DWDM R PROGRAMMING-PRACTICALS

1.List of Programs:

1The intervals and corresponding frequencies are as follows. age frequency

1-5. 200

5-15 450

15-20 300

20-50 1500

50-80 700

80-110 44

Compute an approximate median value for the data

**CODING-**

intervals <- c("1-5", "5-15", "15-20", "20-50", "50-80", "80-110")

frequencies <- c(200, 450, 300, 1500, 700, 44)

cumulative\_freq <- cumsum(frequencies)

median\_interval\_index <- which(cumulative\_freq >= sum(frequencies)/2)[1]

lower\_bound <- as.numeric(strsplit(intervals[median\_interval\_index], "-")[[1]][1])

upper\_bound <- as.numeric(strsplit(intervals[median\_interval\_index], "-")[[1]][2])

cumulative\_freq\_before <- cumulative\_freq[median\_interval\_index] - frequencies[median\_interval\_index]

frequency\_median <- frequencies[median\_interval\_index]

width <- upper\_bound - lower\_bound

median\_value <- lower\_bound + ((sum(frequencies)/2 - cumulative\_freq\_before) / frequency\_median) \* width

print(paste("Approximate Median Value:", median\_value))

**OUTPUT-**

[1] "Approximate Median Value: 32.94"

2. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the mean of the data? What is the median?

(b) What is the mode of the data? Comment on the data’s modality (i.e., bimodal, trimodal, etc.).

(c) What is the midrange of the data?

(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

**Coding:**

**#2a**

x<-c(13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70)

#mean

mean(x)

#median

median(x)

**output:**

mean(x)

[1] 29.96296

> #median

> median(x)

[1] 25

CODING FOR 2b-

**#2b**

#mode

MultipleModes <- function(x) {

uniqx <- unique(x)

freq\_table <- tabulate(match(x, uniqx))

modes <- uniqx[freq\_table == max(freq\_table)]

modes

}

age\_values <- c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

multiple\_modes <- MultipleModes(age\_values)

print(multiple\_modes)

**output:**

25 35

**CODING FOR 2c-**

#midrange

c) age\_values <- c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

X<-(min(age\_values)+max(age\_values))/2

print(X)

OUTPUT-

41.5

**CODING FOR 2d-**

d) #quartile

age\_values <- c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

quantile(age\_values)

**output:** 0% 25% 50% 75% 100%

13.0 20.5 25.0 35.0 70.0

3.Data Preprocessing :Reduction and Transformation

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

**Coding:**

**#3a**

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)

}

**OUTPUT:**

[1] 0

[1] 0.125

[1] 0.25

[1] 0.5

[1] 1

**#3b**

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] -0.9486833

[1] -0.6324555

[1] -0.3162278

[1] 0.3162278

[1] 1.581139

4.Data:11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,

72,73,75

a) Smoothing by bin mean

b) Smoothing by bin median

c) Smoothing by bin boundaries

**CODING-**

#binning

data <- c(11, 13, 13, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 45, 71, 72, 73, 75)

range=6

bin1=c()

bin2=c()

bin3=c()

bin4=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)

}

range5=range4+1

range6=range\*4

for(l in data[range5:range6]){

bin4=append(bin4,l)

}

**#4a**

mean(bin1)

mean(bin2)

mean(bin3)

mean(bin4)

**#4b**

median(bin1)

median(bin2)

median(bin3)

median(bin4)

**OUTPUT:**

**#4a**

> mean(bin1)

[1] 13.83333

> mean(bin2)

[1] 20.16667

> mean(bin3)

[1] 30.66667

> mean(bin4)

[1] 63.5

>

**#4b**

> median(bin1)

[1] 14

> median(bin2)

[1] 20

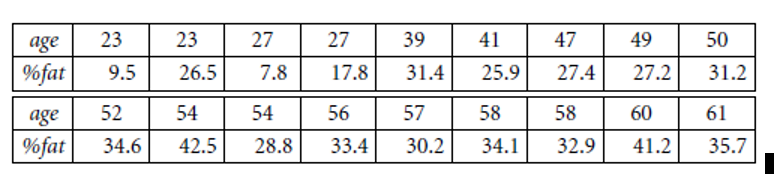
> median(bin3)

[1] 27

> median(bin4)

[1] 71.5

5) 5. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:



**CODING-**

age <- c(23,23,27,27,39,41,47,49,50,52,54,54,56,57,58,58,60,61)

body\_fat\_percent <- c(9.5,26.5,7.8,17.8,31.4,25.9,27.4,27.2,31.2,34.6,42.5,28.8,33.4,30.2,34.1,32.9,41.2,35.7)

**#5.a**

mean(age)

mean(body\_fat\_percent)

median(age)

median(body\_fat\_percent)

sd(age)

sd(body\_fat\_percent)

**#5.b**

#create dataframe

df<-data.frame(age,body\_fat\_percent)

#box plot

boxplot(df)

#scatter plot

plot(df)

#qq plot

qqnorm(age)

qqline(age)

qqnorm(body\_fat\_percent)

qqline(body\_fat\_percent)

**OUTPUT-**

**#5a**

> mean(age)

[1] 46.44444

> mean(body\_fat\_percent)

[1] 28.78333

> median(age)

[1] 51

> median(body\_fat\_percent)

[1] 30.7

> sd(age)

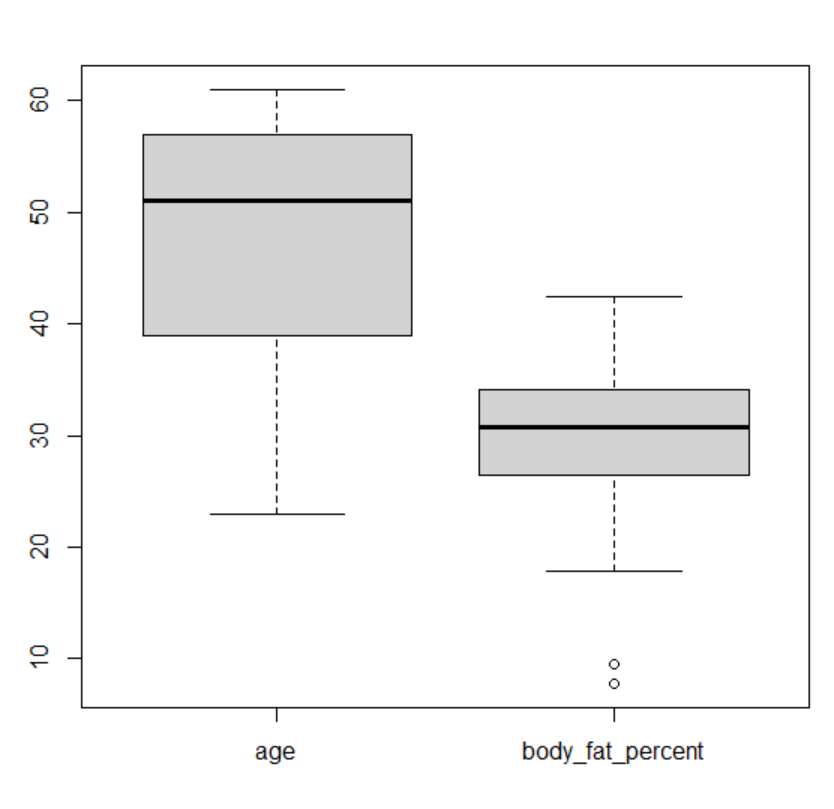
[1] 13.21862

> sd(body\_fat\_percent)

[1] 9.254395

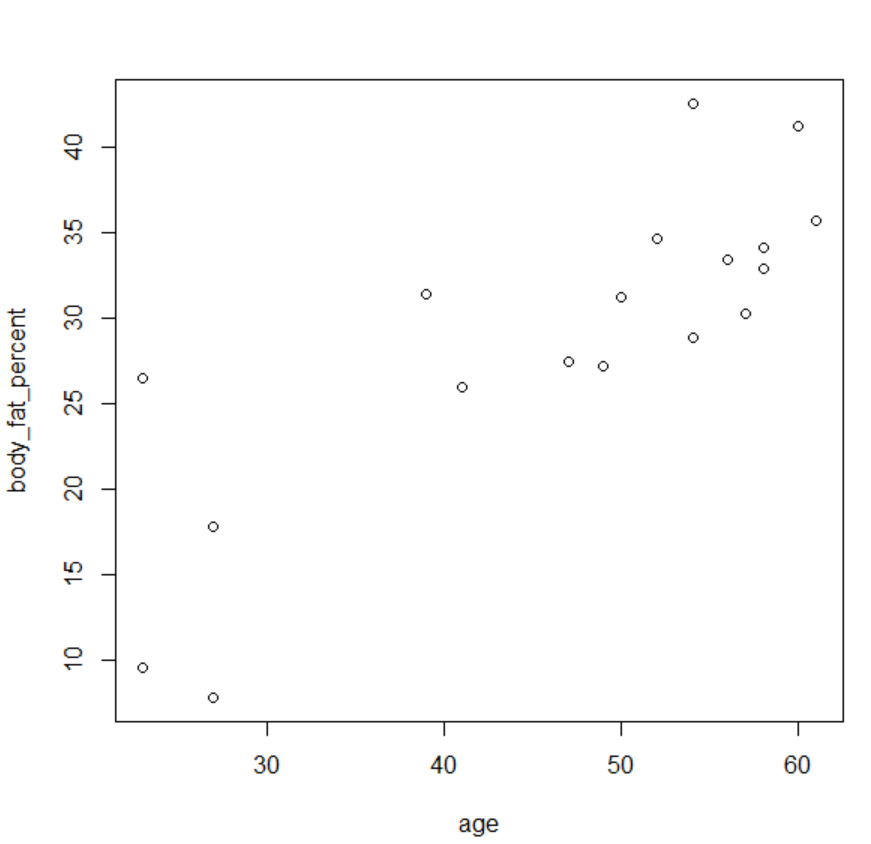
**#5.b**

BOXPLOT-

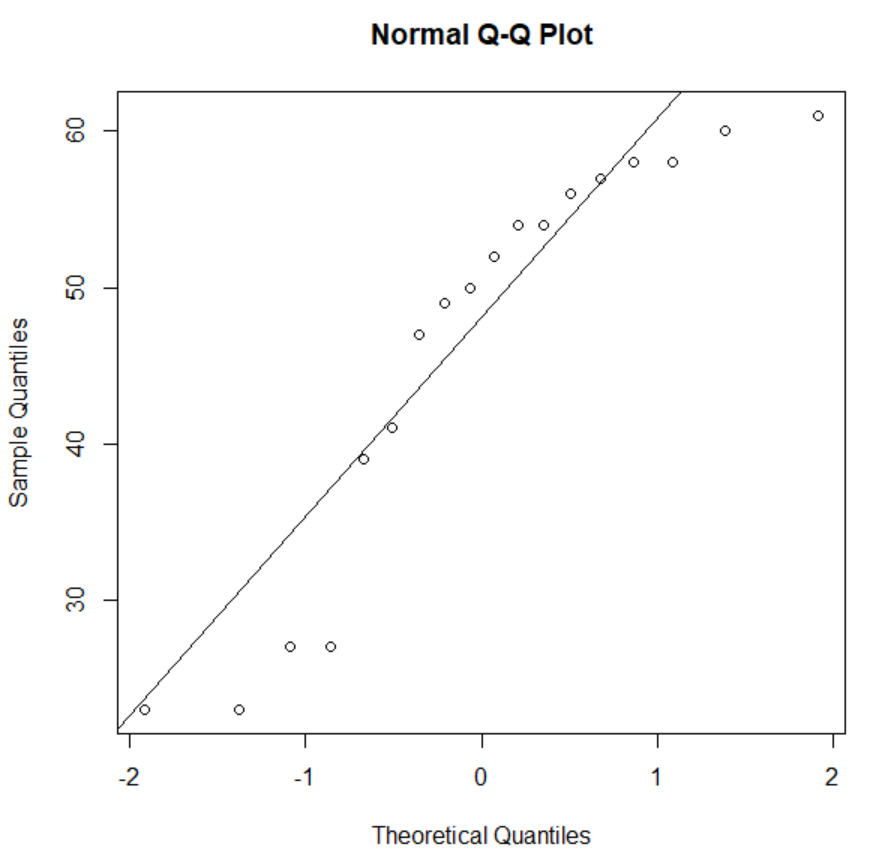


SCATTER PLOT-

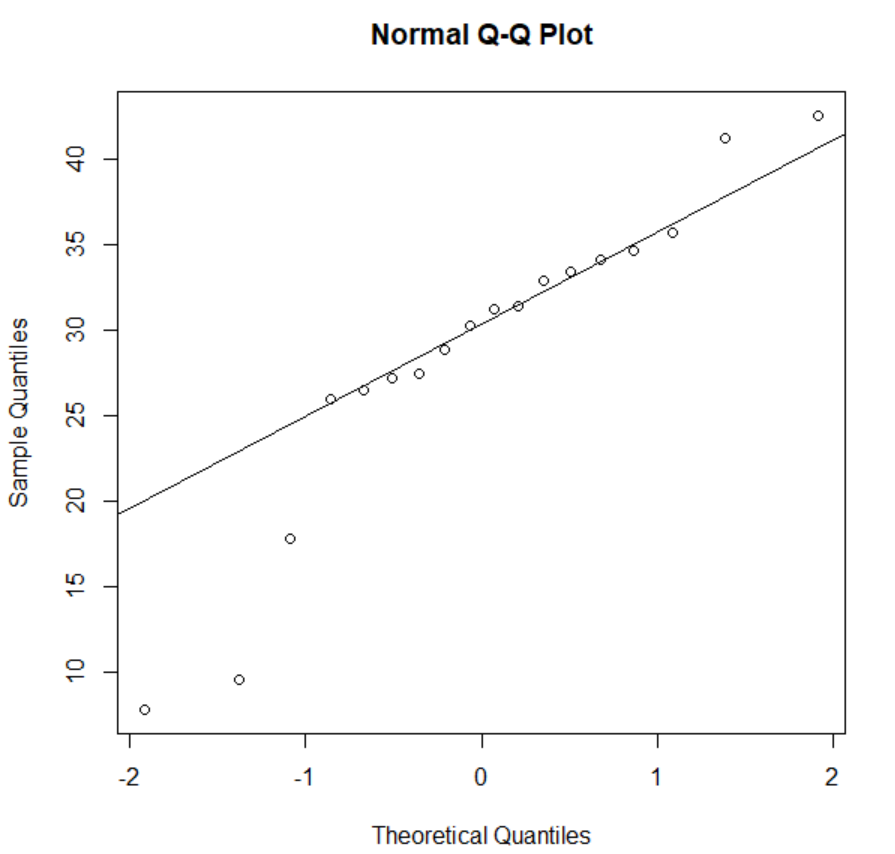
**#5c**

QQ 

QQ PLOT FOR AGE-



QQPLOT FOR BODY FAT PERCENT-



6.Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

(i) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].  
(ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.  
(iii) Use normalization by decimal scaling to transform the value 35 for age. Perform the above functions using R – tool

**CODING-**

age <- c(23,23,27,27,39,41,47,49,50,52,54,54,56,57,58,58,60,61)

new\_age<-c()

for(i in age){

if(i<=35){

new\_age=append(new\_age,i)

}

}

print(new\_age)

**#6a**

#min max normalization

min<-min(new\_age)

max<-max(new\_age)

for (i in new\_age)

{

result1=i-min

result2=max-min

result3=result1/result2

print(result3)

}

**#6b**

#z score normalization

mean1<-mean(new\_age)

for (i in new\_age)

{

result1=i-mean1

result2=result1/12.94

print(result2)

}

**#6c**

#decimal scaling

n=200

j=nchar(y)

scaling=n/10^j

print(scaling)

**OUTPUT-**

6.a MIN MAX NORMALIZATION

[1] 0

[1] 0

[1] 1

[1] 1

6.b Z SCORE NORMALIZATION

[1] -0.8660254

[1] -0.8660254

[1] 0.8660254

[1] 0.8660254

6.c DECIMAL SCALING

[1] 0.2

7.The following values are the number of pencils available in the different boxes. Create a vector and find out the mean, median and mode values of set of pencils in the given data.

Box1 Box2 Box3 Box4 Box5 Box6 Box7 Box8 Box9 Box 10

9 25 23 12 11 6 7 8 9 10

**CODING-**

box\_no=c("box1","box2","box3","box4","box5","box6","box7","box8","box9","box10")

pencil=c(9,25,23,12,11,6,7,8,9,10)

df<-data.frame(box\_no,pencil)

#dataframe

print(df)

#mean

mean(pencil)

#median

median(pencil)

#mode

mode=names(which.max(table(pencil)))

print(mode)

**OUTPUT-**

> data.frame(box\_NO,pencil)

box\_NO pencil

1 box1 9

2 box2 25

3 box3 23

4 box4 12

5 box5 11

6 box6 6

7 box7 7

8 box8 8

9 box9 9

10 box10 10

> mean(pencil)

[1] 12

> median(pencil)

[1] 9.5

> print(mode)

[1] "9"

8. the following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones sold.

x :4 1 5 7 10 2 50 25 90 36

y :12 5 13 19 31 7 153 72 275 110

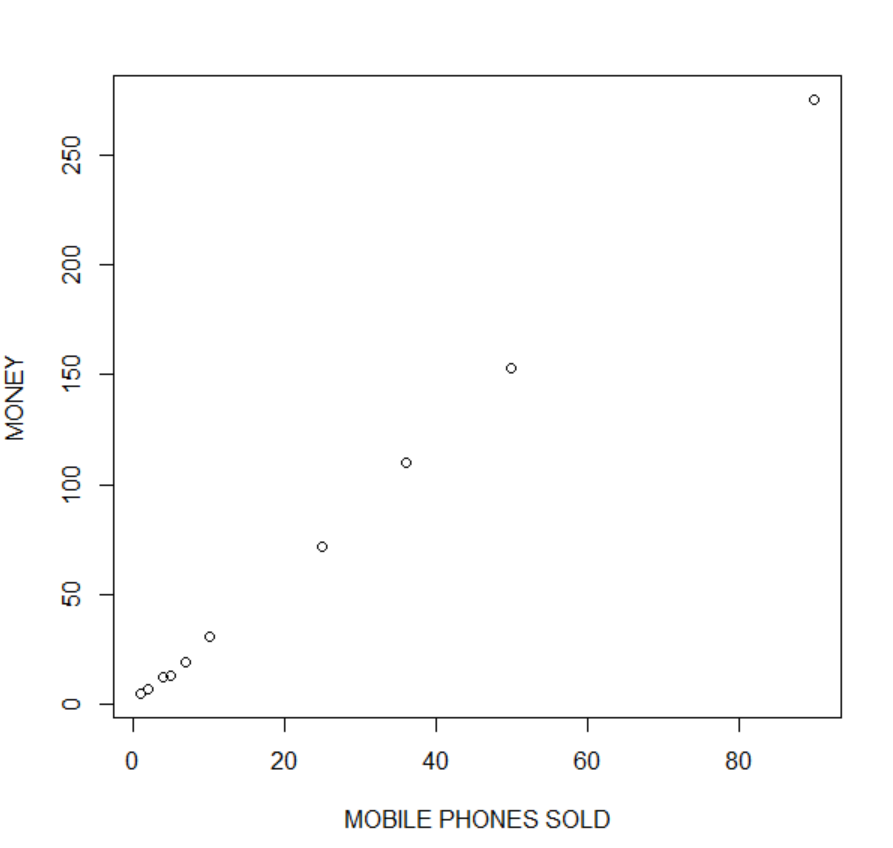
CODING-

x<-c(4, 1, 5, 7, 10, 2, 50, 25, 90, 36)

y<-c(12,5, 13, 19, 31, 7, 153, 72, 275, 110)

plot(x,y,xlab='MOBILE PHONES SOLD',ylab='MONEY')

OUTPUT-



9. Implement of the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75. Partition them into three bins by each of the following methods. Plot the data points using histogram.

(a) equal-frequency (equi-depth) partitioning (b) equal-width partitioning

CODING-

marks<-c(55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75)

binning1=c()

binning2=c()

binning3=c()

class=6

#binning partition

for(a in marks[1:class]){

binning1=append(binning1,a)

}

range1=range+1

range2=range\*2

for(b in marks[range1:range2])

{

binning2=append(binning2,b)

}

range3=range2+1

range4=range\*3

for(c in marks[range3:range4])

{

binning3=append(binning3,c)

}

print(binning1)

print(binning2)

print(binning3)

**#histogram**

hist(binning1)

hist(binning2)

hist(binning3)

**#9a**

**#equal-frequency**

freq=length(marks)/range

print(freq)

**#9b**

**#equal-width**

min<-min(marks)

max<-max(marks)

result<-max-min

width<-result/range

cat("width is",width)

bin1=width+min

print(bin1)

bin2=2\*width+min

print(bin2)

bin3=3\*width+min

print(bin3)

OUTPUT-

> print(binning1)

[1] 55 60 71 63 55 65

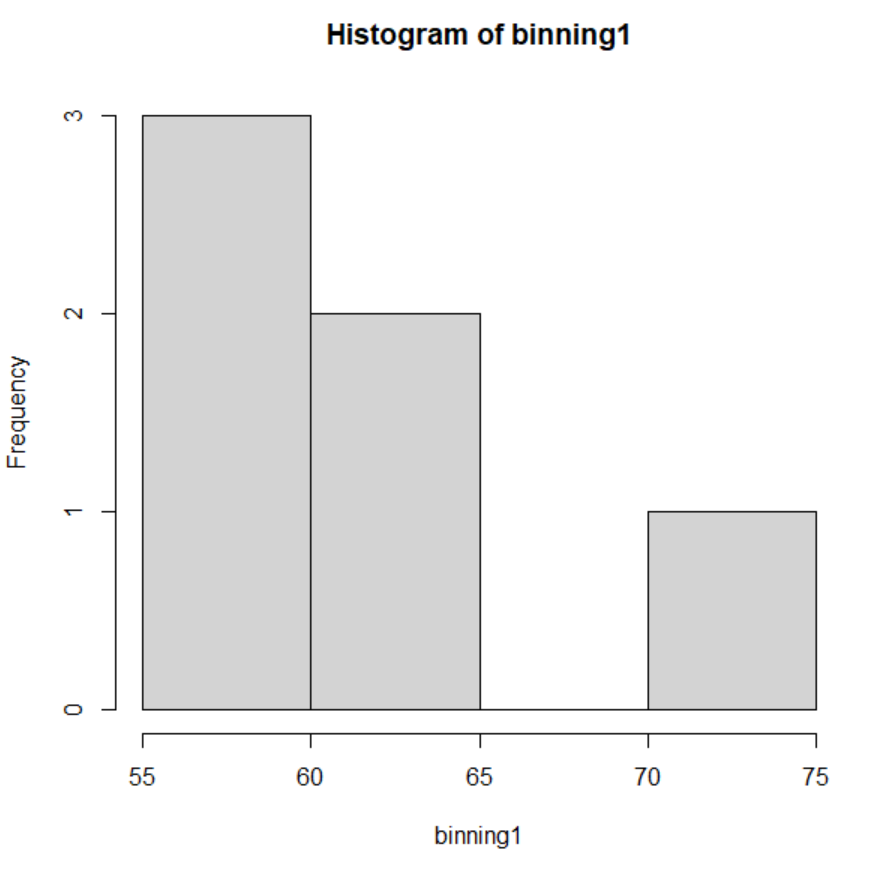
> print(binning2)

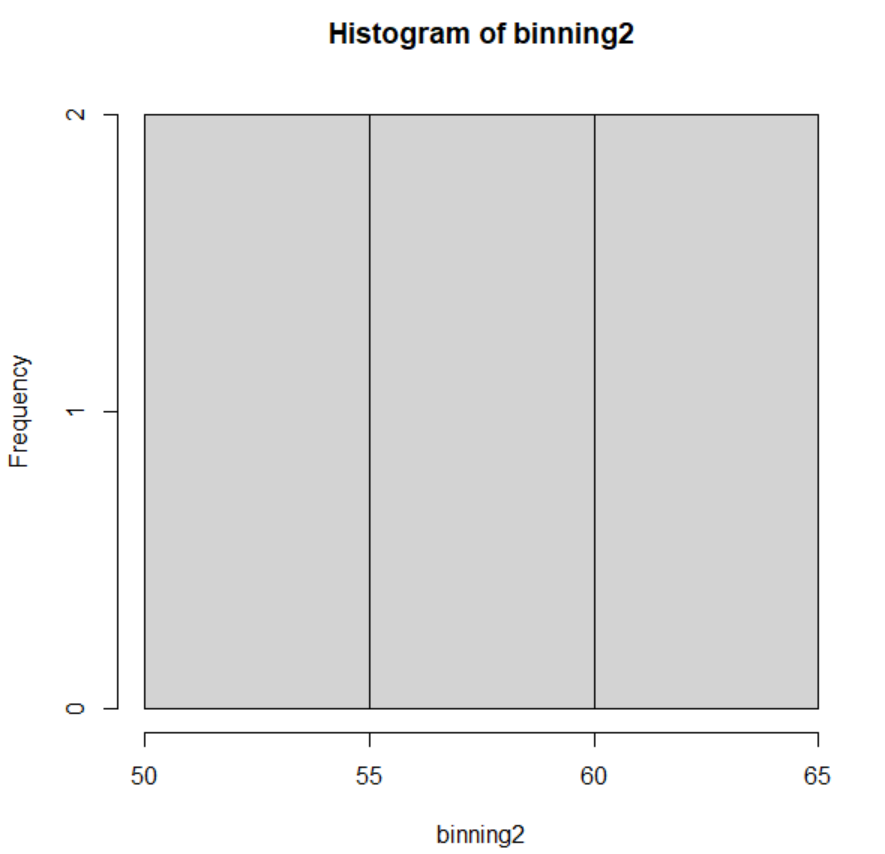
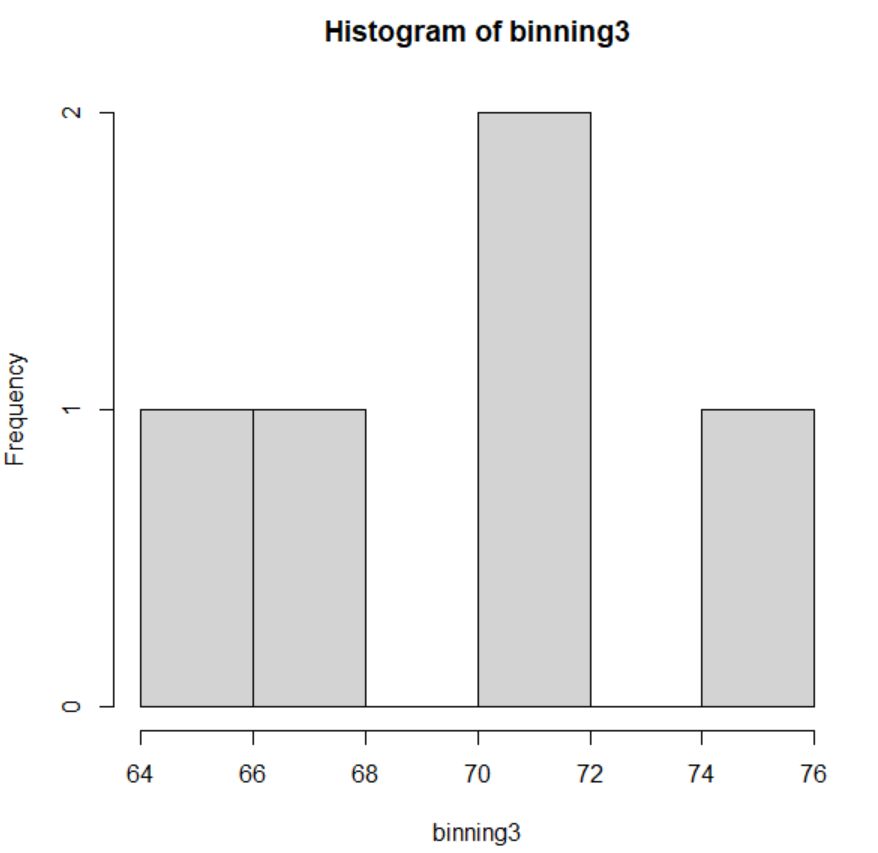
[1] 50 55 58 59 61 63

> print(binning3)

[1] 65 67 71 72 75 NA

HISTOGRAM-





#9a

#equal frequency

> print(freq)

[1] 2.833333

**#9b**

#equal width

width is 4.166667> bin1=width+min

> print(bin1)

[1] 54.16667

> bin2=2\*width+min

> print(bin2)

[1] 58.33333

> bin3=3\*width+min

> print(bin3)

[1] 62.5

10. Suppose that the speed car is mentioned in different driving style.

Regular 78.3 81.8 82 74.2 83.4 84.5 82.9 77.5 80.9 70.6 Speed

Calculate the Inter quantile and standard deviation of the given data.

CODING-

speed<-c(78.3 ,81.8 ,82 ,74.2 ,83.4 ,84.5 ,82.9 ,77.5 ,80.9 ,70.6 )

#interquartile

IQR(speed)

#standard deviation

sd(speed)

OUTPUT-

> IQR(speed)

[1] 4.975

> sd(speed)

[1] 4.445835

11.Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

CODING-x

marks<-c(13,15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

quantile(marks)

OUTPUT-

> quantile(marks)

0% 25% 50% 75% 100%

13.0 20.5 25.0 35.0 70.0

12.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

(i)Use cov() to calculate the sample covariance between B and C.

(ii)Use another call to cov() to calculate the sample covariance matrix for the preferences.

(iii)Use cor() to calculate the sample correlation between B and C.

(iv)Use another call to cor() to calculate the sample correlation matrix for the preferences.

CODE:

(i)b<-c(22, 28, 10)

c<-c(20, 40, 40)

cov(b,c)

(ii)a<-c(18, 2, 20)

b<-c(22, 28, 10)

c<-c(20, 40, 40)

pre<-cbind(a,b,c)

cov(pre)

(iii).b<-c(22, 28, 10)

c<-c(20, 40, 40)

cor(b,c)

(iv)a<-c(18, 2, 20)

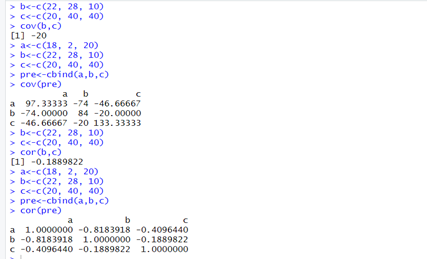
b<-c(22, 28, 10)

c<-c(20, 40, 40)

pre<-cbind(a,b,c)

cor(pre)

OUTPUT:



13.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,

|  |
| --- |
| 18, 18, 18, 20, 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 |

CODE:

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, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30)

bin<-length(data)/3

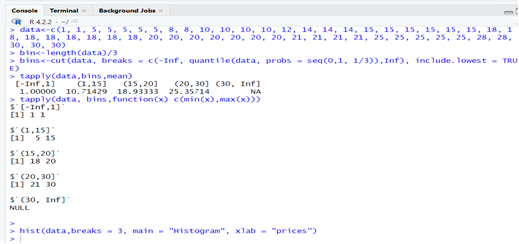
bins<-cut(data, breaks = c(-Inf, quantile(data, probs = seq(0,1, 1/3)),Inf), include.lowest = TRUE)

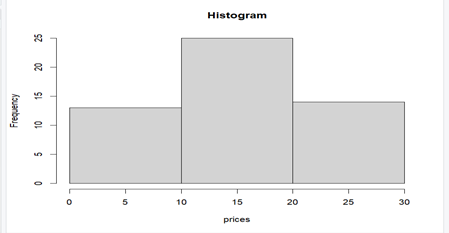
tapply(data,bins,mean)

tapply(data, bins,function(x) c(min(x),max(x)))

hist(data,breaks = 3, main = "Histogram", xlab = "prices")

OUTPUT:





14.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

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

CODE:

A <- c(76, 35, 47, 64, 95, 66, 89, 36, 84)

B <- c(51, 56, 84, 60, 59, 70, 63, 66, 50)

mean\_A <- mean(A)

median\_A <- median(A)

range\_A <- max(A) - min(A)

mean\_B <- mean(B)

median\_B <- median(B)

range\_B <- max(B) - min(B)

combined\_data <- data.frame(Class = c(rep("A", length(A)), rep("B", length(B))), Score = c(A, B))

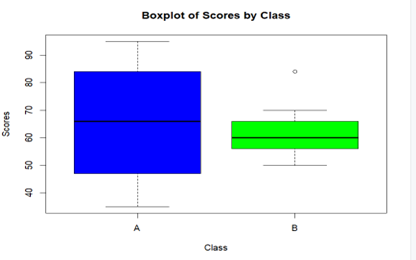
boxplot(Score ~ Class, data = combined\_data, col = c("blue", "green"), xlab = "Class", ylab = "Scores", main = "Boxplot of Scores by Class")

(II)

combined\_data <- data.frame(Class = c(rep("A", length(A)), rep("B", length(B))), Score = c(A, B))

boxplot(Score ~ Class, data = combined\_data, col = c("blue", "green"), xlab = "Class", ylab = "Scores", main = "Boxplot of Scores by Class")

OUTPUT:



15.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

CODE:

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

min\_value <- 50000

max\_value <- 100000

v <- 80

min\_max\_normalized <- (v - min\_value) / (max\_value - min\_value)

min\_max\_normalized

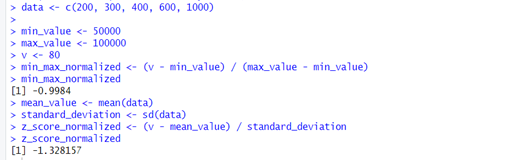
mean\_value <- mean(data)

standard\_deviation <- sd(data)

z\_score\_normalized <- (v - mean\_value) / standard\_deviation

z\_score\_normalized

OUTPUT:



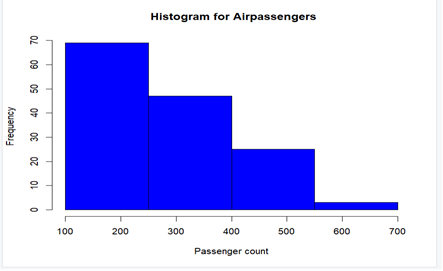
16.Make a histogram for the “AirPassengers “dataset, start at 100 on the x-axis, and from values 200 to 700, make the bins 150 wide

CODE:

data("AirPassengers")

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

OUTPUT:



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

CODE:

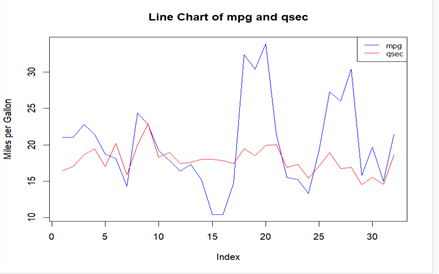
data("mtcars")

plot(mtcars$mpg, type = "l", col = "blue", xlab = "Index", ylab = "Miles per Gallon", main = "Line Chart of mpg and qsec")

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

legend("topright", legend = c("mpg", "qsec"), col = c("blue", "red"), lty = 1, cex = 0.8)

OUTPUT:



18.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.

CODE:

data("iris")

str(iris)

plot(iris$Sepal.Length, iris$Petal.Length, main = "Scatter plot of Sepal.Length vs. Petal.Length",xlab = "Sepal.Length", ylab = "Petal.Length", col = "blue", pch = 16)

model <- lm(Petal.Length ~ Sepal.Length, data = iris)

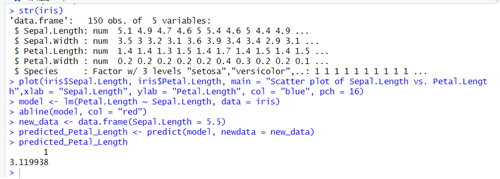
abline(model, col = "red")

new\_data <- data.frame(Sepal.Length = 5.5)

predicted\_Petal\_Length <- predict(model, newdata = new\_data)

predicted\_Petal\_Length

OUTPUT:



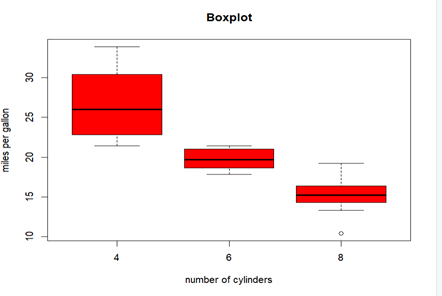
19.Create a Boxplot graph for the relation between "mpg"(miles per galloon) and "cyl"(number of Cylinders) for the dataset "mtcars" available in R Environment.

CODE:

data("mtcars")

boxplot(mpg ~ cyl, data = mtcars, main = "Boxplot", xlab = "number of cylinders", ylab = "miles per gallon", col= "red")

OUTPUT:



20. 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

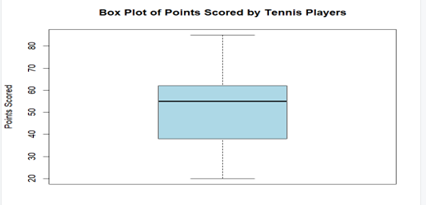
technique.

CODE:

score <- c(20, 25, 30, 32, 35, 38, 40, 45, 50, 52, 55, 56, 58, 59, 60, 62, 65, 70, 75, 80, 85)

boxplot(score, col = "lightblue", main = "Box Plot of Points Scored by Tennis Players", ylab = "Points Scored")

OUTPUT:



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

CODE: w

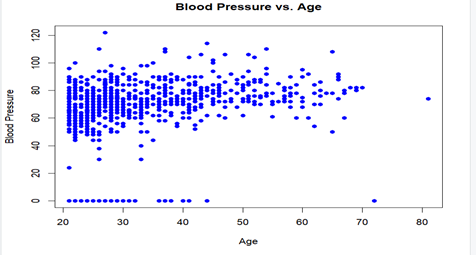
dia<-read.csv("C:/Users/haris/Downloads/diabetes.csv")

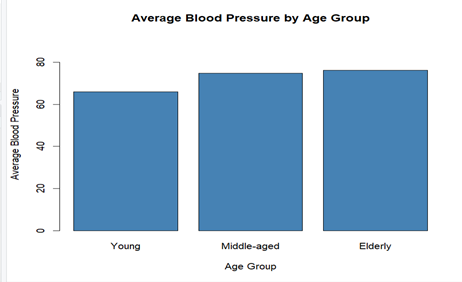
View(dia)

plot(dia$Age, dia$BloodPressure, xlab = "Age", ylab = "Blood Pressure", main = "Blood Pressure vs. Age", col = "blue",pch = 16)

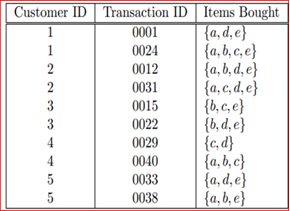
barplot(dia$Age,dia$Blood\_Pressure)

OUTPUT:





22.Consider the data set and perform the Apriori Algorithm and FP algorithm support:3 and confidence=50%



Input:

@relation dataset

@attribute a{true,false}

@attribute b{true,false}

@attribute c{true,false}

@attribute d{true,false}

@attribute e{true,false}

@data

true false false true true

true true true false true

true true false true true

true false true true true

false true true false true

false true false true true

false false true true false

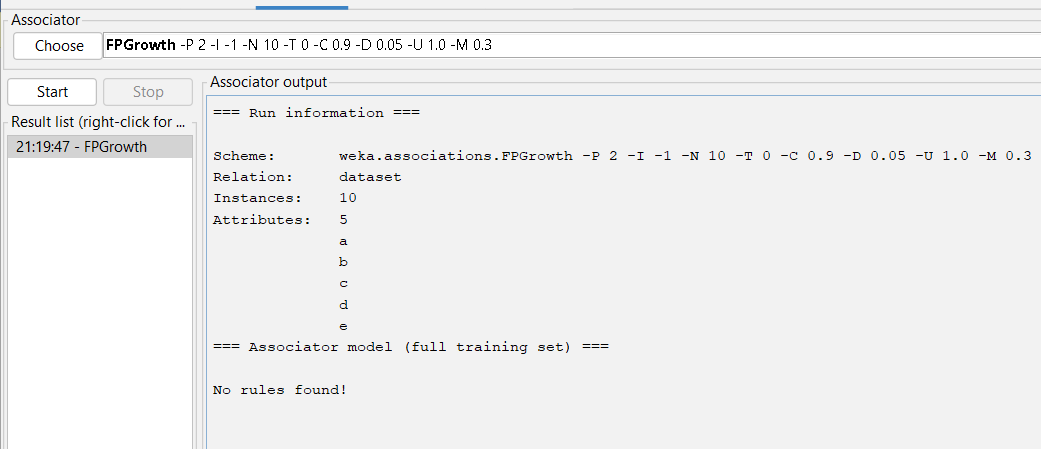
true true true false false

true false false true true

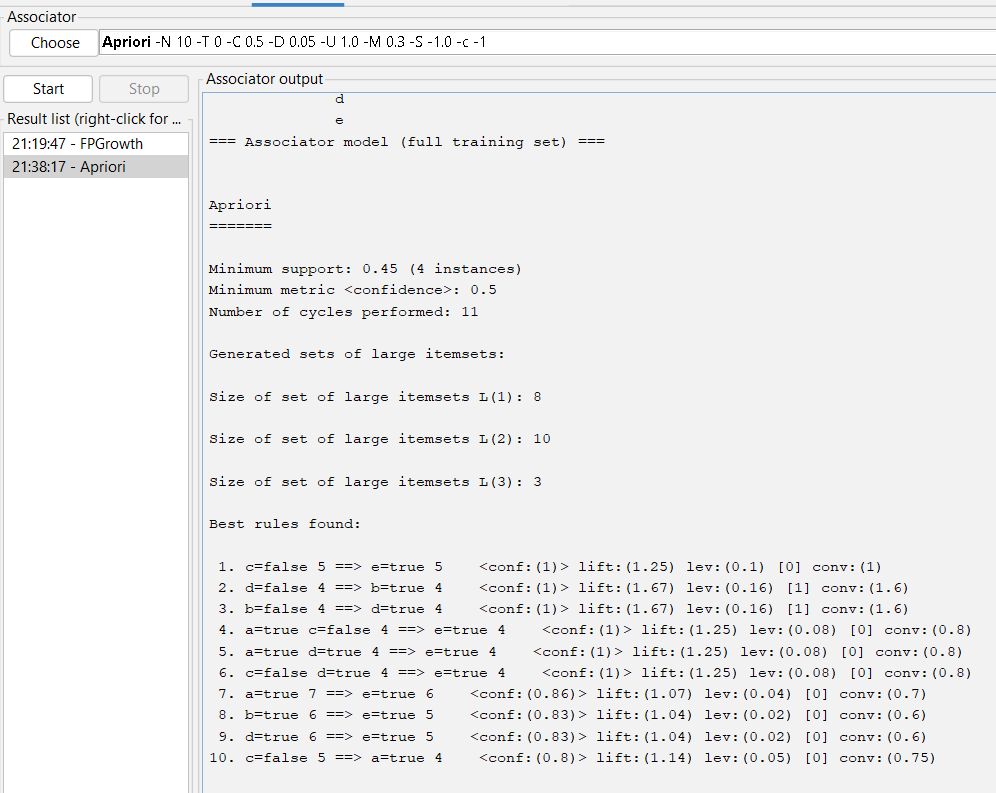
true true false false true

output:

FPGROWTH:



APRIORI ALGORITHM:



23.Consider the data set and perform the Apriori Algorithm and FP algorithm support:3 and confidence=50%

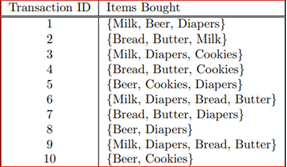
Consider the market basket transactions shown in the above table.

(a) What is the maximum number of association rules that can be extracted

from this data (including rules that have zero support)?

(b) What is the maximum size of frequent itemsets that can be extracted

(assuming minsup > 0)?



@relation transactions

@attribute Milk {0,1}

@attribute Beer {0,1}

@attribute Diapers {0,1}

@attribute Bread {0,1}

@attribute Butter {0,1}

@attribute Cookies {0,1}

@data

1,1,1,0,0,0

0,0,0,1,1,1

1,0,1,0,0,1

0,0,0,1,1,1

0,1,1,0,0,1

1,0,1,1,1,0

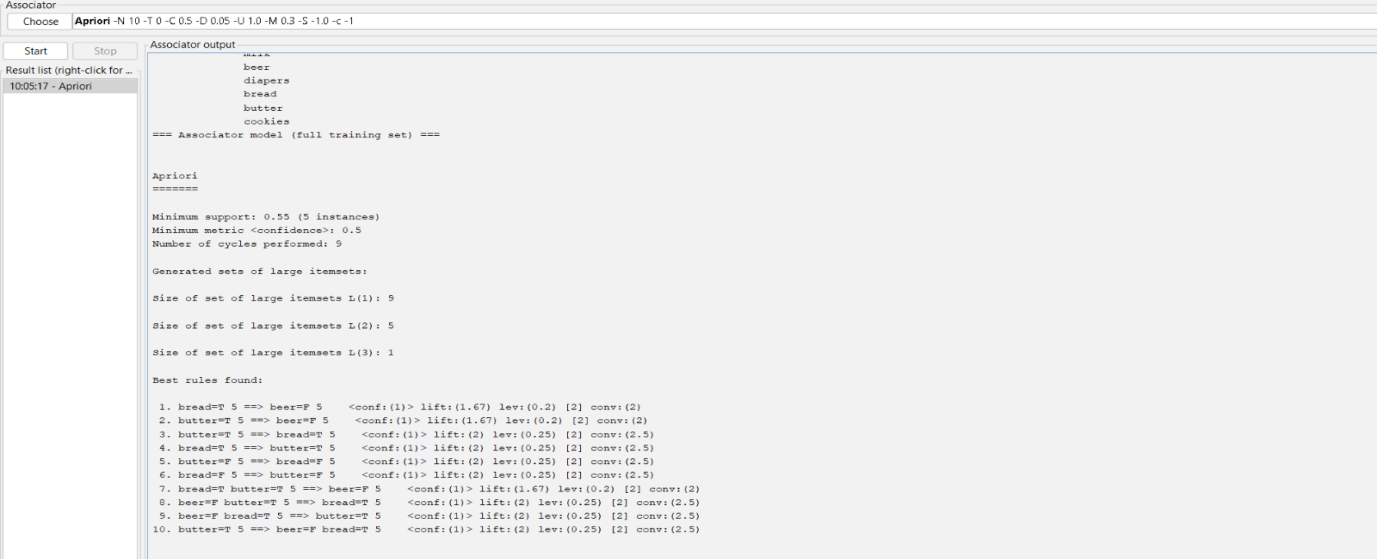
0,0,1,1,1,0

0,1,1,0,0,0

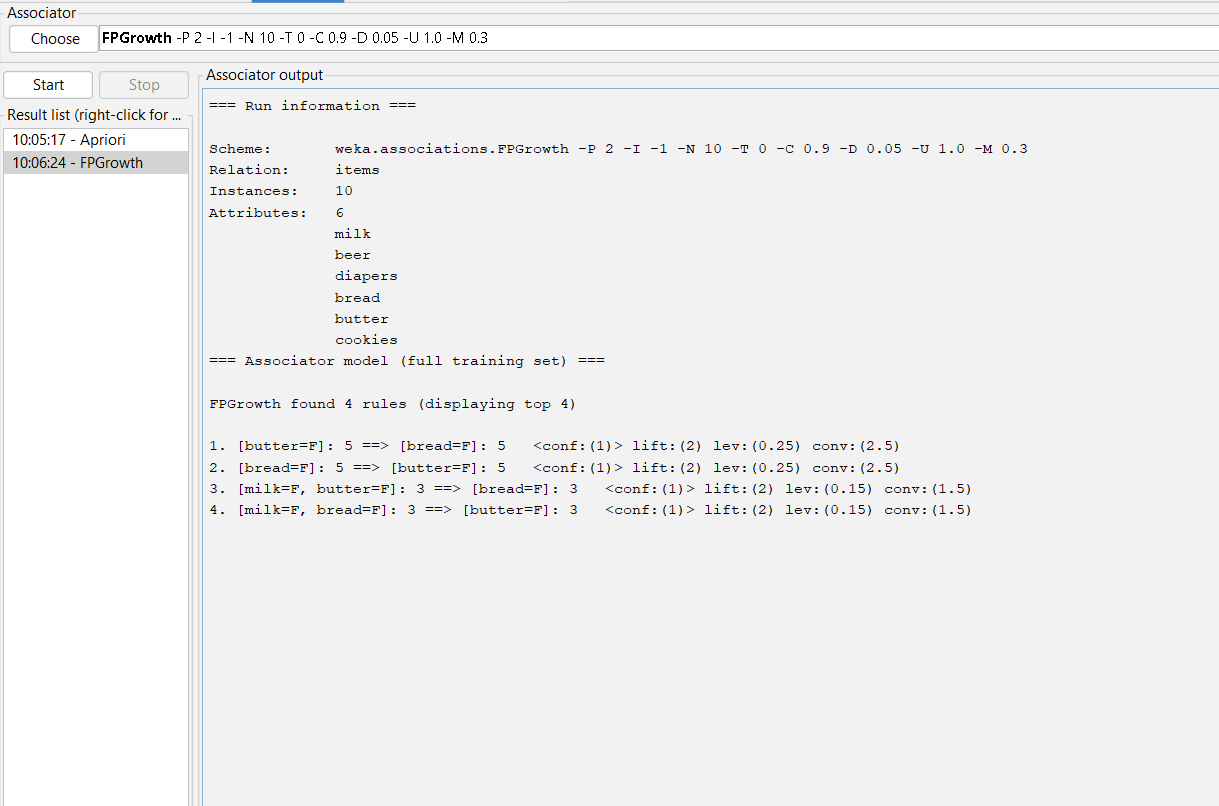
1,0,1,1,1,0

0,1,0,0,0,1

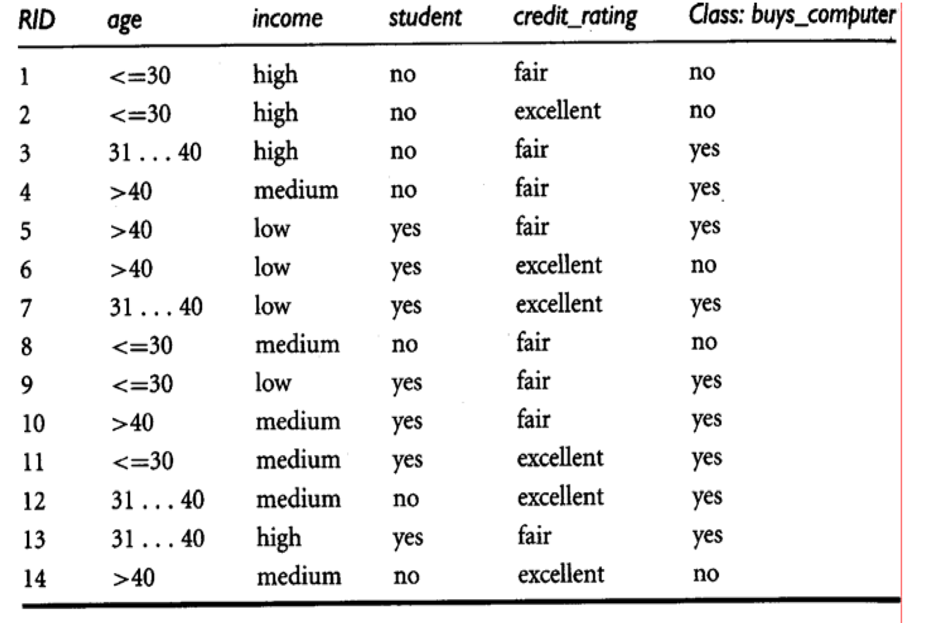
Apriori algorithm:



Fp growth algorithm:



24.Bayes classification and descion tree (using training and test data)



Input:

@relation decision\_tree

@attribute age{young,middle,old}

@attribute income{low,medium,high}

@attribute student{yes,no}

@attribute Creit\_rating{fair,excellent}

@attribute class{yes,no}

@data

young high no fair no

young high no excellent no

middle high no fair yes

old medium no fair yes

old low yes fair yes

old low yes excellent no

middle low yes excellent yes

young medium no fair no

young low yes fair yes

old medium yes fair yes

young medium yes excellent yes

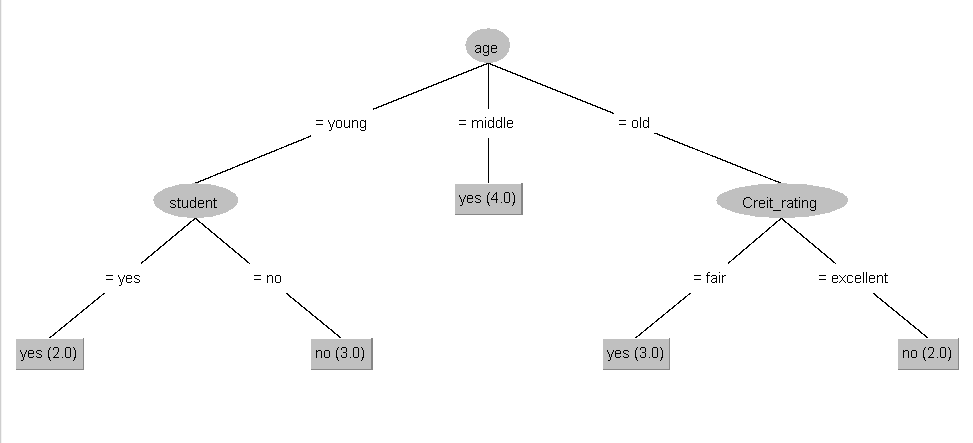
middle medium no excellent yes

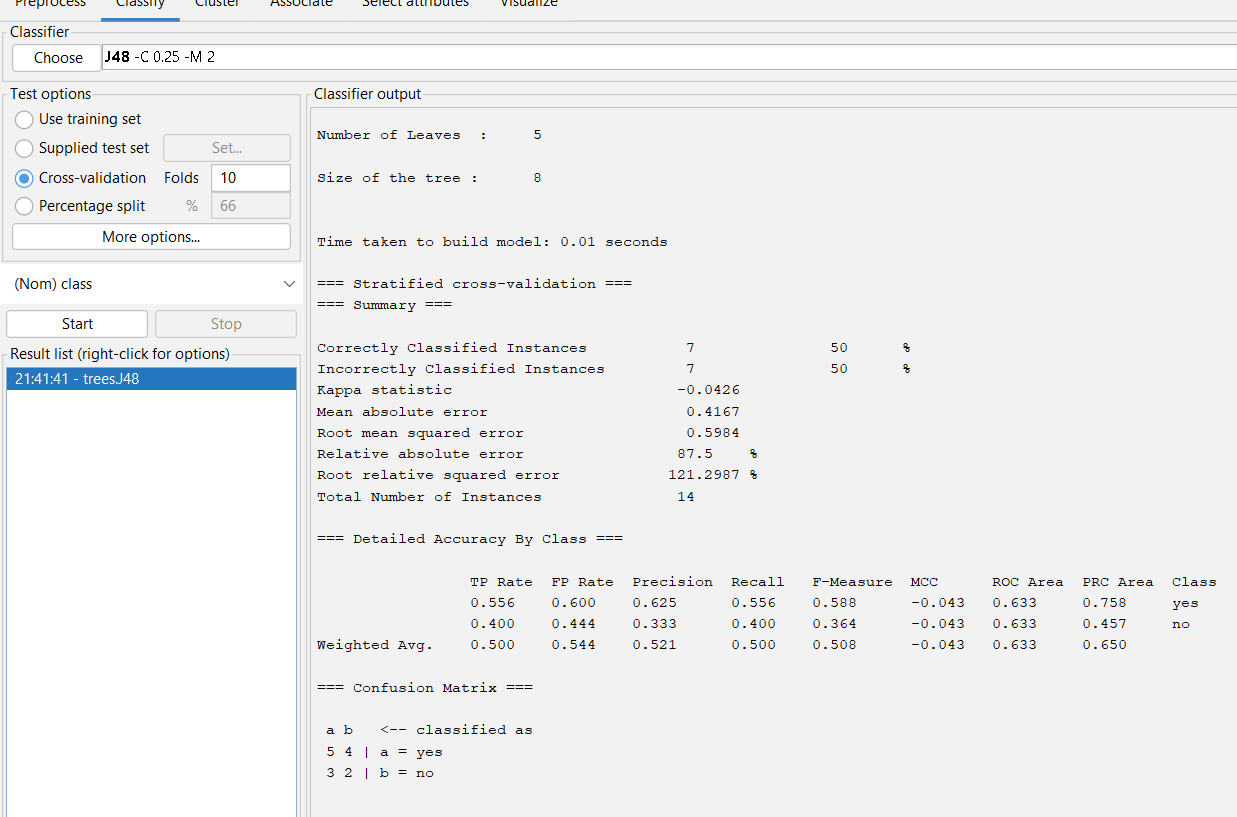
middle high yes fair yes

old medium no  excellent no

output:

tree:





25.Analysis the dataset “diabetes. csv” how the diabetes trend is for different age people, using linear regression and multiple regression.

Input:

data<-read.csv("C:/Users/Hari Naidu/Desktop/POM/download papers/diabetes.csv")

data

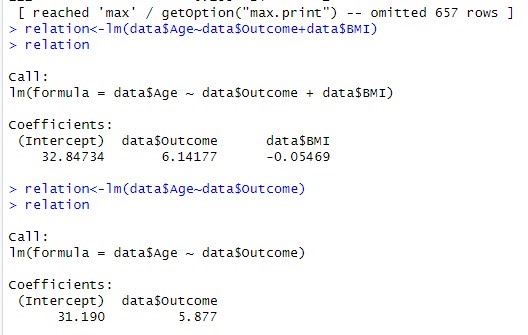
relation<-lm(data$Age~data$Outcome)

relation

relation<-lm(data$Age~data$Outcome+data$BMI)

relation

output:



26.Implement using WEKA for the given Suppose a database has five transactions. Let min sup= 50%(2) and min con f = 80%.

**Transactions** **Items**

T1 (M, O, N, K, E, Y)

T2 (D, O, N, K, E, Y)

T3 (M, A, K, E)

T4 (M, U, C, K, Y)

T5 (C,O, O, K, I ,E)

* Find all frequent item sets using Apriori algorithm
* Also draw FP-Growth Tree

@relation transactions

@attribute M {0,1}

@attribute O {0,1}

@attribute N {0,1}

@attribute K {0,1}

@attribute E {0,1}

@attribute Y {0,1}

@attribute D {0,1}

@attribute A {0,1}

@attribute U {0,1}

@attribute C {0,1}

@attribute I {0,1}

@data

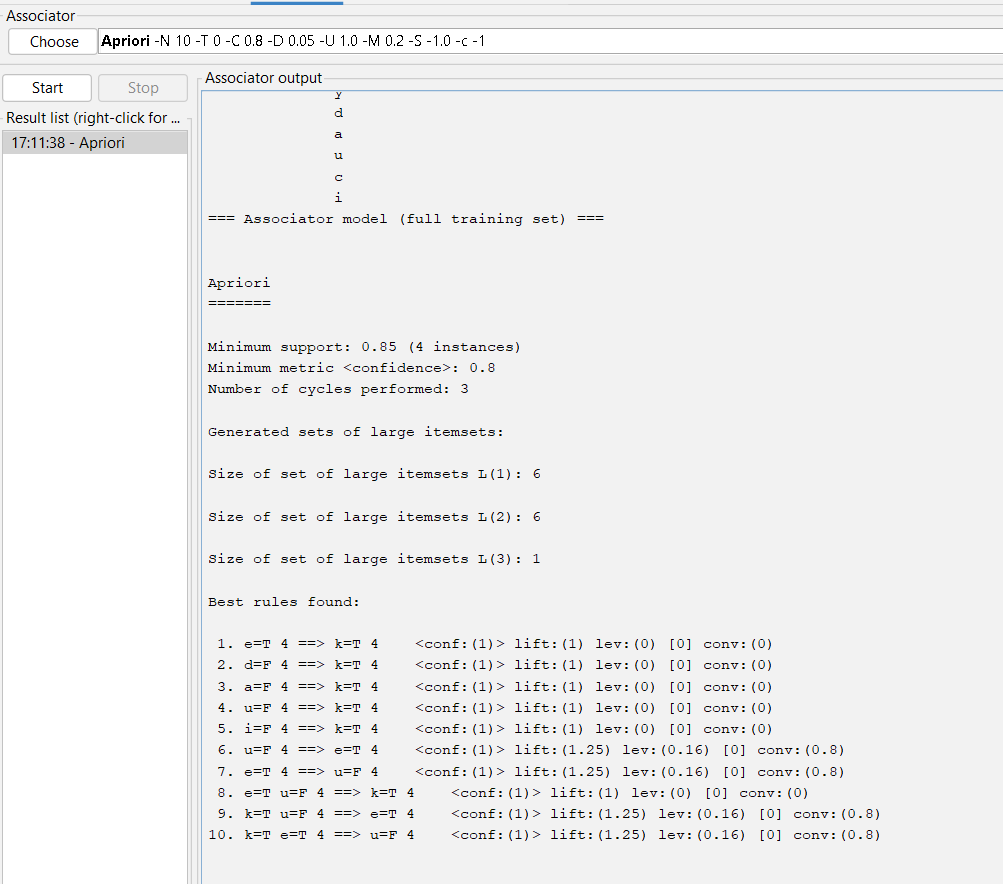
1,1,1,1,1,1,0,0,0,0,0

0,1,1,1,1,1,1,0,0,0,0

1,0,0,1,1,0,0,1,0,0,0

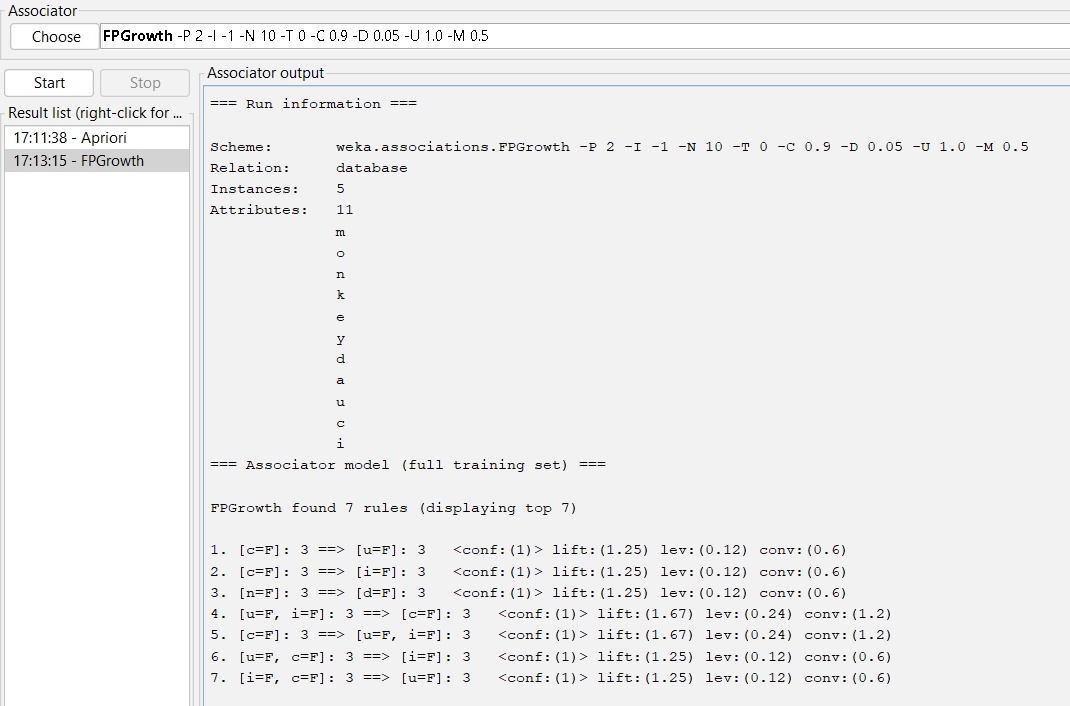
1,0,0,1,0,1,0,0,1,0,0

0,1,0,1,1,0,0,0,0,1,1

Input:

Apriori algorithm:

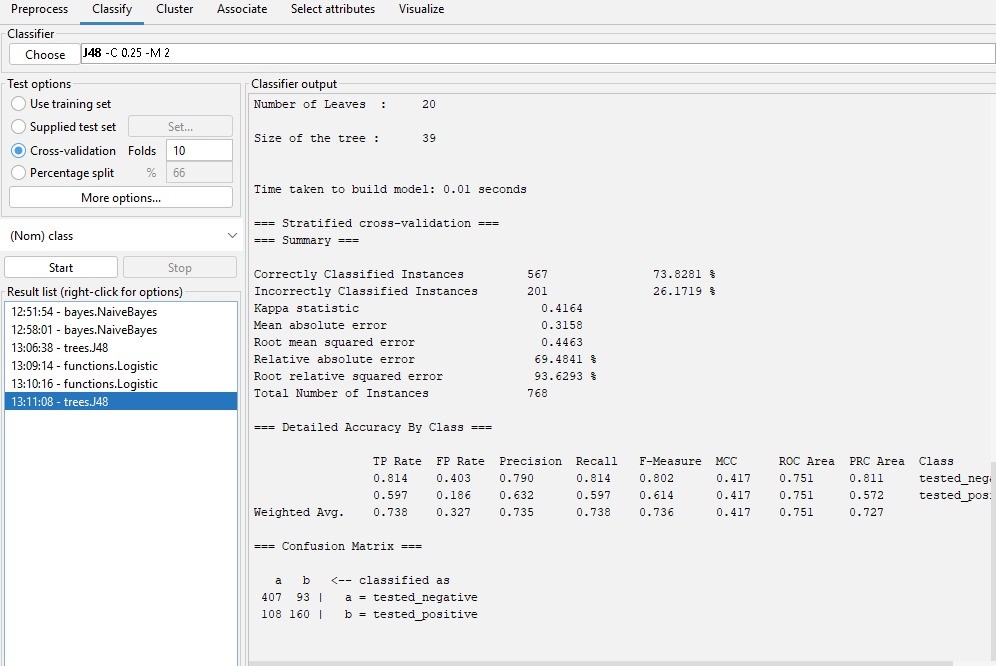
Fpgrowth algorithm:



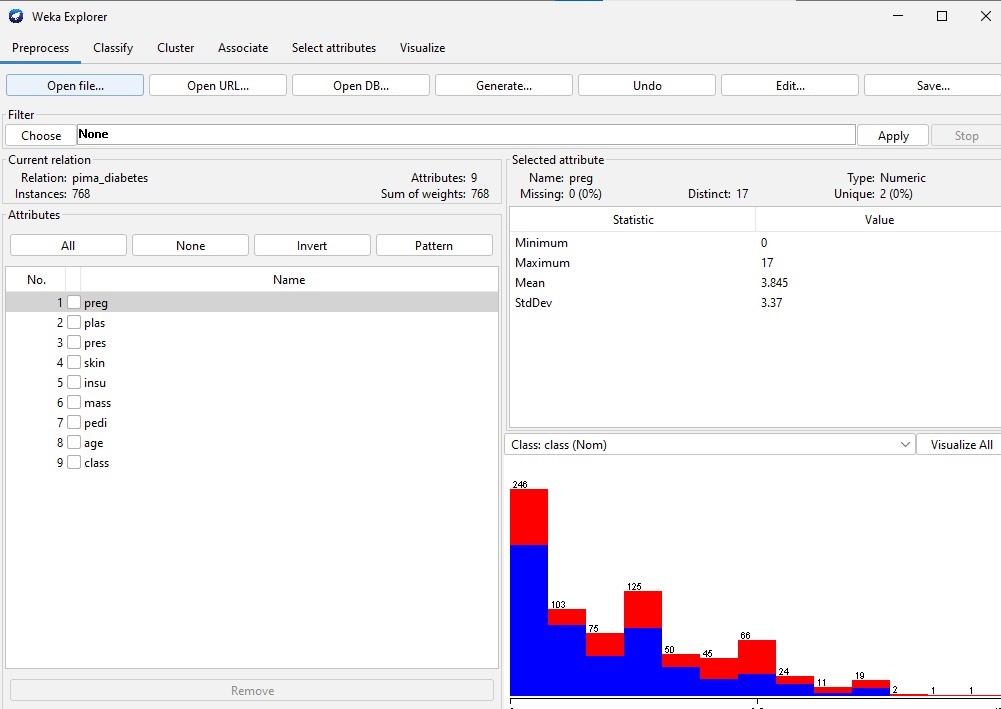
27. Prediction of Categorical Data using Decision Tree Algorithm through WEKA using any datasets. a) Tree b) Preprocess c) Logistic

Output:

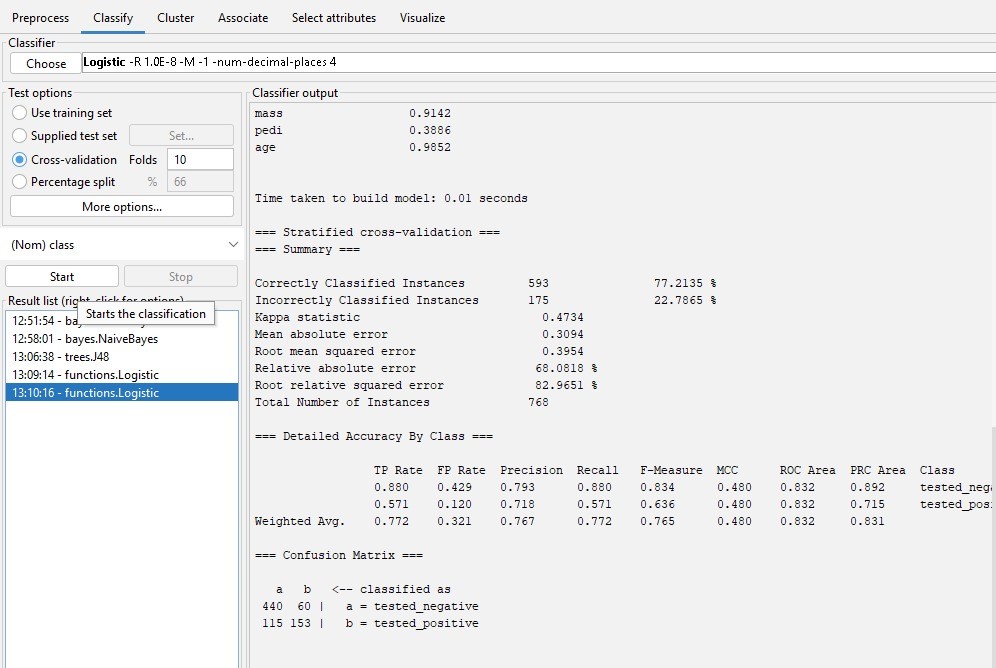
Tree:

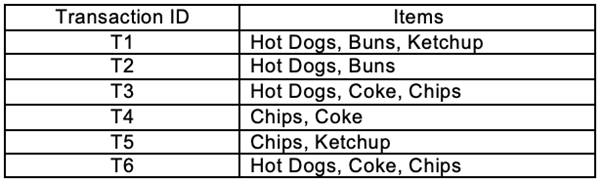


Preprocessor:



Logistic:



28.Create the dataset using ARFF file format:

a.Find the **frequent itemsets** and generate **association rules** on this. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%).

b.List the various rule generated by apriori and FP tree algorthim ,mention wheather accepted or rejcted.

Input:

@relation hotdogs

@attribute hotdogs{t,f}

@attribute buns{t,f}

@attribute ketchup{t,f}

@attribute coke{t,f}

@attribute chips{t,f}

@data

t t t f f

t t f f f

t f f t t

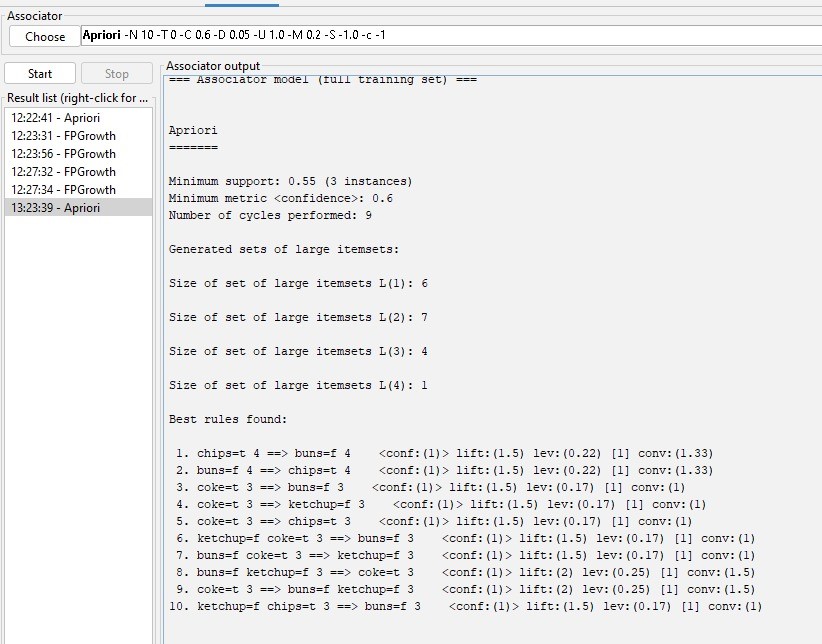
f f f t t

f f t f t

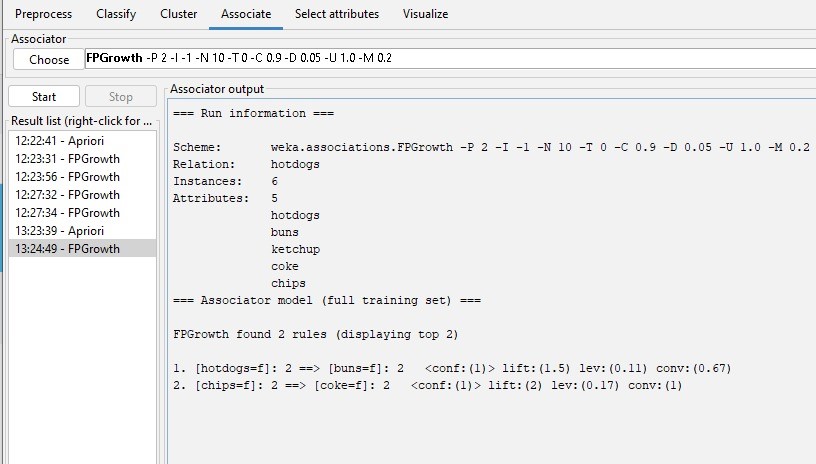
t f f t t

output:

apriori algorithm:



Fp growth:



29.Prediction of Categorical Data using Rule base classification and decision tree classification through WEKA using any datasets. Compare the accuracy using two algorithm and plot the graph

Input:

@relation decision\_tree

@attribute age{young,middle,old}

@attribute income{low,medium,high}

@attribute student{yes,no}

@attribute Creit\_rating{fair,excellent}

@attribute class{yes,no}

@data

young high no fair no

young high no excellent no

middle high no fair yes

old medium no fair yes

old low yes fair yes

old low yes excellent no

middle low yes excellent yes

young medium no fair no

young low yes fair yes

old medium yes fair yes

young medium yes excellent yes

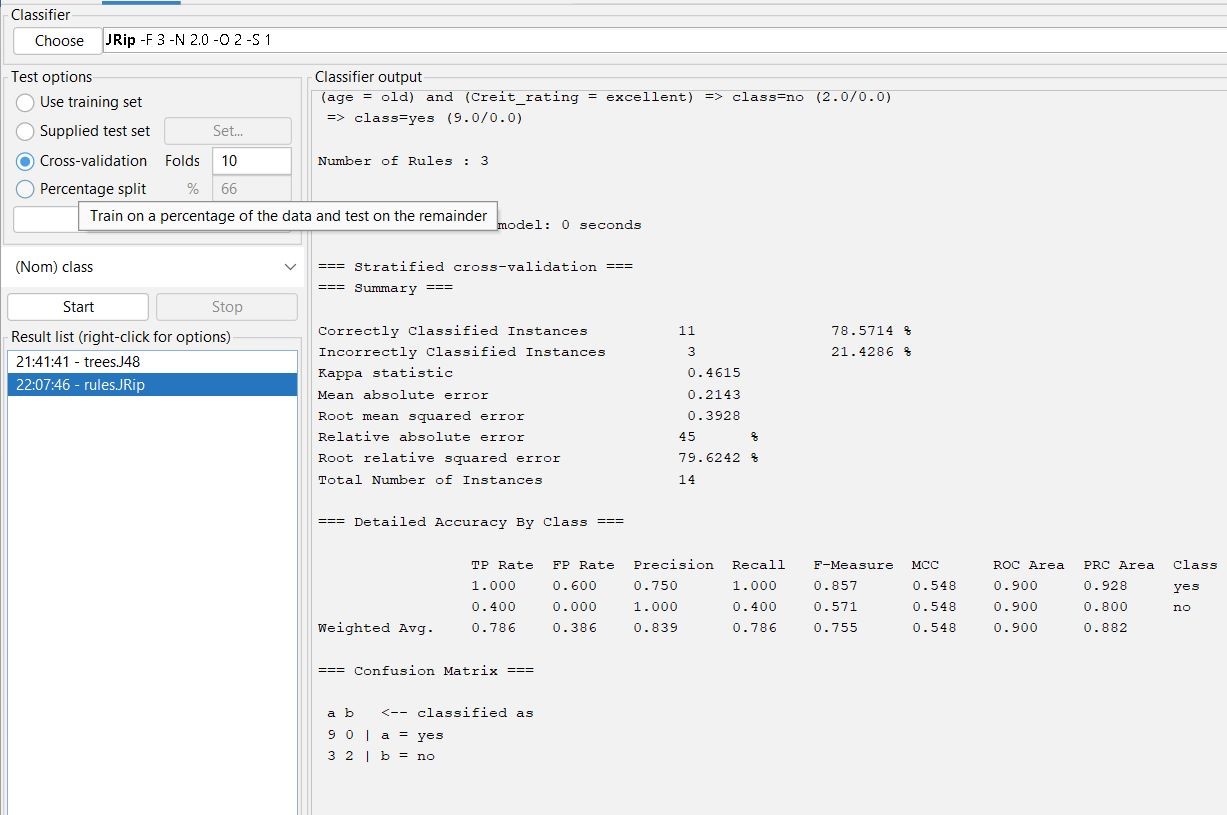
middle medium no excellent yes

middle high yes fair yes

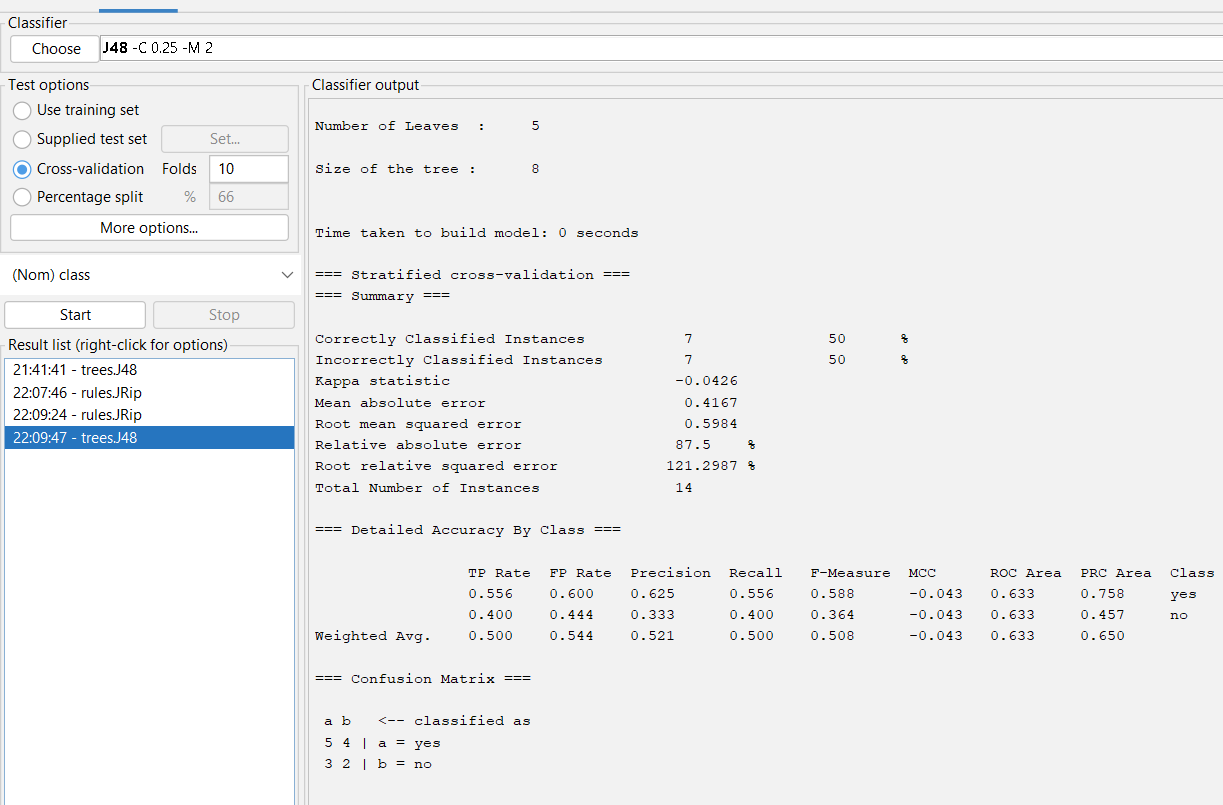
old medium no  excellent no

Output:

Rule based classification:



Decision tree:



30)Implement of the R script using a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215. Partition them into three bins by each of the following methods.

* + 1. equal-frequency (equi depth) partitioning
    2. equal-width partitioning
    3. clustering

CODE:

# Sales price data

sales\_prices <- c(5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215)

# (a) Equal-Frequency (Equi-depth) Partitioning

bins\_eq\_freq <- split(sales\_prices, ceiling(seq\_along(sales\_prices) / (length(sales\_prices) / 3)))

print("Equal-Frequency Bins:")

print(bins\_eq\_freq)

# (b) Equal-Width Partitioning

min\_val <- min(sales\_prices)

max\_val <- max(sales\_prices)

num\_bins <- 3

bin\_width <- (max\_val - min\_val) / num\_bins

bin\_edges <- seq(min\_val, max\_val + bin\_width, by=bin\_width)

bin\_labels <- cut(sales\_prices, breaks=bin\_edges, include.lowest=TRUE)

bins\_eq\_width <- split(sales\_prices, bin\_labels)

print("Equal-Width Bins:")

print(bins\_eq\_width)

# (c) Clustering (K-Means)

set.seed(42)

kmeans\_result <- kmeans(matrix(sales\_prices, ncol=1), centers=num\_bins)

cluster\_bins <- split(sales\_prices, kmeans\_result$cluster)

print("Clustering Bins (K-Means):")

print(cluster\_bins)

# Histograms

par(mfrow=c(1,3)) # Arrange plots in 1 row, 3 columns

hist(sales\_prices, breaks=bin\_edges, col="blue", main="Equal-Width Partitioning",

xlab="Sales Prices", ylab="Frequency")

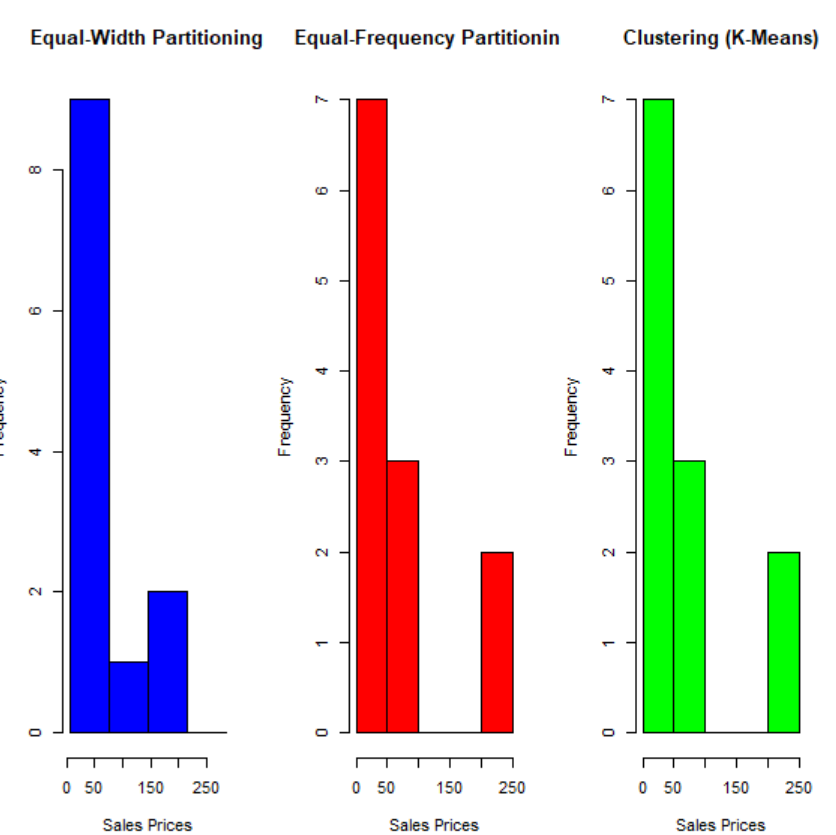
hist(sales\_prices, breaks=length(sales\_prices)/3, col="red", main="Equal-Frequency Partitioning",

xlab="Sales Prices", ylab="Frequency")

hist(sales\_prices, col="green", main="Clustering (K-Means)", xlab="Sales Prices", ylab="Frequency")

par(mfrow=c(1,1)) # Reset plot layout

OUT PUT:



31) 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 Electronicsprices 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,

18, 18, 18, 20, 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

**CODE:**

prices <- 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, 18, 18, 18, 20, 20, 20, 20,

20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30)

bin\_size <- length(prices) / 3

bins\_eq\_freq <- split(prices, ceiling(seq\_along(prices) / bin\_size))

print("Equal-Frequency Bins:")

print(bins\_eq\_freq)

bins\_mean <- lapply(bins\_eq\_freq, mean)

smoothed\_mean <- unlist(lapply(seq\_along(bins\_eq\_freq), function(i) rep(bins\_mean[[i]], length(bins\_eq\_freq[[i]]))))

print("Smoothed Data (Bin Mean):")

print(smoothed\_mean)

bins\_boundary <- lapply(bins\_eq\_freq, function(bin) {

min\_val <- min(bin)

max\_val <- max(bin)

sapply(bin, function(x) ifelse(abs(x - min\_val) < abs(x - max\_val), min\_val, max\_val))

})

smoothed\_boundary <- unlist(bins\_boundary)

print("Smoothed Data (Bin Boundary):")

print(smoothed\_boundary)

par(mfrow=c(1,3))

hist(prices, col="blue", main="Original Data", xlab="Price", ylab="Frequency", breaks=10)

hist(smoothed\_mean, col="red", main="Bin Mean Smoothing", xlab="Price", ylab="Frequency", breaks=10)

hist(smoothed\_boundary, col="green", main="Bin Boundary Smoothing", xlab="Price", ylab="Frequency", breaks=10)

par(mfrow=c(1,1))

OUT PUT:

