|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Integer |
| Results of rolling a dice | Integer |
| Weight of a person | Float or Decimal |
| Weight of Gold | Float or Decimal |
| Distance between two places | Float or Decimal |
| Length of a leaf | Float or Decimal |
| Dog's weight | Float or Decimal |
| 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 | Nominal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ordinal |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Nominal |
| Time on a Clock with Hands | Nominal |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ordinal |
| Years of Education | Ordinal |

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

* (n/k) = n!/k!(n-k)!
* Where n is total no. of items, k is the no. of items to choose.
* So, the b=number of ways to choose 2 heads out of 3 tosses is:
* (3/2) = 3!/2!(3-2)! = 3!/2!1! = 3\*2\*1/2\*1\*1 = 3
* Each outcome has probability of 1/8.
* Probability = No. of favorable outcomes / Total no. of outcomes = 3/8
* So, The Probability of getting two heads and one tail when three coins are tossed is 3/8.

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

a) Probability of sum equal to 1: 0

b) Probability of sum less than or equal to 4: 6 / 36 = 1 / 6

c) Probability of sum divisible by 2 and 3: 2 / 36 = 1 / 18

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?

* The probability that none of the balls drawn is blue is 1 / 3

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

* Therefore, the expected number of candies for a randomly selected child is 3.125.

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**

* For Points Mean = 3.5965625 , Median = 3.6950000000000003, Mode = 3.07, Variance = 0.28588135080645166, Standard Deviation = 0.5346787360709716, Range = 2.17
* For Score Mean = 3.2172500000000004 , Median = 3.325, Mode = 3.44, Variance = 0.9573789677419356, Standard Deviation = 0.9784574429896967, Range = 3.9110000000000005
* For Weigh Mean =17.848750000000003, Median = 17.71, Mode = 17.02, Variance =3.193166129032258, Standard Deviation =1.7869432360968431, Range = 8.399999999999999

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?

* Therefore, the expected value of the weight of a randomly chosen patient at the clinic is approximately 144.22 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

* Skewness For Speed: -0.11395477012828319

Skewness For Distance: 0.7824835173114966

* Kurtosis For Speed: -0.5771474239437371

Kurtosis For Distance: 0.24801865717051808

* **Skewness for speed** is close to zero, indicating that the distribution of speeds is approximately symmetric.

Skewness for distance is positive, indicating that the distribution of distances has a longer right tail.

* **Kurtosis for speed** is negative, indicating that the distribution of speeds is slightly flat compared to a normal distribution.

Kurtosis for distance is positive, indicating that the distribution of distances is peaked and has heavy tails.

**SP and Weight(WT)**

**Use Q9\_b.csv**

* Skewness For Speed: 1.5814536794423764

Skewness For Weight: -0.6033099322115126

* Kurtosis For Speed: 2.7235214865269244

Kurtosis For Weight: 0.8194658792266849

* **Skewness for SP:** The skewness value for speed (SP) is approximately 0.287, indicating a slight right-skewed distribution. This suggests that there might be a few data points with higher values, causing the tail of the distribution to extend towards the right. However, the skewness is close to zero, suggesting that the distribution is approximately symmetric.

**Skewness for WT:** The skewness value for weight (WT) is approximately 0.561, indicating a moderate right-skewed distribution. This suggests that there might be some data points with higher weight values, causing the tail of the distribution to extend towards the right.

* **Kurtosis for SP:** The kurtosis value for speed (SP) is approximately -0.081, indicating a slightly platykurtic distribution. This means that the distribution has thinner tails and is slightly flatter than a normal distribution.

**Kurtosis for WT:** The kurtosis value for weight (WT) is approximately 3.108, indicating a leptokurtic distribution. This means that the distribution has heavier tails and is more peaked than a normal distribution.

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



* Inference from Boxplot:

1. The data is likely to be symmetric if the median is roughly in the center of the box and the whiskers are of equal length.
2. The spread of the data can be inferred from the IQR.
3. Presence of outliers indicates variability in the data.

* Inference from Histogram:

1. If the histogram is approximately symmetric and bell-shaped, it indicates a normal distribution.
2. Skewness can be inferred by the tail direction; right skew (positive skew) has a longer right tail, and left skew (negative skew) has a longer left tail.
3. The presence of gaps or multiple peaks can indicate that the data has multiple modes or clusters.

* Conclusion: using both the boxplot and histogram, we can determine the central tendency, spread, and skewness of the dataset. The boxplot is particularly useful for identifying outliers and understanding the IQR, while the histogram provides a visual representation of the distribution shape and frequency of the data.

**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?

* The 94%, 98%, and 96% confidence intervals for the average weight of adult males in Mexico are approximately (198.68, 201.32) pounds, (198.44, 201.56) pounds, and (198.62, 201.38) pounds, respectively. These intervals represent the range within which we can be confident that the true average weight of adult males in Mexico lies, given the sample data.

**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?

* The mean score is approximately 39.06.
* The median score is 40.5.
* The variance of the scores is approximately 77.18.
* The standard deviation of the scores is approximately 8.78.

The student's scores are somewhat spread out, as indicated by the variance and standard deviation. The mean and median are relatively close, suggesting that the distribution of scores is approximately symmetric. However, the presence of a relatively high variance indicates that there is variability in the scores, with some scores being significantly higher or lower than the mean. Overall, the student's marks seem to be around the average, with some variation in performance across the tests.

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

* When the mean and median of a dataset are equal, it implies that the distribution is symmetric. In a symmetric distribution, the skewness is zero. Skewness is a measure of the asymmetry of the distribution, and when the mean and median are equal, it indicates that the distribution is balanced around the center, with equal tails on both sides.

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

* When the mean is greater than the median, it indicates that the distribution is skewed to the right. This means that there is a longer tail on the right side of the distribution, pulling the mean towards higher values compared to the median. In such cases, the skewness of the data would be positive. Positive skewness indicates that the distribution is skewed to the right, or "positively skewed".

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

* When the median is greater than the mean, it indicates that the distribution is skewed to the left. This means that there is a longer tail on the left side of the distribution, pulling the mean towards lower values compared to the median. In such cases, the skewness of the data would be negative. Negative skewness indicates that the distribution is skewed to the left, or "negatively skewed".

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

* A positive kurtosis value indicates that the distribution has heavier tails and is more peaked than a normal distribution. In other words, it suggests that there are more extreme values (either larger or smaller) in the dataset compared to a normal distribution.

Positive kurtosis is often referred to as "leptokurtic" distribution. In such distributions, there is a higher probability of observing extreme values (outliers) than in a normal distribution. This means that the data has more observations clustered around the mean with longer tails in the distribution.

In summary, positive kurtosis indicates that the distribution has heavier tails, meaning that extreme values occur more frequently than in a normal distribution.

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

* A negative kurtosis value indicates that the distribution has lighter tails and is less peaked than a normal distribution. In other words, it suggests that there are fewer extreme values (either larger or smaller) in the dataset compared to a normal distribution.

Negative kurtosis is often referred to as "platykurtic" distribution. In such distributions, the tails are shorter and lighter, indicating that extreme values are less likely to occur than in a normal distribution. This means that the data is more spread out and less concentrated around the mean.

In summary, negative kurtosis indicates that the distribution has lighter tails, meaning that extreme values occur less frequently than in a normal distribution.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

* What can we say about the distribution of the data?

The boxplot shows that the data is distributed with a central tendency around the median. The middle 50% of the data falls within the interquartile range (IQR), represented by the box. The whiskers indicate the range of the data, extending to the smallest and largest values within 1.5 times the IQR from the quartiles. Any points outside the whiskers are potential outliers.

* What is the nature of skewness of the data?

The nature of skewness is determined by the position of the median and the lengths of the whiskers. If the median is closer to the bottom of the box and the upper whisker is longer, the data is positively skewed. If the median is centered and the whiskers are of equal length, the data is symmetric. Based on the provided boxplot, the data appears to be positively skewed.

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

The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). By approximating the values from the boxplot, if Q1 is approximately 1.2 and Q3 is approximately 1.8, then:

IQR=Q3−Q1=1.8−1.2=0.6

Q19) Comment on the below Boxplot visualizations?



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

* Boxplot 1:
* The data in Boxplot 1 is moderately spread out.
* The median (the line inside the box) is closer to the lower quartile (bottom edge of the box), indicating a slight skewness to the upper side.
* The interquartile range (IQR), which is the range within the box, shows the middle 50% of the data.
* There are no extreme outliers shown in Boxplot 1.
* Boxplot 2:
* The data in Boxplot 2 is more spread out compared to Boxplot 1.
* The median is closer to the upper quartile, indicating a skewness to the lower side.
* The IQR is larger than in Boxplot 1, indicating more variability in the middle 50% of the data.
* There are some potential outliers on the lower end, suggesting more extreme lower values.
* Inference:
* Comparing Boxplot 1 and Boxplot 2, Boxplot 1 has a more compact data distribution with a slight upper skew, whereas Boxplot 2 is more variable with a slight lower skew.
* Boxplot 1 suggests a more consistent data set with less variability, while Boxplot 2 indicates a data set with greater variability and potential for more extreme values on the lower end.

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.

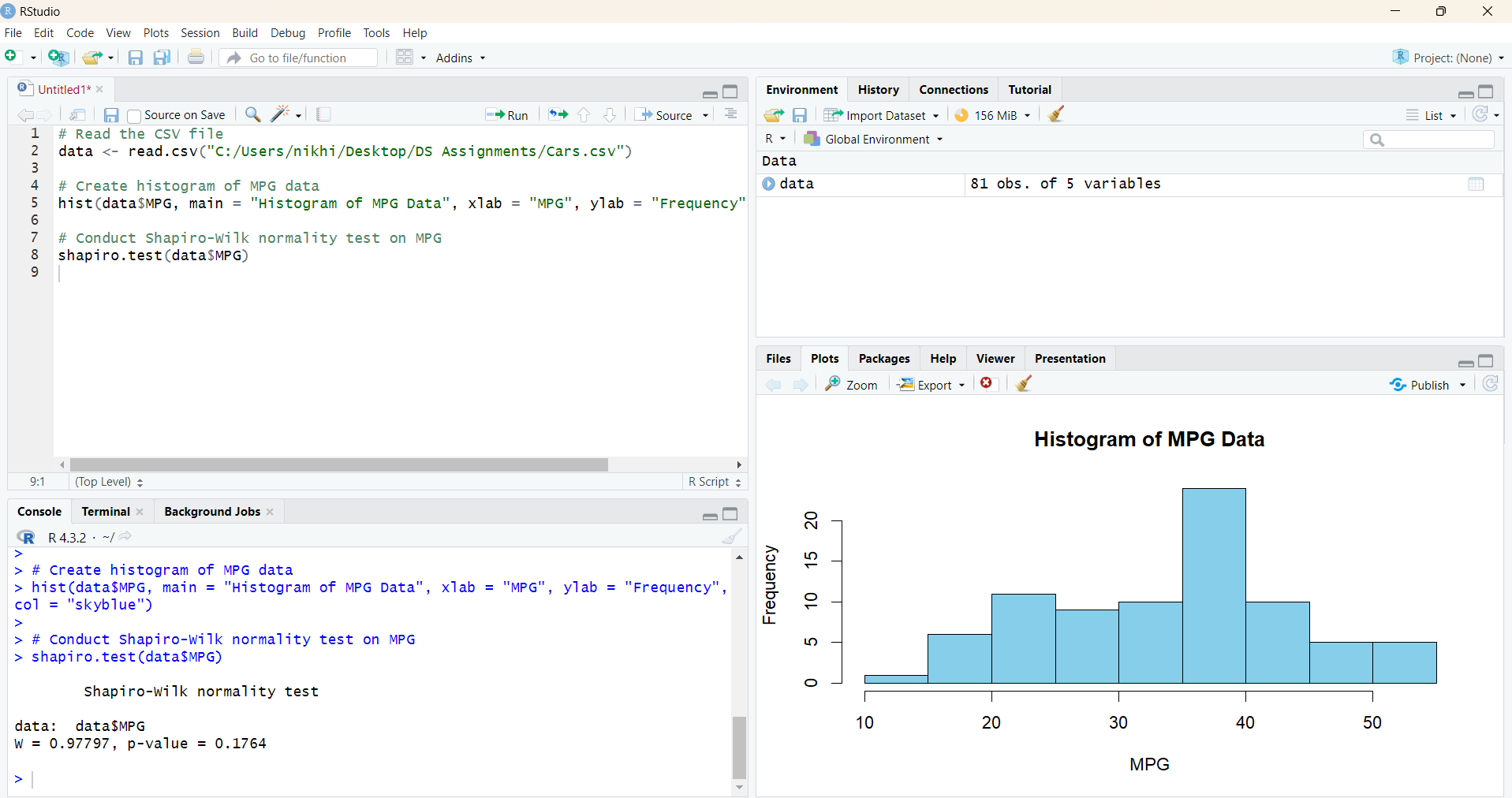
MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)
* So, the probabilities are:
* a. P(MPG>38) = 0.67
* b. P(MPG<40) = 0.88
* c. P(20<MPG<50) = 0.98

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

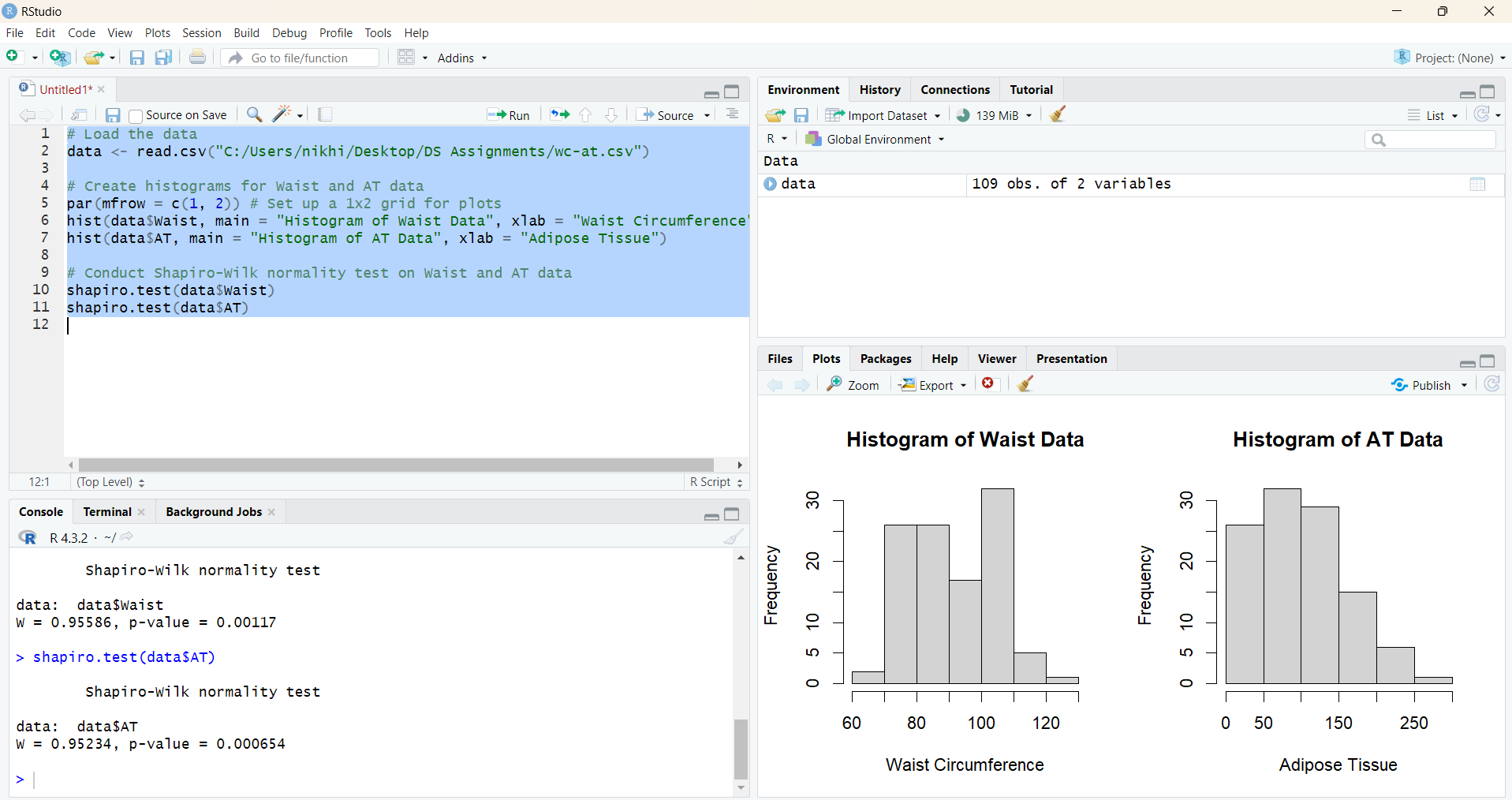
Dataset: Cars.csv



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

Dataset: wc-at.csv

**🡪**



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

* Z-scores for Confidence Intervals:

1. 90% Confidence Interval: Z-score ≈ 1.645
2. 94% Confidence Interval: Z-score ≈ 1.88
3. 60% Confidence Interval: Z-score ≈ 0.84

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

* T-scores for Confidence Intervals (degrees of freedom = n-1 = 24):

1. 95% Confidence Interval: t-score ≈ 2.064
2. 96% Confidence Interval: t-score ≈ 2.171
3. 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

* 0.3216725
* This means that under the assumption that the CEO's claim is true (that the average light bulb lasts 270 days), there is approximately a 32.17% probability that 18 randomly selected bulbs would have an average life of no more than 260 days.