|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Integer (Numerical) Data Type |
| Results of rolling a dice | Integer (Numerical) Data Type |
| Weight of a person | Float(Numerical) Data Type |
| Weight of Gold | Float(Numerical) Data Type |
| Distance between two places | Integer or Float(Numerical) Data Type |
| Length of a leaf | Float(Numerical) Data Type |
| Dog's weight | Float(Numerical)Data Type |
| Blue Color | Character Data Type |
| Number of kids | Integer (Numerical) Data Type |
| Number of tickets in Indian railways | Integer (Numerical) Data Type |
| Number of times married | Integer (Numerical) Data Type |
| Gender (Male or Female) | Boolean Data Type |

Q1) Identify the Data type for the Following:

NAME : M KISHORE ARVIND

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

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

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

**Answer :** Let **‘E’** be the event of Three coins tossed.

The possible outcomes of the event when three coins are tossed are as follows :-

**E** = { HHH , HHT,THH,HTH,THT,HTT,TTH,TTT}

**P(E)** is to find the probability of getting two heads and one tail in the event..

In the event , we have 3 such cases where we get 2 Heads and 1 Tail.

Therefore, the Probability ( **P(E)** ) is equal to:

**P (E) = 3/8 = 0.375**

Hence, from P(E) we infer that the probability of getting 2 heads and 1 tail when 3 coins tossed is **0.375**.

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 :** Let **‘O’** be the possible outcomes of rolling 2 dies at a time .

The possible outcomes of the event are shown below:

**Outcomes (O)** = {(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)}

Now, coming to the parts of the question..

1. Let **‘E1’** be the event of getting the sum **‘1’** in 2 dies when rolled.

Since, it is not possible of getting sum of 1 in two dies .Therefore, the probability of getting 1 is

**P(O(E1)) = 0**

1. Let **‘E2’** be the event of getting the sum of less than (or) equal to **‘4’**.

The possible outcomes of getting of getting **‘E2’** are as follows:

**O(E2)** = {(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}

Count of outcomes of the event **‘E2’** = **N(O(E2))** = 6

Count of total possible outcomes in the event **‘N(O)’** = 36

Hence, the probability of Event **‘E2’** is:

**P(E2) = N(O(E2)) / N(O)**

**P(E2) = 6 / 36**

**P(E2) = 1 / 6 = 0.1667**

Therefore , the Probability of getting a sum of 4 or less than 4 in two dies is **0.1667.**

1. Let **‘E3’** be the event of getting a sum that is divisible by 2 and 3.

The possible outcomes of the event **‘E3’** are as follows:

O(E3) = {(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)}

Count of the event = **N(O(E3))** = 6

Count of total possible outcomes in the event **‘N(O)’** = 36

Hence, the probability of Event **‘E3’** is:

**P(E3) = N(O(E3)) / N(O)**

**P(E3) = 6 / 36**

**P(E3) = 1 / 6 = 0.1667**

Therefore , the Probability of getting a sum that is divisible by 2 and 3 in two dies is **0.1667**.

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 ) :** Given, that a bag contains 2 red balls , 3 green balls , 2 blue balls .

Total no. of balls in the bag = 7 balls .

Since, the balls are drawn at random, here ,we use the concept of Permutation & Combination in this problem.

Let **‘E’** be the event of drawing 2 balls that are not blue in colour.

There can be 3 possibilities in drawing the balls at random for the event (E).They are:

1. **Drawing out 2 red balls.**
2. **Drawing out 2 green balls.**
3. **Drawing 1 red and 1 green ball.**

In **case(i),** Let **” E’ “** be the event of drawing 2 balls at random from the bag.

Therefore , by using concept (P&C).

**P(E’) = 7C2 = 21 .**

1. Let the event of drawing 2 red balls at random be (E1).

Therefore, by using concept of (P&C),the probability of drawing 2 red balls at random from the bag is:

**P( E1) = = 1.** **(Using nCr = (n!) / ((r!) \* (n-r)!) )**

In **case(ii),**Let the probability of drawing 2 green balls at random be (E2).

Therefore, by using concept (P&C).

**P(E2) = = 3.**

In **case(iii),**Let the probability of drawing 1 red ball and 1 green ball at random be **(E3).**

Therefore, by using concept (P&C).

**P(E3) = = 3 \* 2 = 6.**

Therefore, the probability of drawing 2 balls that are not in blue colour is:

**P(E) =( P(E1) + P(E2) + P(E3) ) / P(E’)**

**P(E) =( 3+1+6 ) / 21**

**P(E ) = 10/21 = 0.47619.**

Therefore, the probability of drawing two balls that are neither blue in colour is **0.47619.**

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):** Given are the probabilities for the children A to F .The expected number of candies for a randomly selected child is given by:

**E(x) = ∑ x p(x)**

**x=1 to 6**

We substitute the corresponding no of candies count and the probability of getting it and take the product of each of it and sum it up.

E(x) =( 1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.120)

**E(x) = 3.09** which is approximated to 3.

Therefore, the expected no. of candies that each child gets is **3.**

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

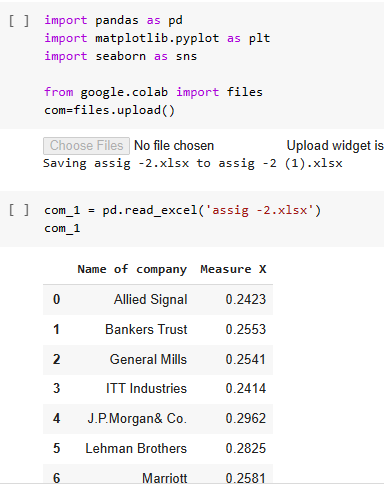
* For Points, Score, Weight

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

**Use Q7.csv file**

**ANS):** To calculate the Mean, Median , Mode , Variance ,Standard deviation ,Range for the datasets Points , Score ,Weight, firstly we have to load the csv file in google colab notebook by using the following commands:

we get the cars.csv table in google colab.

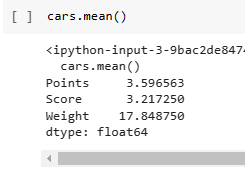




**MEAN:**

To calculate mean for Points,Score,Weight , we have used the in built function

cars.mean( ).



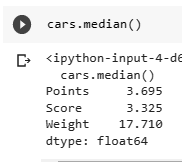
This tells us that the mean of all the data in points column is 3.596563.And the mean of all data points in Score column is 3.217250.And the mean of all the data points in weight column is 17.848750.

**MEDIAN**:

To calculate median for Points,Score,Weight , we have used the in built function

given by:

cars.median( )



This tells us that the median of all the data in points column is 3.695.And the median of all data points in Score column is 3.325.And the median of all the data points in weight column is 17.710.

**MODE:**

To calculate mode for Points,Score,Weight , we have used the in built function given by:

Cars.mode( ) function.



This tells us that AMC Javelin and Cadillac Fleetwood cars have highest frequencies in the cars table.

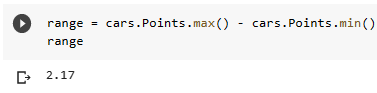
**RANGE:**

Range is calculated by the difference of maximum value and minimum value.

**Range** = maximum value - minimum value.

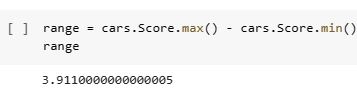
In python , we use max and min function to get the values of maximum and minimum values respectively.

For Points column ,the range is:



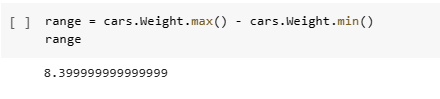
Therefore, the range of Points column in cars is 2.17.

For Score column ,the range is:



Therefore, the range of Score column is 3.911.

For Weight column, the range is:

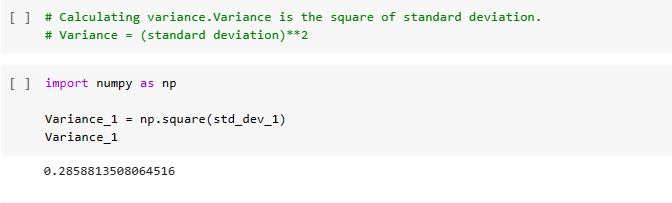


Therefore, the range of weight column is 8.399.

**VARIANCE:**

To calculate Variance, all we need is to square the individual values of Points, Score and Weight. To do the square , we invoke square function from numpy library.

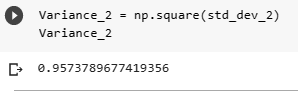
Variance of Points =



Therefore, the variance of Points is 0.28588.

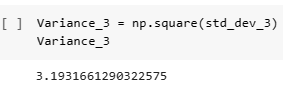
For Score column, we type this code in colab notebook:

Variance of Score =



Therefore, the variance of score column is 0.957.

For Weight column:



Therefore, the variance of Weight column is 3.193.

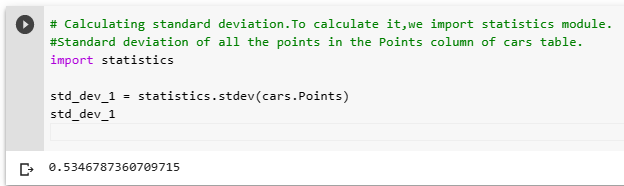
**STANDARD DEVIATION :**

To find the Standard Deviation of the table, we have an inbuilt function in google colab . The general formula is given as:

Standard deviation = (Variance)\*\*0.5 .

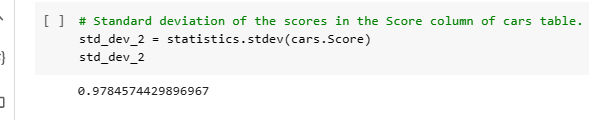
To invoke this formula, we need to import Statistics module.

For Points column, we type the command as:



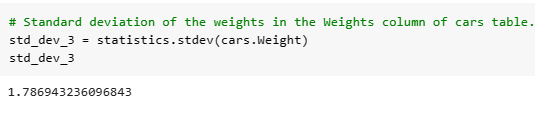
Therefore, the standard deviation of Points column is 0.535.

For Score column, we type the command as:



Therefore, the standard deviation of Score column is 0.97845.

For Weight column, we type the command as:



Therefore, the **standard deviation** of **Weight** column is **1.787**.

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?

ANS): Given are the weights (X) of patients at a clinic (in pounds).

We are required to find the Expected value of the weights .

The Expected value of the weights of the patients is given as:

**E(X) = ∑ x.p(x).**

As all the patients have a frequency ‘ 1’, we can directly sum up all the values.i.e

E(X)=108+110+123+134+135+145+167+187+199

**E(X)=1308.**

Therefore, the expected value of a patient is **1308.**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**ANS):**  To measure the Skewness of a data , we find out using the Skewness formula. It is given as :

**Skewness = ( mean – mode ) / (standard deviation)**

And to find Kurtosis, we use the underlying formula:

**Kurtosis ( β2)= ∑ [[(xi - µ)/σ]^4 ] / N**

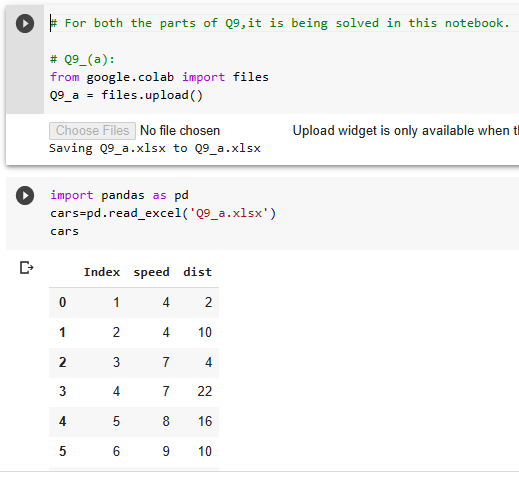
**i=1 to N**

**where ‘xi’ is the data point , ‘µ’ is the mean , ‘ σ ’ is the standard deviation.**

1. In this cars.csv dataset ,we are given with the details of types of cars and their respective speeds and distance they cover.

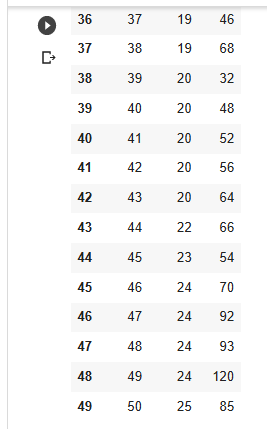
We are required to find Skewness and kurtosis of the data .

To measure the Skewness and Kurtosis , first, we need to load the data set in python,using the import function.



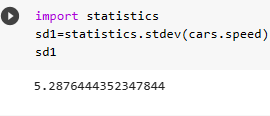




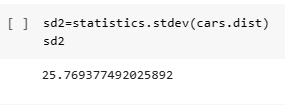


Firstly, we need to find the standard deviation for individual columns .

For Speed column, the Standard Deviation is given as:



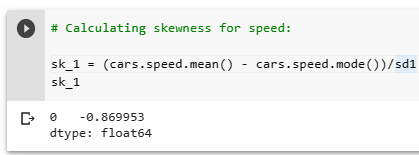
For Distance column, the Standard Deviation is given as:



**SKEWNESS :**

Applying the formula of Skewness :

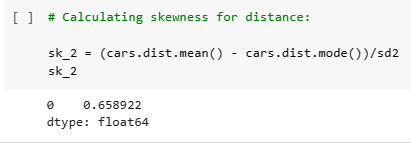
1. Skewness of Speed is =



Therefore, the Skewness of Speed is -0.8699.

The negative sign in Skewness indicates that the data is left skewed data.

ii) Skewness of distance =

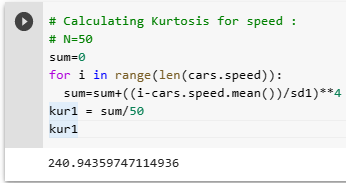


Therefore , the Skewness of distance is 0.658922.

The positive sign in Skewness indicates that the data is right skewed data.

**KURTOSIS:**

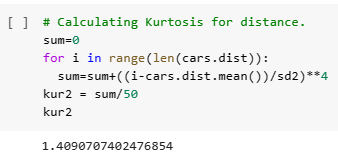
1. Kurtosis of Speed =



Therefore, the Kurtosis of Speed is 240.943.

As the value of Kurtosis is greater than 3 , it is Leptokurtic curve.

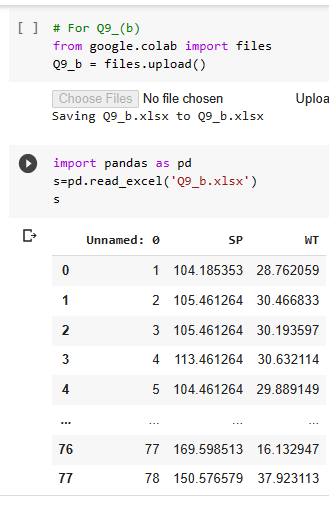
1. Kurtosis of Distance =



As the value of Kurtosis is less than 3, it is a Platykurtic curve.

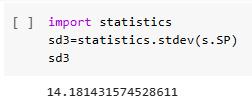
1. In ( b ) part, we are given with the weight dataset.We are required to find the Skewness and Kurtosis of it.

Firstly, we load the dataset in python:

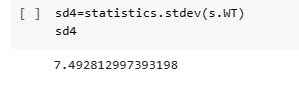


Firstly, we find the standard deviation of SP and WT columns.

Standard Deviation of (SP) =



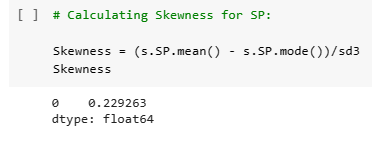
Standard Deviation of (WT) =



**SKEWNESS:**

Applying the Skewness formula , we get

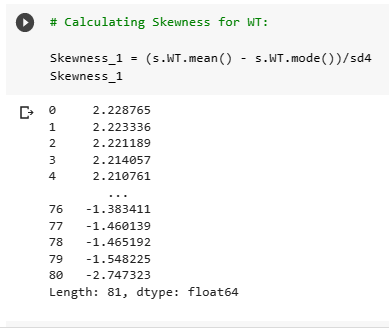
Skewness of SP =



Therefore, the Skewness of SP is 0.229236.

The positive value of Skewness indicates that the data is right skewed data.

Skewness of WT =

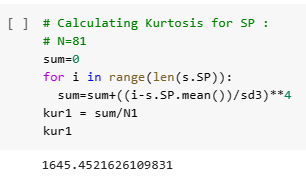


There are positive and negative values of Skewness .

**KURTOSIS:**

Applying the Kurtosis formula, we get:

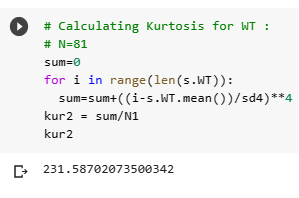
Kurtosis of SP :



Therefore, the Kurtosis of SP is 1645.4521.

As the Kurtosis value is greater than 3 , it is a Leptokurtic curve.

Kurtosis of WT:



Therefore, the Kurtosis of WT is 231.587.

As the value of Kurtosis is greater than 3 , it is Leptokurtic curve.

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



**Ans):** In the above graph of Histogram, it is visualized about the weights of the chicks (vs) the frequency count of the corresponding weights.

In the graph,we see that the **highest** no. of chickens have a weight in the range of

50 - 100 .

The **second highest** no.of chickens have weights in the range of

100 - 150.

**Very least** no.of chickens have weights in the range of

350 – 400.

Now,coming to the box plot, firstly we see that the box plot has an orientation of vertical.Therefore,it is also called as **Vertically – oriented box plot.**

In this box plot,we see that the **median is near to the lower whisker**,that means,that the points are more concentrated to the right of the kurtosis.Therefore , it is called a **right-skewed data.**

We can also see that the upper whisker is greater than the lower whisker that means that the **it has more data points in the Q4(Quartile group 4) region.**

And there are some data that does not lie in the given range .So ,they lie outside the box plot,which are also called as **outliers.**

These are the basic inferences from the above given 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?

**ANS): Given:**

In the above problem ,

i)**The sample size is (n)** = 2000 men.

ii)**The population size is (p)** = 3,000,000 men.

iii)**Sample mean of weights** = 200 pounds.

iv)**Sample standard deviation of weights** = 30 pounds .

Here, in this problem as the population standard deviation is not given,we separately find the **Std Error**. It is given by:

**Std\_Error = Sample standard deviation / (sample size)**

**Std\_Error = 30/**

**Std\_Error = 0.670 pounds.**

We also need to find the degrees of freedom (df) is given by:

**df = n(sample size) – 1**

**df = 2000 - 1 = 1999.**

* We use degrees of freedom (df) to find the corresponding t values from the t –table.

The formula for calculating the confidence level limits are given as:

**Lower\_Limit = Sample mean - ( t % \* std Error )**

**Upper\_Limit = Sample mean + ( t % \* std Error )**

* For **94%** confidence level :-

Lower \_Limit = Sample mean - (t\_94% \* std\_Error)

Upper \_Limit = Sample mean + (t\_94% \* std\_Error)

From the t table, we get **t\_94% = 1.962.**

Substituting the values ,we get

**Lower \_Limit** = 200 - ( 1.962 \* 0.670) = **198.68546.**

**Upper \_Limit** = 200 + (1.962 \* 0.670) = **201.31454.**

**Therefore , the confidence intervals for 94% is 198.68546 - 201.31454.**

* For **96%** confidence level :-

Lower \_Limit = Sample mean - (t\_96% \* std\_Error)

Upper \_Limit = Sample mean + (t\_96% \* std\_Error)

From the t table, we get **t\_96% = 1.962.**

Substituting the values ,we get

**Lower \_Limit** = 200 - ( 1.962 \* 0.670) = **198.68546.**

**Upper \_Limit** = 200 + (1.962 \* 0.670) = **201.31454.**

**Therefore , the confidence intervals for 96% is 198.68546 – 201.31454.**

* For  **98%** confidence level :-

Lower \_Limit = Sample mean - (t\_98% \* std\_Error)

Upper \_Limit = Sample mean + (t\_98% \* std\_Error)

From the t table, we get **t\_98% = 2.330.**

Substituting the values ,we get

**Lower \_Limit** = 200 - ( 2.330 \* 0.670) = **198.4389.**

**Upper \_Limit** = 200 + (2.330\* 0.670) = **201.5611** .

**Therefore , the confidence intervals for 98% is 198.4389 – 201.5611.**

**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**

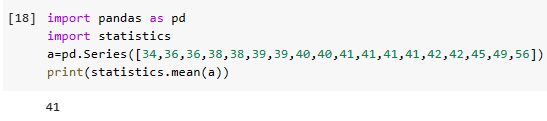
1. Find mean,median,variance,standard deviation.
2. What can we say about the student marks?

**Ans):**1) Given the data of a student ‘s marks in tests .We are required to find the following :

1. **Mean:**
2. **Median:**
3. **Standard Deviation:**
4. **Variance:**
5. **MEAN:**

Mean is the sum of all the numbers divided by the total no.of numbers in the series.

We do it in google colab by first loading the data in pandas and then import mean from statistics module.



From the above screenshot, we see that the mean of the students’s marks is **41.**

1. **MEDIAN**

**Median** is the mid value of the data.The data is required to be in ascending order.Since, the given data is in ascending order,no need to reshuffle the data.

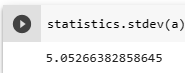
Now.we check the no.of data entries in the data.If there are even number of data sets,then we will have 2 values in the mid region.We take the average of the two number and we will get the median.

If there are odd no.of data entries,then,there is an exact number in the middle of the data which is indeed called the median of the data.

C:\Users\KISHORE\OneDrive\Pictures\Screenshots\Screenshot (205).png

From the above screenshot, we see that the median of the students sore is 40.5.

iii)**Standard deviation(Std)** :



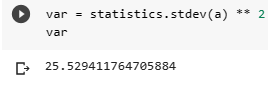
By importing statistics module from python and then using the **standard deviation** function from it.

We get standard deviation = 5.05266.

iv)**Variance**: The variance of the data refers to how a data is variant in the data .To calculate the variance of the data,We use the formula of the variance:

variance is the square of standard deviation.

Variance = standard deviation \*\* 2



From the above screenshot:

Variance = 25.5294.

Therefore, the variance of students marks is 25.5294.

2) We can conclude from the students marks that the marks obtained by the student are less deviated .

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

Ans): If mean and median are equal.then,the nature of the skewness **is zero skewness .**

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

Ans): The nature of skewness **is Right skewed (or) positively skewed data**.

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

Ans): The nature of the skewness **is Left skewed (or) negatively skewed data.**

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

Ans): A positive kurtosis value for a data indicates that the values are more concentrated towards the right side ,and the left tail is spread out.

In positively skewed data , the condition is **Mean>Median>Mode.**

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

Ans): A negative kurtosis value for a data indicates that the values are more concentrated towards the left side,and the right tail is spread out.

In negatively skewed data,the condition is **Mode>Median>Mean.**

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



What can we say about the distribution of the data?

Ans): Basically a box plot uses normal distribution to plot the graph.

What is nature of skewness of the data?

Ans): The nature of the skewness is **Left skewed or negatively** skewed data.

What will be the IQR of the data (approximately)?   
Ans): IQR = Q3-Q1 = 18-10 =8 (approx).

Therefore , the **IQR =8 (approx).**

Q19) Comment on the below Boxplot visualizations?



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

ANS): Given are two box plots in the graph of measuring the (wbs) for various data points.

* In the box plot 1 , we firstly see that the spread of the data is low when compared to the box plot 2,which means that the data taken to visualization done in box plot 1 is very less compared to the values taken in the visualization in box plot 2.
* Box plot 1 takes the range of values from 240(approx) to 287.5 abs.Whereas in boxplot 2,the values range from 150 – 350 abs(approx).
* But ,the median of box plot 1 and 2 are at the same line which is equal to 262.5 abs .Which means that the median of both the box plots are same.
* The Upper and Lower Whiskers of box plot 2 are greater than box plot 1.
* The Inter-Quartile Range(IQR) which is calculated as IQR = Q3-Q1 is also higher for box plot 2 than box plot 1.

These are the few inferences we get by visualizing the box plots 1 and 2.

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

Data \_set: Cars.csv

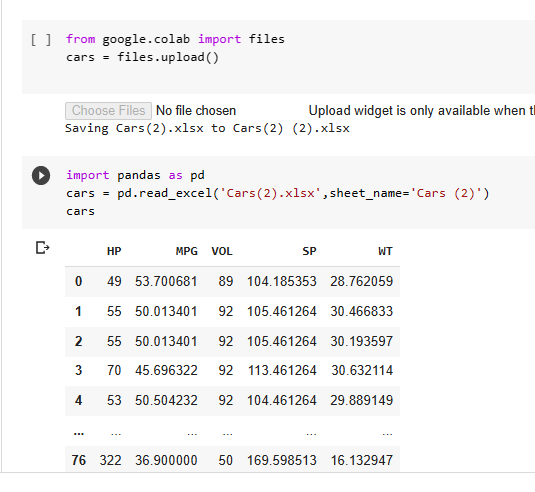
Calculate the probability of MPG of Cars for the below cases.

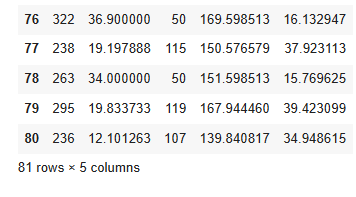
MPG<- Cars$MPG

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

c. P (20<MPG<50)

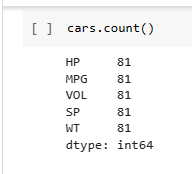
ANS): Given is the dataset of Cars to us to find the probability .First , let us load the Cars.csv in google colab.





After loading the dataset in pandas, we take the total count of each column.

We get it using the count method in python.

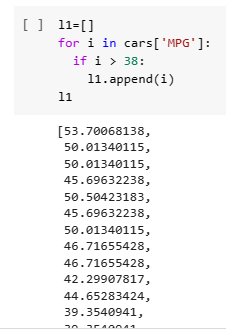


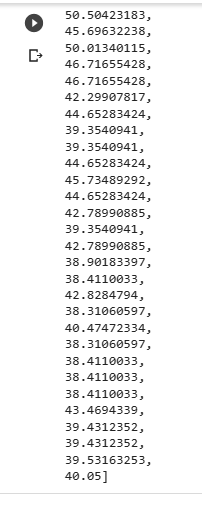
From the above screenshot, we can say that the total no.of values in each column of cars data is 81.

Let, ’P’ be the probability to find the cases. The cases to find the probability are given below:

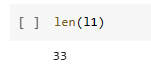
1. P(MPG > 38) :

ANS): To find the probability of MPG > 38, We first create an empty list , and from the dataset set given to us, we extract the values in MPG columns which are greater than 38 and then append it to the empty list (l1).





By using ‘len’ function , we count the no. of elements in the list which are greater than 38 .



Therefore, the probability of MPG column which has elements greater than 38 is

P(MPG>38) = No. of elements greater than 38 / Total no.of elements in MPG

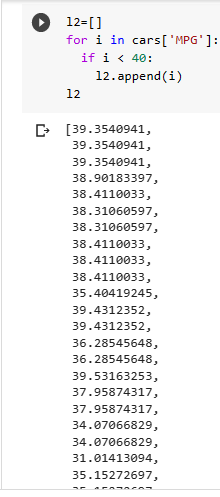
Column.

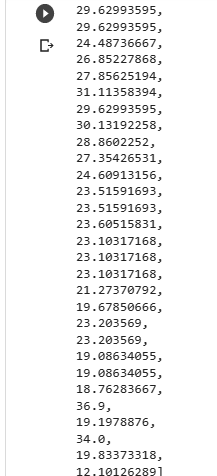
i.e P(MPG > 38 ) = 33/ 81 = 0.407.

Therefore, the probability of MPG > 38 is 0.407.

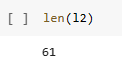
b)P(MPG<40):

ANS): To find the probability of MPG <40, We first create an empty list(l2) , and from the dataset set given to us, we extract the values in MPG columns which are lesser than 38 and then append it to the empty list (l2).





By using ‘len’ function , we count the no. of elements in the list which are lesser than 40 .



Therefore, the probability of MPG column which has elements which are lesser than 40 is

P(MPG<40) = No. of elements greater than 40 / Total no.of elements in MPG

Column.

P(MPG<40) = 61/81 = 0.753.

Therefore, we the probability of MPG having less than 40 is 0.753.

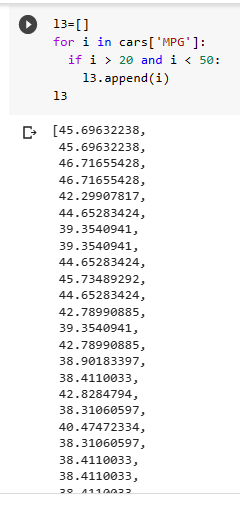
1. P(20<MPG<50):

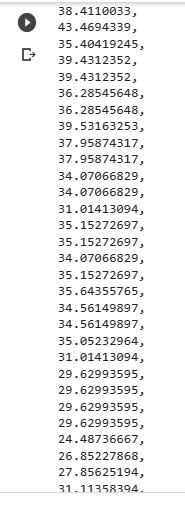
ANS) :To find the probability of 20<MPG <50, We first create an empty list(l3) ,

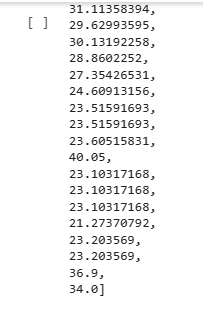
and from the dataset set given to us, we extract the values in MPG

columns which are greater than 20 and less than 50 , then append it to the

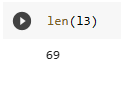
empty list (l3).







By using ‘len’ function , we count the no. of elements in the list which are greater than 20 and less than 50.



Therefore, the probability of MPG column which has elements which are greater than 20 and less than 50 is

P(20<MPG<50) = No. of elements greater than 20 and less than 50 / Total no.of

elements in

MPG

P(20<MPG<50) = 69/81 = 0.851.

Therefore, we the probability of MPG having greater than 20 and less than 50 is 0.851.

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

ANS): We are given with data of different components of cars. In particular we have MPG of cars.We need to find whether the MPG data follows normal distribution or not.

To find whether a data follows normal distribution or not, we use the

**Emperical rule** of Normal Distribution.

**EMPIRICAL RULES :**

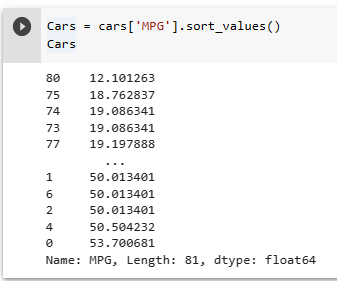
1. **Rule 1 :** Rule 1 states that approximately **(69%)** of the data should lie in the 1 standard deviation away from the mean.
2. **Rule 2** : Rule 2 states that approximately **(95.5%)** of the data should lie in the 2 standard deviation away from the mean.
3. **Rule 3:** Rule 3 states that approximately **(99.8%)** of the data should lie in the 3rd standard deviation away from the mean.

If a curve follows these rules, then, it is confirmed as a normal curve and follows normal distribution.

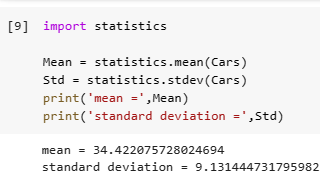
To find this, we first upload our cars.csv file in python



As the values are not in the order, we first sort the values using sort\_values () method.



After sorting the values , we need to find the mean and standard deviation of the MPG column. We find it by importing Statistics module from python. And use the mean and standard deviation methods from it.



From this , we get mean = 34.422 and standard deviation = 9.131.

**1 STANDARD DEVIATION AWAY FROM MEAN.**

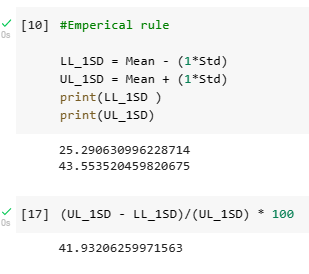
After this, we find the values for 1 standard deviation away from mean.

For this, we need the upper limit and lower limit of 1st standard deviation away from mean.

We find it using the formula for upper and lower limit

Upper\_Limit = mean + (1\* standard deviation)

Lower\_Limit = mean - (1\*standard deviation)



Taking the percentage , we get 41.93 % .Which is less than the Rule1 range.

Let us check for 2 standard deviation away from mean.

**2 STANDARD DEVIATION AWAY FROM MEAN**

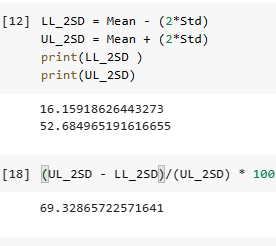
Now, we shall find the values for 2 standard deviation away from mean.

For this, we need the upper limit and lower limit of 2nd standard deviation away from mean.

We find it using the formula for upper and lower limit

Upper\_Limit = mean + (2\* standard deviation)

Lower\_Limit = mean - (2\*standard deviation)

****

Taking the percentage , we get 69.3286 , which is less than Rule 2 range.

Let us check for 3 standard deviation away from mean.

**3 STANDARD DEVIATION AWAY FROM MEAN**

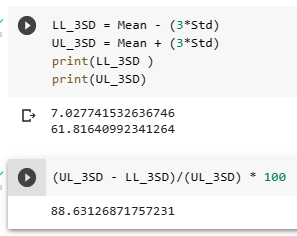
Now, we shall find the values for 3 standard deviation away from mean.

For this, we need the upper limit and lower limit of 3rd standard deviation away from mean.

We find it using the formula for upper and lower limit

Upper\_Limit = mean + (3\* standard deviation)

Lower\_Limit = mean - (3\*standard deviation)



Taking the percentage , we get 88.631 , which is less than Rule 3 range.

Since, all the 3 test cases fail in this dataset , this dataset does not follow the

empirical rule. As a result , it also does not follow Normal Distribution.

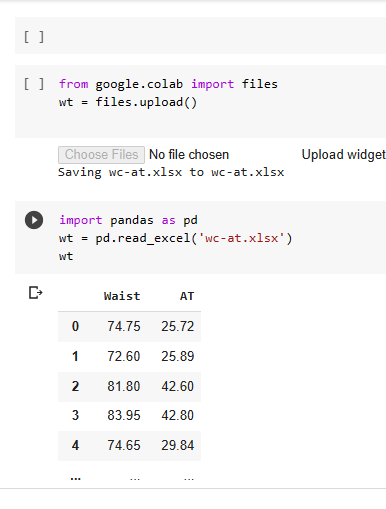
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

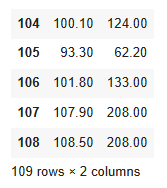
Dataset: wc-at.csv

Ans): In this dataset ,we are given the information of list of values of Adipose Tissue (AT) and Waist Circumference(Waist) .

We need to find whether the dataset follows Normal Distribution or not.

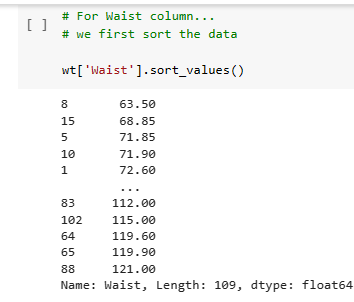
To find this, we first upload our wc-at.csv file in python.





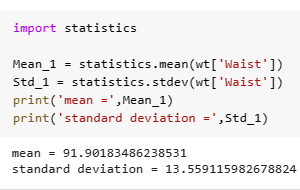
**FOR WAIST COLUMN**

Firstly, we need to sort the data before calculating the mean and standard deviation of WAIST column.



After sorting the data, to calculate mean and standard deviation, we use the inbuilt functions by importing the statistics module.

statistics.mean and statististics.stdev



From the above screenshot, we observe that mean is 91.9018 and standard deviation is 13.559.

**1 STANDARD DEVIATION AWAY FROM MEAN.**

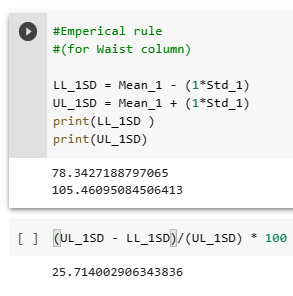
After this, we find the values for 1 standard deviation away from mean.

For this, we need the upper limit and lower limit of 1st standard deviation away from mean.

We find it using the formula for upper and lower limit

Upper\_Limit = mean + (1\* standard deviation)

Lower\_Limit = mean - (1\*standard deviation)



For one standard deviation , we get 25.714 % , which is less than Rule 1

range. Therefore, Rule 1 has failed .

Let us check for 2 standard deviation away from mean.

**2 STANDARD DEVIATION AWAY FROM MEAN**

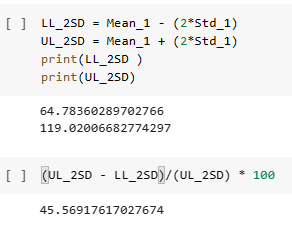
Now, we shall find the values for 2 standard deviation away from mean.

For this, we need the upper limit and lower limit of 2nd standard deviation away from mean.

We find it using the formula for upper and lower limit

Upper\_Limit = mean + (2\* standard deviation)

Lower\_Limit = mean - (2\*standard deviation)



For 2 standard deviation away from mean , we get 45.569 % , which is

less than 95.5% . Therefore, Rule 2 is also failed.

Let us check for 3 Standard deviation away from mean.

**3 STANDARD DEVIATION AWAY FROM MEAN**

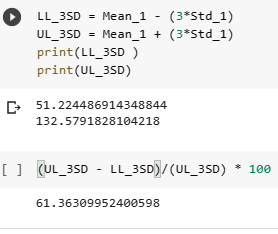
Now, we shall find the values for 3 standard deviation away from mean.

For this, we need the upper limit and lower limit of 3rd standard deviation away from mean.

We find it using the formula for upper and lower limit

Upper\_Limit = mean + (3\* standard deviation)

Lower\_Limit = mean - (3\*standard deviation)

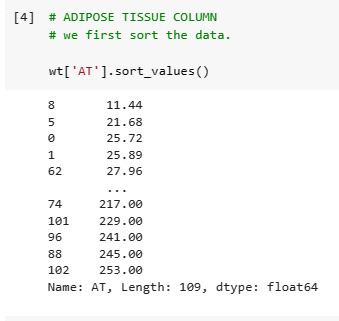


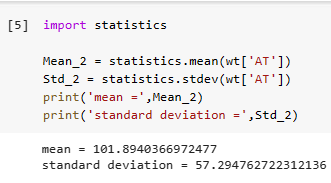
For 3 standard deviations away from mean ,we get 61.363 which is less than 99.8%.

As all the test cases have failed , we conclude that the waist column from wc.xls dataset do not follow Normal Distribution.

**ADIPOSE TISSUE COLUMN**

Firstly, we need to sort the data before calculating the mean and standard deviation of ADIPOSE column.





From the above screenshot, we find that the mean of the AT column of waist data is 101.894 and standard deviation of the AT column is 57.294.

**1 STANDARD DEVIATION AWAY FROM MEAN.**

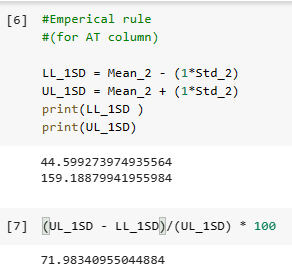
After this, we find the values for 1 standard deviation away from mean.

For this, we need the upper limit and lower limit of 1st standard deviation away from mean.

We find it using the formula for upper and lower limit

Upper\_Limit = mean + (1\* standard deviation)

Lower\_Limit = mean - (1\*standard deviation)



From the above screenshot, we find that the value of 1 standard deviation away from mean is 71.89. which is near to 69%.

Let us check for 2 standard deviation away from mean.

**2 STANDARD DEVIATION AWAY FROM MEAN**

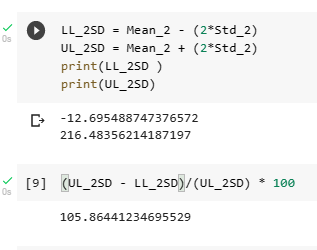
Now, we shall find the values for 2 standard deviation away from mean.

For this, we need the upper limit and lower limit of 2nd standard deviation away from mean.

We find it using the formula for upper and lower limit

Upper\_Limit = mean + (2\* standard deviation)

Lower\_Limit = mean - (2\*standard deviation)



From the above screenshot, we find that the value of 2 standard deviation away from mean is 105.8644%,which is near to 95.5% .

Let us check for 3 Standard deviation away from mean.

**3 STANDARD DEVIATION AWAY FROM MEAN**

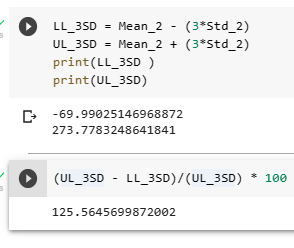
Now, we shall find the values for 3 standard deviation away from mean.

For this, we need the upper limit and lower limit of 3rd standard deviation away from mean.

We find it using the formula for upper and lower limit

Upper\_Limit = mean + (3\* standard deviation)

Lower\_Limit = mean - (3\*standard deviation)



From the above screenshot, we find that the value of 3 standard deviation away from mean is 125.5645% .

From this , we conclude that It does not follow Normal Distribution.

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

ANS) The Z-scores for the respective confidence intervals can be calculated by firstly calculating the area under the curve:

The area under the curve is calculated by:

**Area of the curve = ( 1 + CL ) / 2.** ---🡪 (1)

Where CL is the confidence interval.

i)For 90% confidence interval, CL = 90/100 = 0.90. -🡪 (2)

substitute (2) in (1), we get :

Area = (1+0.90)/2 = 0.95.

Now, looking for this value in normal distribution table , i.e z-table,

But, we see that there is no exact value of 0.95.It lies between 1.64 and

1.65. Therefore, to arrive at an answer ,we take the mean of these

values i.e (1.64+1.65)/2 = 1.645.

Hence,the Z-score for **90%** confidence interval is **1.645.**

ii) For 60% confidence interval, CL = 60/100 = 0.60. --🡪(3)

substitute (3) in(1),we get:

Area = (1+0.60)/2 = 0.80.

Now, looking for this value in normal distribution table , i.e z-table,

But, we see that there is no exact value of 0.80.It lies between 0.84 and

0.85. Therefore, to arrive at an answer ,we take the mean of these

values i.e (0.84+0.85)/2 = 0.845.

Hence,the Z-score for **60%** confidence interval is **0.845**.

iii) For 94% confidence interval,CL = 94/100 = 0.94. --🡪(3)

substitute (3) in(1),we get:

Area = (1+0.94)/2 = 0.97.

Now, looking for this value in normal distribution table , i.e z-table,

But, we see that there is no exact value of 0.97.It lies between 1.88 and

1.89. Therefore, to arrive at an answer ,we take the mean of these

values i.e (1.88+1.89)/2 = 1.885.

Hence,the Z-score for **94%** confidence interval is **1.885**.

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

ANS): Given is the sample size (n) = 25.Firstly , we need to find the degrees of freedom of the sample size,which is given below:

**df=n-1** . Where **df** is **(degrees of freedom).**

i.e **df =25-1=24**.

i) To calculate the t-scores , we infer from the t-table .

For **95%** confidence interval,by looking into the table ,we get:

**t = 2.064.**

Therefore, the t-score for 95% confidence interval is **2.064.**

ii)For 96% confidence interval:

By looking into the table, we get:

**t = 2.064.**

Therefore, the t-score for 96% confidence interval **is 2.064.**

iii)For 99% confidence interval:

By looking into the table, we get:

**t=2.797.**

Therefore,the t-score for 99% confidence interval is **2.797.**

Q 24**)**A Government companyclaims 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): Given:-

Sample Mean = 260 days.

Population mean = 270 days.

Sample Standard deviation = 90 days.

N = 18 bulbs

From Hypothesis testing:

Null Hypothesis : The average bulbs last for more than 270 days.

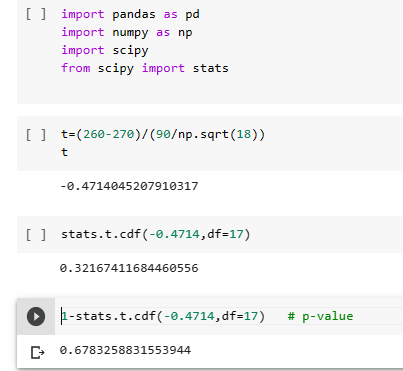
Alternate Hypothesis: The average bulbs lasts less than 270 days.

To check the Hypothesis, We need to find the Hypothesis Test value,which is denoted as ‘ t ‘.

It is given as:

t = (x ’ - µ) / (σ/ (√n))

Using this formula, we find the p value .



We get p = 0.67832 .

Since , p-value > ( significance level of 0.05),there is a high chance of error.

Therefore, the probability of 18 bulbs that last for not more than 270 days is

0.32167.