|  |  |
| --- | --- |
| 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 | Continuous |
| 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 | Interval |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Interval |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Ratio |
| Number of Children | Interval |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Interval |

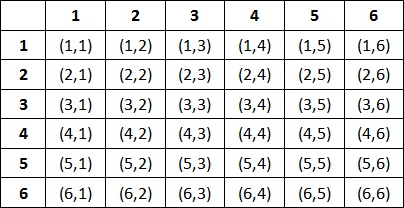
Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

**Answer**- Probability for 2 Heads & 1 Tail = 3/8=**0.375**= {HHT, THH, HTH}

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

**Answer-**



1. Equal to 1

* Probability = **0**

1. Less than or equal to 4

* (1,1), (1,2), (1,3), (2,1), (2,2), (3,1) = 6/36=**0.167**

1. Sum is divisible by 2 and 3

* (1,5), (2,4), (3,3), (4,2), (5,1), (6,6) =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?

**Answer-**

N(S)=Total No. of ways of drawing 2 balls at a time from 7 balls=21

N(B)=No. of ways of drawing 2 balls, none of which is blue=10

P(B)=[N(B)/N(S)]

=10/21

= 0.476

Therefore**,** probability that none of the balls drawn is blue is **0.476**

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

**Answer-**

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

= 3.090

**= 3.09**

Therefore**,** expected number of candies for a randomly selected child are **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**

**Answer-**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **Mean** | 3.5966 | 3.2173 | 17.8488 |
| **Mode** | 3.92 | 3.44 | 17.02 |
| **Median** | 3.695 | 3.325 | 17.71 |
| **Variance** | 0.2769 | 0.9275 | 3.0934 |
| **Standard Deviation** | 0.5263 | 0.9630 | 1.7588 |
| **Range** | 2.17 | 3.911 | 8.4 |

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?

**Answers-**

Expected Value = ∑ (probability \* Value)

Probability of selecting each patient = 1/9

Expected Value = (1/9) (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

=145.33

Therefore, the expected value of the weight of that randomly selected patient is **145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Answer 9\_a=**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Index** | **speed** | **dist** |
| **Skewness** | -3.78E-17 | -0.11751 | 0.806895 |
| **Kurtosis** | -1.2 | -0.50899 | 0.405053 |

Skewness= From the above table we can say that, ‘Index’ and ‘dist’ is positively skewed whereas ‘speed’ is negatively skewed.

Kurtosis= From the above table we can say, kurtosis of ‘dist’ has Heavier tail than ‘Index’ and ‘speed’. Whereas, the tail of ‘speed’ is light and that of ‘Index’ is lightest.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Answer 9\_b=**

|  |  |  |
| --- | --- | --- |
|  | **SP** | **WT** |
| **Skewness** | 1.61145 | -0.61475 |
| **Kurtosis** | 2.977329 | 0.950291 |

Skewness= From the above table we can say that, ’SP’ is positively skewed where as WT is negatively skewed

Kurtosis= From the above table we can say, Kurtosis of ‘SP’ has heavier tail than ‘WT’ which has comparatively lighter tail

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



**Answer –**

a**.** From the above picture we can say, that around 190 chicks have weight in between 50 to 100.

b. Least frequency of chicks has the chick weight ranging between 350 to 400.

c. The distribution is right skewed or positively skewed.



**Answer –**

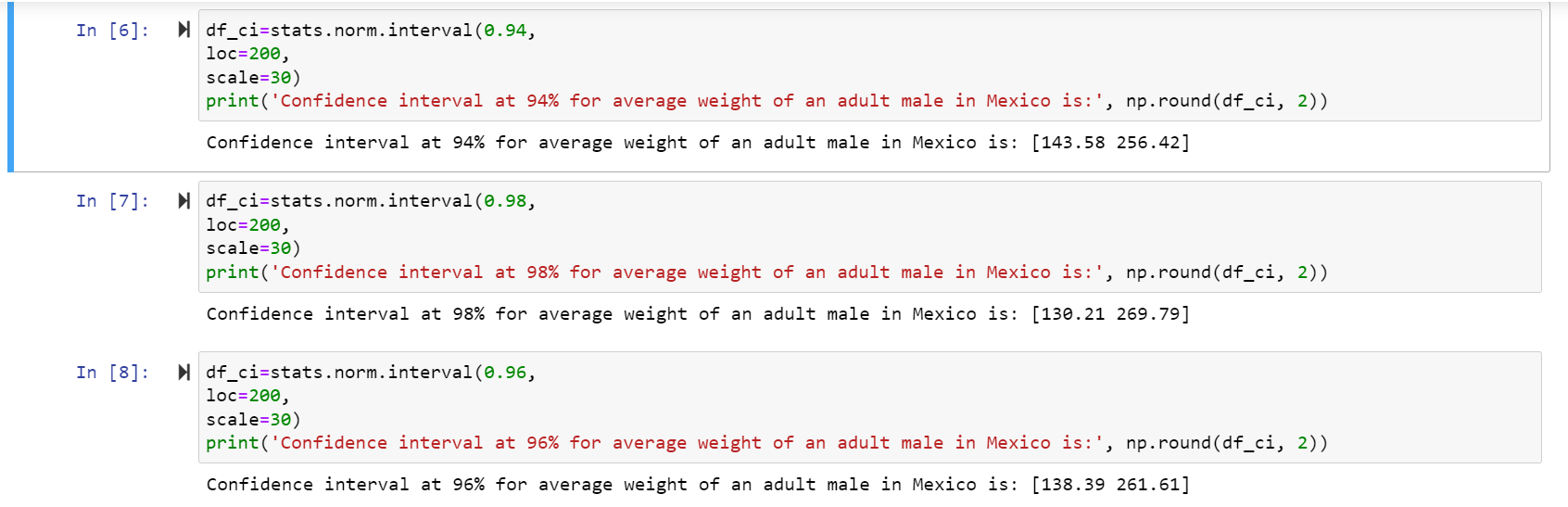
a. There are around 8 outliers in the above box plot

b. The distribution is not equal (Symmetric).

**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?

**Answer –**

1. Confidence interval at 94% for average weight of an adult male in Mexico is: [143.58 256.42]
2. Confidence interval at 98% for average weight of an adult male in Mexico is: [130.21 269.79]
3. Confidence interval at 98% for average weight of an adult male in Mexico is: [130.21 269.79]

****

**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?

**Answer –**

1.

|  |  |
| --- | --- |
| **Mean** | 41 |
| **Median** | 40.5 |
| **Variance** | 24.11111 |
| **Standard Deviation** | 4.910307 |

2.

a. The students have scored marks in between 34 and 56

b. Average marks scored by students are 41.

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

**Answer –** The Skewness is zero i.e., the distribution is symmetric.

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

**Answer –** When the value of mean is greater than median, the distribution is right skewed or positively skewed.

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

**Answer –** When the value of median is greater than median (i.e., mean is lesser than median), the distribution is left skewed or left skewed.

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

**Answer –**

a. Positive kurtosis value indicates higher presence of outliers.

b. The positive kurtosis value also indicates the distribution is peaked and has thick tails in the distribution.

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

**Answer –** The negative kurtosis value indicates the distribution is comparatively flat and has thin tails.

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)?

**Answer –**

a. The distribution of data is not normal as it is not equally distributed.

b. The data is negatively skewed or left skewed.

c. In the above box plot, Q1=10, Q3=18, Median=15

We know, IQR= Q3-Q1= 18-10

**Therefore, IQR=8.**

Q19) Comment on the below Boxplot visualizations?



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

**Answer –**

a. The data of both the box plots is normal as it is equally distributed.

b. There are no presence of outliers in both the box plots.

c. Median of both the plots are similar=262.5(approximately).

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)

**Answer –** P(MPG>38) = **0.34748702501304063**

****

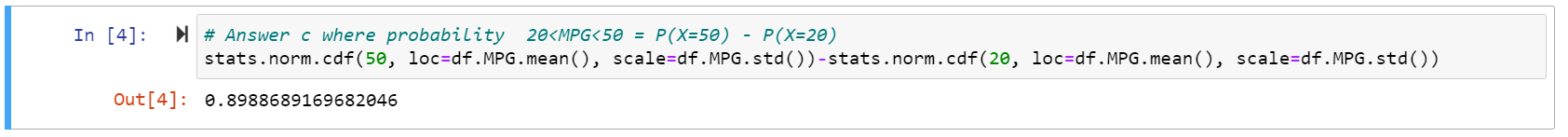
* 1. P(MPG<40)

**Answer –** P(MPG<40) = **0.7293498762151616**



* 1. P (20<MPG<50)

**Answer –** P (20<MPG<50) = **0.8988689169682046**

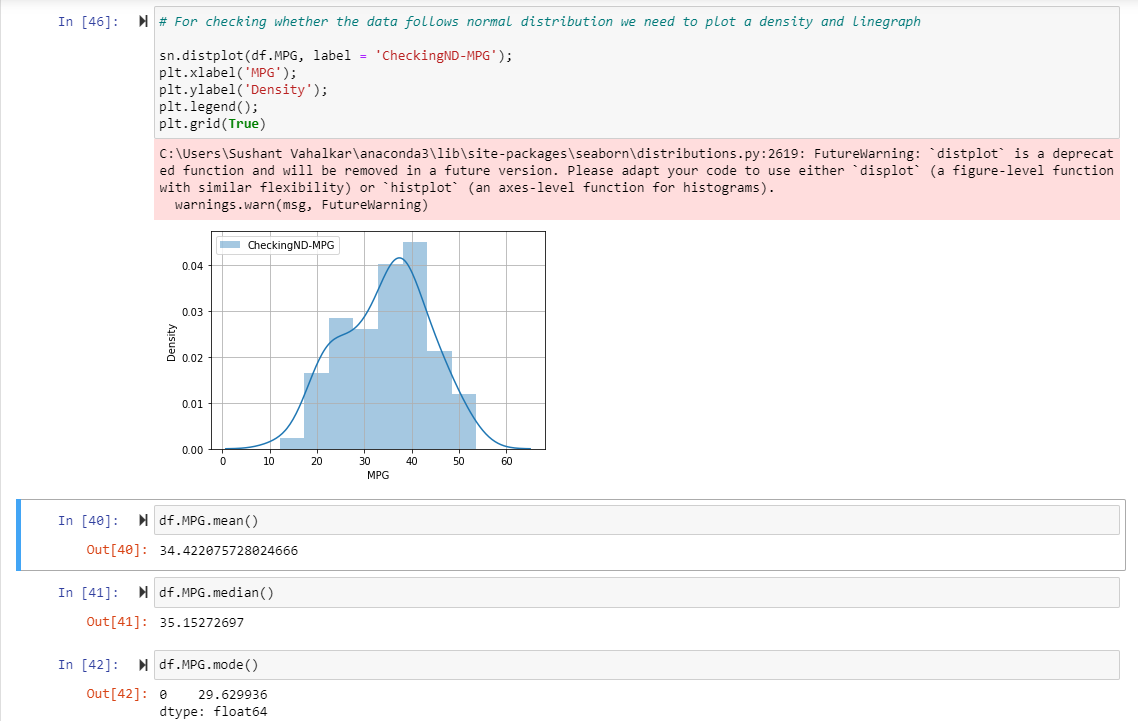


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

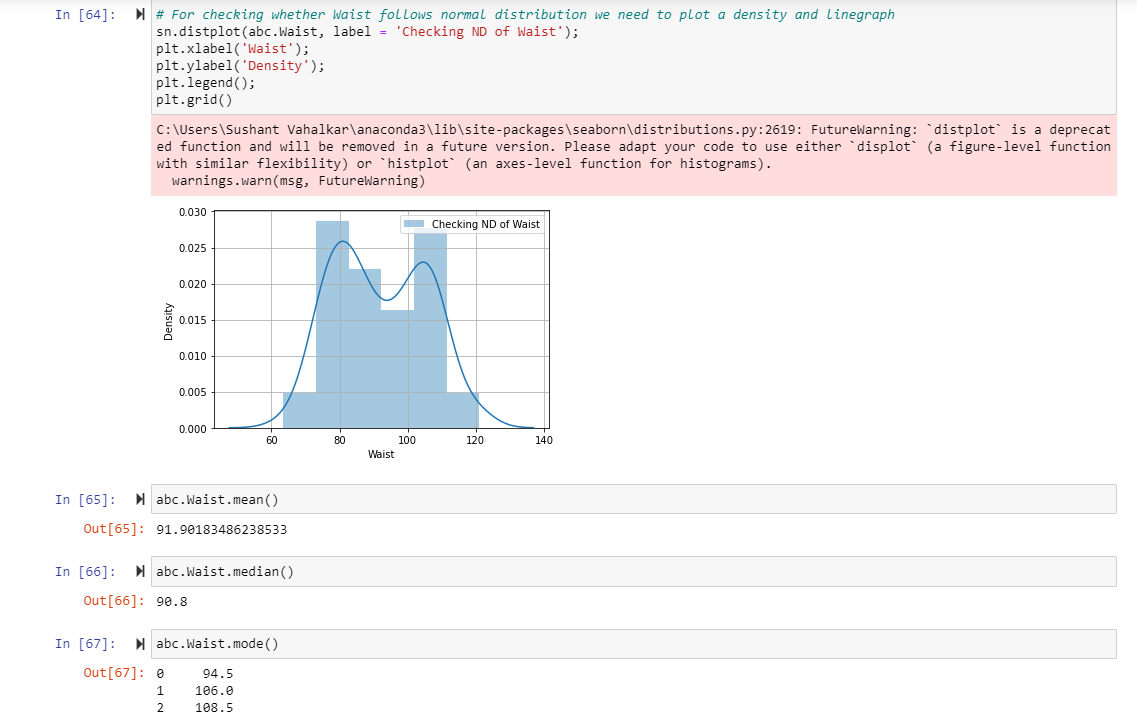
**Answer-** From the values and graph below, we can say that MPG of Cars is not normally distributed as mean, median and mode are not exactly equal (Partially equal)



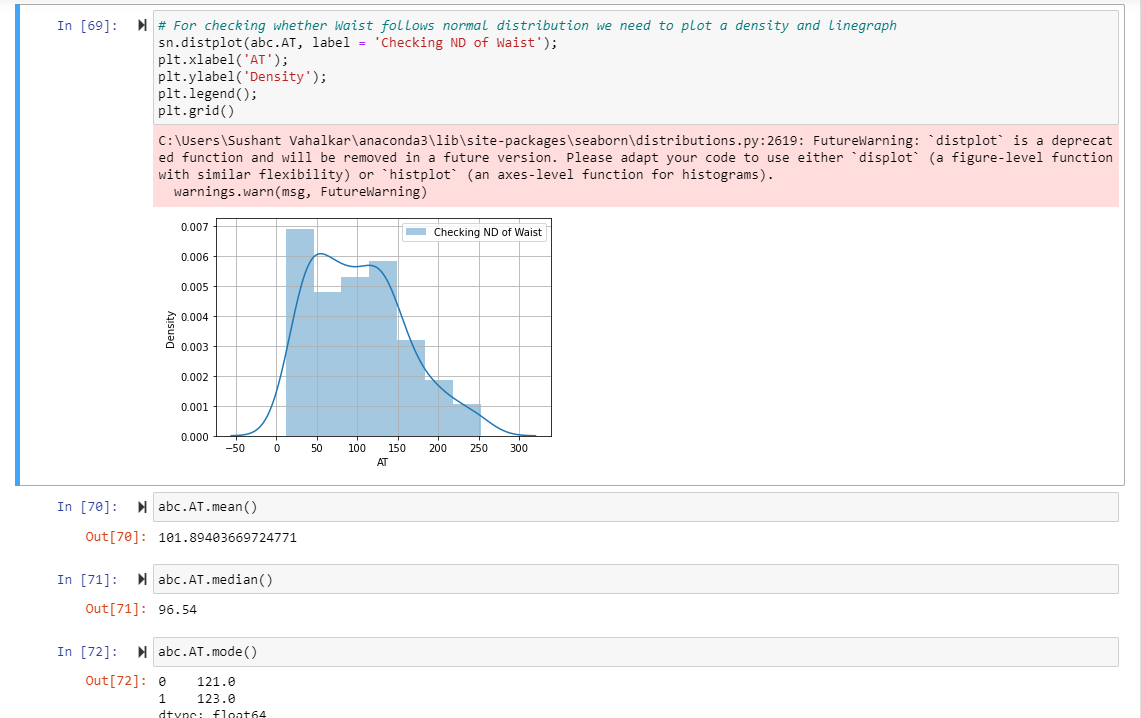
1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Answer- 1) From the values and graph below, we can say that Waist of wc-at is not normally distributed as mean, median and mode are not exactly equal (Partially equal)



Answer- 2) From the values and graph below, we can say that AT of wc-at is not normally distributed as mean, median and mode are not exactly equal.



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

**Answer –** We have considered Computer\_Data as the reference Dataset

**We have Calculated Z scores of hd at 90%, 94% and 60% confidence interval respectively.**

1. Z score of hd at 90% confidence interval is: 747.9449
2. Z score of hd at 94% confidence interval is: 818.586
3. Z score of hd at 60% confidence interval is: 482.1042

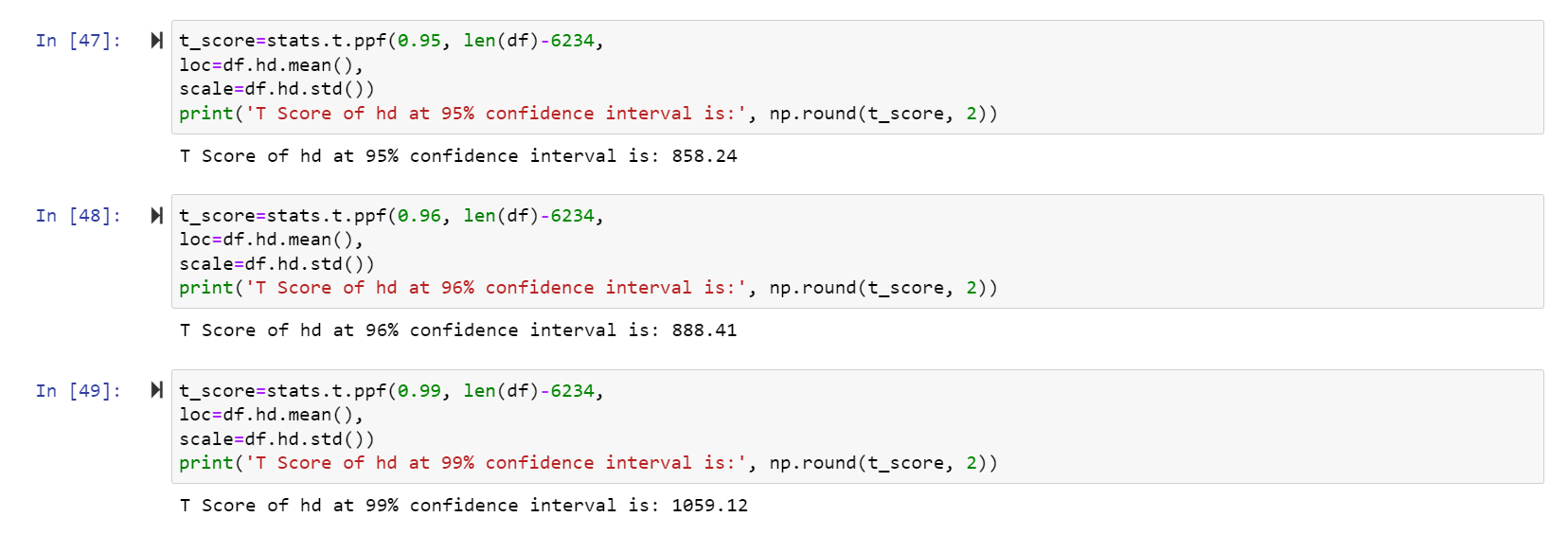
****

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

**Answer-** We have considered Computer\_Data as the reference Dataset

**We have Calculated T scores of hd at 95%, 96% and 99% confidence interval respectively.**

1. T Score of hd at 95% confidence interval is: **858.24**
2. T Score of hd at 96% confidence interval is: **888.41**
3. T Score of hd at 99% confidence interval is: **1059.12**

****

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

**Answer-** Given data: -

1. Population Mean(μ) = 270
2. Sample Mean(x̄)= 260
3. Sample SD(S) = 90
4. n = 18

Find t-scores

t = (x̄- μ)/(S/n^0.5)

= (260-270)/ (90/sqrt (18))

= **-0.4714045**

Now, we find P value: -

P value = stats.t.cdf (-0.4714045, df=17)

= **0.32167**

Hence, Probability that 18 randomly selected bulbs would have an average life of no more than 260 days is **32.17%**

Assuming significance value α = 0.05 (Standard Value) (If p value < α; Reject Ho and accept Ha or vice-versa)

**Thus, as p-value > α, we accept Ho i.e. The CEO claims are false and the avg. life of bulb > =260 days**