

Mahatma Education Society’s

**Pillai College of Arts, Commerce & Science**

# (Autonomous)

**Affiliated to University of Mumbai**

NAAC Accredited 'A' grade (3 cycles) Best College Award by University of Mumbai

ISO 9001:2015 Certified

**CERTIFICATE**

*This is to certify that Mr.*  ***Rohit Malviya*** *of* ***M.Sc. D.A. Part-I Semester I*** *has completed the practical work in the Subject of* ***Research Methodology*** *during the academic year 2024-25 under the guidance of Prof****.*** ***Soly Zachariah*** *being the partial requirement for the fulfillment of the curriculum of* ***Degree of Master of Science in Data Analytics****,* ***University of Mumbai****.*

***Place:***

***Date:***

*Name & Signature of faculty Name & Signature of external*

Dr. K.M. Vasudevan Pillai Campus, Sector 16, New Panvel - 410206. Tel : 27456100 / 1700 | Fax : 27483208 | Website : www.pcacs.ac.in

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| **Practical No.** | **Detail** | **Date** | **Sign** |
| 1. | Using R execute the basic commands. | 23-07-2024 |  |
| 2. | Import the data from Excel / .CSV find mean median mode, standard deviation variance. | 29-07-2024 |  |
| 3. | Perform R program for making Diagrams. (Bar Diagram, Multiple Bar Diagram, Pie Chart) | 01-08-2024 |  |
| 4. | Perform R program for making Graphs. (Histogram, Frequency Polygon, Ogive) | 09-08-2024 |  |
| 5. | Import the data from Excel / .CSV and perform the Chi-square Test,goodness of fit, Independence of attributes. | 16-08-2024 |  |
| 6. | Perform an R program on z-test- one population mean, Two population means.One population proportion, two population proportion. | 30-08-2024 |  |
| 7. | Perform an R program on t test- one sample, paired and unpaired. | 13-09-2024 |  |
| 8. | Perform an R program on Non Parametric Test -Sign test, wilcoxon signed rank test. | 13-09-2024 |  |
| 9. | Perform an R program on One way ANOVA and Two way ANOVA. | 25-09-2024 |  |
| 10. | Perform an R program on Friedman Test and Kruskal Wallis test. | 21-09-2024 |  |

**PRACTICAL 1**

**Aim:** Using R execute the basic commands

**Q1) Assign the data in R .**

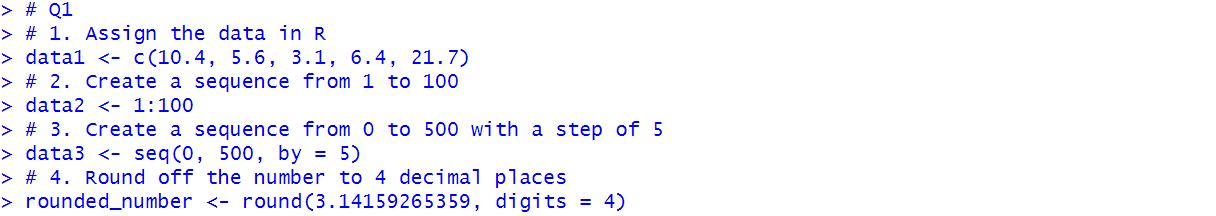
1. 10.4, 5.6, 3.1,6.4 ,21.7

2. 1 to 100

3. 0,5,10,......500

4. Round off the no 3.14159265359 into 4 decimal places.

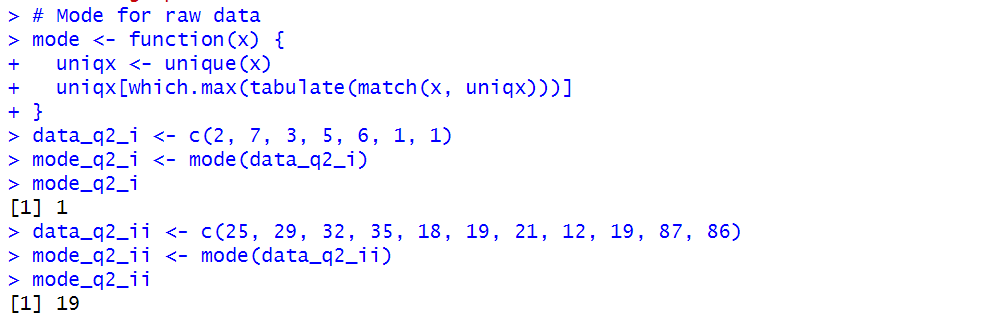
Code:



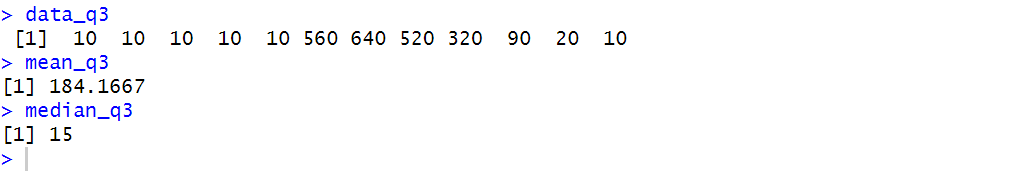
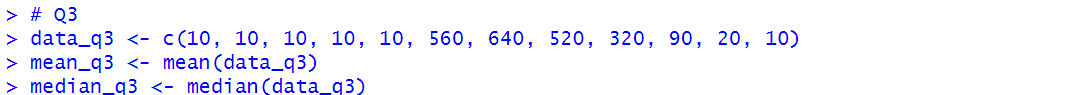
**Q2)Mode for raw data**

(i) 2,7,3,5,6,1,1

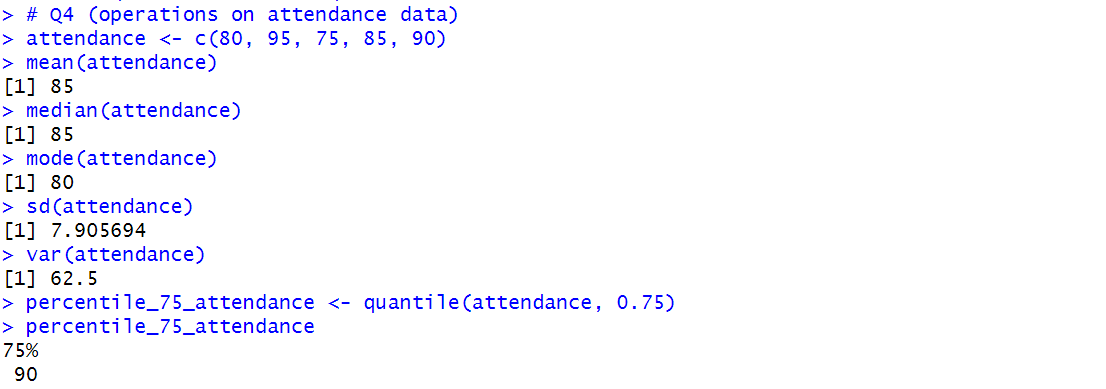
(ii)25,29,32,35,18,19,21,12,19,87,86



**Q3)** Find mean,and median of the data 10 10 10 10 10 560 640 520 320 90 20 10



**Q4)** Consider your attendance and find its mean, median , mode , standard deviations, variance , and 75th percentile .

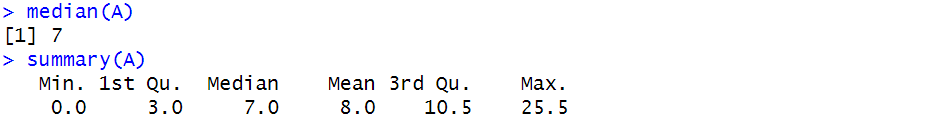
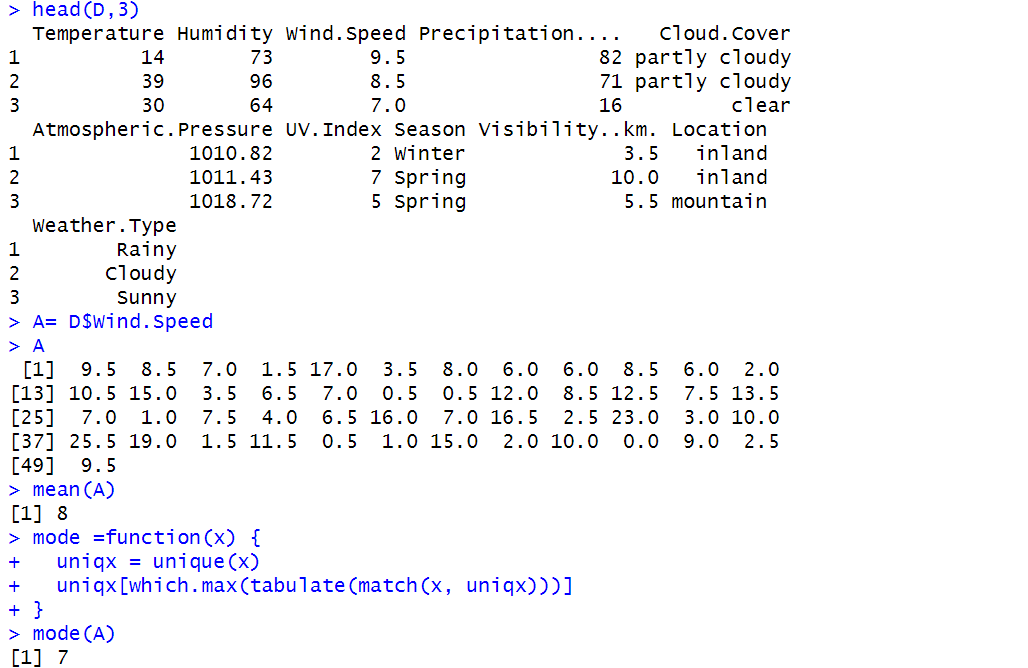


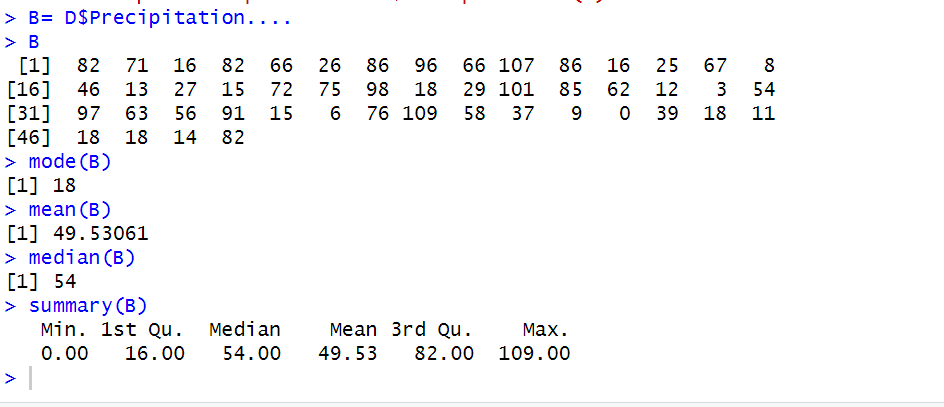
**PRACTICAL 2**

**Aim:** Import the data from Excel / .CSV find mean median mode,standard deviation variance.

Dataset: [Data set- Temp - Sheet1.csv](https://github.com/graveoffirefly/MMRR/blob/main/Data%20set-%20Temp%20-%20Sheet1.csv)

**mean median mode,standard deviation variance of the column “wind.speed”**





**PRACTICAL 3**

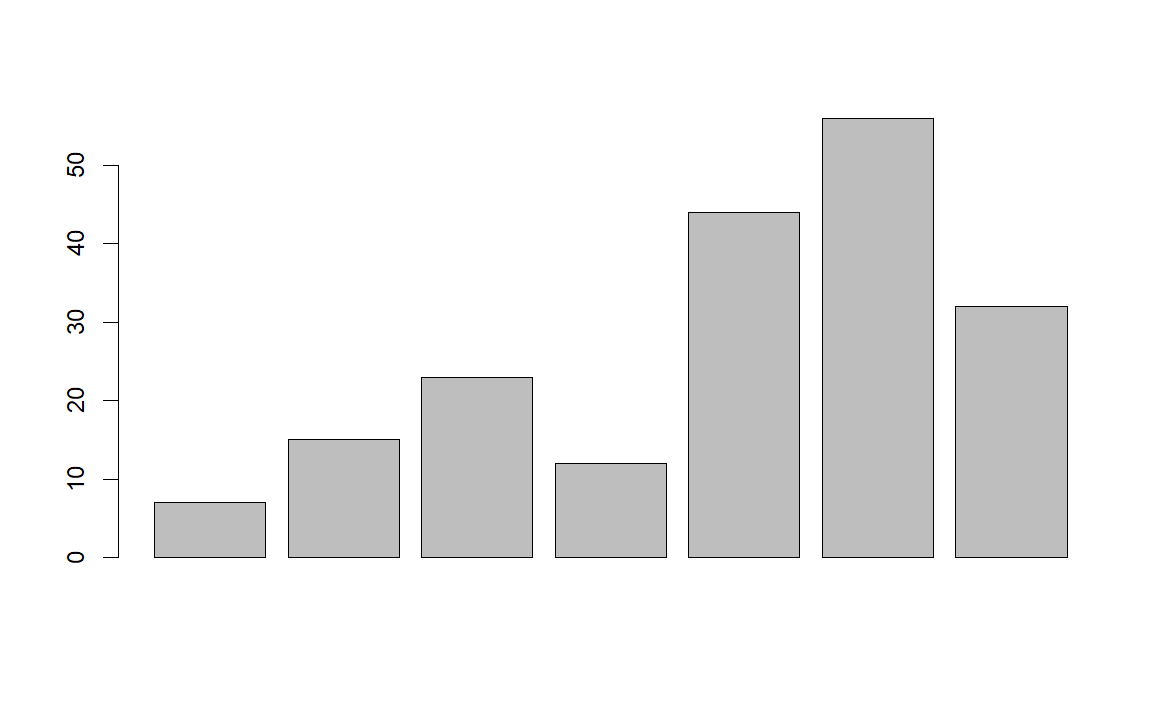
**Aim:** Perform R program for making Diagrams(Bar Diagram, Multiple Bar Diagram, Pie Chart)

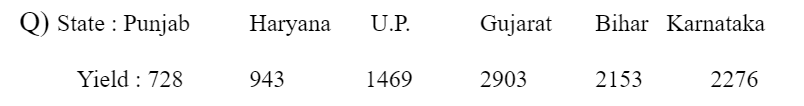
Bar Diagram :

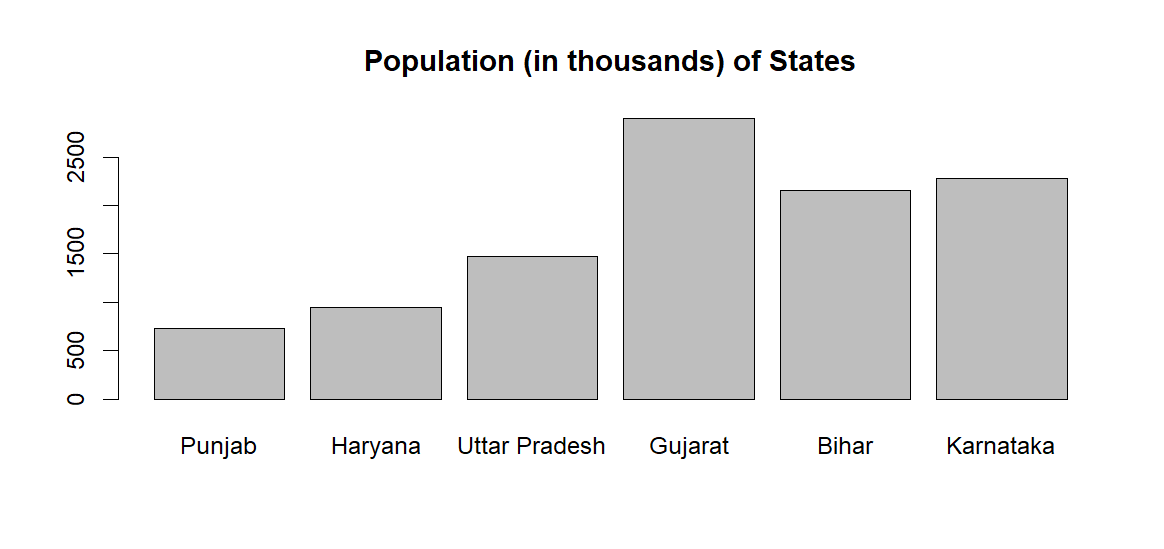
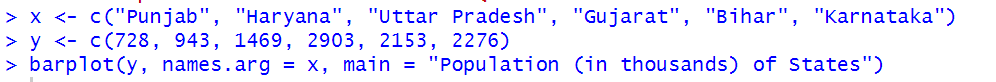
barplot(x)

barplot(y, names.arg = x)

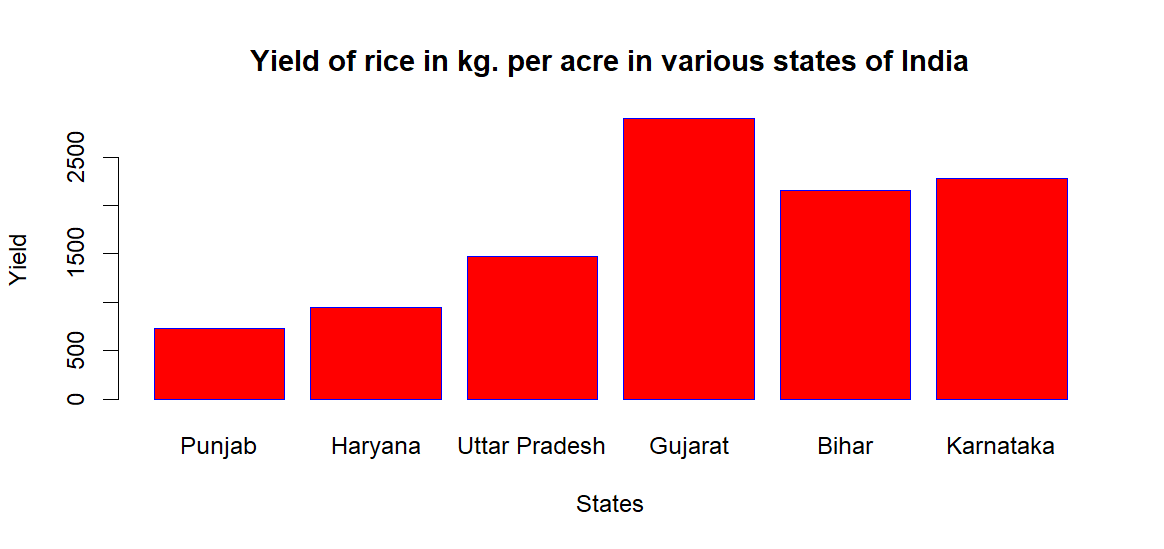
****

****

****

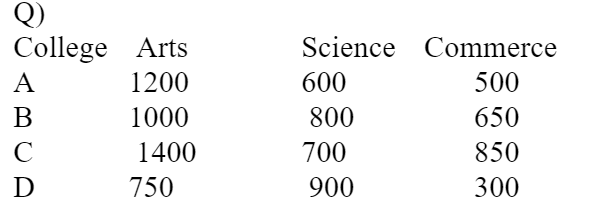


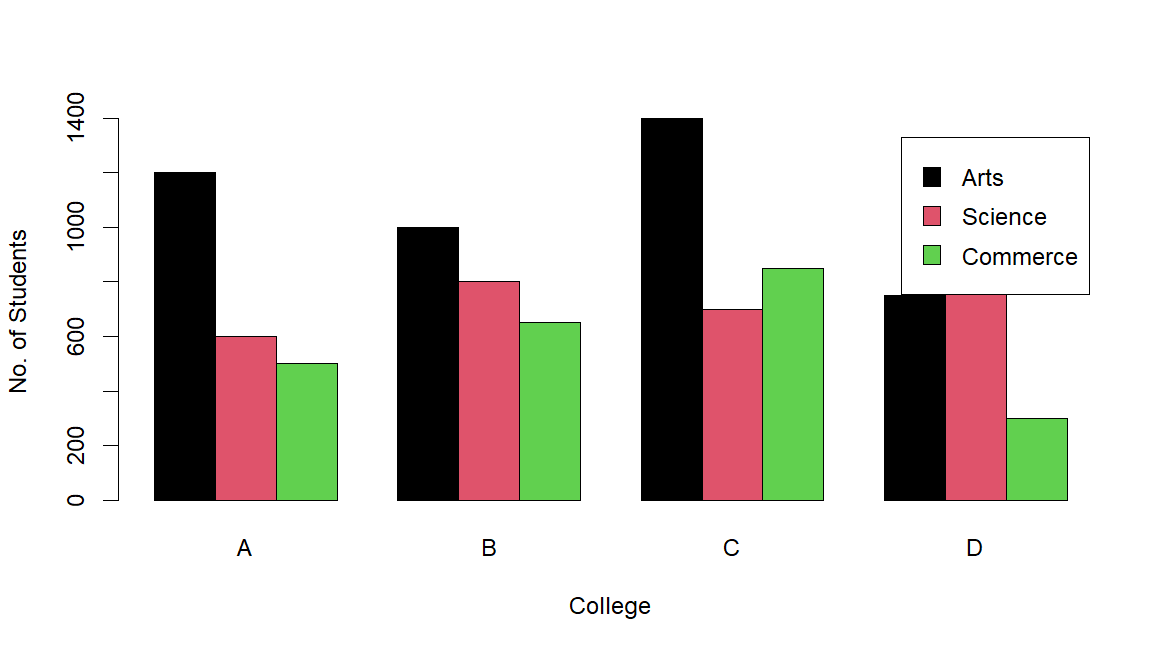
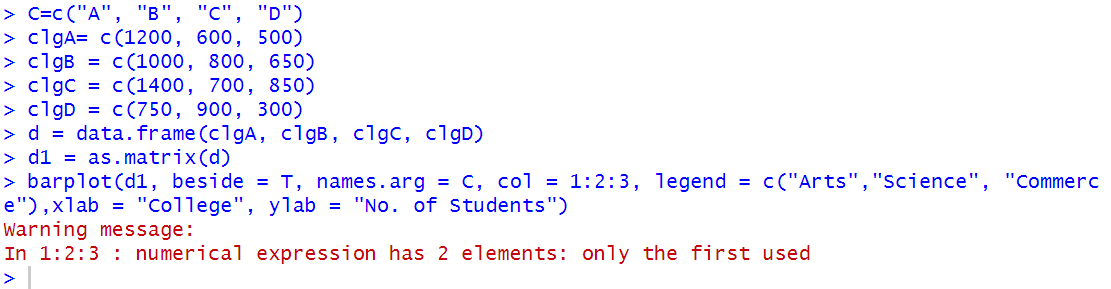


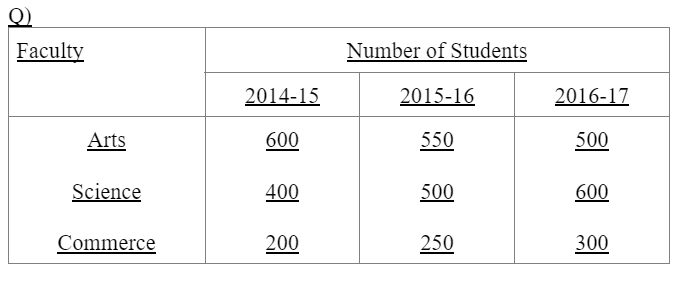


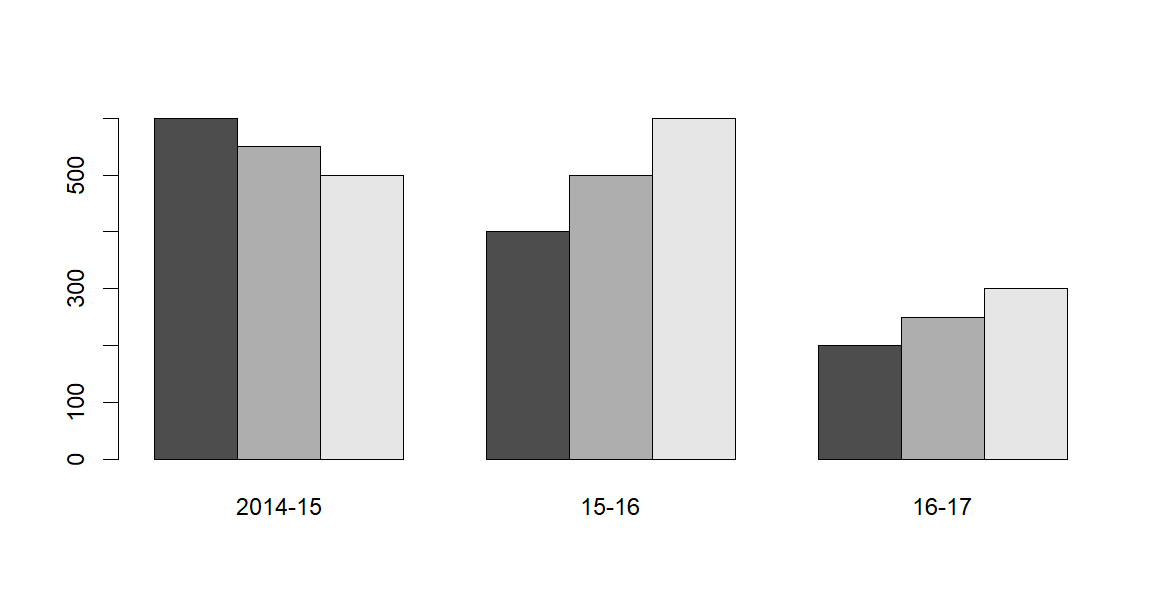
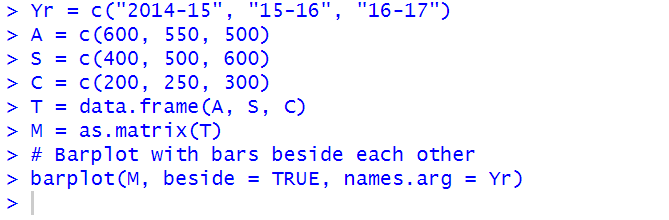
**Multiple bar diagrams syntax:**

barplot(...., beside = T, names.arg = …., col = 1:2:3, legend = c("...",".....", ….),xlab = "....", ylab = ".....")

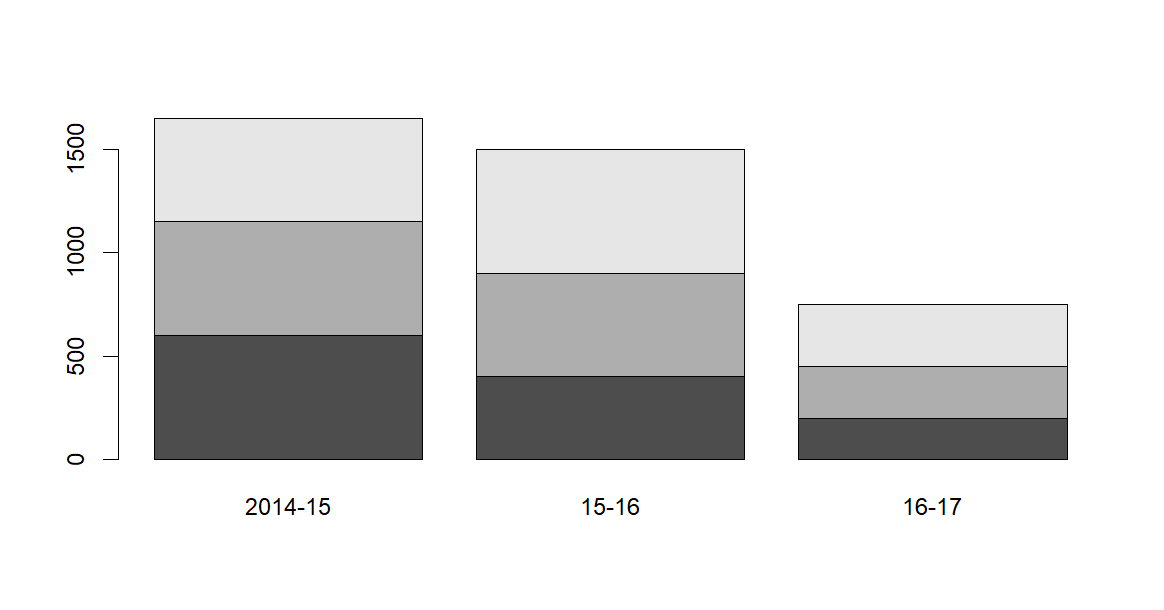




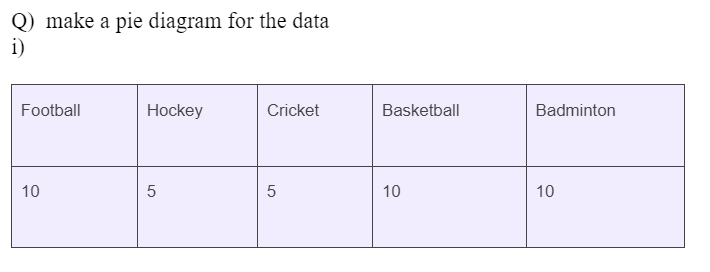


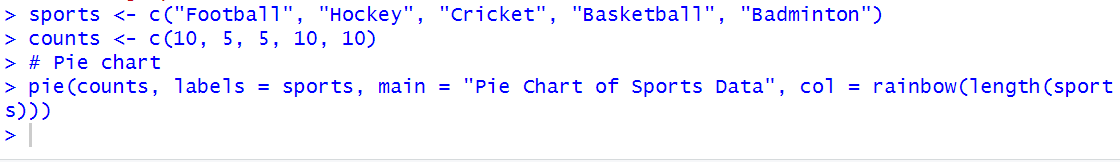


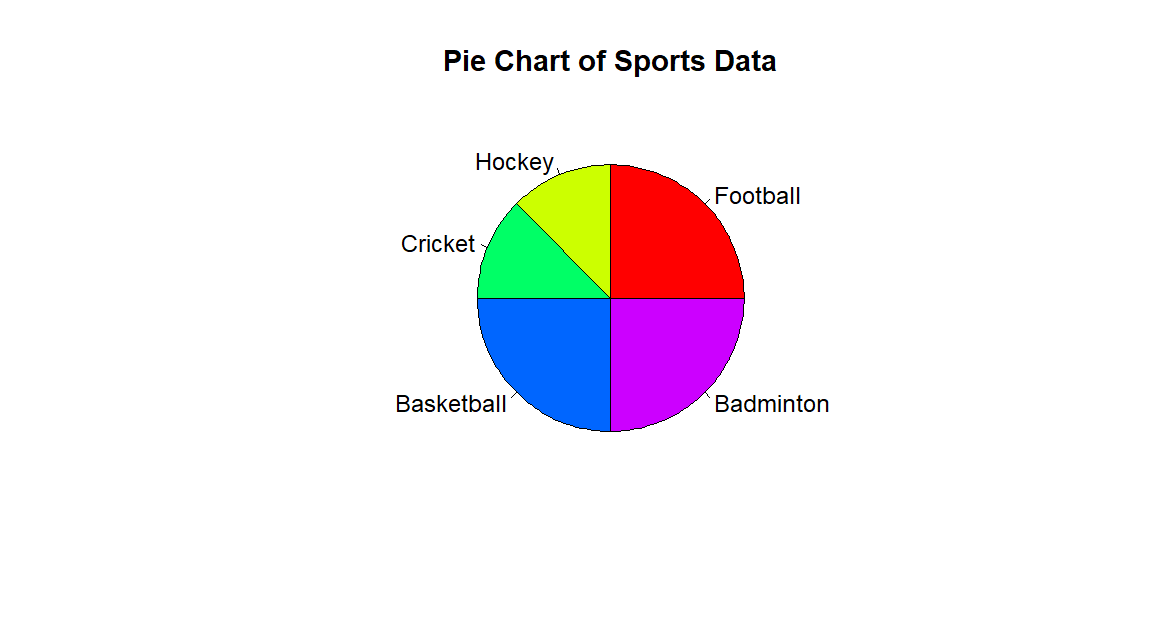


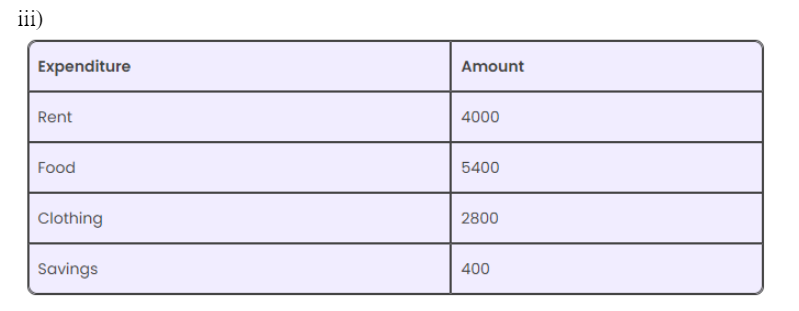


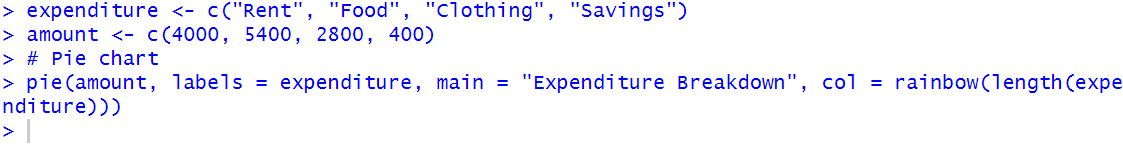
**PIE DIAGRAM**



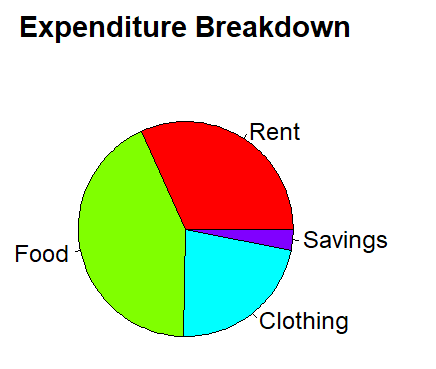




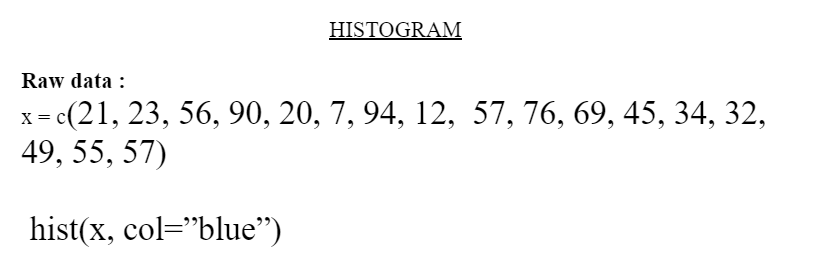


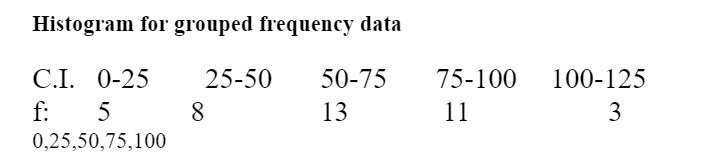


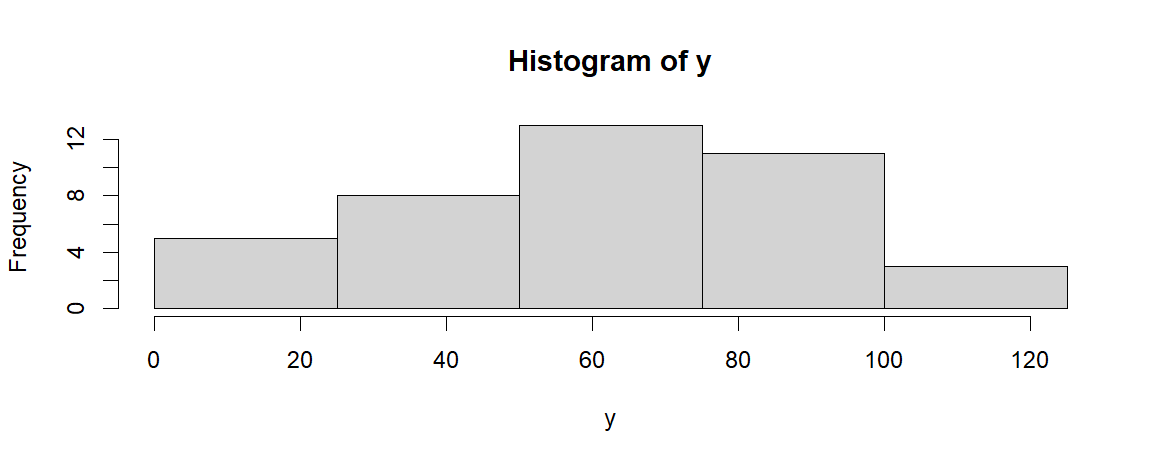
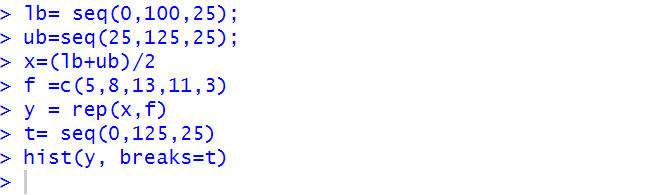
Output:



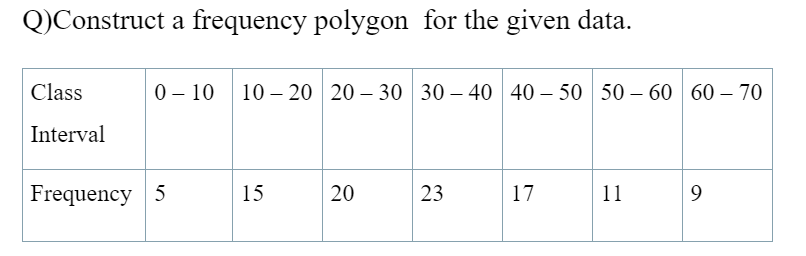
**PRACTICAL 4**

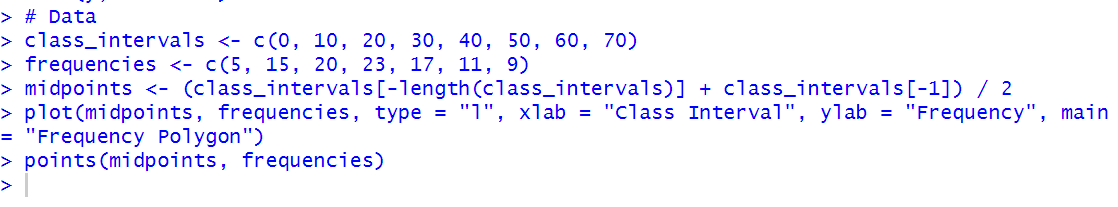


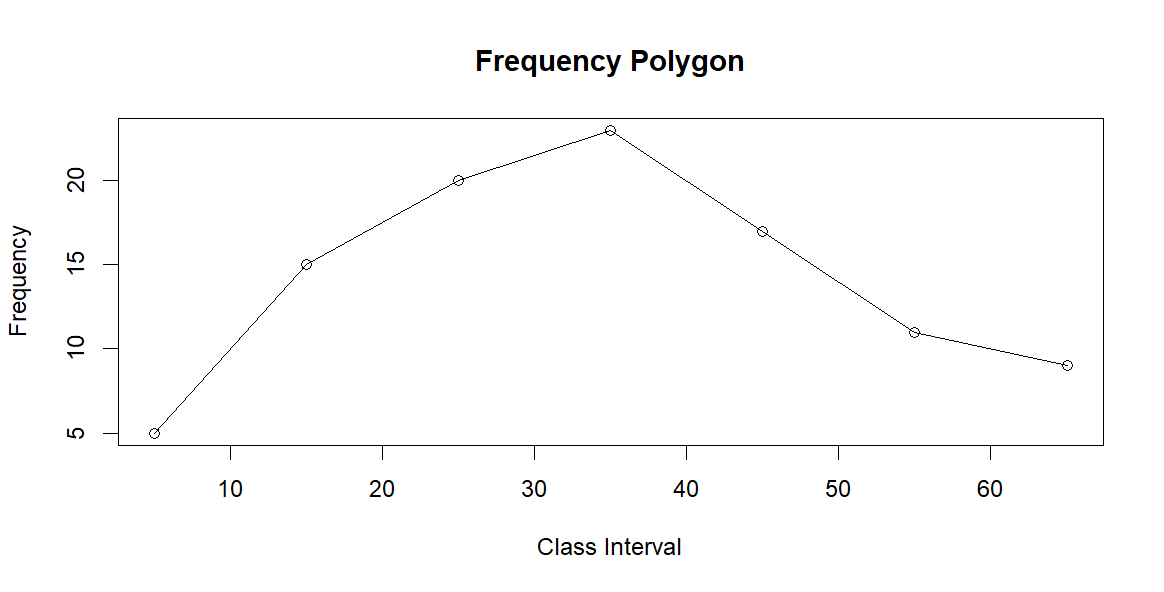




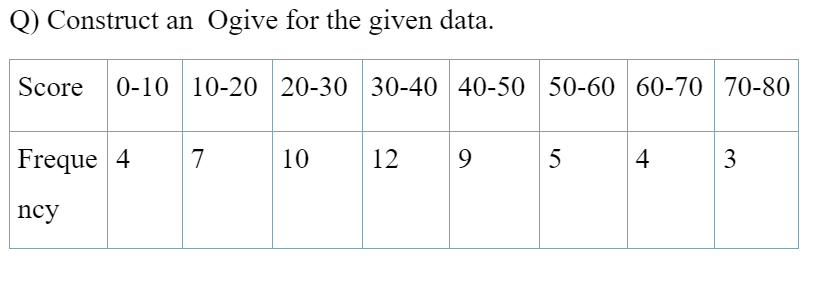
Frequency polygon

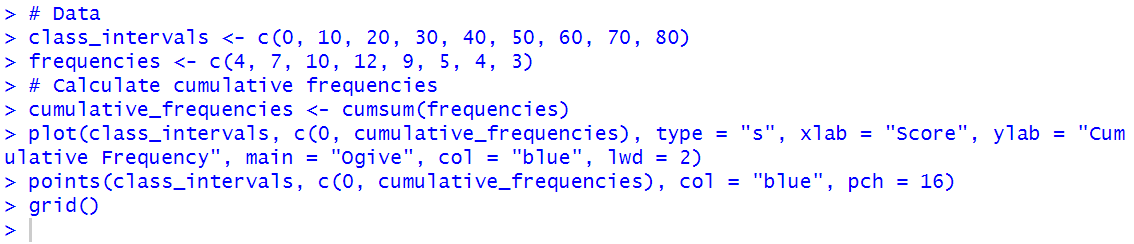


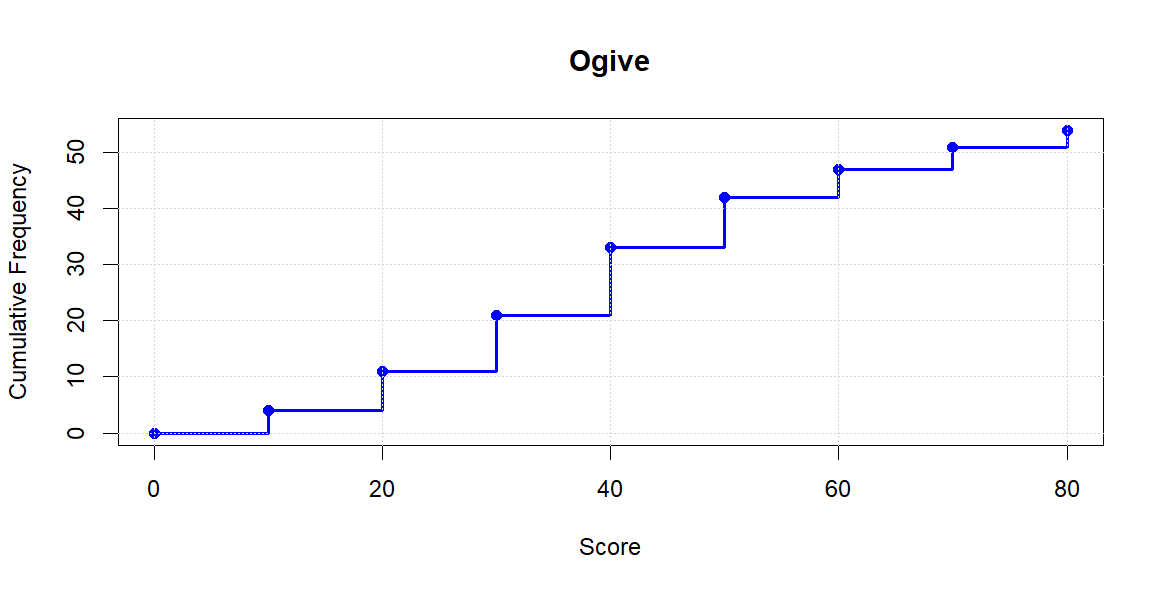




Ogives







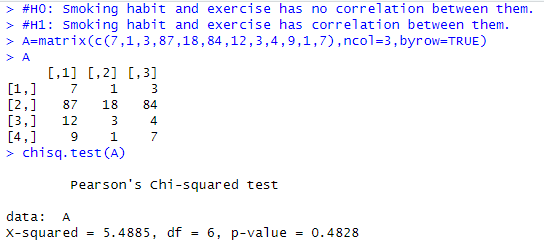
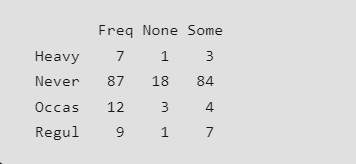
**PRACTICAL 5**

Aim : Import the data from Excel / .CSV and perform the Chi-square Test,goodness of fit

,Independence of attributes.

1. Chi-Square Test For Independence Of Attribute

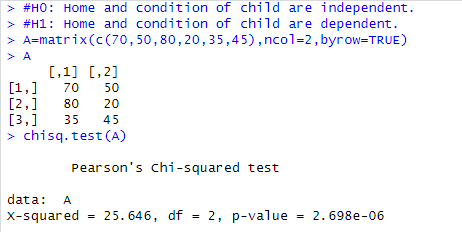
Q) Examine the relation between smoking habit and exercise at 5 % loss.



Conclusion: P-value > 0.05 .Hence accept H0:Smoking habit and exercise has no correlation between them.

Q) Use the following data to test whether the attributes-condition at home (A)and condition of child(B)are independent .

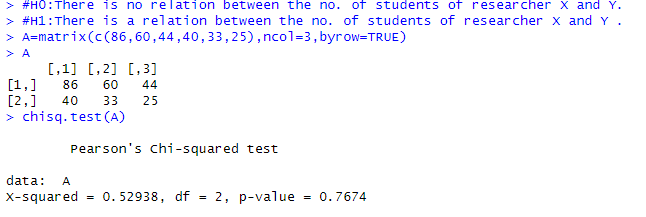
|  |  |  |
| --- | --- | --- |
|  | Clean | Dirty |
| Clean | 70 | 50 |
| Fairly clean | 80 | 20 |
| Dirty | 35 | 45 |



Conclusion: P-value < 0.05 , hence accept H1: Home and condition of child are dependent.

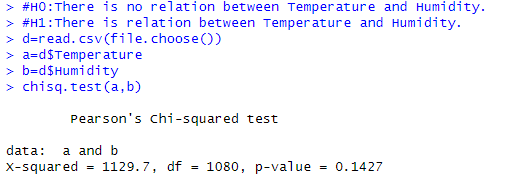
Q) Two Researchers Adopted Different Sampling Techniques While Investigating The Same Group Of Students To Find The Number Of Students Falling In Different Intelligence Level.The Results Are As Follows.

|  |  |  |  |
| --- | --- | --- | --- |
| Researcher | Intelligence Level | | |
| Below Avg | Avg | Above Avg |
| X | 86 | 60 | 44 |
| Y | 40 | 33 | 25 |



Conclusion: P-value > 0.05 ,Hence accept H0:There is no relation between the no. of students of researcher X and Y.

Q) Export the data and examine the dependency of temperature and humidity.



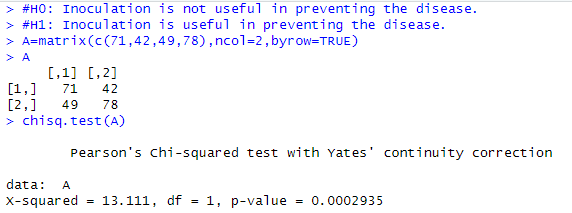
Conclusion :P-value > 0.05, hence accept H0:There is no relation between Temperature and Humidity.

Q) The Following table shows in occultation against cholera .Test inoculation is useful in preventing the disease

Attacked Not Attacked

Inoculated 71 42

Not Inoculated 49 78



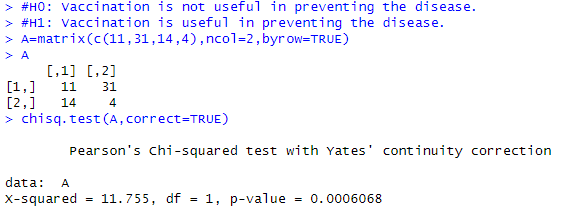
Conclusion: P-value < 0.05 , hence accept H1: Inoculation is useful in preventing the disease.

Q)The Following table shows inoculation against cholera .Test Vaccination is useful in preventing the disease

Attacked Not Attacked

Vaccinated 11 31

Not Vaccinated 14 4



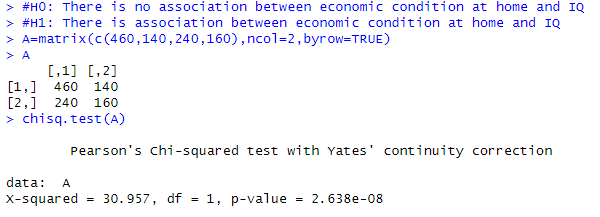
Conclusion : P-value < 0.05 , hence accept H1: Vaccination is useful in preventing the disease.

Q. Thousands of students at college level were gathered according to their I.Q and the economic condition of their home. Use chi square TEST to find out whether there is any association between economic condition at home and IQ

Economic Condition/Iq High Low

Rich 460 140

Poor 240 160



Conclusion: P-value < 0.05 , hence accept H1: There is association between economic condition at home and IQ.

Q. A random sample of 600 students from 12 th ,FY,SY,TY classes were asked the opinion about whether they like Autonomous college or oppose Autonomous college .On the basis of the following data can we say that class of student and opinion are independent

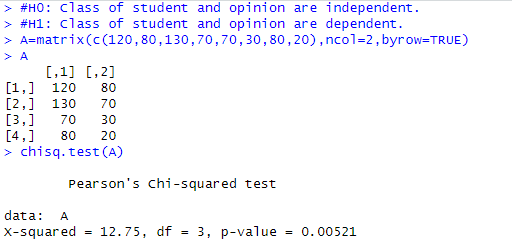
Class Favored Opposed

XII 120 80

FY 130 70

SY 70 30

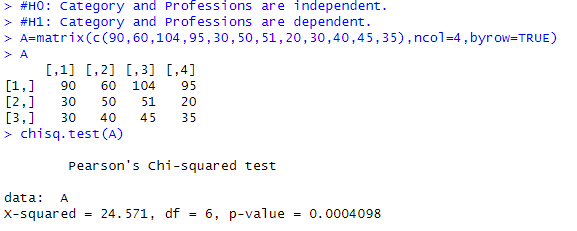
TY 80 20



Conclusion: P-value < 0.05 , hence accept H1: Class of student and opinion are dependent.

Q) Are categories and professions independent or not . comment on the given data

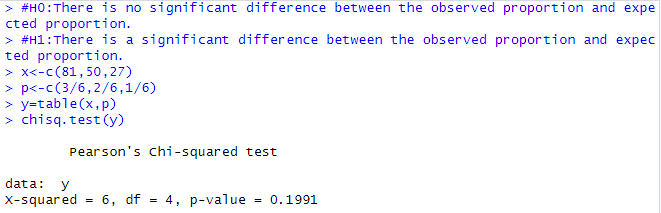
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Categories | P | Q | R | S |
| Doctors | 90 | 60 | 104 | 95 |
| Engineers | 30 | 50 | 51 | 20 |
| Teachers | 30 | 40 | 45 | 35 |



Conclusion : P-value < 0.05 , hence accept H1: Category and Professions are dependent.

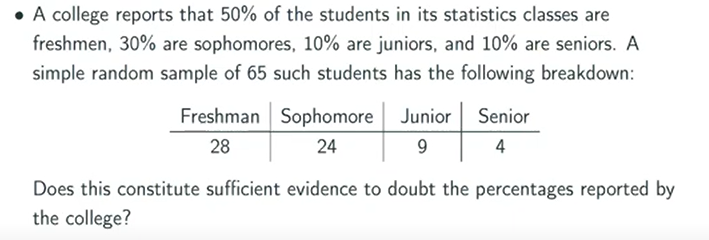
2. Chi Square For Goodness Of Fit

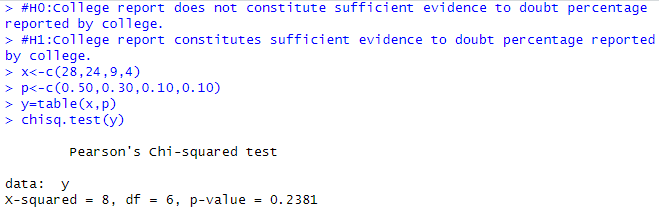
Q)We collected wild tulips and found that 81 were red, 50 were yellow and 27 were white. The ratio of red, yellow and white tulips is 3:2:1. Is there any significant difference between the observed proportions and the expected proportions?



Conclusion : P-value>0.05 ,Hence accept #H0:There is no significant difference between the observed proportion and expected proportion.

Q)





Conclusion : P-value>0.05 ,Hence accept #H0:College report does not constitute sufficient evidence to doubt percentage reported by college.

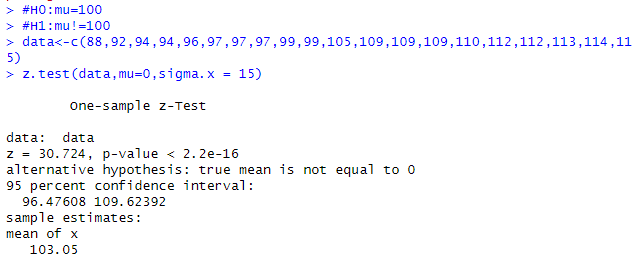
**PRACTICAL 6**

Aim : Perform an R program on z-test- one population mean, Two population means. One population proportion, two population proportion.

1. One population mean

Q) Suppose the IQ in a certain population is normally distributed with a mean of μ = 100 and standard deviation of σ = 15. A scientist wants to know if a new medication affects IQ levels, so she recruits 20 patients to use it for one month and records their IQ levels at the end of the month. Determine if the new medication causes a significant difference in IQ levels:

88,92,94,94,96,97,97,97,99,99,105,109,109,109,110,112,112,113,114,115

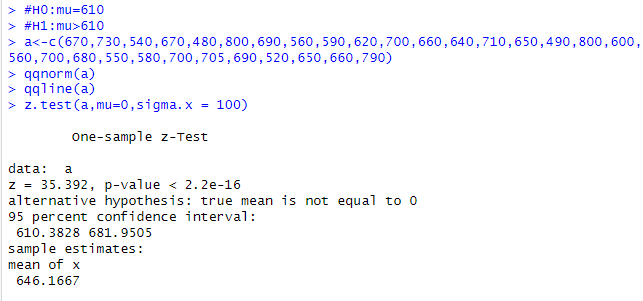


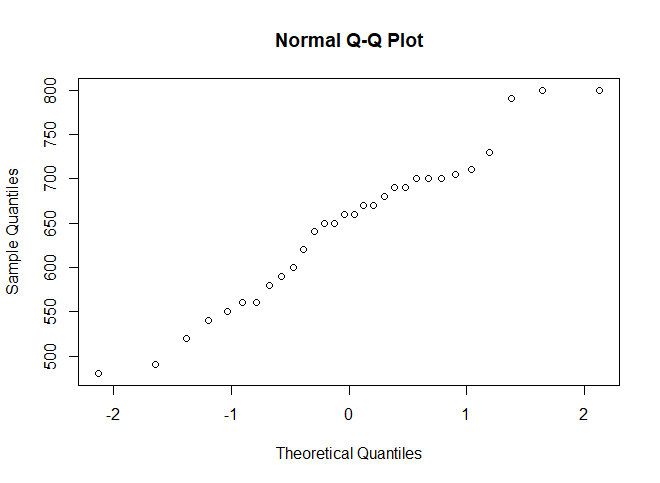
Conclusion : P-value<0.05 ,Hence accept H1:mu!=100

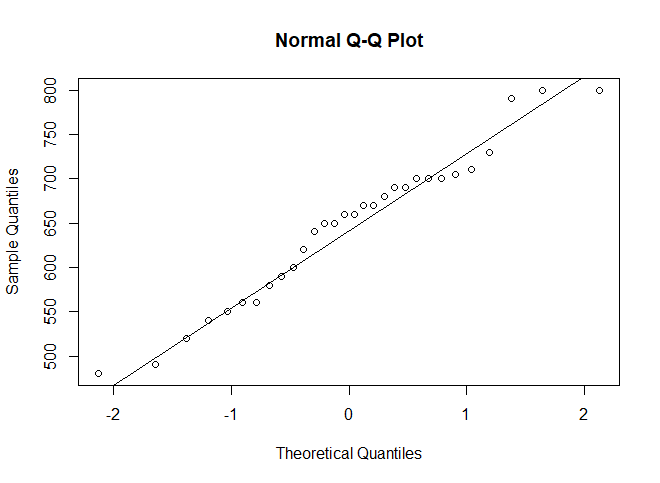
Q)Determine whether the average score of students is higher than 610 in the exam or not. We have the information that the standard deviation for students’ scores is 100. So, we collect the data of 32 students by using random samples and gets following data:

670,730,540,670,480,800,690,560,590,620,700,660,640,710,650,490,800,600,560,700,680,550,580,700,705,690,520,650,660,790

Assume that the score follows a normal distribution. Test at 2% level of significance.







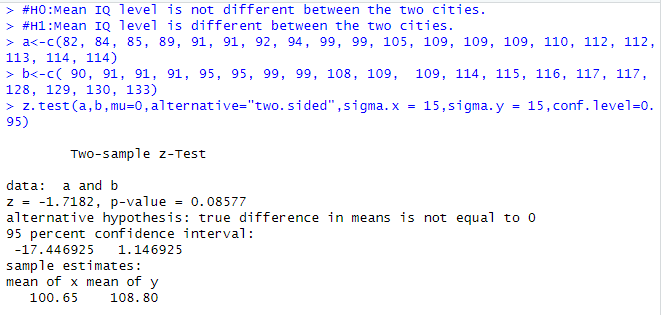
Conclusion : P-value<0.02 ,Hence accept H1:mu >610

2. Two Population mean

Q) Suppose the IQ levels among individuals in two different cities are known to be normally distributed each with population standard deviations of 15. A scientist wants to know if the mean IQ level between individuals in city A and city B are different, so she selects a simple random sample of 20 individuals from each city and records their IQ levels. Determine if the mean IQ level is different between the two cities:

cityA = 82, 84, 85, 89, 91, 91, 92, 94, 99, 99, 105, 109, 109, 109, 110, 112, 112, 113, 114, 114

cityB = 90, 91, 91, 91, 95, 95, 99, 99, 108, 109, 109, 114, 115, 116, 117, 117, 128, 129, 130, 133



Conclusion: P-value>0.05, hence accept #H0:Mean IQ level is not different between the two cities.

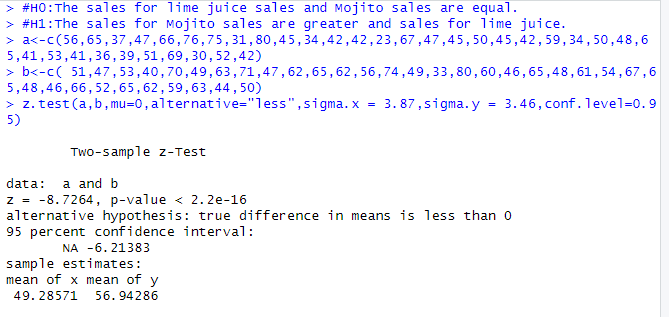
Q) The two independent populations are taken from two shops in a small town.The first shop A sells “traditional” lime juice. However the second shop B is selling “Special” Mojito. We select the two random samples of sales for each drink(shop) and record their sales for 35 days to determine if sales for “Special” Mojito out performed sales of “traditional” lime juice at 5% level of significance. The population variances for lime juice sales is 15 and for Mojito is 12.

lime juice sales = 56,65,37,47,66,76,75,31,80,45,34,42,42,23,67,47,

45,50,45,42,59,34,50,48,65,41,53,41,36,39,51,69,30,52,42

Mojitosales= 51,47,53,40,70,49,63,71,47,62,65,62,56,74,49,33,80,60,46,

65,48,61,54,67,65,48,46,66,52,65,62,59,63,44,50



Conclusion: P-value<0.05, hence accept #H1:The sales for Mojito sales are greater and sales for lime juice.

**PRACTICAL 7**

Aim : Perform an R program on t test- one sample, paired and unpaired

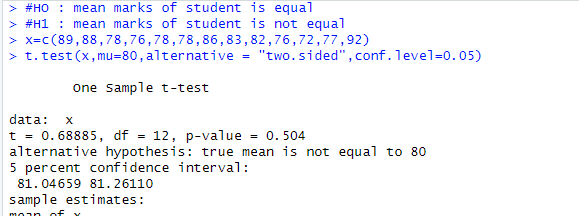
Q1) One Sample T -test

A sample of 13 students from a government school has the following scores in a test.

89 88 78 76 78 78 86 83 82 76 72 77 92.

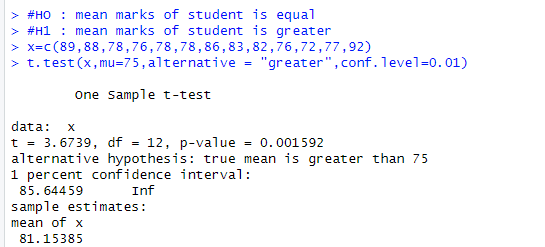
Do this data support that

i) the mean mark of the school students is 80? Test at 5% level.



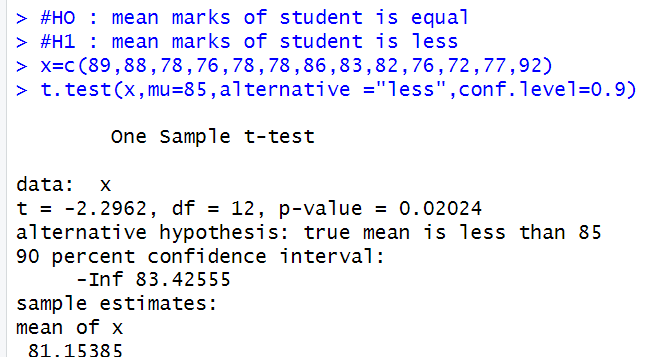
Conclusion : Here P-value >0.05, hence we Accept H1

ii) the mean mark of the school students is more than 75? Test at 1% level.



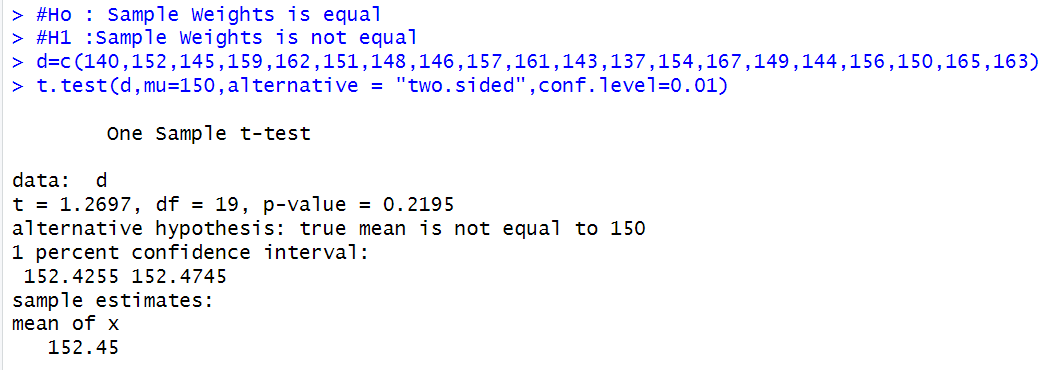
Conclusion : Here P-value < 0.01, hence we Reject H0

iii)the mean mark of the school students is less than 85? Test at 10% level.



Conclusion : Here P-value < 0.9, hence we Reject H0

Q) A sample of weights of 20 apples has weight as follows.140, 152, 145, 159, 162, 151, 148, 146, 157, 161, 143, 137, 154, 167, 149, 144, 156, 150, 165, 163 and you want to test if the average weight of apples in this sample is significantly different from the known average weight of 150 grams. (Lof 99 %)



Conclusion : Here P-value >0.01, hence we Accept H0

Q2) Two Sample T -test (Unpaired Test)

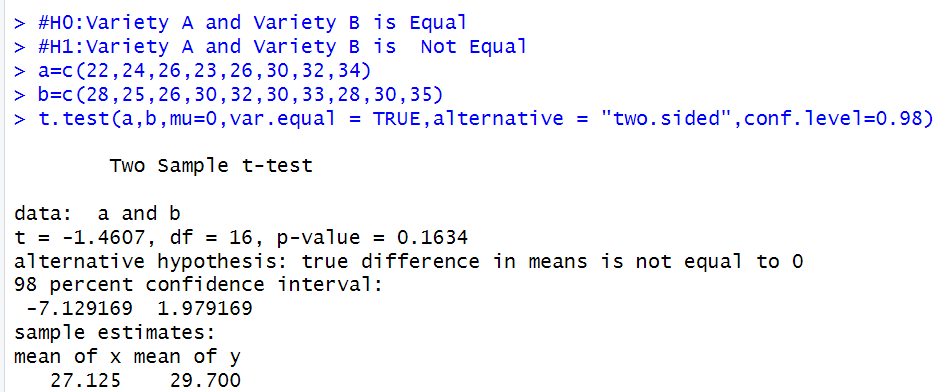
Q) The yield of two varieties of mango (in tons) on two independent samples of 10 and 12 plants are given below.

Variety-A: 22 24 26 23 26 30 32 34

Variety-B: 28 25 26 30 32 30 33 28 30 35

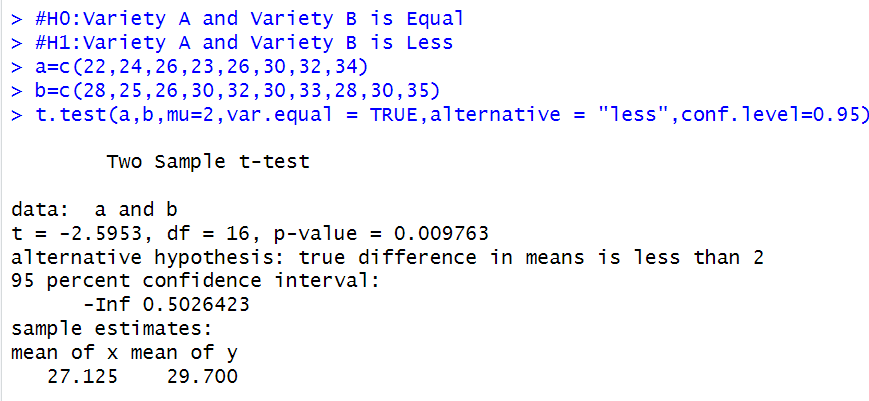
1. Test whether the yield of Variety-A is not equal to Variety-B at 2% level of significance.

yield of Variety-A = Variety-B



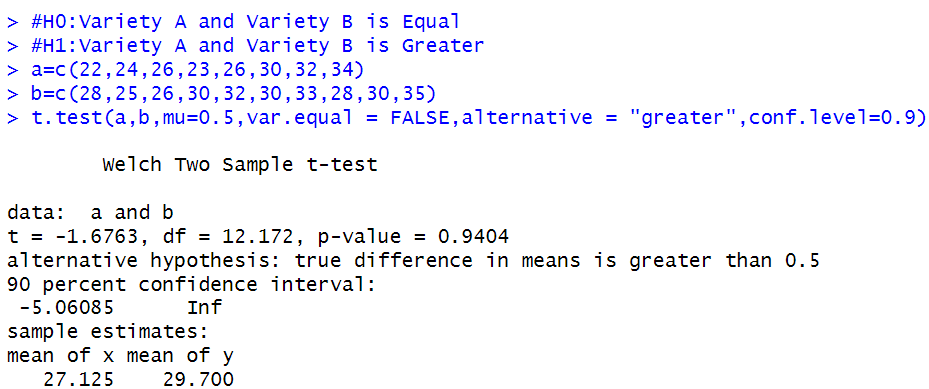
Conclusion : Here P-value >0.02, hence we Accept H0

1. Test whether the difference between yield of Variety-A is less than Variety-B by 2 tones at 5% level of significance.



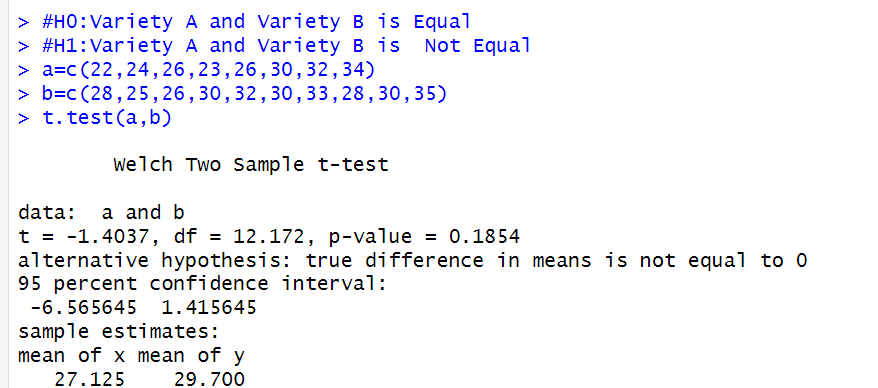
Conclusion : Here P-value <0.05, hence we Reject H0

1. Test whether the difference between yield of Variety-A is more than Variety-B by 0.5 tones at 10% level of significance.



Conclusion : Here P-value >0.01, hence we Accept H0

1. Test whether the yield of Variety-A is not equal to Variety-B at 5% level of significance assume unequal variances of both samples.



Conclusion : Here P-value >0.05, hence we Accept H0

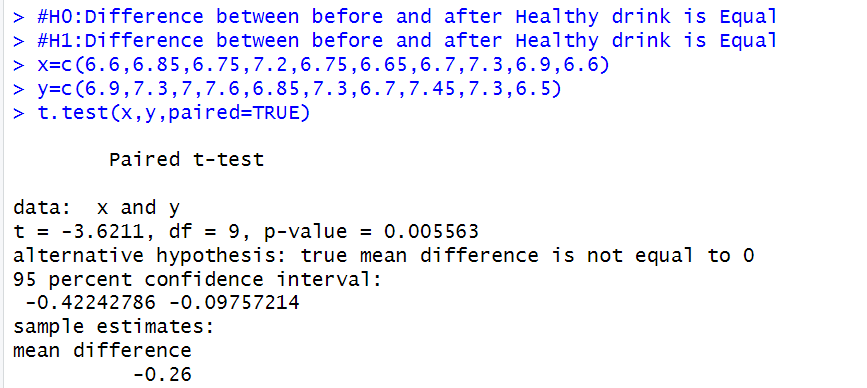
Q3) Paired t-test

Q) A new variety of health drinks in the market for the weight of infants. A sample of 10 babies was selected and was given the above diet for a month and the weights were observed before (X) and after (Y) the diet given.

X : 6.6 6.85 6.75 7.2 6.75 6.65 6.7 7.3 6.9 6.6

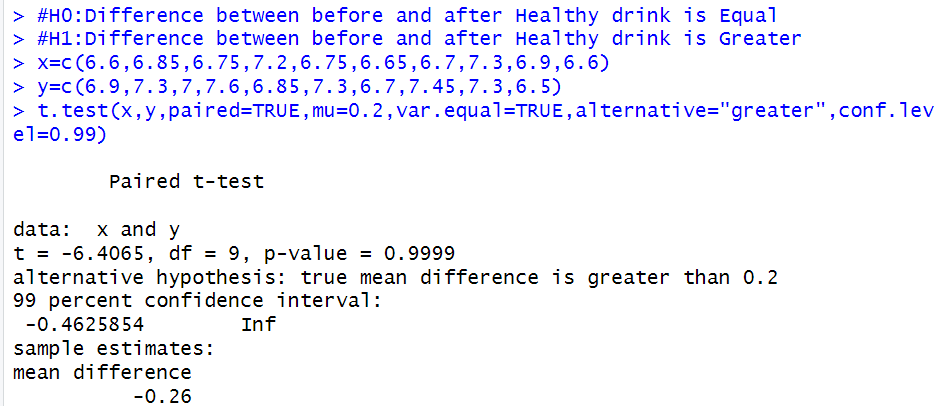
Y : 6.9 7.3 7 7.6 6.85 7.3 6.7 7.45 7.3 6.5

* 1. Examine whether there is significant difference between before and after the healthy drink diet at 5% level of significance.



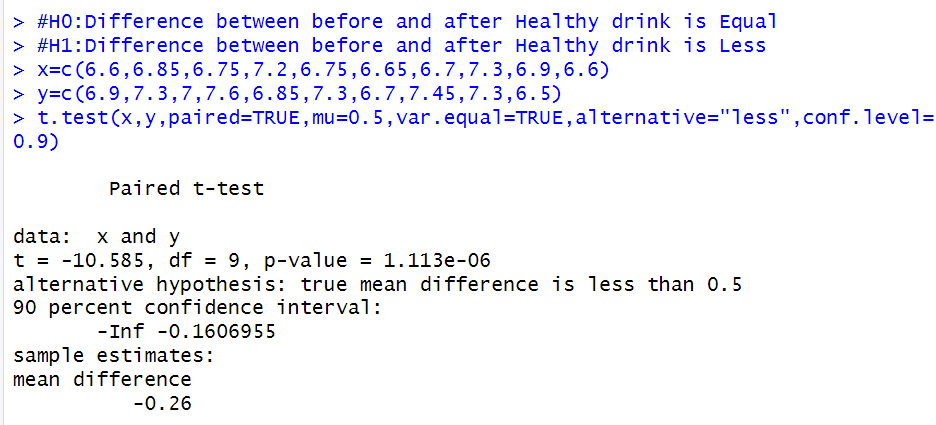
Conclusion : Here P-value <0.05, hence we Reject H0

* 1. Examine whether the weight gain after the healthy drink diet is more than 0.2 kg at 1% level of significance.



Conclusion : Here P-value >0.01, hence we Accept H0

* 1. Examine whether the weight loss after the healthy drink diet is less than 0.5 kg at 10% level of significance.

****

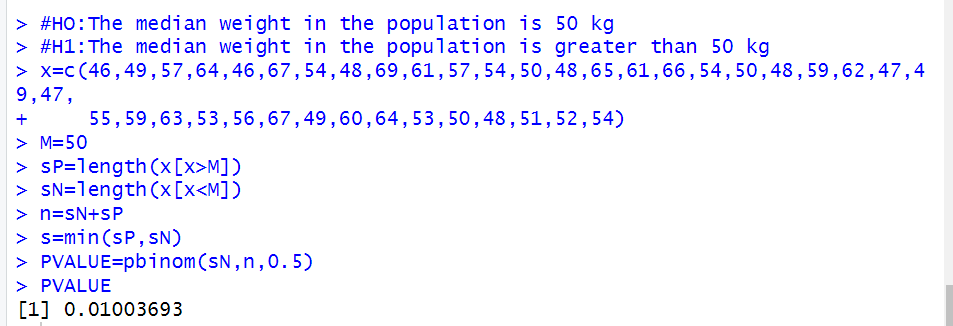
Conclusion : Here P-value <0.1, hence we Reject H0

**PRACTICAL 8**

Aim : Perform an R program on Non ParametricTest -Sign test, wilcoxon signed rank test.

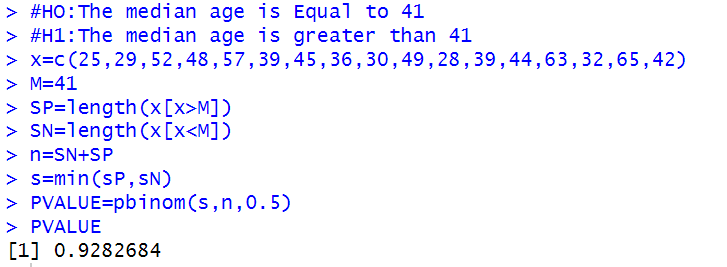
Non ParametricTest -Sign test

Q1. The following data gives the weights (in kg) of 40 students in a random sample. Use a sign test to test whether the median weight in the population is 50 kg against alternatives it is greater than 50kg 46,49,57,64,46,67,54,48,69,61,57,54,50,48,65,61,66,54,50,48,59,62,47,49,47,55,59,63,53,56,67,49,60,64,53,50,48, 51,52,54



Conclusion : Here P-value <0.05, hence we Reject H0

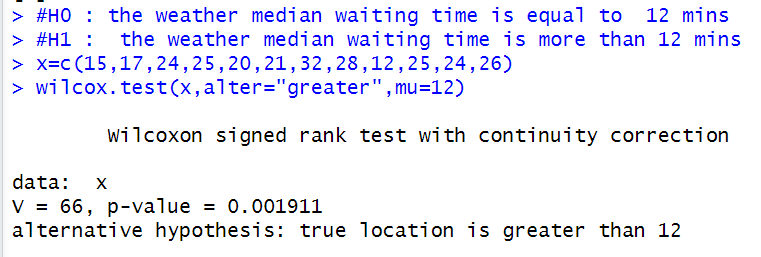
Q2. The median age of a tourist visiting a certain place is 41 years . A random sample of 17 tourists have ages 25,29,52,48,57,39,45,36,30,49,28,39,44,63,32,65,42. Use sign test the claim at 5% l.o.s for the median age is greater than 41.



Conclusion : Here P-value >0.05, hence we Accept H0

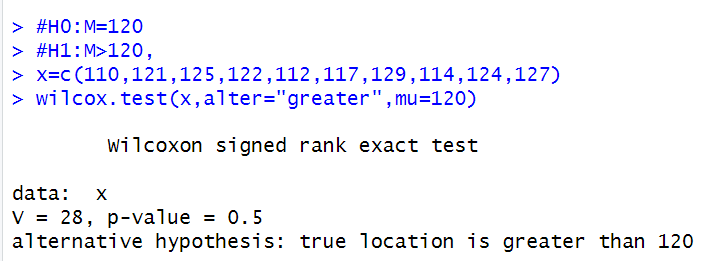
**Wilcoxon Signed Rank Test.**

Q1. The time in minutes that a patient has to wait in a consulting room is recorded for 12 patients 15,17,24,25,20,21,32,28,12,25,24,26. Use wilcoxon signed rank to test the weather median waiting time is more than 12 mins at 5%l.o.s.



Conclusion : Here P-value <0.05, hence we Reject H0

Q2.A random sample of 10 infants showed the following pulse rate: minute: 110,121,125, 122, 112, 117, 129, 114, 124, 127.Assuming that the distribution of pulse rates is symmetric. Is there any evidence to suggest the median pulse rate of infants is more than 120 beats per minute Wilcoxon Signed Rank Test At 5%l.o.s.



Conclusion : Here P-value >0.05, hence we Accept H0

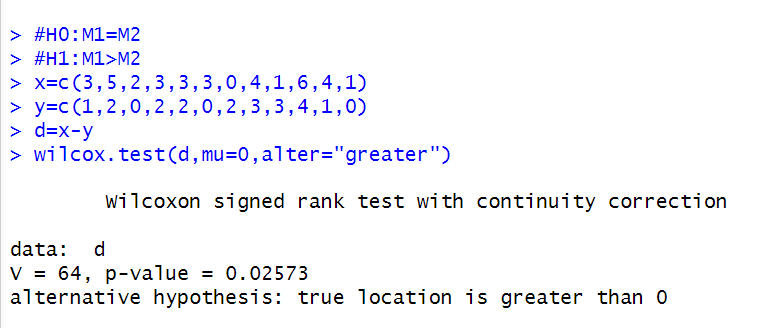
Q3. To determine the effectiveness of a new traffic control system the number of accidents occurred at 12 different locations during four weeks before and after the installation of new system were observed and recorded data is as follows .

location : 1 2 3 4 5 6 7 8 9 10 11 12

Before : 3 5 2 3 3 3 0 4 1 6 4 1

After : 1 2 0 2 2 0 2 3 3 4 1 0

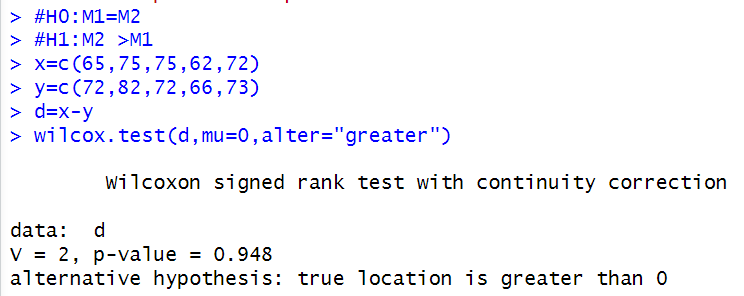
use wilcoxon sign test at5%l.o.s to test whether the traffic control system is effective ?



Conclusion : Here P-value <0.05, hence we Reject H0

Q4.The weights (in kgs)of five persons before and after stopping smoking are as follows .Use wilcoxon sign test at5% l.o.s to test whether a person increases after stopping the smoking.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Person | 1 | 2 | 3 | 4 | 5 |
| before | 65 | 75 | 75 | 62 | 72 |
| after | 72 | 82 | 72 | 66 | 73 |



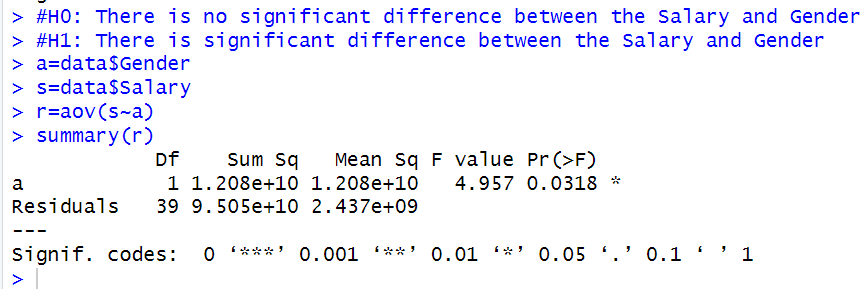
Conclusion : Here P-value > 0.05, hence we Accept H0

**PRACTICAL 9**

Aim: Perform an R program on One way ANOVA and Two way ANOVA.

ONE WAY ANOVA

Q1) Using Dataset



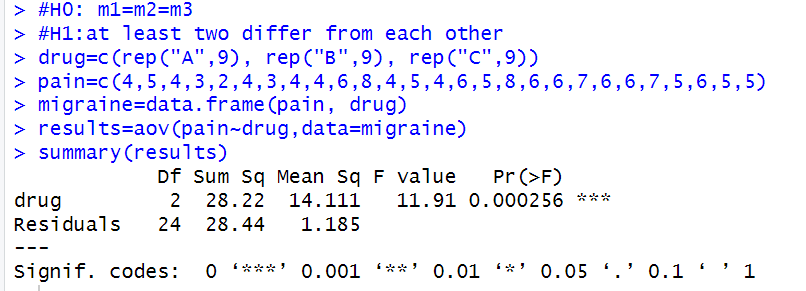
Conclusion : Here P-value <0.05, hence we Reject H0

Q1. A drug company tested three formulations of a relief medicine for migraine headache sufferers for the experiment 27 volunteers were selected and 9 were randomly assigned the one of three drugs formulations the subjects were instructed to take during their next migraine headache and report their pain on the scale of 1 to 10 .

DRUGA:4,5,4,3,2,4,3,4,4

DRUGB:6,8,4,5,4,6,5,8,6

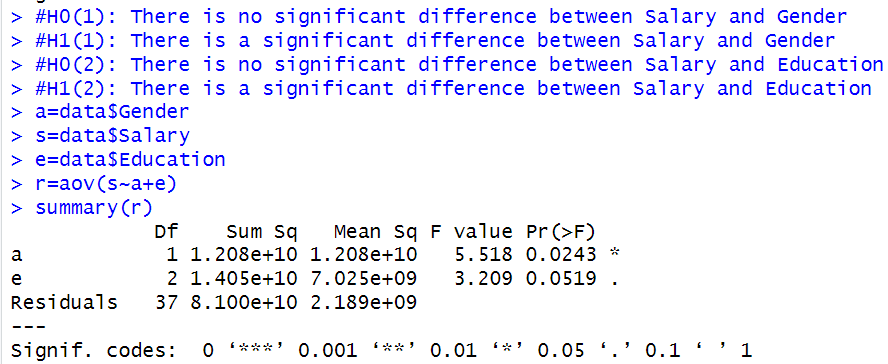
DRUGC:6,7,6,6,7,5,6,5,5



Conclusion : Here P-value <0.05, hence we Reject H0

TWO WAY ANOVA

Q1) Using Dataset



Conclusion : Here P-value of gender=0.0243 <0.05, Reject H0

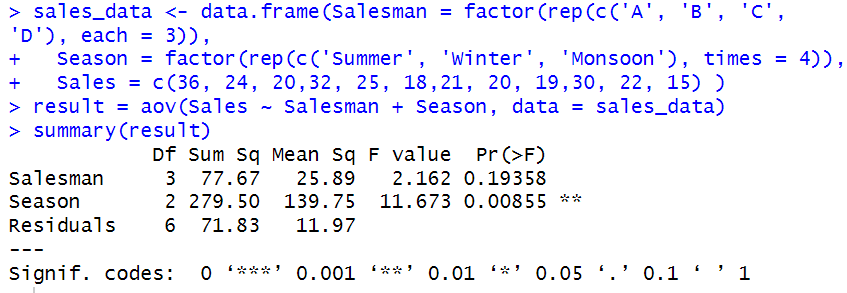
P-value of Education =0.0519 >0.05, Accept H0

A tea company appoints four salesmen a,b,c,and d and observes their sales in three seasons-summer, winter, and monsoon, the figures (in lakhs) of sales are given in the following table .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | B | C | D |
| Summer | 36 | 32 | 21 | 30 |
| winter | 24 | 25 | 20 | 22 |
| monsoon | 20 | 18 | 19 | 15 |

(i)Do the salesmen significantly differ in performance?

(ii)Is there significant difference between the seasons?

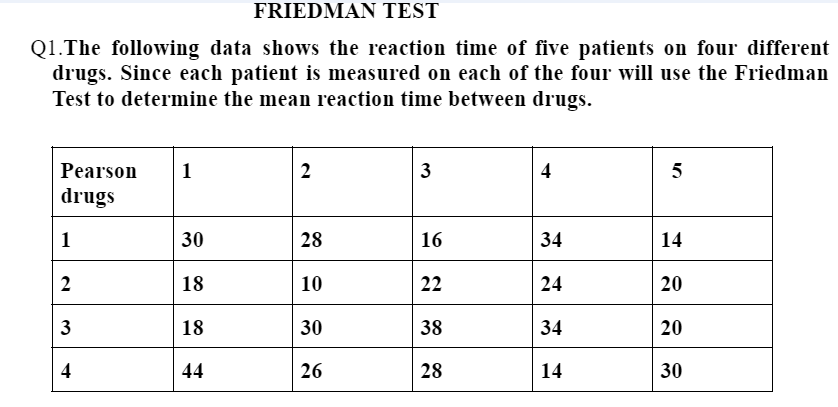


Conclusion : pvalue of season =0.0085 <0.05 ,Reject HO

pvalue of salesman =0.19358 >0.05 , Accept HO

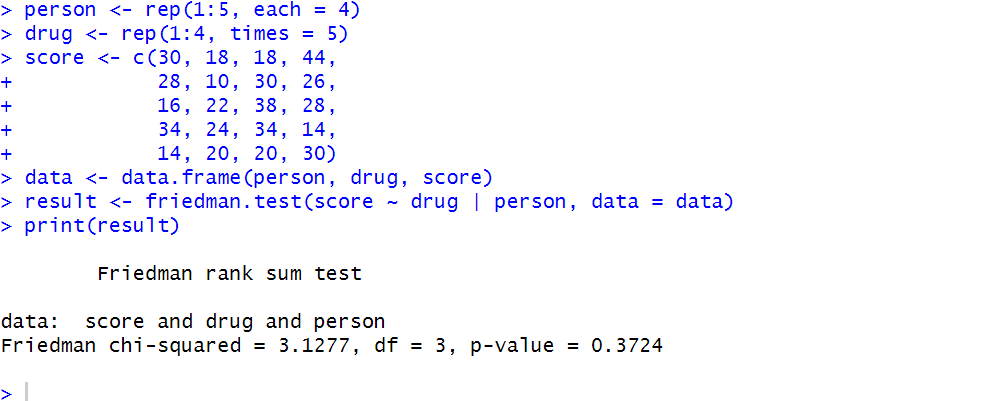
**PRACTICAL 10**

**Aim:** Perform an R program on Friedman Test and Kruskal Wallis test

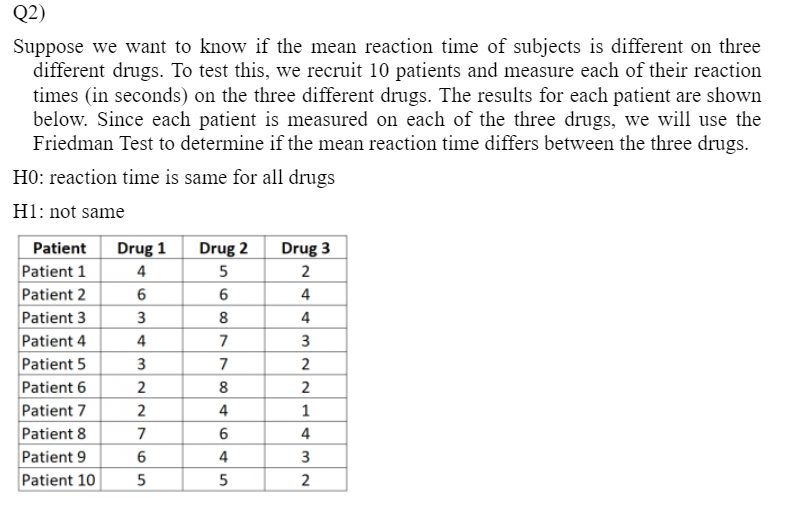
****

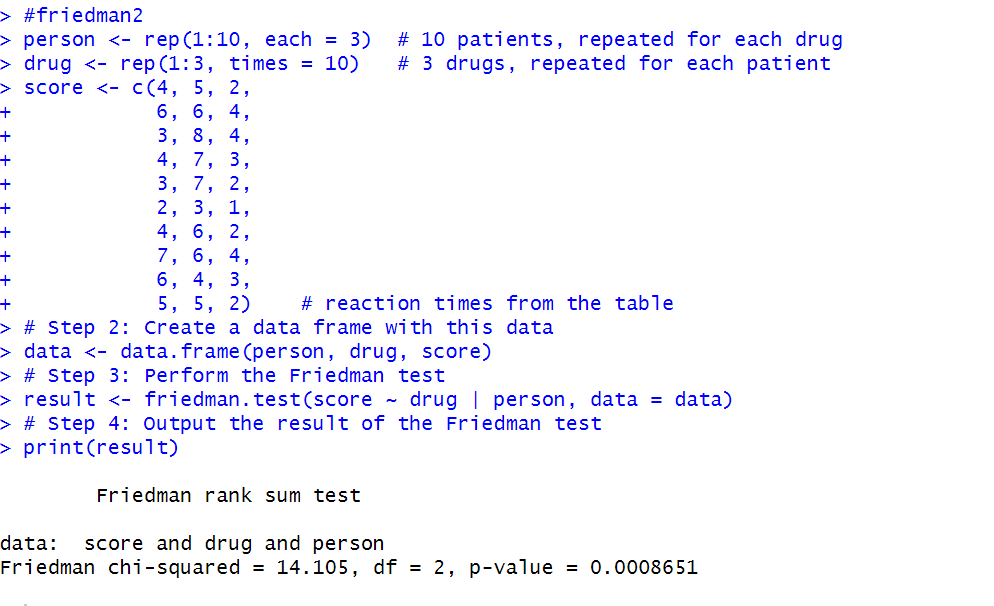
**H0 (null hypothesis): There is no difference in the mean reaction time between the four drugs.**

**H1 (alternative hypothesis): There is a difference in the mean reaction time between the four drugs.**

****

**p-value>0.05 hence we fail to reject the null hypothesis. This means that there is no significant difference in the mean reaction time between the four drugs.**





the **p-value** is less than 0.05, we can reject the null hypothesis H0

which means there is a significant difference in reaction times across the three drugs.

**Kruskal wallis test**

**Q1. The time taken to complete the job on three machines are noted. Test the hypothesis that there is no significant difference between average time taken on these machines to complete the job.**

**x= 2.9, 3.0, 2.5, 2.6,3 .2**

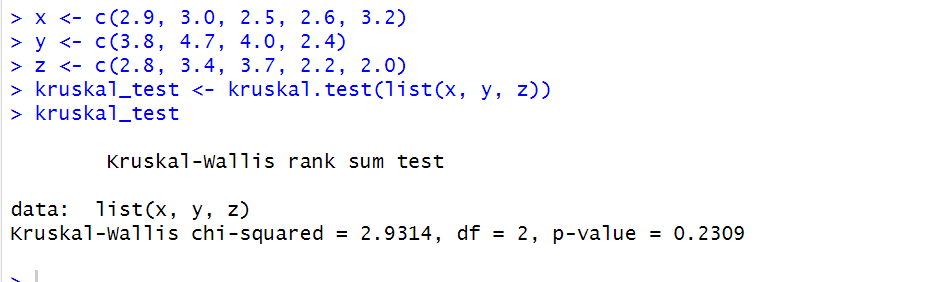
**y= 3.8, 2.7, 4.0, 2, 4**

**z= 2.8, 3.4, 3.7, 2.2, 2.0**

**H0: There is no significant difference between average time taken on three machines to**

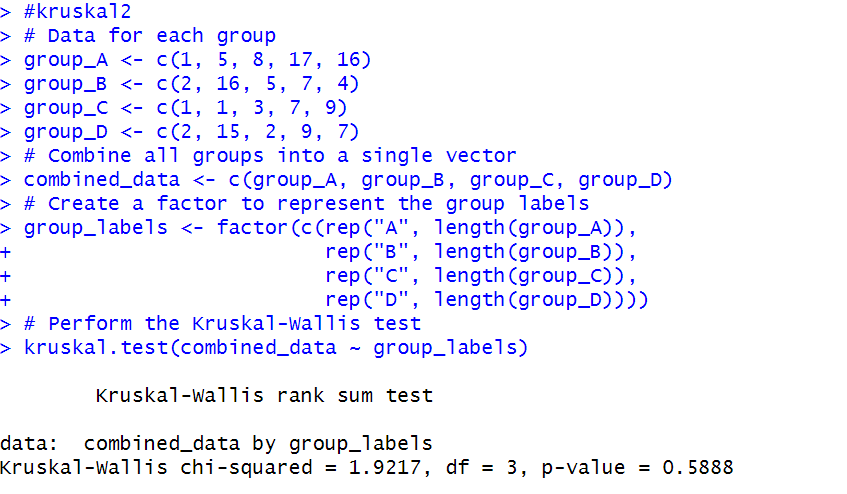
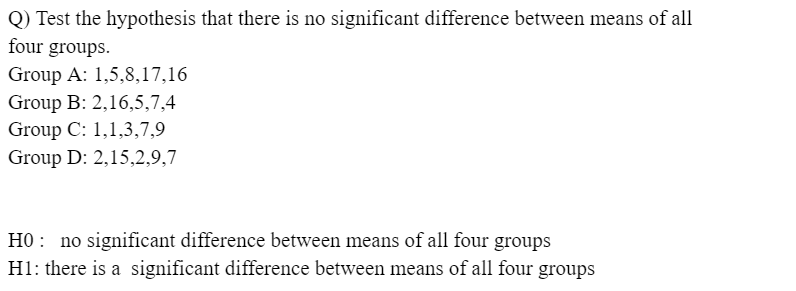
**complete the job.**

**H1: There is a significant difference between the average time taken on three machines to complete the job.**

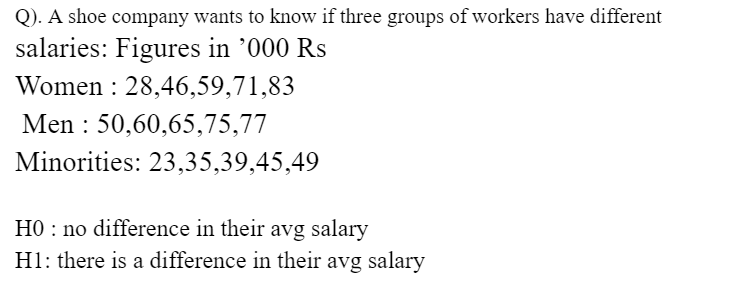


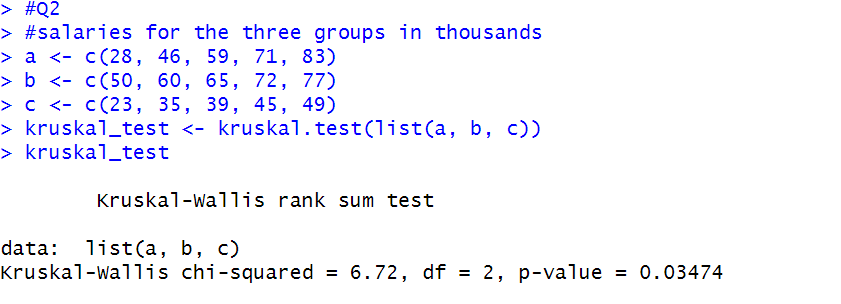
**pval is greater than 0.05 therefore we accept Ho  
Hence There is no significant difference between average time taken on three machines to**

**complete the job.**



the p-value is greater than the chosen significance level 0.05, hence we accept the null hypothesis, indicating that there is no significant difference between the means of given groups.





#pval is less than 0.05, therefore reject Ho (there is a difference in their average salaries)