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	Nominal
Number of kids	Discrete
Number of tickets in Indian railways	Discrete
Number of times married	Discrete
Gender (Male or Female)	Nominal

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	Nominal
Hair Color	Nominal
Socioeconomic Status	Ordinal
Fahrenheit Temperature	Interval
Height	Nominal
Type of living accommodation	Ordinal
Level of Agreement	Ordinal
IQ(Intelligence Scale)	Interval
Sales Figures	Ratios
Blood Group	Interval
Time Of Day	Ratio
Time on a Clock with Hands	Ratio

Number of Children	Interval
Religious Preference	Nominal (As it can't be ranked, its not ordinal)
Barometer Pressure	Interval
SAT Scores	Ordinal
Years of Education	Nominal

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

Ans: When 3 coins are tossed, there are the following 8 possibilities:

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

Probability of two heads and one tail: 3/8

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

a) Equal to 1

Ans: Ans: When a Dice is rolled, the least number we get is 1. So When two Dice are rolled, The probability that sum is equal to 1 is "0".

b) Less than or equal to 4
Ans: We know that, when two Dice are rolled, The total number of possibilities are $6 \times 6 = 36$.
Outcomes having the sum is less than or equal to 4 =
(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)
Probability: 6/36 = 1/6

c) Sum is divisible by 2 and 3
Ans: Total number of possibilities = 36
Outcomes having the sum which is divisible by 2 and 3 :
(1,1),(1,2),(1,3),(1,5),(2,1),(2,2),(2,4),(2,6),(3,1),(3,3),(3,5),(3,6),(4,2),(4,4),
(4,5),(4,6),(5,1),(5,3),(5,4),(5,5),(6,1),(6,2),(6,3),(6,4),(6,6)
Probability: 25/36

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?

Ans: Total number of balls in a bag = 7
Total number of balls drawn at random = 2
Number of balls which are not blue (No. of red and green balls) = 5

The probability when first ball is drawn = 5/7
The probability when second ball is drawn = 4/6 = 2/3
The probability that none of the balls drawn is blue: (5/7) ×(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-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:
Let X be the number of candies.

Expected number of candies = $E(X) = \sum(n \times p)$

$= 1\times0.015 + 4\times0.20 + 3\times0.65 + 5\times0.005 + 6\times0.01 + 2\times0.120$

$=0.015 + 0.8 + 1.65 + 0.025 + 0.06 + 0.24$

$=2.79$

Hence, Expected number of candies for a randomly selected child = 2.79

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

Ans:

	Points	Score	Weigh
Mean	3.596	3.217	17.848
Median	3.695	3.325	17.710
Mode	3.07 and 3.92	3.44	17.02 and 18.90
Variance	0.286	0.957	3.193
Standard deviation	0.535	0.978	1.787
Range	2.17	3.91	8.39

Mean of Score <mean of points < mean of weigh

Here as Points has less standard deviation, the data set is less spread out from the mean. So Points is more consistant.

Points and scores both have positively skewed distribution as mean<median<mode.

Weigh is negatively skewed as mean>median>mode.

Even the range talks about the variability of data. So weigh is more spread out and point is the least spread out.

Q8) Calculate Expected Value for the problem below

a) The weights (X) of patients at a clinic (in pounds), are 108, 110, 123, 134, 135, 145, 167, 187, 199.

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

Ans:

The expected value for a discrete random variable is the mean

The mean of given values of patients' weight is: 145.33

Therefore, expected value of the weight of randomly chose patient is 145.33

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:

1. Skewness:
   Speed = -0.117
   Distance = 0.806

As the skewness of speed is negative as it's mean is lesser than the median is lesser than the mode and it is negatively(left) skewed. and skewness of distance is positive as the mean of all the distances is more than the mode and the distribution is positively (right)skewed.

2. Kurtosis:

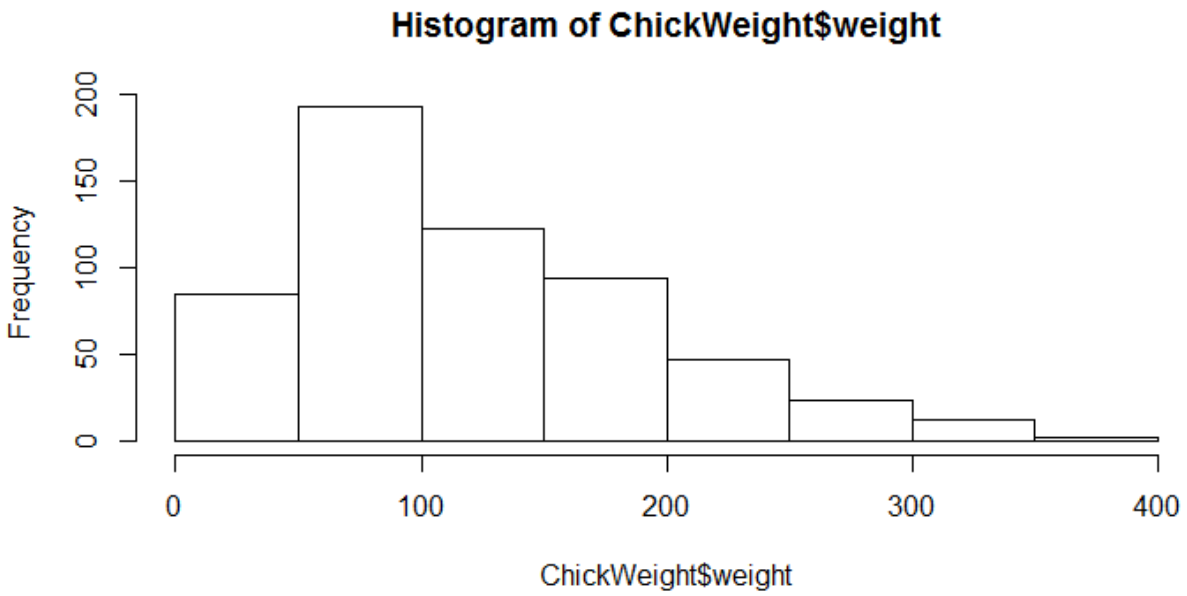
Speed = -0.509

Distance = 0.405

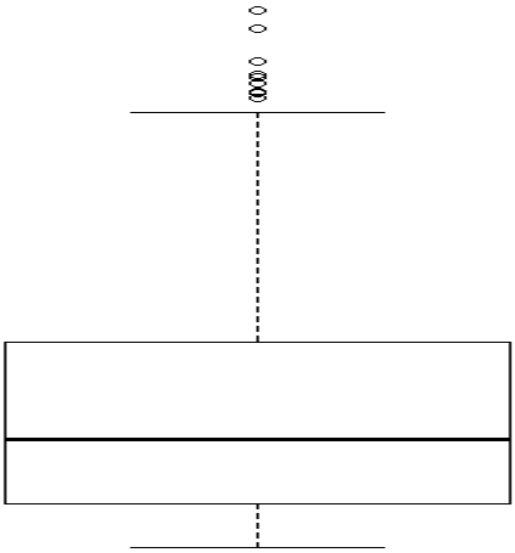
Kurtosis of the speed is negative. So, it has wider peak and thinner tails

Kurtosis of the distance is positive. So, it has wider tails and thinner peaks

Q10) Draw inferences about the following boxplot & histogram



Ans: In the above histogram, the data of the range 0 to 400 is given. In that, the 50-100 chickWeight$weight has highest frequency. That is, 200. So, we can say that most of the values of the data lie in that range. Least number of values are in the range of 350 to 400 lies around the frequency 10. As most of the values are in the left side of the histogram, and the tail is towards right side it is right skewed.



Ans: As the median is closer to the bottom of the box and the lower whisker is shorter, the distribution is positively skewed or right skewed. So the 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?

Ans: Given:

Sample size, n = 2000

Mean, $\bar{x}$ = 200 pounds

Standard deviation, σ = 30 pounds

To calculate:

94%, 98%, 96% confidence intervals

As we are given the standard deviation for the sample, we are using t distribution for solving this question.

Formula: Confidence Interval = $\bar{x} \pm t\frac{\sigma}{\sqrt{n}}$

For 94% confidence interval,

t-value is = 1.896

$\bar{x} + t\frac{\sigma}{\sqrt{n}}$ = 200+ 1.272 = 201.27

$\bar{x} - t\frac{\sigma}{\sqrt{n}}$ = 200-1.272 = 198.73

Therefore, the 94% confidence interval is (198.73, 201.27)

For 96% confidence interval,

t-value is = 2.0732

$\bar{x} + t\frac{\sigma}{\sqrt{n}}$ = 200+ 1.391 = 201.39

$\bar{x} - t\frac{\sigma}{\sqrt{n}}$ = 200 – 1.391 = 198.61

Therefore, the 96% confidence interval is (198.61, 201.39)

For 98% confidence interval,

t-value is = 2.3535

$\bar{x} + t\frac{\sigma}{\sqrt{n}}$ = 200 + 1.578 = 201.58

$\bar{x} - t\frac{\sigma}{\sqrt{n}}$ = 200 - 1.578 = 198.42

Therefore, the 98% confidence interval is (198.42, 201.58)

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) Mean = 41.0
Median = 40.5
Variance = 25.53
Standard deviation = 5.053

2) As students' average marks is slightly more than median, we can say that the curve is moderately symmetrical around mean. It means, more than half of the students scored below average. The marks is in the range of 34 to 56.

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

Ans: Symmetrical (Has normal distribution)

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

Ans: Right skewed (Positively skewed)

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

Ans: Left skewed (Negatively skewed)

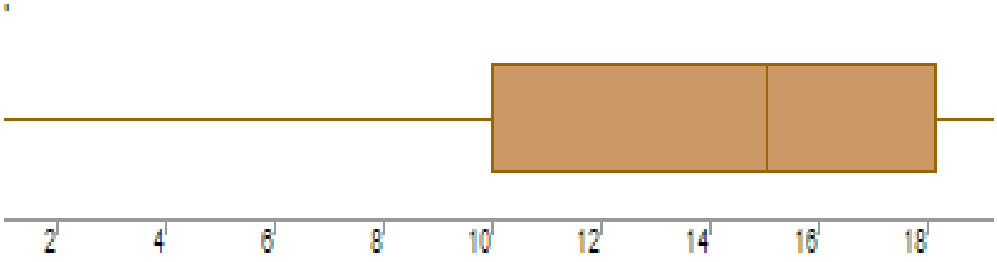
Q16) What does positive kurtosis value indicates for a data?

Ans: Positive kurtosis means the distribution is peaked and possesses thick tails. If it is >3 then it has more outliers.

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

Ans: Negative kurtosis means the distribution is flat and possesses thin tails. If it is <3 then it has less outliers.

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



What can we say about the distribution of the data?

Ans: About 25% of the data are having value less than 10. 75% of the data is more than 10. Median is around 15. Here mean is less than the median. Lower

What is nature of skewness of the data?

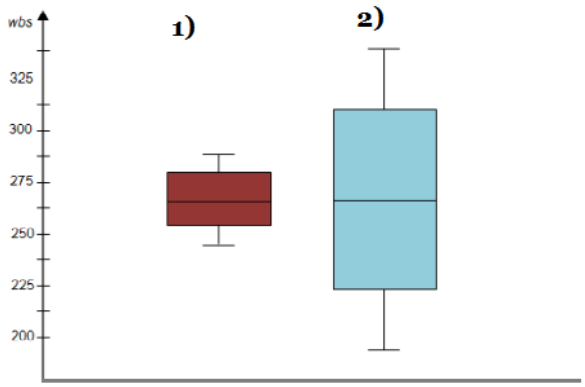
Ans: As the median is closer to the upper quartile, the distribution is negatively skewed.

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

The Interquartile range of the data will be between 10 to 18 (18-10 = 8).

Q19) Comment on the below Boxplot visualizations?

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



Ans: The median of box plots 1 and 2 are the same. The range of 2nd box plot is more than that of 1st box plot. So, the 2nd box plot has wider distribution and hence more scattered data. As the box 2 is taller than the box 1, it has more variable data.

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

Data _set: Cars.csv

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

MPG <- Cars$MPG

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

Ans: P(MPG>38) =
stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std()) = 0.347

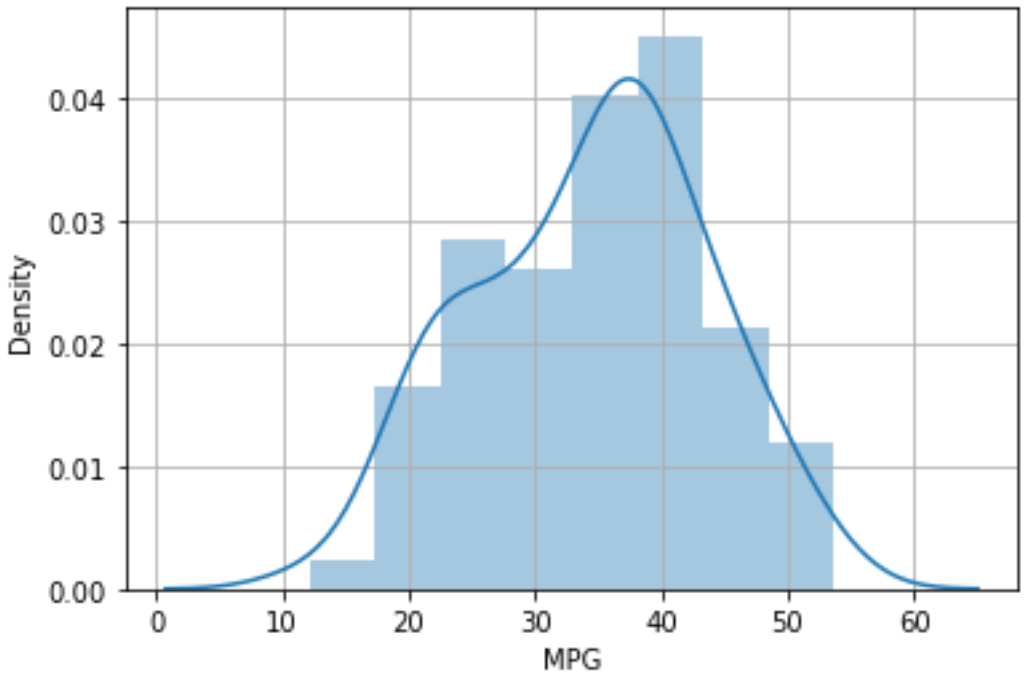
P(MPG<40) =
stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std()) = 0.729

P (20<MPG<50) =
stats.norm.cdf(50,cars.MPG.mean(),cars.MPG.std()) -
stats.norm.cdf(20,cars.MPG.mean(),cars.MPG.std()) = 0.899

Q 21) Check whether the data follows normal distribution

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

Ans: Graph showing Density V/s MPG



From the above plot, although, we feel like left tail is slightly long, we can say that the data is fairly symmetrical. Also, By calculating the mean, median and skewness of the data set,

Mean = 34.42

Median = 35.15

Skewness = -0.17

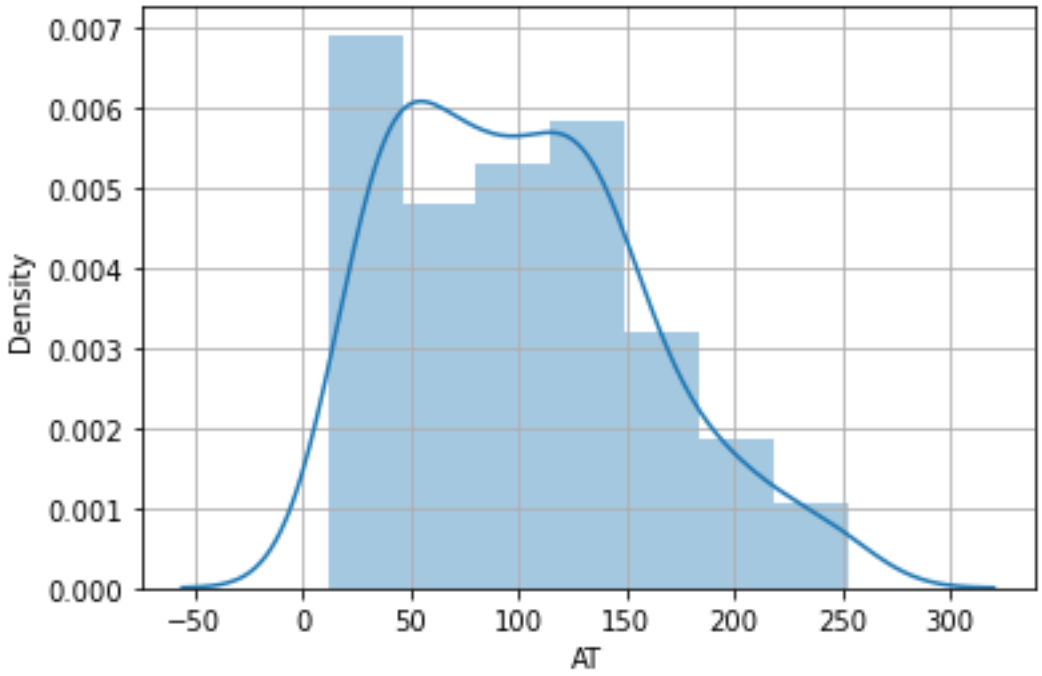
We get to know that, mean ≈ Median and also skewness although negative, lies in the range of 0 to -0.5.

So, we can say that the data is normally distributed.

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

Ans:

For Adipose Tissue (AT), Graph of Density V/s AT



From the above plot, we can say that the data is right symmetric. Also, By calculating the mean, median and skewness of the data set,

Mean = 101.89

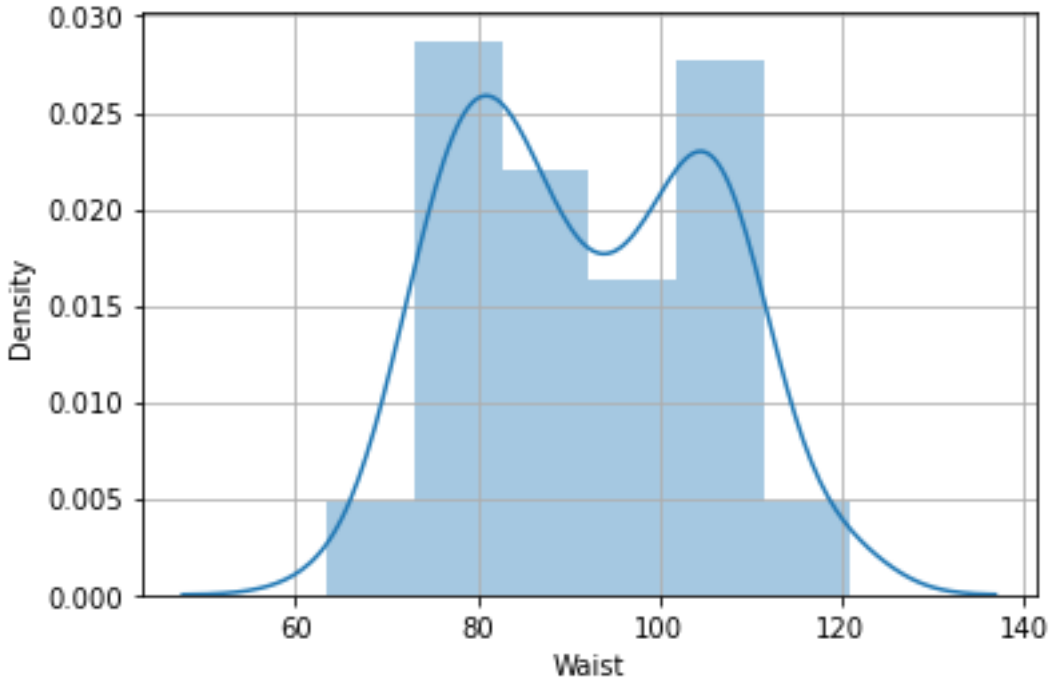
Median = 96.54

Skewness = 0.58

We get to know that, mean > Median and also skewness is positive and more than 0.5

So we can say that the data is right skewed

For WC, Density V/s waist



From the above plot, we can say that the data is fairly symmetrical. Also, By calculating the mean, median and skewness of the data set,

Mean = 91.90

Median = 90.8

Skewness = 0.13

We get to know that, mean ≈ Median and also skewness although negative, lies in the range of 0 to -0.5.

So, we can say that the data is normally distributed.

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

Ans: Z – score for

90% confidence interval = stats.norm.ppf(1-(1-0.90)/2) = 1.645

94% confidence interval = stats.norm.ppf(1-(1-0.94)/2) = 1.881

60% confidence interval = stats.norm.ppf(1-(1-0.60)/2) = 0.842

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

Ans: degree of freedom = sample size-1 =24

t-score for

95% = stats.t.ppf((1-(1-0.95)/2),df = 24) = 2.06

96% = stats.t.ppf((1-(1-0.96)/2),df = 24) = 2.17

99% = stats.t.ppf((1-(1-0.99)/2),df = 24) = 2.79

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

Ans: number of samples(n) = 18

Population mean(μ) = 270

sample mean ($\bar{x}$) = 260

Standard deviation(σ) = 90

Degrees of freedom = 17

$t = \frac{\bar{x} - \mu}{\frac{\sigma}{n}} = \frac{260-270}{\frac{90}{\sqrt{18}}} = -0.471$

The probability that t < - 0.471 with 17 degrees of freedom assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of 0.3218 assuming the mean life of the bulbs is 300 days.