VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Digital Assignment – 01

R Lab

Name – ILIYAS ANSARI

Reg. No. – 22MCA1079

Lab Submission – 01

Submitted to – Dr. A. David Maxim Gururaj

Lab Slot – L5+L6

Exp. 1: Introduction: Understanding Data types; Importing/Exporting data

Import the data file 'Health_data.csv' and perform the following

- data=read.csv("C:/Users/ilyas/Desktop/R/DA-1/Health_data.csv")
- data

1. Display height and weight attribute separately.

```
> a1=data$HEIGHT
> a1
 [1] 154.5 173.5 154.0 184.5 184.5
 [6] 180.5 181.0 176.0 191.0 178.5
[11] 166.5 160.0 180.5 155.5 181.0
[16] 162.0 175.5 159.0 164.0 172.0
[21] 179.0 160.5 140.5 148.5 151.0
[26] 160.0 163.5 158.0 153.5 160.5
[31] 179.0 160.0 168.0 178.5 177.0
[36] 172.0 162.0 167.0 163.0 180.0
[41] 162.5 179.0 180.0 165.0 168.5
[46] 164.5 153.5 148.5 165.5 167.5
```

```
> a2=data$WEIGHT
> a2
 [1] 67.75 72.25 66.25 72.25 71.25
 [6] 74.75 69.75 72.50 74.00 73.50
[11] 74.50 76.00 69.50 71.25 69.50
[16] 66.00 71.00 71.00 67.75 73.50
[21] 68.00 69.75 68.25 70.00 67.75
[26] 71.50 67.50 67.50 64.75 69.00
[31] 73.75 71.25 71.25 71.00 73.50
[36] 65.00 70.00 68.25 72.25 67.00
[41] 68.75 29.50 70.00 71.50 68.00
[46] 73.25 67.50 71.25 68.50 66.75
```

2. Create a data frame for height and weight attributes.

- b1=data.frame(data$HEIGHT, data$WEIGHT)
- b1

```
> b1=data.frame(data$HEIGHT, data$WEIGHT)
> b1
   data.HEIGHT data.WEIGHT
1        154.5        67.75
2        173.5        72.25
3        154.0        66.25
4        184.5        72.25
5        184.5        71.25
6        180.5        74.75
7        181.0        69.75
8        176.0        72.50
9        191.0        74.00
10       178.5        73.50
11       166.5        74.50
12       160.0        76.00
13       180.5        69.50
14       155.5        71.25
15       181.0        69.50
16       162.0        66.00
17       175.5        71.00
18       159.0        71.00
19       164.0        67.75
20       172.0        73.50
21       179.0        68.00
22       160.5        69.75
23       140.5        68.25
24       148.5        70.00
25       151.0        67.75
26       160.0        71.50
27       163.5        67.50
28       158.0        67.50
29       153.5        64.75
30       160.5        69.00
31       179.0        73.75
32       160.0        71.25
33       168.0        71.25
34       178.5        71.00
35       177.0        73.50
36       172.0        65.00
37       162.0        70.00
38       167.0        68.25
39       163.0        72.25
40       180.0        67.00
41       162.5        68.75
42       179.0        29.50
43       180.0        70.00
44       165.0        71.50
45       168.5        68.00
46       164.5        73.25
47       153.5        67.50
48       148.5        71.25
49       165.5        68.50
50       167.5        66.75
```

3. Construct a 2 x 2 contingency table for the categorical attributes Gender and Exercise.

```
> c1=table(data$GENDER, data$Exercise)
> c1

   N  Y
F 15 10
M 10 15
```

4. Four moments about the origin and mean (BMI)

```
> m1=moment(data$BMI, order = 1)
> m1
[1] 28.83746
>
> m2=moment(data$BMI, order = 2)
> m2
[1] 1239.356
>
> m3=moment(data$BMI, order = 3)
> m3
[1] 110445.4
>
> m4=moment(data$BMI, order = 4)
> m4
[1] 15632472
```

5. Skewness and kurtosis(BMI)

```
> sk=skewness(data$BMI)
> sk
[1] 6.216821
> kur=kurtosis(data$BMI)
> kur
[1] 42.11238
```

Exp.2: Computing Summary Statistics / plotting and visualizing data using Tabulation and Graphical Representations.

1. Create a data frame with the following descriptions

- Employee_info=data.frame(
- Emp_id=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15),
- Age=c(30,37,45,32,50,60,35,32,34,43,32,30,43,50,60),
- Sex=c('F','M','F','M','M','M','F','F','M','F','F','M','M','F','F'),
- Status=c(1,1,2,2,1,1,1,2,2,1,2,1,2,1,2)
-)

```
> Employee_info
   Emp_id Age Sex Status
1       1  30   F      1
2       2  37   M      1
3       3  45   F      2
4       4  32   M      2
5       5  50   M      1
6       6  60   M      1
7       7  35   F      1
8       8  32   F      2
9       9  34   M      2
10     10  43   F      1
11     11  32   F      2
12     12  30   M      1
13     13  43   M      2
14     14  50   F      1
15     15  60   F      2
```

2. Find the Summary statistics for male and female employees' data.

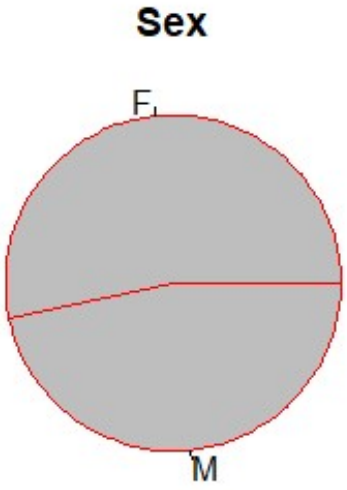
```
> maleinfo=subset(Employee_info, Employee_info$Sex=='M')
> summary(maleinfo)
     Emp_id           Age            Sex                Status
 Min.   : 2.000   Min.   :30.00   Length:7           Min.   :1.000
 1st Qu.: 4.500   1st Qu.:33.00   Class :character   1st Qu.:1.000
 Median : 6.000   Median :37.00   Mode  :character   Median :1.000
 Mean   : 7.286   Mean   :40.86                      Mean   :1.429
 3rd Qu.:10.500   3rd Qu.:46.50                      3rd Qu.:2.000
 Max.   :13.000   Max.   :60.00                      Max.   :2.000
> femaleinfo=subset(Employee_info, Employee_info$Sex=='F')
> summary(femaleinfo)
     Emp_id           Age            Sex                Status
 Min.   : 1.000   Min.   :30.00   Length:8           Min.   :1.0
 1st Qu.: 6.000   1st Qu.:32.00   Class :character   1st Qu.:1.0
 Median : 9.000   Median :39.00   Mode  :character   Median :1.5
 Mean   : 8.625   Mean   :40.88                      Mean   :1.5
 3rd Qu.:11.750   3rd Qu.:46.25                      3rd Qu.:2.0
 Max.   :15.000   Max.   :60.00                      Max.   :2.0
```

3. Draw a line graph for empid and age attributes.

```
plot(Employee_info$Emp_id, Employee_info$Age, type="o",
        main="Age of Employees",
         xlab="Employee's ID",
          ylab='Age', col="red")
```

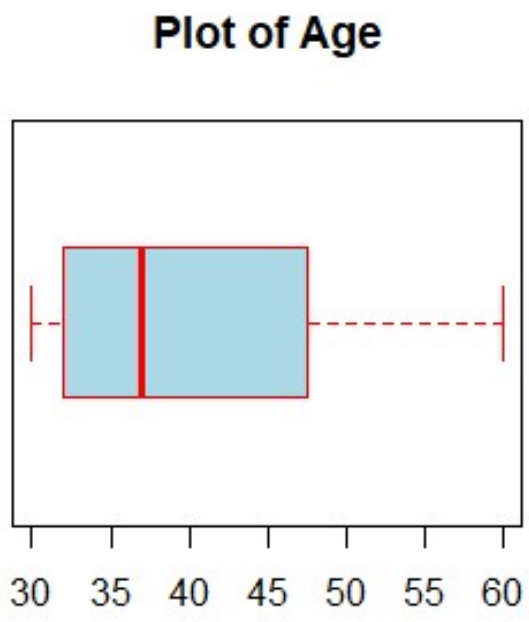
4. Draw a pie chart for sex attribute.

```
gr1=table(Employee_info$Sex)
pie(gr1, main = 'Sex', col="gray",
          border = 'red', radius=1.0)
```



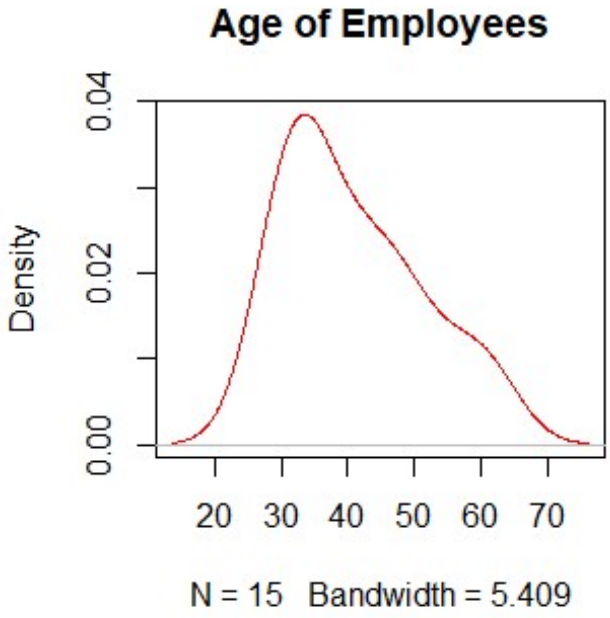
5. Draw a **box plot**, **histogram** and **density** for age attribute.

- boxplot(Employee_info$Age, main="Plot of Age",
- horizontal=T, col='light blue', border='red')



- hist(Employee_info$Age, main="Graph of Age", xlab='Age of Employees',
- col='sky blue', border='red')

- d=density(Employee_info$Age)
- plot(d, col='red')



Exp.3: Applying correlation and simple linear regression model to real dataset; computing and interpreting coefficient of determination.

1. The following data refers to the daily sales of tomatoes (in kg) at different prices (in Rupees) observed on different days in a market.

Price	4.5	5.5	4.5	4.5	4.0	5.5	5.5	6.5	5.0
Qty sold	125	115	140	140	150	150	130	120	130

Price	5.5	6.0	4.5
Qty sold	100	105	150

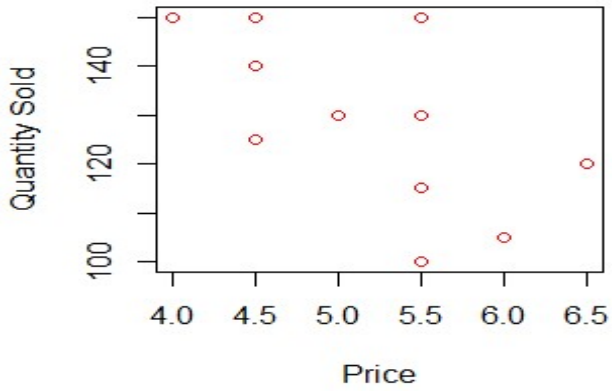
Analyse the data and fit a linear model.

- obj=data.frame(Price=c(4.5,5.5,4.5,4.5,4.0,5.5,5.5,6.5,5.0,5.5,6.0,4.5),
- Qty_Sold=c(125,115,140,140,150,150,130,120,130,100,105,150)
-)

```
> model=lm(Price~Qty_Sold, data=obj)
> model

Call:
lm(formula = Price ~ Qty_Sold, data = obj)

Coefficients:
(Intercept)     Qty_Sold
    8.66359     -0.02731

> summary(model)$r.squared
[1] 0.4025737
```

2. Draw the scatter diagram for price and Qty sold.

- plot(obj$Price, obj$Qty_Sold, xlab="Price",
- ylab="Quantity Sold", col='red')



3. Find the coefficient of correlation for price and Qty Sold (Karl Pearson, Spearman and Kendal)

```
> Karl_Pearson_cor=cor(obj$Price, obj$Qty_Sold, method='pearson')
> Karl_Pearson_cor
[1] -0.634487
> Spearman_cor=cor(obj$Price, obj$Qty_Sold, method='spearman')
> Spearman_cor
[1] -0.650397
> #By Default
> Kendall_cor=cor(obj$Price, obj$Qty_Sold)
> Kendall_cor
[1] -0.634487
```

4. Calculate coefficient of determination and interpret it

```
> #Coefficient of determination
> #saving the linear regression model in a new variable model
> model=lm(Price~Qty_Sold, data=obj)
> model

Call:
lm(formula = Price ~ Qty_Sold, data = obj)

Coefficients:
(Intercept)     Qty_Sold
    8.66359     -0.02731

> #extracting the coefficient of determination from the r.squared attribute of its summary
> summary(model)$r.squared
[1] 0.4025737
```

4: Applying multiple linear regression model to real dataset, computing and interpreting the multiple coefficient of determination

Import admission data:

- data4=read.csv("C:/Users/ilyas/Desktop/R/DA-1/admission.csv")
- data4

1. Construct a multiple linear regression model for the admitted attributes on GREScore, TOFELScore and CGPA.

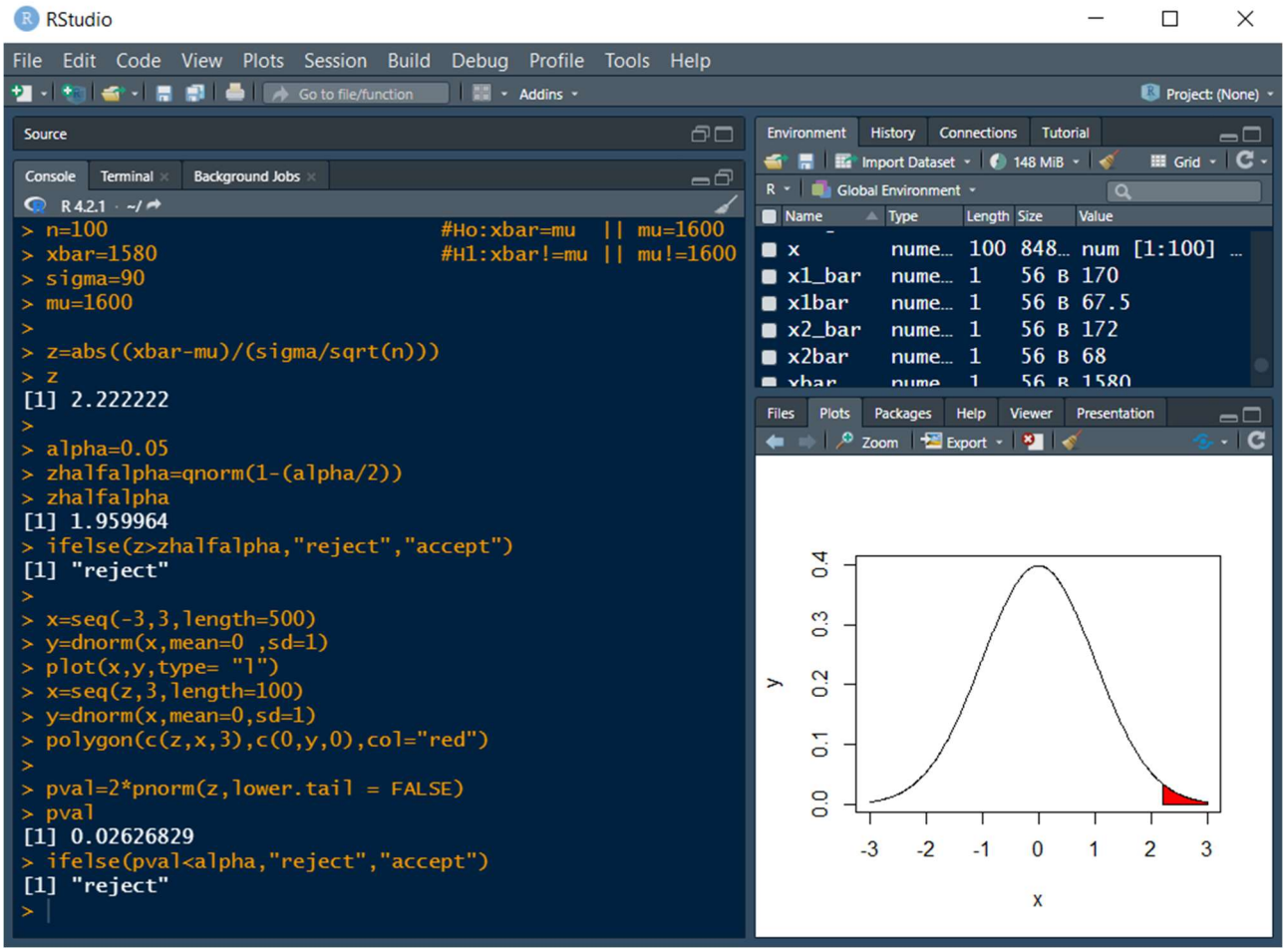
```
> data4=read.csv("C:/Users/ilyas/Desktop/R/DA-1/admission.csv")
> #1
> model4=lm(Admitted~GREScore+TOEFLScore+CGPA, data=data4)
> model4

Call:
lm(formula = Admitted ~ GREScore + TOEFLScore + CGPA, data = data4)

Coefficients:
(Intercept)     GREScore   TOEFLScore         CGPA
  -3.453797     0.002585     0.010339     0.192334
```

2. Find the summary of the model and interpret.

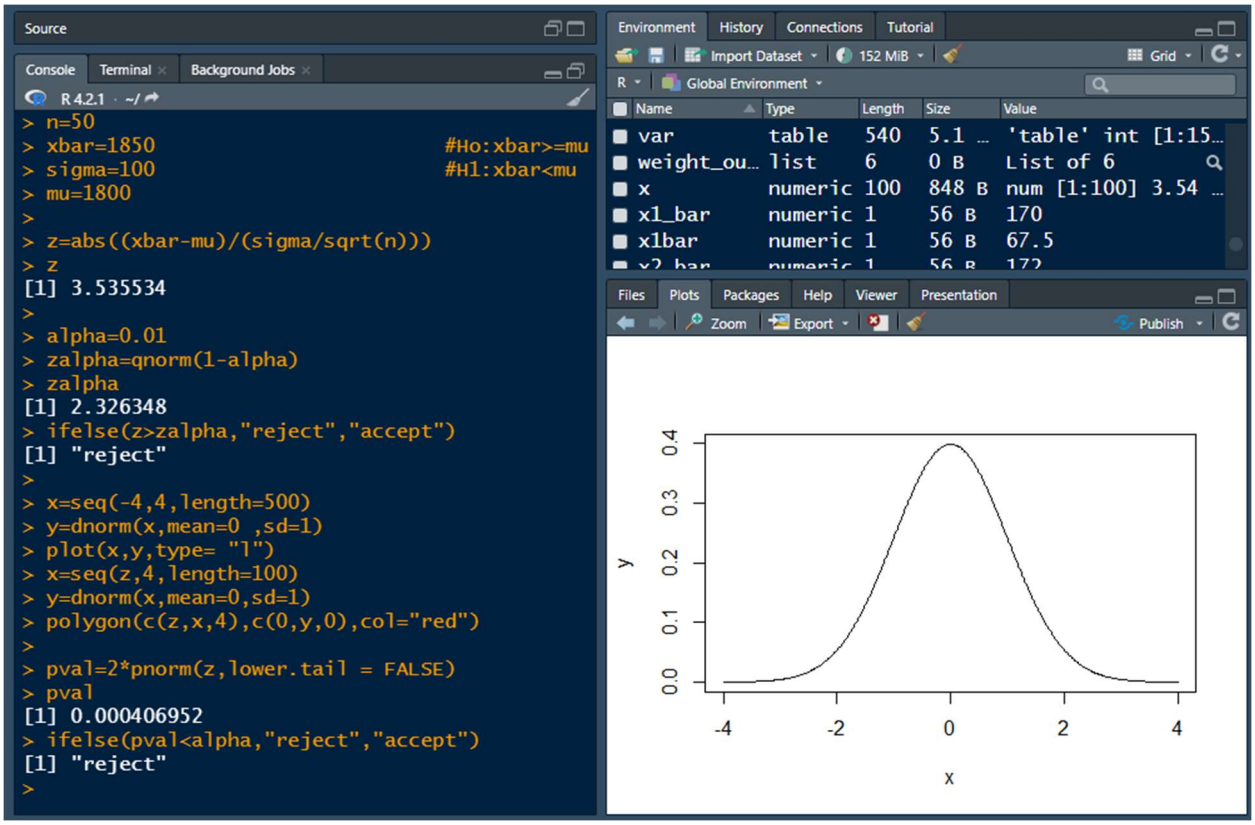
```
> summary(model4)$r.squared
[1] 0.3619033
> summary(model4)

Call:
lm(formula = Admitted ~ GREScore + TOEFLScore + CGPA, data = data4)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51841 -0.19172 -0.04612  0.10572  0.71607

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.453797   0.381320  -9.057  < 2e-16 ***
GREScore     0.002585   0.002100   1.231  0.21906
TOEFLScore   0.010339   0.003755   2.753  0.00612 **
CGPA         0.192334   0.037634   5.111 4.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2625 on 496 degrees of freedom
Multiple R-squared:  0.3619,    Adjusted R-squared:  0.358
F-statistic: 93.77 on 3 and 496 DF,  p-value: < 2.2e-16
```

Exp.5: Testing of hypothesis for one sample mean and proportion from real-time problems
Using R, perform the testing of hypothesis and interpret your results for the following scenarios:

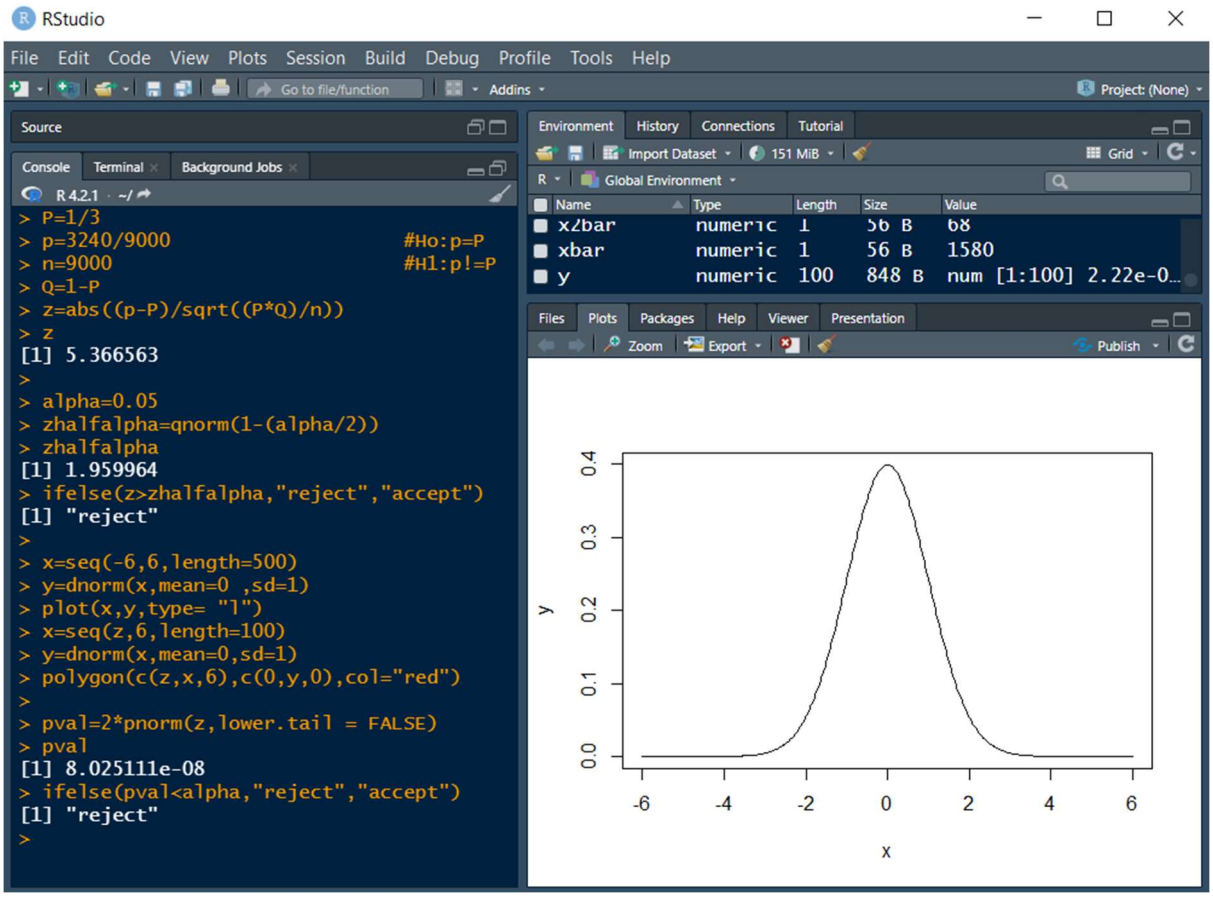
1. The mean lifetime of a sample of 100 light tubes produced by a company is found to be 1580 hours with S.D. of 90 hours. Test the hypothesis that the mean lifetime of the tubes produced by the company is 1600 hours.



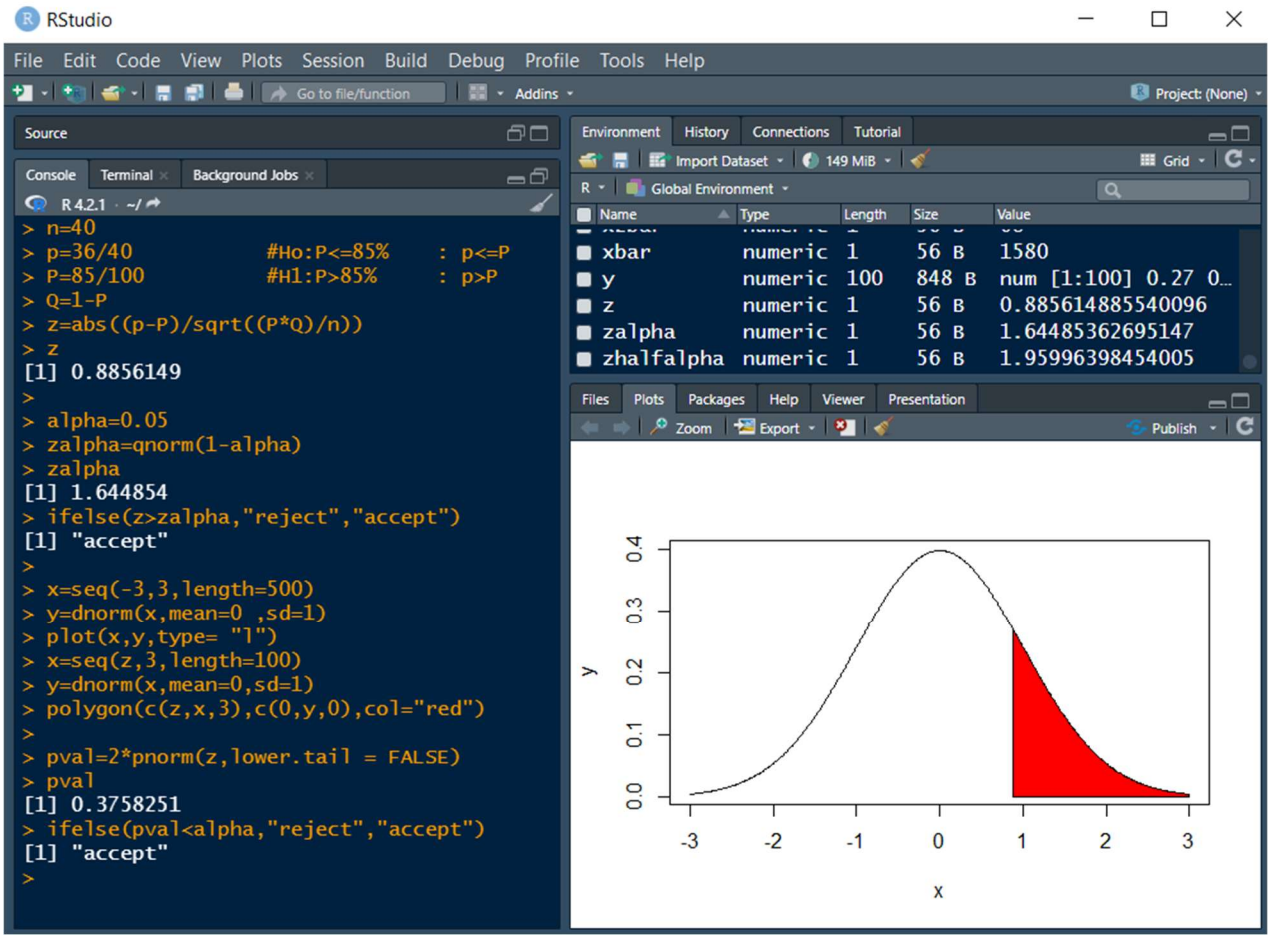
2. The mean breaking strength of the cables supplied by a manufacturer is 1800 with a S.D of 100. By a new technique in the manufacturing process, it is claimed that the breaking strength of the cable has increased. To test this claim a sample of 50 cables is tested and is found that the mean breaking strength is 1850. Can we support the claim at 1% L.O.S.?



3. A die is thrown 9000 times and throw of 3 or 4 is observed 3240 times. Show that the die cannot be regarded as an unbiased one using 5% L.O.S.

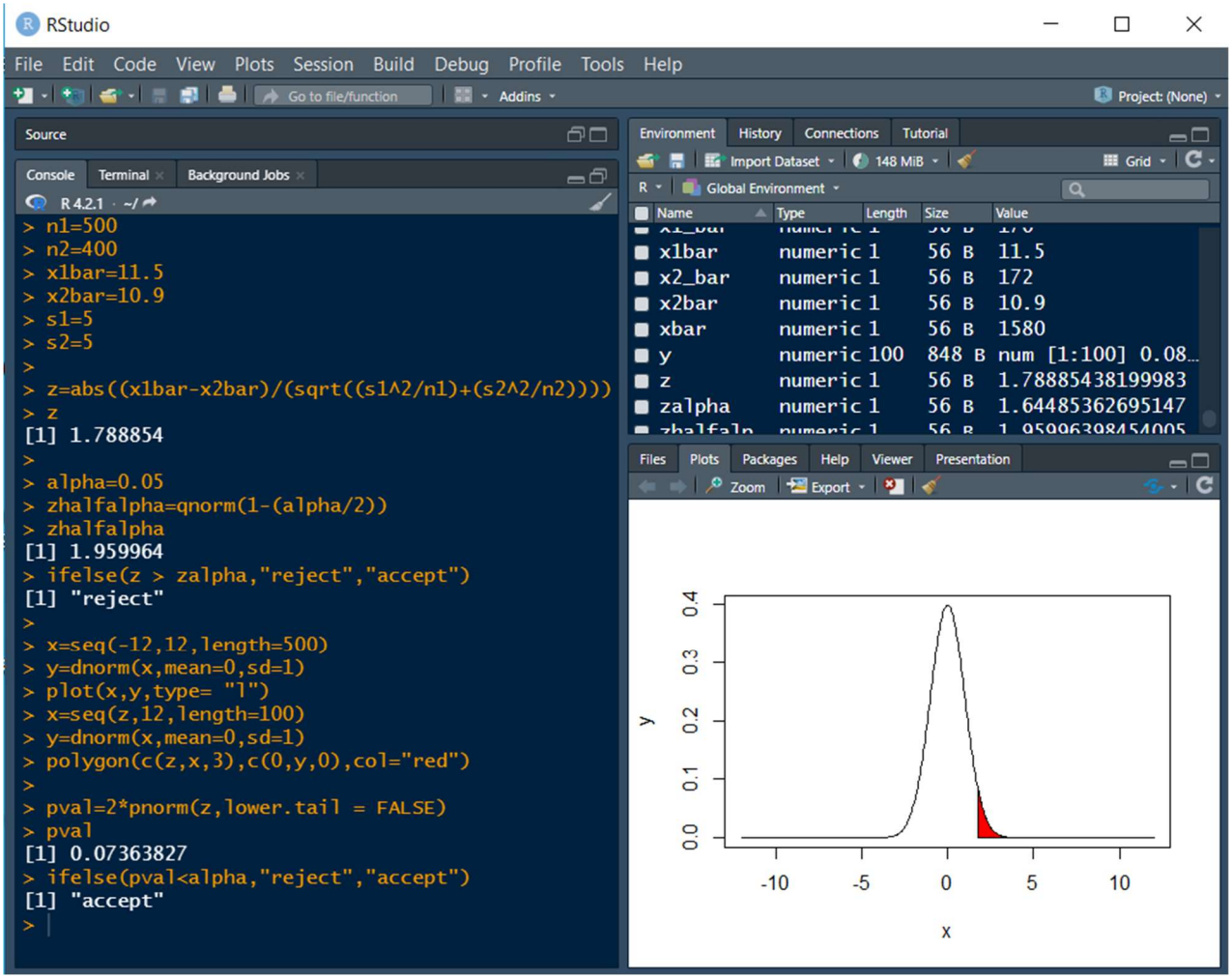


4. 40 people were attacked by a disease and only 36 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more, at 5% L.O.S.?

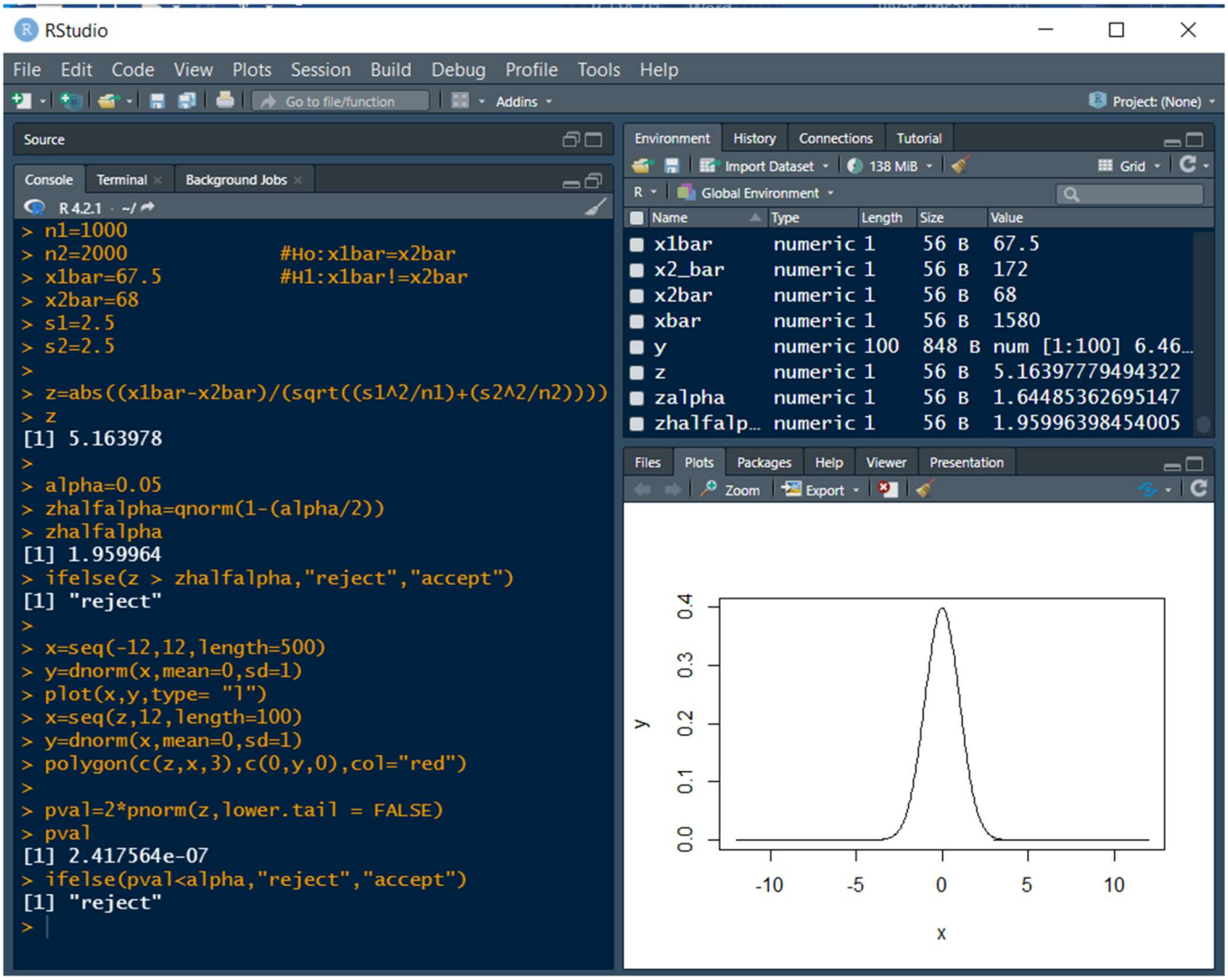


Exp.6: Testing of hypothesis for two sample mean and proportion from real-time problems Using R, perform the testing of hypothesis and interpret your results for the following scenarios:

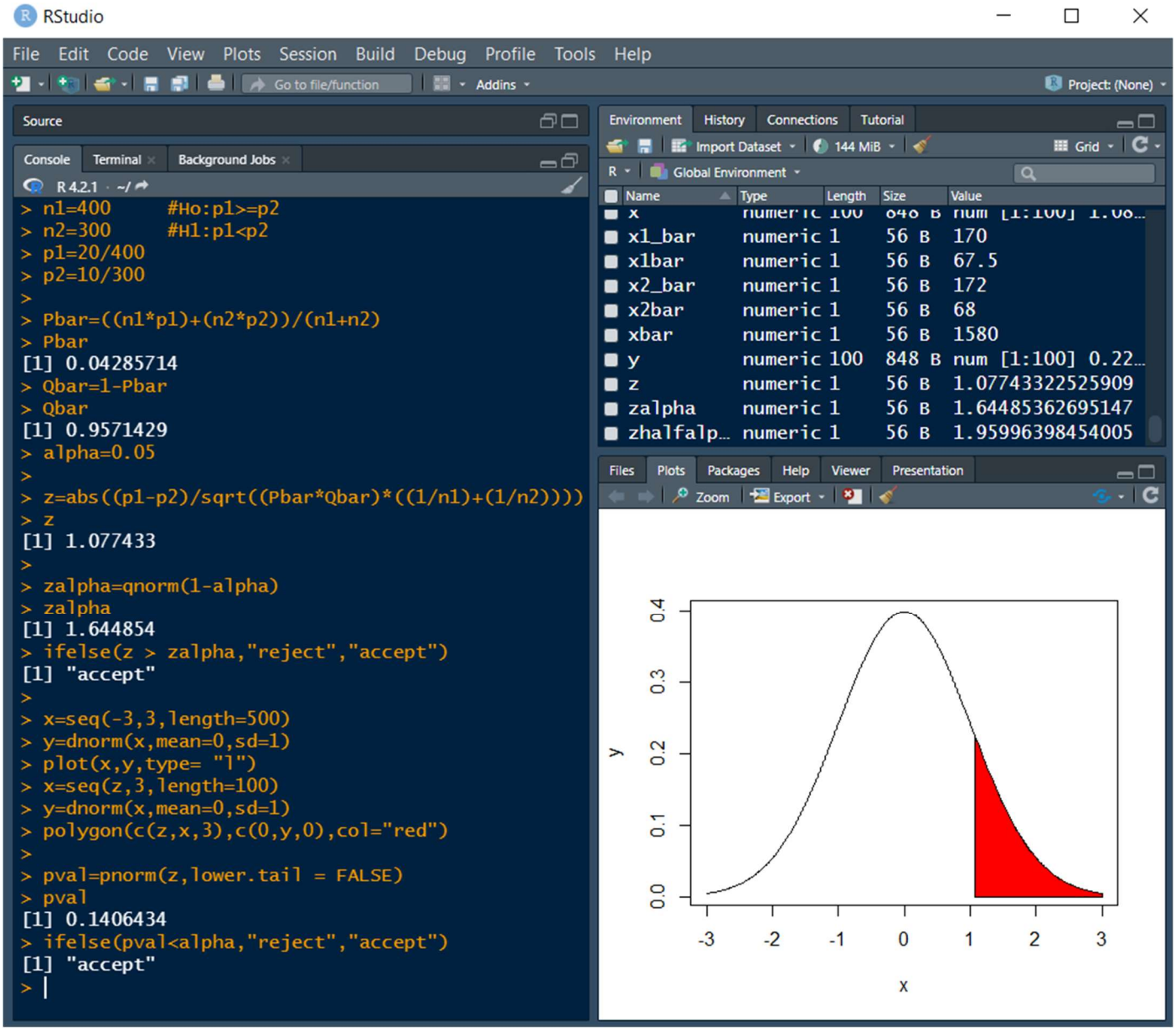
1. Two random samples of sizes 400 and 500 have mean 10.9 and 11.5 respectively. Can the samples be regarded as drawn from the same population with variance 25 at 5% L.O.S?



2. The means of two samples of 1000 and 2000 members are respectively 67.5 and 68 inches. Can they be regarded as drawn from the same population with S.D. 2.5 inches?



3. A machine produced 20 defectives articles in a batch of 400. After overhauling it produced 10 defectives in a batch of 300. Has the machine improved? Use 5% L.O.S.



4. Random samples of 400 men and 600 women were asked whether they would like to have a fly-over near their residence. 200 men and 325 women were in favour of it. Test the equality of proportion of men and women in the proposal? Use 5% L.O.S.