DAY 2- LAB PROGRAMS

1.Covariance and correlation

Children of three ages are asked to indicate their preference for three photographs of adults. Do the data suggest that there is a significant relationship between age and photograph preference? What is wrong with this study?

**Photograph:**

**Age of child** A B C

5-6 years: 18 22 20

7-8 years: 2 28 40

9-10 years: 20 10 40

1. Use cov() to calculate the sample covariance between B and C.
2. Use another call to cov() to calculate the sample covariance matrix for the preferences.
3. Use cor() to calculate the sample correlation between B and C.
4. Use another call to cor() to calculate the sample correlation matrix for the preferences.

**CODING-**

age\_of\_child<-c("5-6 years","7-8 years","9-10 years")

A<-c(18,2,20)

B<-c(22,28,10)

C<-c(20,40,40)

df<-data.frame(age\_of\_child,A,B,C)

df

#1a

covar<-cor(B,C)

covar

#1b

cov(df[c(2,3,4)])

#1c

cor(B,C)

#1d

cor(df[c(2,3,4)])

**OUTPUT-**

**#1a**

[1] -0.1889822

**#1b**

> cov(df[c(2,3,4)])

A B C

A 97.33333 -74 -46.66667

B -74.00000 84 -20.00000

C -46.66667 -20 133.33333

**#1c**

[1] -0.1889822

**#1d**

> cor(df[c(2,3,4)])

A B C

A 1.0000000 -0.8183918 -0.4096440

B -0.8183918 1.0000000 -0.1889822

C -0.4096440 -0.1889822 1.0000000

2.Imagine that you have selected data from the All Electronics data warehouse for analysis. The data set will be huge! The following data are a list of All Electronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, ,20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30)

|  |
| --- |
| 30, 30.  (i) Partition the dataset using an equal-frequency partitioning method with bin equal to 3 (ii) apply data smoothing using bin means and bin boundary. (iii) Plot Histogram for the above frequency division |

**CODING-**

#2

data<-c(1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18,20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30)

bin=3

**#2a-equal frequency**

freq<-length(data)/bin

print(freq)

**#2b-bins mean**

range=ceiling(freq)

bin1=c()

bin2=c()

bin3=c()

for(i in data[1:range]){

bin1=append(bin1,i)

}

range1=range+1

range2=range\*2

for(j in data[range1:range2])

{

bin2=append(bin2,j)

}

range3=range2+1

range4=range\*3

for(k in data[range3:range4])

{

bin3=append(bin3,k)

}

mean(bin1)

mean(bin2)

new\_bin3<-bin3[! bin3%in% c(NA)]

mean(new\_bin3)

**#2c**

hist(bin1)

hist(bin2)

hist(bin3)

**OUTPUT-**

**#2a**

[1] 15.33333

**#2b**

> mean(bin1)

[1] 7.6875

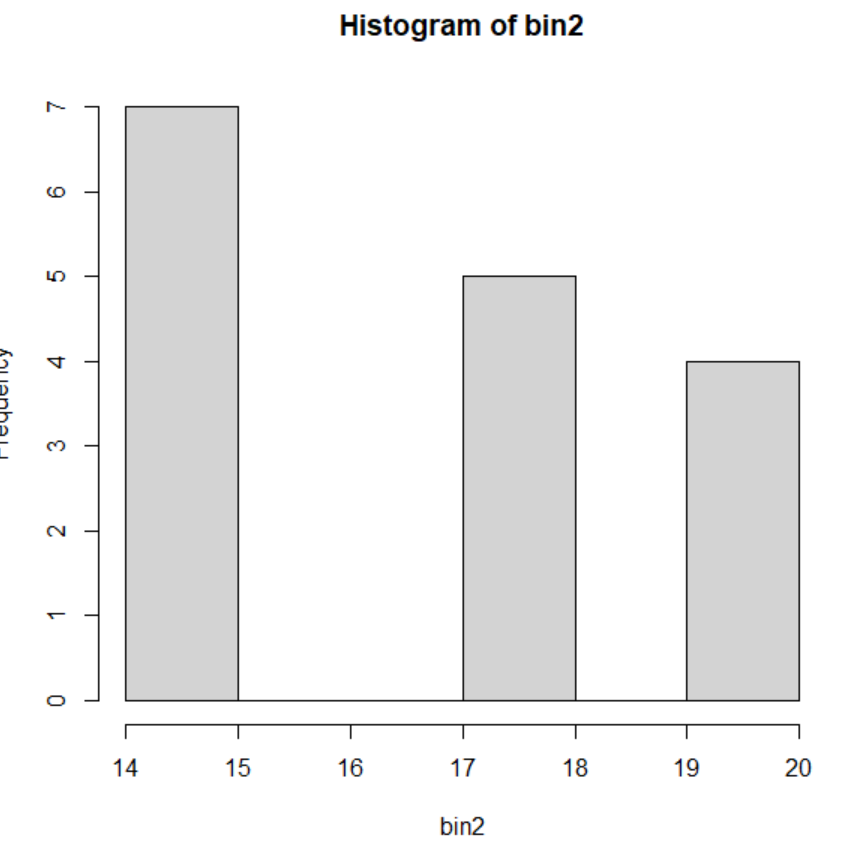
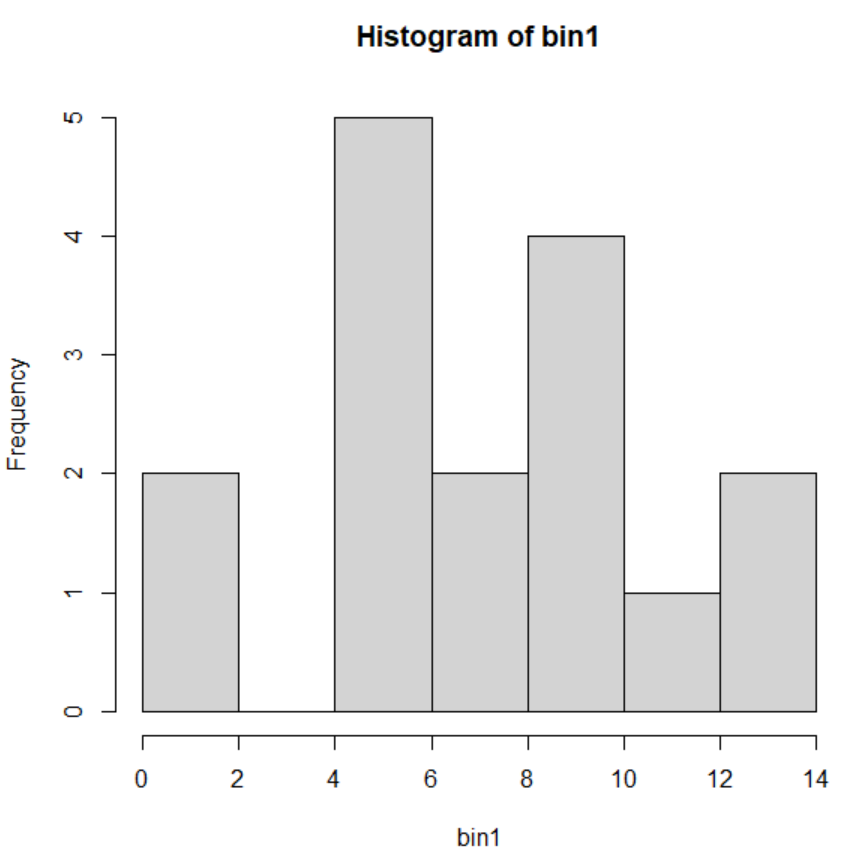
> mean(bin2)

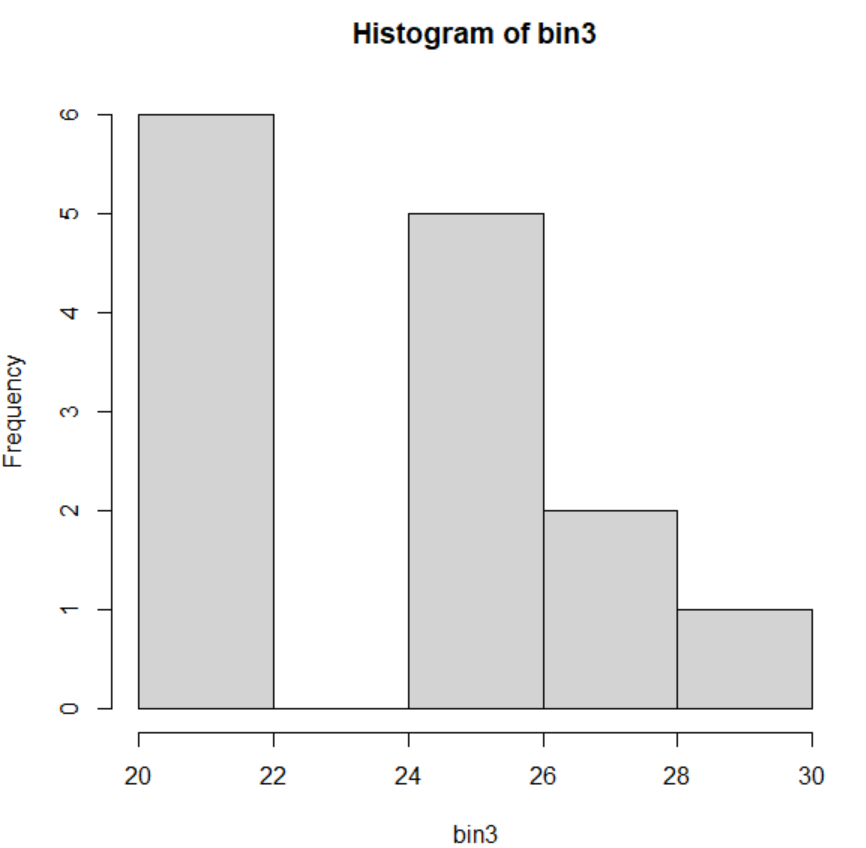
[1] 17.125

> mean(new\_bin3)

[1] 23.92857

**#2c**





3.Two Maths teachers are comparing how their Year 9 classes performed in the end of year exams. Their results are as follows:  
Class A: 76, 35, 47, 64, 95, 66, 89, 36, 8476,35,47,64,95,66,89,36,84

Class B: 51, 56, 84, 60, 59, 70, 63, 66, 5051,56,84,60,59,70,63,66,50

(i) Find which class had scored higher mean, median and range.  
(ii) Plot above in boxplot and give the inferences

**CODING-**

#3a

class\_A<-c(76, 35, 47, 64, 95, 66, 89, 36, 8476,35,47,64,95,66,89,36,84)

class\_B<-c(51, 56, 84, 60, 59, 70, 63, 66, 5051,56,84,60,59,70,63,66,50)

mean\_a<-mean(class\_A)

mean\_a

mean\_b<-mean(class\_B)

mean\_b

median\_a<-median(class\_A)

median\_a

median\_b<-median(class\_B)

median\_b

range\_a<-max(class\_A)-min(class\_A)

range\_a

range\_b<-max(class\_B)-min(class\_B)

range\_b

print(range\_a)

print(range\_b)

#3b-boxplot

boxplot(class\_A)

boxplot(class\_B)

#3b-inferences

if (mean\_a > mean\_b) {

cat("Class A had a higher mean score.\n")

} else if (mean\_a < mean\_b) {

cat("Class B had a higher mean score.\n")

} else {

cat("Both classes had the same mean score.\n")

}

if (median\_a > median\_b) {

cat("Class A had a higher median score.\n")

} else if (median\_a < median\_b) {

cat("Class B had a higher median score.\n")

} else {

cat("Both classes had the same median score.\n")

}

if (range\_a > range\_b) {

cat("Class A had a higher range of scores.\n")

} else if (range\_a < range\_b) {

cat("Class B had a higher range of scores.\n")

} else {

cat("Both classes had the same range of scores.\n")

}

**OUTPUT-**

> mean\_a

[1] 558.8235

> mean\_b

[1] 356.9412

> median\_a

[1] 66

> median\_b

[1] 63

> print(range\_a)

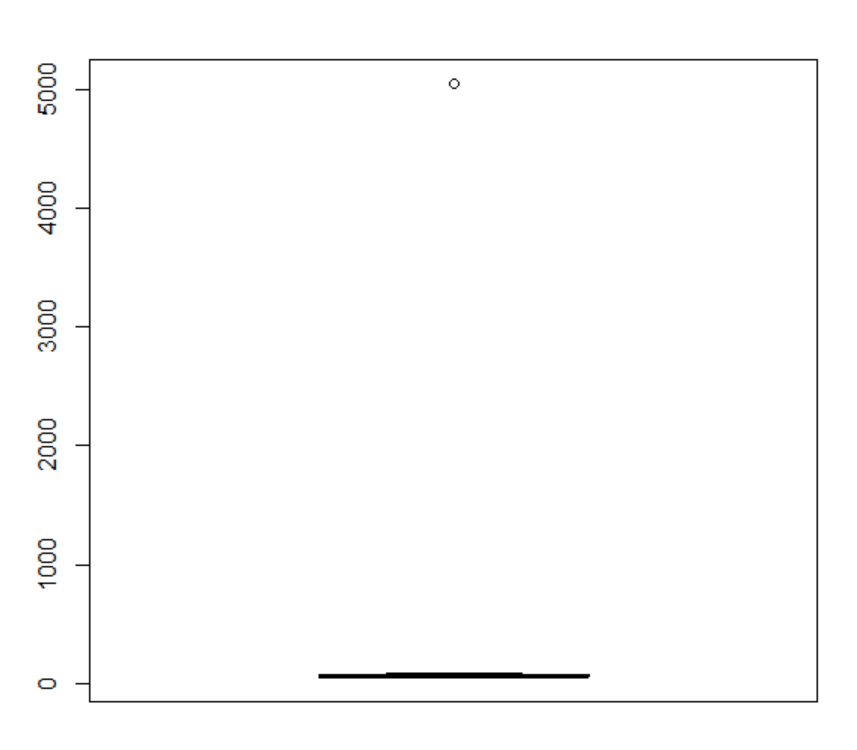
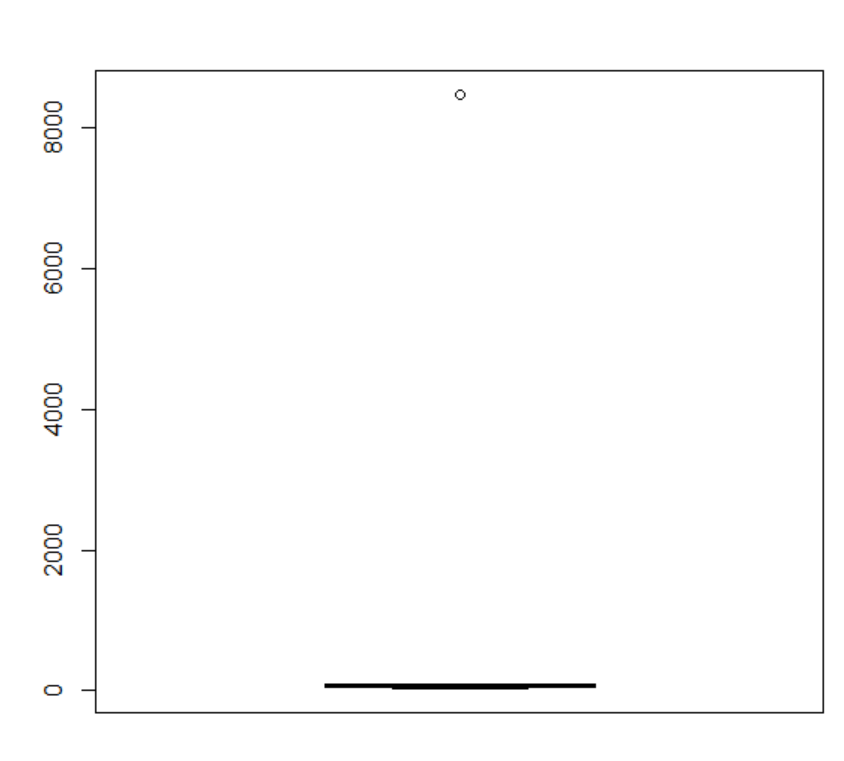
[1] 8441

> print(range\_b)

[1] 5001

**#3b**

**Class a class b**



**#3C**

Class A had a higher mean of scores

Class A had a higher median of scores

Class A had a higher range of scores

**4) Let us consider one example to make the calculation method clear. Assume that the minimum and maximum values for the feature F are $50,000 and $100,000 correspondingly. It needs to range F from 0 to 1. In accordance with min-max normalization, v = $80,**

**b) Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000**

**(a) min-max normalization by setting min = 0 and max = 1**

**(b) z-score normalization**

**#4**

v<-80

min<-50000

max<-100000

result1=v-min

result2=max-min

result3=result1/result2

print(result3)

#min max normalization

data <- c(200, 300, 400, 600, 1000)

min<-min(data)

max<-max(data)

for (i in data)

{

result1=i-min

result2=max-min

result3=result1/result2

print(result3)

}

#z score

data <- c(200, 300, 400, 600, 1000)

mean1<-mean(data)

deviation<-sd(data)

for (i in data)

{

result1=i-mean1

result2=result1/deviation

print(result2)

}

**OUTPUT-**

[1] -0.9984

#min max normalization

[1] 0

[1] 0.125

[1] 0.25

[1] 0.5

[1] 1

#z score normalization

[1] -0.9486833

[1] -0.6324555

[1] -0.3162278

[1] 0.3162278

[1] 1.581139

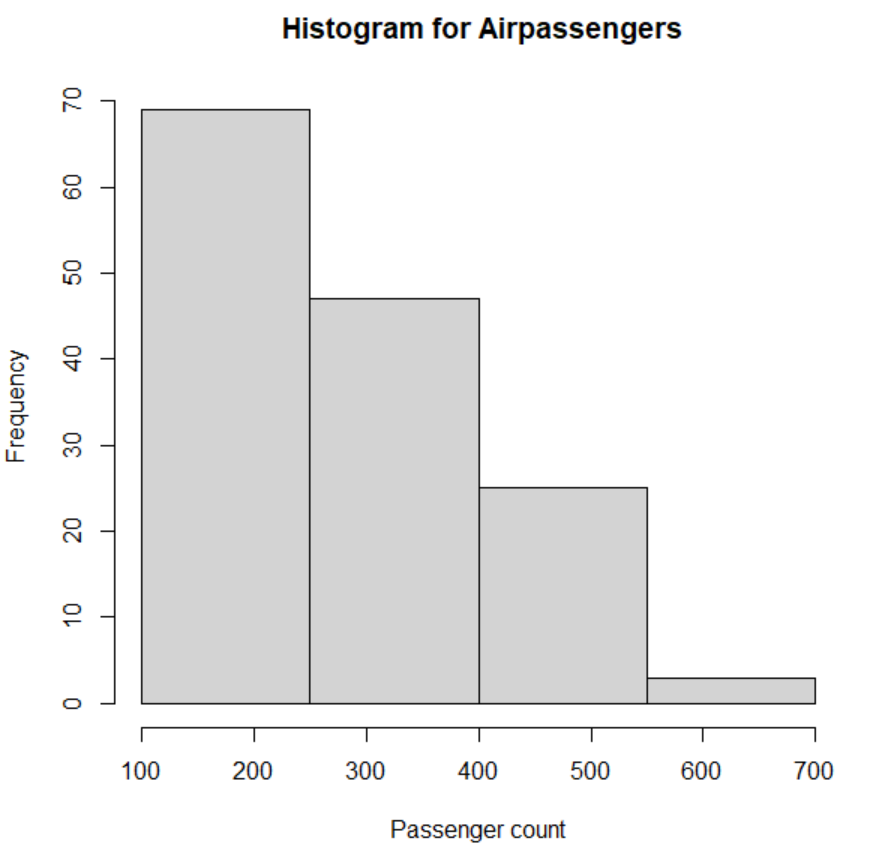
5.Make a histogram for the “AirPassengers “dataset, start at 100 on the x-axis, and from values 200 to 700, make the bins 150 wide

**CODING-**

data("AirPassengers")

hist(AirPassengers, breaks = seq(100, 700, by = 150), main=" Histogram for Airpassengers", xlab = "Passenger count", ylab = "Frequency")

**OUTPUT-**

****

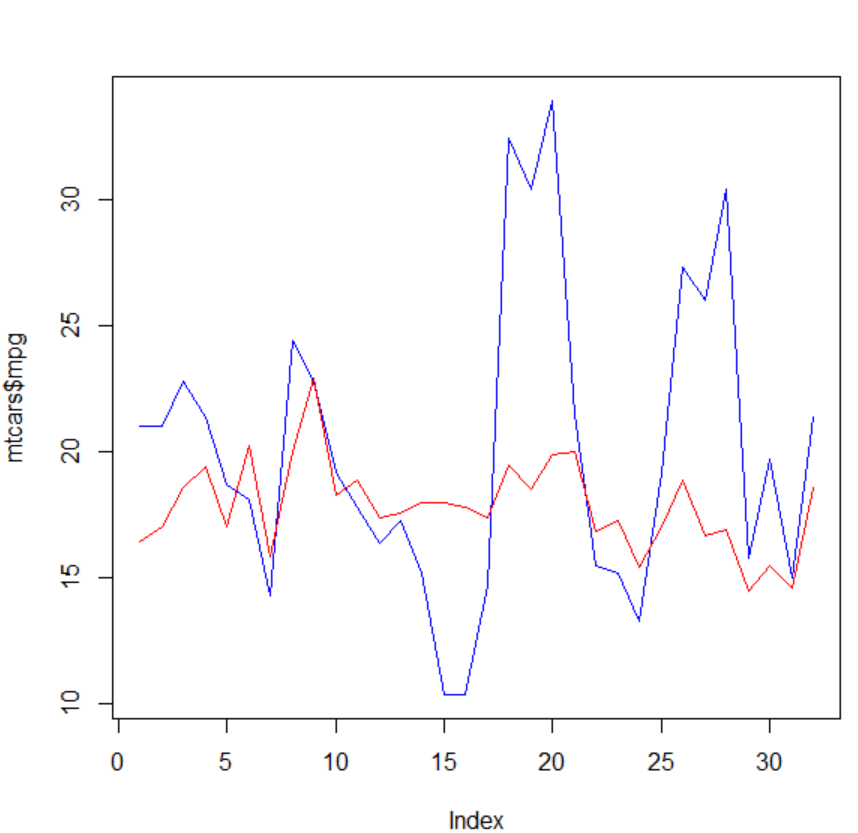
6.Obtain Multiple Lines in Line Chart using a single Plot Function in R.Use attributes“mpg”and“qsec”of the dataset “mtcars”

**CODING-**

plot(mtcars$mpg,type = "l",col="blue")

lines(mtcars$qsec,type="l",col="red")

**OUTPUT-**

****

7.Download the Dataset "water" From R dataset Link.Find out whether there is a linear relation between attributes"mortality" and"hardness" by plot function.Fit the Data into the Linear Regression model.Predict the mortality for the hardness=88.

**CODING-**

path <- "/Users/kadiv/OneDrive/Desktop/water\_potability.csv"

content1 <- read.csv(path)

# contents of the csv file

print(content1)

#linear regression

linear\_reg<-lm(ph~Hardness,data=content1)

new\_var<-data.frame(Hardness = 88)

#predicted value with newdata

predict(linear\_reg,newdata=new\_var)

**output-**

1

6.650547

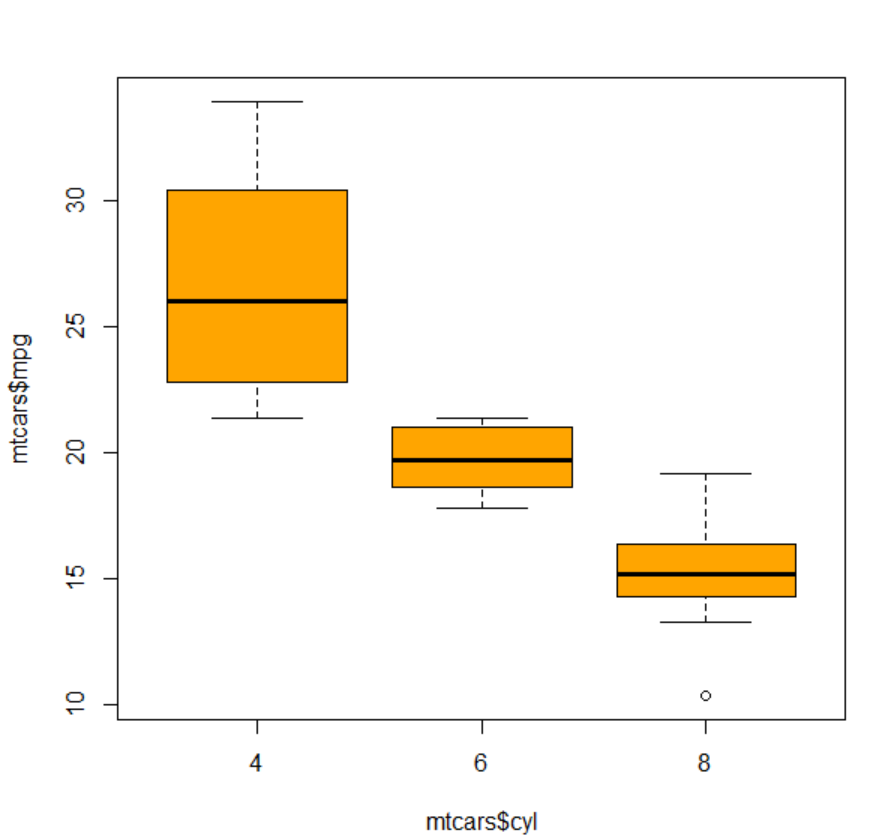
**8.Create a Boxplot graph for the relation between "mpg"(miles per galloon) and "cyl"(number of Cylinders) for the dataset "mtcars" available in R Environment.**

**CODING-**

mtcars

boxplot(mtcars$mpg~mtcars$cyl,col='orange')

**OUTPUT-**

****

9. Assume the Tennis coach wants to determine if any of his team players are scoring

outliers. To visualize the distribution of points scored by his players, then how can he

decide to develop the box plot? Give suitable example using Boxplot visualization

technique.

**CODING-**

players<-c("player1","player2","player3","player4","player5","player6","player7","player8","player9","player10")

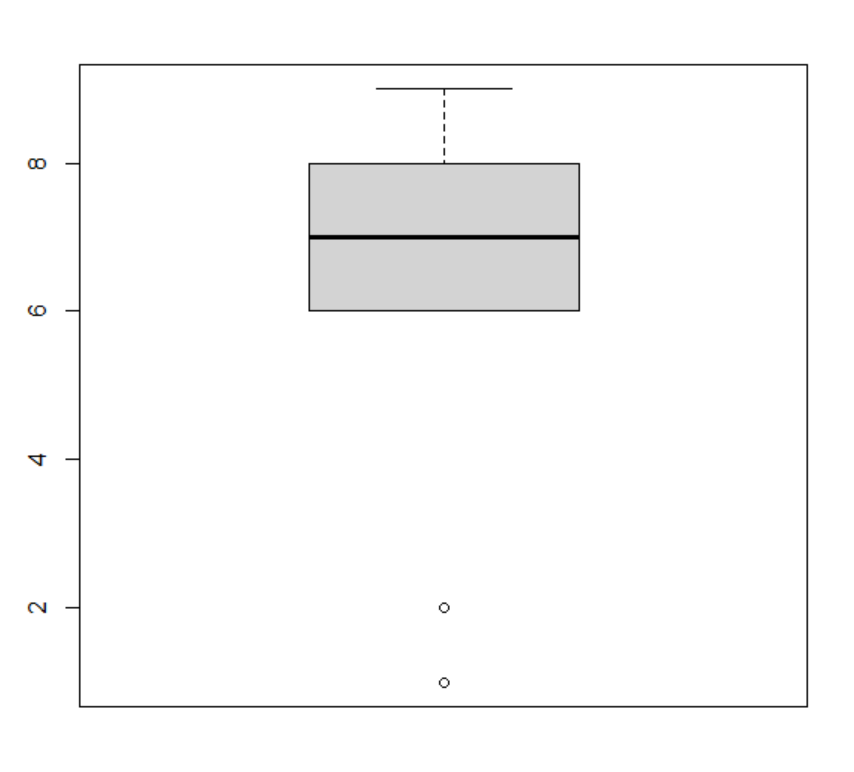
points<-c(8,9,6,7,7,1,2,6,8,9)

data.frame(players,points)

boxplot(points)

#outliers are points-(1,2)

**OUTPUT-**

****

10. Implement using R language in which age group of people are affected byblood pressure based on the diabetes dataset show it using scatterplot and bar chart (that is BloodPressure vs Age using dataset “diabetes.csv”)

**CODING-**

**#enter your copy path to diabetes.csv**

path <- "/Users/kadiv/OneDrive/Desktop/Tools DWDM/diabetes.csv"

content <- read.csv(path)

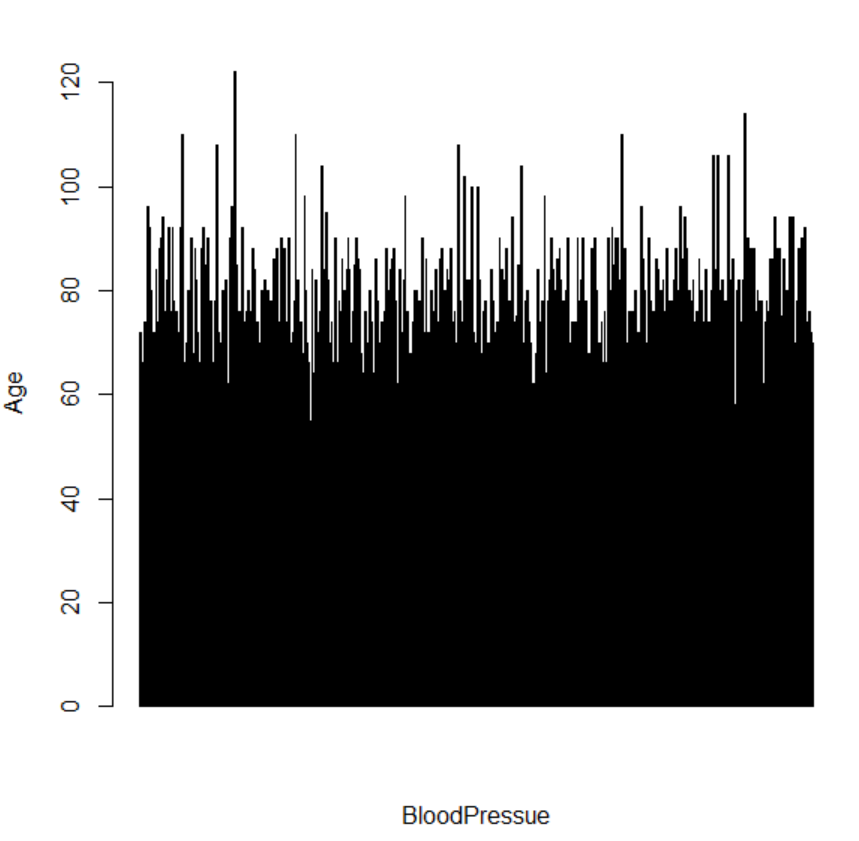
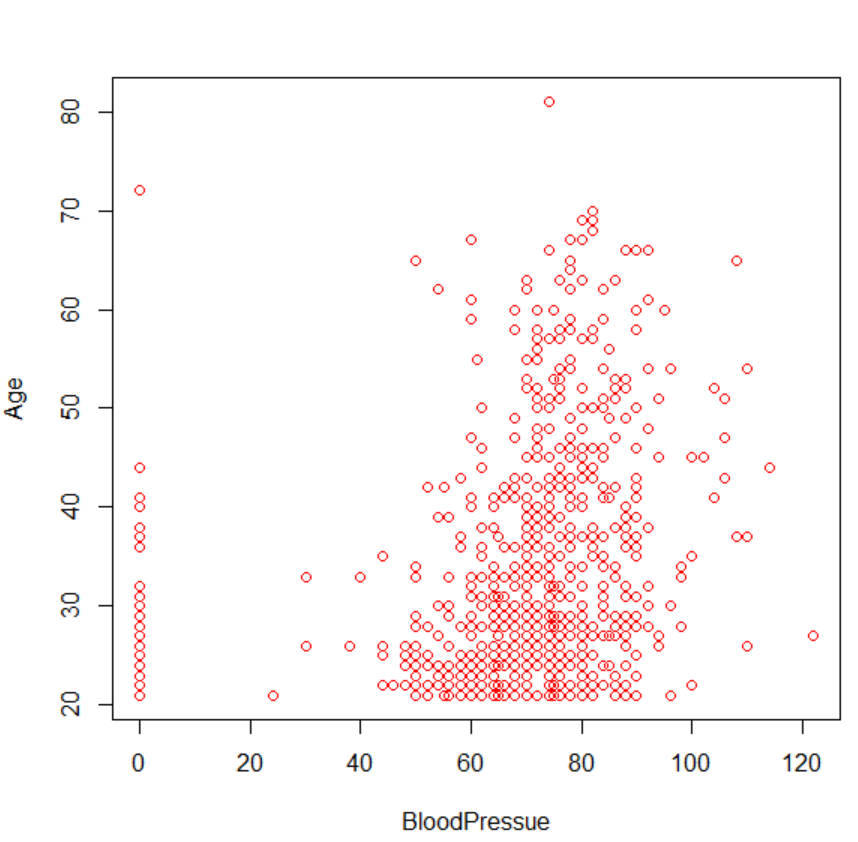
# contents of the csv file

print(content)

plot(content$BloodPressure,content$Age,col='red',xlab='BloodPressue',ylab='Age')

barplot(content$BloodPressure,content$Age,xlab='BloodPressue',ylab='Age')

**OUTPUT-**

****