|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete Data (Numerical – Quantitative) |
| Results of rolling a dice | Discrete Data (Numerical – Quantitative) |
| Weight of a person | Continuous Data (Numerical – Quantitative) |
| Weight of Gold | Continuous Data (Numerical – Quantitative) |
| Distance between two places | Continuous Data (Numerical – Quantitative) |
| Length of a leaf | Continuous Data (Numerical – Quantitative) |
| Dog's weight | Continuous Data (Numerical – Quantitative) |
| Blue Color | Nominal Data (Categorical – Qualitative) |
| Number of kids | Discrete Data (Numerical – Quantitative) |
| Number of tickets in Indian railways | Discrete Data (Numerical – Quantitative) |
| Number of times married | Discrete Data (Numerical – Quantitative) |
| Gender (Male or Female) | Nominal Data (Categorical – Qualitative) |

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

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

1. HHH, HHT, HTH, THH, TTT, TTH, THT, HTT

Total No. of Possible Outcomes = 8

Event of 2 H and 1T = HHT, HTH, THH = 3 outcomes out of 8

i.e.; Probability = 3/8

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

1. Equal to 1
2. Probability = 0; as the minimum sum will be 2
3. Less than or equal to 4
4. Total no. of outcomes = 62 = 36

Event of <=4 is = (1,1), (1,2), (1,3), (2,1), (2,2), (3,1) = 6 outcomes out of 36

i.e.; Probability = 6/36 = 1/6

1. Sum is divisible by 2 and 3
2. Total no. of outcomes = 62 = 36

Event of Sum is divisible by 2 and 3 = (1,5), (2,4), (3,3), (4,2), (5,1), (6,6) = 6 outcomes out of 36

i.e.; Probability = 6/36 = 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?

1. RR GGG BB

Total number of balls = 7

Total no. of blue balls = 2

The probability of drawing non-blue balls in the 1st draw, P 1st Draw = 5/7

Total no. of balls left = 6

The probability of drawing non-blue balls in the 2nd draw, P 2nd Draw = 4/6 = 2/3

The probability that none of the balls drawn is blue =P 1st Draw x P 2nd Draw

=5/7 x 2/3 = 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-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



Expected Value = ƩPiXi

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies Count  a (Xi) | Probability  b (Pi) | Weight  a x b (Pi x Xi) |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.20 | 0.8 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.120 | 0.24 |
| Total |  | 1 | **Mean, x̅ = 3.09** |

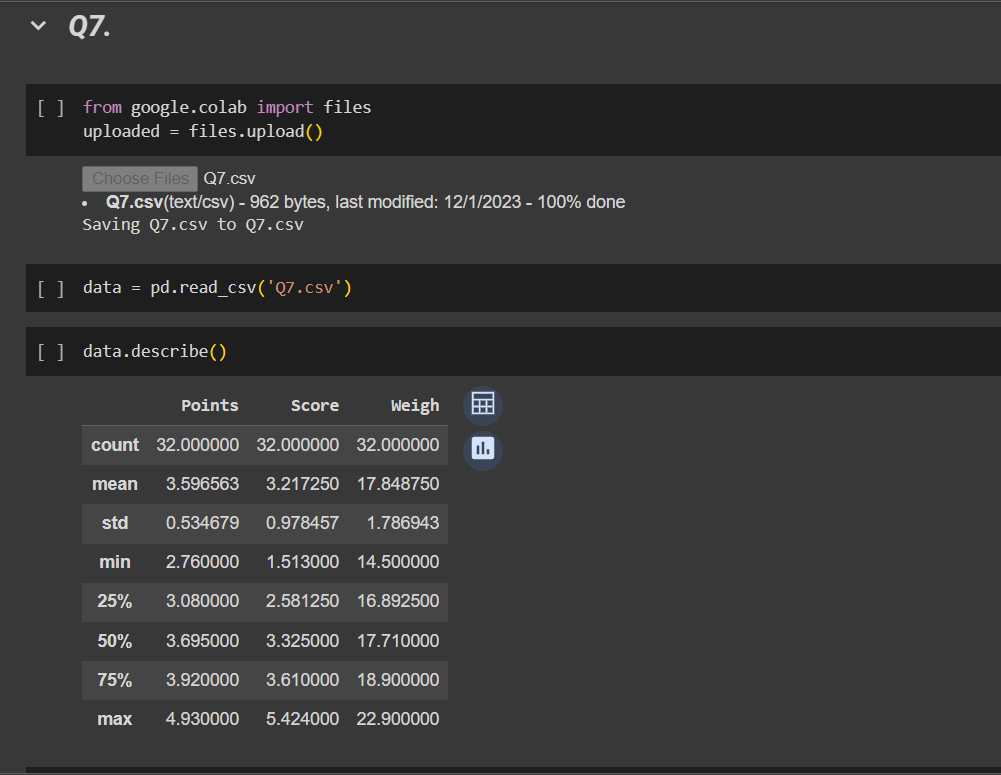
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

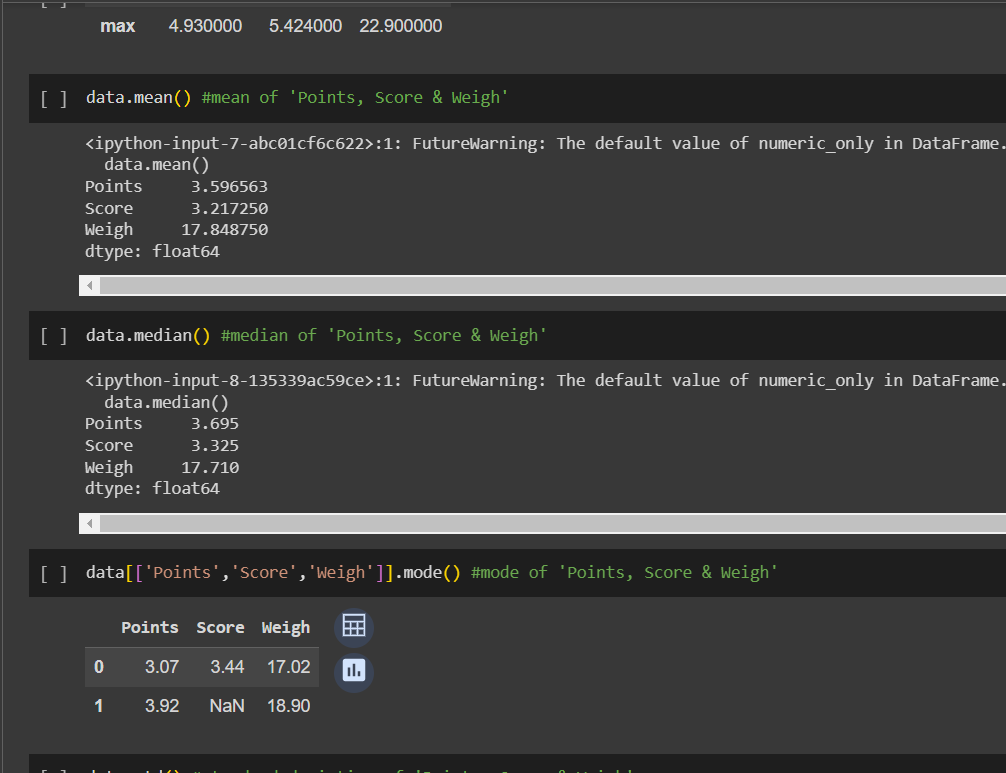
* For Points, Score, Weigh>

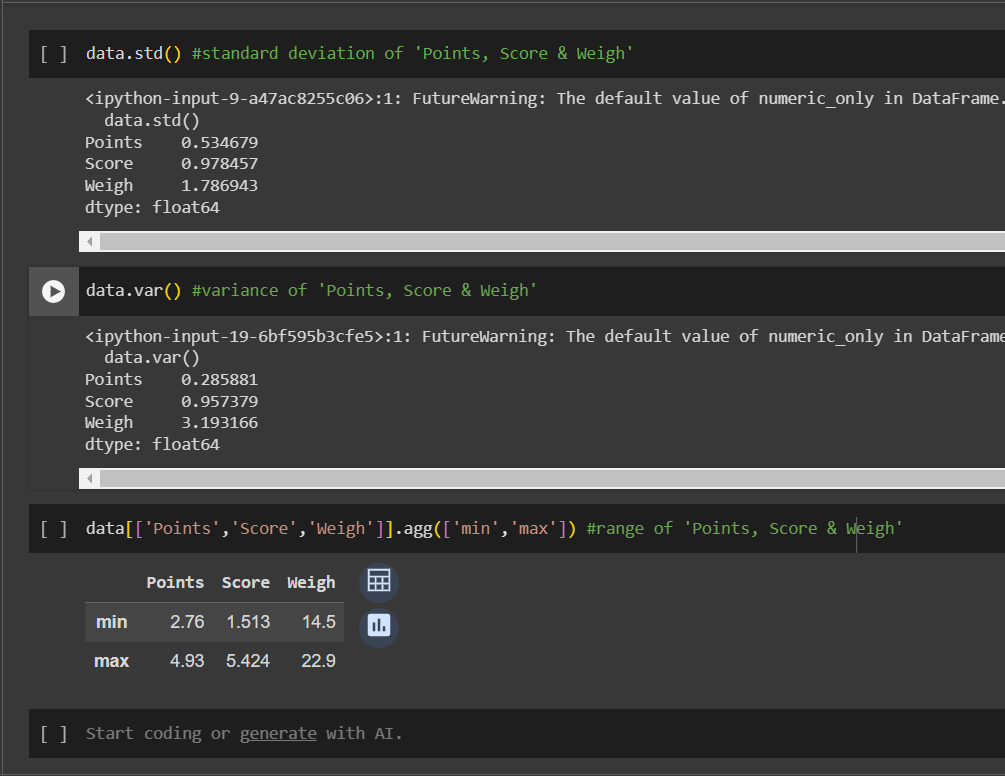
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

1. Please find the attached screenshots for the Python codes and my inference derived from the descriptive statistics.







From the above-derived values, the following are my inferences –

1. Points –
   1. The mean and the median values are close, therefore it’s a roughly symmetric distribution. (as mean<median marginally, it hints towards left-tailed data – negative skewness)
   2. As there are 2 modes, it’s a bimodal distribution
   3. The standard deviation is relatively low, indicating moderate variability.
2. Score –
   1. The mean and the median values are close, therefore it’s a roughly symmetric distribution. (as mean<median marginally, it hints towards left-tailed data – negative skewness)
   2. It has a single mode, it’s an unimodal distribution
   3. The standard deviation is relatively higher, indicating higher variability.
3. Weigh –
   1. The mean and the median values are close, therefore it’s a roughly symmetric distribution. (as mean>median marginally, it hints towards right-tailed data – positive skewness)
   2. As there are 2 modes, it’s a bimodal distribution.
   3. The standard deviation is relatively higher, indicating higher variability.

Q8) Calculate the 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?

1. Expected value (a.k.a) Mean = Ʃxi / n

n = 9

Ʃxi/n, x̅ = (108+110+123+134+135+145+167+187+199)/9

= 1308 / 9

∴**mean, x̅ = 145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

1. Please find the attached screenshots for the Python codes and my inference derived from them.

From the above-derived values, the following are my inferences –

Distance –

1. The mean is greater than the median values, therefore it’s right-tailed data.
2. Also, the skewness value is 0.806. Hence, it’s a positively skewed data
3. The kurtosis value is less than 3 (0.405) meaning it is platykurtic or negative kurtosis.
4. This means that the distribution has a flatter peak with thinner tails, i.e., the data is more spread out

Speed –

1. There is not much difference between the mean and the median values (mean = 15.4 and median = 15).
2. The skewness value is -0.117. Hence, it’s slightly negatively skewed data.
3. The kurtosis value is less than 3 (-0.508) meaning it is platykurtic or negative kurtosis.
4. This means that the distribution has a flatter peak with thinner tails, i.e., the data is more spread out



**SP and Weight (WT)**

**Use Q9\_b.csv**

1. Please find the attached screenshots for the Python codes and my inference derived from them.

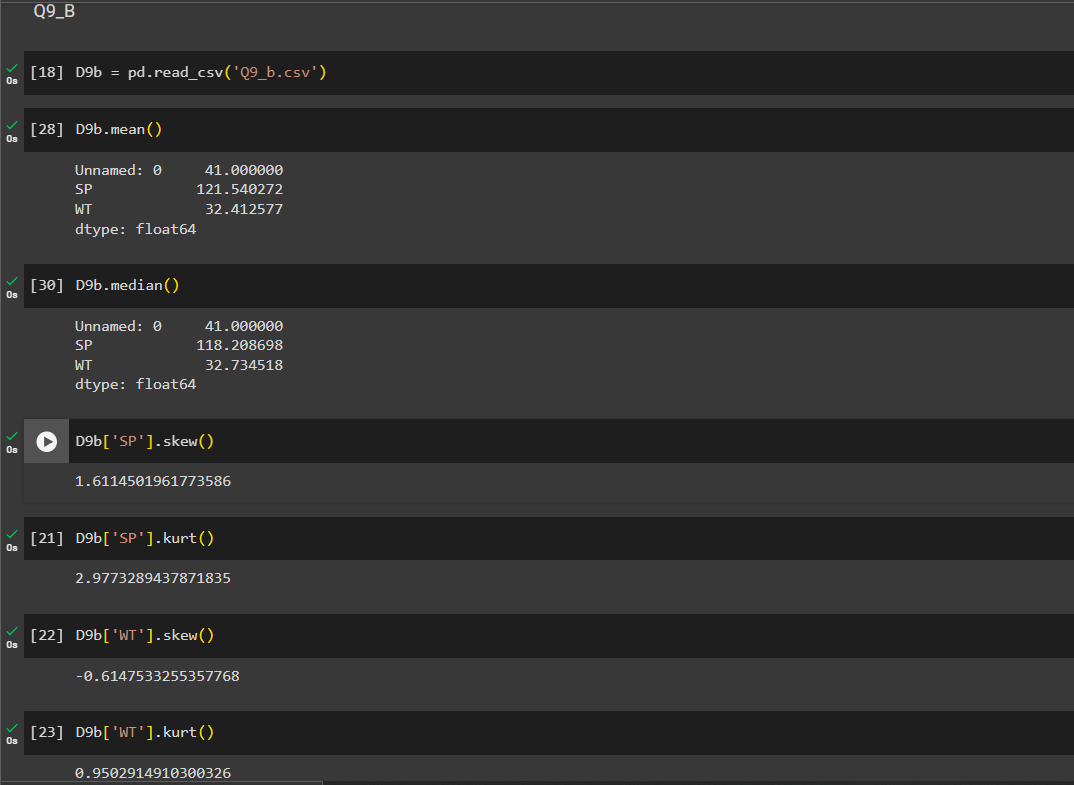
From the below-derived values, the following are my inferences –

SP –

1. The mean is greater than the median values, therefore it’s right-tailed data. (mean = 121.5 and median = 118.2)
2. Also, the skewness value is 1.611. Hence, it’s a positively skewed data
3. The kurtosis value is marginally less than 3 (2.977) meaning it is tending towards platykurtic or negative kurtosis.
4. This means that the distribution has a slightly flatter peak with thinner tails, i.e., the data is a little spread out

Weight –

1. There is not much difference between the mean and the median values (mean = 32.4 and median = 32.7).
2. The skewness value is -0.614. Hence, it’s slightly negatively skewed data.
3. The kurtosis value is less than 3 (0.950) meaning it is platykurtic or negative kurtosis.
4. This means that the distribution has a flatter peak with thinner tails, i.e., the data is more spread out



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





1. Following are the inferences –
   1. This Histogram is right skewed.
   2. The skewness value is positive.
   3. The number of observations on the right side is less.
   4. This data doesn't follow a normal distribution.
   5. There are multiple outliers beyond the upper quartile value.

**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?

1. Sample size (n) = 2000

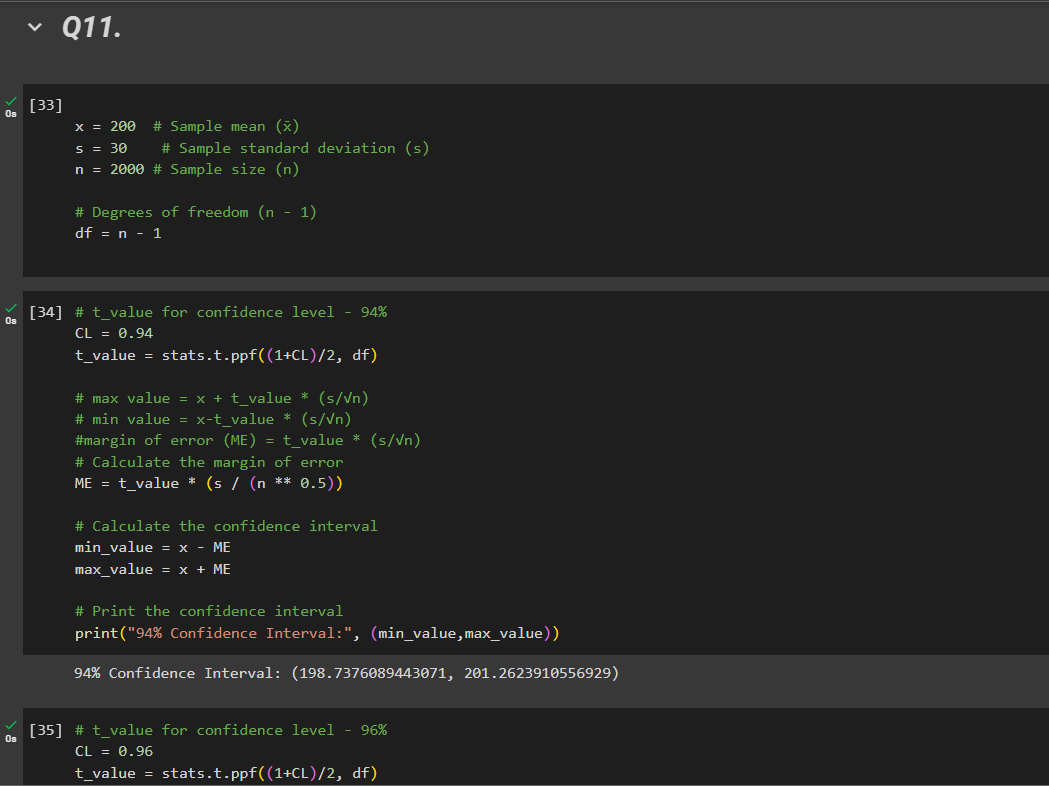
Population size (N) = 3,000,000

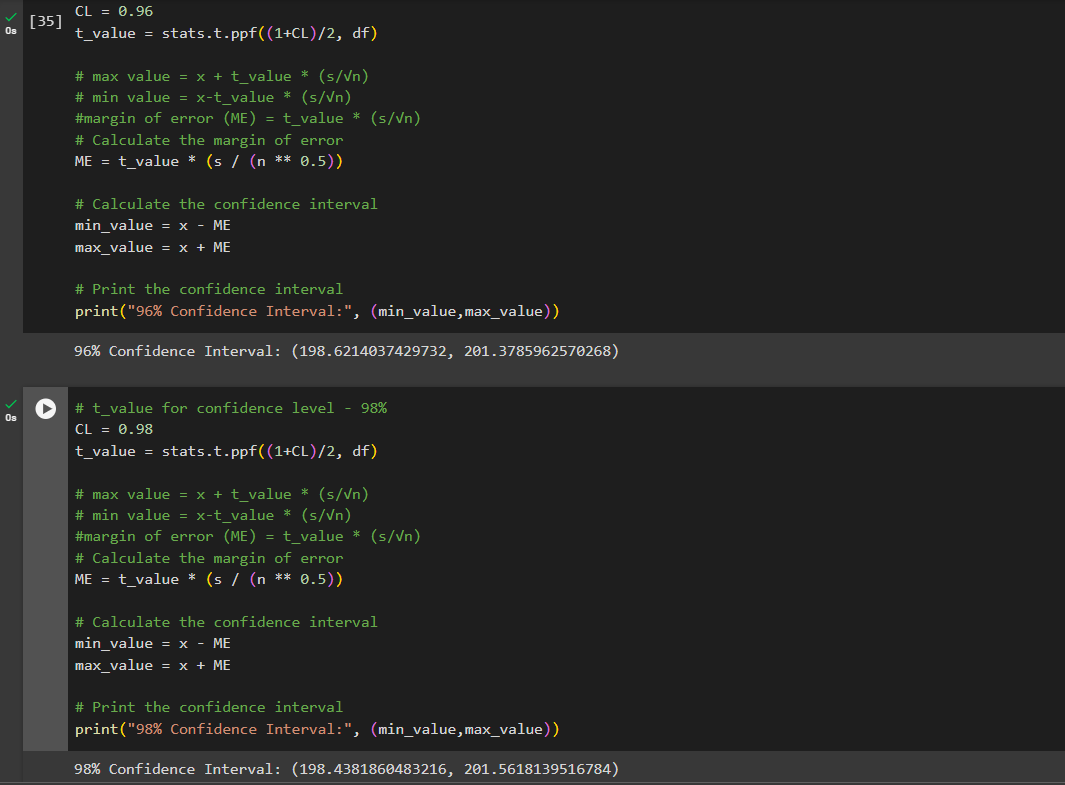
Mean weight (x̅) = 200

Standard deviation (σ) = 30

Confidence levels (CL) = 94%, 98%, 96%

Please find the attached screenshots for the Python codes for confidence interval calculations –





∴ 94% Confidence interval = 198.737 – 201.262

∴ 96% Confidence interval = 198.621 – 201.378

∴ 98% Confidence interval = 198.438 – 201.561

**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?
3. Please find the attached screenshots for the Python codes and my inference derived from them.

From the below-derived values, the following are my inferences –

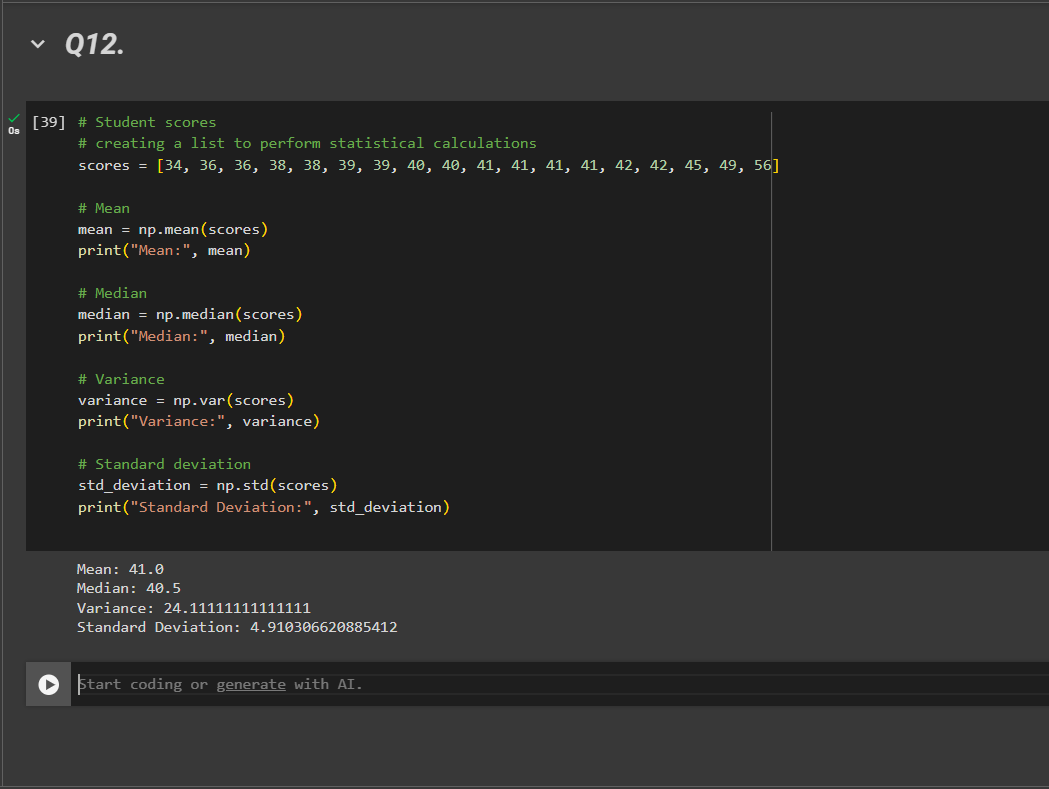
∴ Mean = 41.0

∴ Median = 40.5

∴ Variance = 24.11

∴ Standard Deviation = 4.91

1. On average, the student scored 41 marks in the tests
2. 40.5 is the middle value, i.e., half the scores are above the value and half the scores are below the value.
3. The median value is close to the mean, indicating a symmetric distribution of scores.
4. The variance and standard deviation values indicate the variability in the score, i.e., the scores are more spread-out ranging from 34 to 56.

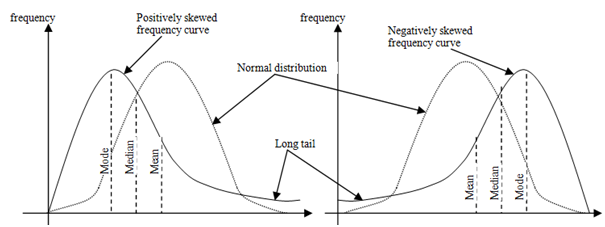


Q13) What is the nature of skewness when mean, median of data are equal?

1. If the mean and median are equal, it suggests that the data is symmetrically distributed around the center. The skewness for symmetric distribution is 0.

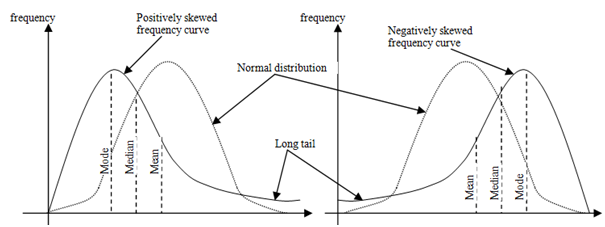
Q14) What is the nature of skewness when mean > median?

1. If the mean is greater than the median, it suggests that the data is concentrated towards the left side of the distribution with very large values &/ outliers stretching out the right tail. It suggests that the distribution is positively skewed and the distribution is not symmetric.



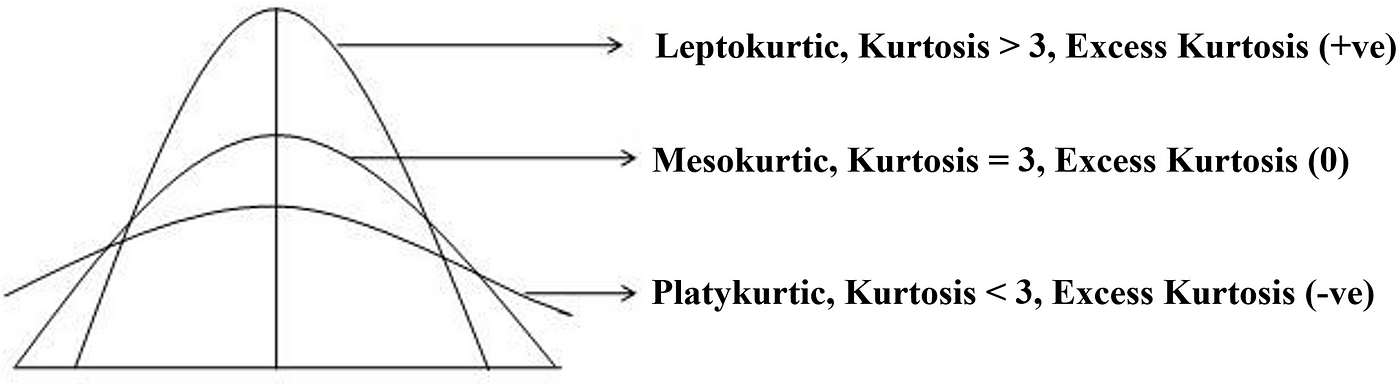
Q15) What is the nature of skewness when median > mean?

1. If the median is greater than the mean, it suggests that the data is concentrated towards the right side of the distribution with very small values &/ outliers stretching out the left tail. It suggests that the distribution is negatively skewed and the distribution is not symmetric.



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

1. This distribution has sharper peaks and fatter tails than the normal distribution. This distribution is referred to as leptokurtic.
   1. Fatter tails indicate more frequent extreme values.
   2. As the peak is higher and sharper, more data is concentrated around the mean



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

1. This distribution has flatter peaks and thinner tails than the normal distribution. This distribution is referred to as platykurtic.
   1. Thinner tails indicate less frequent extreme values.
   2. As the peak is lower and broader, the data is more spread out.

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)?

1. The distribution of the data is skewed.
   1. The nature of skewness is left skewed, i.e., there is a longer tail on the left-hand side
   2. Based on the above image Q1=10 and Q3=18. ∴ IQR = Q3-Q1, i.e., 8

Q19) Comment on the below Boxplot visualizations?



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

1. The following are the inferences from the distribution of data from boxplot 1 w.r.t. boxplot 2
   1. The medians of both the boxplots are similar.
   2. The IQR for boxplot 2 is greater than boxplot 1, i.e., the data is more spread out in 2 as compared to 1.
   3. Range of values for boxplot 2 is greater than 1 as it has a smaller min value and larger max value.
   4. Neither of the box plots have any outliers.
   5. Both the box plots are following normal distribution as the distributions are symmetric around the center and both the side tails 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.

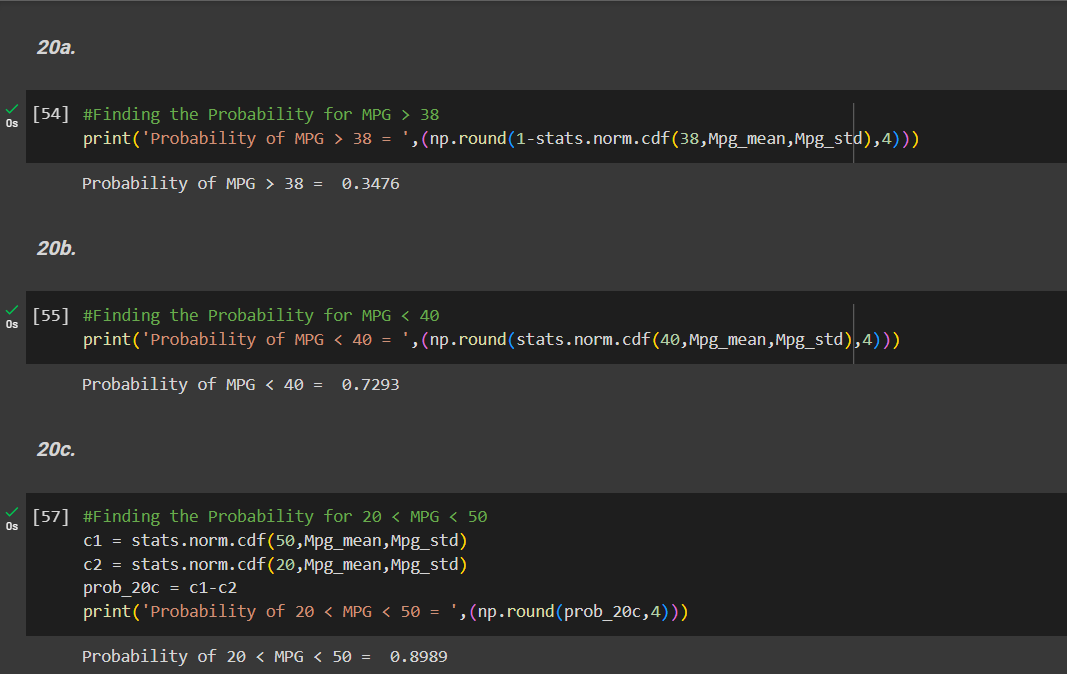
MPG <- Cars$MPG

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

c. P (20<MPG<50)

1. The following are the probabilities using python code –





Q 21) Check whether the data follows normal distribution

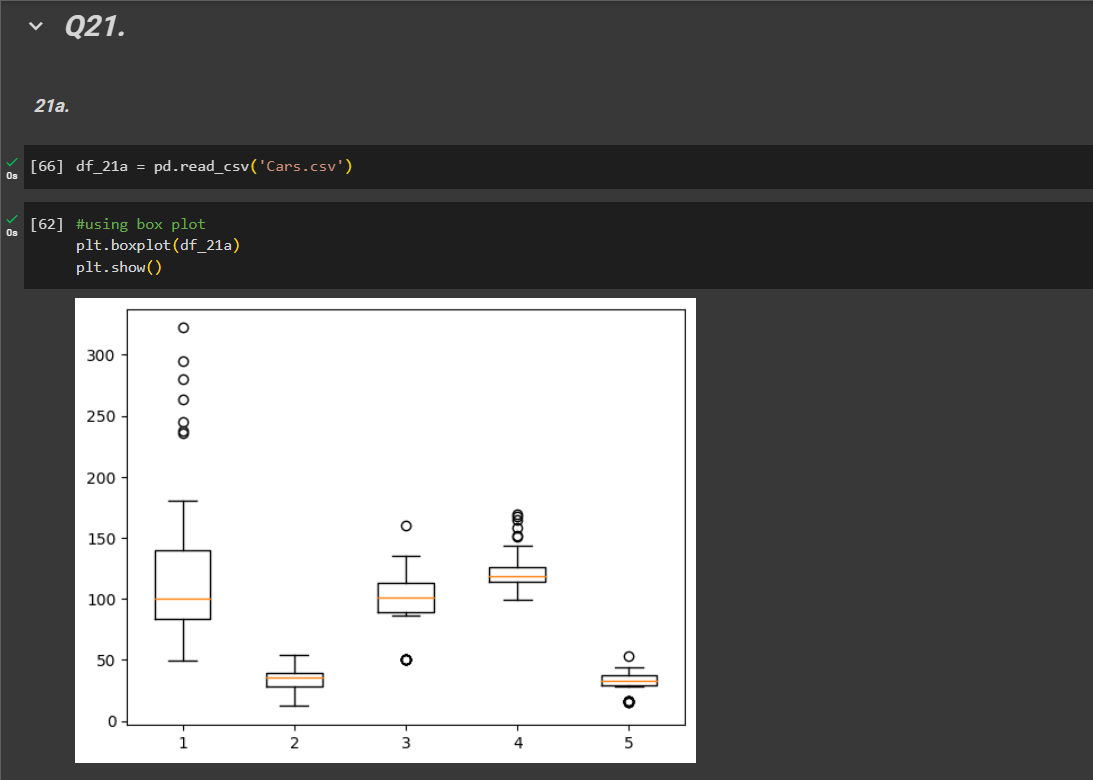
1. Check whether the MPG of Cars follows Normal Distribution

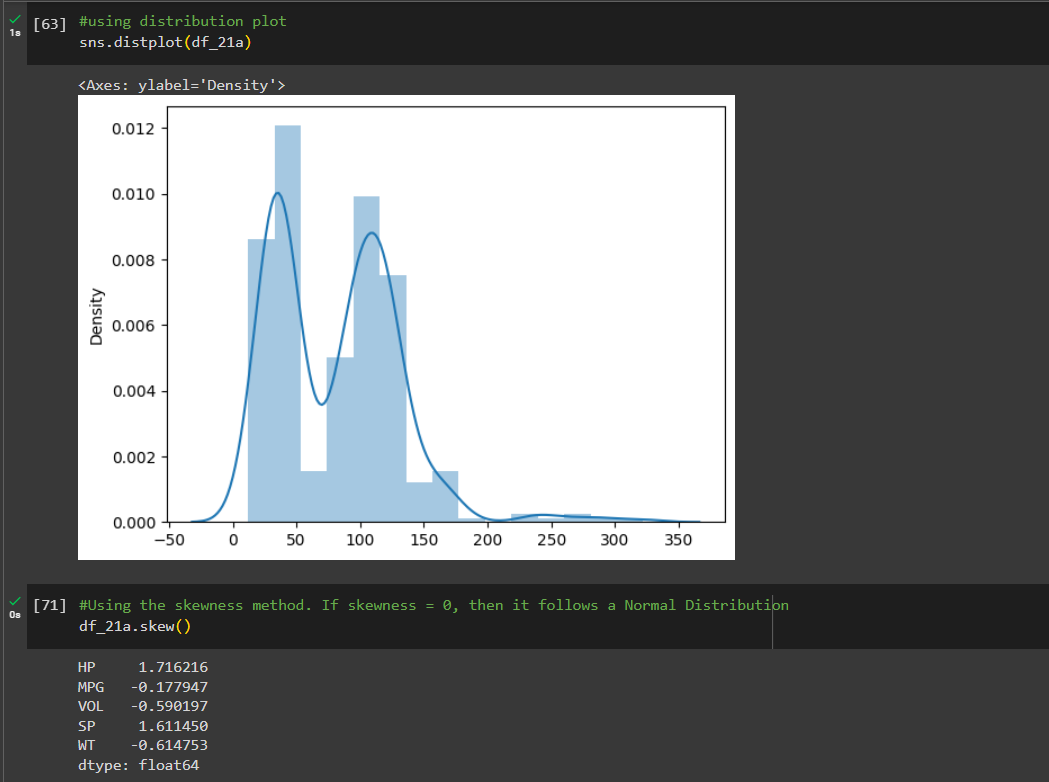
Dataset: Cars.csv

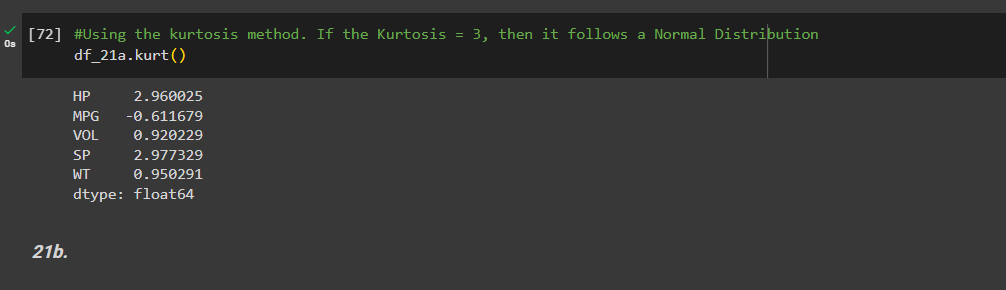
1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

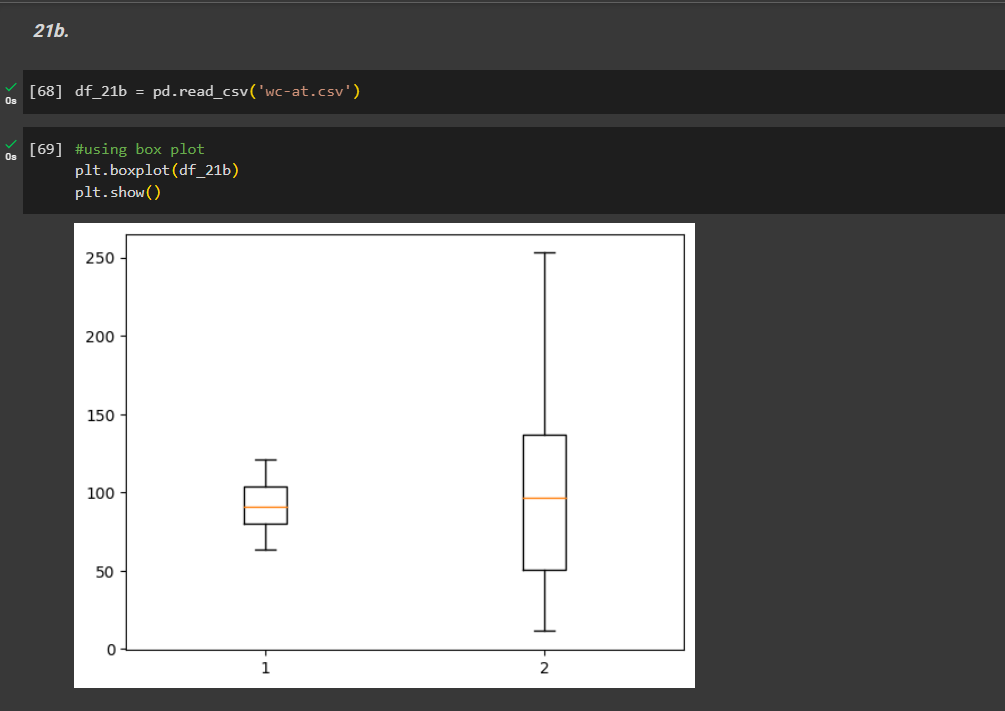
Dataset: wc-at.csv

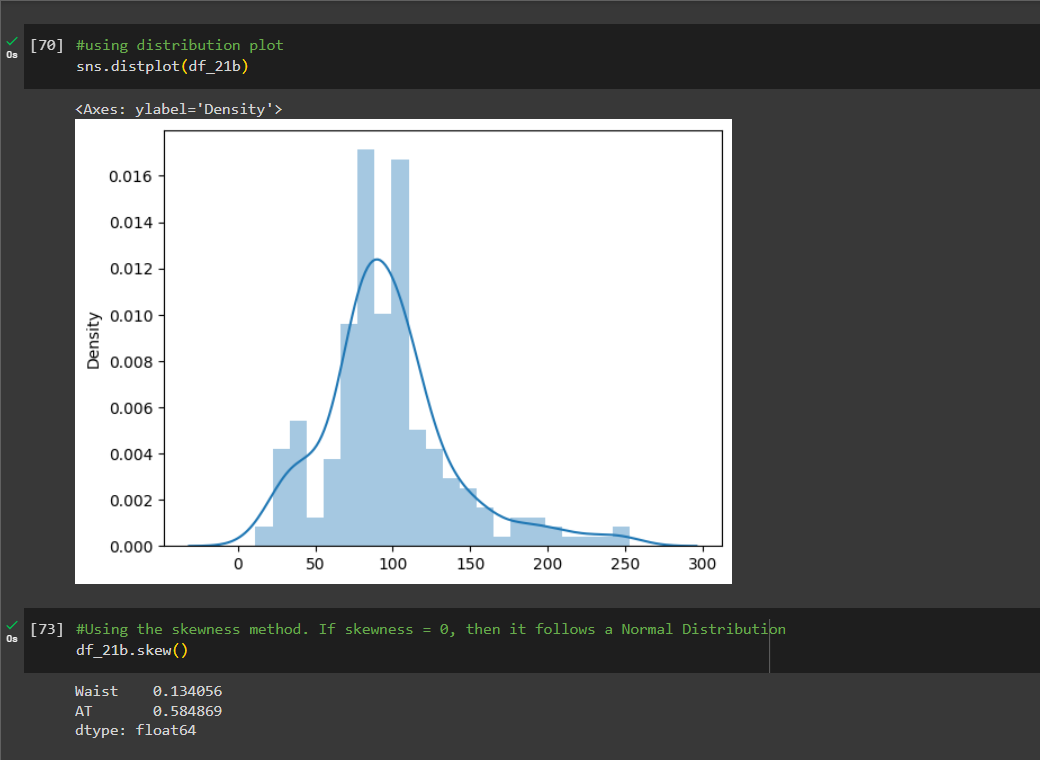
1. Please find the attached screenshots for the Python codes and my inference derived from them.

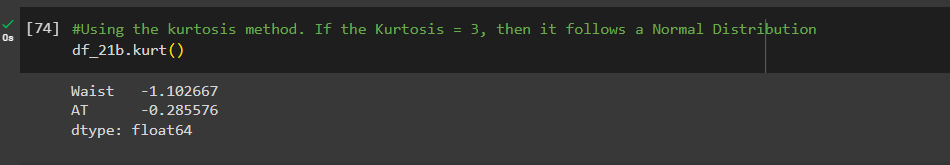










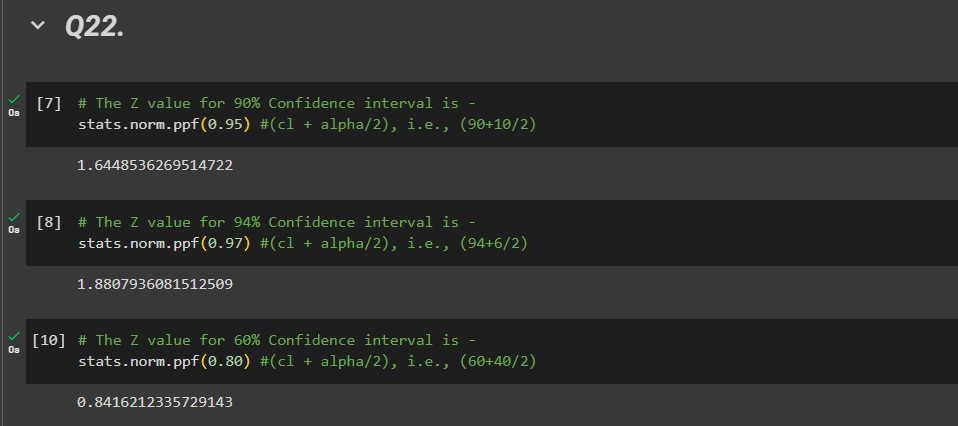


1. Dataset A – Cars.csv
   1. Looking at the box plots and the distribution plot we understand that the dataset is not following a normal distribution.
   2. The skewness values are all either greater than 0 or below 0. Again this indicates that it doesn’t follow normal distribution
   3. Similarly, the kurtosis values are all below 3. Stating that they are Platykurtic. HP and SP columns are close to 3 (2.9) but the dataset overall doesn’t follow normal distribution.
2. Dataset B – wc-at.csv
   1. Looking at the box plots and the distribution plot we understand that the dataset follows a somewhat normal distribution as compared to cars dataset.
   2. The skewness values are slightly higher than 0.
   3. The kurtosis values are below 3. Stating that they are platykurtic.

∴Both the distributions do not follow normal distribution.

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

1. The Z-score value python codes are as follows –

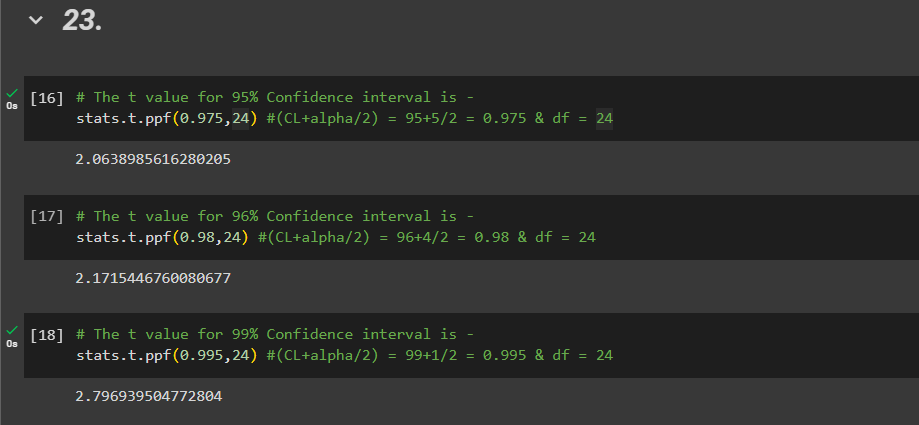


Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

1. The t-score value python codes are as follows –

n (sample size) = 25

degrees of freedom = n - 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

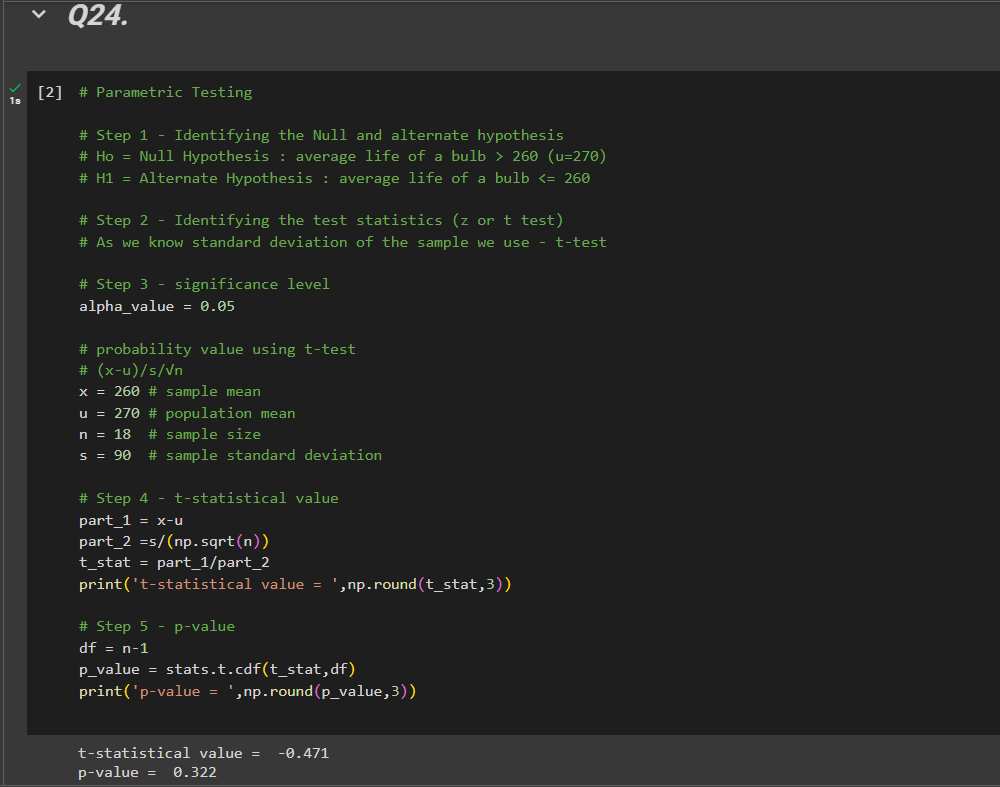
1. Sample size (n) = 18

Sample Mean (x̅) = 260

Population Mean (μ) = 270

Sample Standard deviation (s) = 90

The python codes are as follows –



As p-value > 0.05, we accept the null hypothesis, i.e., μ>270.

∴, Probability of the sample bulb with μ <=260 = 0.322