**S.BHAVYA SRI**

**192211953**

**CSA1679**

**DATA WAREHOUSING AND DATA MINING FOR INDUSTRY**

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

1-5. 200

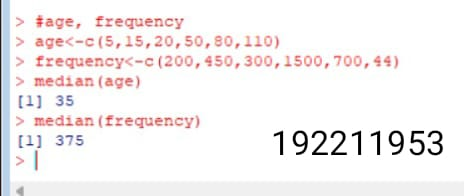
5-15 450

15-20 300

20-50 1500

50-80 700

80-110 44 Compute an approximate median value for the data



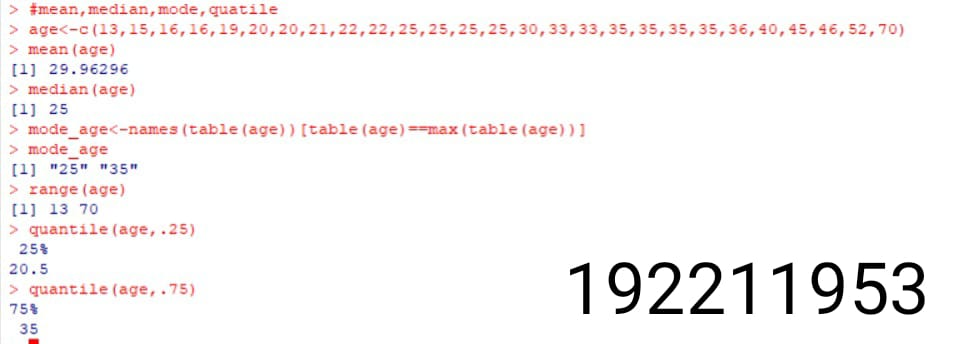
2.Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(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?



3.Data Preprocessing :Reduction and Transformation

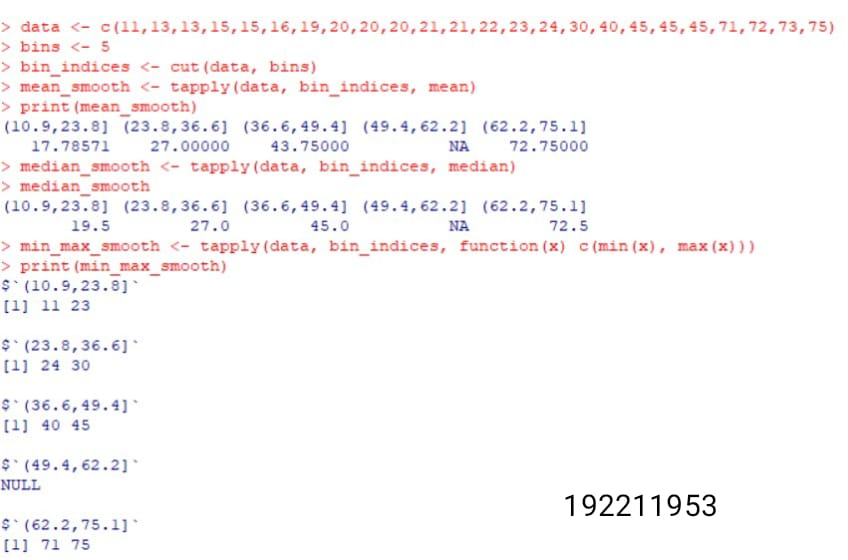
Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000 (a) min-max normalization by setting min = 0 and max = 1 (b) z-score normalization4.Data:11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,

72,73,75

a) Smoothing by bin mean

b) Smoothing by bin median

c) Smoothing by bin boundaries

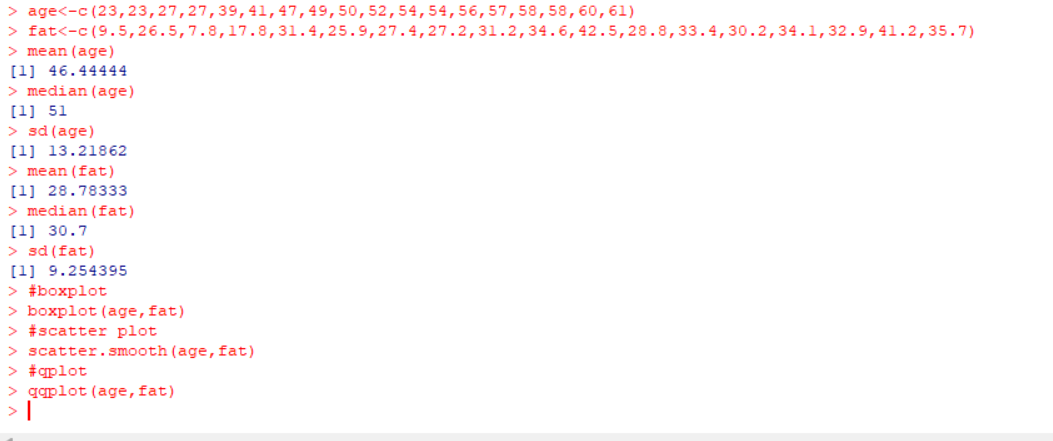


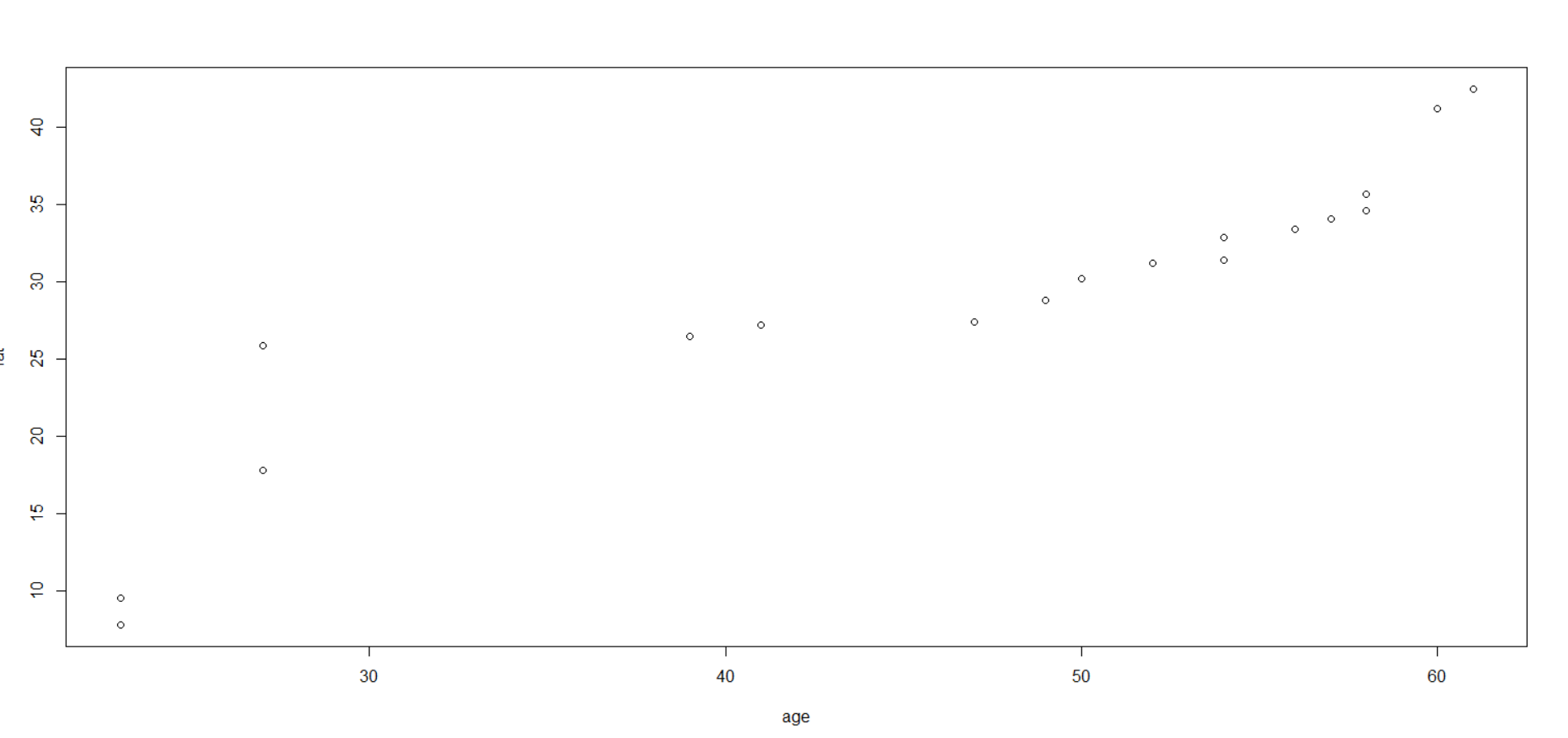
4 question:Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

(a) Calculate the mean, median, and standard deviation of age and %fat.

(b) Draw the boxplots for age and %fat.

(c) Draw a scatter plot and a q-q plot based on these two variables.



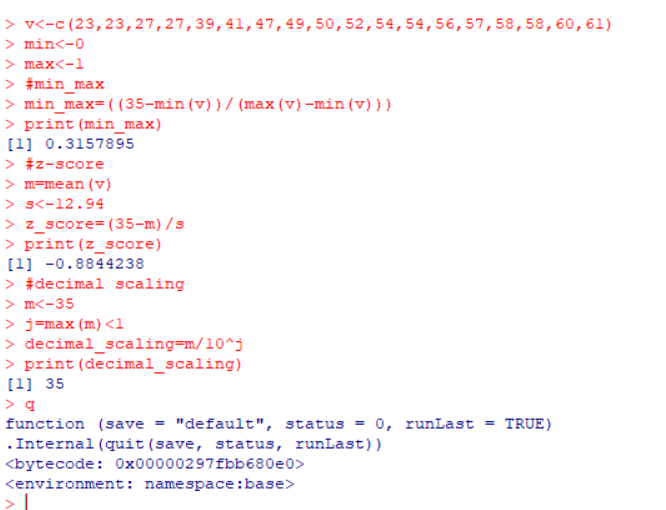


5 question:.Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

(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.

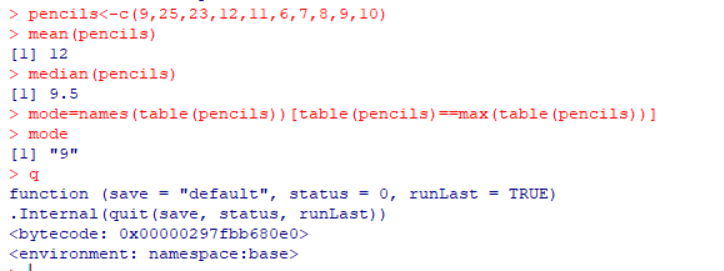
(iii) Use normalization by decimal scaling to transform the value 35 for age. Perform the above functions using R – tool



6 question: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

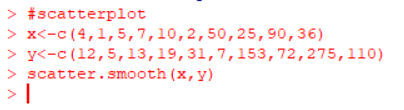
9 25 23 12 11 6 7 8 9 10

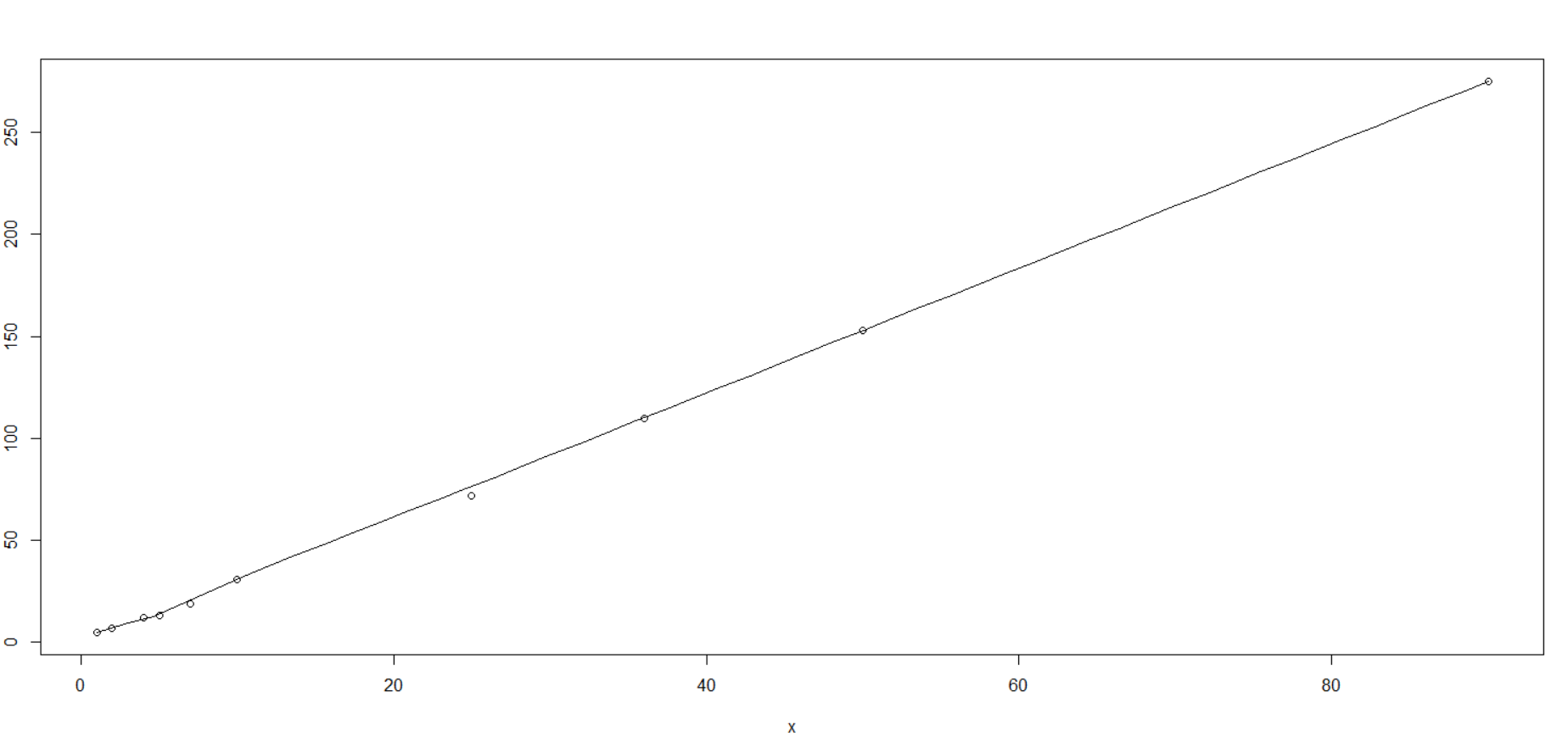


7 question:. the following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones sold.

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

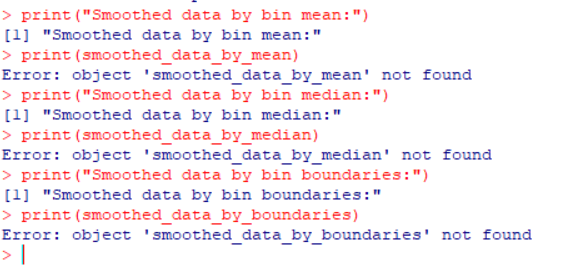
y :12 5 13 19 31 7 153 72 275 110





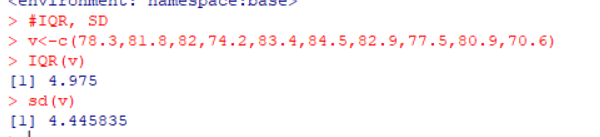
8 question:Implement of the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75. Partition them into three bins by each of the following methods. Plot the data points using histogram.

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



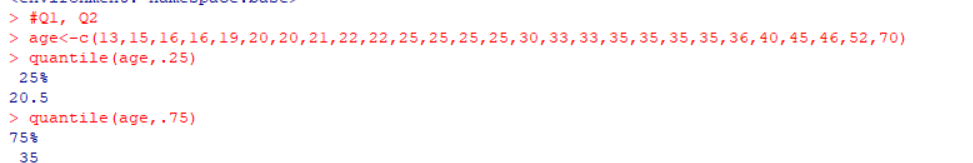
9 question:Implement of the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75. Partition them into three bins by each of the following methods. Plot the data points using histogram.

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



10 question: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?



**EXERCISE 2:**

1 question: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

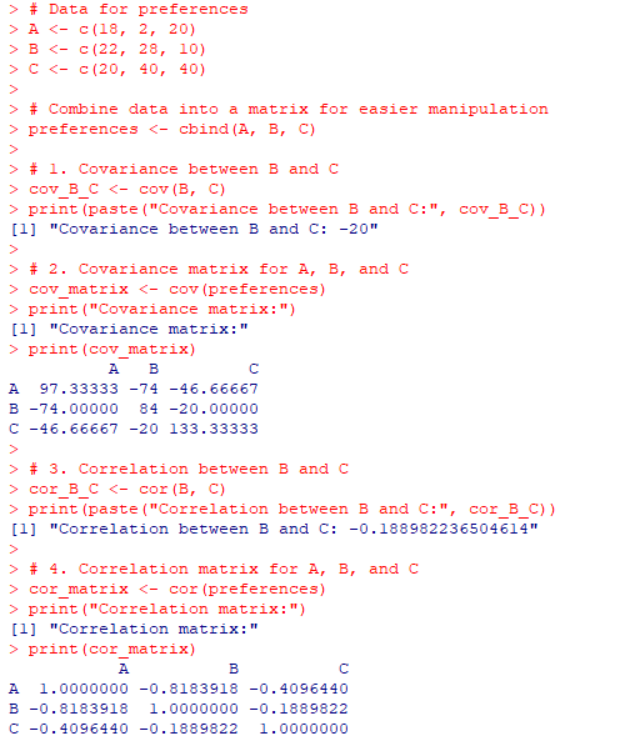
9-10 years: 20 10 40

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.

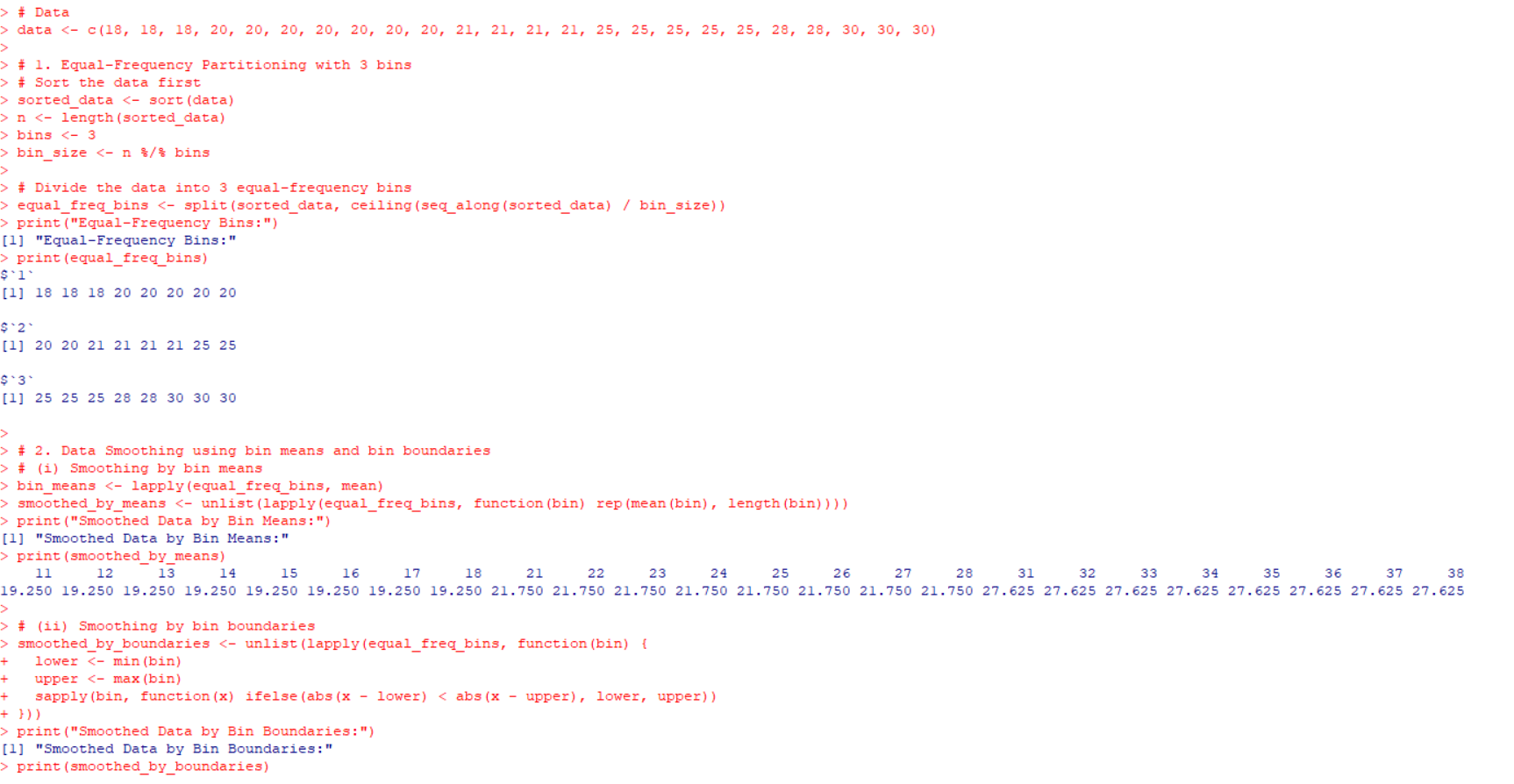
3. Use cor() to calculate the sample correlation between B and C.

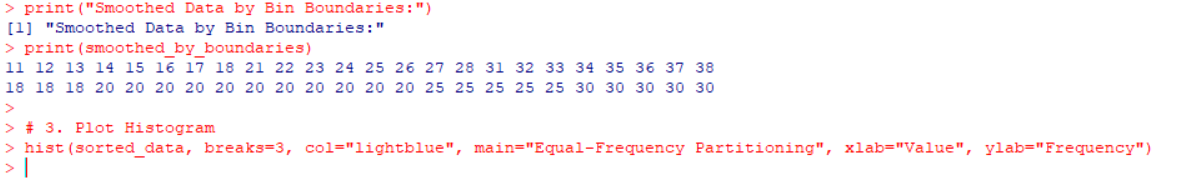
4. Use another call to cor() to calculate the sample correlation matrix for the preferences.

****

2 question:.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 |

****

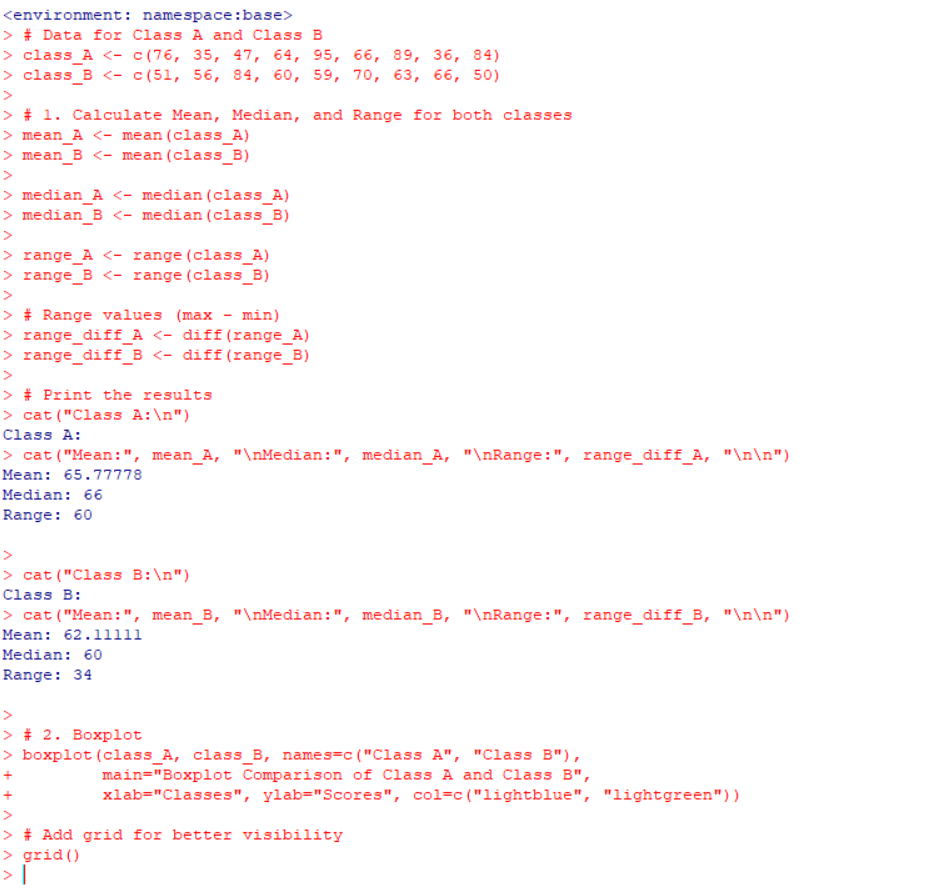
****

3 question: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

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

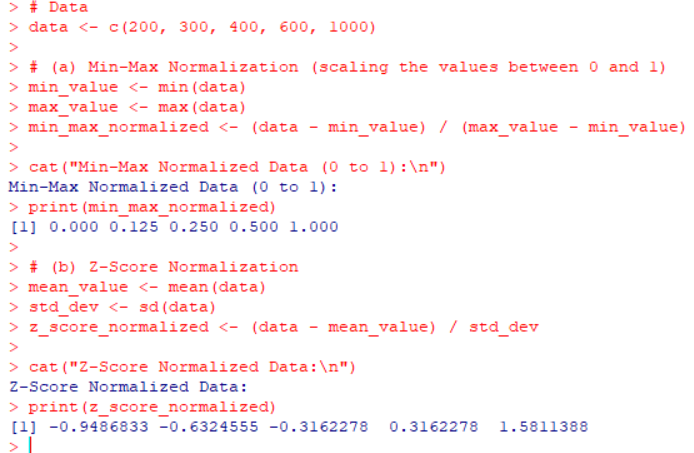
****

4 question: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,

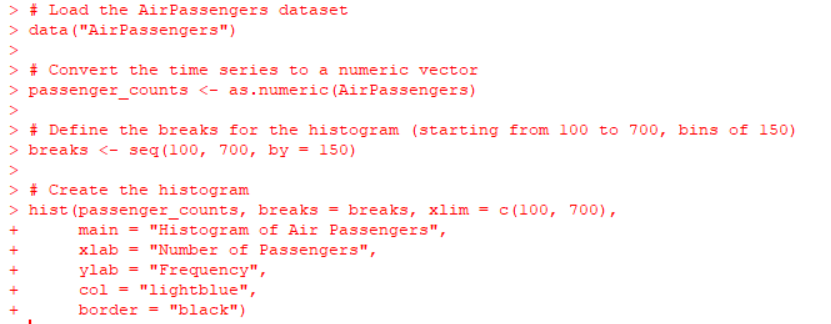
b) Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000

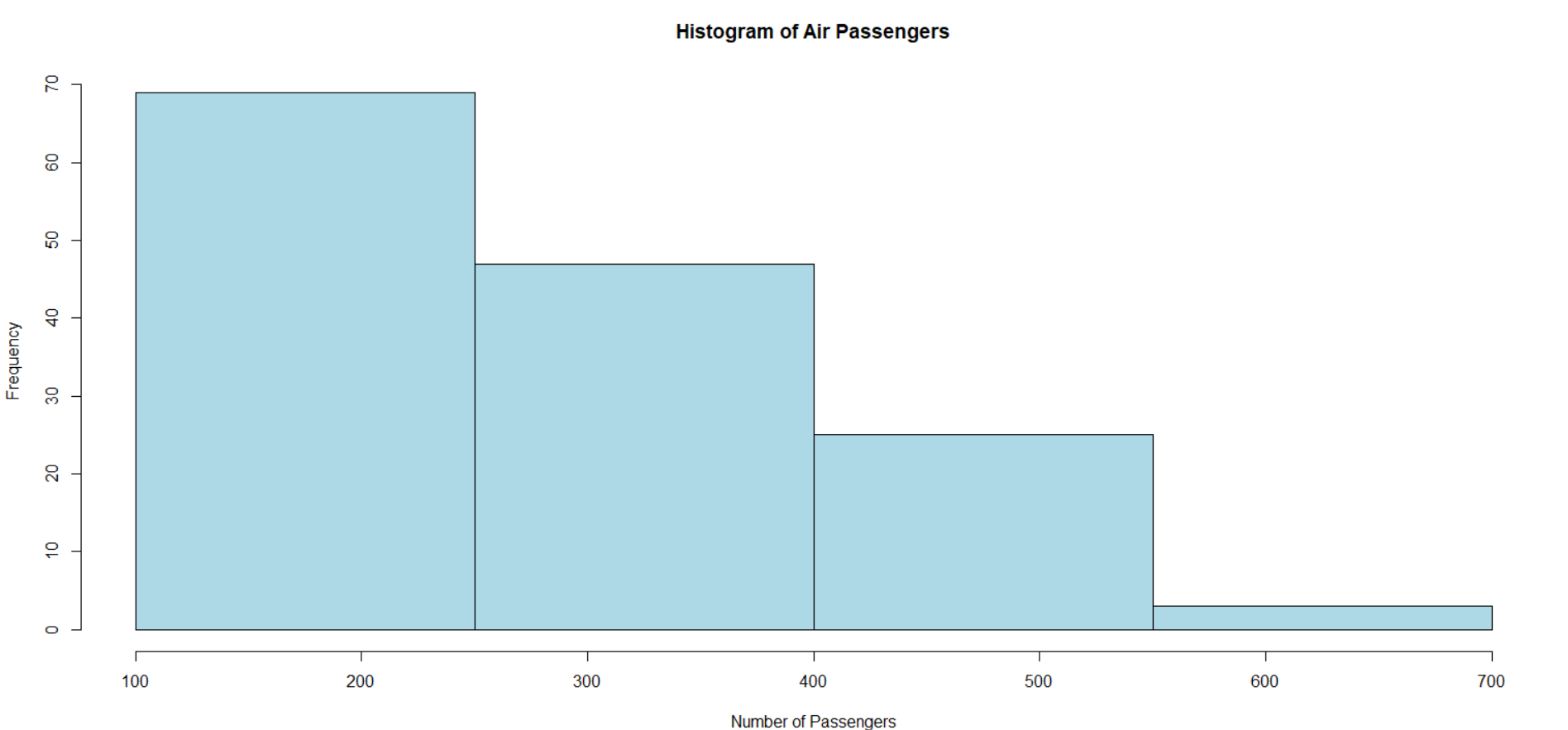
(a) min-max normalization by setting min = 0 and max = 1

(b) z-score normalization

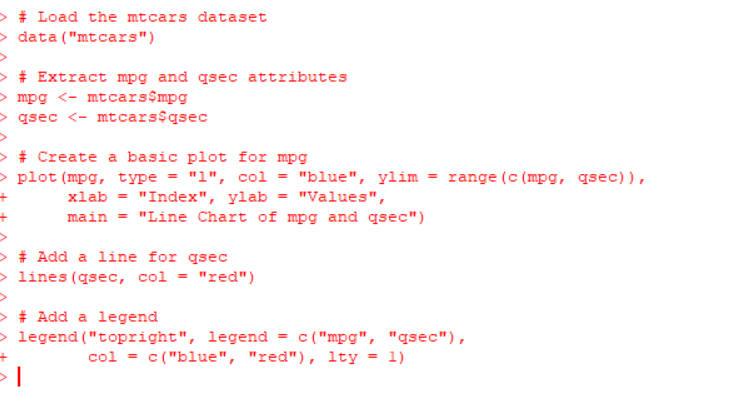
****

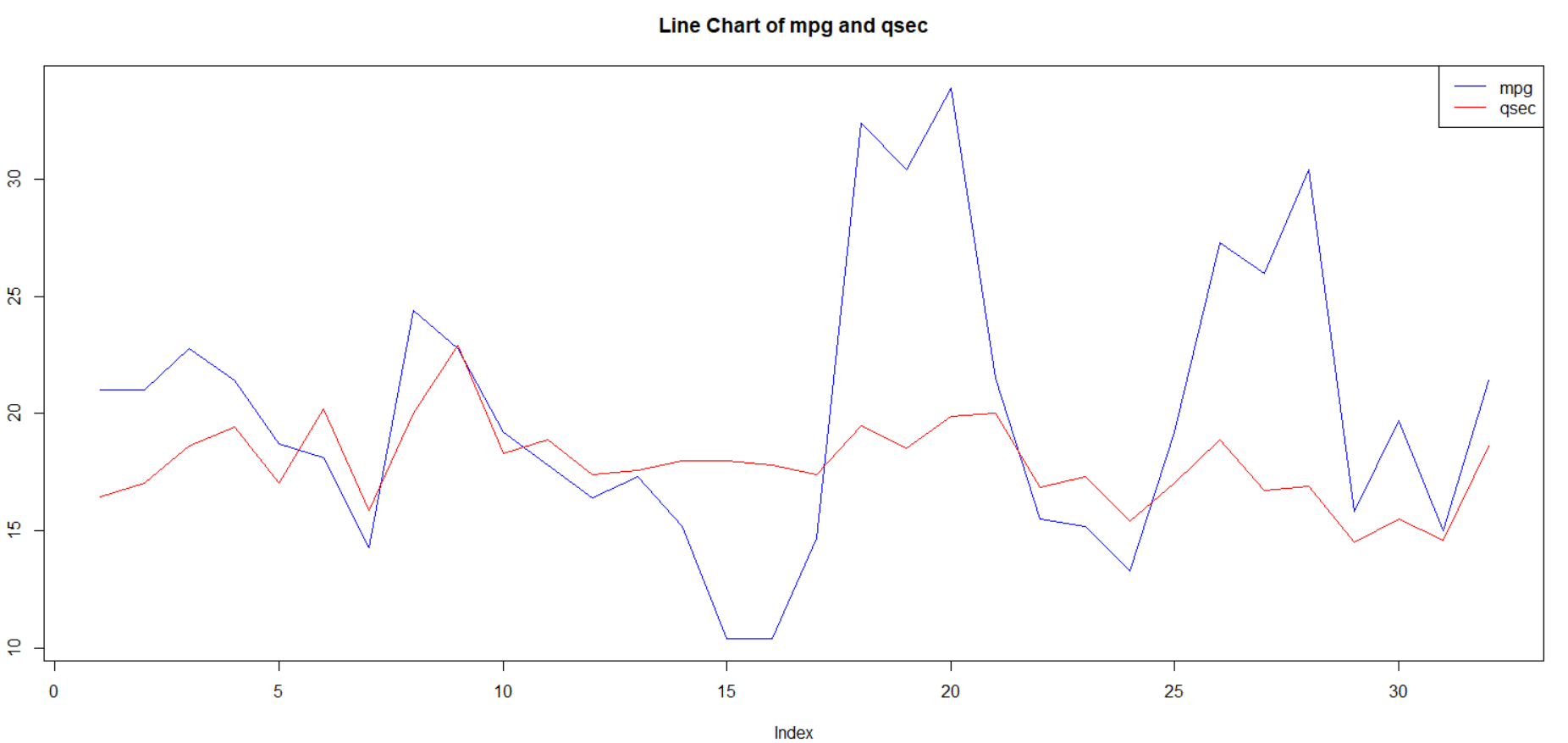
5 question:Make a histogram for the “AirPassengers “dataset, start at 100 on the x-axis, and from values 200 to 700, make the bins 150 wide

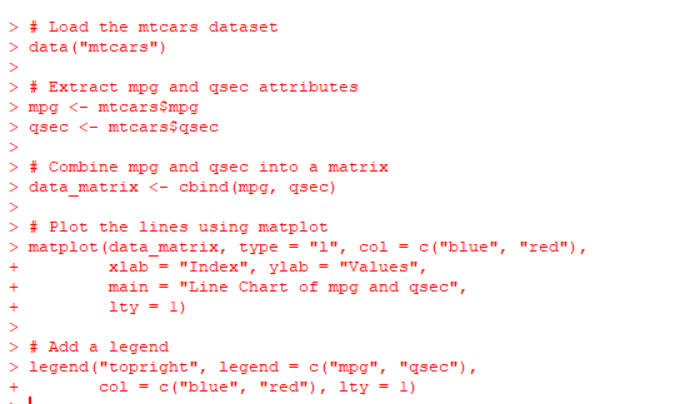
****

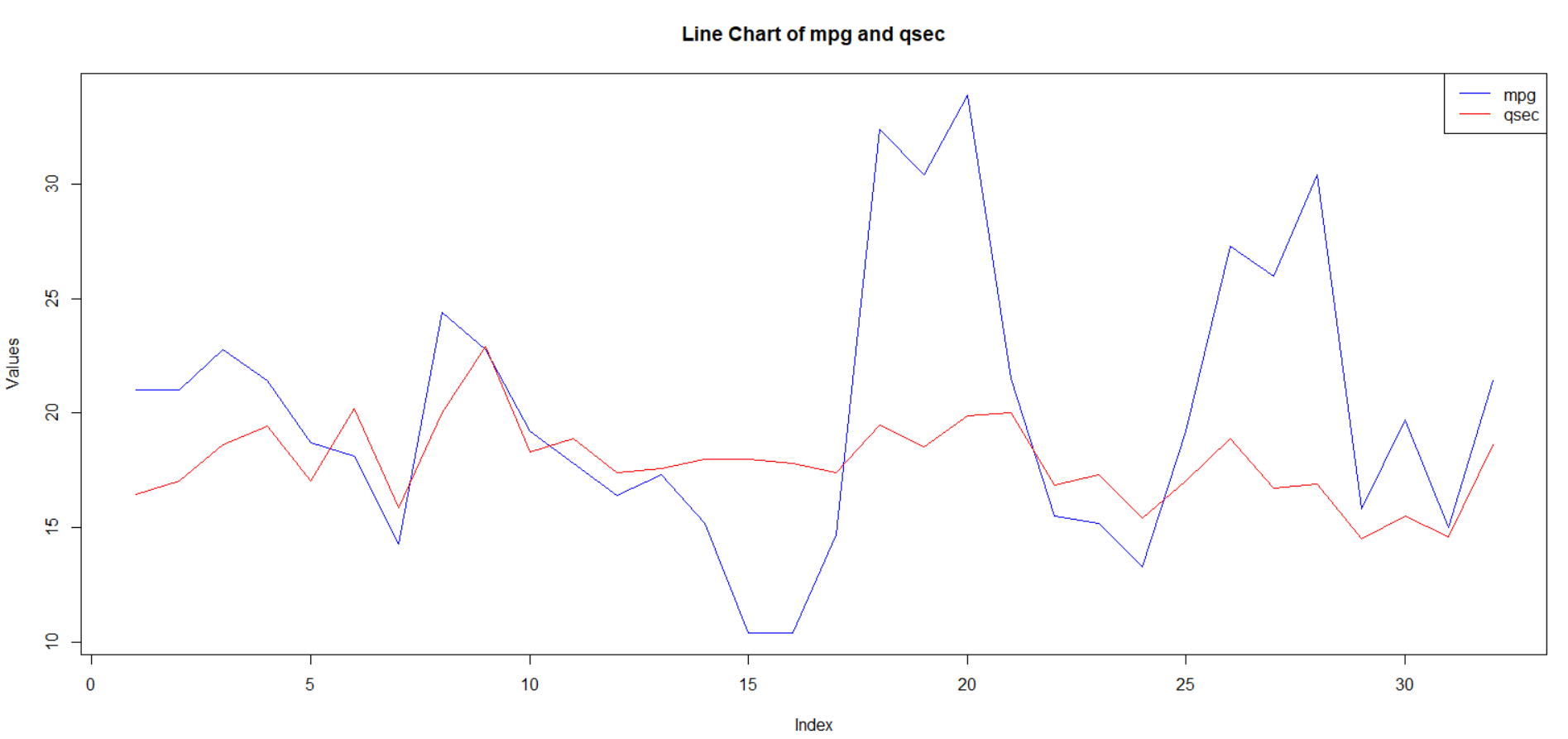
****

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

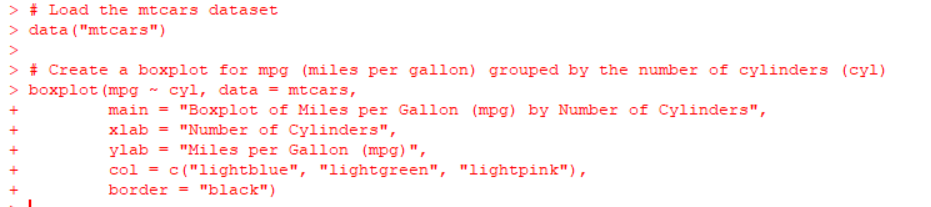
****

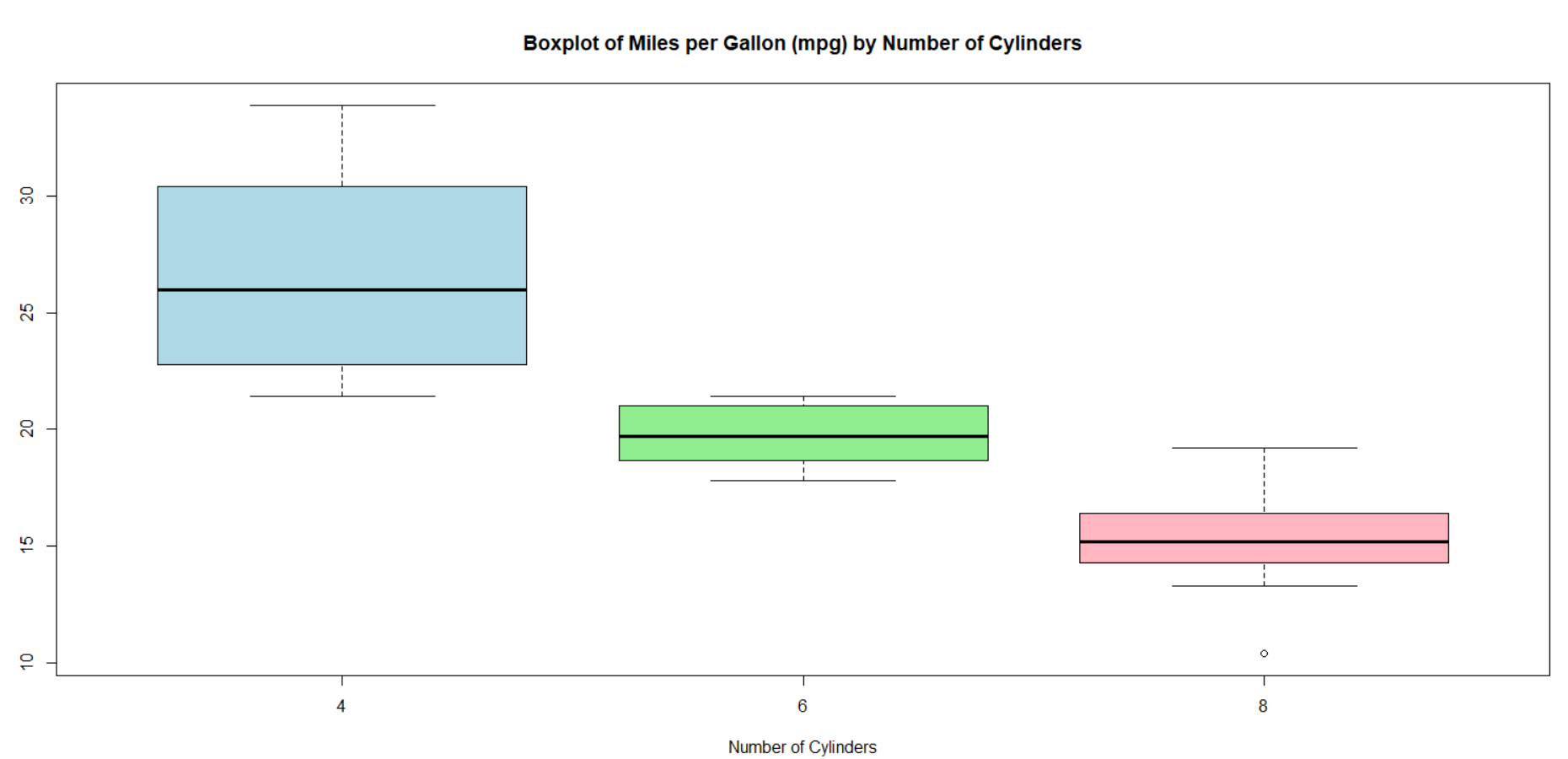
****

****

****

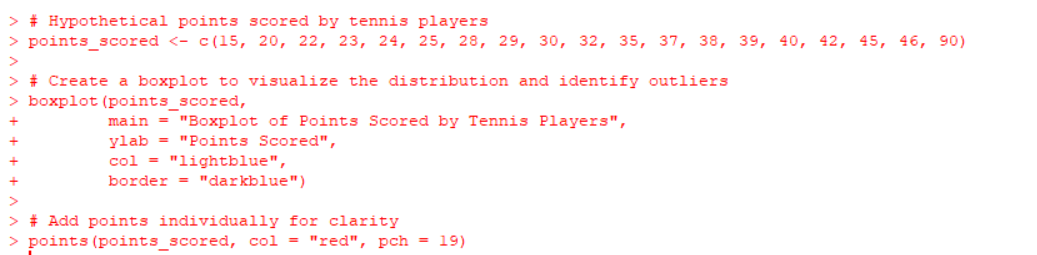
8 question:Create a Boxplot graph for the relation between "mpg"(miles per galloon) and "cyl"(number of Cylinders) for the dataset "mtcars" available in R Environment.

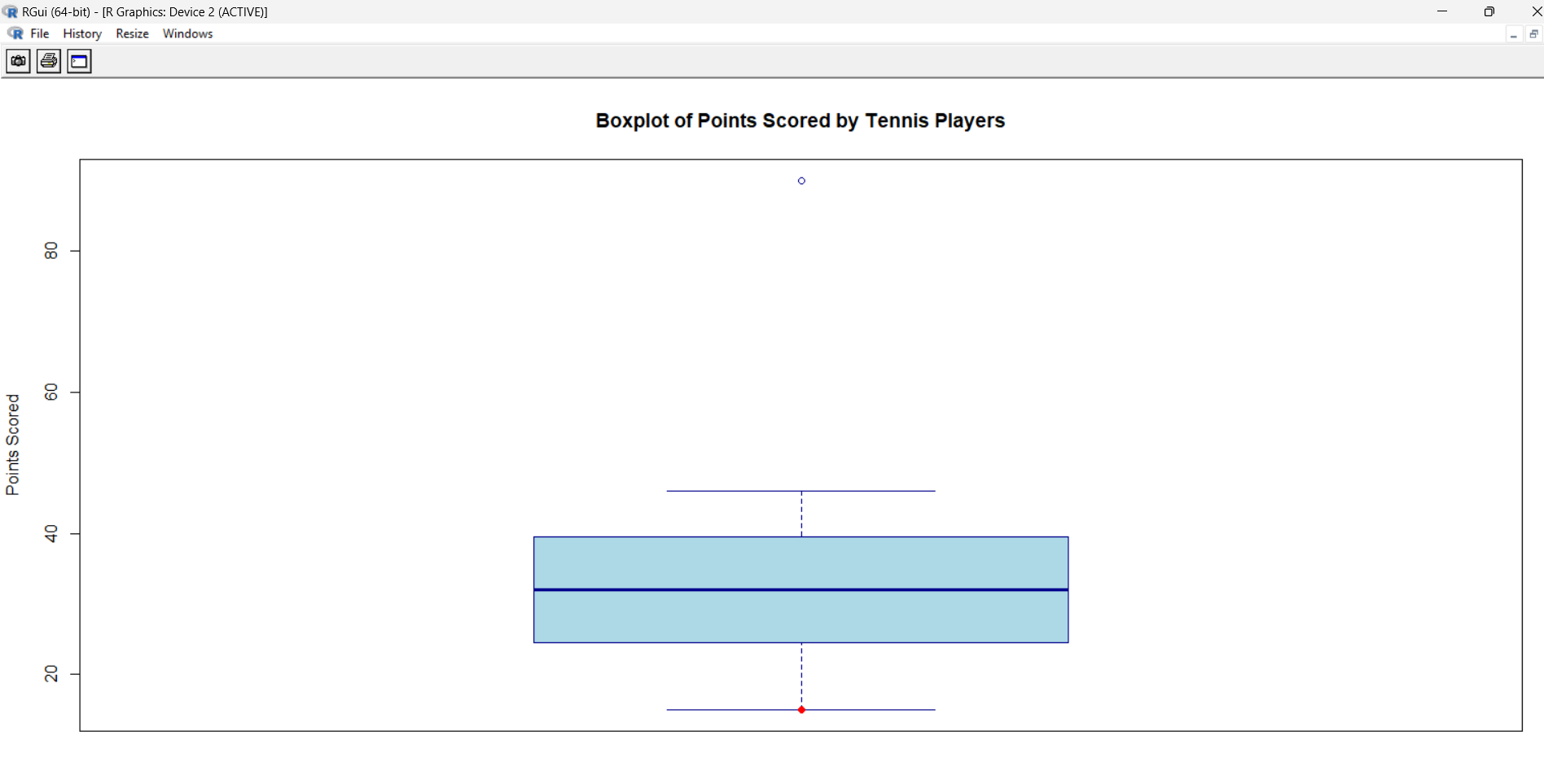
****

****

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

technique.

****

****

10 question: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”)

# Load any necessary library

> # (Uncomment this if the dataset is in CSV and needs to be imported)

> # diabetes\_data <- read.csv("path\_to\_diabetes\_dataset.csv")

>

> # Example synthetic diabetes dataset

> set.seed(123)

> age <- sample(20:80, 100, replace = TRUE) # Ages between 20 and 80

> blood\_pressure <- rnorm(100, mean = 120, sd = 15) # Random blood pressure values

> diabetes\_status <- sample(c("Diabetic", "Non-Diabetic"), 100, replace = TRUE)

>

> # Create a dataframe to hold the dataset

> diabetes\_data <- data.frame(age, blood\_pressure, diabetes\_status)

>

> # Scatterplot: Age vs Blood Pressure

> plot(diabetes\_data$age, diabetes\_data$blood\_pressure,

+ main = "Scatterplot of Age vs Blood Pressure",

+ xlab = "Age",

+ ylab = "Blood Pressure",

+ col = ifelse(diabetes\_data$diabetes\_status == "Diabetic", "red", "blue"),

+ pch = 19)

>

> # Add a legend to the scatterplot

> legend("topright", legend = c("Diabetic", "Non-Diabetic"), col = c("red", "blue"), pch = 19)

>

> # Bar chart: Count of people in different age groups

> # Define age groups (bins)

> diabetes\_data$age\_group <- cut(diabetes\_data$age, breaks = c(20, 30, 40, 50, 60, 70, 80),

+ labels = c("20-30", "30-40", "40-50", "50-60", "60-70", "70-80"))

>

> # Create a bar plot of the age groups and the count of people

> barplot(table(diabetes\_data$age\_group),

+ main = "Bar Chart of Age Group Distribution",

+ xlab = "Age Group",

+ ylab = "Count of People",

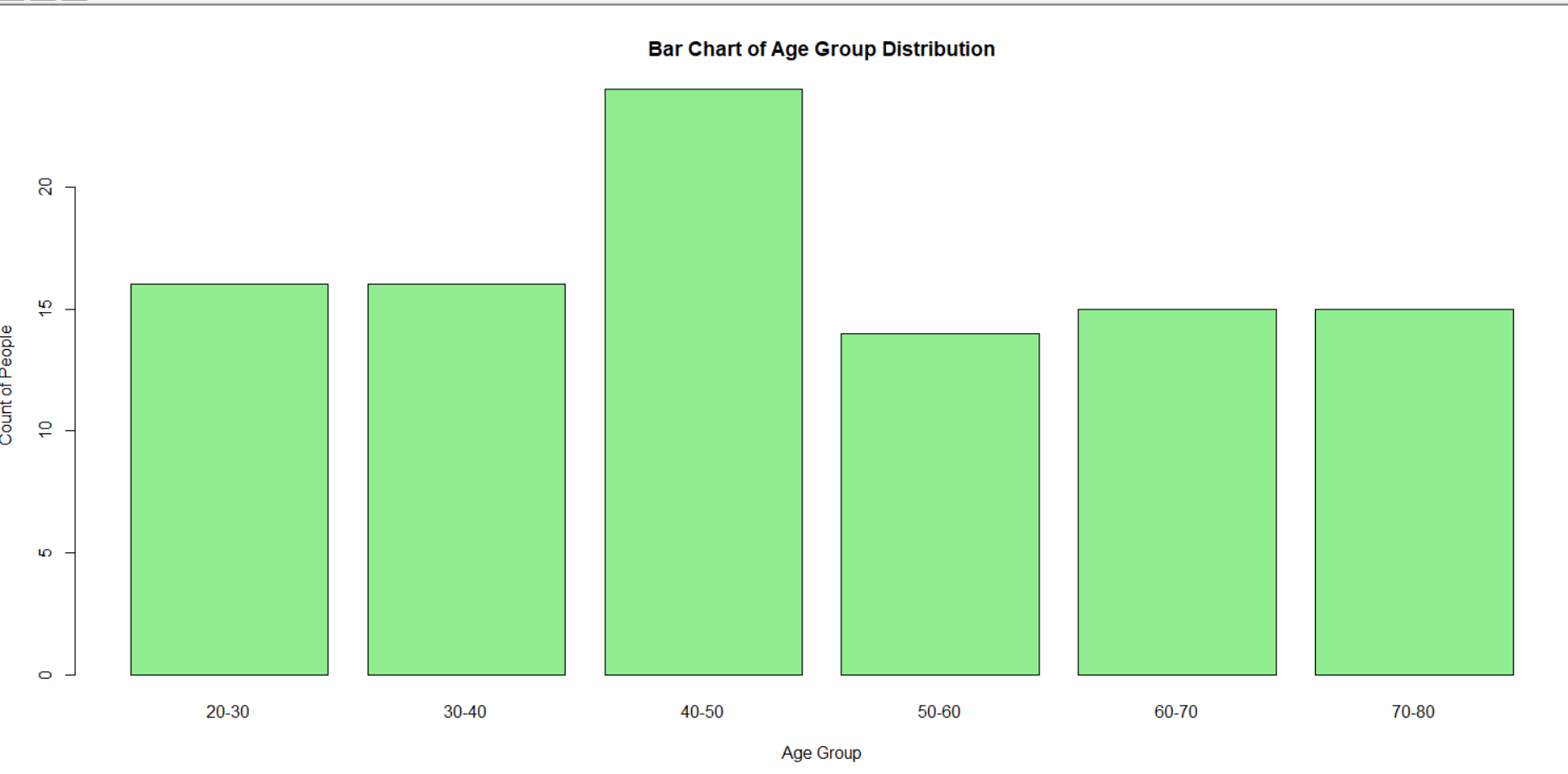
+ col = "lightgreen",

+ border = "black")

>

> # Add grid for better readability

> grid()

****

**EXERCISE 3:**

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

@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

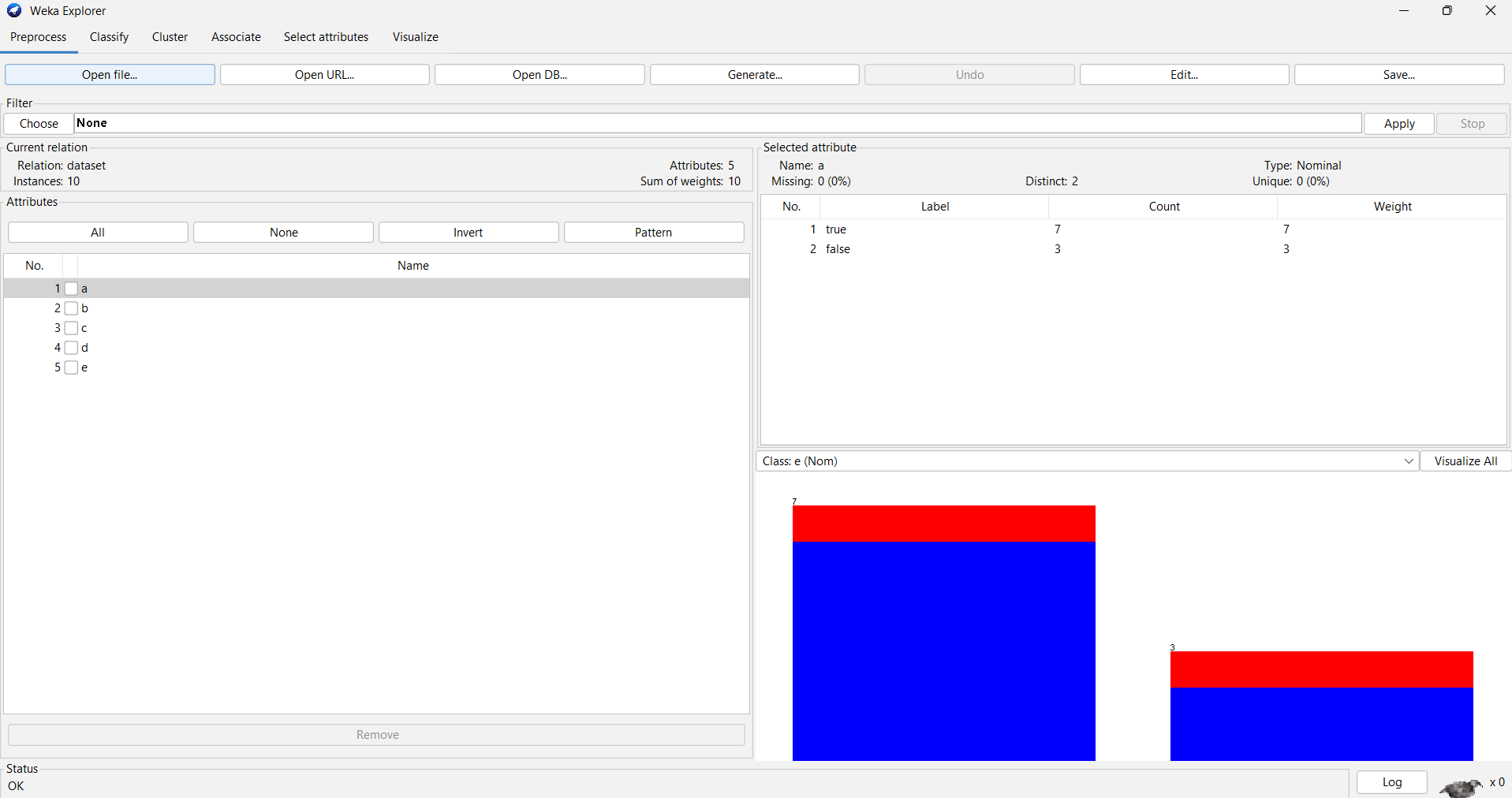
false true false true true

false false true true false

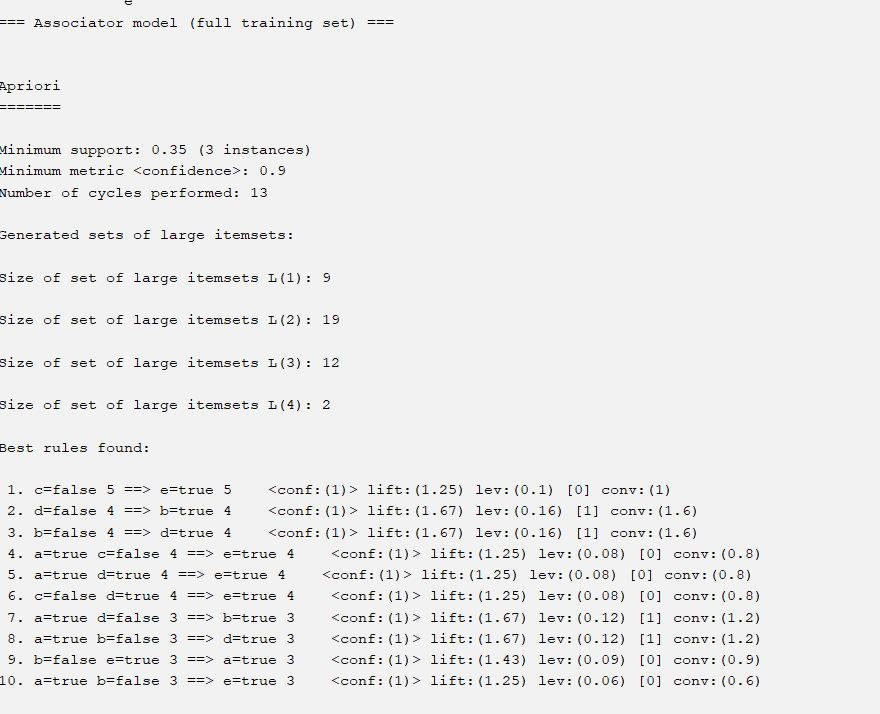
true true true false false

true false false true true

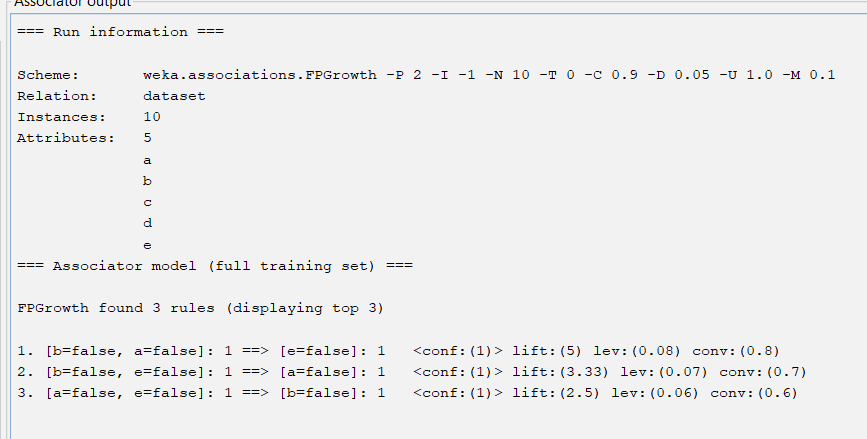
true true false false true



**Apriori algorithm:**



**F growth:**

****

2 question: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 dataset

@attribute Milk{true,false}

@attribute Bread{true,false}

@attribute Diapers{true,false}

@attribute Butter{true,false}

@attribute Beer{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

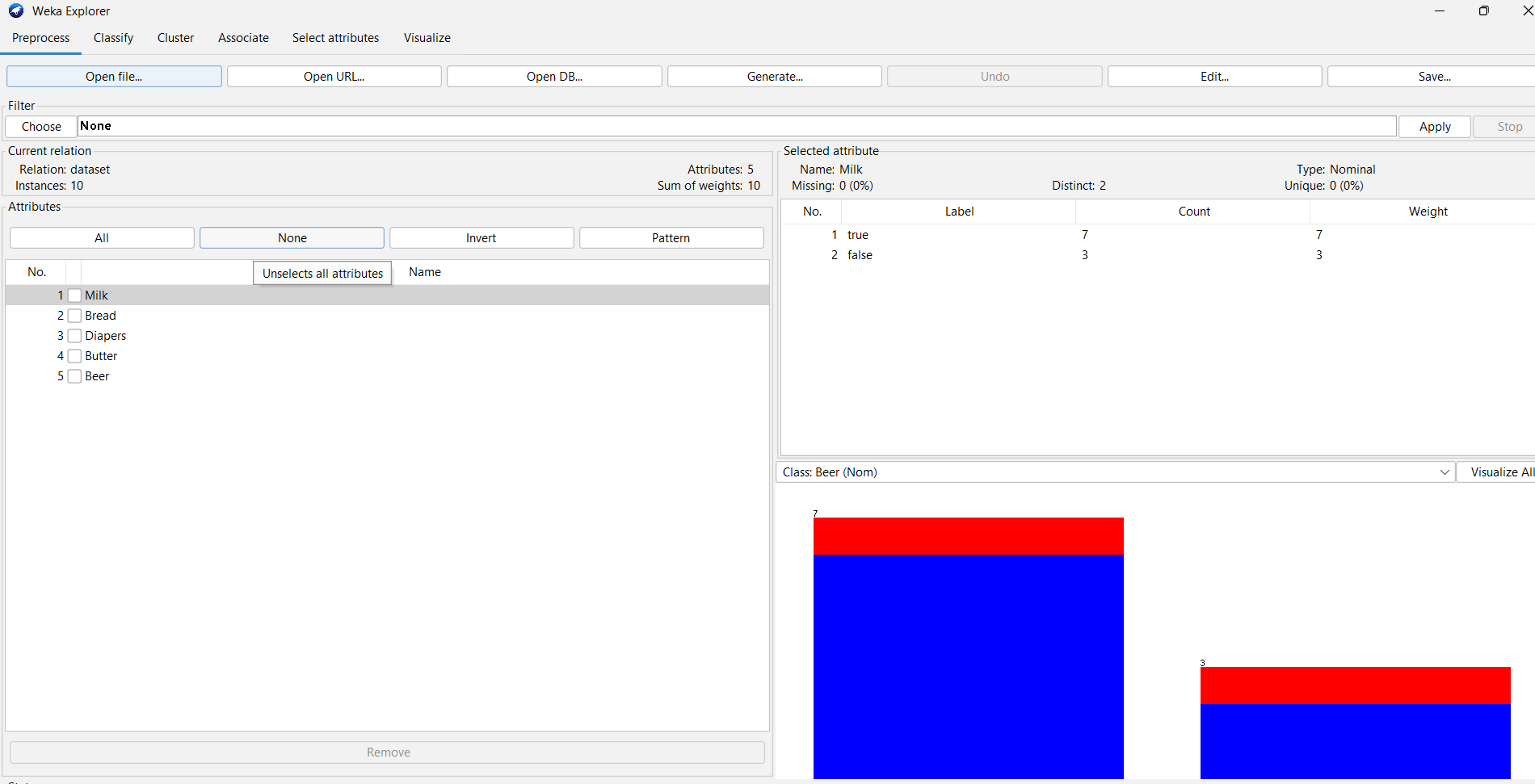
false true false true true

false false true true false

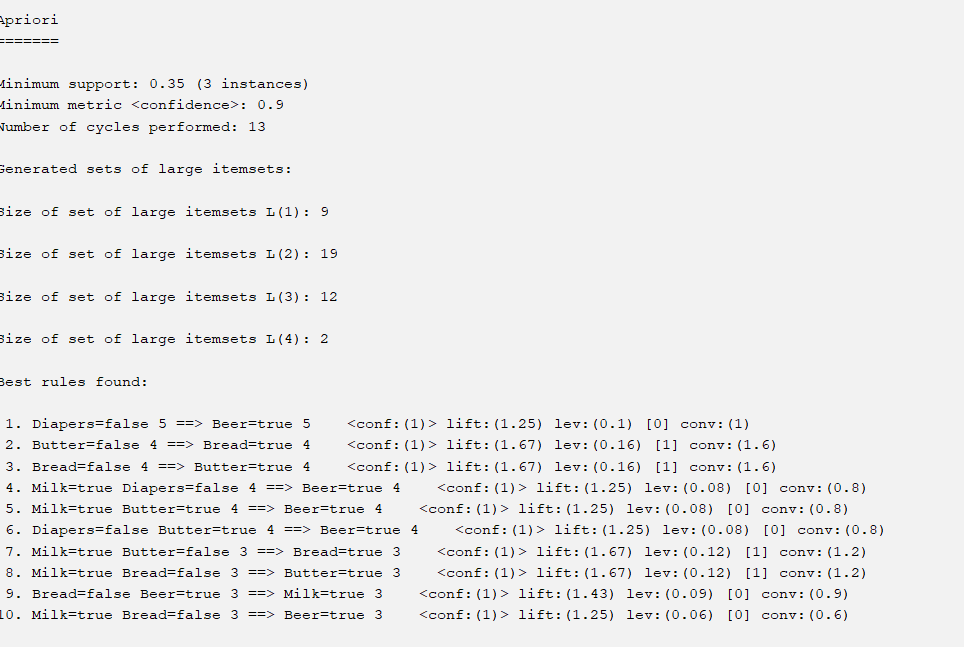
true true true false false

true false false true true

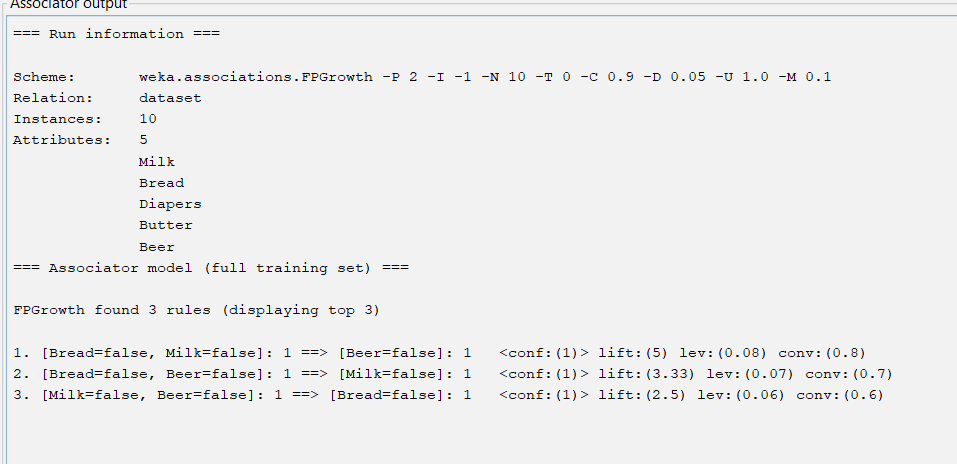
true true false false true

****

**Apriori algorithm:**

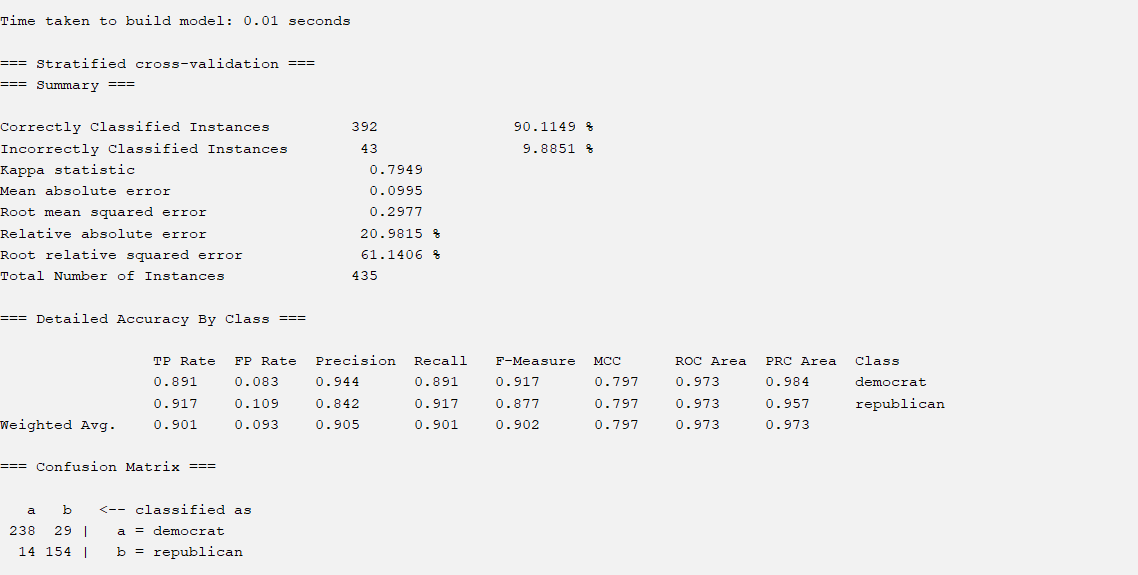
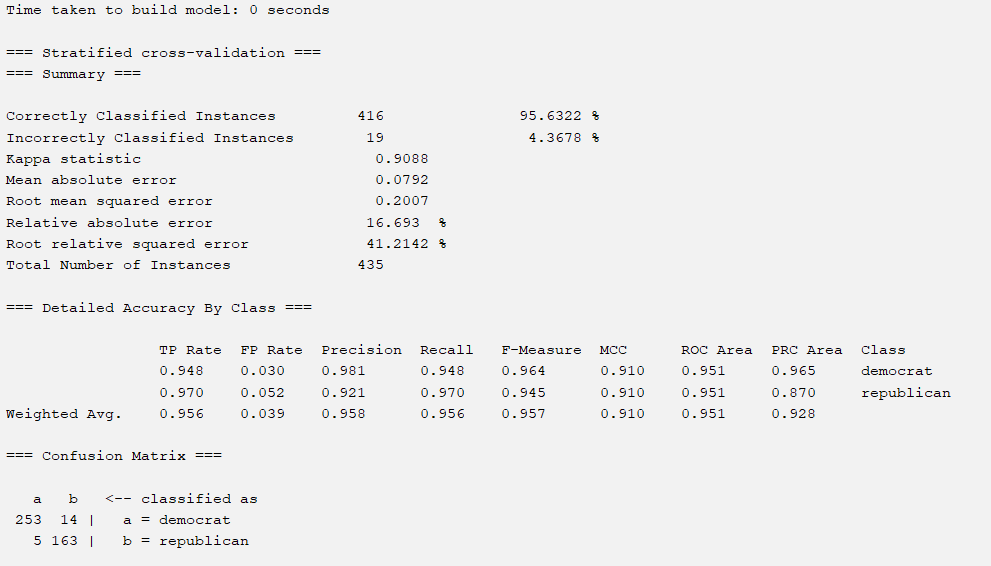
****

**F growth:**

****

**3 question:Bayes classification and descion tree (using training and test data)**

**BAYES:**

****

4 question: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

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

@relation dataset

@attribute Monkey{true,false}

@attribute donkey{true,false}

@attribute make{true,false}

@attribute mucky{true,false}

@attribute cookie{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

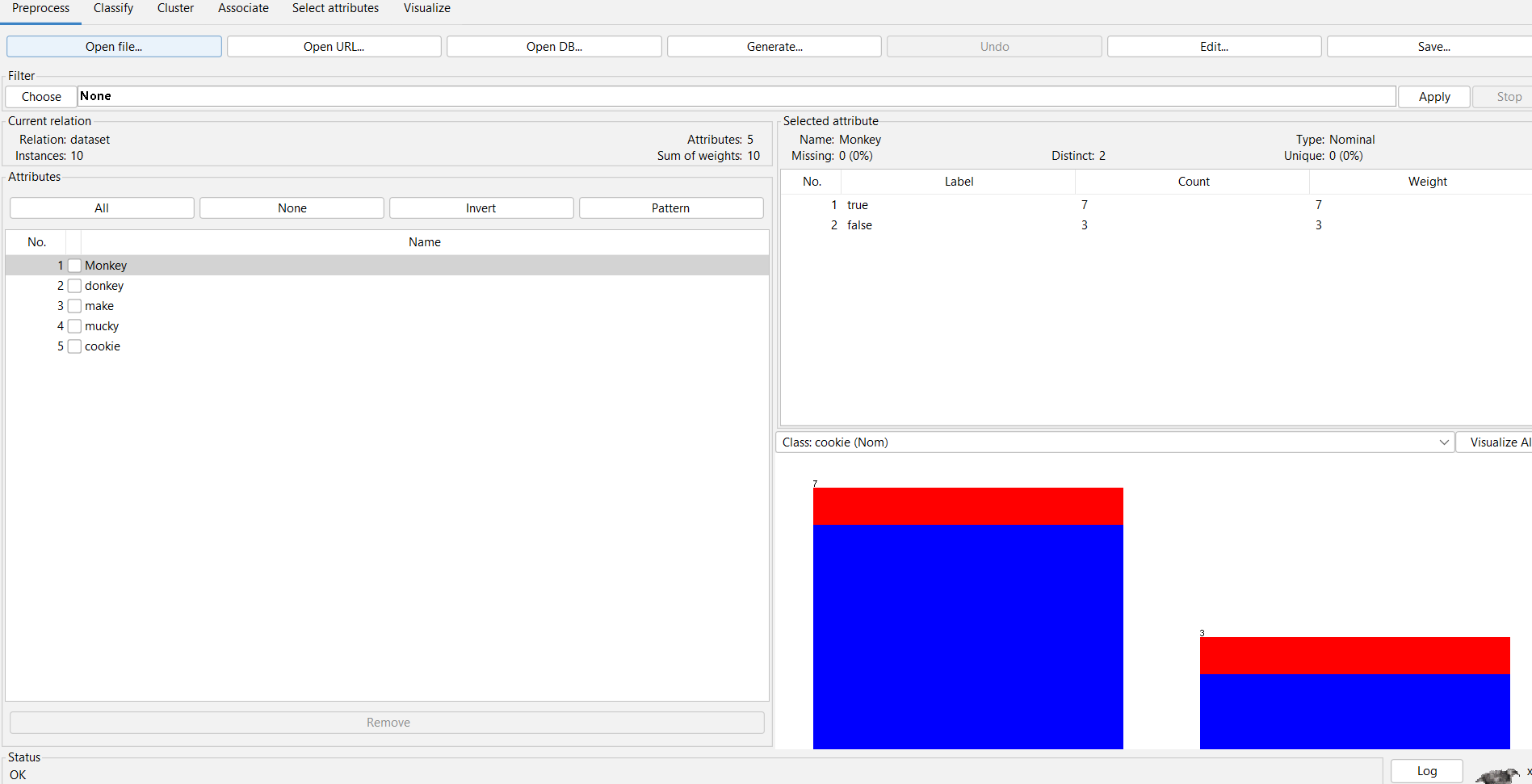
false true false true true

false false true true false

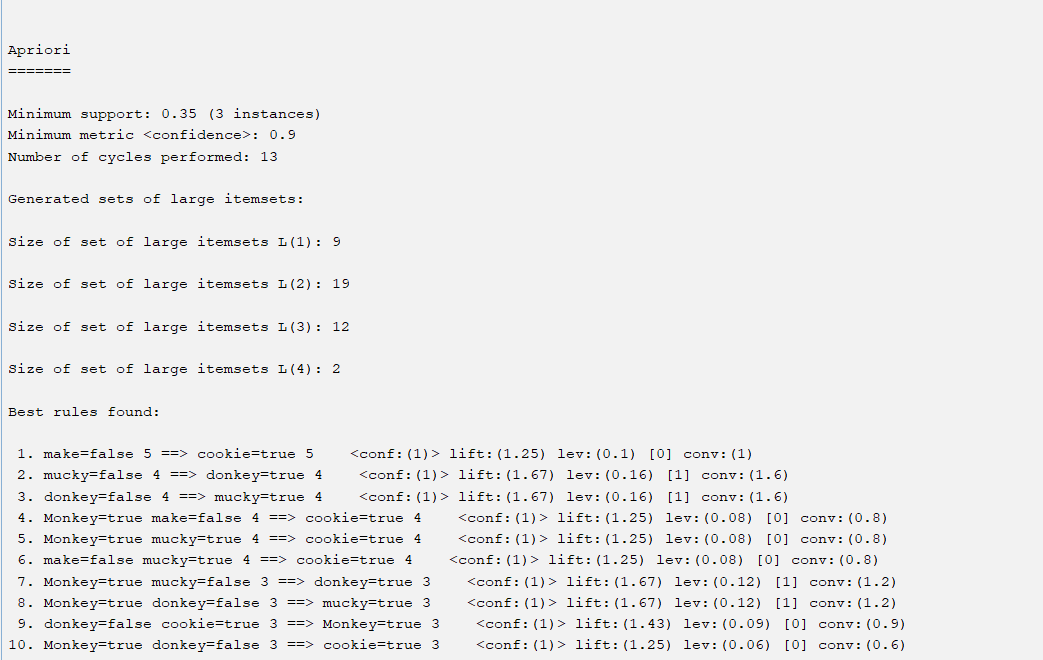
true true true false false

true false false true true

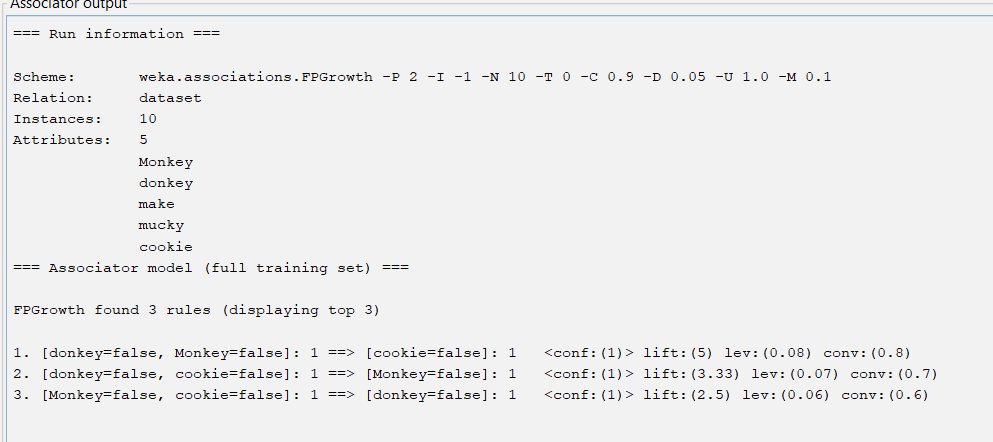
true true false false true

****

**Apriori algorithm:**

****

**F growth:**

****

5 question: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.

@relation dataset

@attribute hotdogs{true,false}

@attribute buns{true,false}

@attribute coke{true,false}

@attribute chip{true,false}

@attribute ketchup{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

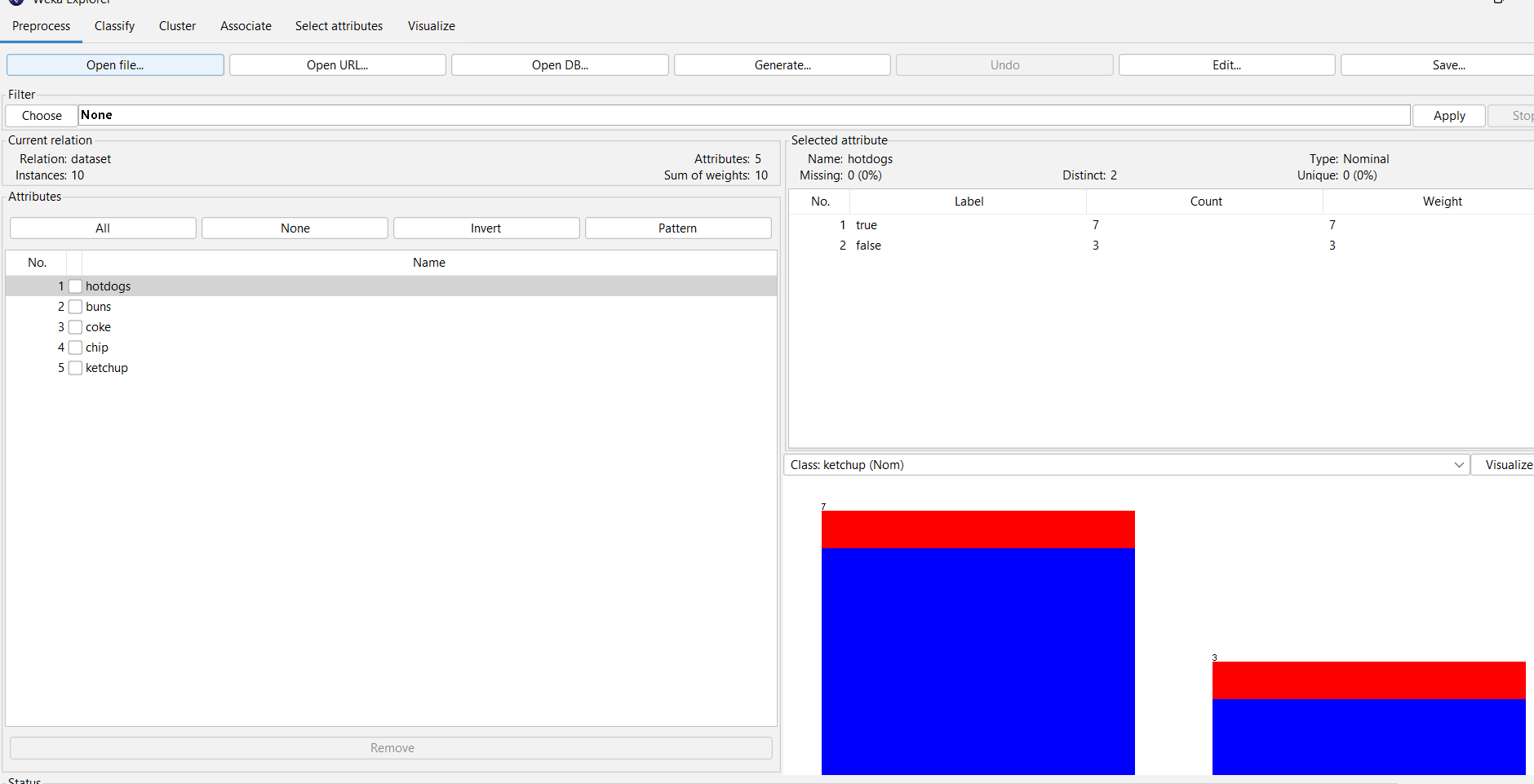
false true false true true

false false true true false

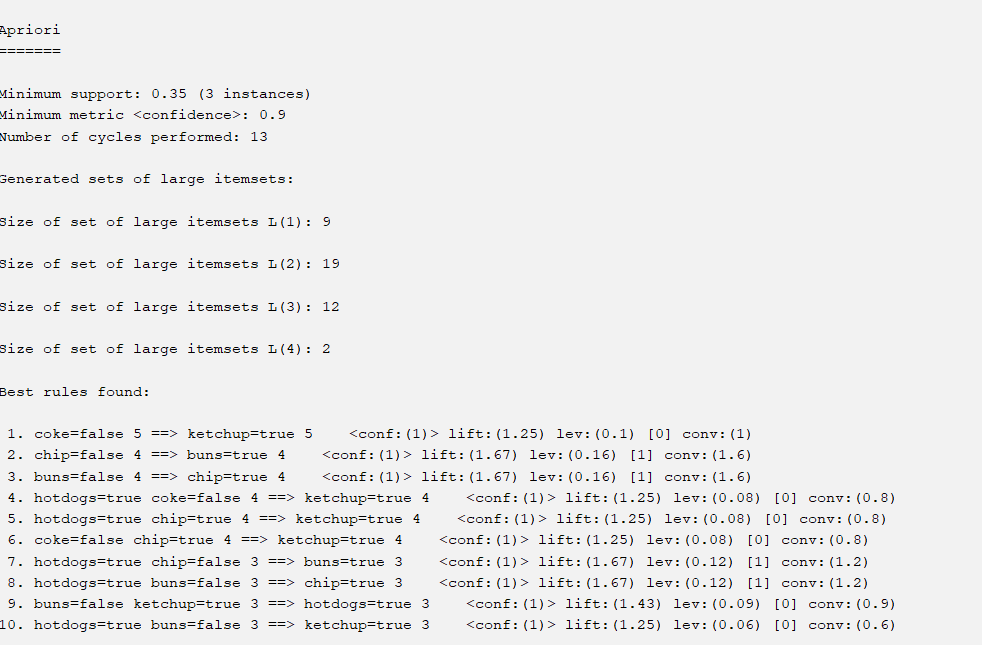
true true true false false

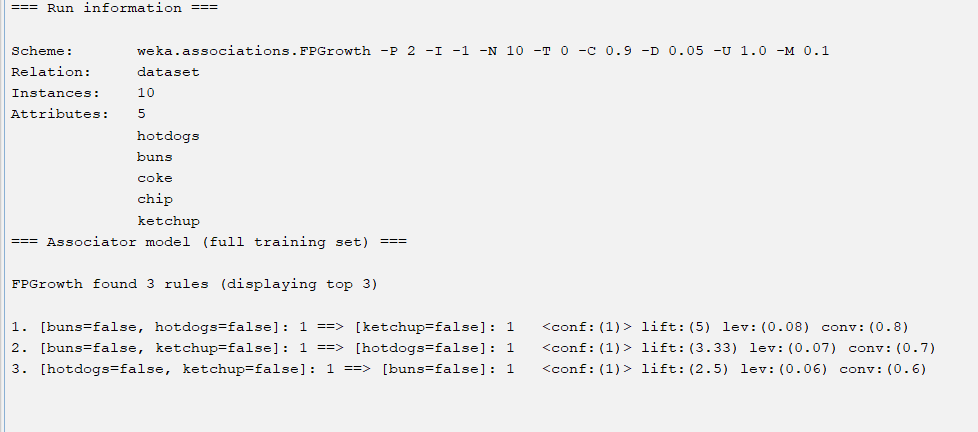
true false false true true

true true false false true

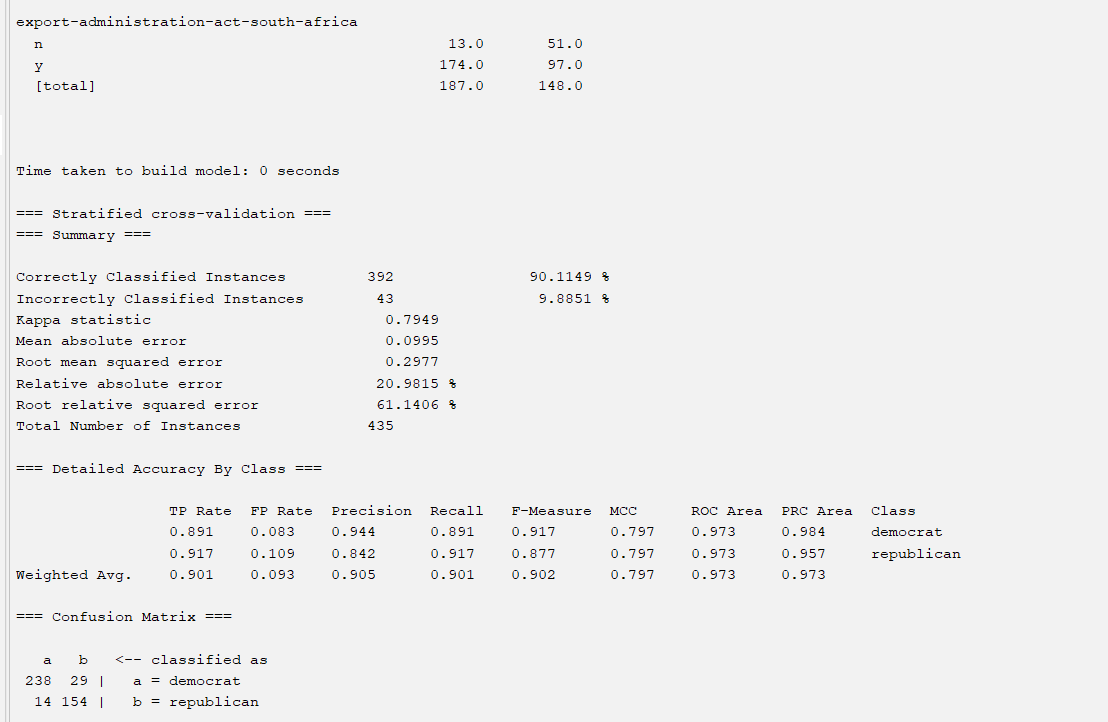
****

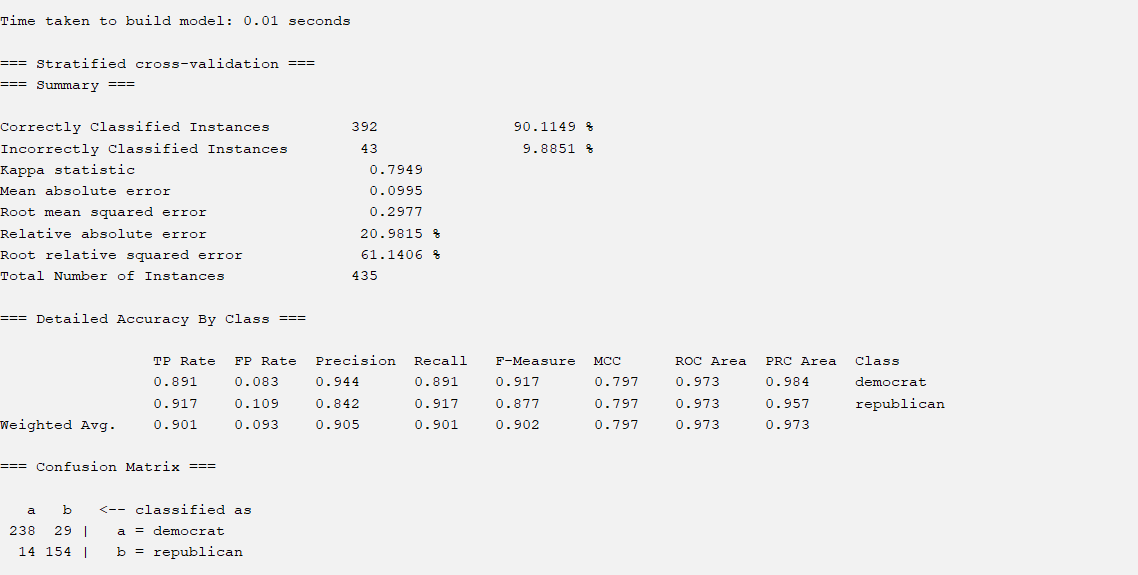
**Apriori algorithm:**

**F GROWTH:**

****6 QUESTION: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

NATIVE BAYER:

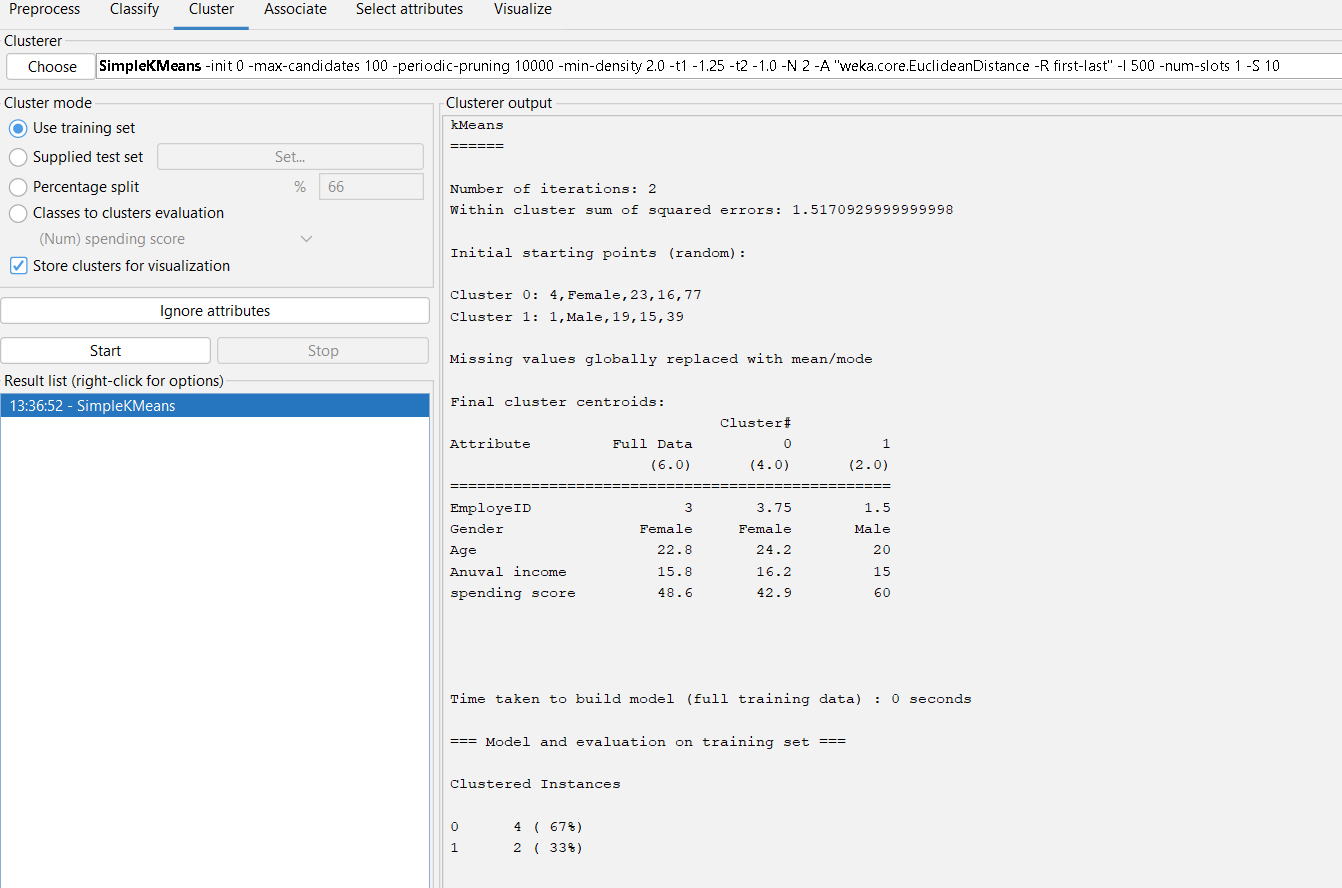
****

DECISION TREE:  ****

**EXERCISE 4**

**1 QUESTION:**

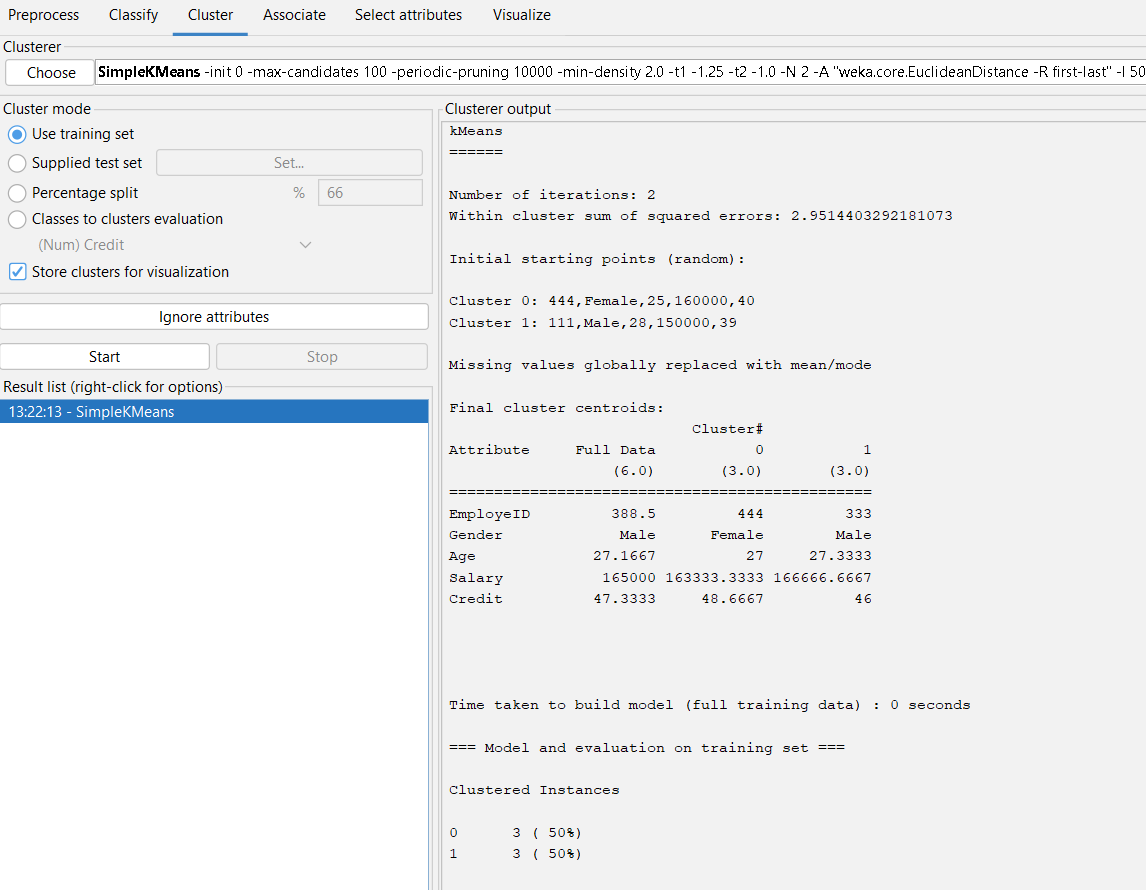
| **EmployeID** | **Gender** | **Age** | **Anuval income** | **spending score** |
| --- | --- | --- | --- | --- |
| **1** | **Male** | **19** | **15** | **39** |
| **2** | **Male** | **21** | **15** | **81** |
| **3** | **Female** | **20** | **16** | **6** |
| **4** | **Female** | **23** | **16** | **77** |
| **5** | **Female** | **31** | **17** | **40** |

****

**2 QUESTION:**

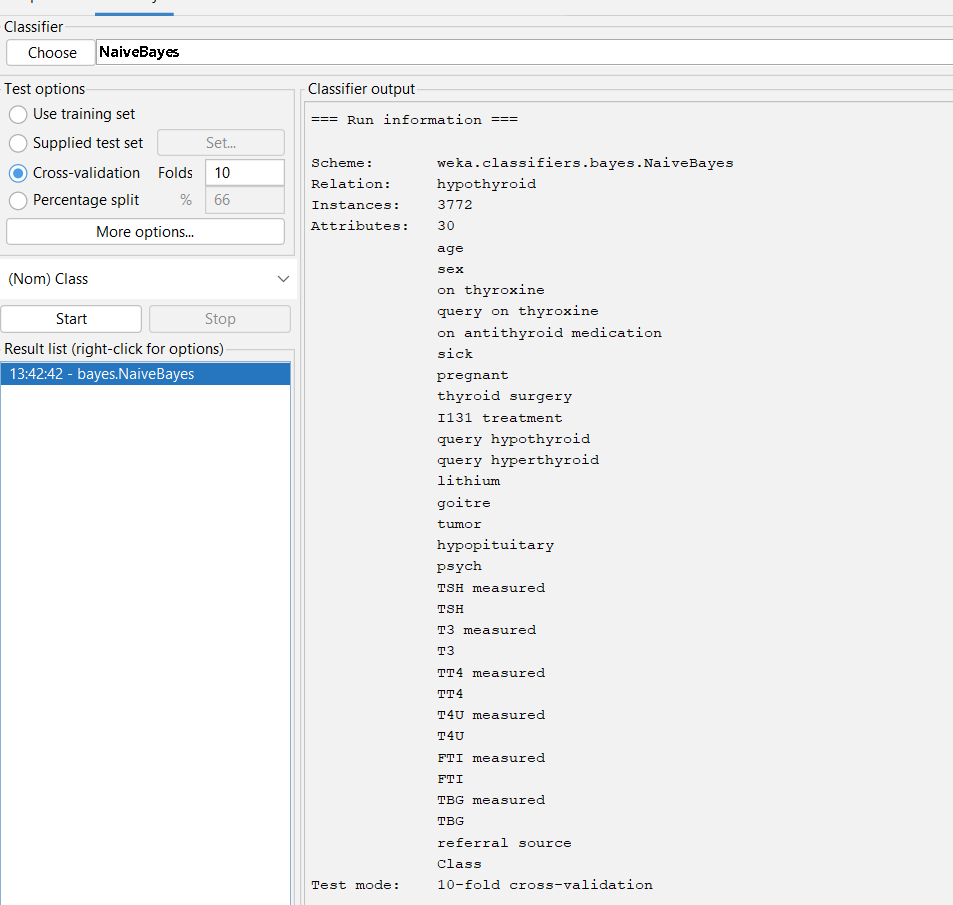
| **EmployeID** | **Gender** | **Age** | **Salary** | **Credit** |
| --- | --- | --- | --- | --- |
| **111** | **Male** | **28** | **150000** | **39** |
| **222** | **Male** | **25** | **150000** | **27** |
| **333** | **Female** | **26** | **160000** | **42** |
| **444** | **Female** | **25** | **160000** | **40** |
| **555** | **Female** | **30** | **170000** | **64** |
| **666** | **Male** | **29** | **200000** | **72** |

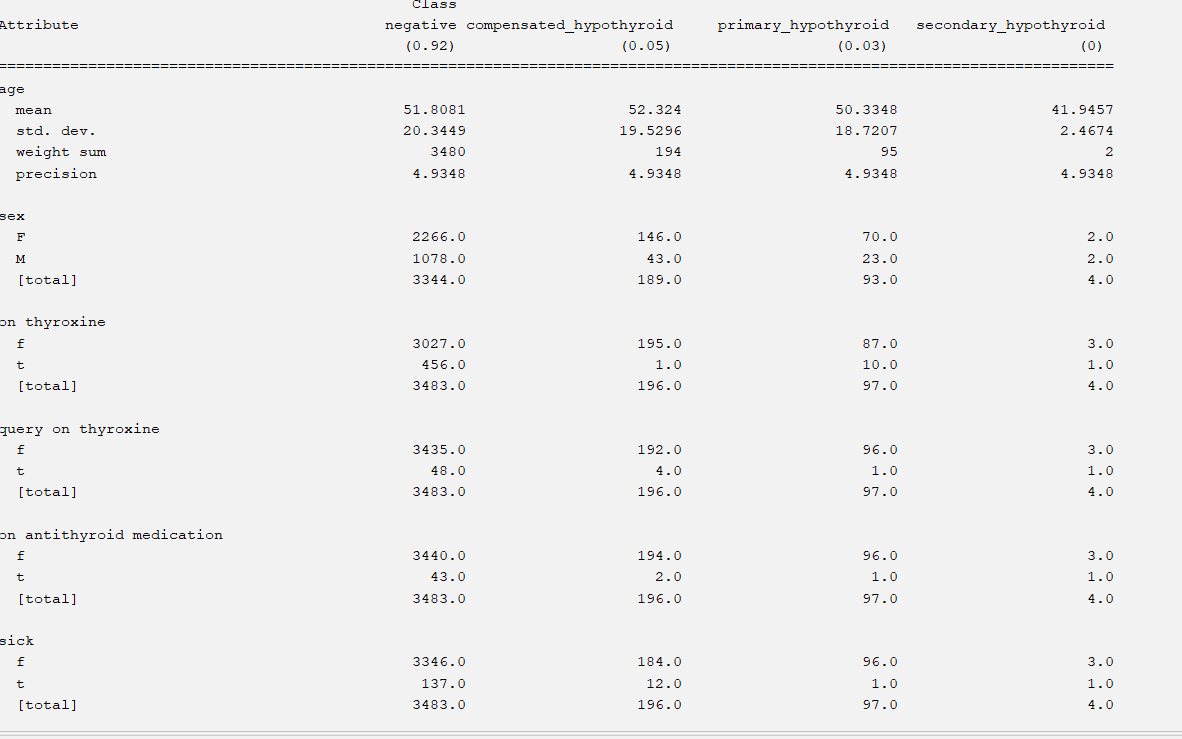
**K MEANS CLUSTERING:**

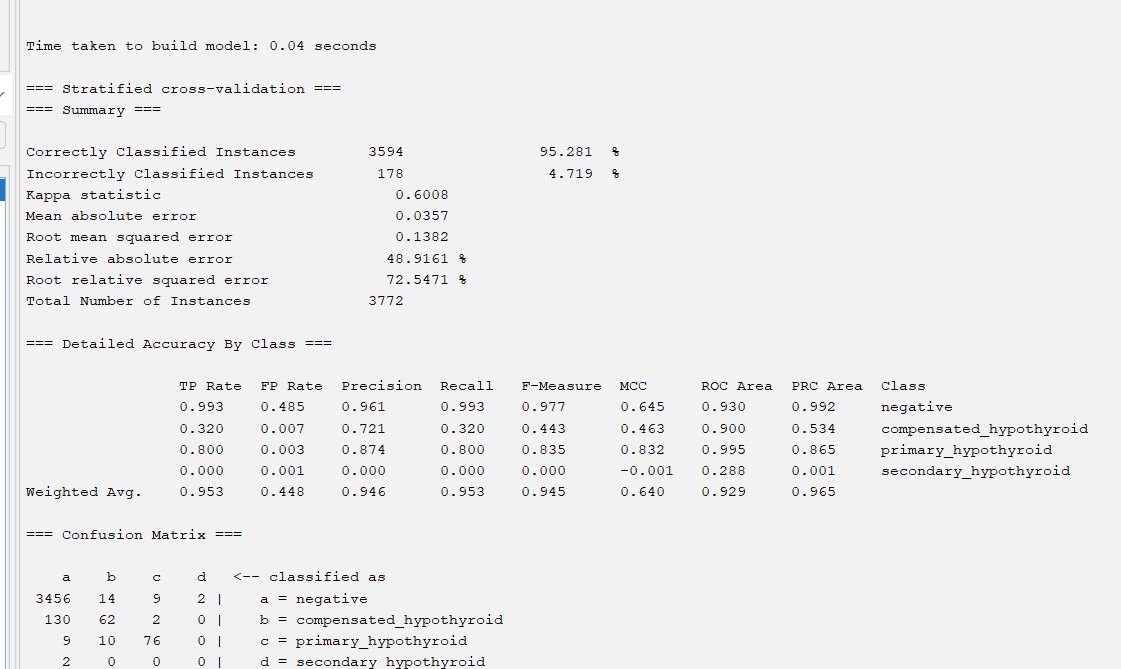
****

3 question:.Prediction of categorical data using Naïve Bayes classification through WEKA using any datasets. Compare the Naïve Bayes algorithm with SVM using the summary of results given by the classifiers and plot the graph.

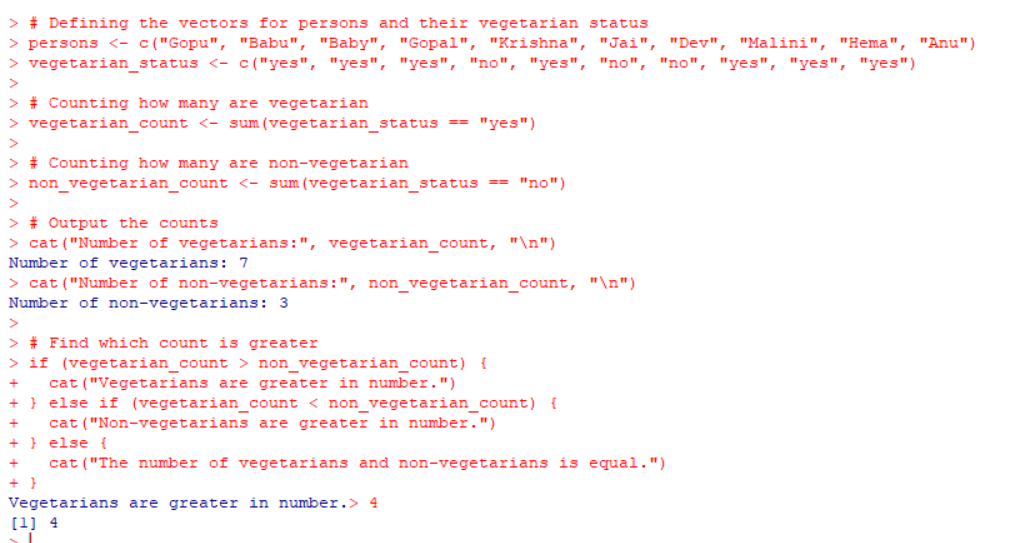
**NAVI BAYES:**

****

****

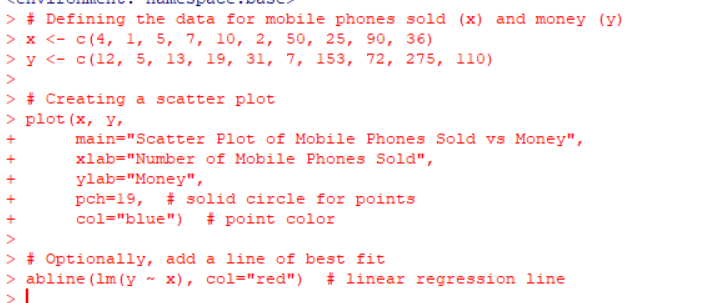
****

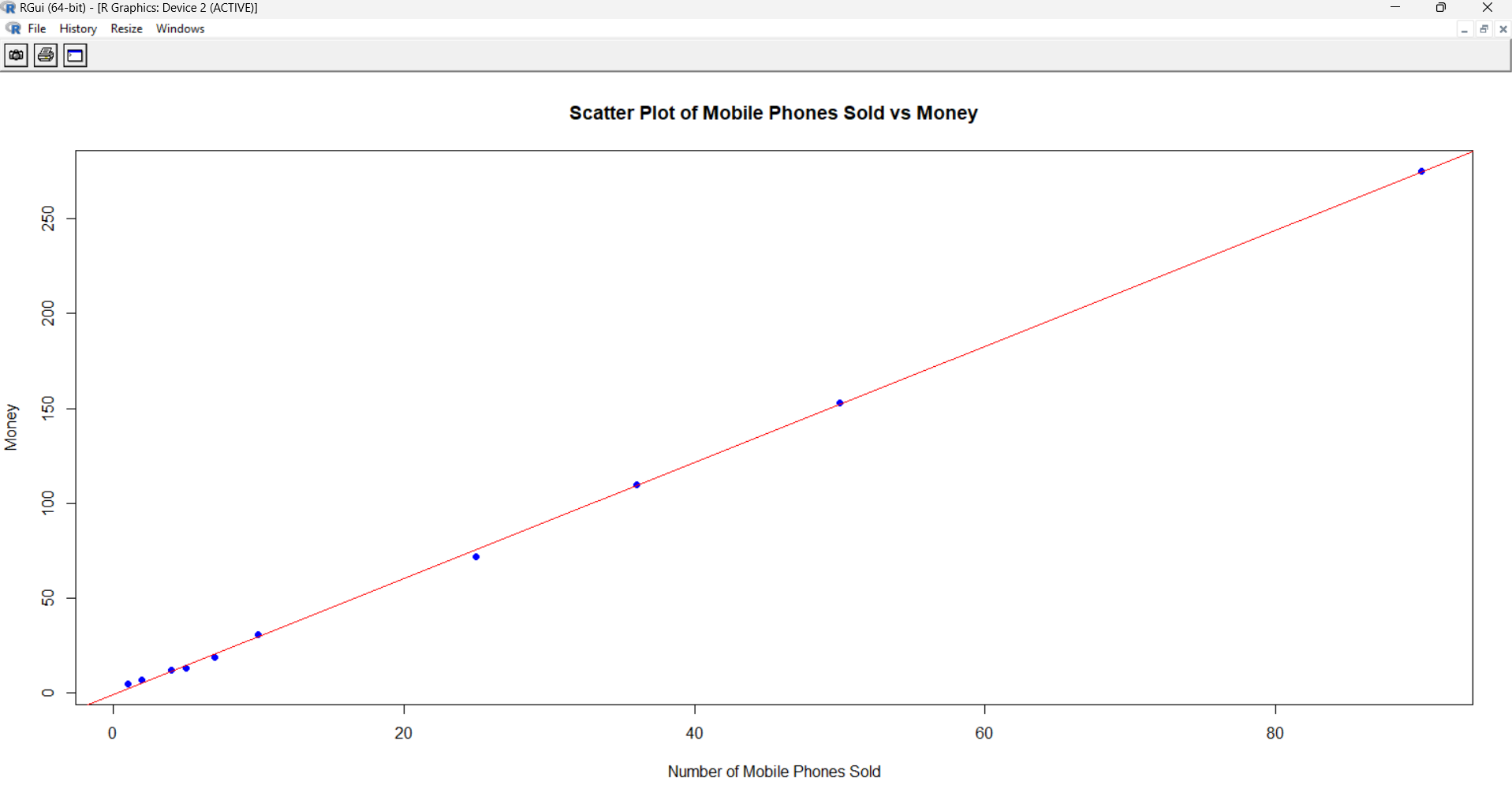
4 QUESTION:The following list of persons with vegetarian or not details given in the table. How will you find out how many of them are vegetarian and how many of them are non-vegetarian? Which type of the person total count is greater value?

****

5 question: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** |

****

****

6 question:.Generate rules using FP growth algorithm using the given dataset which has the following transactions with items purchased: Consider the values as support=50% and confidence=75%.

@relation dataset

@attribute bread{true,false}

@attribute cheese{true,false}

@attribute egg{true,false}

@attribute jucie{true,false}

@attribute yogurat{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

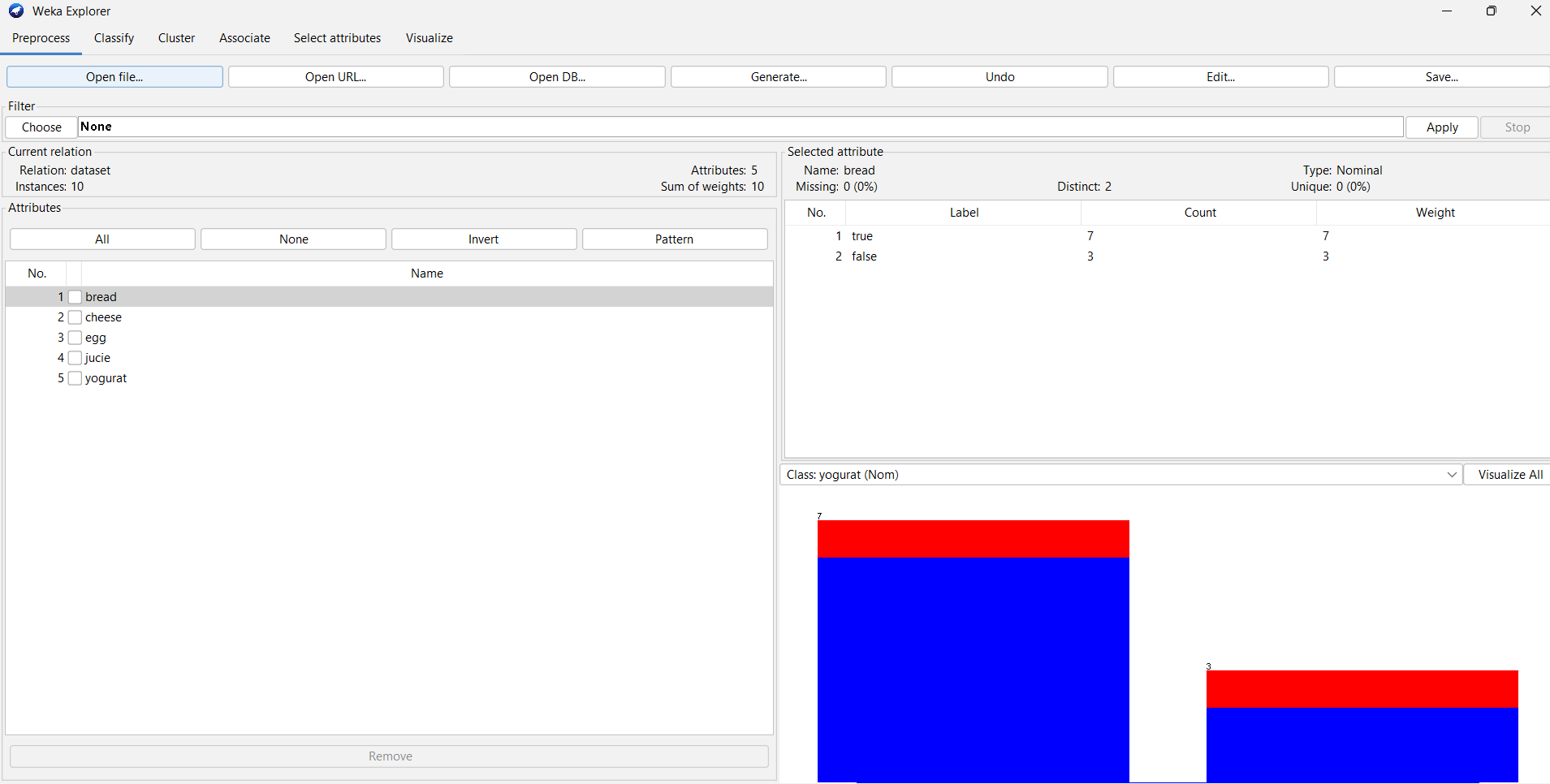
false true false true true

false false true true false

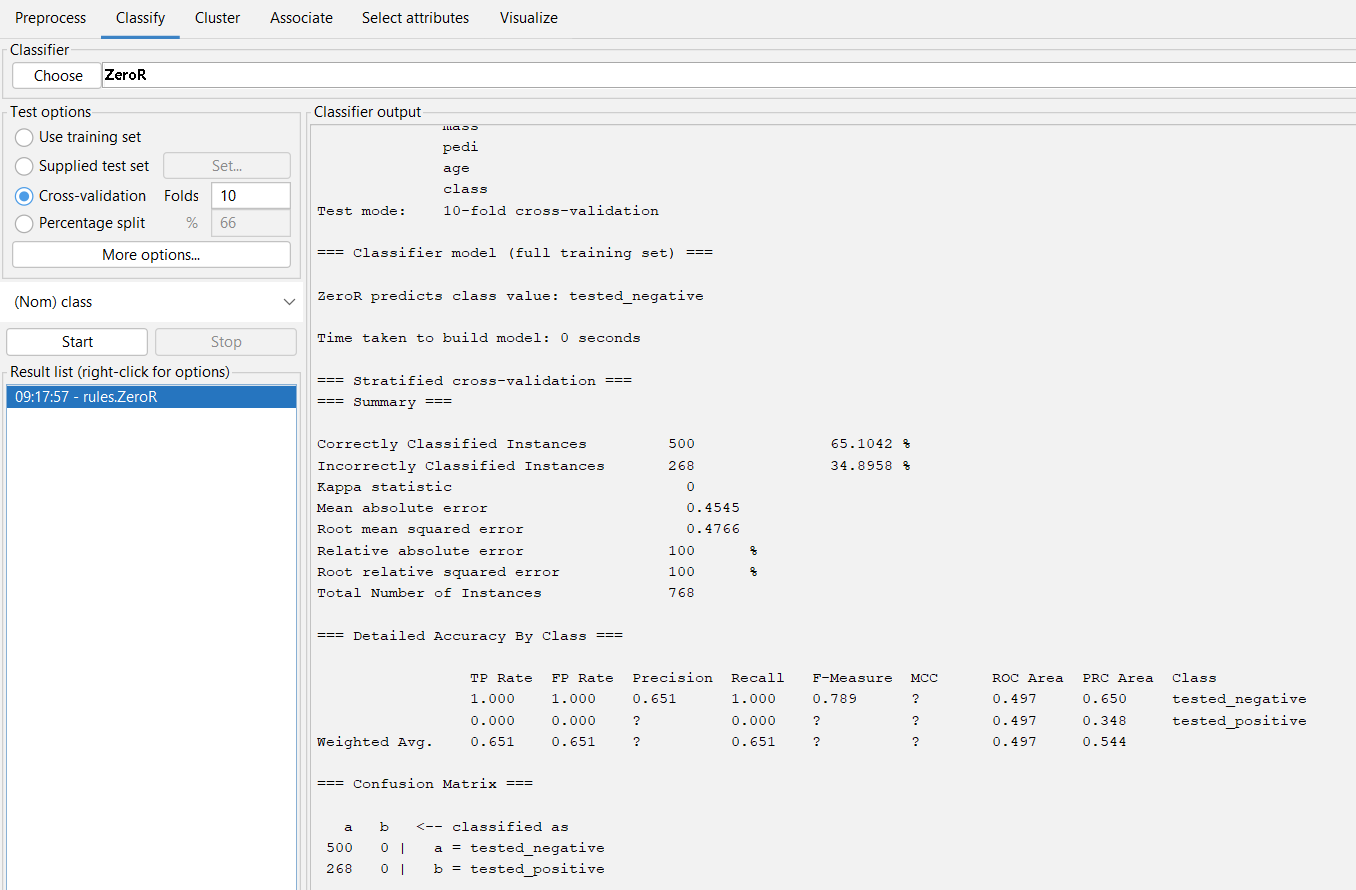
true true true false false

true false false true true

true true false false true

****

**7 question:7.Prediction of Diabetes Data using Decision tree classifier in WEKA. Compare it with Support Vector Machine classifier. Show the result accuracy and F1 measure calculation .Plot the graph and explain the summary of results.**

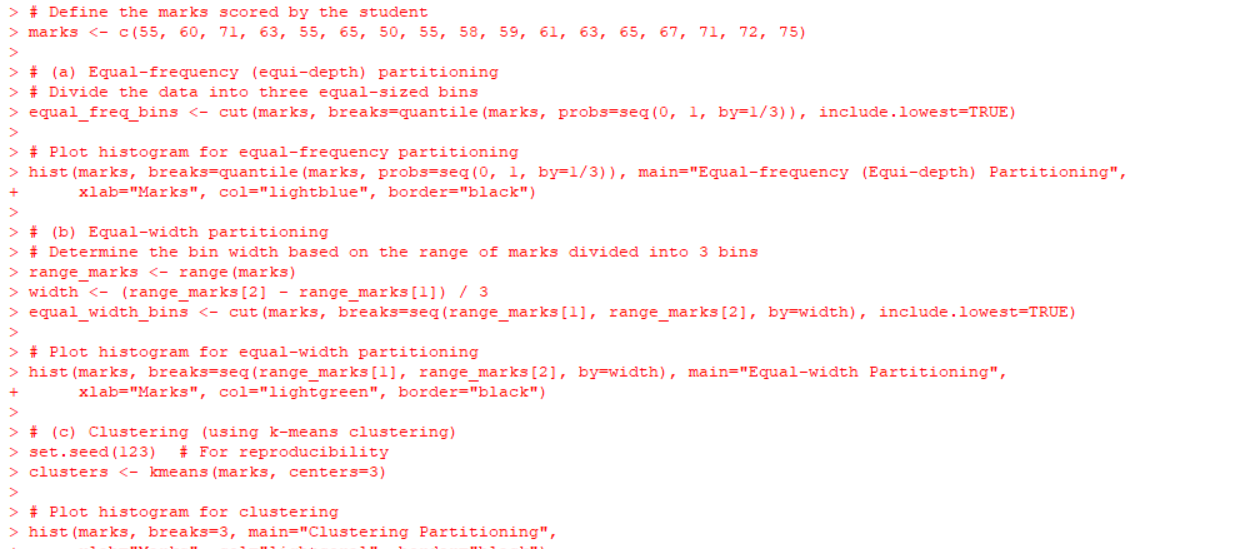
****

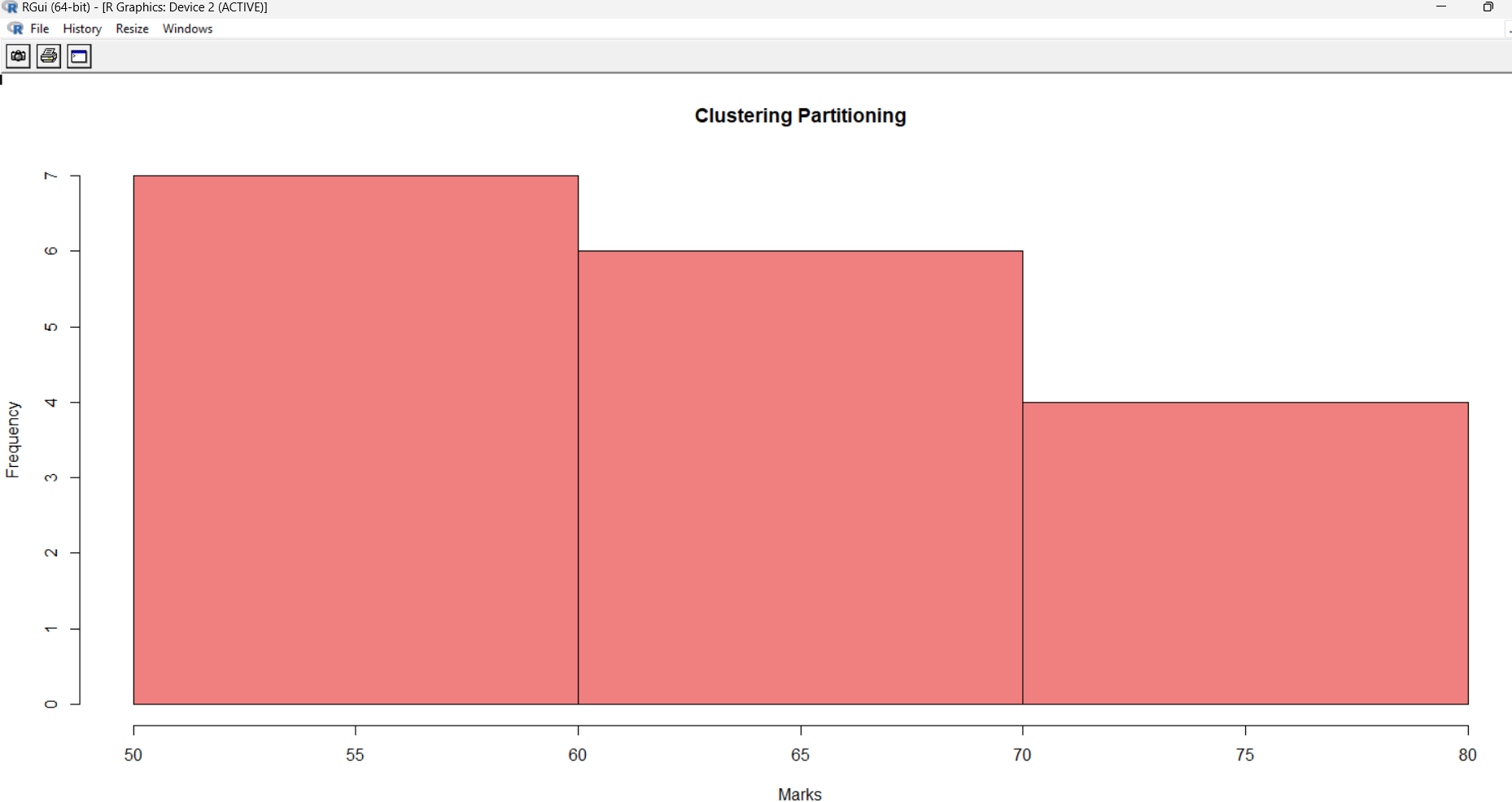
8 question:8.Implement of the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75. Partition them into three bins by each of the following methods. Plot the data points using histogram.

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

(b) equal-width partitioning

(c) clustering

****

****

**9 question:**

10 question.Create an ARFF file for the table below and implement for the Apriori Algorithm and FP growth algorithm and compare the rules generated by both the algorithms. Identify the unique rules generated by the above algorithms.

@relation dataset

@attribute sony{true,false}

@attribute bpl{true,false}

@attribute lg{true,false}

@attribute samsung{true,false}

@attribute onida{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

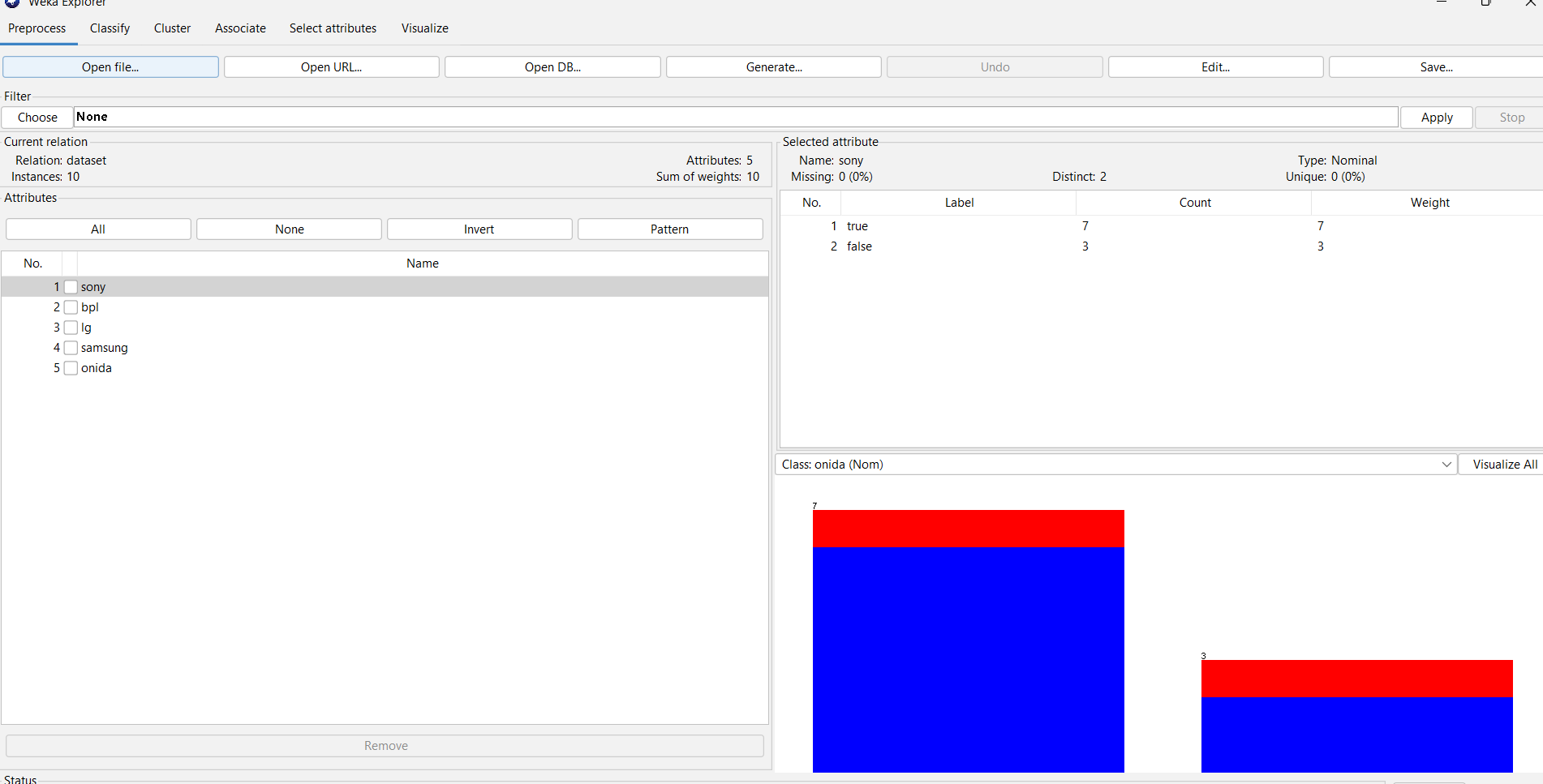
false true false true true

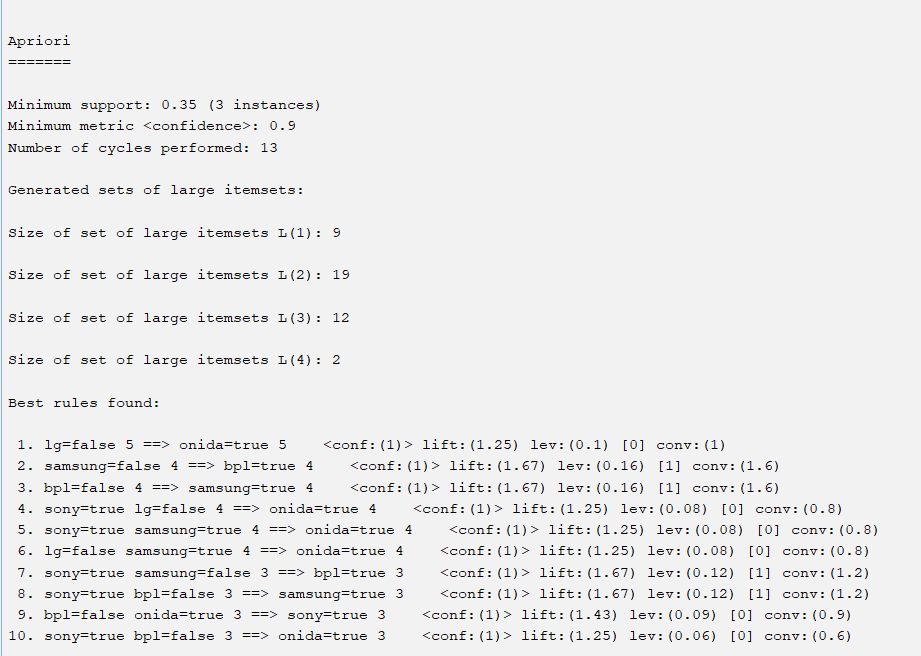
false false true true false

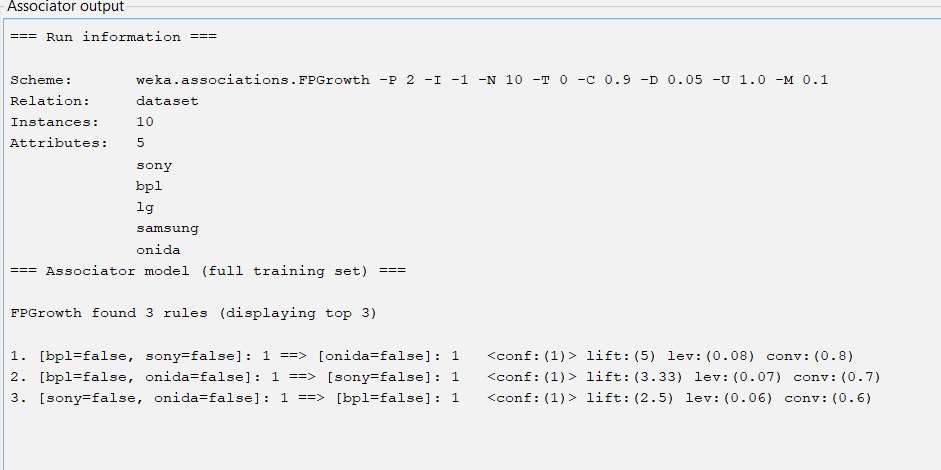
true true true false false

true false false true true

true true false false true

****

****

****

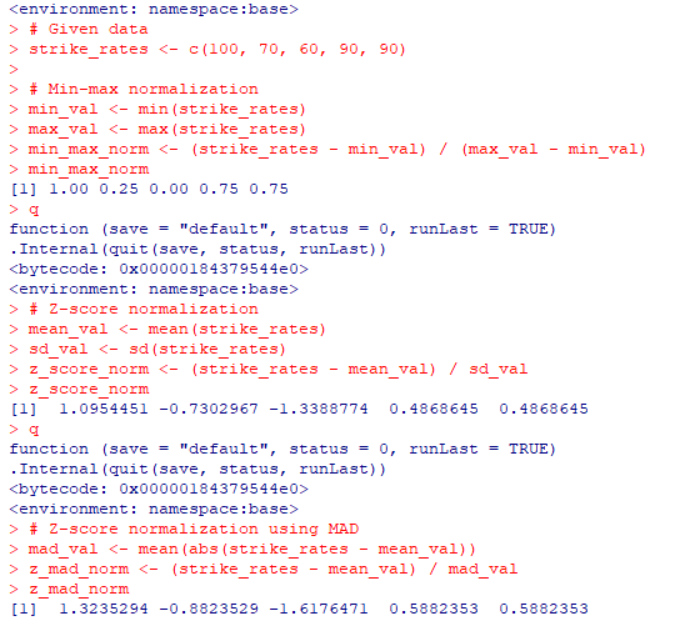
11 question:The given are the strike-rates scored by a batsman in season 1 in different tournaments. 100, 70, 60, 90, 90

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

(b) z-score normalization

(c) z-score normalization using the mean absolute deviation instead of standard deviation

(d) normalization by decimal scaling

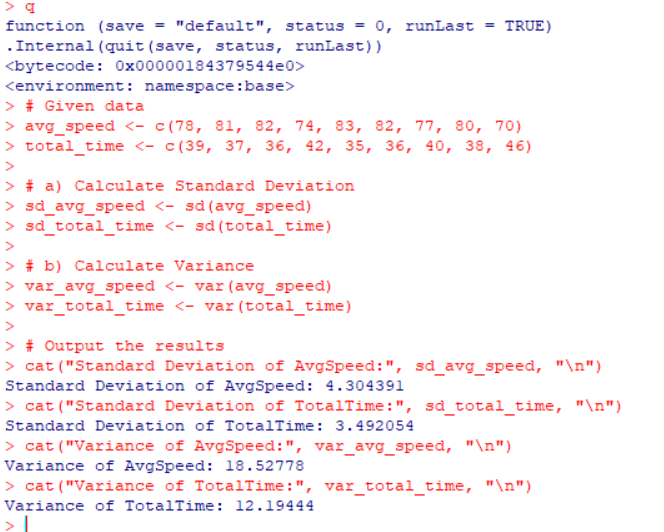
****

12 question: Suppose some car is tested for the AvgSpeed and TotalTime data for 9 randomly selected car with the following result

| AvgSpeed  (in kph) | 78 | 81 | 82 | 74 | 83 | 82 | 77 | 80 | 70 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TotalTime  (in mins) | 39 | 37 | 36 | 42 | 35 | 36 | 40 | 38 | 46 |

a) Calculate the standard deviation of AvgSpeed and TotalTime.

b) Calculate the Variance of AvgSpeed and TotalTime for the above dataset.

****

13 question:Consider this table

a) TID items bought

b) T100 {M, O, N, K, E, Y}

c) T200 {D, O, N, K, E, Y }

d) T300 {M, A, K, E}

e) T400 {M, U, C, K, Y}

f) T500 {C, O, O, K, I ,E}

g) (a) Find all frequent item set using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

h) (b) List all of the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and itemi denotes variables representing items (e.g., “A”, “B”, etc.):

i) ∀x ∈ transaction, buys(X, item1) ∧ buys(X, item2) ⇒ buys(X, item3)

@relation dataset

@attribute Monkey{true,false}

@attribute donkey{true,false}

@attribute make{true,false}

@attribute mucky{true,false}

@attribute cookie{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

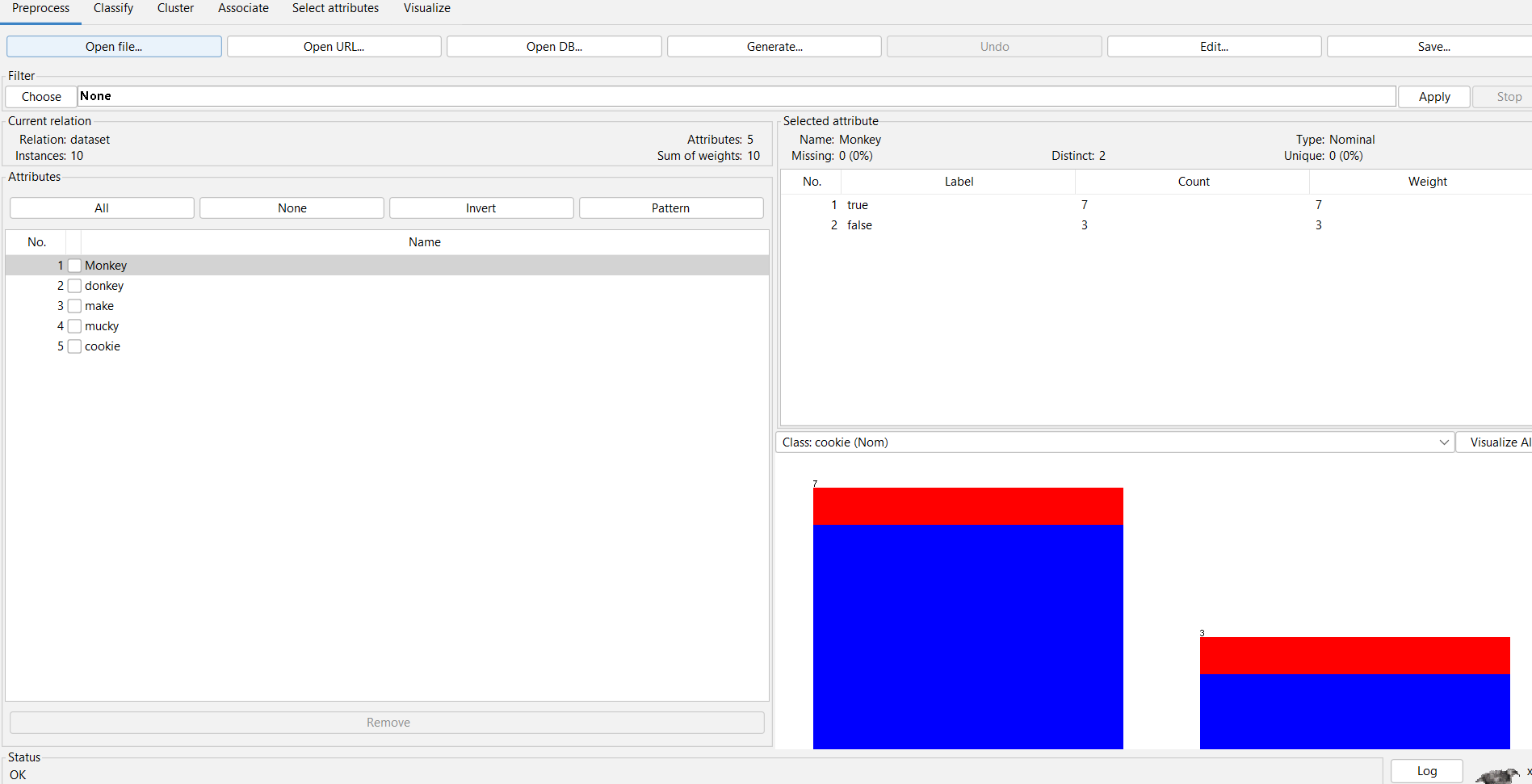
false true false true true

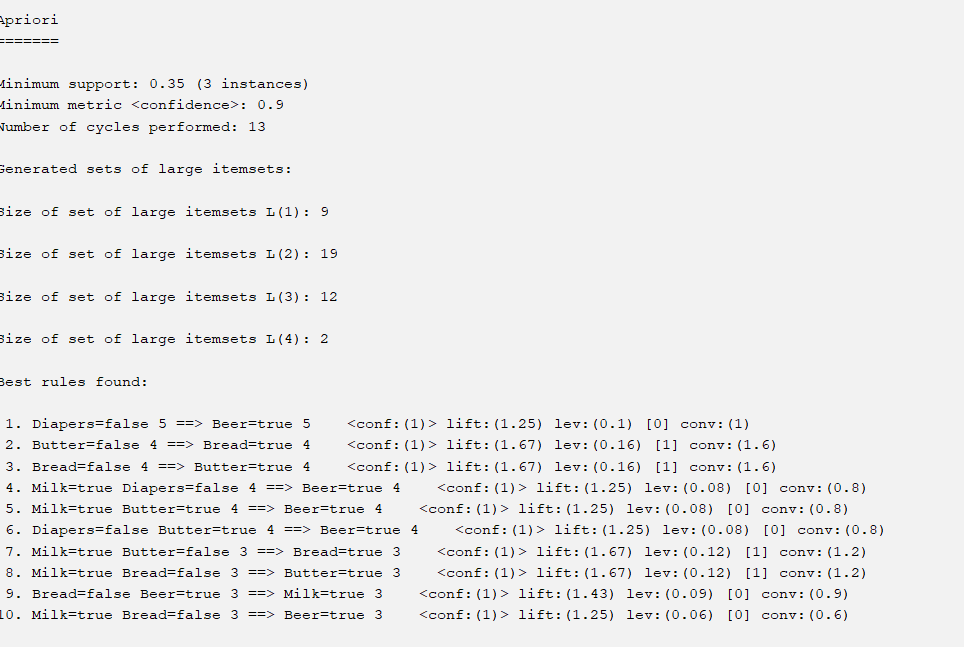
false false true true false

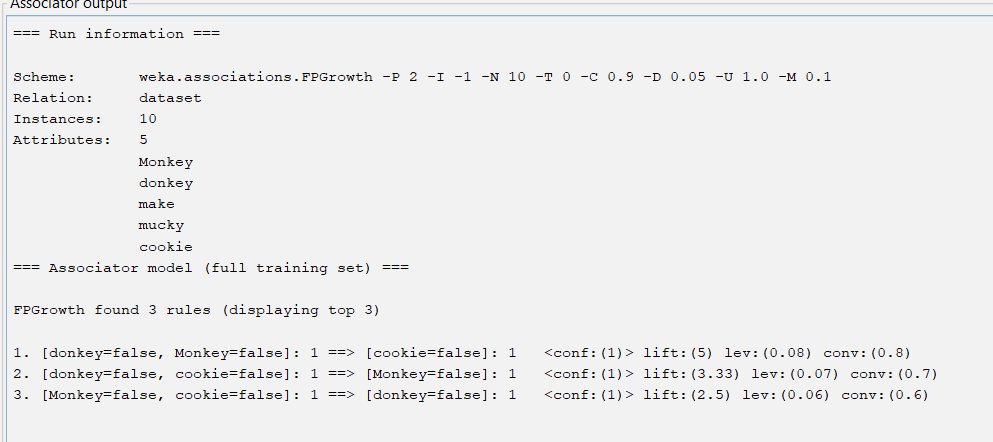
true true true false false

true false false true true

true true false false true

****

****

****