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

Q1) Identify the Data type for the Following:

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

Nominal, Ordinal, Interval, Ratio.

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

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

**Ans**- 3 coins are tossed = {HHH, TTT, HHT,

HTT, THH, HTH,

TTH, THT}

We get no of outcomes = 8

Now we used probability formula to find the probability that two heads and one tail are obtained.

P(c)= No of ways it can happen/ Total no of outcomes -------(1)

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

Thus,the probability of getting two heads and one tail when three coins are tossed is **0.375**.

Top of Form

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dice 1** | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
| **Dice 2** | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 6 | 6 | 6 | 6 | 6 | 6 |
| **Sum** | 2 | 3 | 4 | 5 | 6 | 7 | 3 | 4 | 5 | 6 | 7 | 8 | 4 | 5 | 6 | 7 | 8 | 9 | 5 | 6 | 7 | 8 | 9 | 10 | 6 | 7 | 8 | 9 | 10 | 11 | 7 | 8 | 9 | 10 | 11 | 12 |

**Ans-** We get total no of outcomes is equals to **36**.

**a) Equal to 1**

It is not possible, cause of we cannot get sum is equal to 1 from total 36 no of outcomes.

**b) Less than or equal to 4**

Less than or equal to 4 is 6 from 36 outcomes.

P(p<=4)=6/36 = 1/6 = **0.1666**

**c) Sum is divisible by 2 and 3**

P **=**6/36 = 1/6 = **0.1666**

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 no of ways to choose the 2 balls =**21**

Now, we choose the balls from non blue = **10**

Now we find out probability

P(c) = (none of the balls drawn is blue) / (Total no of ways to choose the 2 balls)

P(c) = 10/21 = **0.4761**

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-** Expected number of candies for RS child =(candies\*probability)

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

= **3.115**

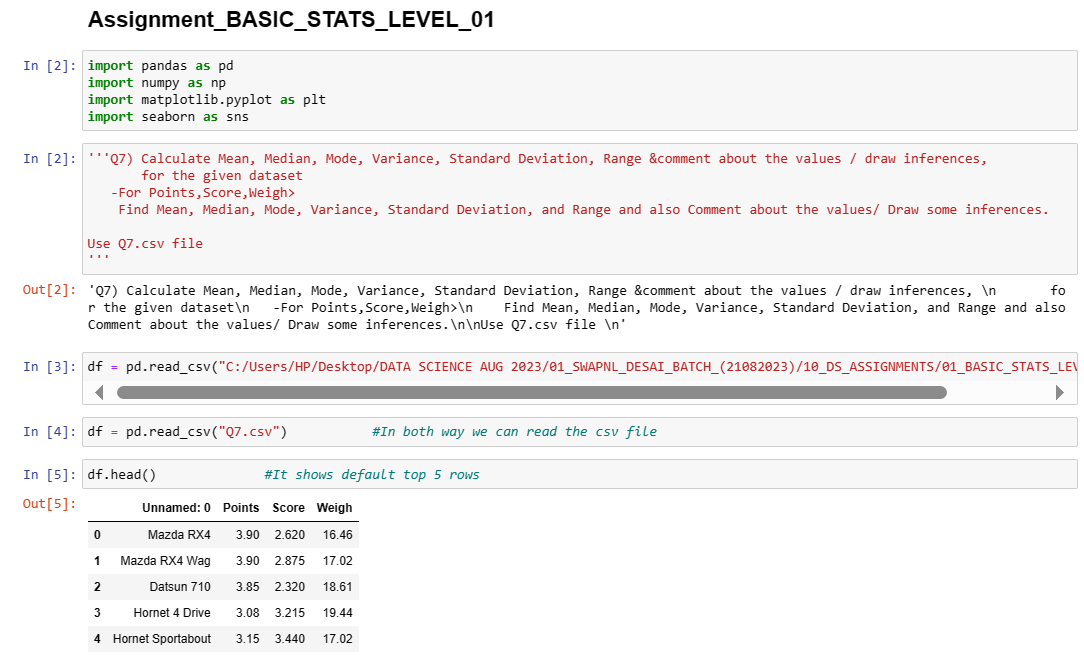
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

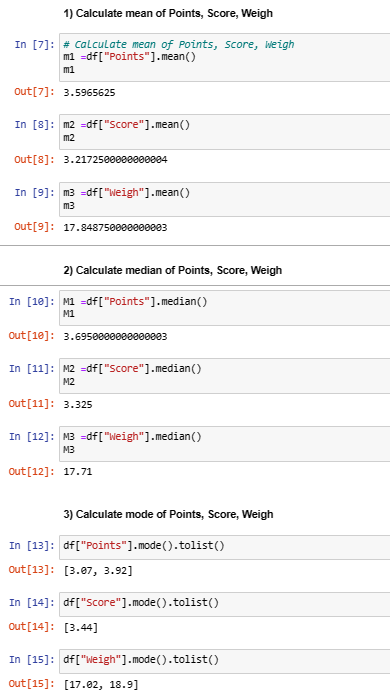
* For Points, Score, Weigh>

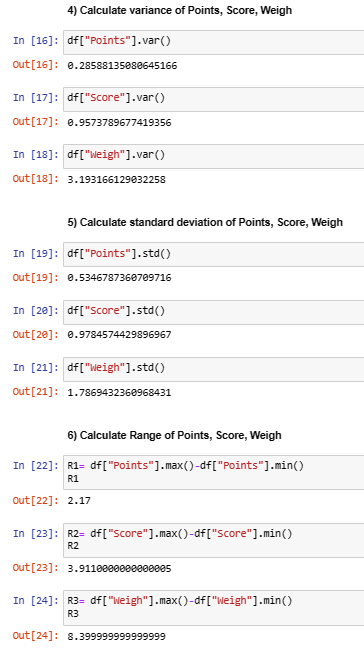
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

**Ans-** Solved in ipython notebook.

****

****

****

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-** Expected value means mean of the total no of observations.

X = [108, 110, 123, 134, 135, 145, 167, 187, 199]

We can use the mean formula to find out the Expected Value of the Weight

of that patient?

Expected value(mean) = (sum(X)/(n)) -----------------(1)

=((108+110+123+134+135+145+167+187+199)/9)

= 1308 /9

= **145.33**

So, the expected value of the weight of a patient randomly chosen from this list is approximately **145.33** pounds.

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

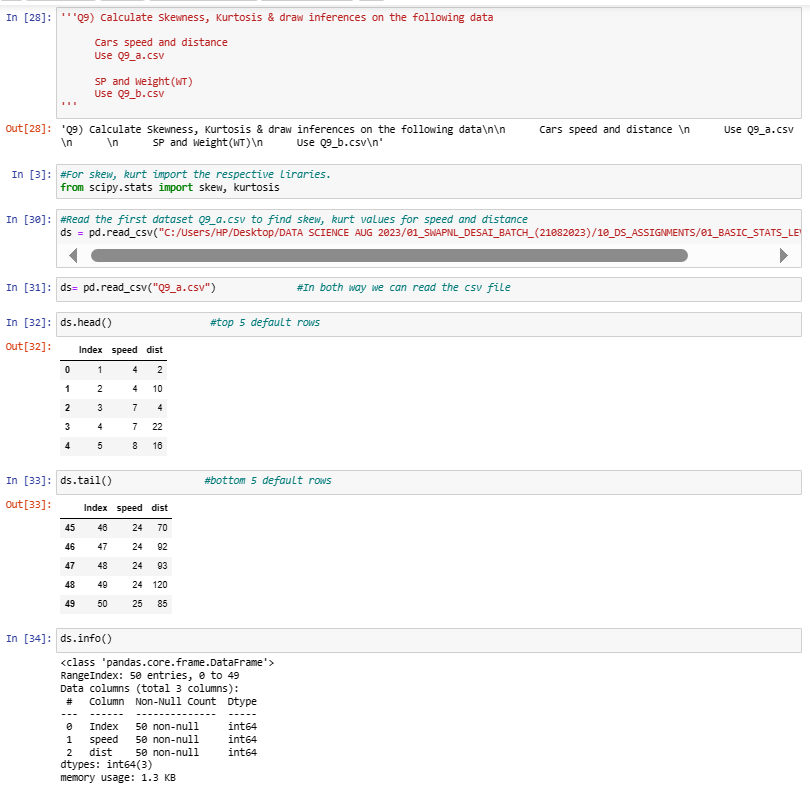
**Cars speed and distance**

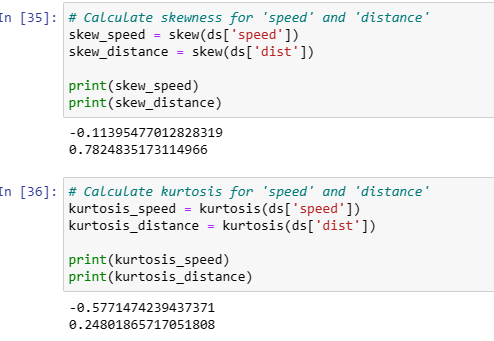
**Use Q9\_a.csv**

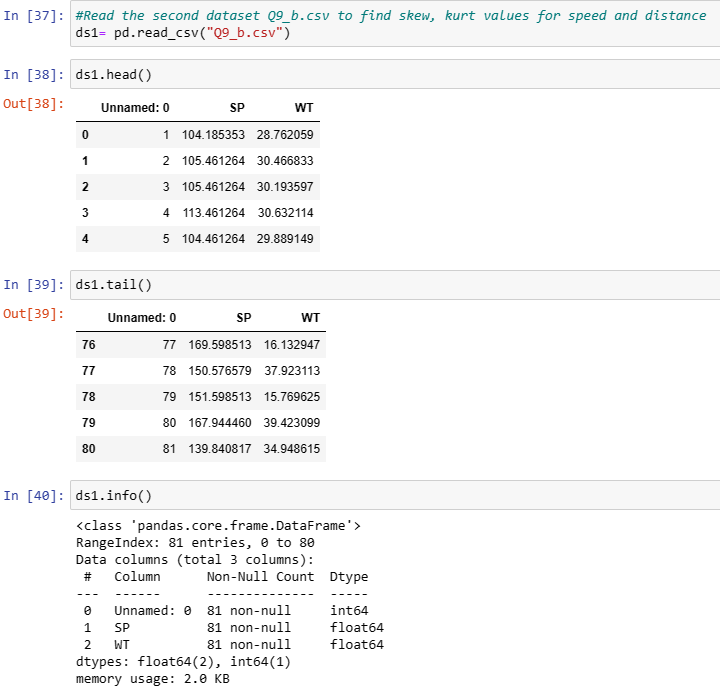
**SP and Weight (WT)**

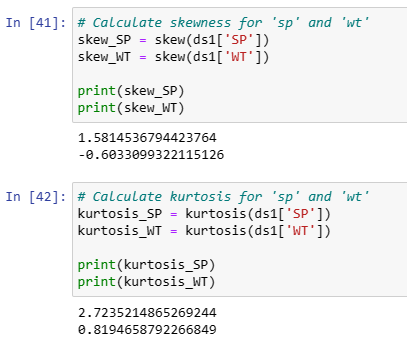
**Use Q9\_b.csv**

**Ans-** Solved in ipython notebook.

****

****

****

****

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





**Ans –**

1. **Histogram –**

**Shape –** From this above histogram we can say that the data is distributed maximumly in left side of the graph and tail flowing towards the right side it means that the shape of the data is a “**positively skewed**”.

**Outliers –** Extreme value in hist is 200.

1. **Boxplot –**

**Lower Quartile (Q1):** The below line of the box represents lower quartile and

25% of the data.

**Median (Q2):** The line inside the box represents the median and 50% of the data.

**Upper Quartile (Q3):** The above line of the box represents upper quartile and 75% of the data.

**Interquartile Range (IQR):** The box represents the middle 50% of the data, and its length is the interquartile range (Q3 - Q1).

**Whiskers:** The lines extending from the box indicate the range of the data.

Whiskers = 1.5\*IQR

**Outliers:** Points beyond the whiskers are considered outliers.

**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 -** Solved in ipython notebook.

Formula of confidence interval is

CI = x +( t \* (s/sqrt(n)) ----------1)

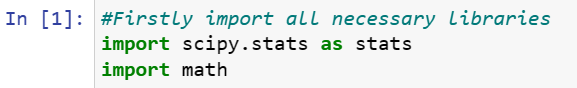
Where,

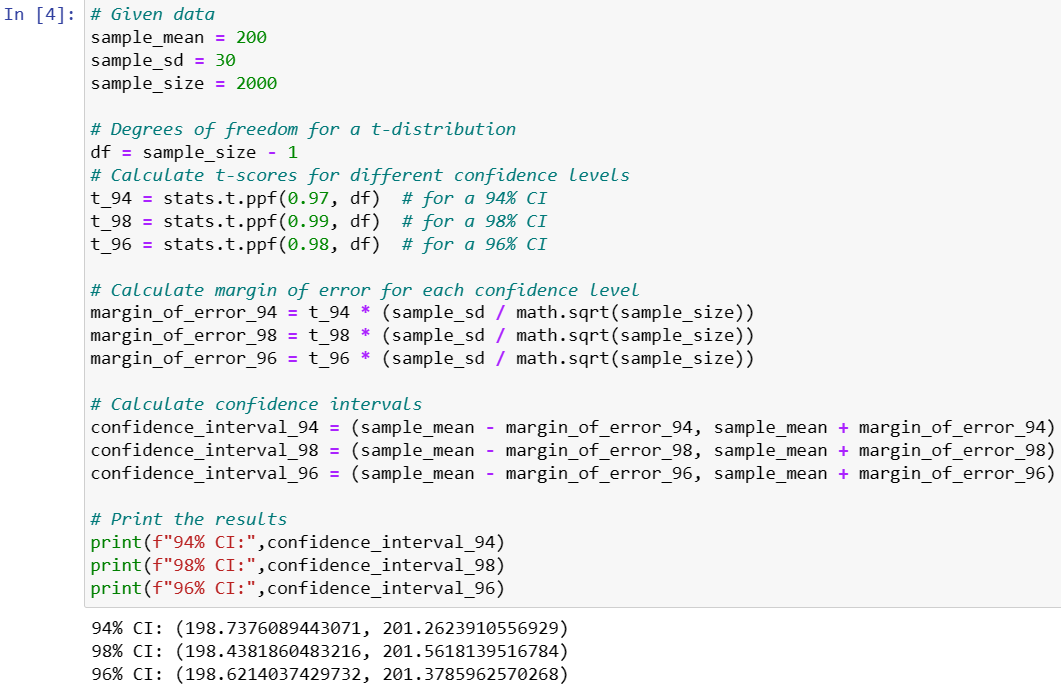
X = sample mean = 200

t = Degree of freedom (n-1) = 2000-1 =1999

s = sample deviation = 30

n = sample size = 2000





**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** – Here N=no of observations = 18

1) Mean = xi/N = 738/ 18 = **41**

Median = 40+41 / 2 = **40.5**

Variance = ∑ (xi-x)^2/n = **24.11**

Standard deviation =sqrt ((xi-x)^2/n-1) = **4.910**

2) From these above observations, we can say that **mean>median** that’s why it is positively skewed data and there is no outlier observed.

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

**Ans** –

* When the mean and median of a data are equal, it means that the data is distributed in symmetric manner.
* There is a balance between the left and right tails.
* 50% of the data distributed to the left and 50% to the right of the graph.

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

**Ans** –

* When the mean > median, it means that data is maximumly distributed on the left and tail spread longer towards right. Hence, we called as “Positively skewed data”.
* Tails flows towards right side of the graph.

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

**Ans** –

* When the median > mean, it means that data is maximumly distributed on the right and tail spread longer towards left. Hence, we called as “Negatively skewed data”.
* Tails flows towards left side of the graph.

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

**Ans** –

* When kurtosis is positive it means its “leptokurtic”.
* Indicates a distribution with wider tails than a normal distribution.
* The distribution has more outliers in the tails.
* The distribution of data at the peak is sharper.

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

**Ans** –

* When kurtosis is negative it means its “platykurtic”.
* Indicates a distribution with lesser tails than a leptokurtic.
* The distribution has less outliers in the tails.
* The distribution of data at the peak is flat.

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



What can we say about the distribution of the data?

**Ans**-

* Data is max distributed to the right and tail flows towards left, hence it is left skewed data.
* It is called “**Negative skewness**”.
* Upper quartile(Q1) = 18
* Middle quartile/median (Q2) = Near 15.5
* Lower quartile(Q3) = 10

What is nature of skewness of the data?

**Ans**-

* Data is max distributed to the right and tail flows towards left, hence it is left skewed data.
* Nature of skewness is negatively skewed.

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

**Ans**- Using formula we can calculate IQR.

IQR = Q3-Q1 = 18-10 = 8

IQR = 8

Q19) Comment on the below Boxplot visualizations?



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

**Ans** –

* From these boxplots we can say that both the data is normally distributed.
* Boxplot 1 median and Boxplot 2 median is nearly same.
* Boxplot 2 shows max data distribution as compare to Boxplot 1.
* Boxplot 1 shows sample data and Boxplot 2 shows population data for larger data.
* Boxplot 2 Q1= 225 value is less than Boxplot 1 (Q1=250 approx.) and Boxplot 2 Q3= 312 value is greater than Boxplot 1 (Q3=287 approx.)

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

Data \_set: Cars.csv

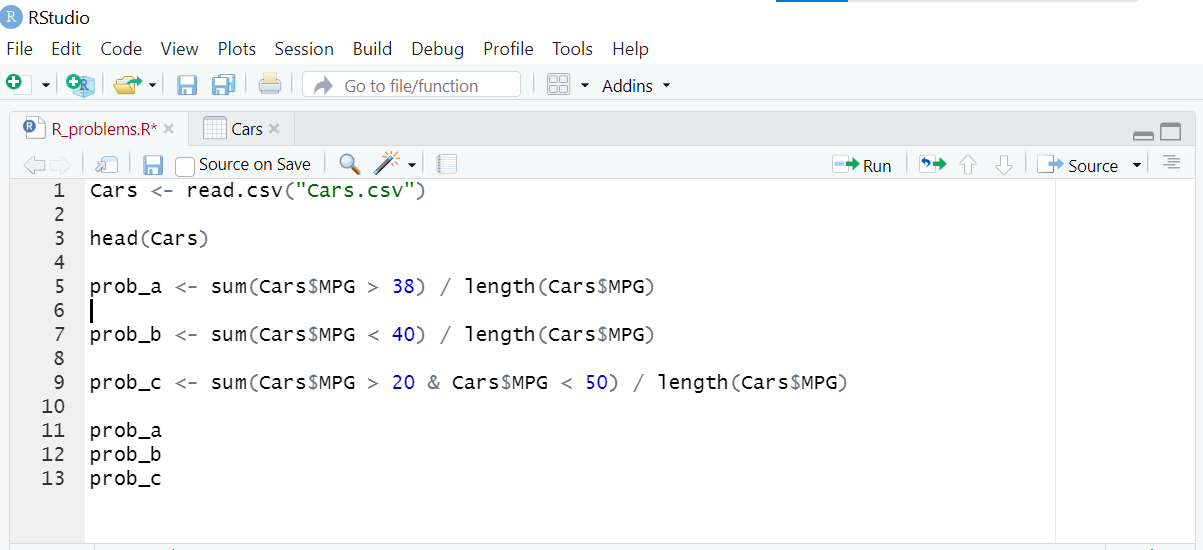
Calculate the probability of MPG of Cars for the below cases.

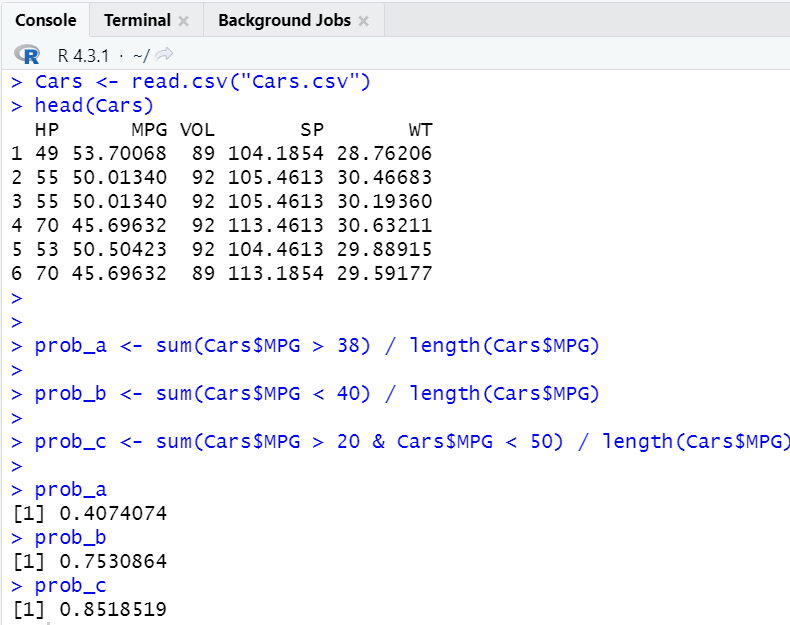
MPG <- Cars$MPG

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

c. P (20<MPG<50)

**Ans** – Solved in R File.



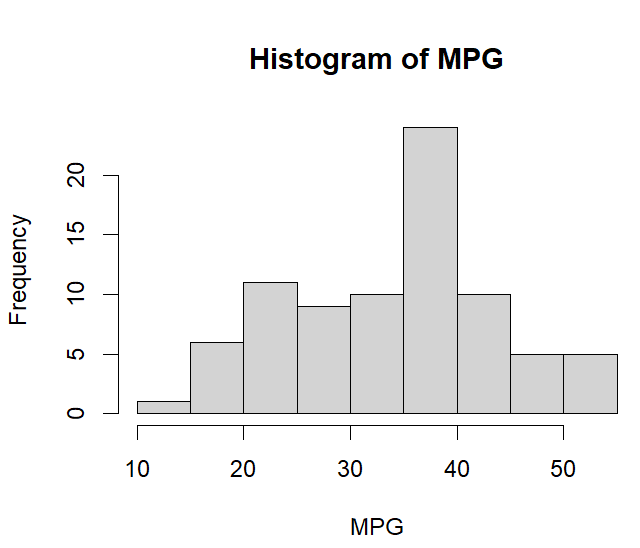


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

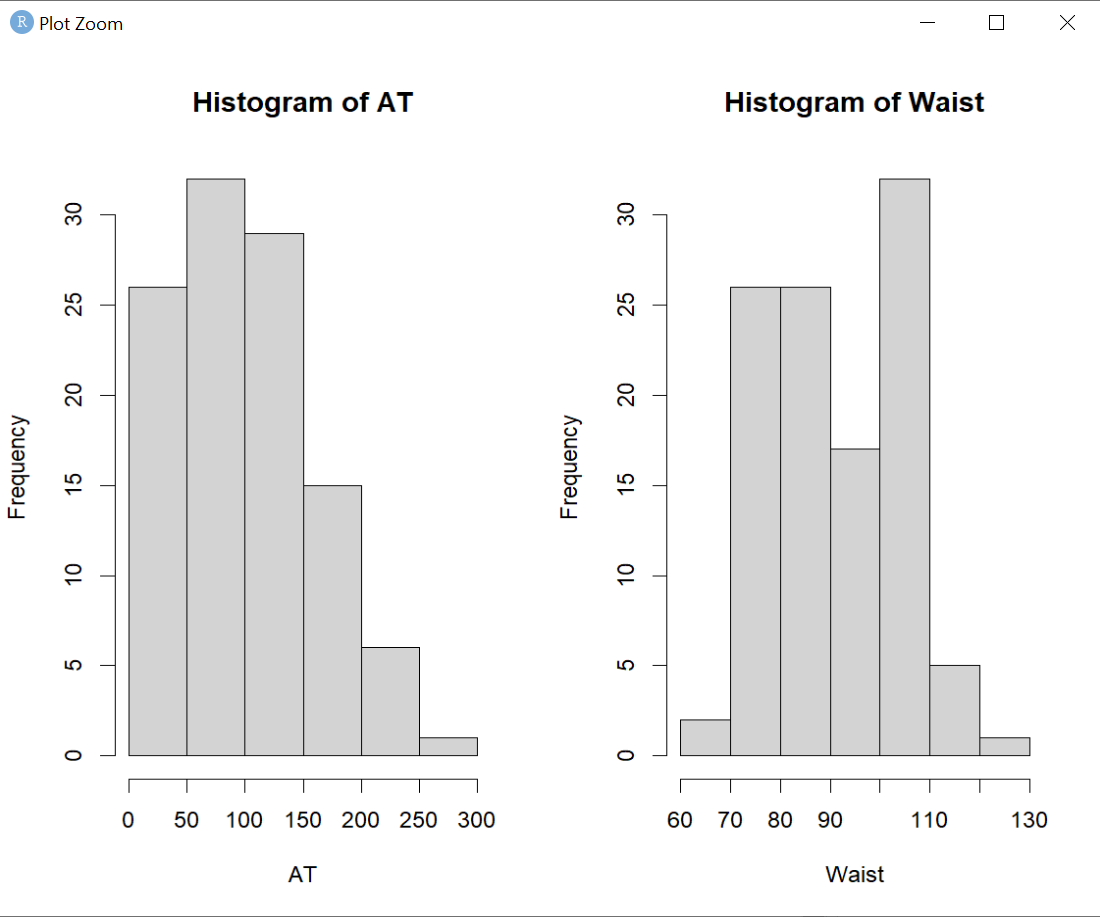
**Ans**- Solved in R File.



* From this above histogram we can say that the data is not normally distributed.

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

Dataset: wc-at.csv



•From this above histogram we can say that the data is not normally

distributed.

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

**Ans** – we can use following formula to find Z\_scores for respective CI.

Z=±Z

α/2

where,

α/2 is the significance level divided by 2.

* z\_90 <- qnorm(0.95) # Z scores of 90% confidence interval is ±1.645
* z\_94 <- qnorm(0.97) #Z scores of 94% confidence interval is ±1.881
* z\_60 <- qnorm(0.9) #Z scores of 60% confidence interval is ±0.842

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

**Ans** – For a given confidence level and degrees of freedom (df), the t-score can be calculated using the formula:

t= ±qt((confidence level+1​/2),df) --------1)

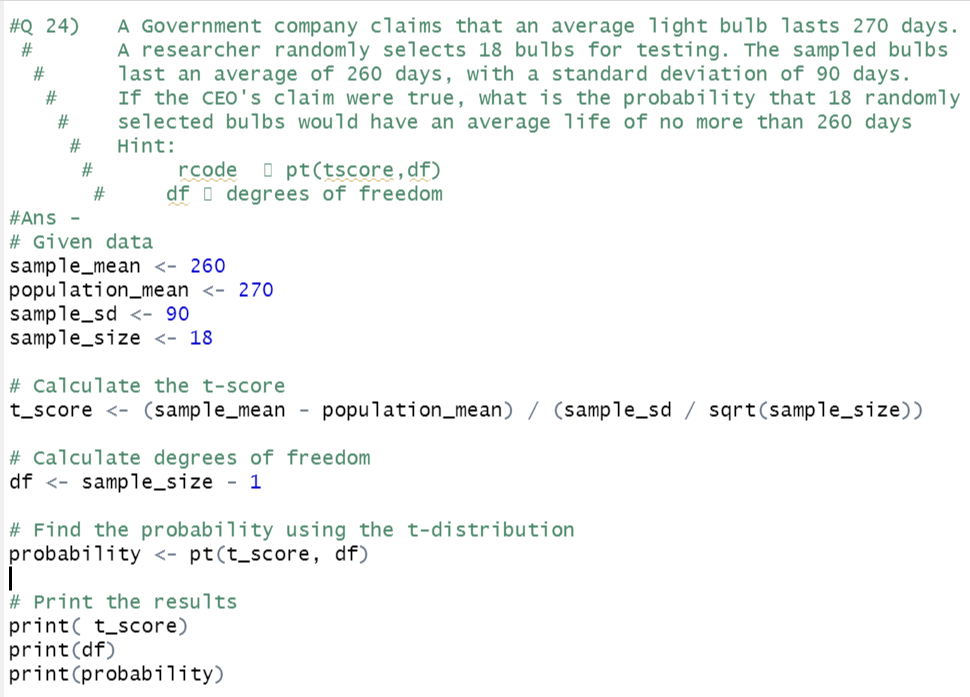
* t score of 95% confidence interval is 2.0638
* t score of 96% confidence interval is 2.171
* t score of 99% confidence interval is 2.796

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

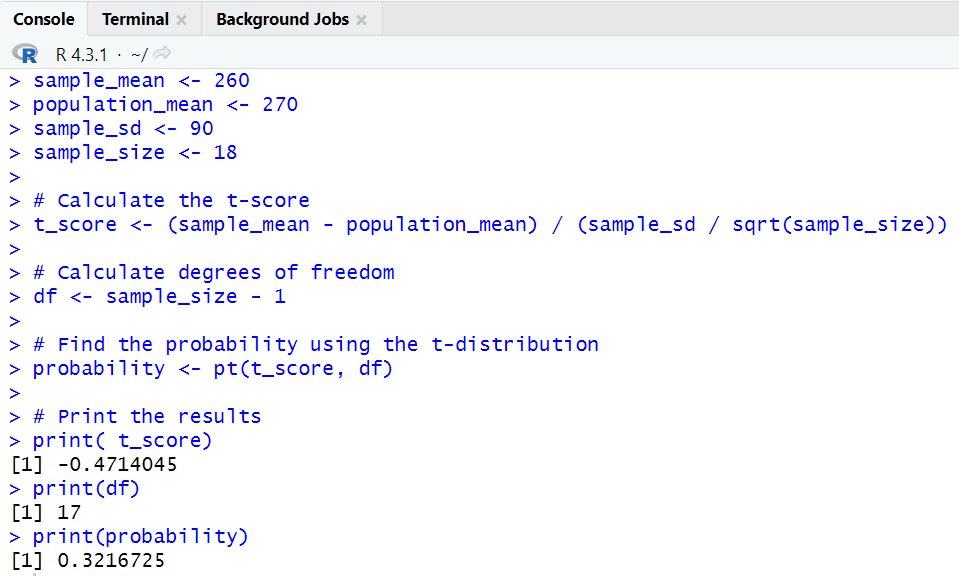
Hint: rcode 🡪 pt(tscore,df) , df 🡪 degrees of freedom

**Ans –** Solved in R File.

**INPUT QUADRANT IN R**

****

**OUTPUT QUADRANT IN R**

****

Thus**,** the probability that 18 randomly selected bulbs would have an average life of no more than 260 days is **0.3216**