SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
Kattankulathur - 603203.

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF MATHEMATICS

Academic Year (2021 – 2022, Even Semester)

18MAB167J -STATISTICAL MODELLING

SEMESTER–II

Name : **SOURABH SHARMA**

Reg. No : **RA2111042010007**

Class : **B. Tech. (CSE)**

Year : **I Year**

Section : **CSBS-A**

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
Kattankulathur – 603203.

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF MATHEMATICS

BONAFIDE CERTIFICATE

Reg. No:

R	A	2	1	1	1	0	4	2	0	1	0	0	0	7

Certified that this is the bonafide record of work done by SOURABH SHARMA of II semester B.Tech. COMPUTER SCIENCE AND ENGINEERING during the academic year 2021-2022 in the 18MAB167J - STATISTICAL MODELLING Laboratory.

Dr.R.Varadharajan
Staff In charge

HOD/Mathematics

Submitted for the practical examination held on_____ at SRM Institute of Science and Technology, Kattankulathur, Chennai-603203.

Examiner-1

Examiner-2

INDEX

OF THE RECORD

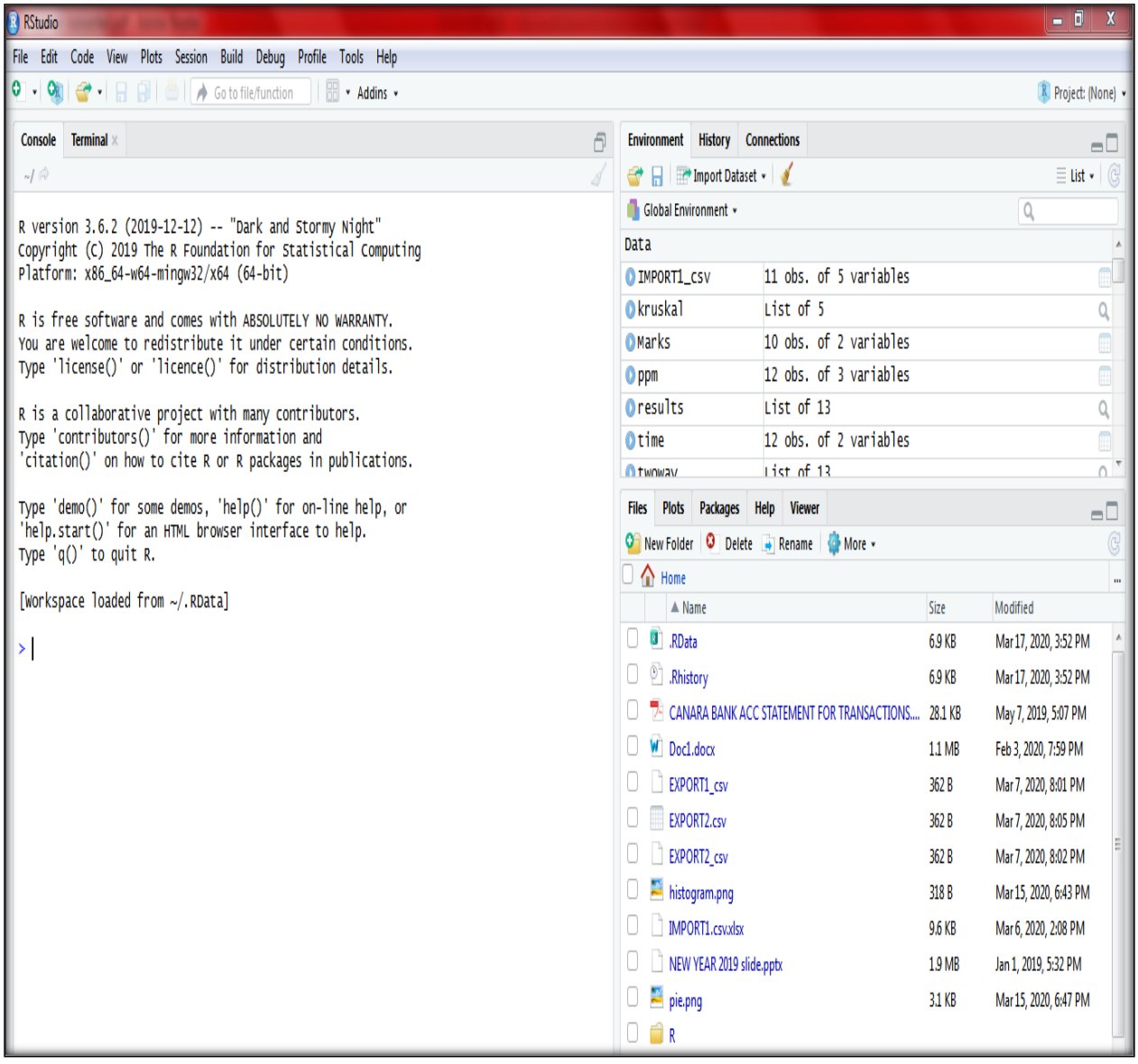
SR.NO.	EXPT.NO.	DATE	EXPT NAME	FACULTY SIGN
1	1	08.03.2022	R CONSOLE	
2	2	08.03.2022	R EDITOR	
3	3	08.03.2022	VECTOR AND ITS OPERATIONS	
4	4	15.03.2022	MATRIX AND ITS MANIPULATIONS	
5	5	15.03.2022	DATA FRAME AND ITS MANIPULATIONS	
6	6	22.03.2022	DATA FRAME AND ITS MANIPULATIONS	
7	7	22.03.2022	DATA HANDLING IN R	
8	8	05.04.2022	DATA HANDLING IN R	
9	9	05.04.2022	IMPORTING DATA	
10	10(a)	12.04.2022	BAR DIAGRAM	
11	10(b)	12.04.2022	SUBDIVIDED BAR DIAGRAM	
12	11	21.04.2022	HISTOGRAM	
13	12	21.04.2022	PIE CHART	
14	13	06.05.2022	BOX PLOT	
15	14	06.05.2022	DESCRIPTIVE STATISTICS	
16	15	13.05.2022	DESCRIPTIVE STATISTICS (GROUPED DATA)	
17	16	13.05.2022	CORRELATION AND REGRESSION	
18	17	20.05.2022	MANN WHITNEY U TEST	
19	18	20.05.2022	ONE WAY ANOVA	
20	19	27.05.2022	ONE WAY ANOVA	
21	20	27.05.2022	TWO WAY ANOVA	
22	21	03.06.2022	SINGLE LINEAR REGRESSION	
23	22	03.06.2022	MULTIPLE LINEAR REGRESSION	
24	23	17.06.2022	T TEST	

R CONSOLE

EXPERIMENT NO. :- 1

DATE:- 08.03.2022

R CONSOLE

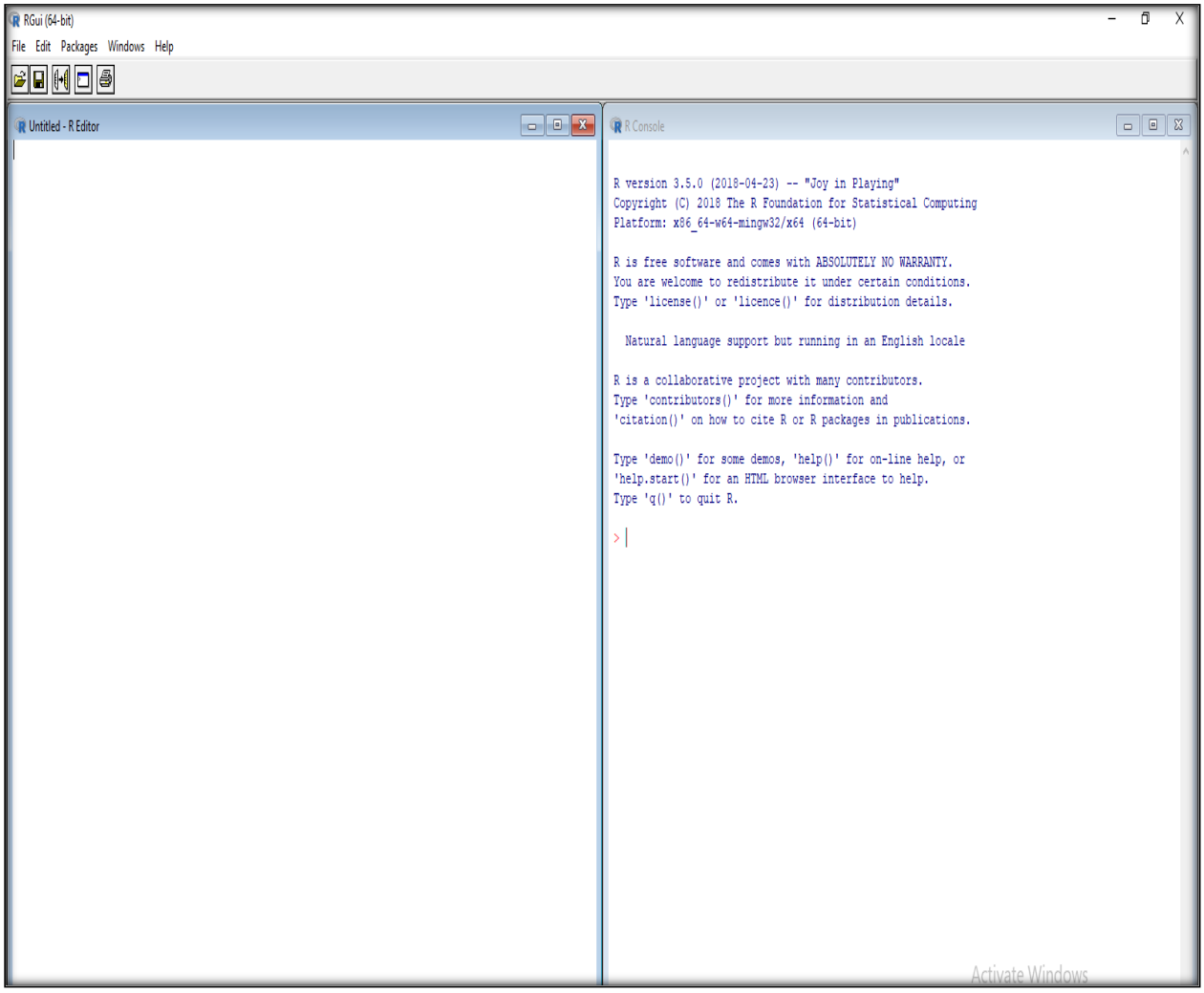


R EDITOR

EXPERIMENT NO. :-2

DATE:-08.03.2022

R EDITOR



VECTOR AND IT'S OPERATIONS

EXPERIMENT NO. :-3

DATE:-08.03.2022

PROBLEM:-

Creating a vector of profit percentage, i.e., gain of 20 companies using scan() function and assign it in a variable profit percentage.

1)List all the profit percentages below 30.

2)List all the profit percentages above 30 .

AIM:-

To create a vector of profit percentages between 10 and 50 of 20 companies and to list all the profit percentages specified above by using R Studio.

PROCEDURE:-

STEP1: Open R console window.

STEP2: Enter the data using scan() function and assign it to the variable PROFIT PERCENTAGE.

STEP3: Now the data set has been created between the intervals 10 and 50.

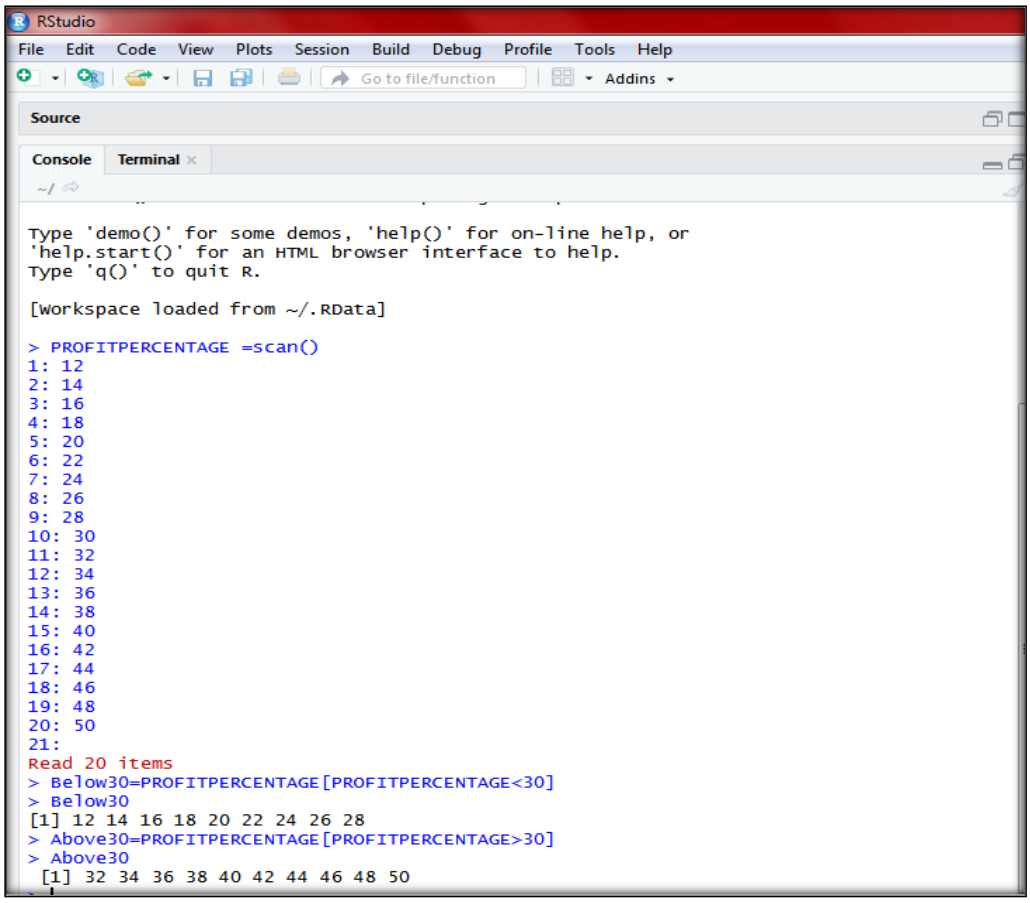
STEP4: To list the profit percentages below 30, use the operand "<" and to list above 30, use the operand ">".

STEP5: The results will be displayed in the R console window.

R INPUT:-

```
PROFITPERCENTAGE = scan()

Below30=PROFITPERCENTAGE[PROFITPERCENTAGE<30]

Below30

Above30=PROFITPERCENTAGE[PROFITPERCENTAGE>30]

Above30
```

R OUTPUT:-



RESULTS:-

The Profit percentages below 30 are 12,14,16,18,20,22,24,26, 28.

The Profit percentages above 30 are 32,34,36,38,40,42,44,46, 48,50.

INFERENCE:-

From the results, we can conclude that there are 9 companies whose profit percentage is below 30 and there are 10 companies whose profit percentage is above 30.

MATRIX AND ITS OPERATIONS

EXPERIMENT NO.:- 4

DATE:-15.03.2022

PROBLEM:-

Construct two 2x2 matrices with your own data and verify the following

tr (B^-1AB) = tr(A),

where tr () represents the trace of the matrix.

AIM:-

To construct two 2x2 matrices and to verify the following

tr (B^-1AB) = tr(A) using R Studio.

PROCEDURE:-

STEP1: Open R console window.

STEP2: Enter elements of matrix using matrix () and c () commands.

STEP3: The trace of B^-1 is found by sum (diag (BINVAB)).

STEP4: The trace of A is found by sum (diag (A)).

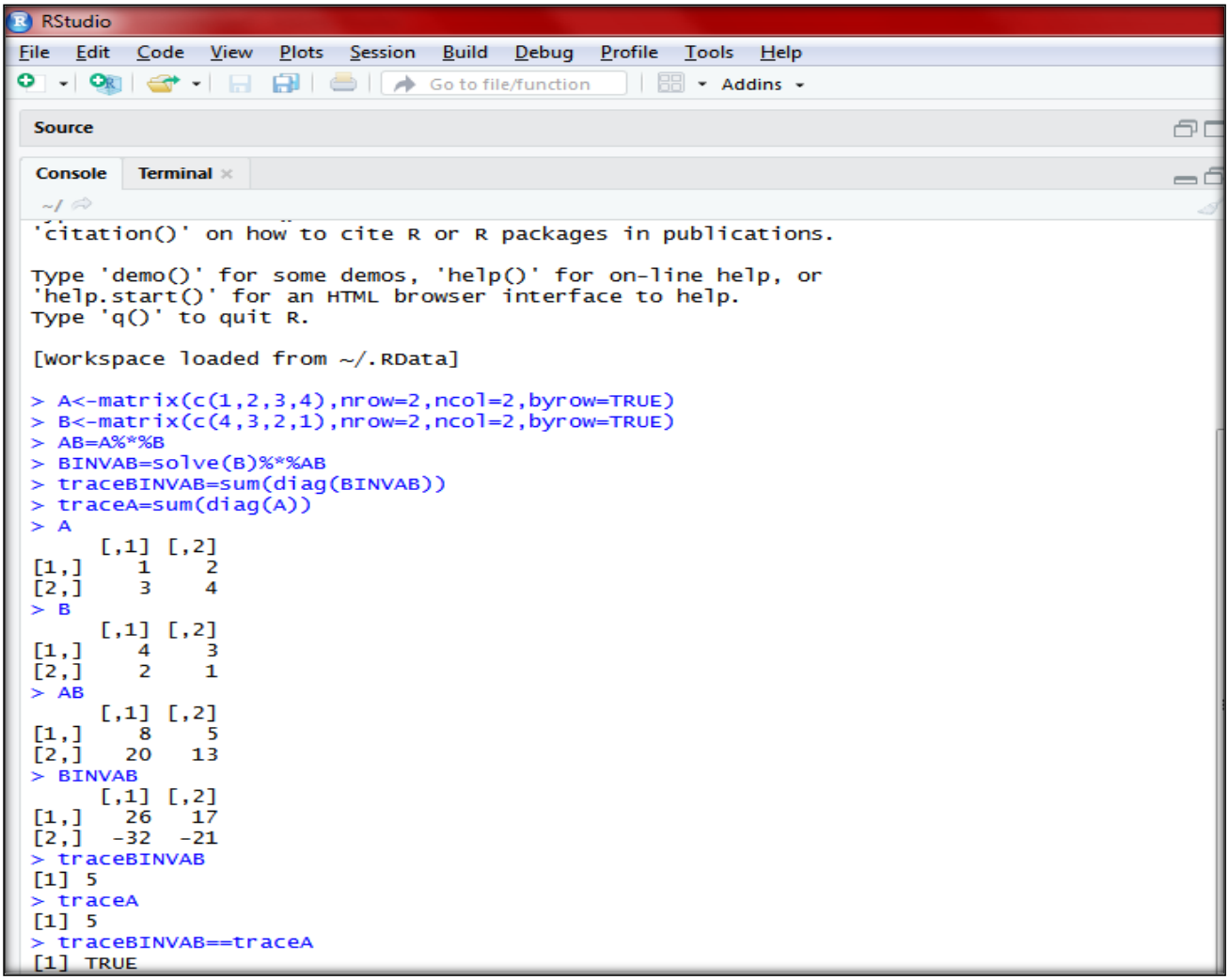
STEP5: Verify both results using "==" command and the result will be displayed in R console window.

R INPUT:-

```
A<-matrix (c (1, 2, 3, 4), nrow=2, ncol=2, byrow=TRUE)

B<-matrix (c (4, 3, 2, 1), nrow=2, ncol=2, byrow=TRUE)

AB=A%*%B

BINVAB = solve(B)%*%AB

traceBINVAB = sum (diag (BINVAB))

traceA=sum (diag (A))

traceBINVAB==traceA

A

B

BINVAB

traceBINVAB
```

```
traceA

traceBINVAB==traceA
```

R OUTPUT:-



RESULTS:-

The trace of matrices A and BINVAB is 5.

INFERENCE:-

From the result, we can conclude that the sum of diagonal elements of the matrices BINVAB and A are equal.

DATA FRAME AND IT'S MANIPULATIONS

EXPERIMENT NO.:-5

DATE:-15.03.2022

PROBLEM:-

Construct a data frame with height and weight of 20 students. Find the BMI value and also to create subsets by

1) BMI<20

2)20<=BMI<=25

3) BMI>25 using R studio.

AIM:-

To construct a data frame with height and weight of 20 students. Find BMI value and also to create subsets by

1) BMI<20

2)20<=BMI<=25

3) BMI>25 using R studio.

PROCEDURE:-

STEP1: Open R console window.

STEP2: The data is entered using data.frame () command.

STEP3: The formula to calculate BMI is Weight/(Height)^2. Add another column using cbind command.

STEP4: The subset is created using the students whose BMI<20,20<=BMI<=25,BMI>25 using subsets() command.

STEP5:The results will be displayed in the R console window.

R INPUT:-

```
STUDENTS<-edit(data.frame())

attach(STUDENTS)

BMI=WEIGHT/(HEIGHT)^2

BMIres=cbind(STUDENTS,BMI)

 RES1=subset(STUDENTS,BMI<20)

 RES2=subset(STUDENTS,BMI>=20&BMI<=25)
```

```
RES3=subset(STUDENTS,BMI>25)
```

BMI

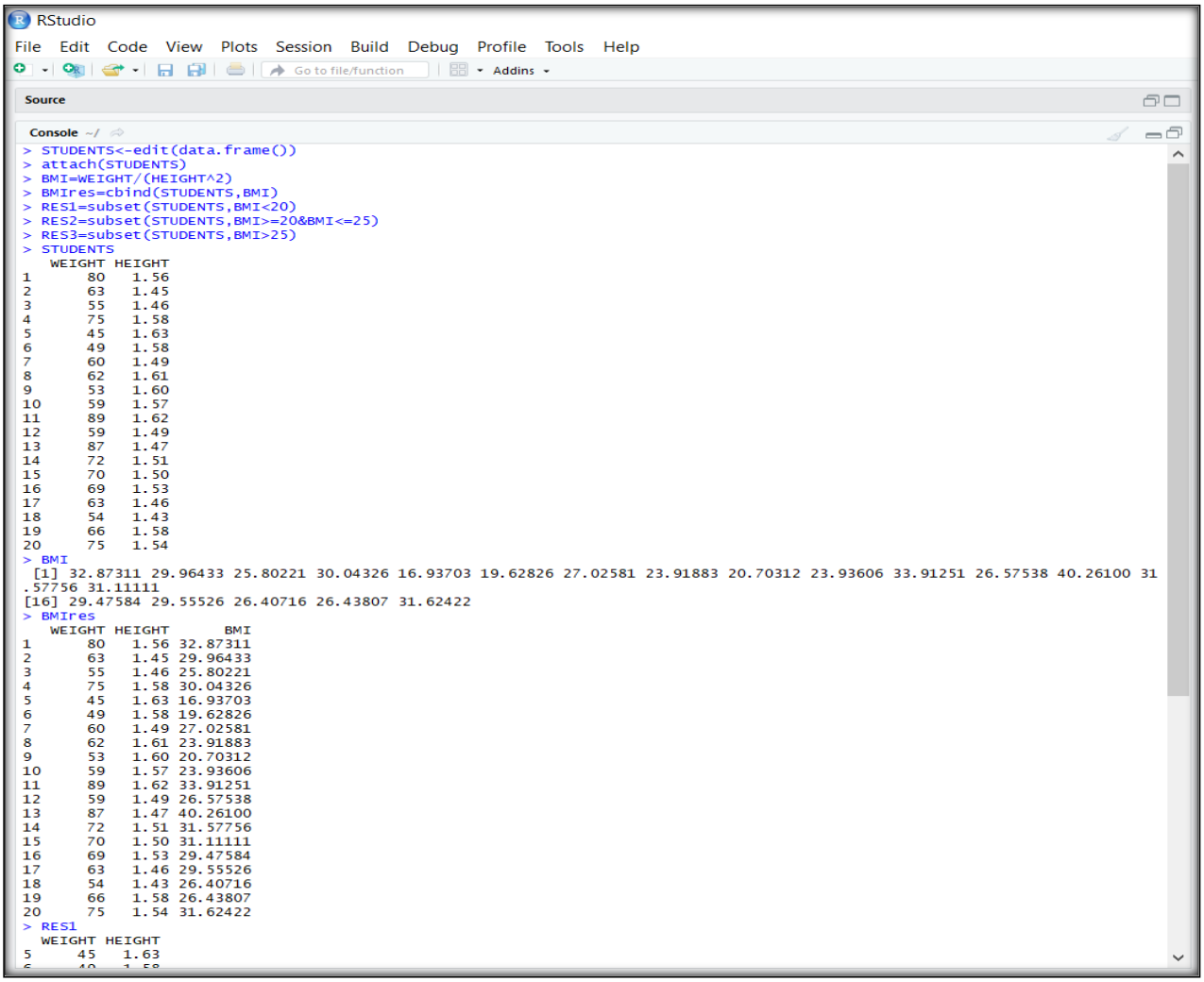
BMIres

RES1

RES2

RES3

R OUTPUT:-



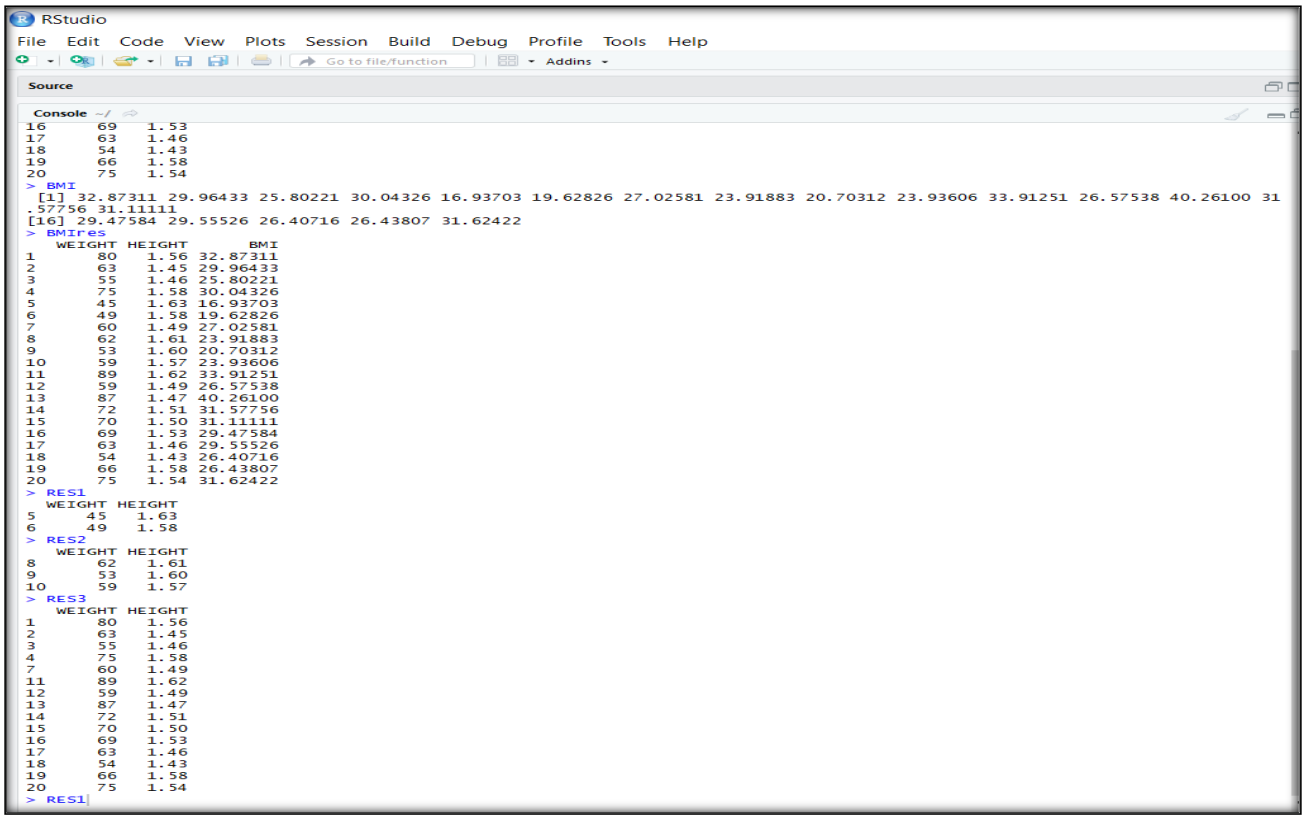
RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source

Console ~/

```
> STUDENTS<-edit(data.frame())
> attach(STUDENTS)
> BMI=WEIGHT/(HEIGHT^2)
> BMIres=cbind(STUDENTS,BMI)
> RES1=subset(STUDENTS,BMI<20)
> RES2=subset(STUDENTS,BMI>=20&BMI<=25)
> RES3=subset(STUDENTS,BMI>25)
> STUDENTS
   WEIGHT HEIGHT
1      80   1.56
2      63   1.45
3      55   1.46
4      75   1.58
5      45   1.63
6      49   1.58
7      60   1.49
8      62   1.61
9      53   1.60
10     59   1.57
11     89   1.62
12     59   1.49
13     87   1.47
14     72   1.51
15     70   1.50
16     69   1.53
17     63   1.46
18     54   1.43
19     66   1.58
20     75   1.54
> BMI
 [1] 32.87311 29.96433 25.80221 30.04326 16.93703 19.62826 27.02581 23.91883 20.70312 23.93606 33.91251 26.57538 40.26100 31
.57756 31.11111
[16] 29.47584 29.55526 26.40716 26.43807 31.62422
> BMIres
   WEIGHT HEIGHT      BMI
1      80   1.56 32.87311
2      63   1.45 29.96433
3      55   1.46 25.80221
4      75   1.58 30.04326
5      45   1.63 16.93703
6      49   1.58 19.62826
7      60   1.49 27.02581
8      62   1.61 23.91883
9      53   1.60 20.70312
10     59   1.57 23.93606
11     89   1.62 33.91251
12     59   1.49 26.57538
13     87   1.47 40.26100
14     72   1.51 31.57756
15     70   1.50 31.11111
16     69   1.53 29.47584
17     63   1.46 29.55526
18     54   1.43 26.40716
19     66   1.58 26.43807
20     75   1.54 31.62422
> RES1
  WEIGHT HEIGHT
5     45   1.63
```



RESULTS:-

The data frame with height and weight of 20 students have been created successfully. Also the BMI values have been found and also the subsets have been created successfully.

INFERENCE:-

From the result, we observe that 2 students whose BMI<20,3 students whose BMI is between 20 and 25 and 15 students whose BMI>25.

DATA FRAME AND IT'S MANIPULATIONS

EXPERIMENT NO.:-6

DATE:-22.03.2022

PROBLEM:-

Create a data frame with the first column should be numeric and the second column should be character and the third column should be either numeric or logical by using vectors of equal length and find the summary of the data.

AIM:-

To create a data frame of equal length using vectors and to find the summary of the data.

PROCEDURE:-

STEP1: Open R console window.

STEP2: Enter data using vector functions.

STEP3: Print data.

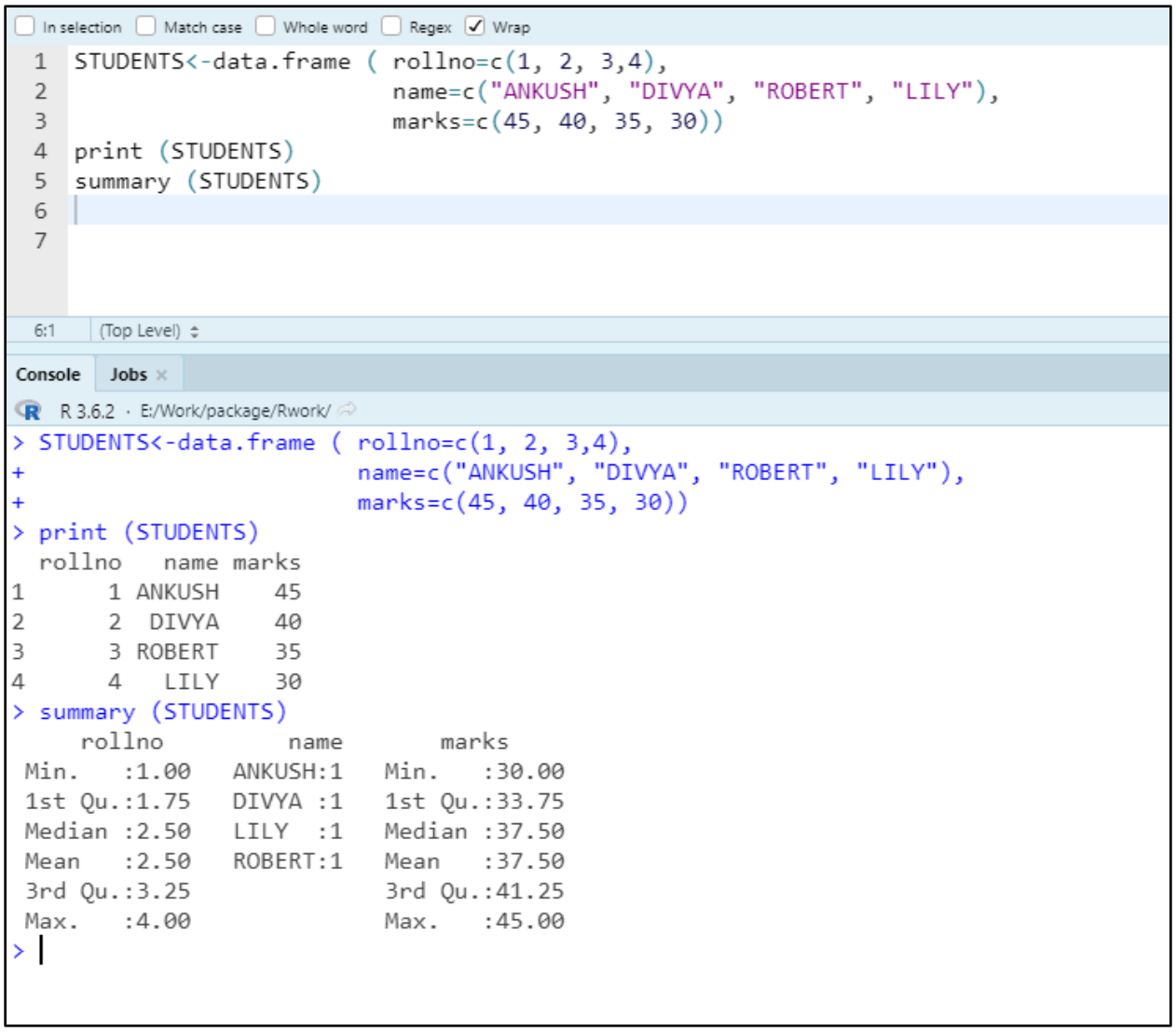
STEP4: By using the function summary (), we can find the summary of the data.

STEP5: The results will be displayed in the R console window.

R INPUT:-

```
STUDENTS<-data.frame(rollno=c(1, 2, 3,4),
        name=c("ANKUSH", "DIVYA", "ROBERT", "LILY"),
        marks=c(45, 40, 35, 30))
print (STUDENTS)
summary (STUDENTS)
```

R OUTPUT:-



```
In selection  Match case  Whole word  Regex  Wrap

1  STUDENTS<-data.frame ( rollno=c(1, 2, 3,4),
2                         name=c("ANKUSH", "DIVYA", "ROBERT", "LILY"),
3                         marks=c(45, 40, 35, 30))
4  print (STUDENTS)
5  summary (STUDENTS)
6
7

6:1  (Top Level)

Console  Jobs

R 3.6.2 · E:/Work/package/Rwork/

> STUDENTS<-data.frame ( rollno=c(1, 2, 3,4),
+                        name=c("ANKUSH", "DIVYA", "ROBERT", "LILY"),
+                        marks=c(45, 40, 35, 30))
> print (STUDENTS)
  rollno   name marks
1      1 ANKUSH    45
2      2  DIVYA    40
3      3 ROBERT    35
4      4   LILY    30
> summary (STUDENTS)
     rollno         name       marks
 Min.   :1.00   ANKUSH:1   Min.   :30.00
 1st Qu.:1.75   DIVYA :1   1st Qu.:33.75
 Median :2.50   LILY  :1   Median :37.50
 Mean   :2.50   ROBERT:1   Mean   :37.50
 3rd Qu.:3.25              3rd Qu.:41.25
 Max.   :4.00              Max.   :45.00
>
```

RESULTS:-

The data frame using vectors of equal length has been constructed successfully. Also the summary of the data has been created successfully.

INFERENCE:-

From the results, we can conclude that the average of marks scored by those 4 students is 37.5.

DATA HANDLING IN R

EXPERIMENT NO.:- 7

DATE:-22.03.2022

PROBLEM:-

Handle the data using R Studio using various functions.

I. **Knowing working directory.**
II. **Setting working directory.**
III. **Listing variables.**

AIM:-

To handle data in R Studio using various functions.

PROCEDURE:-

STEP1: Open R console window.

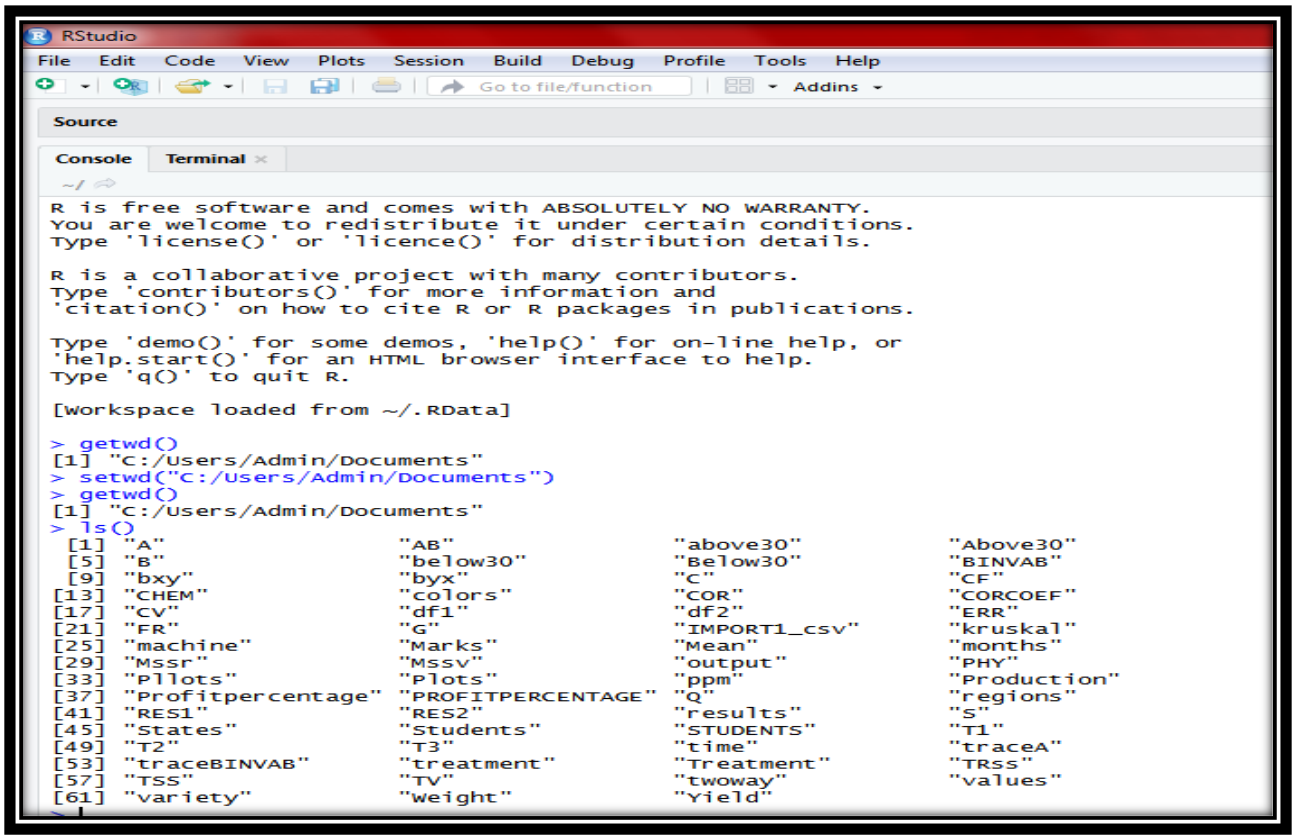
STEP2: Enter the code corresponding to each function.

STEP3: The result will be displayed on the R console window.

R INPUT:-

```
>getwd ()

>setwd("C:/Admin/MY BOOK/Documents")

>getwd ()

>ls()
```

R OUTPUT:-



RESULTS:-

We constructed data handling techniques in R using various R commands.

INFERENCE:-

From the result, the working directory is “C:/Users/ Admin/ Documents” and the listing variables is found to be the previously saved workspace restored.

DATA HANDLING IN R

EXPERIMENT NO.:- 8

DATE:-05.04.2022

PROBLEM:-

Handle the data using various R commands in R Studio.

I. **Removing variables.**
II. **Sequence with single interval.**
III. **Sequence with different interval.**

AIM:-

To handle the data in R Studio using various R commands.

PROCEDURE:-

STEP1: Open R console window.

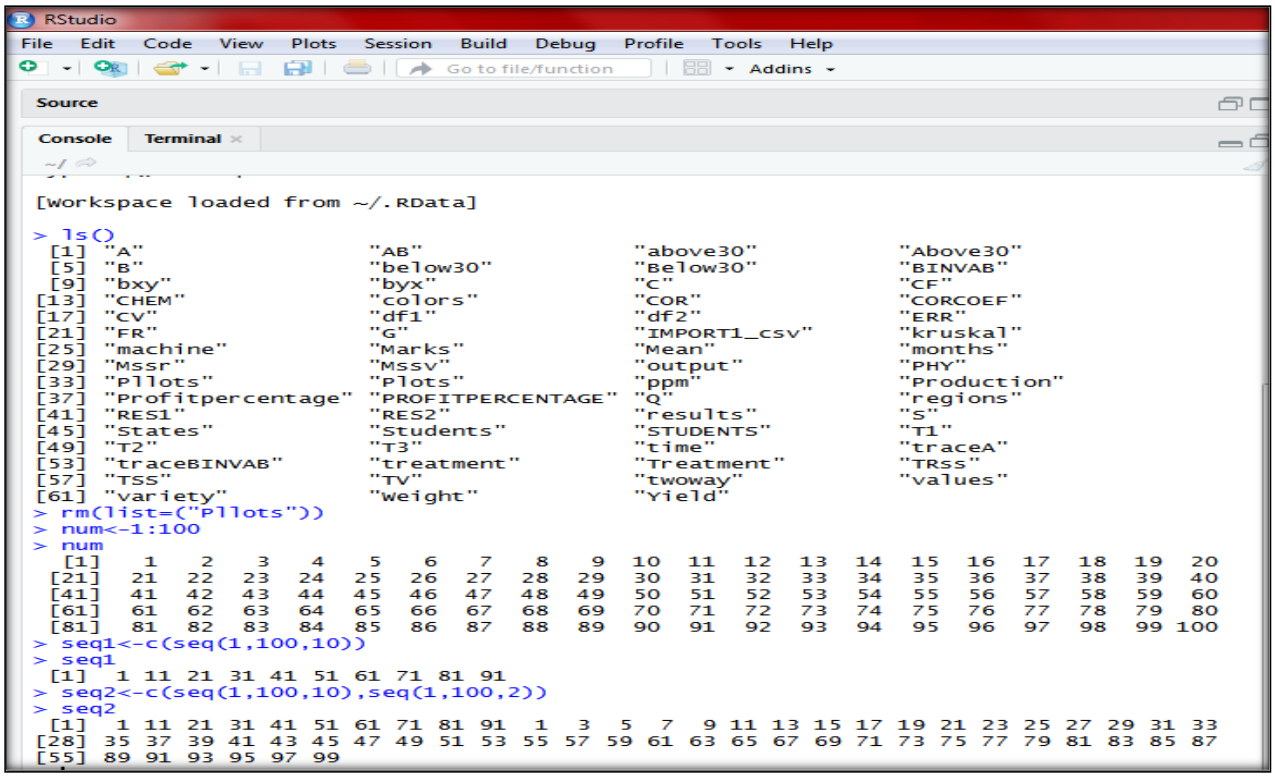
STEP2: Enter the code corresponding to each function.

STEP3: The results will be displayed in the R console window.

R INPUT:-

```
ls ()

rm(list="")

num<-1:100

num

seq1<-c (seq(1,100,10))

seq1

seq2<-c (seq(1,100,10),seq(1,100,2))

seq2
```

R OUTPUT:-



RESULTS:-

We have constructed data handling using various R commands.

INFERENCE:-

From the results, we can conclude that we removed “Plots” from the listing variables and the sequences with single and different intervals have been created successfully.

IMPORTING DATA FROM EXCEL

EXPERIMENT NO.:- 9

DATE:-05.04.2022

PROBLEM:-

Create an excel file related to marks of 10. Import the data from the excel into R and find the summary of the data.

AIM:-

To import data from excel and to find the summary of that data.

PROCEDURE:-

STEP1: Create Excel sheet.

STEP2: Save the file into documents.

STEP3: Open R console window.

STEP4: Enter the R command read_excel () function from readxl library to import the data.

STEP4: Print data.

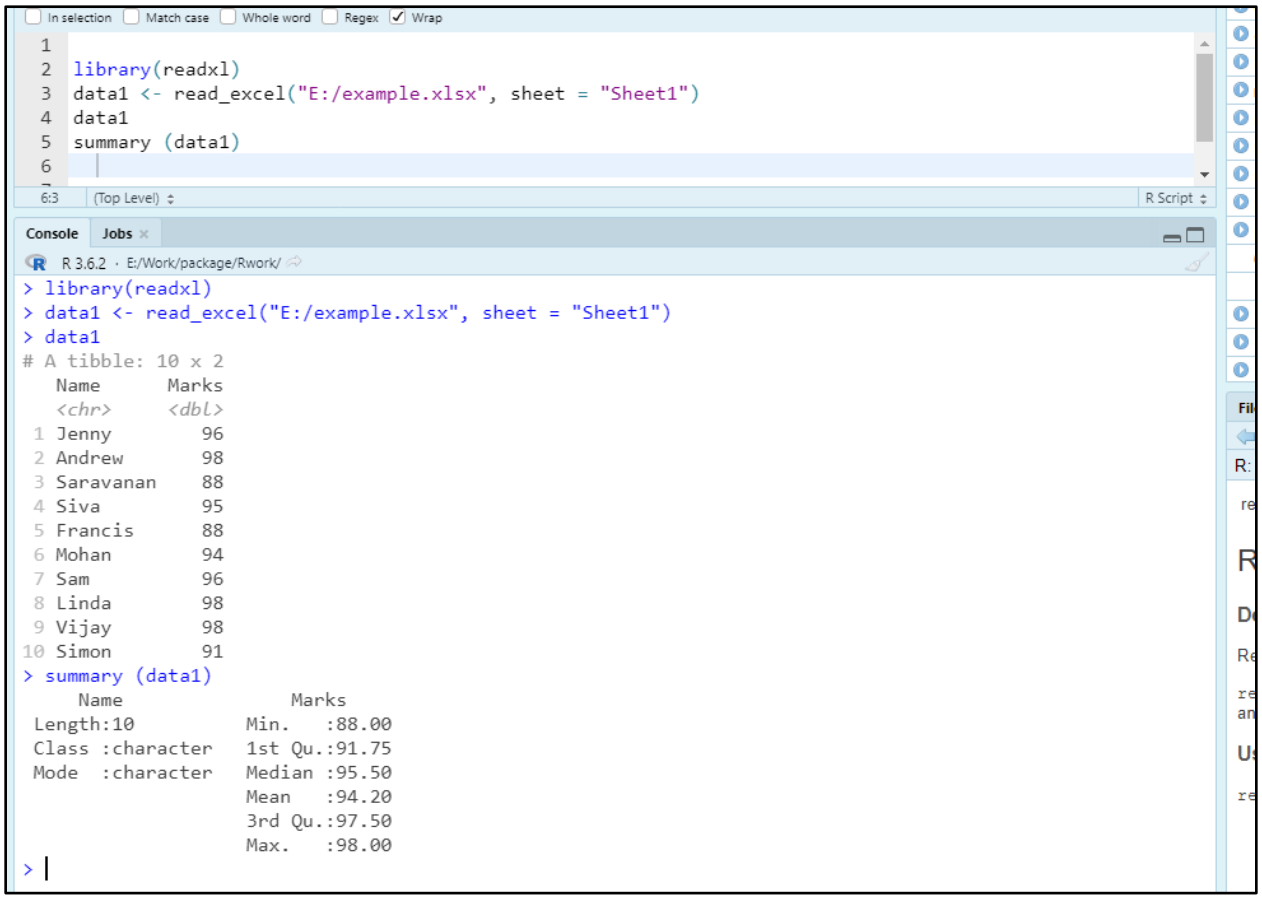
STEP6: Use the command summary () to find the summary of the data.

STEP7: The result will be displayed on the R console window.

R INPUT:-

```
library(readxl)

data1 <- read_excel("E:/example.xlsx", sheet = "Sheet1")

data1

summary (data1)
```

R OUTPUT:-



```
In selection  Match case  Whole word  Regex  Wrap

1
2  library(readxl)
3  data1 <- read_excel("E:/example.xlsx", sheet = "Sheet1")
4  data1
5  summary (data1)
6

6:3  (Top Level)                                   R Script

Console  Jobs
R 3.6.2 · E:/Work/package/Rwork/

> library(readxl)
> data1 <- read_excel("E:/example.xlsx", sheet = "Sheet1")
> data1
# A tibble: 10 x 2
   Name      Marks
   <chr>     <dbl>
 1 Jenny        96
 2 Andrew       98
 3 Saravanan    88
 4 Siva         95
 5 Francis      88
 6 Mohan        94
 7 Sam          96
 8 Linda        98
 9 Vijay        98
10 Simon        91
> summary (data1)
     Name               Marks
 Length:10          Min.   :88.00
 Class :character   1st Qu.:91.75
 Mode  :character   Median :95.50
                    Mean   :94.20
                    3rd Qu.:97.50
                    Max.   :98.00
>
```

RESULT:-

The data has been imported successfully from the excel sheet and also the summary of the data has been created.

INFERENCE:-

From the results, we can conclude that the mark scored is 88 and the highest is 98

BAR DIAGRAM

EXPERIMENT NO.:- 10(a)

DATE:-12.04.2022

PROBLEM:-

Construct a bar diagram for the following data related to the rice production in different states of India.

STATES	TN	KL	KA	AS	BR
QUANTITY	20	10	15	5	10

AIM:-

To construct a bar diagram related to the rice production in different states of the country India.

PROCEDURE:-

STEP1: Open R console window.

STEP2: Enter the data related to the bar graph.

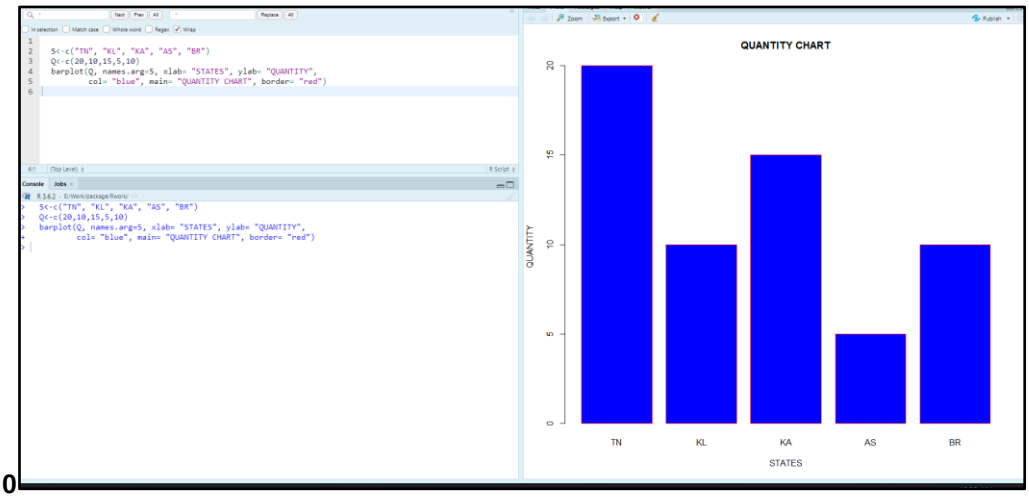
STEP3: Plot the values using barplot () command.

STEP4: The results will be displayed in the R console window.

R INPUT:-

```
S<-c("TN", "KL", "KA", "AS", "BR")

Q<-c(20,10,15,5,10)

barplot(Q, names.arg=S, xlab= "STATES", ylab=   QUANTITY", col= "blue", main=
"QUANTITY CHART", border= "red")
```

R OUTPUT:-



RESULTS:-

For the given data, bar diagram has been constructed successfully.

INFERENCE:-

From the results, we conclude that rice production is high in Tamil Nadu (TN) and less in Assam (AS).

SUBDIVIDED BAR DIAGRAM

EXPERIMENT NO. :- 10(b)

DATE:- 12.04.2022

PROBLEM:-

Construct a sub divided bar diagram for the following data.

	MARCH	APRIL	MAY	JUNE
DEL	29	36	39	39
MAA	33	36	38	37
BOM	31	32	33	32
CCU	34	35	35	34

AIM:-

To construct a sub divided bar diagram related to the temperature of 4 different cities in different months.

PROCEDURE:-

STEP1: Open R console window.

STEP2: Enter the data using matrix () and c () commands.

STEP3: Use barplot() function for plotting the data.

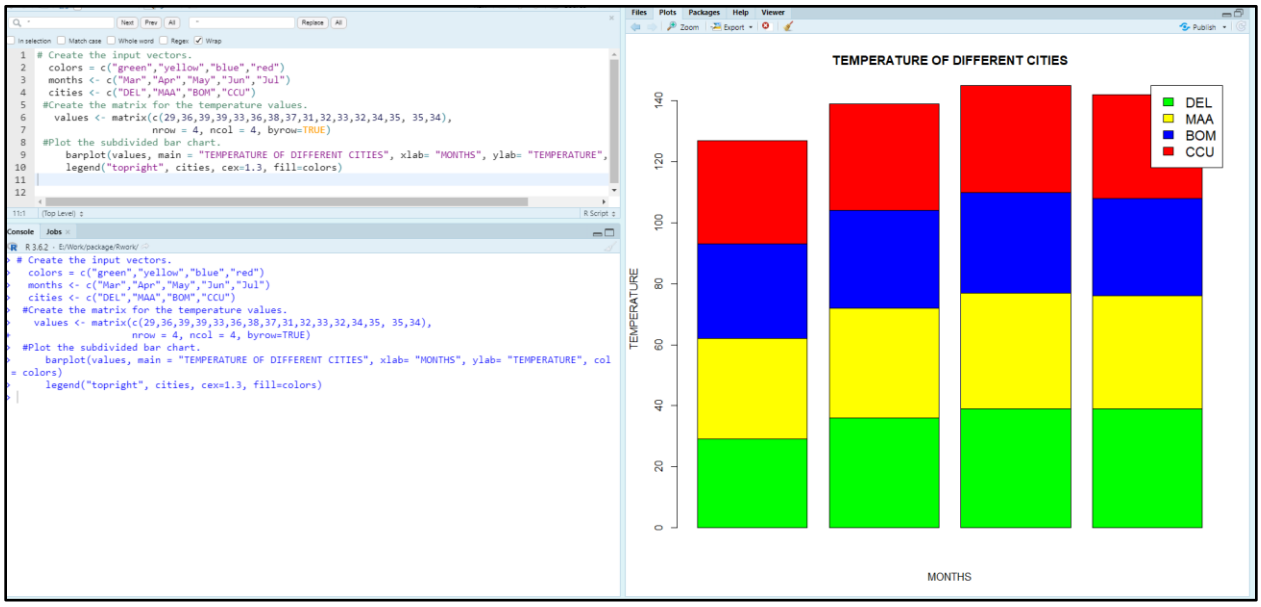
STEP4: The results will be displayed on the R console window.

R INPUT:-

```
# Create the input vectors.

  colors = c("green","yellow","blue","red")

  months <- c("Mar","Apr","May","Jun","Jul")

  cities <- c("DEL","MAA","BOM","CCU")

#Create the matrix for the temperature values.

values <- matrix (c(29, 36, 39, 39, 33, 36, 38, 37, 31, 32, 33, 32, 34, 35, 35, 34),
nrow = 4, ncol = 4, byrow=TRUE)

#Plot the subdivided bar chart.

barplot(values, main = "TEMPERATURE OF DIFFERENT CITIES", xlab= "MONTHS", ylab=
"TEMPERATURE", col = colors)

legend("topright", cities, cex=1.3, fill=colors)
```

```
dev.off()
```

R OUTPUT:-



RESULTS:-

The sub divided bar diagram has been constructed successfully.

INFERENCE:-

From the results, we can conclude that the temperature is high in the month of May and is less in the month of March.

HISTOGRAM

EXPERIMENT NO.:- 11

DATE:- 21.04.2022

PROBLEM:-

The height of 20 students are given below.Draw a histogram for the given data.

160,164,170,175,180,183,162,173,158,172,176,166,181,163,155,160,173,177,169,180.

AIM:-

Construct a histogram for the given height of 20 students.

PROCEDURE:-

STEP1: Open R console window.

STEP2: Enter the data using c () function.

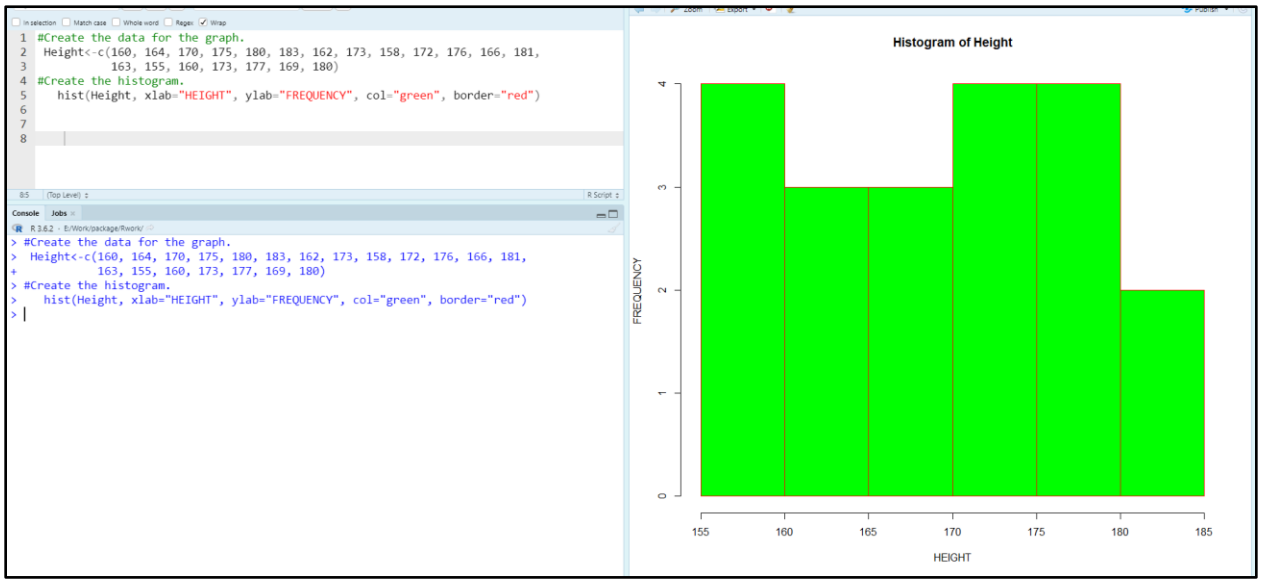
STEP3:Construct histogram using hist() commands.

STEP4:The outputs will be displayed in R console window.

R INPUT:-

```
#Create the data for the graph.

Height<- c(160, 164, 170, 175, 180, 183, 162, 173, 158, 172, 176, 166, 181, 163, 155, 160, 173, 177, 169, 180)

#Create the histogram.

 hist(Height, xlab="HEIGHT", ylab="FREQUENCY", col="green", border="red")

dev.off()
```

R OUTPUT:-



RESULTS:-

The histogram has been constructed successfully.

INFERENCE:-

From the results, we can conclude in the intervals 155-160,170-175 and 175-180, there are same number of students.

PIE DIAGRAM

EXPERIMENT NO.:- 12

DATE:- 21.04.2022

PROBLEM:-

Construct a pie chart for the following data

GRADES	NO. OF STUDENTS
O	5
A+	10
A	15
B	20

AIM:-

To construct the pie diagram for the given data.

PROCEDURE:-

STEP1:Open R console window.

STEP2: Enter the data using c() command.

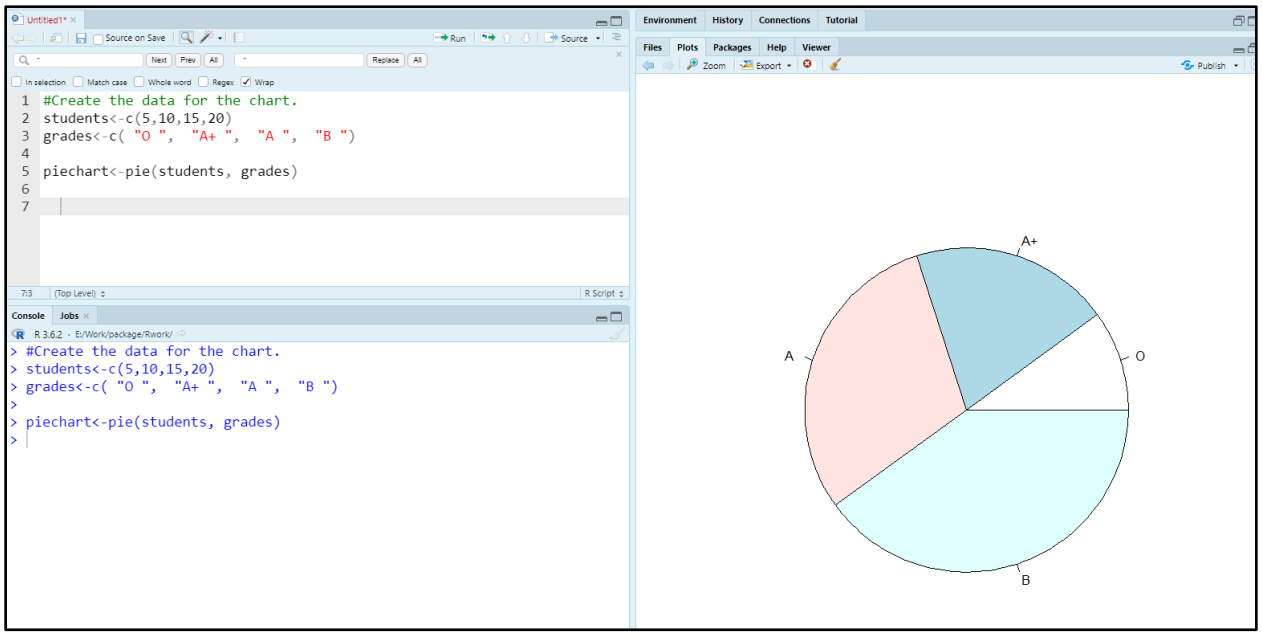
STEP3: Use R code pie() for constructing pie chart.

STEP4: The results will be displayed on the R console window.

R INPUT:-

```
#Create the data for the chart.

students<-c(5,10,15,20)

grades<-c( "O ",  "A+ ",  "A ",  "B ")

piechart<-pie(students, grades)

dev.off()
```

R OUTPUT:-



RESULTS:-

The pie chart has been constructed successfully.

INFERENCE:-

From the results, we can conclude that less students score O grade and more students score B grade.

BOXPLOT

EXPERIMENT NO.:- 13

DATE:-06.05.2022

PROBLEM:-

The income of 10 employees in a company has been given below.

EMPLOYEE NO.	1	2	3	4	5	6	7	8	9	10
GENDER	M	F	M	M	F	M	M	M	F	F
INCOME(in k)	95	90	85	80	75	70	65	60	55	50

AIM:-

To construct the boxplot for the given data.

PROCEDURE:-

STEP1:Open R console window.

STEP2:Enter the data using c() command.

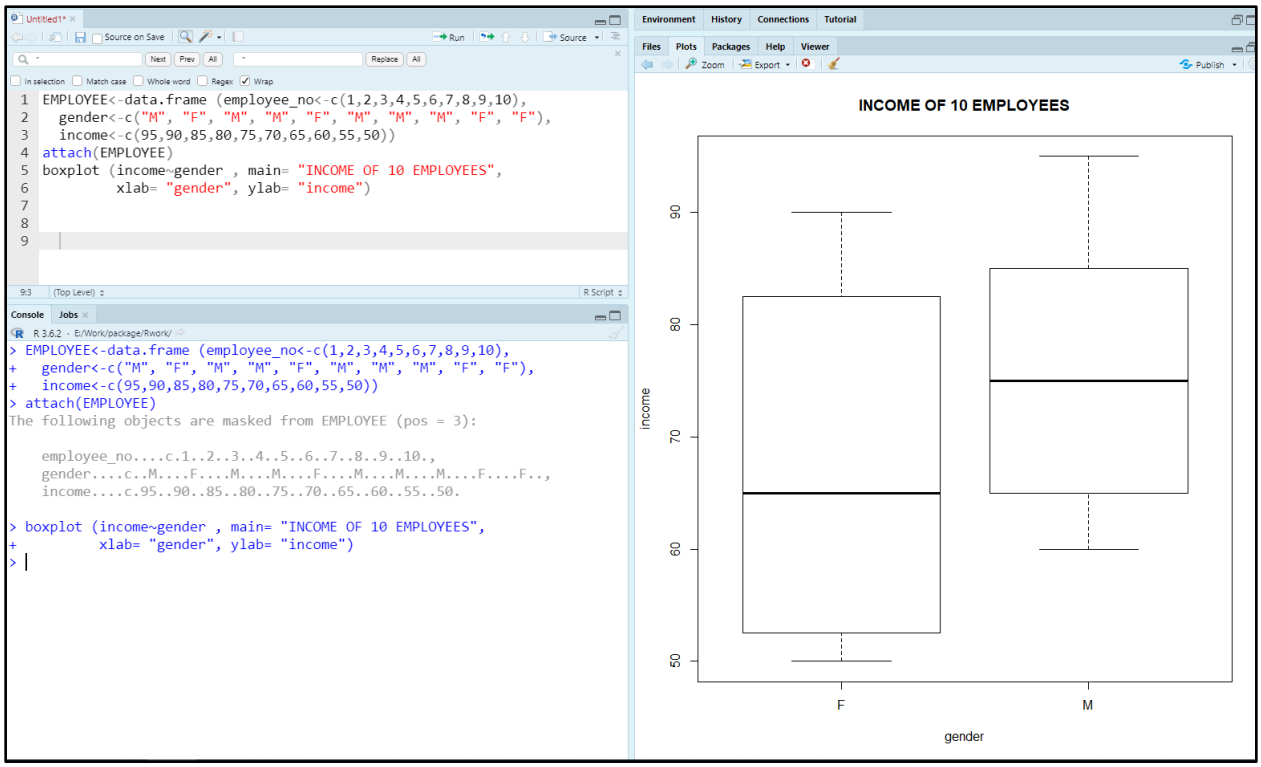
STEP3:Construct boxplot using boxplot() command.

STEP4:The output will be displayed on R console window.

R INPUT:-

```
EMPLOYEE<-data.frame (employee_no<-c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10), gender<-c("M", "F", "M", "M", "F", "M", "M", "M", "F", "F"), income<-c(95, 90, 85, 80, 75, 70, 65, 60, 55, 50))

attach(EMPLOYEE)

boxplot (income~gender , main= "INCOME OF 10 EMPLOYEES", xlab= "gender", ylab= "income")
```

R OUTPUT:-



RESULT:-

The boxplot for the income of 10 employees in a company has been constructed successfully.

INFERENCE:-

From the results, we can conclude that there are no employees whose income is above or below the outliers.

DESCRIPTIVE STATISTICS

EXPERIMENT NO.:- 14

DATE:- 06.05.2022

PROBLEM:-

The data regarding performance of 10 students in a subject are given below. Compute the descriptive statistics for the variable performance with reference to gender.

GENDER	M	F	M	M	F	M	M	M	F	F
MARKS	50	45	40	35	37	42	33	24	25	22

AIM:-

To find whether the female students performance is more consistent than male students performance or not.

PROCEDURE:-

STEP1: Open R console window.

STEP2: Enter the data "GENDER" and "MARKS" using the command data.frame ( ).

STEP3: Compute coefficient of variation using an user defined function.

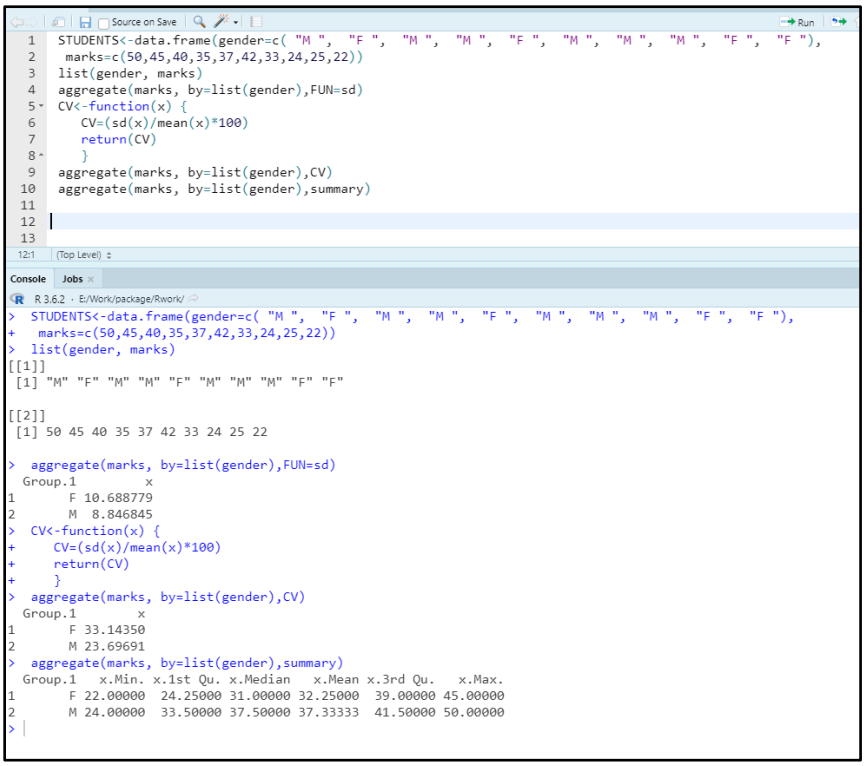
STEP4:By using the R command "aggregate()",gender wise marks can be calculated.

STEP5:The output will be displayed in the R console window.

R INPUT:-

```
STUDENTS<-data.frame(gender=c( "M ",  "F ",  "M ",  "M ",  "F ",  "M ",  "M ",  "M ",  "F ",  "F "),

marks=c(50, 45, 40, 35, 37, 42, 33, 24, 25, 22))

list(gender, marks)

aggregate(marks, by=list(gender),FUN=sd)

CV<-function(x) {

    CV=(sd(x)/mean(x)*100)

    return(CV)

    }

aggregate(marks, by=list(gender),CV)

aggregate(marks, by=list(gender),summary)
```

R OUTPUT:-



```
STUDENTS<-data.frame(gender=c( "M ",  "F ",  "M ",  "M ",  "F ",  "M ",  "M ",  "M ",  "F ",  "F "),
 marks=c(50,45,40,35,37,42,33,24,25,22))
list(gender, marks)
aggregate(marks, by=list(gender),FUN=sd)
CV<-function(x) {
   CV=(sd(x)/mean(x)*100)
   return(CV)
   }
aggregate(marks, by=list(gender),CV)
aggregate(marks, by=list(gender),summary)
```

```
> STUDENTS<-data.frame(gender=c( "M ",  "F ",  "M ",  "M ",  "F ",  "M ",  "M ",  "M ",  "F ",  "F "),
+  marks=c(50,45,40,35,37,42,33,24,25,22))
> list(gender, marks)
[[1]]
 [1] "M" "F" "M" "M" "F" "M" "M" "M" "F" "F"

[[2]]
 [1] 50 45 40 35 37 42 33 24 25 22

> aggregate(marks, by=list(gender),FUN=sd)
  Group.1         x
1       F 10.688779
2       M  8.846845
> CV<-function(x) {
+    CV=(sd(x)/mean(x)*100)
+    return(CV)
+    }
> aggregate(marks, by=list(gender),CV)
  Group.1        x
1       F 33.14350
2       M 23.69691
> aggregate(marks, by=list(gender),summary)
  Group.1   x.Min. x.1st Qu. x.Median   x.Mean x.3rd Qu.   x.Max.
1       F 22.00000  24.25000 31.00000 32.25000  39.00000 45.00000
2       M 24.00000  33.50000 37.50000 37.33333  41.50000 50.00000
>
```

RESULTS:-

GENDER	CV	MINIMUM	MAXIMUM	MEAN	STANDARD DEVIATION
F	33.1435	22	45	32.25	10.688779
M	23.6969	24	50	37.33	8.846845

INFERENCE:-

From the results, we may conclude that the female student's performance is more consistent than the male student's performance.

DESCRIPTIVE STATISTICS (GROUPED DATA)

EXPERIMENT NO.:- 15

DATE:- 13.05.2022

PROBLEM:-

Calculate the mean, median, mode, standard deviation, skewness and kurtosis for the following grouped data.

MARKS	10-20	20-30	30-40	40-50	50-60
NO. OF PERSONS	8	22	43	17	10

AIM:-

To calculate the mean, median, mode, standard deviation, skewness and kurtosis for the given grouped data related to the marks scored by 100 students.

PROCEDURE:-

STEP1: Open R console window.

STEP2: Enter the mid value of marks and number of students using c() command.

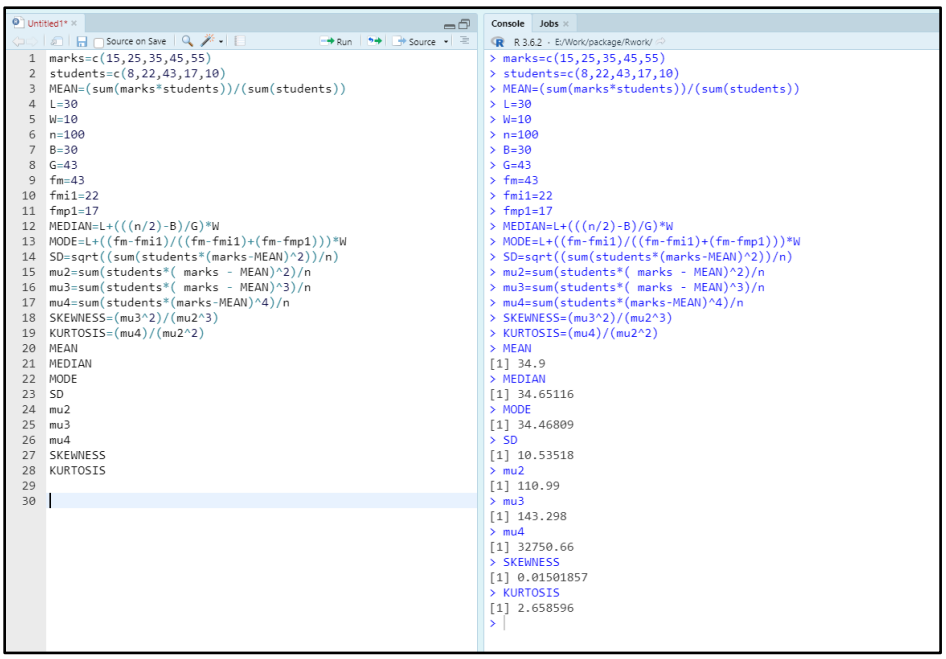
STEP3: The mean is calculated by using the ∑fx/∑f, where f is the number of students and x is the mid value.
STEP4: Similarly formulate the median, mode, standard deviation, skewness and kurtosis using the equations.
STEP5: By using the formulae, we calculate the values and the outputs will be displayed in the R console window.

R INPUT:-

```
marks=c(15,25,35,45,55)
students=c(8,22,43,17,10)
MEAN=(sum(marks*students))/(sum(students))
L=30
W=10
n=100
B=30
G=43
fm=43
fmi1=22
fmp1=17
MEDIAN=L+(((n/2)-B)/G)*W
MODE=L+((fm-fmi1)/((fm-fmi1)+(fm-fmp1)))*W
SD=sqrt((sum(students*(marks-MEAN)^2))/n)
mu2=sum(students*( marks - MEAN)^2)/n
mu3=sum(students*( marks - MEAN)^3)/n
mu4=sum(students*(marks-MEAN)^4)/n
SKEWNESS=(mu3^2)/(mu2^3)
KURTOSIS=(mu4)/(mu2^2)
MEAN
```

MEDIAN
MODE
SD
mu2
mu3
mu4
SKEWNESS
KURTOSIS

R OUTPUT:-



```
marks=c(15,25,35,45,55)
students=c(8,22,43,17,10)
MEAN=(sum(marks*students))/(sum(students))
L=30
W=10
n=100
B=30
G=43
fm=43
fmi1=22
fmp1=17
MEDIAN=L+(((n/2)-B)/G)*W
MODE=L+((fm-fmi1)/((fm-fmi1)+(fm-fmp1)))*W
SD=sqrt((sum(students*(marks-MEAN)^2))/n)
mu2=sum(students*( marks - MEAN)^2)/n
mu3=sum(students*( marks - MEAN)^3)/n
mu4=sum(students*(marks-MEAN)^4)/n
SKEWNESS=(mu3^2)/(mu2^3)
KURTOSIS=(mu4)/(mu2^2)
MEAN
MEDIAN
MODE
SD
mu2
mu3
mu4
SKEWNESS
KURTOSIS
```

```
> marks=c(15,25,35,45,55)
> students=c(8,22,43,17,10)
> MEAN=(sum(marks*students))/(sum(students))
> L=30
> W=10
> n=100
> B=30
> G=43
> fm=43
> fmi1=22
> fmp1=17
> MEDIAN=L+(((n/2)-B)/G)*W
> MODE=L+((fm-fmi1)/((fm-fmi1)+(fm-fmp1)))*W
> SD=sqrt((sum(students*(marks-MEAN)^2))/n)
> mu2=sum(students*( marks - MEAN)^2)/n
> mu3=sum(students*( marks - MEAN)^3)/n
> mu4=sum(students*(marks-MEAN)^4)/n
> SKEWNESS=(mu3^2)/(mu2^3)
> KURTOSIS=(mu4)/(mu2^2)
> MEAN
[1] 34.9
> MEDIAN
[1] 34.65116
> MODE
[1] 34.46809
> SD
[1] 10.53518
> mu2
[1] 110.99
> mu3
[1] 143.298
> mu4
[1] 32750.66
> SKEWNESS
[1] 0.01501857
> KURTOSIS
[1] 2.658596
>
```

RESULTS:-

The mean, median, mode, standard deviation, skewness and kurtosis have been calculated successfully for the grouped data.

MEAN=34.9
MEDIAN=34.65116
MODE=34.46809

STANDARD DEVIATION=10.53518
SKEWNESS=0.01501857
KURTOSIS=2.658596

INFERENCE:-

From the results, we can conclude that the mean, median and mode are normally distributed since their values are approximately same. Since the skewness value is greater than 0, its positively skewed. The kurtosis value is lesser than 3, so its platykurtic.

CORRELATION AND REGRESSION

EXPERIMENT NO.:- 16

DATE:-13.05.2022

PROBLEM:-

The marks scored by 9 students in two subjects has been given. Compute two regression and a correlation coefficient.

PHY	**19**	**25**	**24**	**10**	**37**	**18**	**29**	**14**	**21**
CHEM	**10**	**12**	**13**	**8**	**36**	**23**	**27**	**14**	**12**

AIM:-

To compute the regression and correlation coefficients for the data related to the marks scored by 9 students in two subjects.

PROCEDURE:-

STEP1: Open R console window.

STEP2: Enter the data by the c() command.

STEP3: Compute the regression and correlation coefficients using the formulae.

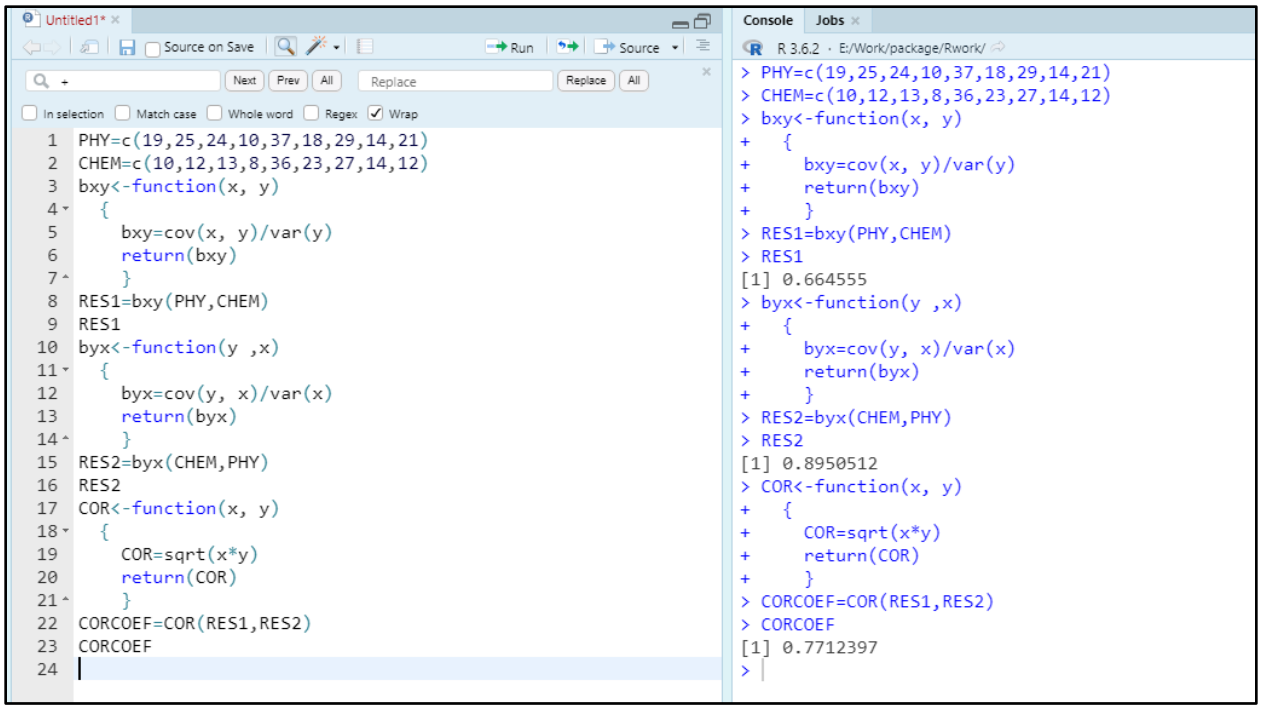
STEP4: The output will be displayed on the R console window.

R INPUT:-

```
PHY=c(19,25,24,10,37,18,29,14,21)
CHEM=c(10,12,13,8,36,23,27,14,12)
bxy<-function(x, y)
  {
    bxy=cov(x, y)/var(y)
    return(bxy)
    }
RES1=bxy(PHY,CHEM)
RES1
byx<-function(y ,x)
  {
    byx=cov(y, x)/var(x)
    return(byx)
    }
```

```
RES2=byx(CHEM,PHY)
RES2
COR<-function(x, y)
  {
    COR=sqrt(x*y)
    return(COR)
    }
CORCOEF=COR(RES1,RES2)
CORCOEF
```

R OUTPUT:-



RESULTS:-

The two regression and correlation coefficients has been found out successfully.

1. **bxy= 0.664555**
2. **byx= 0.8950512**
3. **COR= 0.7712397**

INFERENCE:-

From the results, we can conclude that the PHY and CHEM marks are positively correlated since the correlation coefficient lies between 0 and 1.

MAN WHITNEY U TEST

EXPERIMENT NO.:- 17

DATE:- 20.05.2022

PROBLEM:-

The following data is the mark obtained by male and female students in a class for the Mathematics test.

Male	**59**	**47**	**63**	**78**	**69**	**64**	**53**	**72**	**67**	**74**	**67**	**61**			
Female	**67**	**60**	**64**	**84**	**91**	**48**	**81**	**93**	**93**	**59**	**61**	**92**	**45**	**91**	**59**

Is there enough evidence to conclude that there is a difference between in the mean marks scored by male and female students?

AIM:-

To test there is enough evidence to conclude that there is a significant difference between in the mean marks scored by male and female students.

PROCEDURE:-

STEP1: Setting the hypothesis

H0: There is no significant difference in the mean marks scored by male and female students.

H1: There is a significant difference in the mean marks scored by male and female students.

STEP2: Open R console window.

STEP3: Enter the data by using c() command.

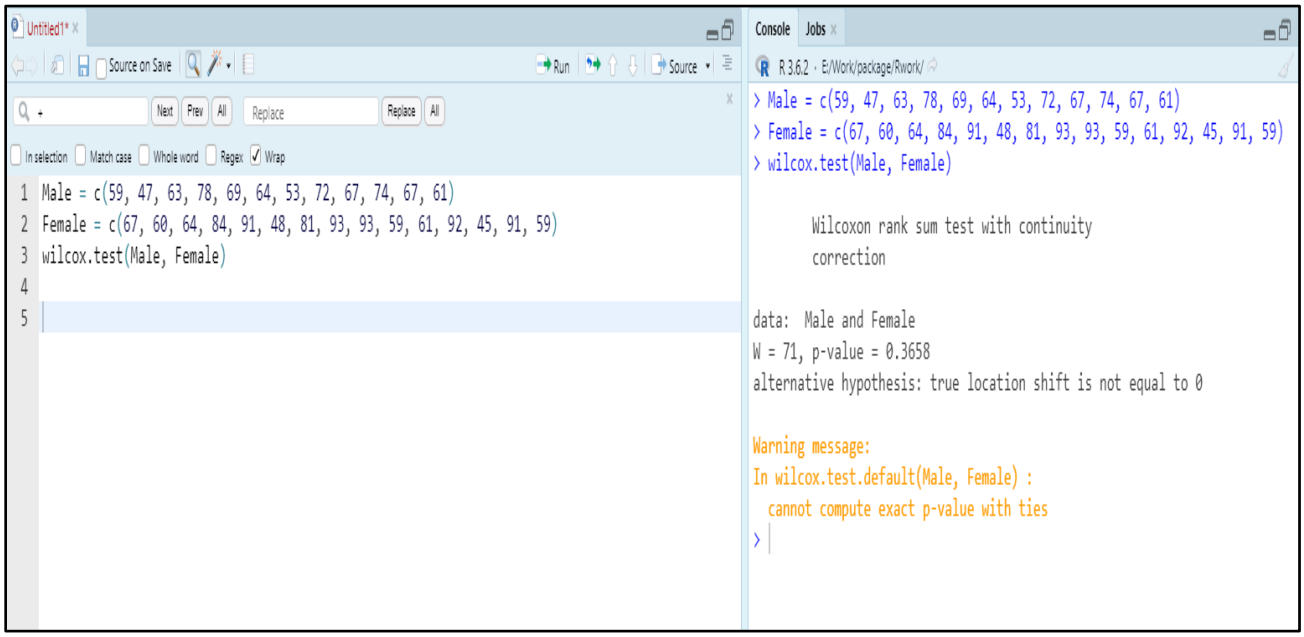
STEP4: Perform "MAN WHITNEY U TEST" by the command wilcox.test()

STEP5: The output will be displayed in the R console window.

R INPUT:-

```
Male = c(59, 47, 63, 78, 69, 64, 53, 72, 67, 74, 67, 61)

Female = c(67, 60, 64, 84, 91, 48, 81, 93, 93, 59, 61, 92, 45, 91, 59)

wilcox.test(Male, Female)
```

R OUTPUT:-



RESULTS:-

The Man Whitney U test has been performed successfully.

The test statistic W = 109

p-value = 0.3658

α=0.05

Since p> α, null hypothesis(H0) may be accepted at 5% level of significance.

INFERENCE:-

From the result, we can conclude that there is no significant difference between the mean marks scored by male and female students.

ONE WAY ANALYSIS OF VARIANCE (ANOVA)

EXPERIMENT NO.:- 18

DATE:- 20.05.2022

PROBLEM:-

The following data describes the lives of four brands of electric fans. Test whether the mean electric lives of those four brands are same using one way analysis of variance (ANOVA).

BRAND	LIVES OF THE ELECTRIC FANS (IN HOURS)							
PHILIPS	1600	1610	1650	1680	1700	1720	1800	-
LG	1580	1640	1640	1700	1750	-	-	-
CROMPTON	1460	1550	1600	1620	1640	1660	1740	1820
SURYA	1510	1520	1570	1570	1600	1680	-	-

AIM:-

To test whether each of the factors independently affect the sales and there is a combined effect of pack design and price by using R Studio.

PROCEDURE:-

STEP1: Setting the hypothesis.

H0: There is no significant difference between the average life of the electric fans, i.e. ,

$\mu1 = \mu2 = \mu3 = \mu4$

STEP2:Open R console window.

STEP3: Enter data using c() command.

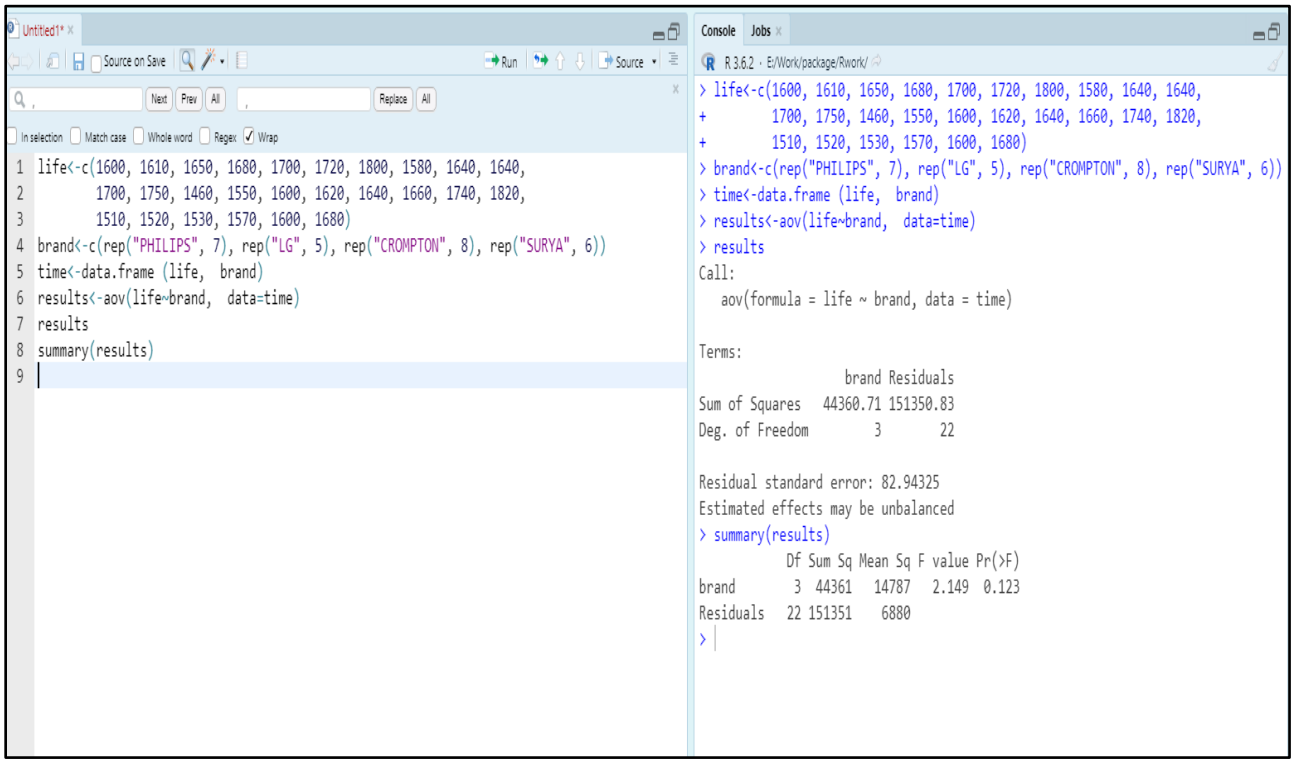
STEP4:Perform analysis of variance using the command "aov()".

STEP5:The outputs will be displayed on the R console window.

R INPUT:-

```
life<-c(1600, 1610, 1650, 1680, 1700, 1720, 1800,
1580, 1640, 1640,
        1700, 1750, 1460, 1550, 1600, 1620, 1640,
1660, 1740, 1820,
        1510, 1520, 1530, 1570, 1600, 1680)
brand<-c(rep("PHILIPS", 7), rep("LG", 5),
rep("CROMPTON", 8), rep("SURYA", 6))
time<-data.frame (life,  brand)
results<-aov(life~brand,  data=time)
results
summary(results)
```

R OUTPUT:-



```
life<-c(1600, 1610, 1650, 1680, 1700, 1720, 1800, 1580, 1640, 1640,
        1700, 1750, 1460, 1550, 1600, 1620, 1640, 1660, 1740, 1820,
        1510, 1520, 1530, 1570, 1600, 1680)
brand<-c(rep("PHILIPS", 7), rep("LG", 5), rep("CROMPTON", 8), rep("SURYA", 6))
time<-data.frame (life,  brand)
results<-aov(life~brand,  data=time)
results
summary(results)
```

```
> life<-c(1600, 1610, 1650, 1680, 1700, 1720, 1800, 1580, 1640, 1640,
+         1700, 1750, 1460, 1550, 1600, 1620, 1640, 1660, 1740, 1820,
+         1510, 1520, 1530, 1570, 1600, 1680)
> brand<-c(rep("PHILIPS", 7), rep("LG", 5), rep("CROMPTON", 8), rep("SURYA", 6))
> time<-data.frame (life,  brand)
> results<-aov(life~brand,  data=time)
> results
Call:
   aov(formula = life ~ brand, data = time)

Terms:
                    brand Residuals
Sum of Squares   44360.71 151350.83
Deg. of Freedom         3        22

Residual standard error: 82.94325
Estimated effects may be unbalanced
> summary(results)
            Df Sum Sq Mean Sq F value Pr(>F)
brand        3  44361   14787   2.149  0.123
Residuals   22 151351    6880
>
```

RESULTS:-

Since p-value is equal to 0.123 which is greater than 0.05 (p>α), null hypothesis may be accepted at 5% level of significance.

INFERENCE:-

From the results, we can conclude that there is no significant difference between the average life of electric fans of different brands.

ONE WAY ANALYSIS OF VARIANCE (ANOVA)

EXPERIMENT NO.:- 19
DATE:-27.05.2022
PROBLEM:-

The following table gives the yield of wheat of 10 plots , all of approximately equal fertility, when v arieties of wheat are cultivated completely . Test the hypothesis the varieties of wheat ar e not significantly different.

PLOTS	1	2	3	4	5	6	7	8	9	10
VAREITY	A	B	D	C	B	C	A	D	B	D
YIELD	32	34	29	31	33	34	34	26	36	30

AIM:-

To test whether the varieties of wheat are significantly different or not.

PROCEDURE:-

STEP1:Setting the hypothesis.
H0:There is no significant difference between the varieties of wheat.

STEP2:Open R console window.

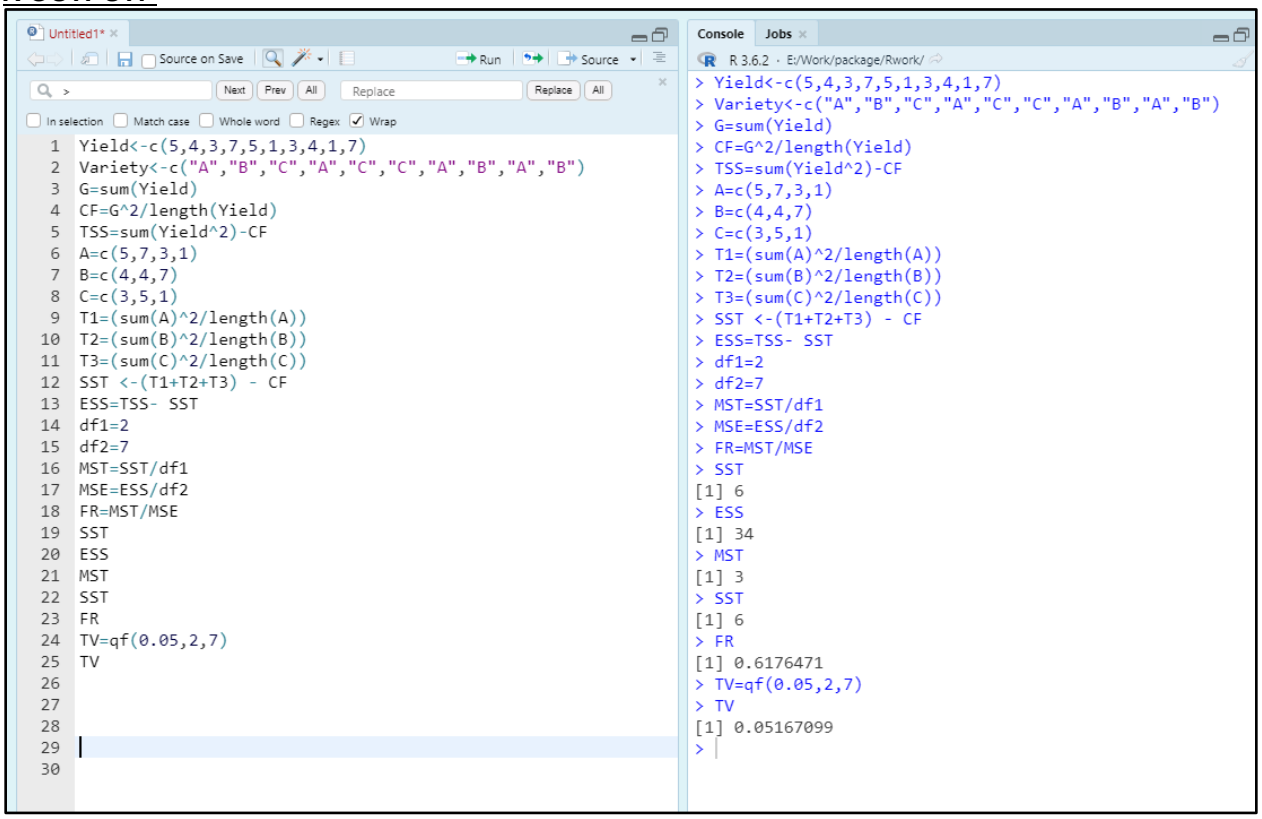
STEP3: Enter the data using c() command.

STEP4:Construct the formula for finding F-ratio using user defined functions.

STEP5: Then the results will be displayed in the R console window.
R INPUT:-

```
Yield<-c(5,4,3,7,5,1,3,4,1,7)
Variety<-c("A","B","C","A","C","C","A","B","A","B")
G=sum(Yield)
CF=G^2/length(Yield)
TSS=sum(Yield^2)-CF
A=c(5,7,3,1)
B=c(4,4,7)
C=c(3,5,1)
T1=(sum(A)^2/length(A))
T2=(sum(B)^2/length(B))
T3=(sum(C)^2/length(C))
SST <-(T1+T2+T3) - CF
ESS=TSS- SST
df1=2
df2=7
MST=SST/df1
MSE=ESS/df2
FR=MST/MSE
SST
ESS
MST
SST
FR
TV=qf(0.05,2,7)
TV
```

R OUTPUT:-



```
Yield<-c(5,4,3,7,5,1,3,4,1,7)
Variety<-c("A","B","C","A","C","C","A","B","A","B")
G=sum(Yield)
CF=G^2/length(Yield)
TSS=sum(Yield^2)-CF
A=c(5,7,3,1)
B=c(4,4,7)
C=c(3,5,1)
T1=(sum(A)^2/length(A))
T2=(sum(B)^2/length(B))
T3=(sum(C)^2/length(C))
SST <-(T1+T2+T3) - CF
ESS=TSS- SST
df1=2
df2=7
MST=SST/df1
MSE=ESS/df2
FR=MST/MSE
SST
ESS
MST
SST
FR
TV=qf(0.05,2,7)
TV
```

```
> Yield<-c(5,4,3,7,5,1,3,4,1,7)
> Variety<-c("A","B","C","A","C","C","A","B","A","B")
> G=sum(Yield)
> CF=G^2/length(Yield)
> TSS=sum(Yield^2)-CF
> A=c(5,7,3,1)
> B=c(4,4,7)
> C=c(3,5,1)
> T1=(sum(A)^2/length(A))
> T2=(sum(B)^2/length(B))
> T3=(sum(C)^2/length(C))
> SST <-(T1+T2+T3) - CF
> ESS=TSS- SST
> df1=2
> df2=7
> MST=SST/df1
> MSE=ESS/df2
> FR=MST/MSE
> SST
[1] 6
> ESS
[1] 34
> MST
[1] 3
> SST
[1] 6
> FR
[1] 0.6176471
> TV=qf(0.05,2,7)
> TV
[1] 0.05167099
>
```

RESULTS:-

F Ratio= 0.6176471
Tabled value= 0.0567099

Since F value > Tabled value, we reject null hypothesis.
INFERENCE:-

From the results, we can conclude that there is no significant difference between the varieties of wheat.

TWO WAY ANALYSIS OF VARIANCE (ANOVA)

EXPERIMENT NO.:- 20
DATE:- 27.05.2022
PROBLEM:-

Perform two way analysis of variance for the following data.

PLOTS	TREATMENTS			
	A	B	C	D
I.	38	40	41	39
II.	45	42	49	36
III.	40	38	42	42

AIM:-
To test whether there is a significant difference between the plots and the treatments.

PROCEDURE:-

STEP1: Setting the hypothesis.

H01:There is no significant difference between the three plots.

H02:There is no significant difference between the four treatments.

STEP2:Open R console window.

STEP3:Enter the data using c() command.

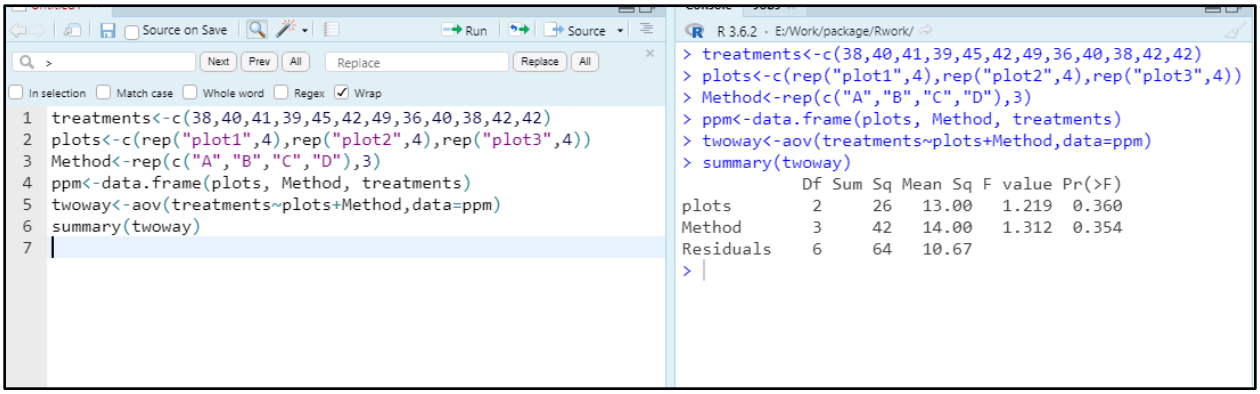
STEP4:Perform analysis of variance using aov() command.

STEP5:The results will be displayed on the R console window.

R INPUT:-

```
treatments<-c(38,40,41,39,45,42,49,36,40,38,42,42)
plots<-c(rep("plot1",4), rep("plot2",4), rep("plot3",4))
Method<-rep(c("A","B","C","D"),3)
ppm<-data.frame(plots, Method, treatments)
twoway<-aov(treatments~plots+Method,data=ppm)
summary(twoway)
```

R OUTPUT:-



RESULT:-

For H01, since F value is 0.360 which is greater than 0.05, the null hypothesis(H01) is accepted at 5% level of significance.

For H02, since F value is 0.354 which is greater than 0.05, the null hypothesis(H02) is accepted at 5% level of significance.

INFERENCE:-

From the result, we can conclude that there is no significant difference between the row as well as column means, i.e., there is no significant difference between the three plots as well as the four tre atments.

SIMPLE LINEAR REGRESSION

EXPERIMENT NO.:- 21

DATE:-03.06.2022

PROBLEM:-

A teacher collected the following data on annual marks and number of students. Fit a suitable regression model for the data.

	1	2	3	4	5	6	7	8	9	10
NO. OF STUDENTS	1	3	4	4	6	8	10	10	11	13
ANNUAL MARKS	80	97	92	102	103	111	119	123	117	136

AIM:-

To fit a Simple linear Regression Model for the given data.

PROCEDURE:-

STEP1: Setting the hypothesis
Ho: The fitted model is not significant.
H1: The fitted model is significant.

STEP 2: Click R console window.

STEP 3: Enter the given using c() function.

STEP 4: Use R code lm() for fitting regression model.

STEP 5: Find out summary of model using the function summary().

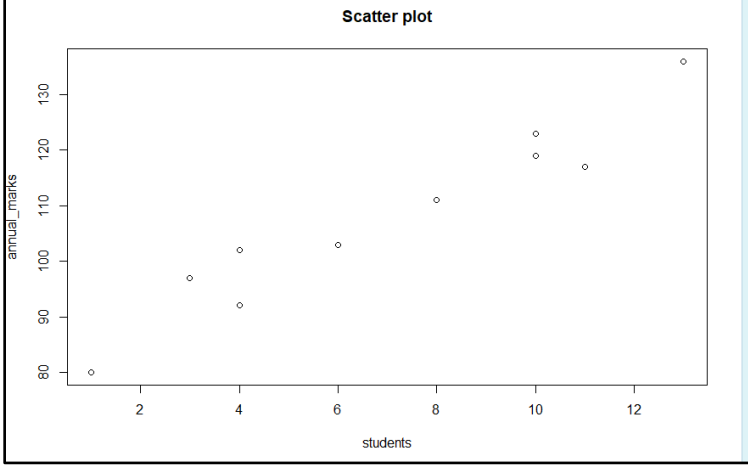
STEP 6: Then the results will be displayed in the R console window.

R INPUT:-

```
rollno<-c(1:10)
students<-c(1,3,4,4,6,8,10,10,11,13)
annual_marks<-c(80,97,92,102,103,111,119,123,117,136)
marksdetails<-data.frame(rollno, students, annual_marks)
marksdetails
plot(students, annual_marks, main="Scatter plot")
cor.test (students, annual_marks)
regmodel<-lm(annual_marks~students)
summary(regmodel)
```

R OUTPUT:-

```
rollno<-c(1:10)
students<-c(1,3,4,4,6,8,10,10,11,13)
annual_marks<-c(80,97,92,102,103,111,119,123,117,136)
marksdetails<-data.frame(rollno, students, annual_marks)
marksdetails
plot(students, annual_marks, main="Scatter plot")
cor.test (students, annual_marks)
regmodel<-lm(annual_marks~students)
summary(regmodel)
```




```
> plot(students, annual_marks, main="Scatter plot")
> cor.test (students, annual_marks)

        Pearson's product-moment correlation

data:  students and annual_marks
t = 10.34, df = 8, p-value = 6.609e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8529446 0.9918346
sample estimates:
      cor
0.9645646

> regmodel<-lm(annual_marks~students)
> summary(regmodel)

Call:
lm(formula = annual_marks ~ students)

Residuals:
   Min     1Q Median     3Q    Max
 -7.00  -3.25  -1.00   3.75   6.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.0000     3.0753   26.01 5.12e-09
students      4.0000     0.3868   10.34 6.61e-06

(Intercept) ***
students    ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.61 on 8 degrees of freedom
Multiple R-squared:  0.9304,    Adjusted R-squared:  0.9217
F-statistic: 106.9 on 1 and 8 DF,  p-value: 6.609e-06
```

RESULTS:-

1. From scatter plot, we can conclude that Annual marks and number of students had linear relationship.
2. The best fitted simple linear Regression model is Y= 80+4X
3. 92.17% variations in Y are explained by the variable X.
4. Since p<0.05 (0.000006609 < 0.05) at 5% level, we reject H0.

INFERENCE:-

From the result, we can conclude that the fitted regression model is significant.

MULTIPLE LINEAR REGRESSION

EXPERIMENT NO.:- 22
DATE:-03.06.2022
PROBLEM:-

Find suitable multiple regression model for the following data.

X1	**4**	**6**	**7**	**9**	**13**	**15**
X2	**15**	**12**	**8**	**6**	**4**	**8**
X3	**30**	**24**	**20**	**14**	**10**	**4**

AIM:-

To fit a Multiple linear Regression Model for the given data.

PROCEDURE:-

STEP 1: Setting the hypothesis.

Ho: The fitted model is not significant.

H1: The fitted model is significant.

STEP 2: Click R console window.

STEP 3: Enter the given using c() function.

STEP 4: Use R code lm() for fitting regression model.

STEP 5: Find out summary of model using the function summary().
STEP 6: Use the R code plot() for plotting the data.

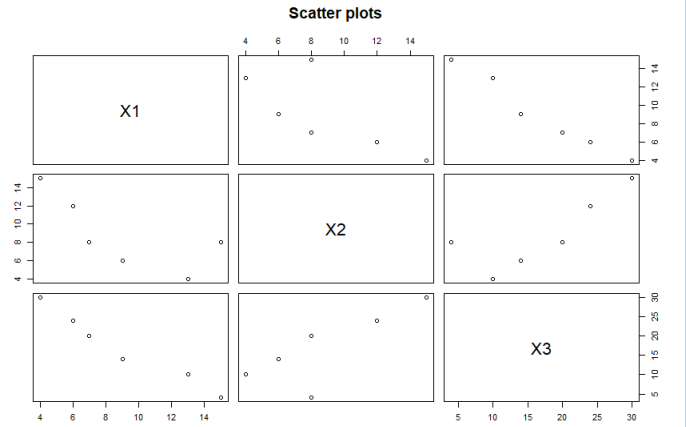
STEP 7: Then the results will be displayed in the R console window.

R INPUT:-

```
X1<-c(4,6,7,9,13,15)
X2<-c(15,12,8,6,4,8)
X3<-c(30,24,20,14,10,4)
Details<-data.frame(X1,X2,X3)
Model<-lm(X1~X2+X3)
Model
summary(Model)
plot(~X1+X2+X3,data=Details, main="Scatter plots")
```

R OUTPUT:-

```
X1<-c(4,6,7,9,13,15)
X2<-c(15,12,8,6,4,8)
X3<-c(30,24,20,14,10,4)
Details<-data.frame(X1,X2,X3)
Model<-lm(X1~X2+X3)
Model
summary(Model)
plot(~X1+X2+X3,data=Details, main="Scatter plots")
```




```
> X1<-c(4,6,7,9,13,15)
> X2<-c(15,12,8,6,4,8)
> X3<-c(30,24,20,14,10,4)
> Details<-data.frame(X1,X2,X3)
> Model<-lm(X1~X2+X3)
> Model

Call:
lm(formula = X1 ~ X2 + X3)

Coefficients:
(Intercept)           X2           X3
    16.0189       0.1273      -0.4790

> summary(Model)

Call:
lm(formula = X1 ~ X2 + X3)

Residuals:
       1        2        3        4        5        6
 0.44226 -0.04996 -0.45681 -1.07635  1.26217 -0.12130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.01886    1.09443  14.637 0.000692 ***
X2           0.12732    0.19009   0.670 0.550932
X3          -0.47903    0.08021  -5.972 0.009394 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.028 on 3 degrees of freedom
Multiple R-squared:  0.9647,    Adjusted R-squared:  0.9412
F-statistic: 41.05 on 2 and 3 DF,  p-value: 0.00662

> plot(~X1+X2+X3,data=Details, main="Scatter plots")
>
```

RESULTS:-

1. From the matrix scatter plot, we can conclude that by X1, X2 and X3 has linear relationship.
2. The best fitted multiple linear Regression model is X1= 16.01886+0.12732X2-0.47903X3.
3. 96.47% variations in X1 are explained by the variable X2 & X3.
4. Since p<0.05 (0.00662 < 0.05) at 5% level, we reject H0 and conclude that the fitted model is significant.

Result:

INFERENCE: -

From the result, we conclude that the fitted multiple regression model is significant.

T – TEST

EXPERIMENT NO.:- 23

DATE:-17.06.2022

PROBLEM:-

The lifetime of electronic bulbs for a random sample of ten from a large consignment gave the following data.

ITEM	1	2	3	4	5	6	7	8	9	10
LIFE	4.2	4.6	3.9	4.1	5.2	3.8	3.8	3.9	4.3	5.6

Can we accept the hypothesis that the average lifetime of bulbs is 4000hrs.

AIM:-

To test the average life time is equal to 4000hrs by using one sample t test.

PROCEDURE:-

STEP 1: Setting the hypothesis

Ho: The average lifetime of bulbs is 4000hrs.

H1: The average lifetime of bulbs is not 4000hrs.

STEP 2: Click R console window.

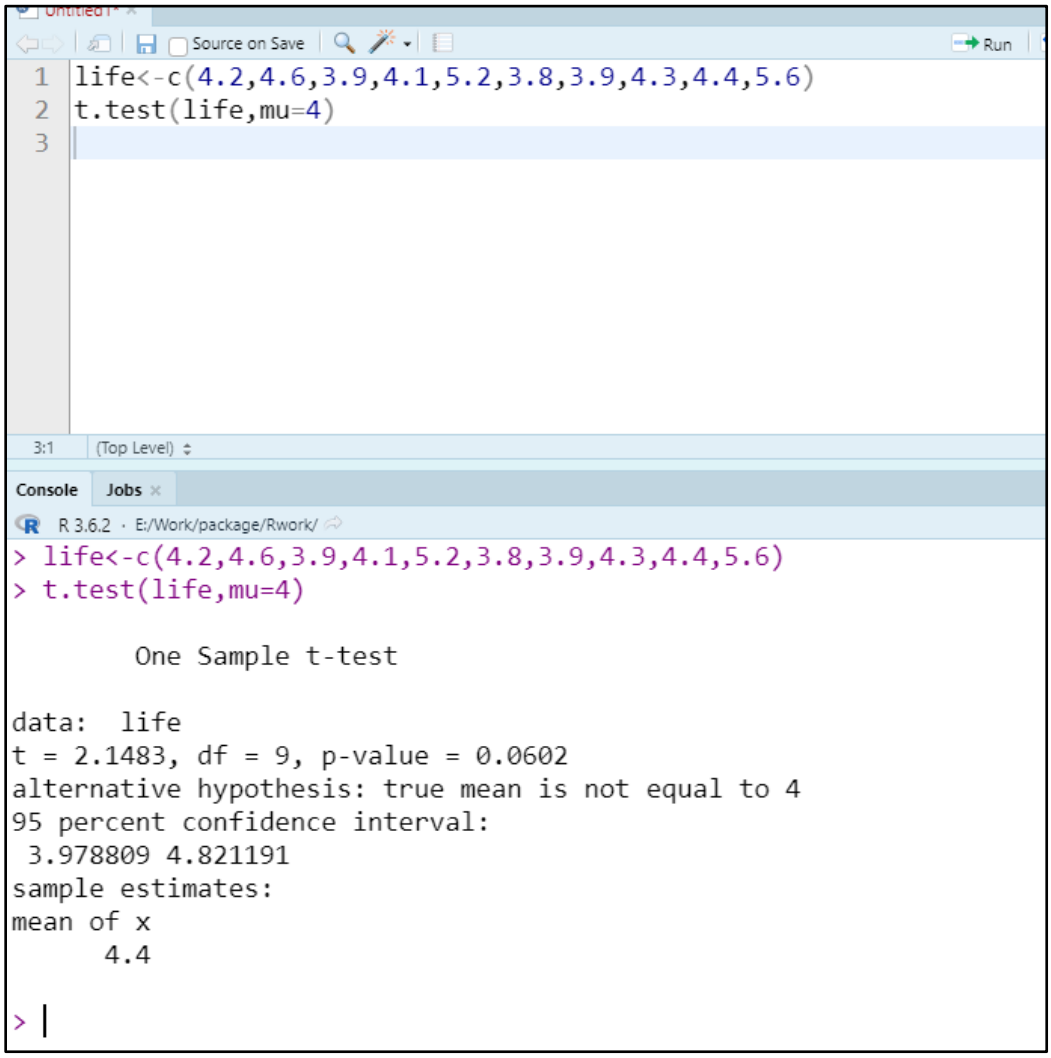
STEP 3: Enter the given data using c() function.

STEP 4: Use the R code t.test() for performing t test.
STEP 5: Then the results will be displayed in the R console window.

R INPUT:-

```
life<-c(4.2,4.6,3.9,4.1,5.2,3.8,3.9,4.3,4.4,5.6)
t.test(life,mu=4)
```

R OUTPUT:-



RESULTS:-

From the given data, we may observed that the p value is (.0602) greater than 5% significance level (1.96) So the null hypothesis is accepted.

INFERENCE:-

From the result, we conclude that the average life of bulbs is 4000hrs.