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

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

Ans- Two heads and one tail=(HHT, THH,HTH)

So probability = 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

Ans:- a) Equal to 1

0/36

b) Less than equal to 4

6/36

1. Sum is divisible by 2 and 3

3/36

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?

2 RED: R1 R2

3 GREEN: G1 G2 G3

2 BLUE: B1 B2

2 balls are drawn randomly.

Starting with R1 which is drawn along with either of the remaining 6 balls. Means possible 6 ways for R1.… (1)

Then R2 is drawn with either of the remaining 5 balls. Means possible 5 ways for R2…….(2)

Similarly, we draw G1, G2, G3, B1, B2 . For each possible ways = 4, 3. 2, 1, 0 ………(3)

So, total possible ways = 6+5+4+3+2+1 = 21

So, this is how we calculate by permutation Combination too..

by 7C2 = combination of 7, objects taken 2

= 7! /{ (7–2) ! 2!} = 21

Now excluding B1 B2 , we find the possible ways out.. Starting from R1, paired with R2, G1, G2, G3… ie 4 + 3 + 2+ 1+ 0 = 10

Or, 5C2 = 5! / {(5–2)! 2! } = 10

Hence, probability( event, none of them blue) = (no of outcomes favourable to the event) /( total outcomes)

Ans 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

**Ans:-**

Expected number of candies for a randomly selected child

=  1 \* 0.015  + 4\*0.20  + 3 \*0.65  + 5\*0.005  + 6 \*0.01  + 2 \* 0.12

= 0.015 + 0.8  + 1.95 + 0.025 + 0.06 + 0.24

=       3.090

=  3.09

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**

**install.packages("readr") # To install paCKAGE TO IMPORT CSV FILES#**

**library(readr)**

**install.packages("readxl") # We used to install to import .xlsw file**

**library(readxl) # Revoke readxl function#**

**getwd()**

**setwd("E:\\Tej\\Assignments")**

**Q7<-read.csv("E:\\Tej\\Assignments\\Q7.csv")**

**View(Q7)**

**mean(Q7$Points)**

**mean(Q7$Score)**

**mean(Q7$Weigh)**

**median(Q7$Points)**

**median(Q7$Score)**

**median(Q7$Weigh)**

**install.packages("NCmisc")**

**library(NCmisc)**

**mode(Q7$Points)**

**mode(Q7$Score)**

**mode(Q7$Weigh)**

**summary(Q7)**

**var(Q7$Points)**

**var(Q7$Score)**

**var(Q7$Weigh)**

**sd(Q7$Points)**

**sd(Q7$Score)**

**sd(Q7$Weigh)**

**range(Q7$Points)**

**range(Q7$Score)**

**range(Q7$Weigh)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.59 | 3.22 | 17.85 |
| Median | 3.69 | 3.33 | 17.71 |
| Variance | 0.29 | 0.96 | 3.19 |
| Standard Deviation | 0.53 | 0.98 | 1.79 |
| Range | 2.76 - 4.93 | 1.513 - 5.424 | 14.5 - 22.9 |

Inference Drawn:

* The mean is useful for spotting trends in the data because we can compare means over a time period to spot trends. The mean is the most common measure of central tendency.
* The **median** divides a sample of data in half; it is the middle score. The median is a useful statistic if we think our data have some extreme cases. The median is not impacted by extreme cases, but the mean is.

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:**

Find the mean. The mean is: Expected Value=

**EV = ∑X/n**

(108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)/9 = 145.333.

It is not mandatory for the expected value to be present in sample space. As calculated above Expected Value of 145.33 is not present in our sample space.

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

**Cars speed and distance**

**Use Q9\_a.csv**

Skewness of Speed = Negative Skewness . Value is -0.11751

Skewness of Distance = Positive Skewness. Value is 0.806895

Kurtosis of Speed = Negative Kurtosis . Value is -0.50899

Kurtosis of Distance = Positive Kurtosis . Value is -0.50899

**SP and Weight(WT)**

**Use Q9\_b.csv**

Skewness of SP = Positive Skewness. Value is 1.61145

Skewness of WT = Negative Skewness. Value is -0.61475

Kurtosis of SP = Positive Kutosis. Value is 2.977329

Kurtosis of WT = Positive Kurtosis. Value is 0.950291

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



1. By the above histogram we observed that data is not normal distribution. Positive Skewness is observed. Right hand side tail is observed.
2. We can see highest frequency of ChickWeight$weight range from 50 to 100
3. As the ChickWeight$weight increases the frequency is less.
4. It is almost zero when of ChickWeight$weight is 400

Lower Extreme

Lower Quartile (Q3)

Median (Q2)

Upper Quartile (Q1)

Upper Extreme

Outliers



The above boxplot suggests that the distribution has lots of outliers towards upper extreme.

It is not normal distribution.

**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**

We will go for T-score calculation as sigma is not known.

P = 3,000,000

s = 2000

sample mean = 200

sample standard deviation = 30

94%

qt(0.97,1999)

Confidence Interval varies from [198.7376 to 201.2624]

98%

Qt(0.99,1999)

Confidence Interval varies from [198.4382 to 201.5618]

96%

Qt(0.98,1999)

Confidence Interval varies from [198.6214 to 201.3786]

**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.

**Ans:-** Mean=41

Mode=41

Median=40.5

Variance=25.53

Standard deviation=5.052

1. What can we say about the student marks?

**Ans:-** Mostly students get scored As Average score as 41 .

Highest score is 56.09

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

**Ans :-** If the mean, median and the mode of a set of numbers are equal, it means, the distribution is symmetric. If there is symmetric distribution  like the bell-shaped normal curve then the mean = median = mode. Means there is no skewness

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

**Ans:-** If the mean is greater than median then, the mean reflects the skewing the most. Data is skewed to the right most means it is positively skewed. (Right skewed)

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

**Ans:-** If the median is greater than the mean, the distribution is negatively skewed. (Left skewed)

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

**Ans :-** Positive values of kurtosis indicate that a distribution is peaked and possess thick tails. Leptokurtic distributions have positive kurtosis values. A leptokurtic distribution has a higher peak and taller (i.e. fatter and heavy) tails than a normal distribution.

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

**Ans:-** A distribution with a negative kurtosis value indicates that the distribution has lighter tails than the normal distribution. For example, data that follow a beta distribution with first and second shape parameters equal to 2 have a negative kurtosis value.

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



What can we say about the distribution of the data?

**Ans:-** There are outliers lying left side of in the distribution of the data.

What is nature of skewness of the data?

**Ans :-** Distribution data is Negatively skewed.

What will be the IQR of the data (approximately)?   
**Ans:-** IQR =Q3-Q1

= 18-10

Approximately IQR= 8

Q19) Comment on the below Boxplot visualizations?



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

1. **Both are Normally Distributed**
2. **First boxplot is looking like a subset of second Box plot**
3. **For first box IQR is 25 and for second Box plot IQR is 75**
4. **Boxplot 1 – 50% of data is distributed between 250 to 275.**
5. **Boxplot 2 – 50% of data is distributed between 225 to 300**

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

Data \_set: Cars.csv

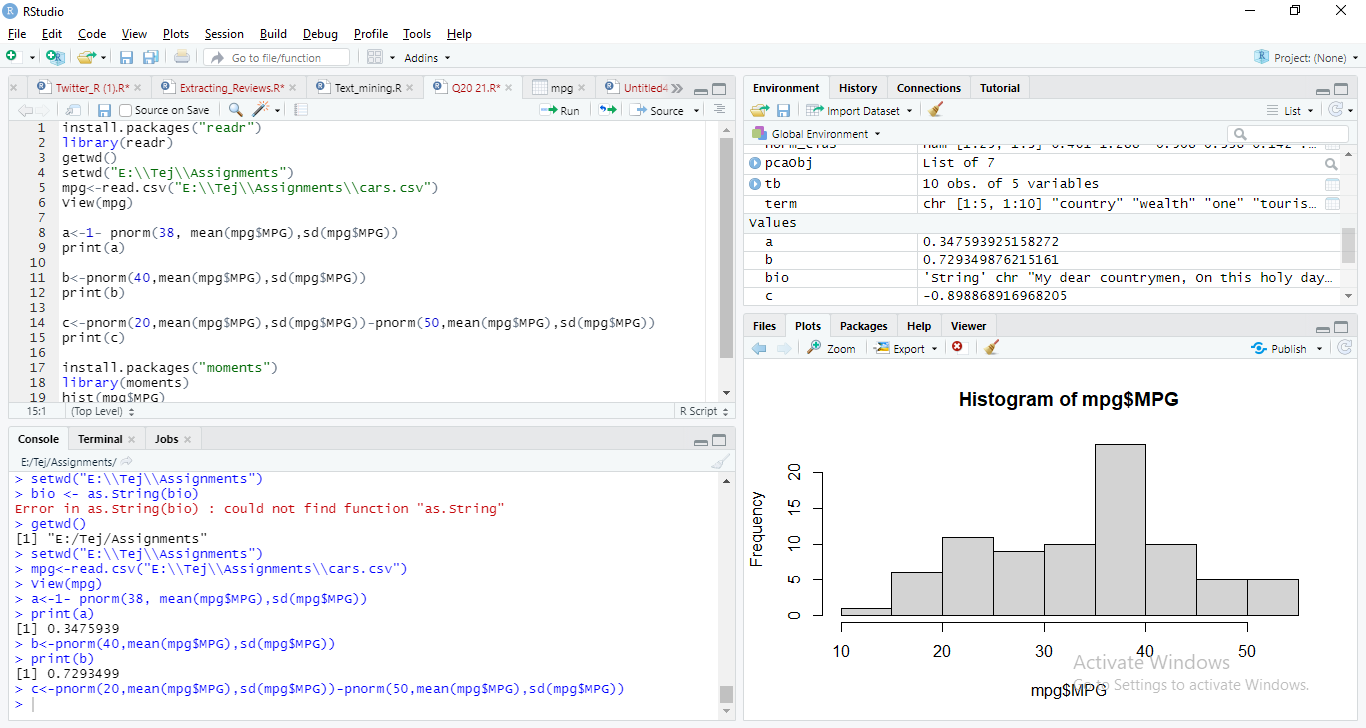
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

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

c. P (20<MPG<50)

**Answer:**



P(MPG>38 = 1 - pnorm(38, mean=mean(Cars$MPG), stddev = sd(Cars$MPG))

Answer = 0.3475939

P(MPG<40 = pnorm(40, mean=mean(Cars$MPG), stddev = sd(Cars$MPG))

Answer = 0.7293499

P(20<MPG<50) = pnorm(50,mean=mean(Cars$MPG),stddev = sd(Cars$MPG)) –

Pnorm(20,mean=mean(Cars$MPG),stddev= sd(Cars$MPG)

Answer = 0.8988689

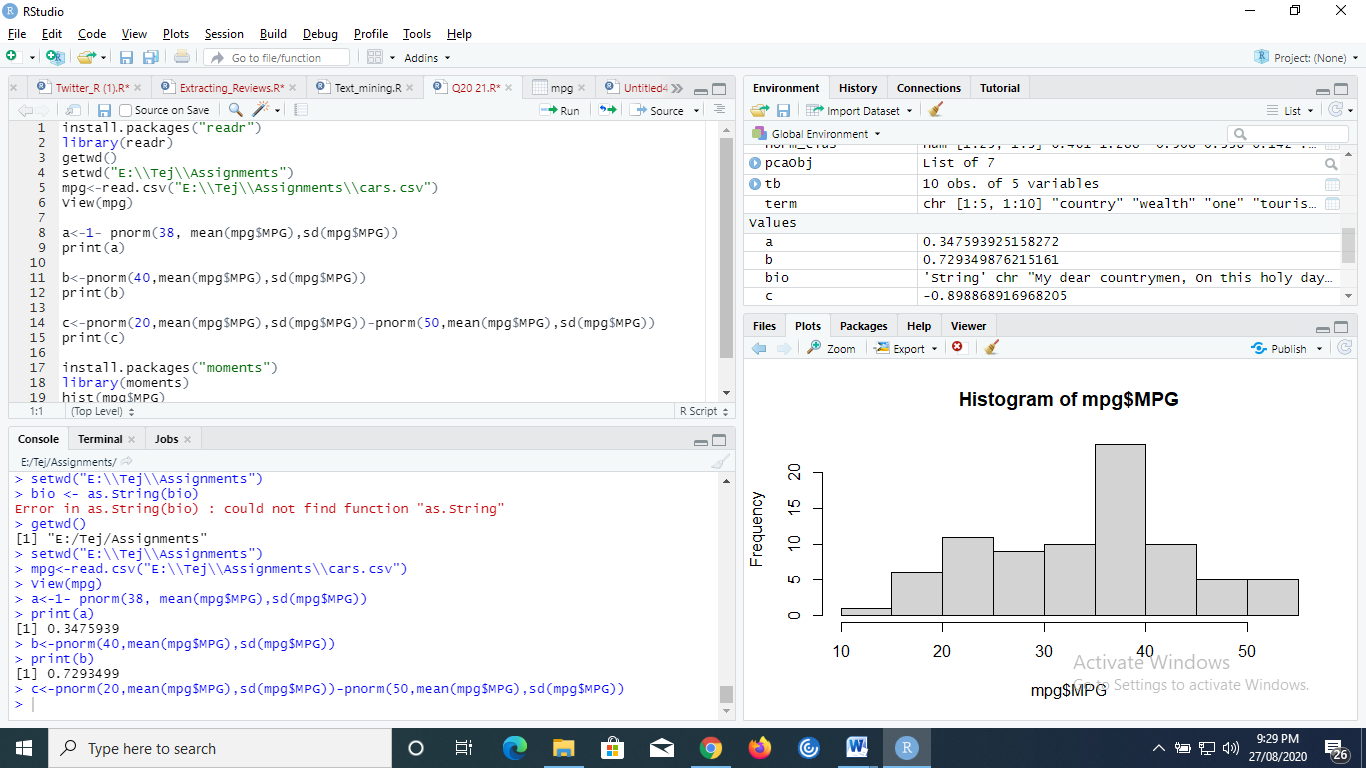
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Answer :-**

MPG of Cars does not follows Normal Distribution as per the Histogram in below print screen by using Rcode.



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

Dataset: wc-at.csv

Ans:-

> WC=read.csv(file.choose())

> hist(WC$AT)

> qqnorm(WC$AT)

> hist(WC$Waist)

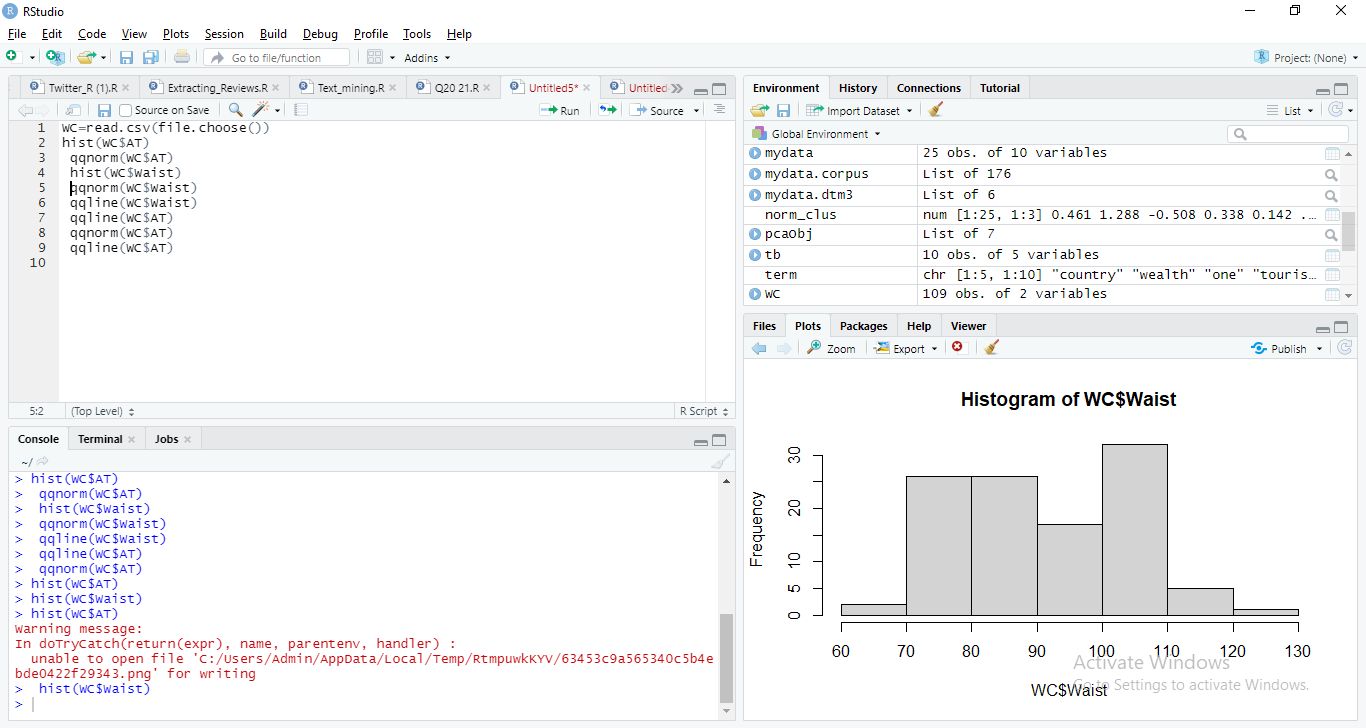
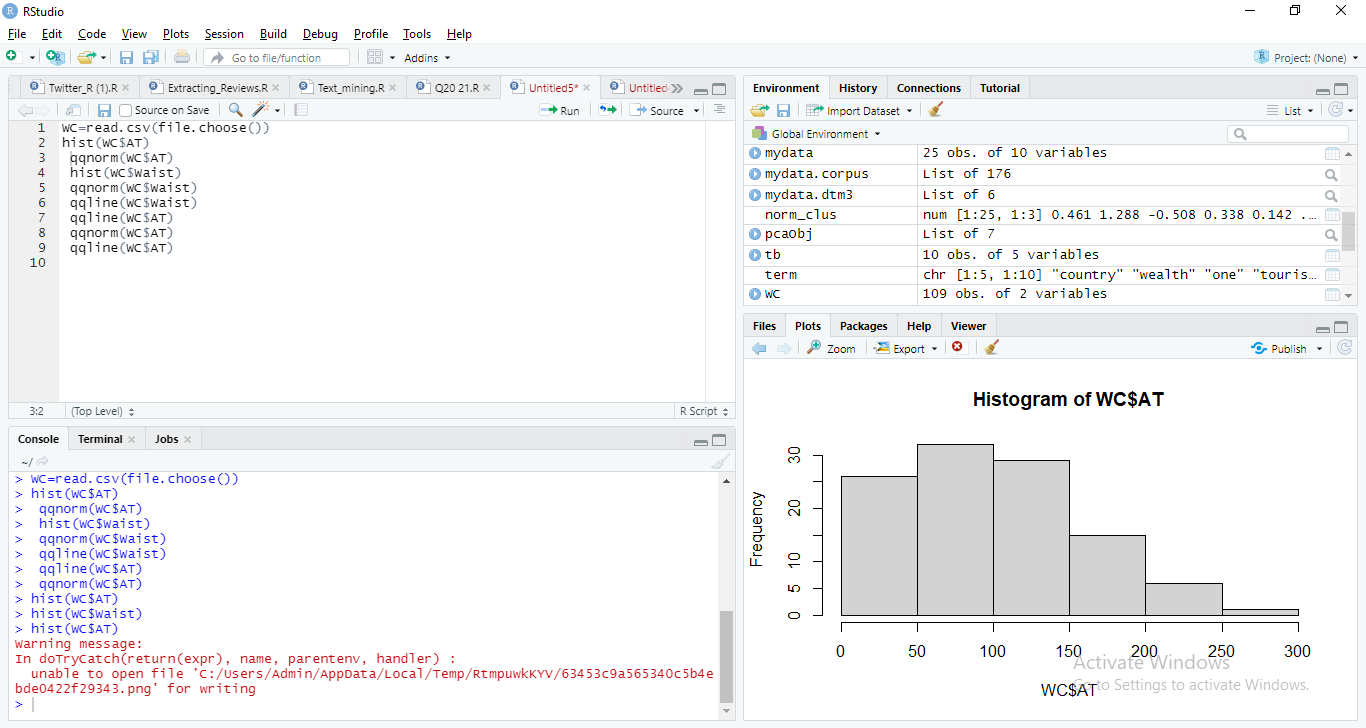
> qqnorm(WC$Waist)

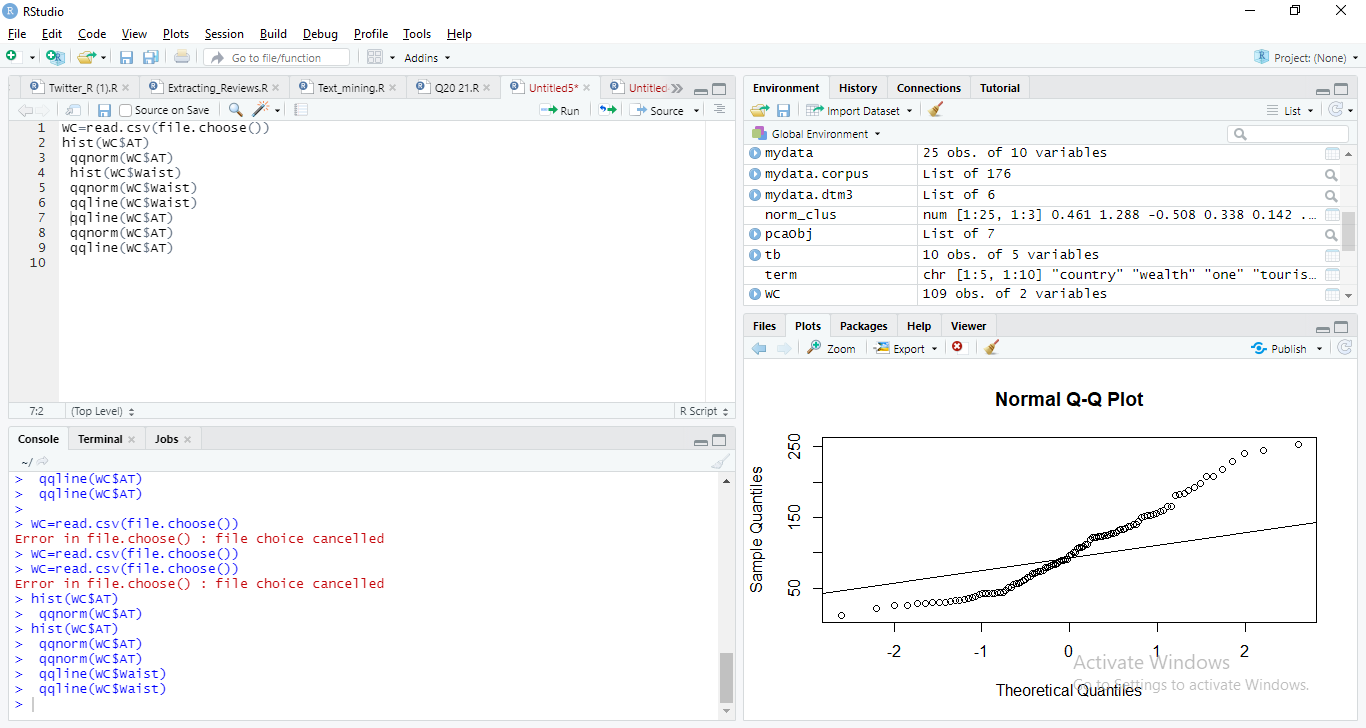
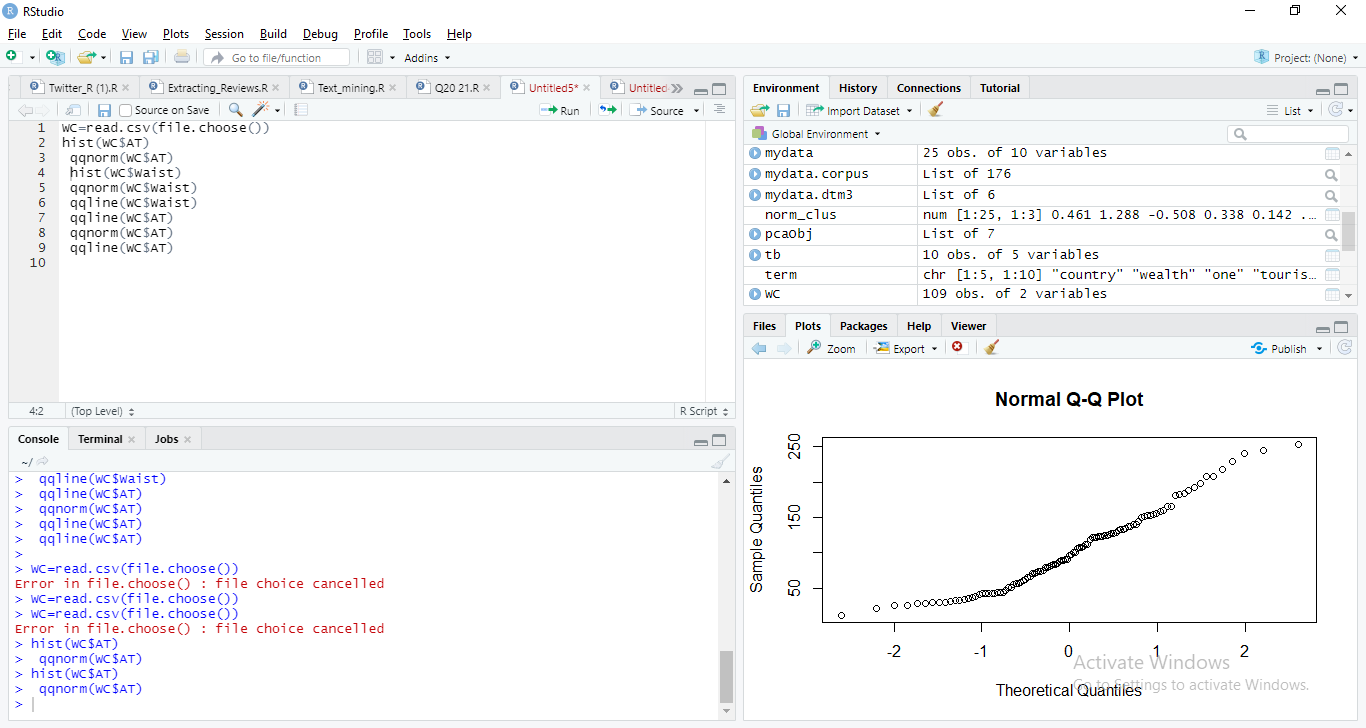
> qqline(WC$Waist)

> qqline(WC$AT)

> qqnorm(WC$AT)

> qqline(WC$AT)





The Adipose Tissue (AT) and Waist Circumference (Waist) does not follow normal distribution.

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

**Ans:-**

#Z-score calculation

> qnorm(0.950) #90%

[1] 1.644854

> qnorm(0.970) #94%

[1] 1.880794

> qnorm(0.80) #60#

[1] 0.8416212

Z Scores of 90% confidence interval is 1.645

Z Scores of 94% confidence interval is 1.88

Z scores of 60% confidence interval is 0.841

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

**Ans:-** > #T-Score calculation

> qt(0.975,24) #95%

[1] 2.063899

> qt(0.980,24) #96%

[1] 2.171545

> qt(0.995,24) #99%

[1] 2.79694

t scores of 95% confidence interval is 2.063

t scores of 96% confidence interval is 2.171

t scores 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

**Answer**

Hypothesis Test will be performed.

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

n = 18

sample\_mean x ̅= 260

sample\_sd = 90

population\_mean µ= 270

T score = 0.5286

Pvalue = 0.6242969

Probability is 62%