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

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 | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

**Answer:**

When 3 coins are tossed then there are 23 = 8 possible outcomes, as each coin can have either Head (H) or Tail (T). The Possible outcomes are

Possible Outcomes = (HHH, HHT, HTH, THH, HTT, THT, TTH, TTT)

The outcome with 2 Heads & 1 Tail are = (HHT, HTH, THH)

Probability of 2 Heads & 1 Tail = Number of favorable outcome (divide by) Total number of outcomes.

Therefore, Probability of 2 Heads & 1 Tail = 3/8 = 0.375 = 37.5%

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:**

Total Probable Outcome = {(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)}

1. **Equal to 1**

There is no outcome where the outcome is 1, so the probability is 0

P (equal to 1) = 0

1. **Less than or equal to 4**

Number of outcomes where the sums are less than or equal to 4 = (1,1), (1.2), (1,3), (2,1), (2,2), (3,1) = 6

Probability that outcome is less than or equal to 4 = (Total Number of outcome that are less than or equal to 4) divide by (Total number of outcome with rolling 2 dice)

P (less than or equal to 4) = 6/36 = 1/6 = 0.166

1. **Sum is divisible by 2 and 3**

**To make it divisible by 2 and 3, the sum should be divisible by 2 and 3, which means it must be a multiple of 6.**

Therefore, Number of outcome where the sum is divisible by 2 and 3 = (1,5), (2,4), (3,3), (4,2), (5,1) = 5

Probability of the outcome where the sum is divisible by 2 and 3 = 5/36

**Probability that the sum is divisible by 2 and 3 = 5/36 = 0.138**

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:**

Total number of ways to draw two balls from the bag without considering their color = 2 balls from 7 (2 red + 3 green + 2 blue) = 7C2 = (7!/(2!(7-2)!) = 21 ways.

The number of ways to draw two balls that are not blue. There are 5 non-blue balls (2 red + 3 green), and you want to choose 2 of them:

Ways to choose 2 non-blue balls from 5 balls = 5C2 = ((5!/(2!(5-2)!)) =10 ways.

Probability that both balls are not blue = (Number of ways to choose 2 non-blue balls) / (Total number of ways to choose 2 balls) = (((5!/(2!(5-2)!))/ (7!/(2!(7-2)!))) = 10/21 = 0.476

**So, the probability that none of the balls drawn is blue is 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

**Answer**

The formula for calculating the expected value (mean) E(X) = ∑I (Xi \* P(Xi))

Where:

(E(X)) is the expected number of candies

Xi is the candy count for child i.

(P(Xi)) is the probability of child (i) having that candy count.

Therefore, E(X) = (1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.0120) = 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24 = 3.09

**So, the expected number of candies for a randomly selected child is 3.09 candies.**

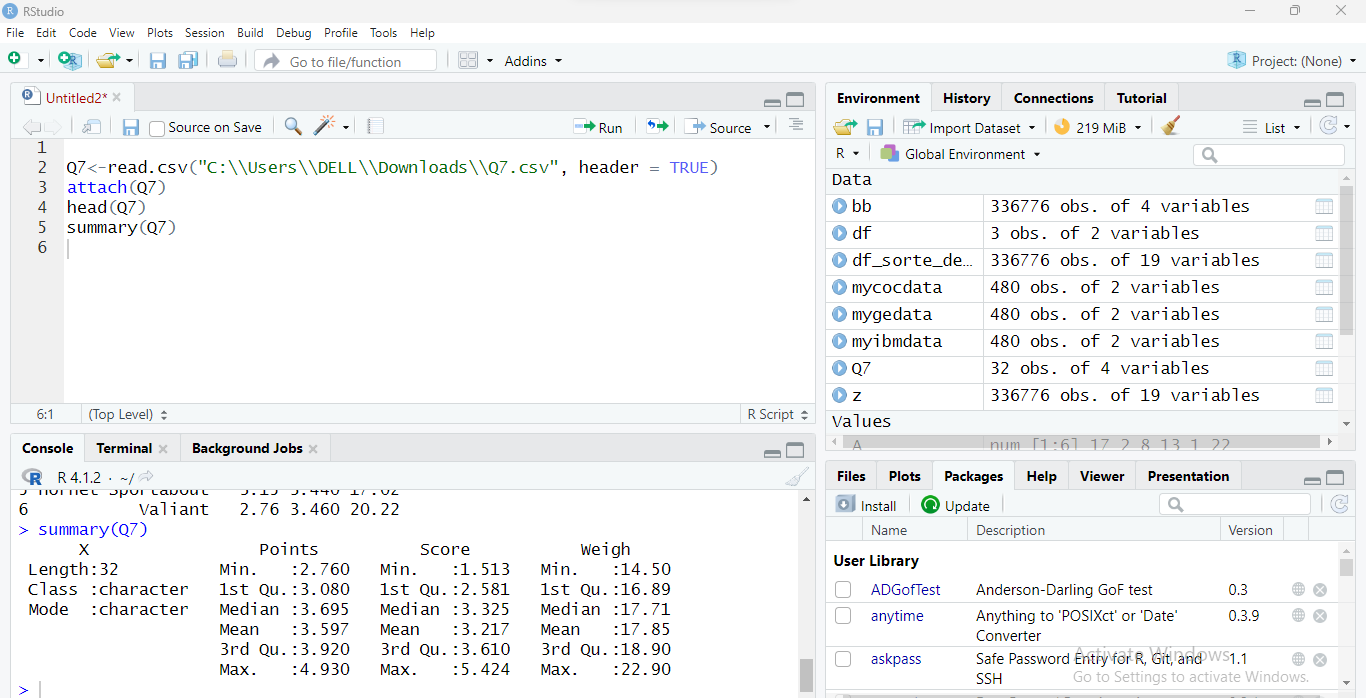
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

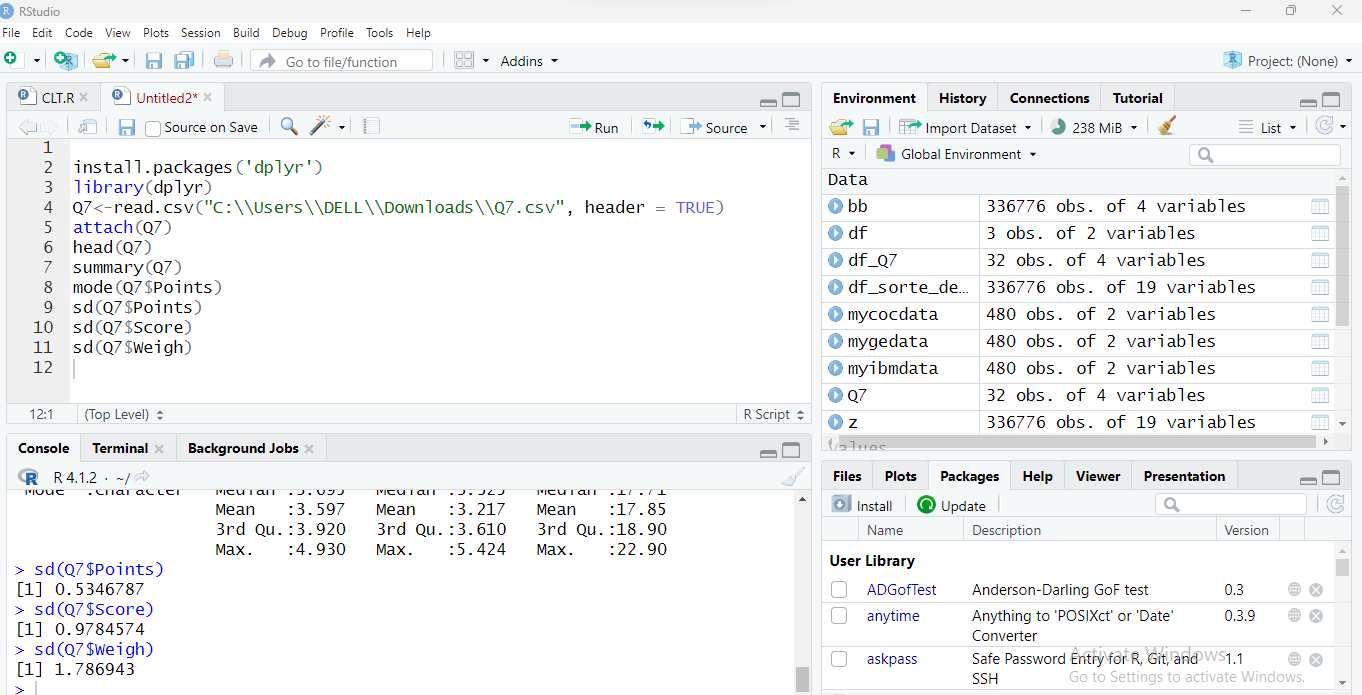
* For Points, Score, Weigh>

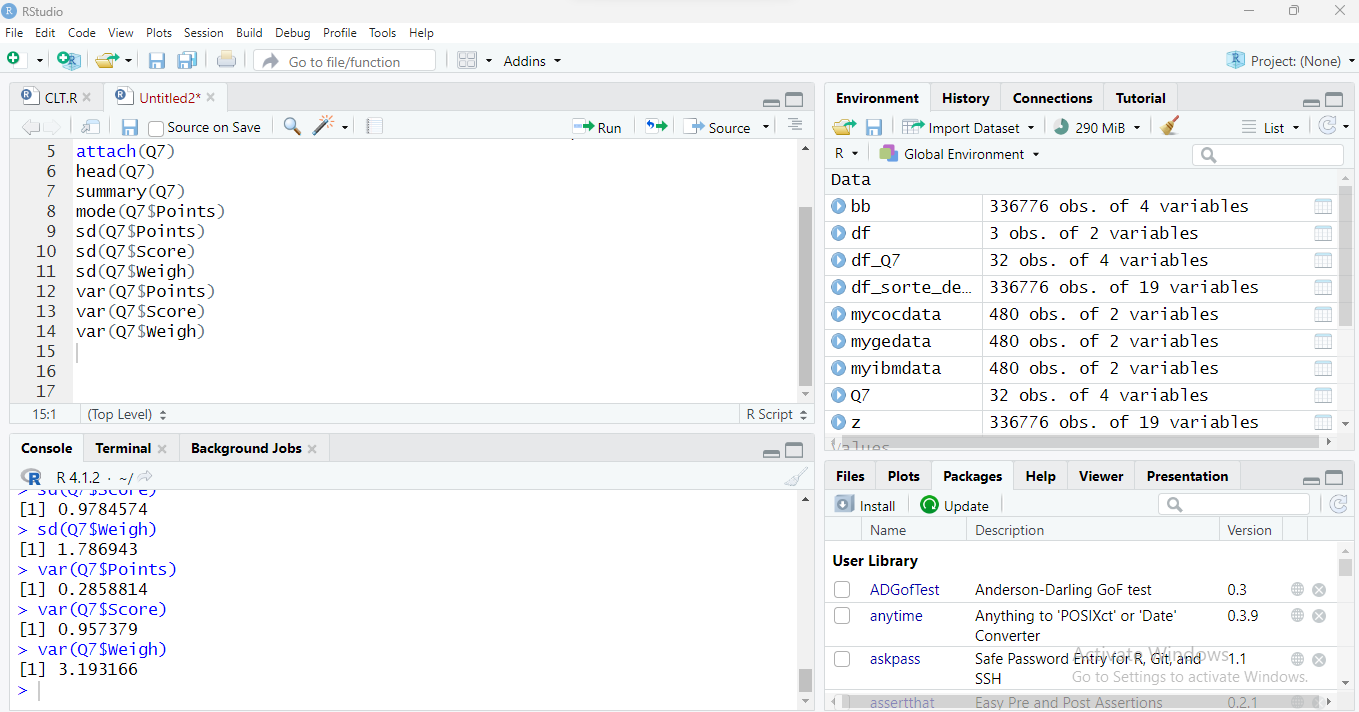
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

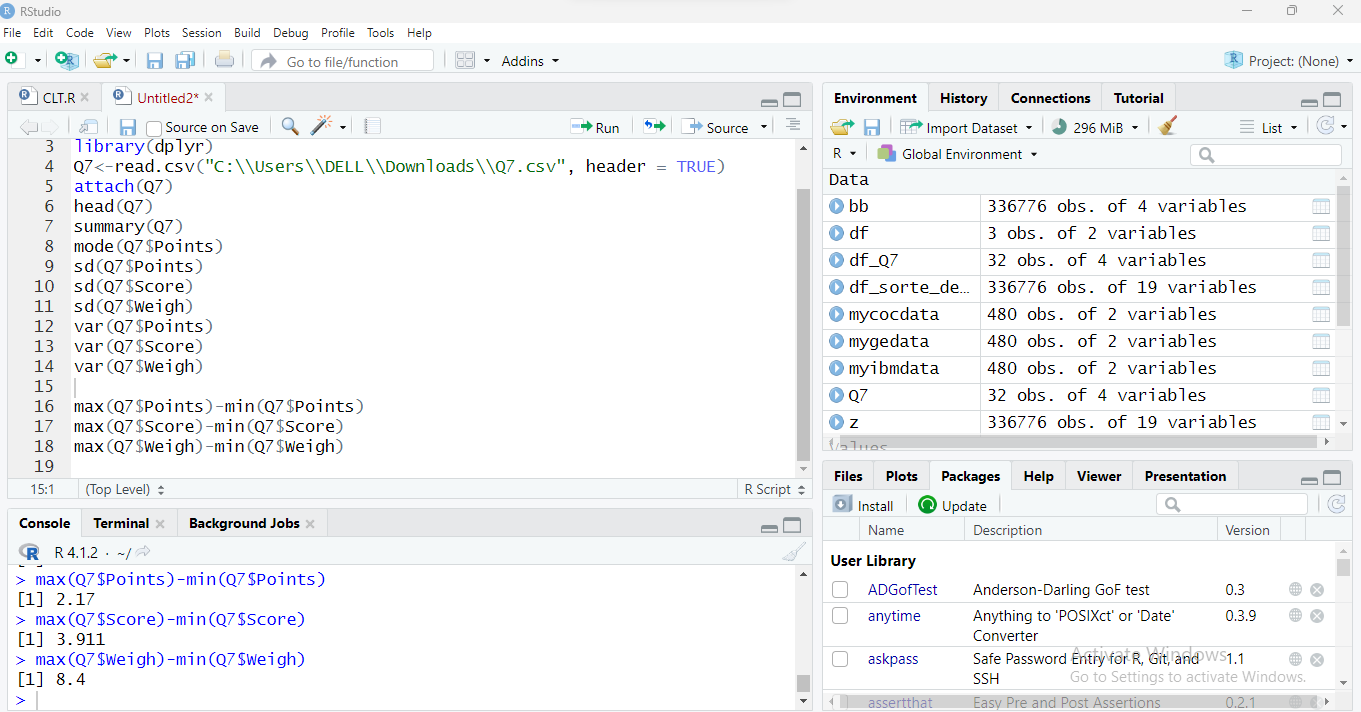
**Use Q7.csv file**

**Answer:**









|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.597 | 3.217 | 17.85 |
| Median | 3.695 | 3.325 | 17.71 |
| Standard Deviation | 0.534 | 0.978 | 1.786 |
| Variance | 0.285 | 0.957 | 3.193 |
| Range | 2.17 | 3.911 | 8.4 |
| Mode | Numeric | Numeric | Numeric |

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?

**Answer:**

For Expected value (Mean) we sum up the weights of all the patients and divide it by total number of observations.

Expected Value = (108+110+123+134+145+167+187+199)/9 = 1173/9 = 130.33

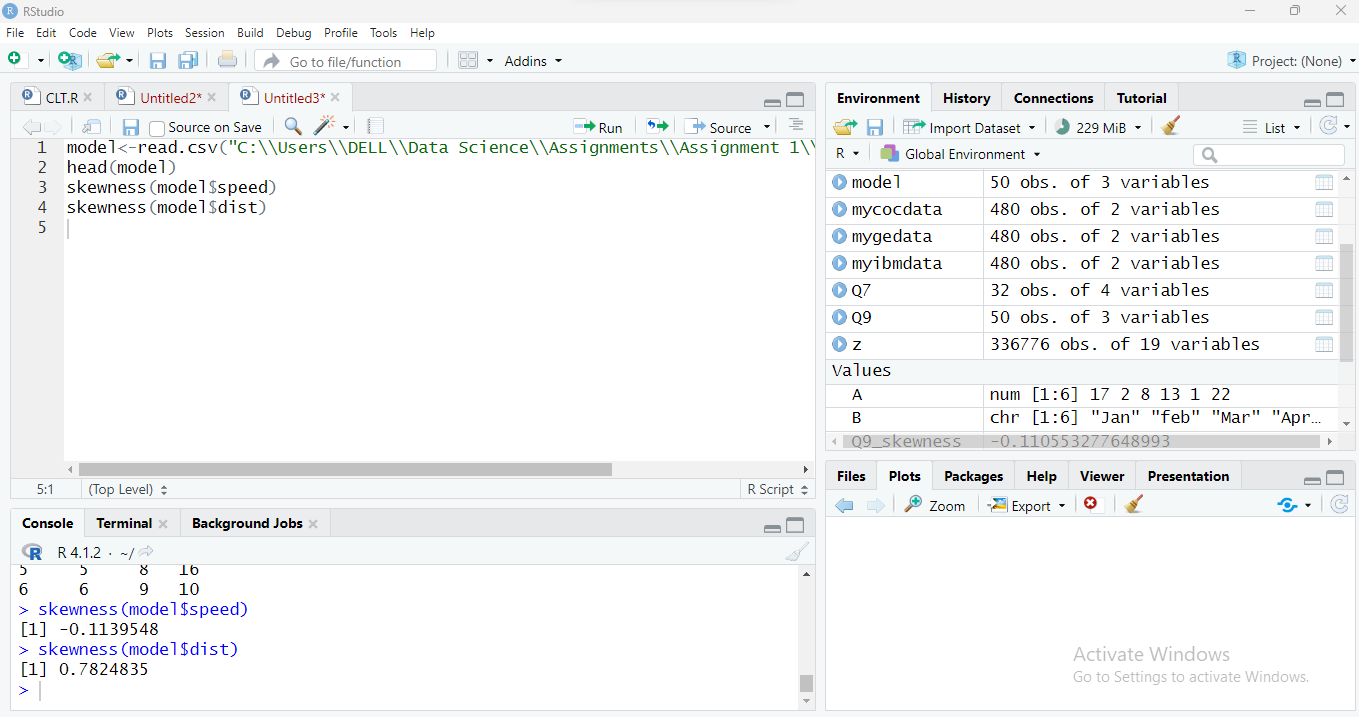
Therefore, the Expected Value of the randomly chosen patient is 130.33 pounds

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Answer:**

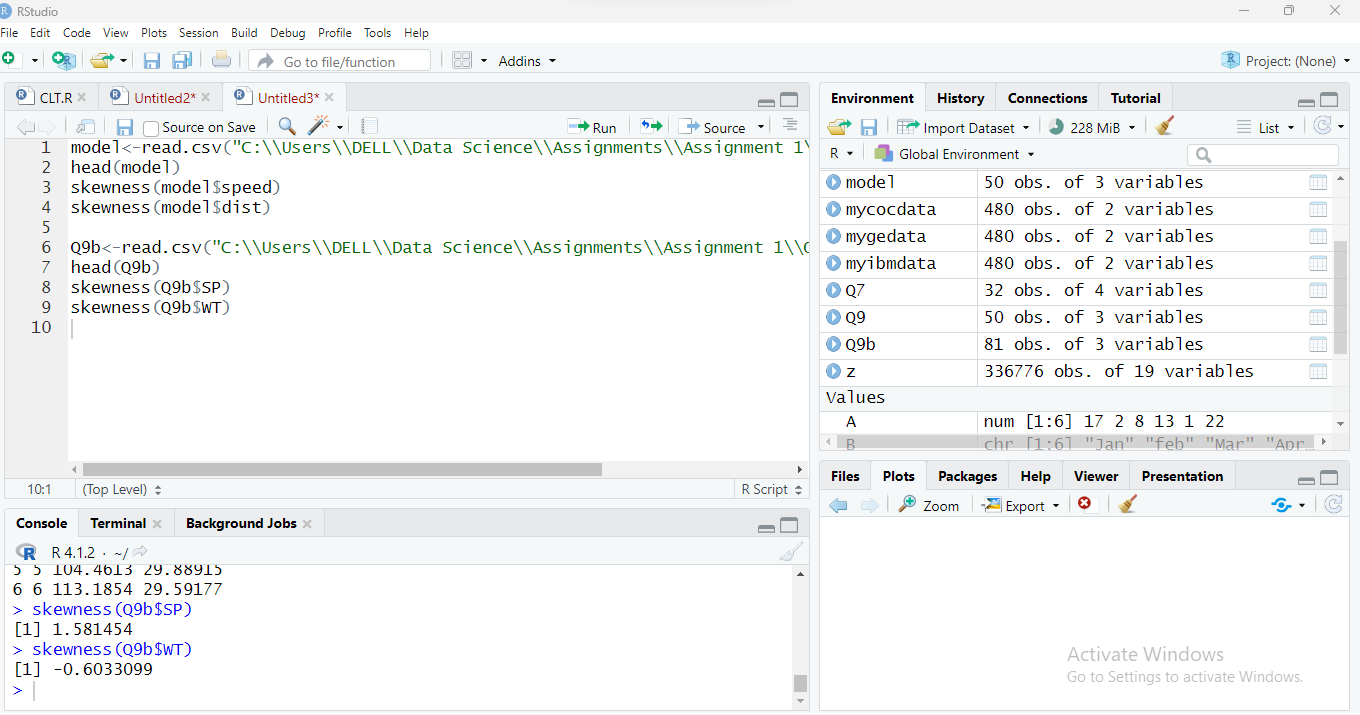
****

**Skewness for Speed:** The skewness for speed is -0.113 which is negative. This implies that the distribution is left skewed.

**Skewness for Distance:**  The skewness for distance is 0.782 which is positive. This implies that the distribution is slightly skewed towards right.

**SP and Weight(WT)**

**Use Q9\_b.csv**

****

**Skewness for Speed:**  The skewness for speed is 1.581 which is positive. Hence it is right skewed.

**Skewness for Weight:**  The skewness for weight is -0.6033 which is negative. This implies that the it is left skewed.

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



**Answer:**

The above histogram represents the distribution of Chick weights with median weight of 200. There is a slight right skew in the distribution reflecting that large number of chicks are below the median wieght. There are also few outliers with few chicks weighing belo 100 and few wieghing above 400.



**Answer:**

The above box plot suggests that the median is less than mean reflecting that the data is right skewed and there are outliers too in the given data. Also, there is a less data point between Q1 and bottom point.

**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:**

Sample Size = 2000

Sample Mean (Average weigh of sample) = 200

Standard Deviation of the Sample = 30

Confidence Interval = Sample Mean + & - (Z score \* Sample Standard Deviation)/Square root of Sample Size

For 94% confidence interval, Z score is approximately 1.881

Confidence interval for 94% = 200 + 1.881 (30/√2000) & 200 - 1.881 (30/√2000)

Confidence interval for 94% = 200+1.26 & 200-1.26 = 201.26 & 198.74

For 98% Confidence Interval Z score is approximately 2.33

Confidence Interval for 98% = 200 + 2.33(30/√2000) & 200-2.33 (30)/√2000)

Confidence Interval for 98% = 200+1.56 & 200 – 1.56 = 201.56 & 198.44

For 96% confidence interval Z score is 2.05

Confidence interval for 96% = 200 + 2.05 (30/√2000) & 200 – 2.05 (30/√2000)

Confidence interval for 96% = 200+1.37 & 200-1.37 = 201.37 & 198.63

**Therefore,**

**Confidence interval for 94% = 201.26 & 198.74**

**Confidence Interval for 98% = 201.56 & 198.44**

**Confidence interval for 96% = 201.37 & 198.63**

**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. Find mean, median, variance, standard deviation.

Mean = (34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56)/18

**Mean = 738/18 = 41**

For median, arranging the scores in ascending order

34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

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

Variance = ∑ (X- µ)2/N

Sum of the squared difference from the mean for each data point (∑ (X- µ)2 ) = 434

|  |  |
| --- | --- |
| Data Point | Sum of Squared Difference from the mean for each data point |
| 34 | (34-41)^2 = 49 |
| 36 | (36-41)^2 = 25 |
| 36 | (36-41)^2 = 25 |
| 38 | (38-41)^2 = 9 |
| 38 | (38-41)^2 = 9 |
| 39 | (39-41)^2 = 4 |
| 39 | (39-41)^2 = 4 |
| 40 | (40-41)^2 = 1 |
| 40 | (40-41)^2 = 1 |
| 41 | (41-41)^2 = 0 |
| 41 | (41-41)^2 = 0 |
| 41 | (41-41)^2 = 0 |
| 41 | (41-41)^2 = 0 |
| 42 | (42-41)^2 = 1 |
| 42 | (42-41)^2 = 1 |
| 45 | (45-41)^2 = 16 |
| 49 | (49-41)^2 = 64 |
| 56 | (56-41)^2 = 225 |
|  | Sum = 434 |

**Variance = ∑ (X- µ)2/N = 434/18 = 24.11**

**Standard Deviation = √∑ (X- µ)2/N = √24.11 = 4.91**

1. What can we say about the student marks?

**Answer:**

The average marks of the students are 41 and the with a variability of 24.11 and standard deviation 4.91 in the scores. The data is not highly skewed.

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

**Answer:**

When mean and median are equal, its suggests that data is not skewed and symmetrically distributed equally on both the left & right side of the central value in the distribution. It also means there is no distance between mean and median, that is mean and median are located at the centre of the distribution.

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

**Answer:**

The nature of the skewness will be Positive. That is the dataset is distributed towards right side and the distribution is positively skewed with tail on the right side being longer.

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

**Answer**

The nature of the skewness will be Negative. That is the dataset is distributed towards left side and the distribution is negatively skewed with tail on the left side being longer.

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

**Answer:**

Positive Kurtosis value indicates that the dataset has a higher peak than the normal distribution. It also suggests that the dataset has higher number of outliers than that would have been expected in a normal distribution.

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

**Answer**

Negative Kurtosis indicates that the dataset has a lower & broader peak than the normal distribution. It also suggests that the dataset has fewer number of outliers than that would have been expected in a normal distribution.

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



What can we say about the distribution of the data?

**Answer:** In the above box plot, more than 50% of the data are between 10 to 18. And the data is concentrated between Q3 and Q4.

What is nature of skewness of the data?

**Answer:** The nature of the skewness of the above data is left skewed as the median is greater than mean.

What will be the IQR of the data (approximately)?   
**Answer:** IQR = Q3 – Q1 =14-6 = 8

Q19) Comment on the below Boxplot visualizations?



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

**Answer:**

In the above two plot, it is shown that the second plot has taller whiskers indicating greater spread in the data. Also, there doesn’t seem any outlier in the above box plot and the data is symmetrically distributed.

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

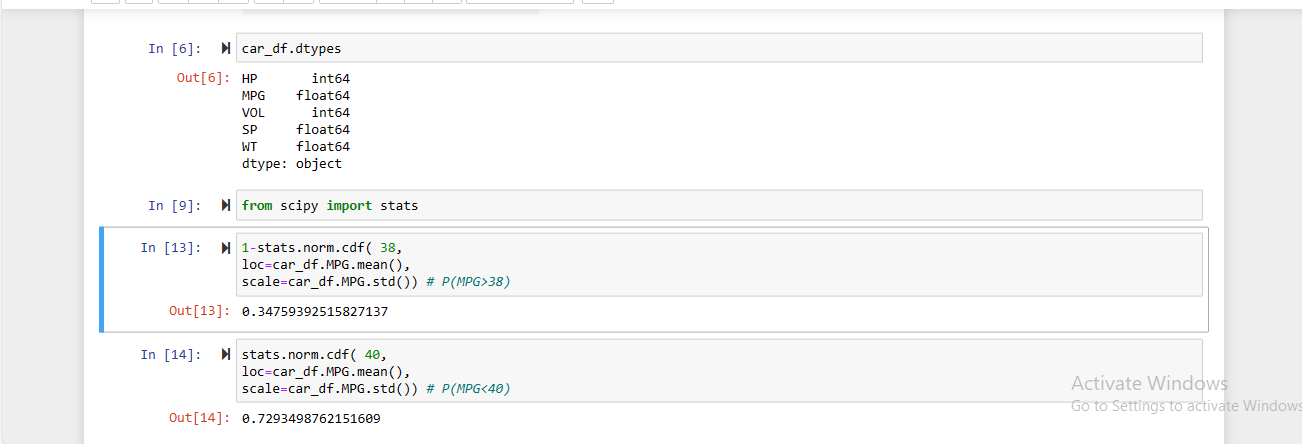
Data \_set: Cars.csv

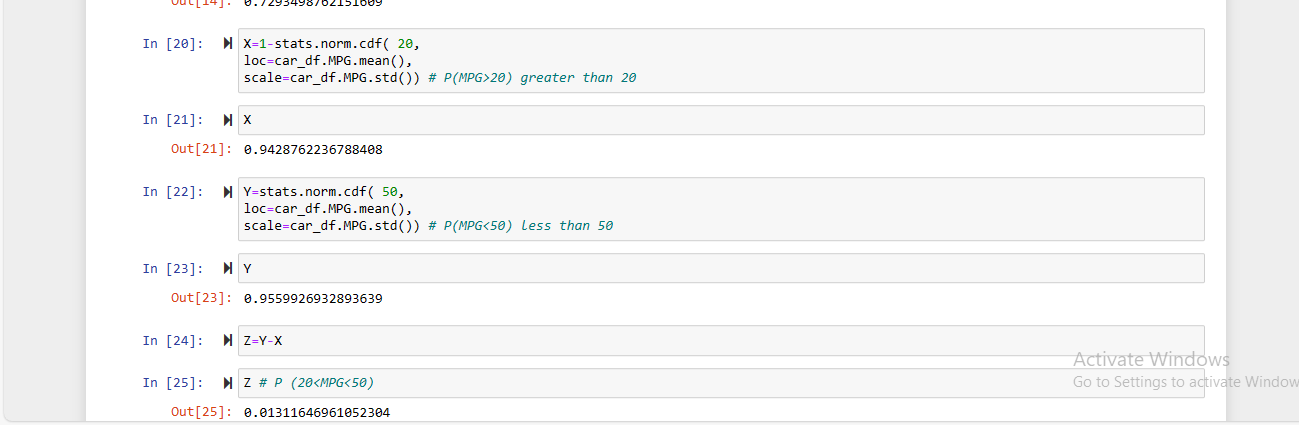
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

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

c. P (20<MPG<50)





**Answers:**

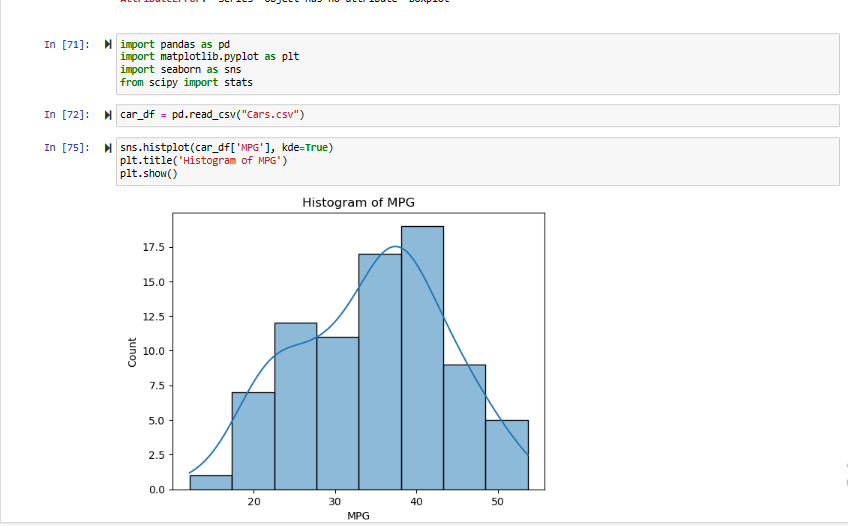
1. P(MPG>38) = 0.34759
2. P(MPG<40) = 0.7293
3. P (20<MPG<50) = 0.0131

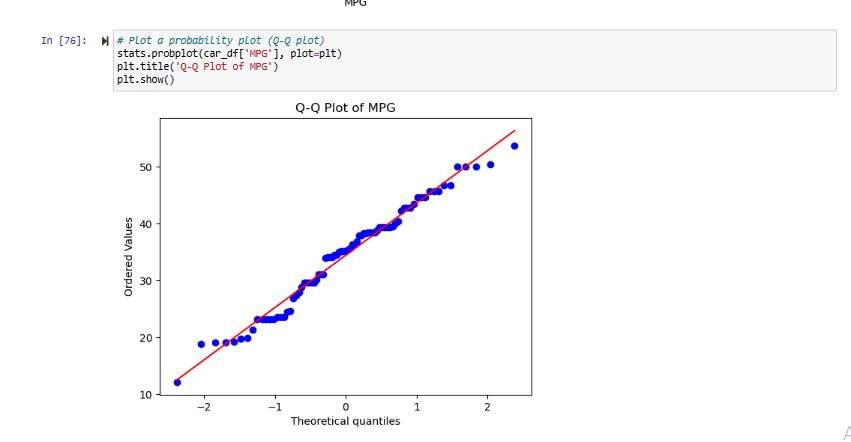
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Answers:**



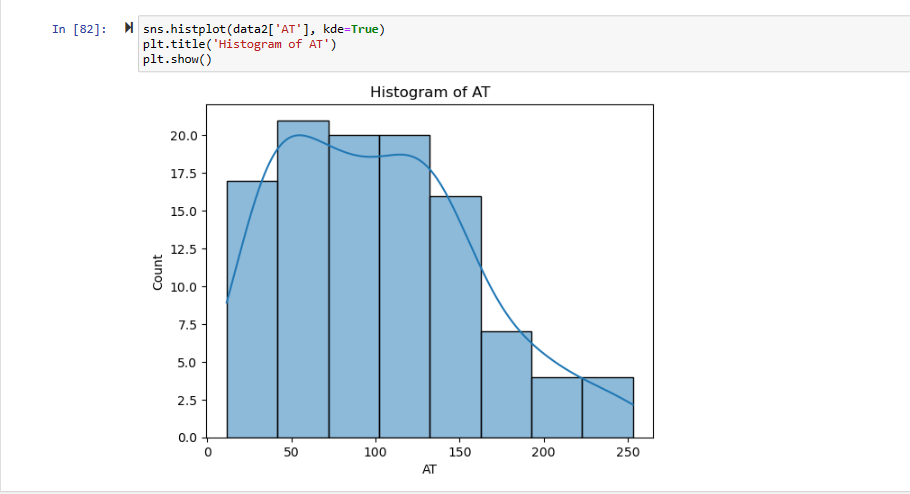


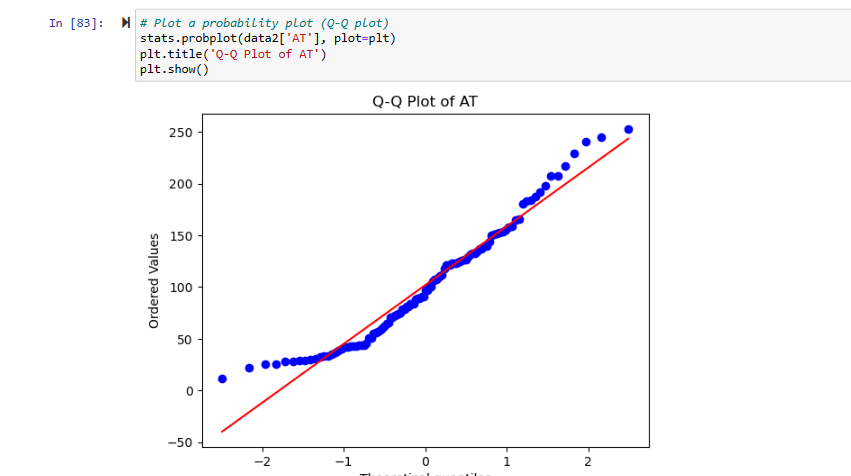
The above histogram & QQ plot show that the data is more or less normally distributed with majority of the data concentrated in the left side of the histogram.

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

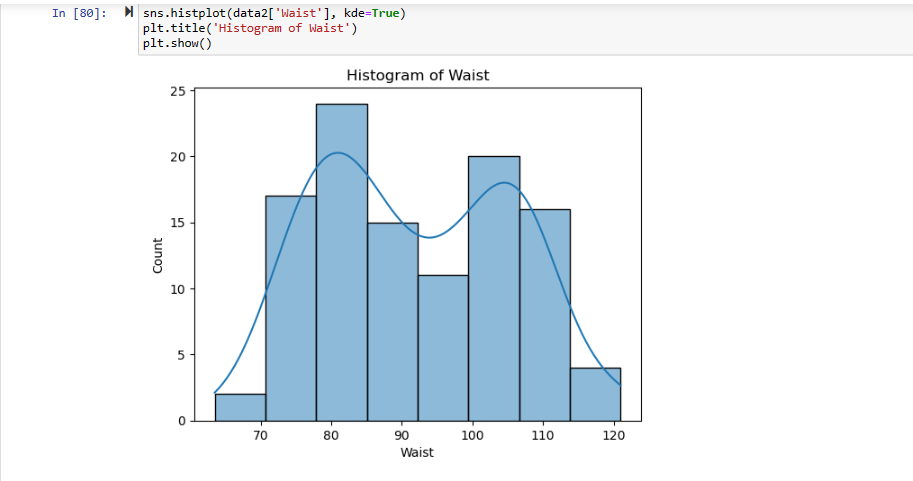
Dataset: wc-at.csv

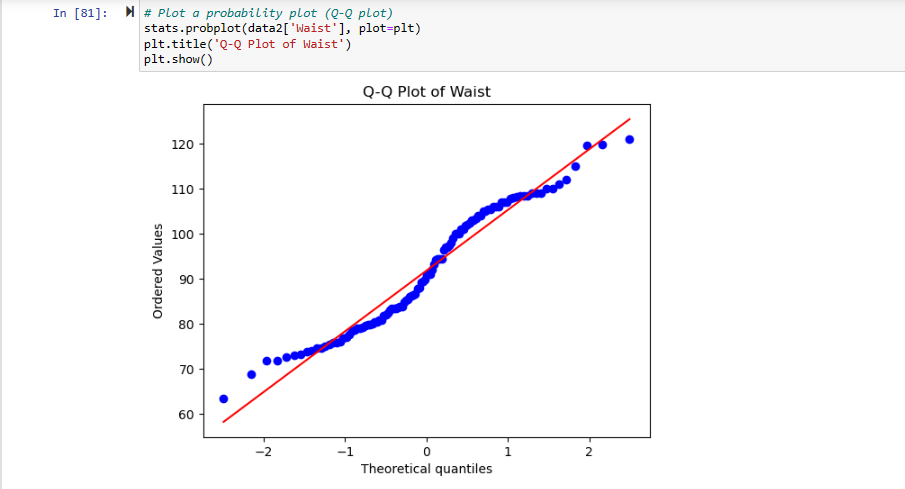
**Answers:**

****

****

The above histogram & QQ plot of the AT clearly suggests that the AT does not follow a normal distribution. Majority of the data are concentrated in the left side of the histogram indicating that the data is negatively skewed.





The above two chart clearly reflects that the Waist does not follow a normal distribution.

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

**Answers:**

The Z-scores of 90% confidence interval is 1.65

The Z-scores of 94% confidence interval is 1.88

The Z-scores of 60% confidence interval is 0.84

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

**Answer:**

The t-score can be found using the t-tables.

The given sample size is 25, therefore the degree of freedom will be = n-1 = 25-1=24

For 95% confidence interval, t-score = 2.064

For 96% confidence interval, t-score = 2.177

For 99% confidence interval, t-score = 2.797

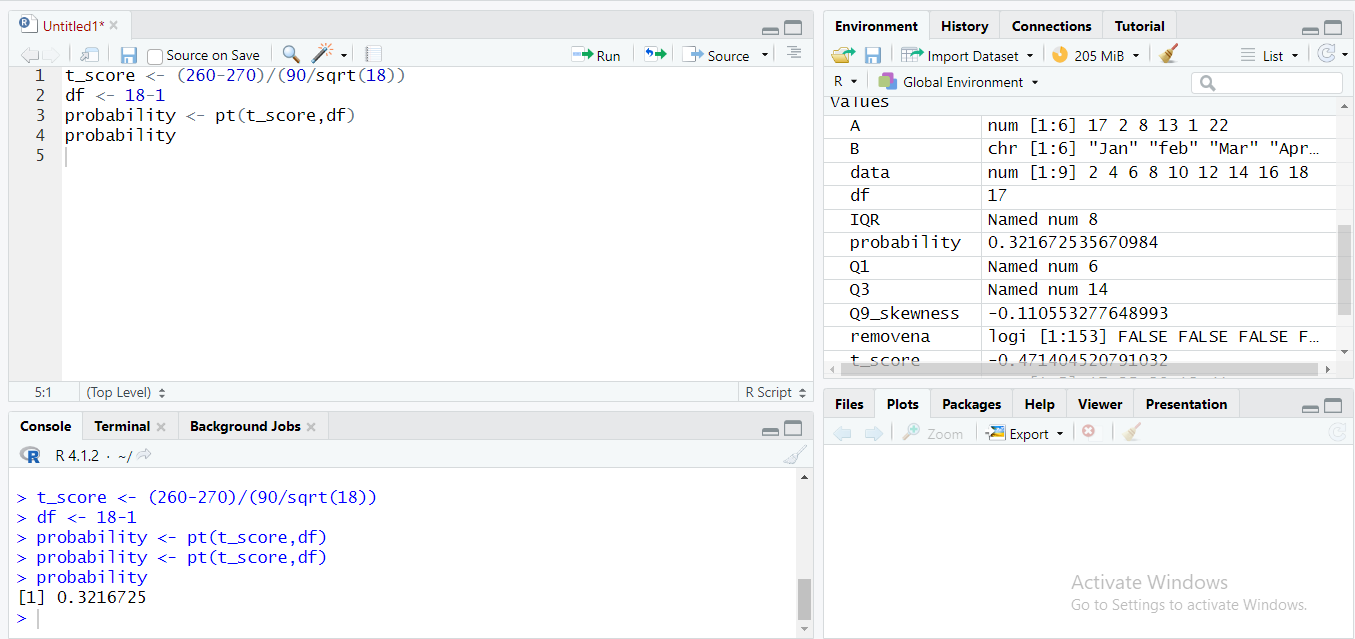
Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

**Answers:**



Therefore, the probability of the bulbs that are lasting less than 260 days on average is 0.3216