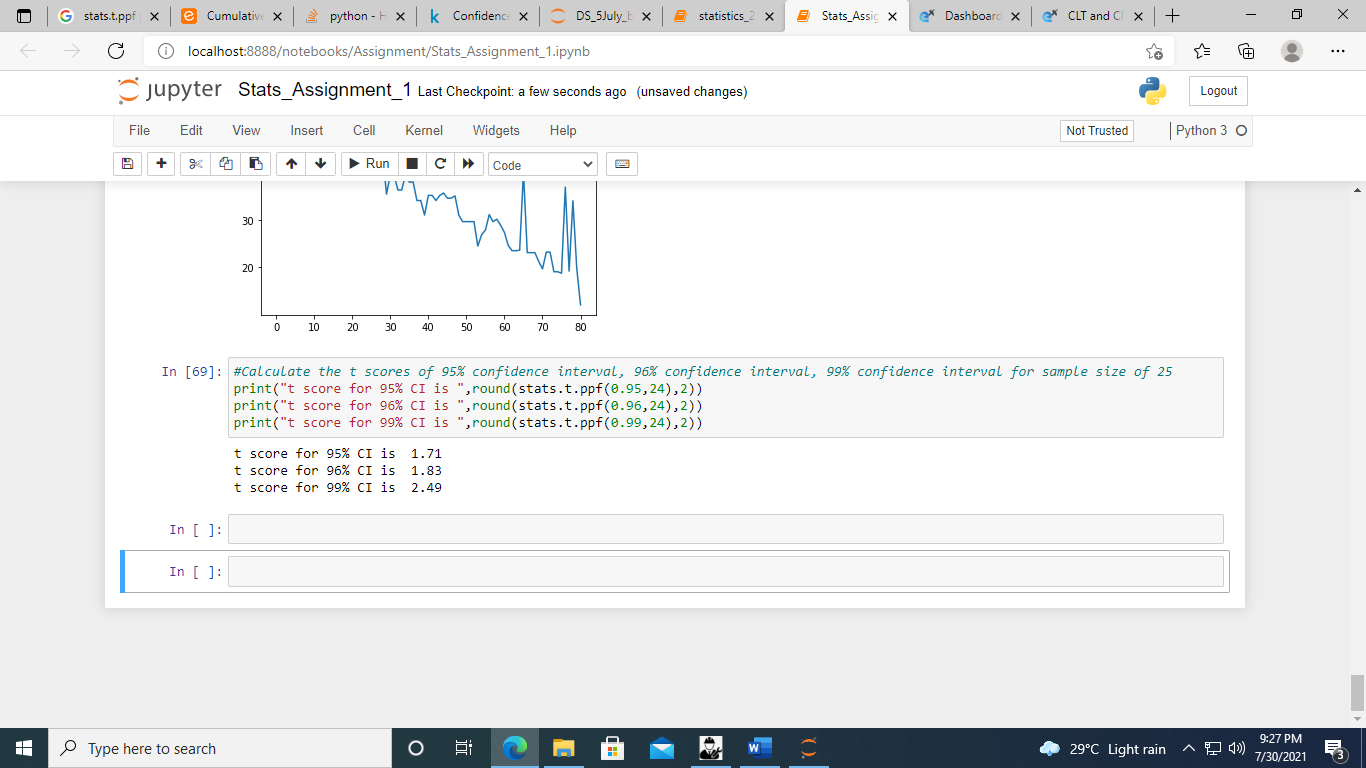
Therefore, Z-critical = 1.88

A confidence of 60% interval means

It has 2 tails for 40/2 = 20%, so it goes from 20% to 80%

Therefore, z-critical = 0.84

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



Degree of freedom = sample size -1

= 25-1

=24

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

Sol.

Population mean = 270

sample size = 18

sample mean = 260

sample SD = 90 days

Degree of freedom (df) = 18-1 =17

t = (sample mean) – (population mean) / sample SD/square root (sample size)

= 260 -270 / 90/4.24

= -10 / 21.23

t score = - 0.47

The probability of 18 randomly selected bulbs would have an average life of no more than 260 days is **32 %.** (calculated through online calculator by passing t score and df “[Quick P Value from T Score Calculator (socscistatistics.com)](https://www.socscistatistics.com/pvalues/tdistribution.aspx)”)

#online calculator for probability using t-score and df(degree of freedom)

#https://www.socscistatistics.com/pvalues/tdistribution.aspx

#we can calculate probability from tscore and df.

from scipy import stats

stats.t.sf(abs(-0.47), df=17)