**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

**Input:**

#age, frequency

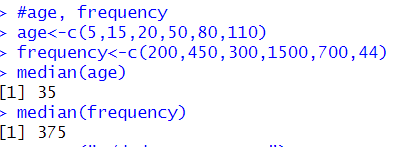
age<-c(5,15,20,50,80,110)

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

median(age)

median(frequency)

**output:**



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?

**Input:**

#mean,median,mode,quatile

age<-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(age)

median(age)

mode\_age<-names(table(age))[table(age)==max(table(age))]

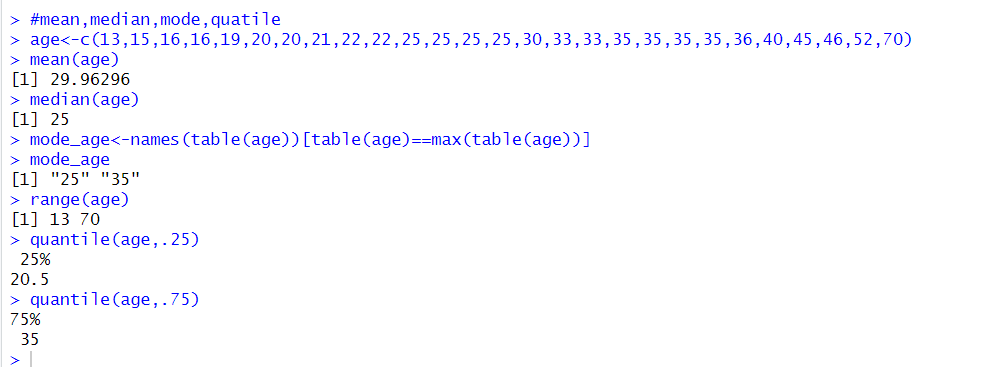
mode\_age

range(age)

quantile(age,.25)

quantile(age,.75)

**output:**



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

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

**input:**

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)

bins <- 5

bin\_indices <- cut(data, bins)

mean\_smooth <- tapply(data, bin\_indices, mean)

print(mean\_smooth)

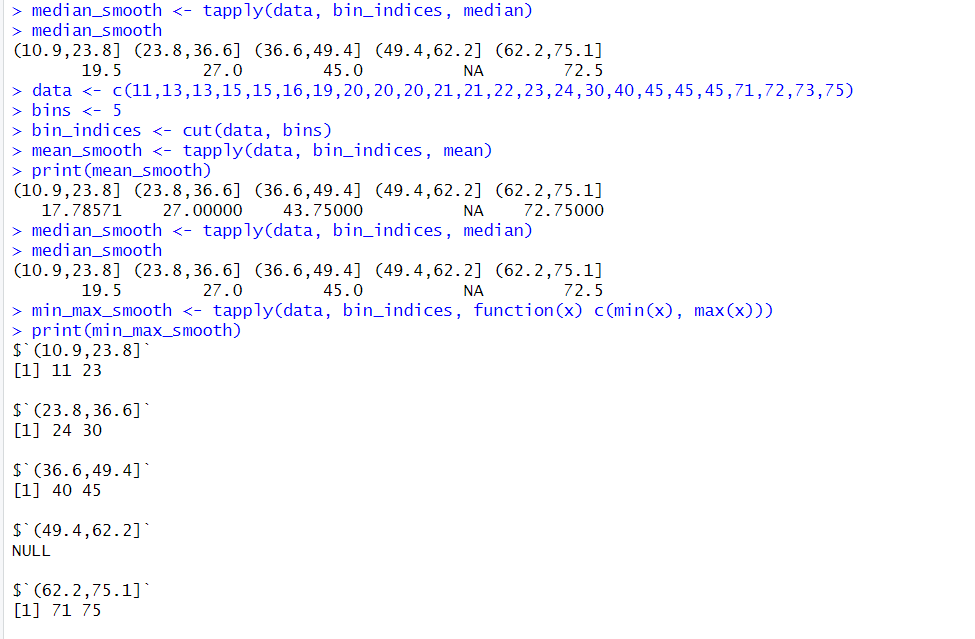
median\_smooth <- tapply(data, bin\_indices, median)

median\_smooth

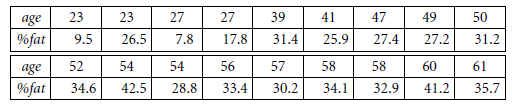
min\_max\_smooth <- tapply(data, bin\_indices, function(x) c(min(x), max(x)))

print(min\_max\_smooth)

**output:**



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



1. 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.

**Input:**

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

fat<-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)

mean(age)

median(age)

sd(age)

mean(fat)

median(fat)

sd(fat)

#boxplot

boxplot(age,fat)

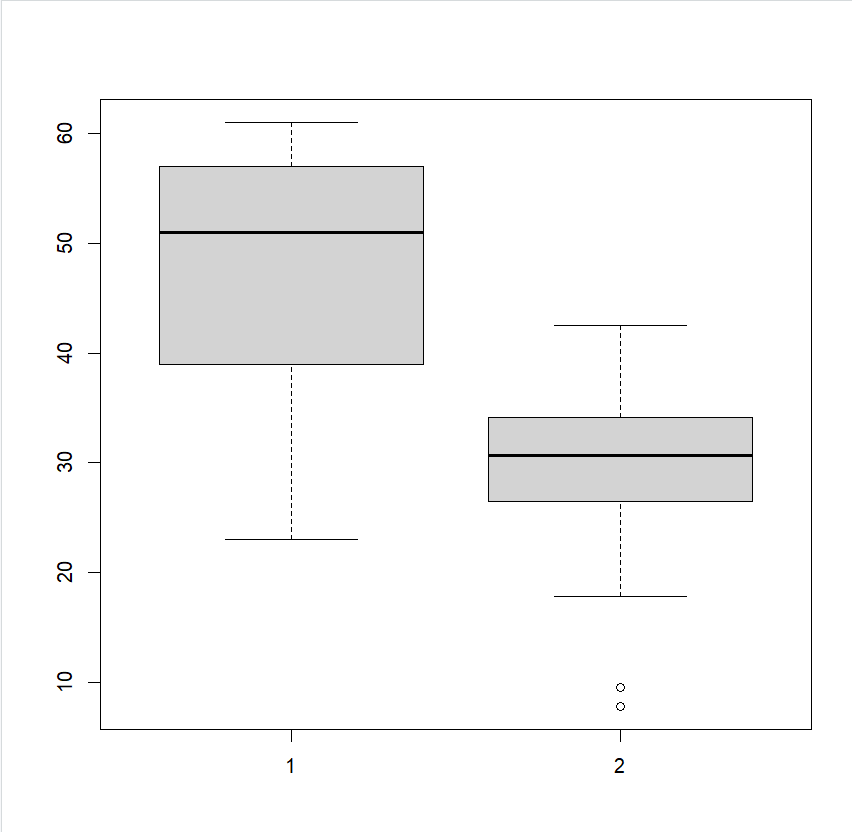
#scatter plot

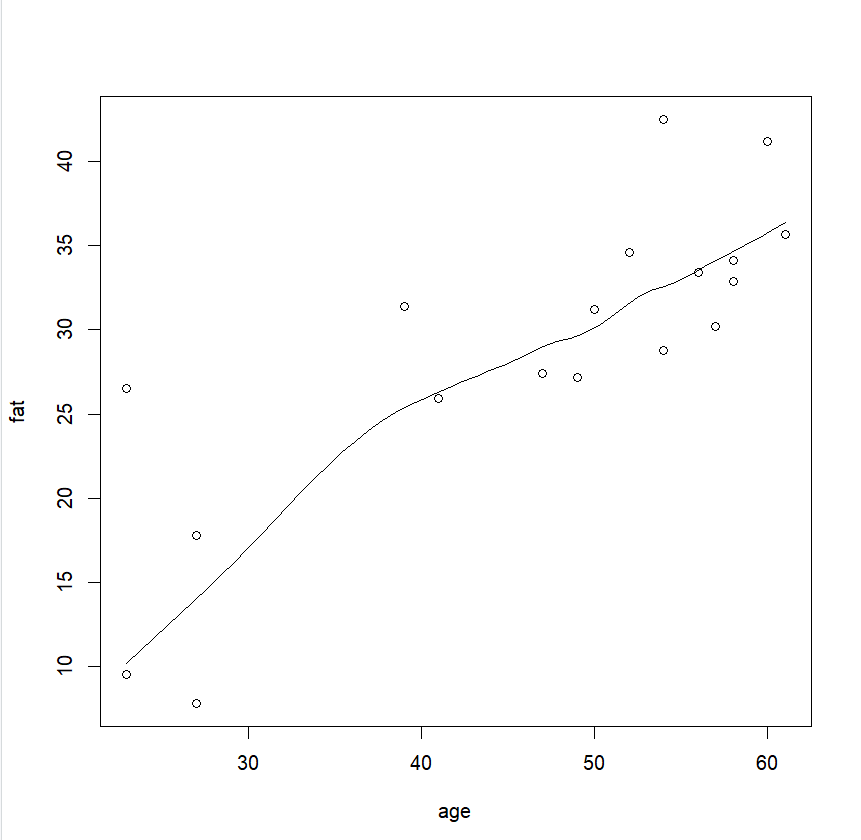
scatter.smooth(age,fat)

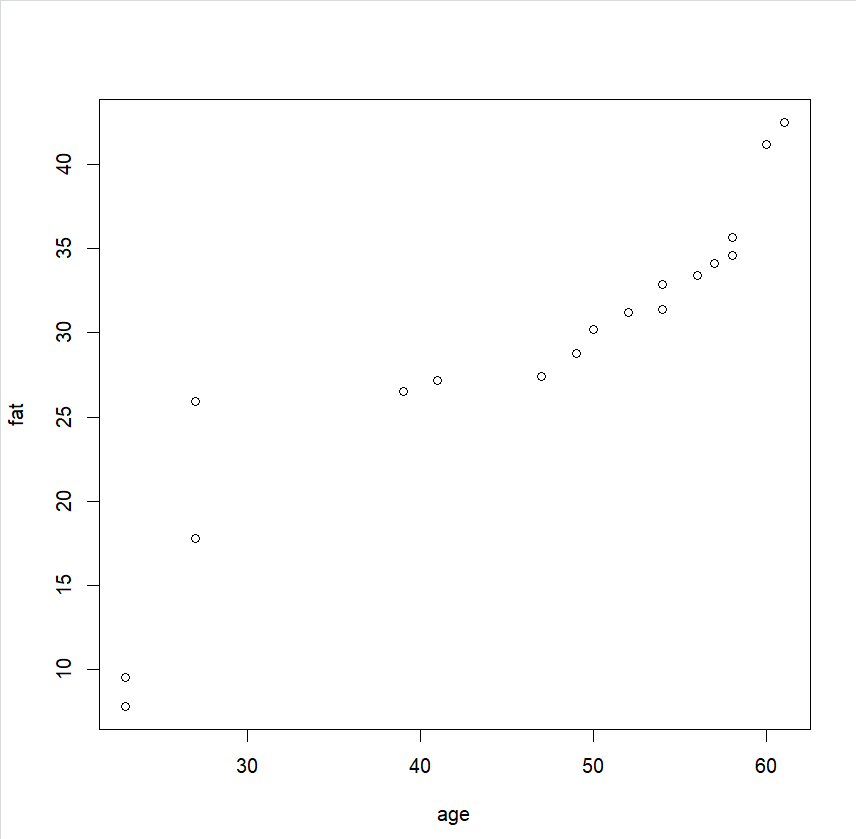
#qplot

qqplot(age,fat)

**output:**







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

**Input:**

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

min<-0

max<-1

#min\_max

min\_max=((35-min(v))/(max(v)-min(v)))

print(min\_max)

#z-score

m=mean(v)

s<-12.94

z\_score=(35-m)/s

print(z\_score)

#decimal scaling

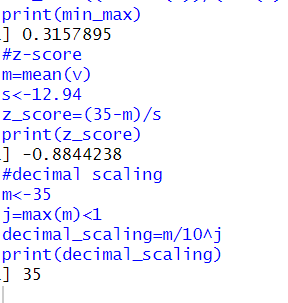
m<-35

j=max(m)<1

decimal\_scaling=m/10^j

print(decimal\_scaling)

**output:**



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

**Input:**

pencils<-c(9,25,23,12,11,6,7,8,9,10)

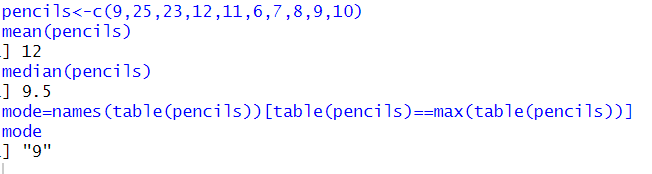
mean(pencils)

median(pencils)

mode=names(table(pencils))[table(pencils)==max(table(pencils))]

mode

**output:**



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

**input:**

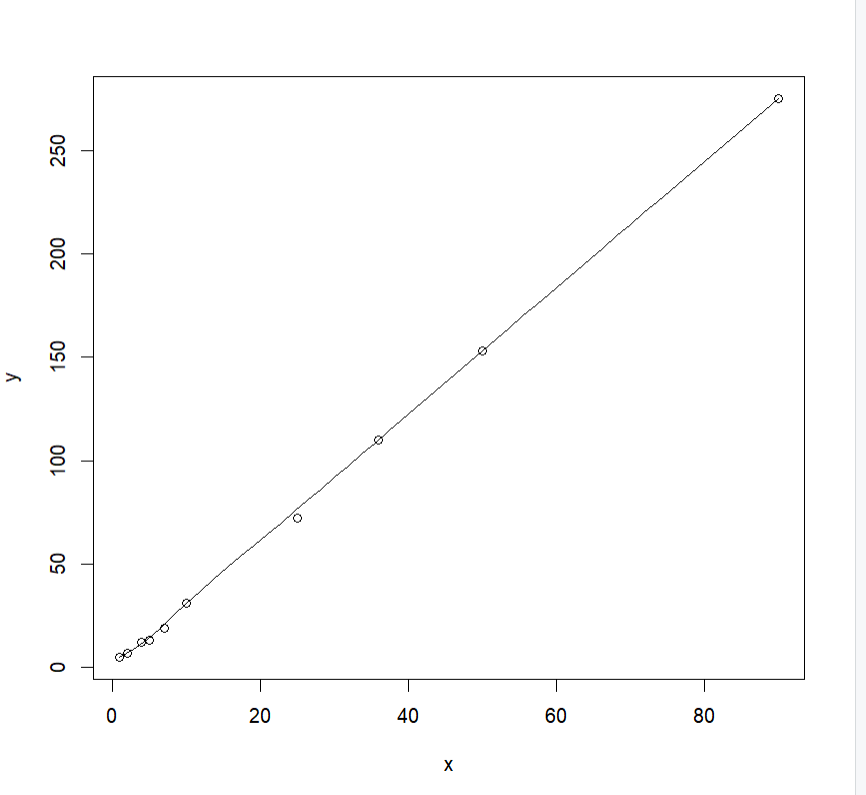
#scatterplot

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

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

scatter.smooth(x,y)

**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

**Input:**

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

num\_bins <- 3

bins\_eq\_frequency <- cut(marks, breaks = num\_bins, labels = FALSE)

hist(marks, breaks = num\_bins, col = "lightblue", xlab = "Marks", main = "Equal-Frequency (Equi-Depth) Partitioning")

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

bin\_mean <- tapply(data, cut(data, num\_bins), mean)

smoothed\_data\_by\_mean <- unname(bin\_mean[as.character(cut(data, num\_bins))])

bin\_median <- tapply(data, cut(data, num\_bins), median)

smoothed\_data\_by\_median <- unname(bin\_median[as.character(cut(data, num\_bins))])

bin\_boundaries <- tapply(data, cut(data, num\_bins), function(x) c(min(x), max(x)))

smoothed\_data\_by\_boundaries <- unlist(bin\_boundaries[as.character(cut(data, num\_bins))])

print("Original data:")

print(data)

print("Smoothed data by bin mean:")

print(smoothed\_data\_by\_mean)

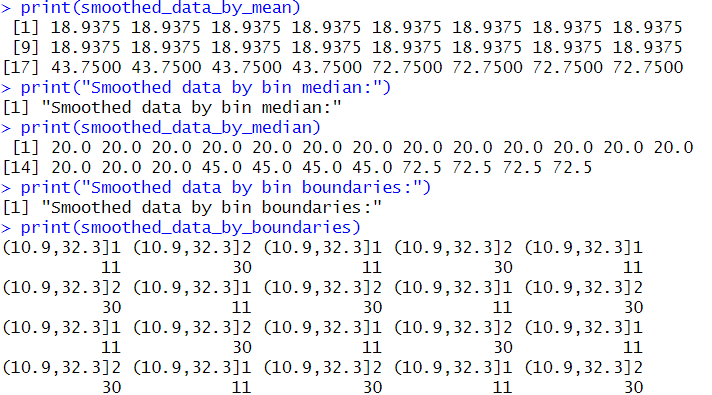
print("Smoothed data by bin median:")

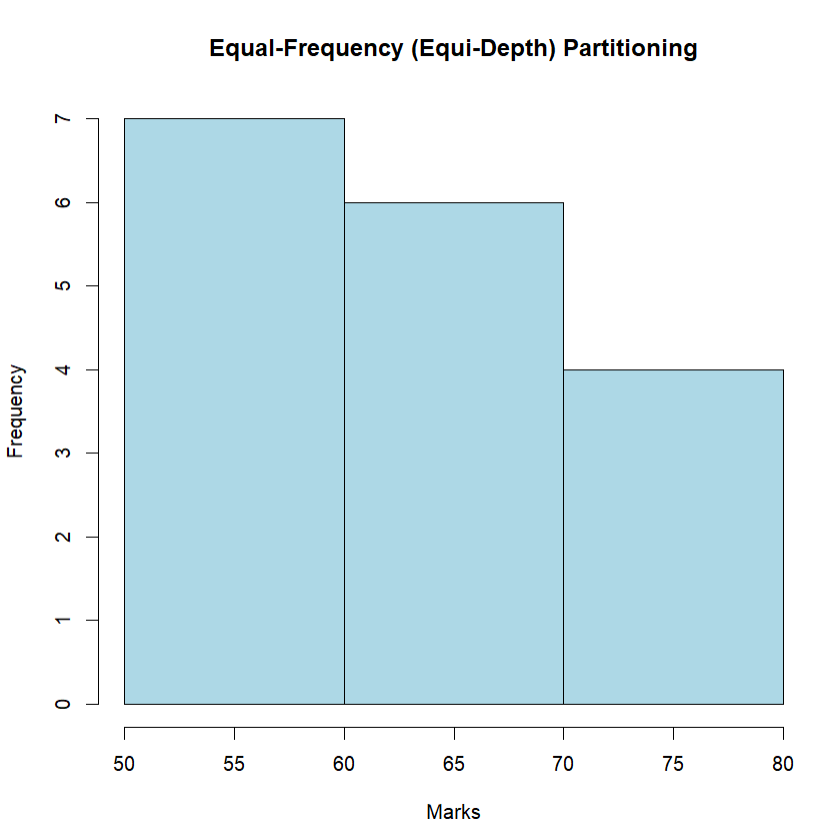
print(smoothed\_data\_by\_median)

print("Smoothed data by bin boundaries:")

print(smoothed\_data\_by\_boundaries)

**output:**





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.

**Input:**

#IQR, SD

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

IQR(v)

sd(v)

**output:**



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?

**Input:**

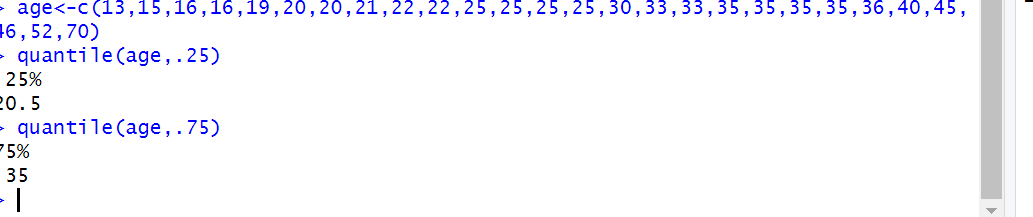
#Q1, Q2

age<-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,.25)

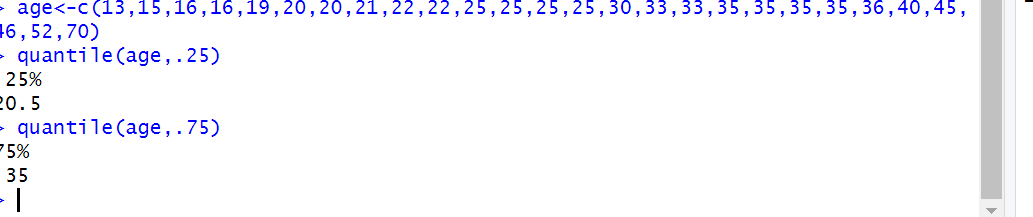
quantile(age,.75)

**output:**

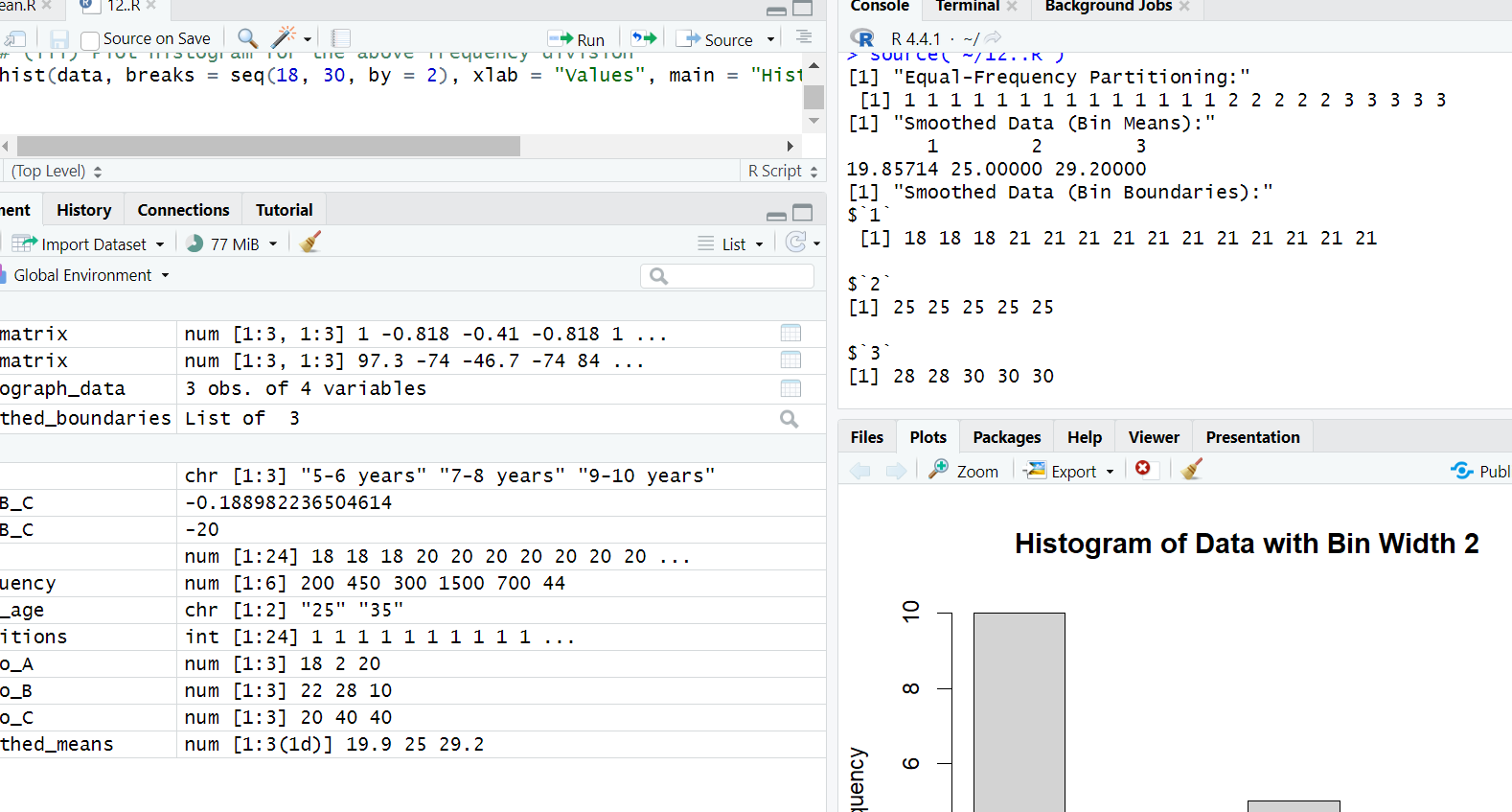


**Day 2**

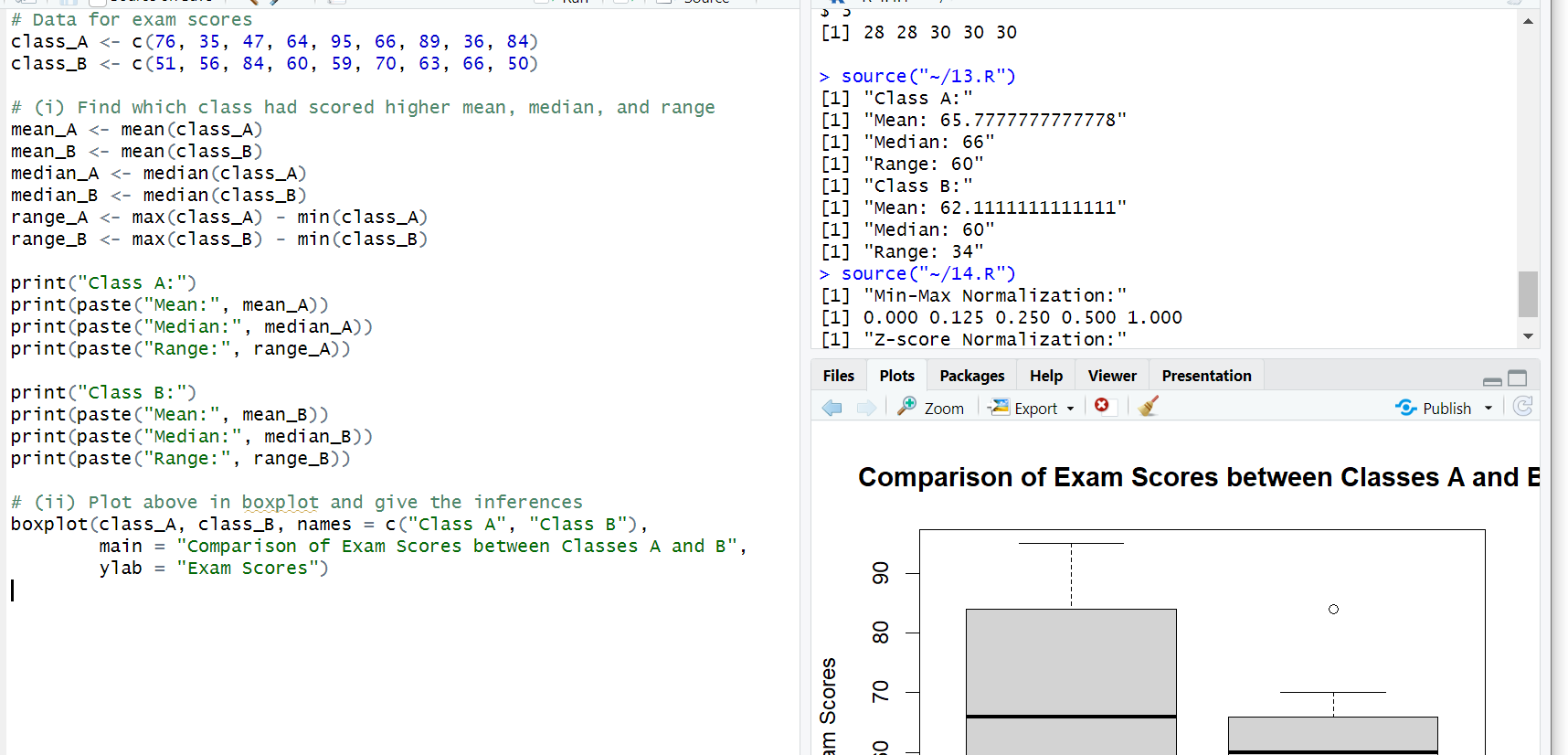
1



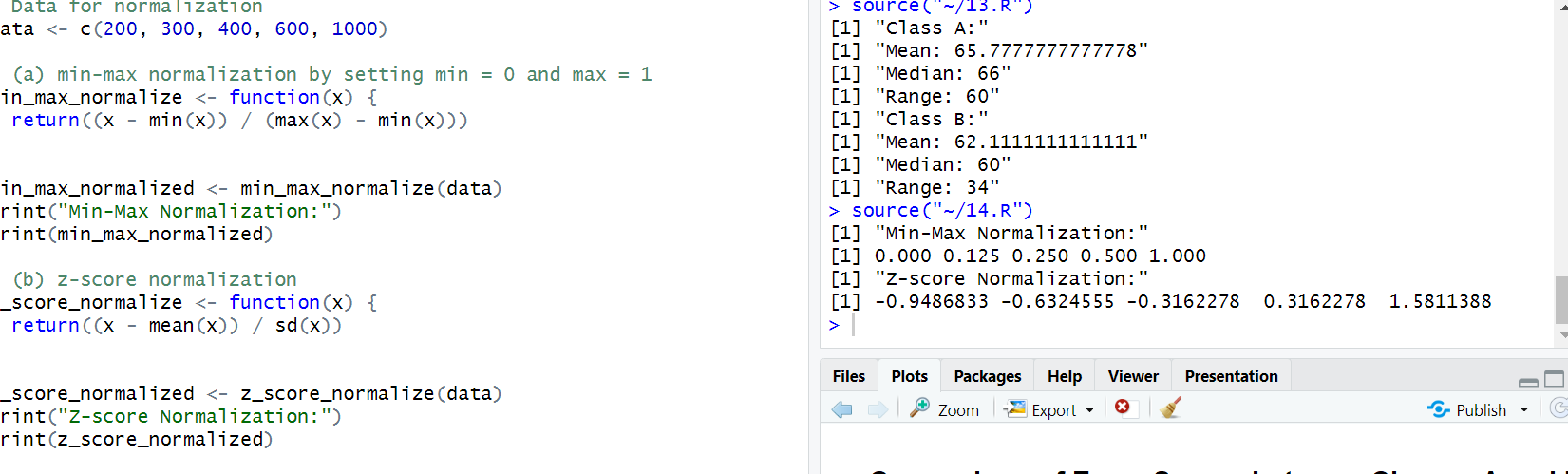
2



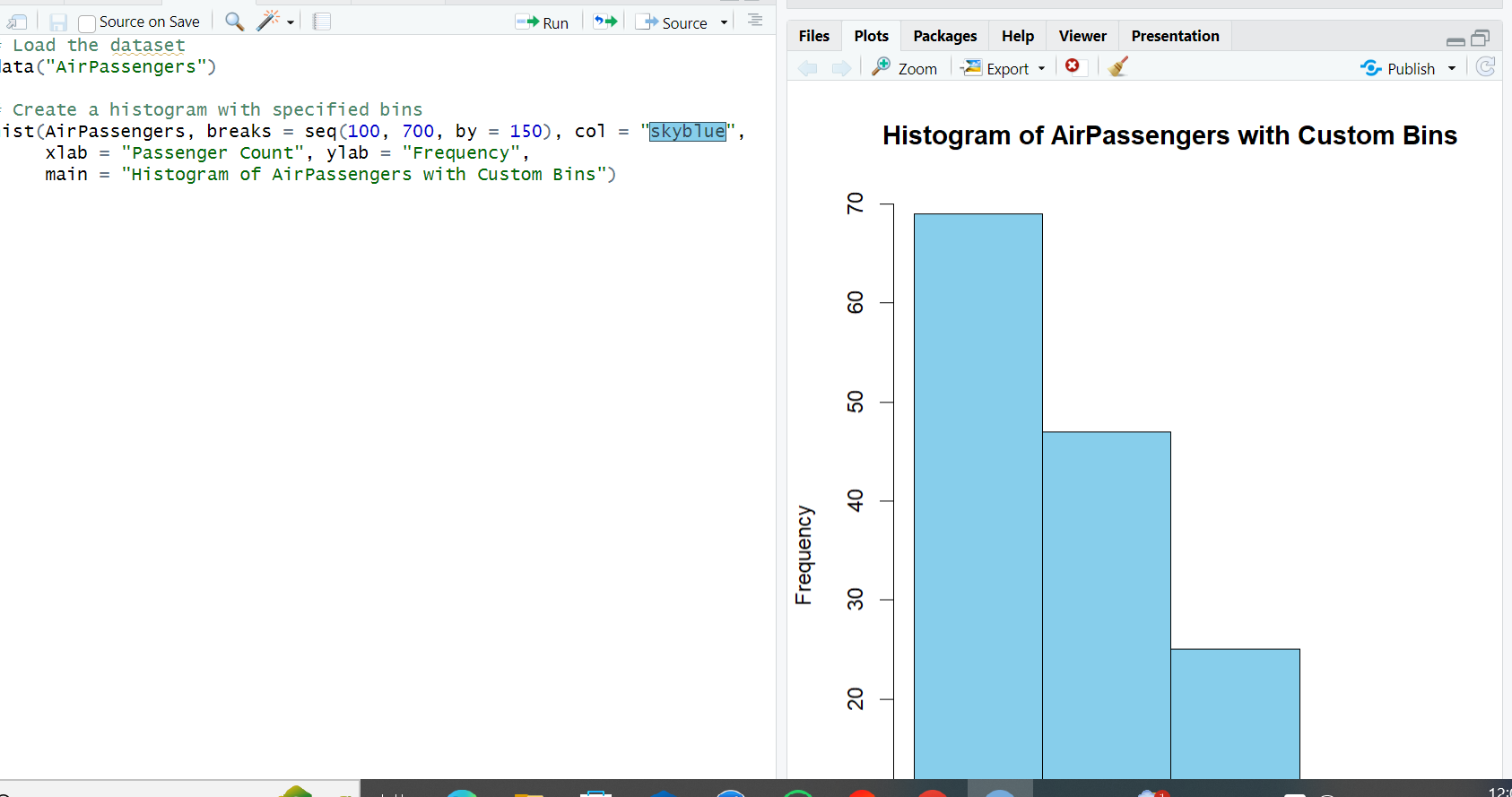
3



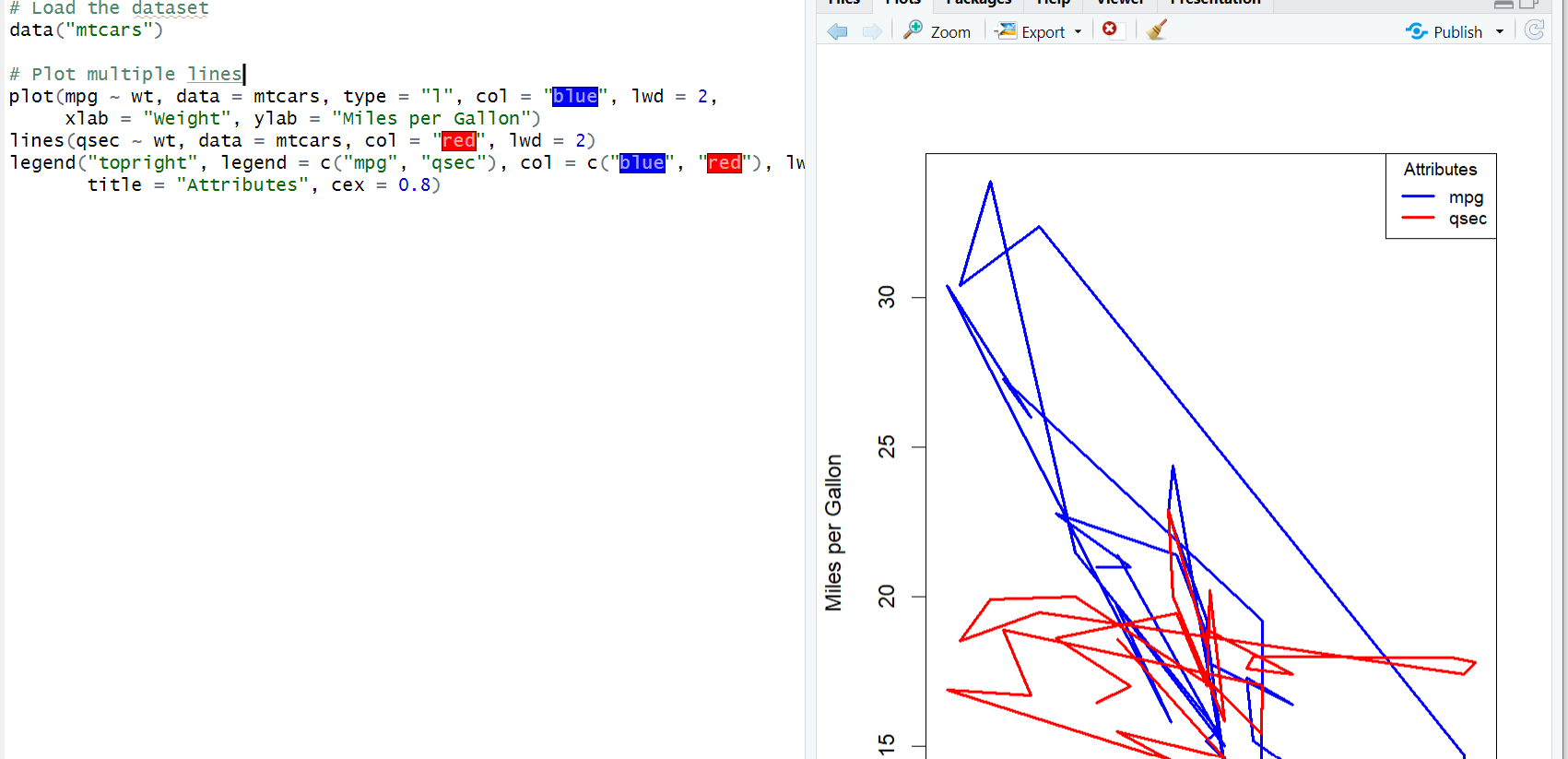
4



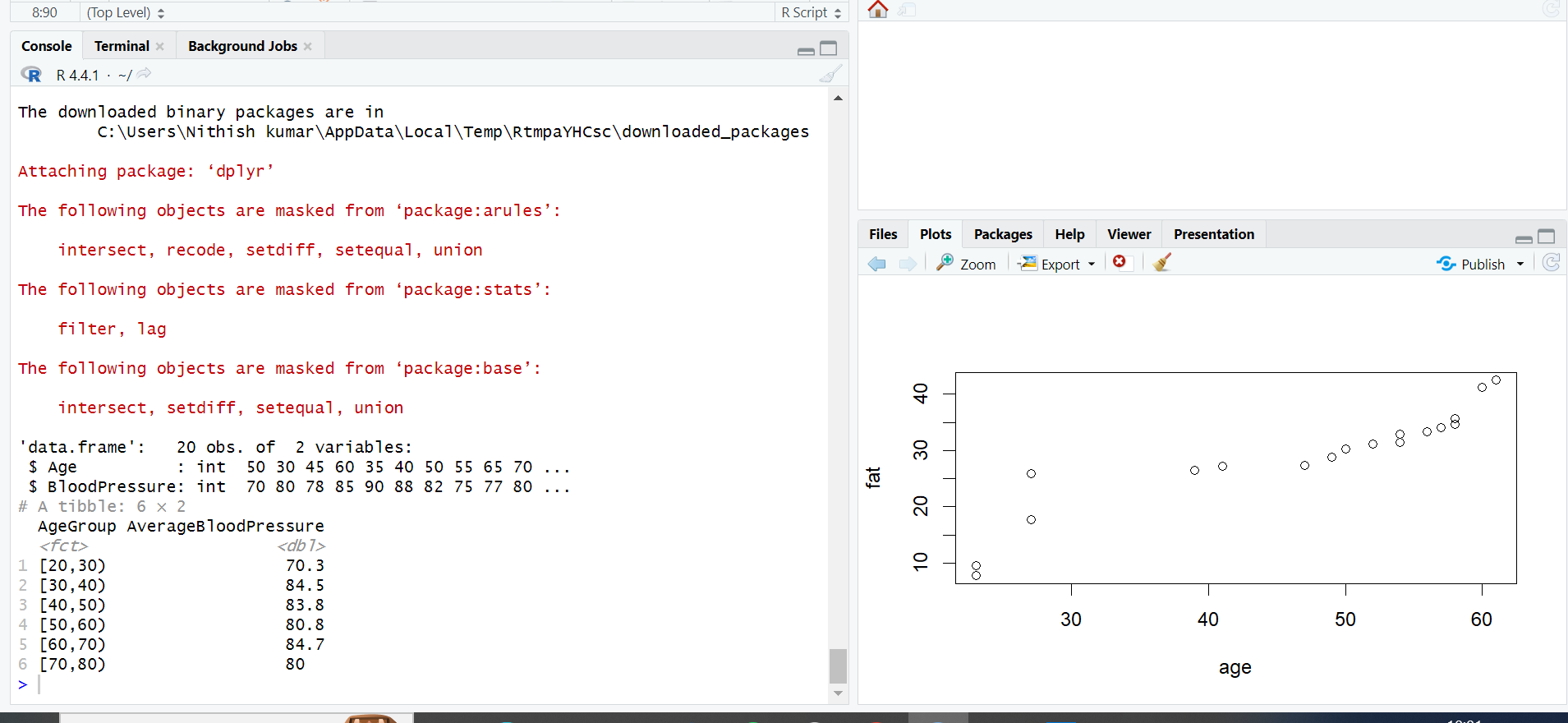
5



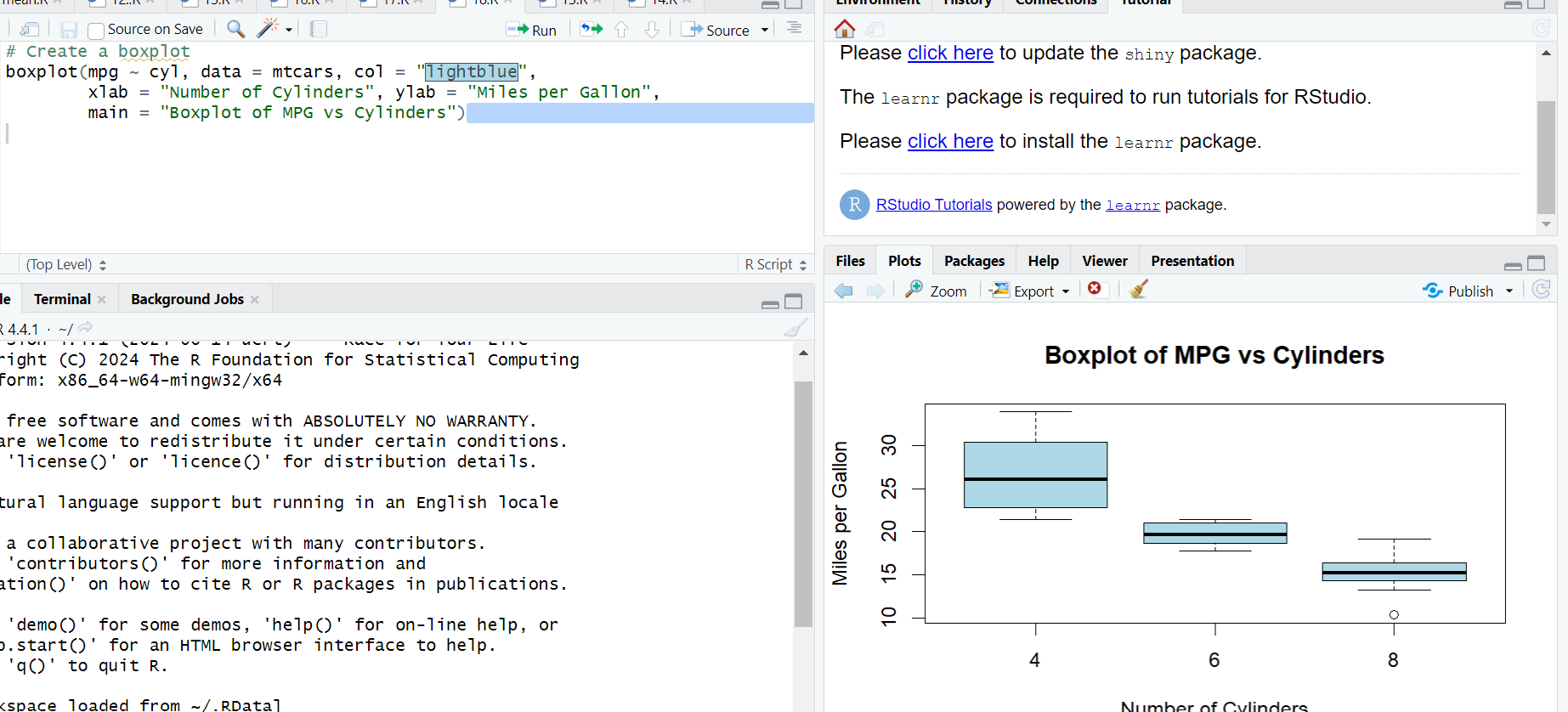
6



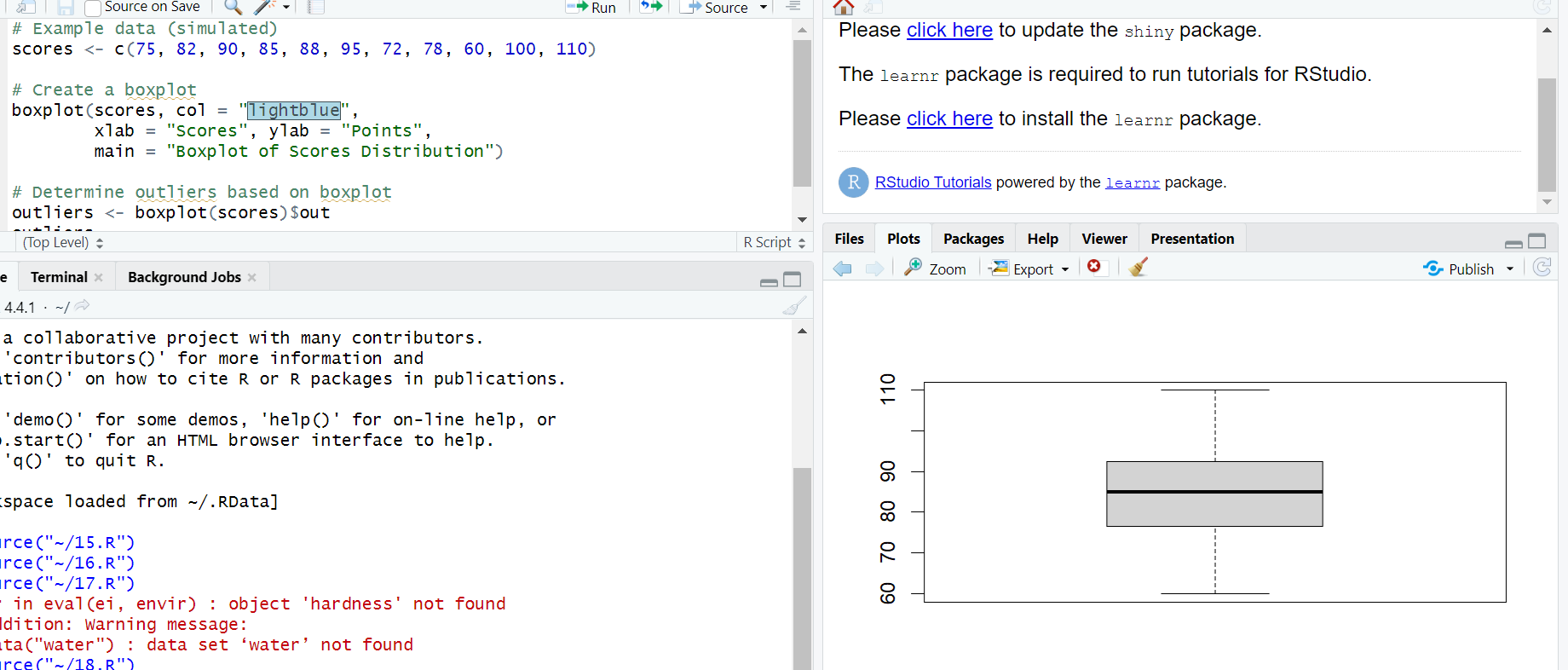
7



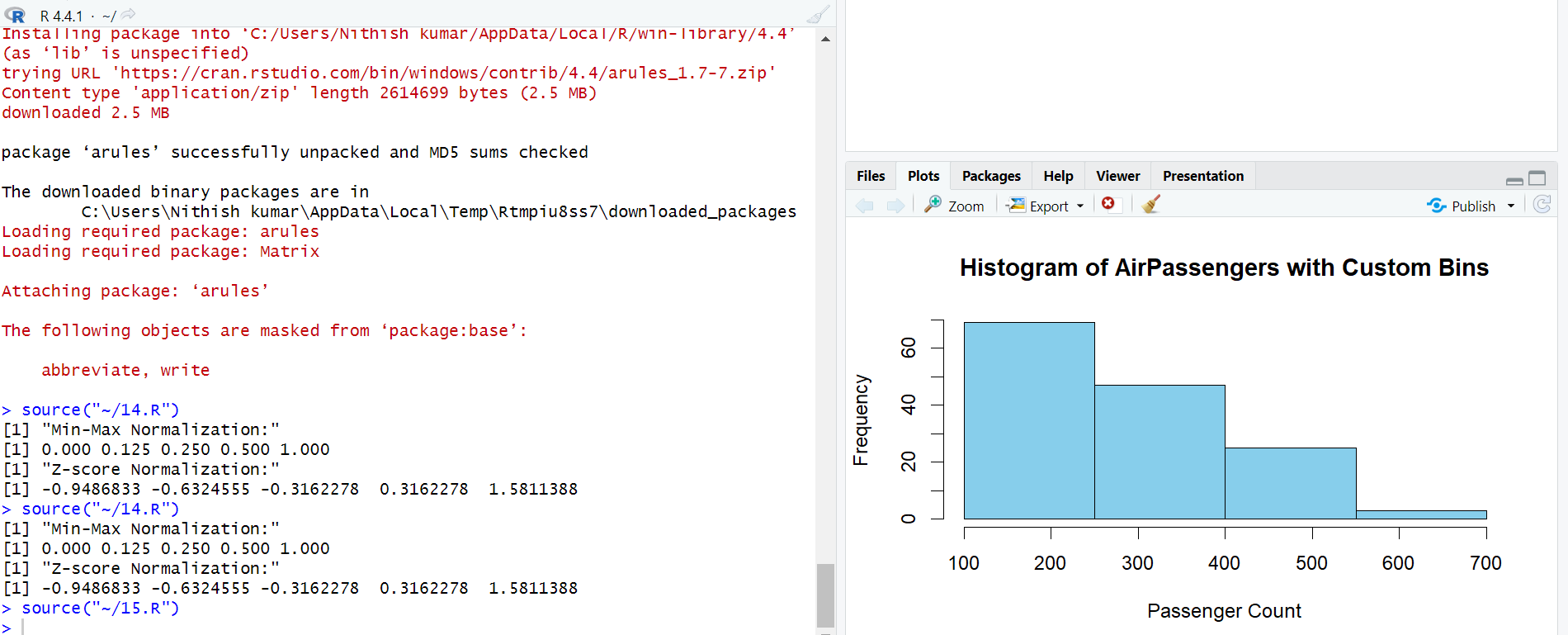
8



9

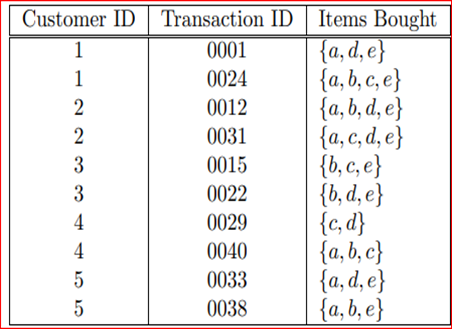


10



**DAY 3**

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



**Dataset:** (sample.arff)

@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

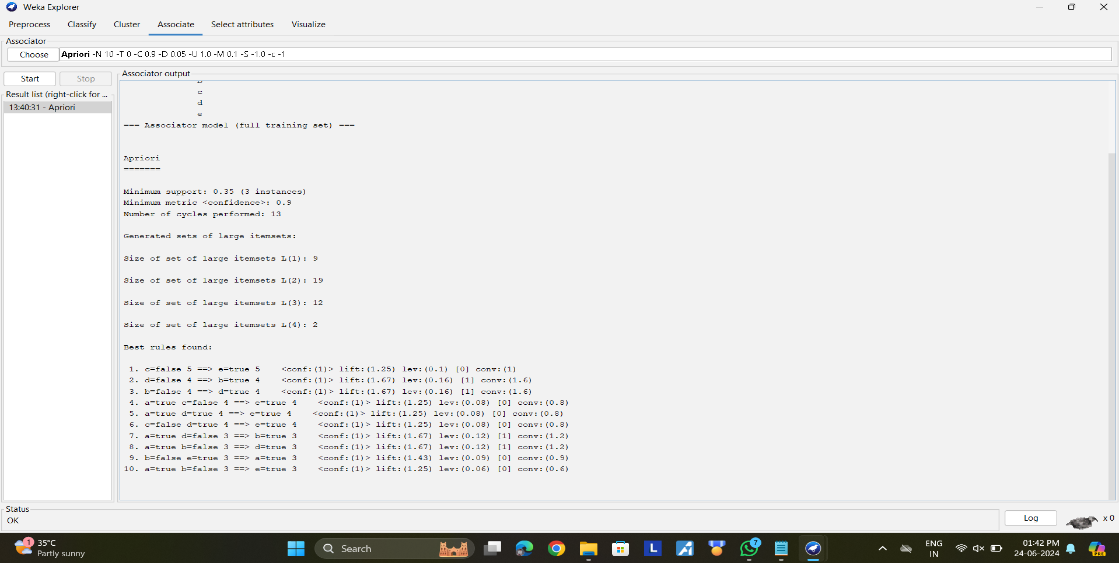
false false true true false

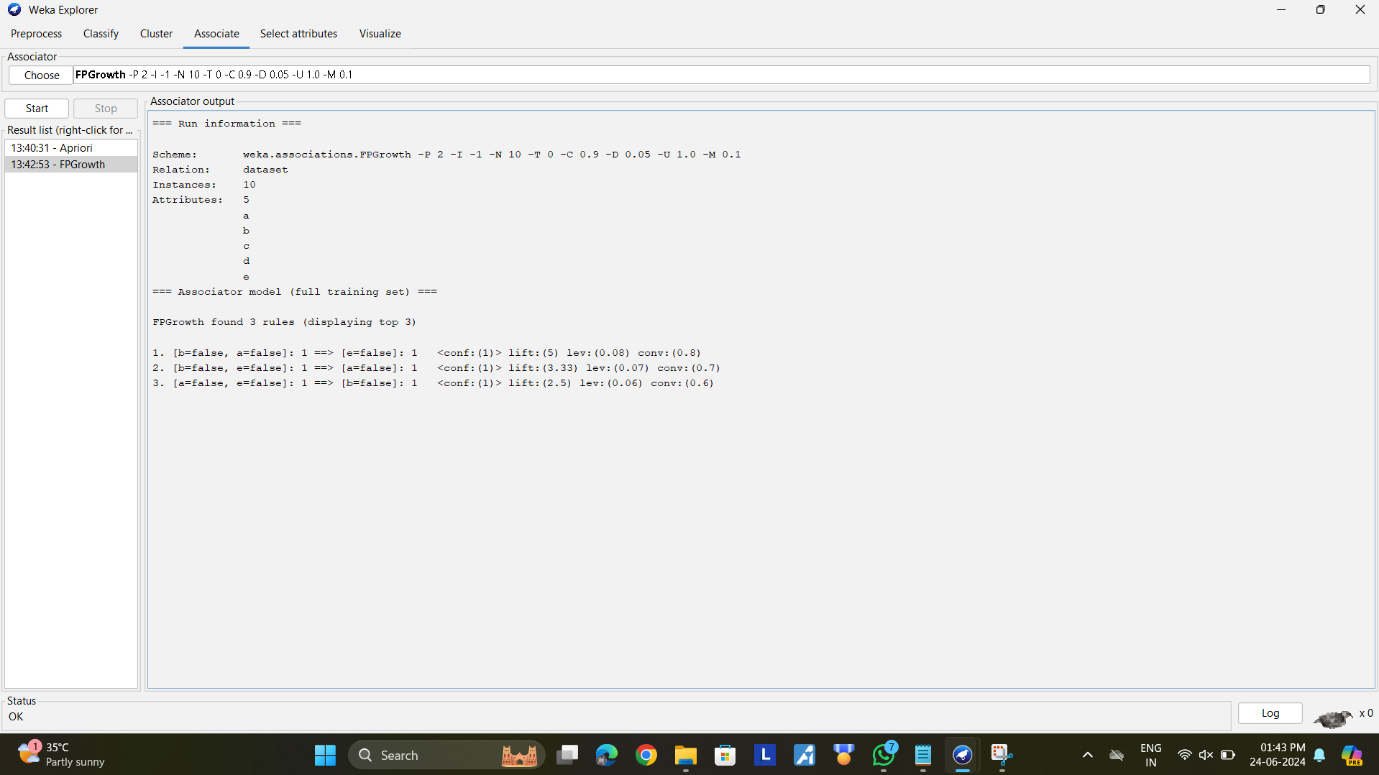
true true true false false

true false false true true

true true false false true

**Output:**





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

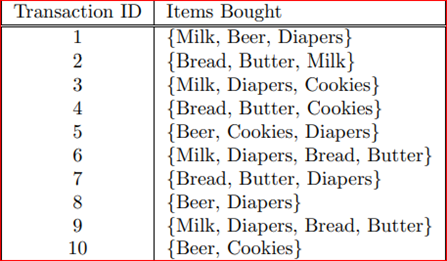
Consider the market basket transactions shown in the above table.

(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)?



**Dataset:** (sample.arff)

@relation dataset

@attribute milk{true,false}

@attribute beer{true,false}

@attribute diapers{true,false}

@attribute bread{true,false}

@attribute butter{true,false}

@attribute cookies{true,false}

@data

true true true false false false

true false false true true false

true false true false false true

false false false true true true

false true true false false true

true false true true true false

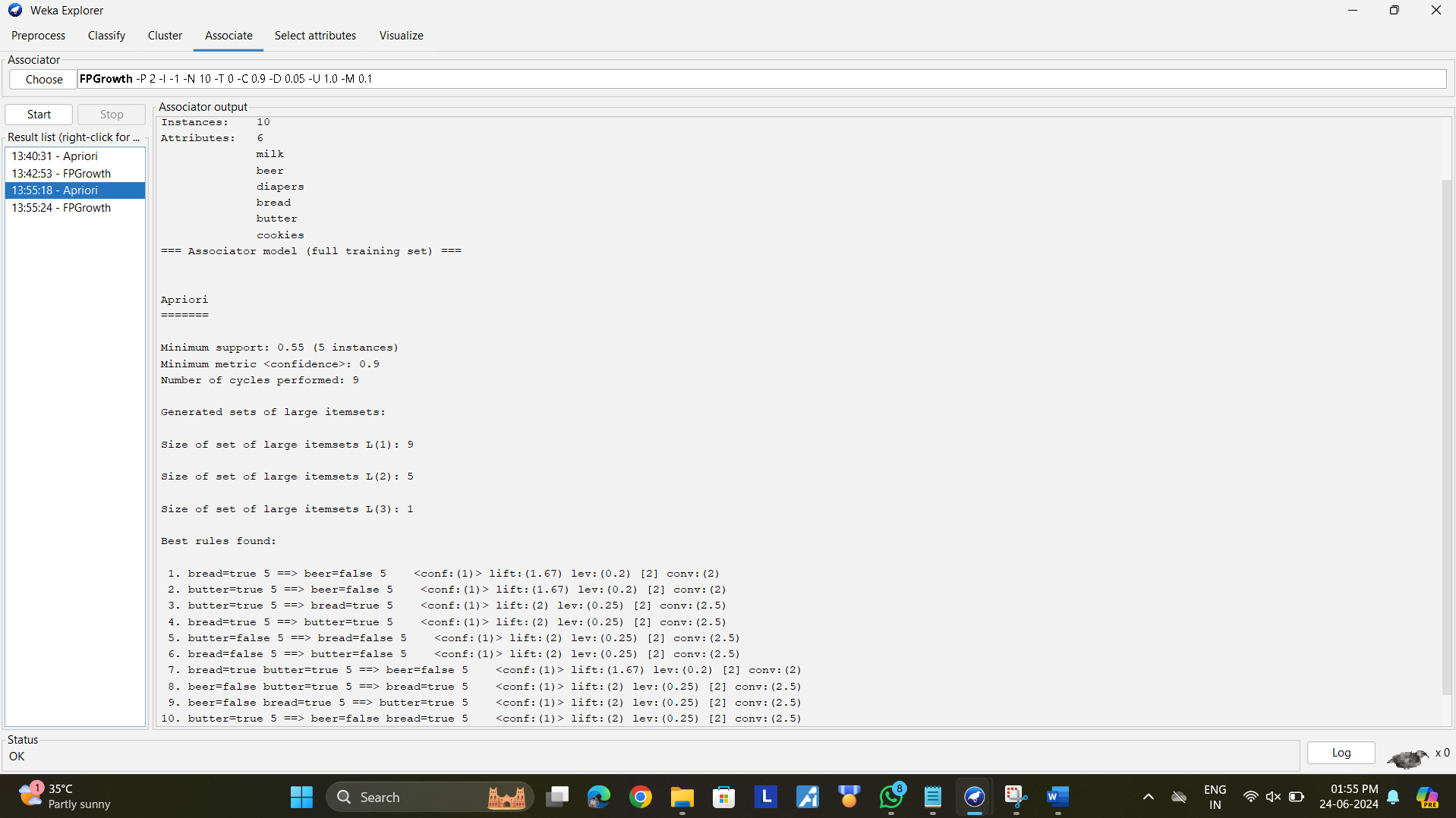
false false true true true false

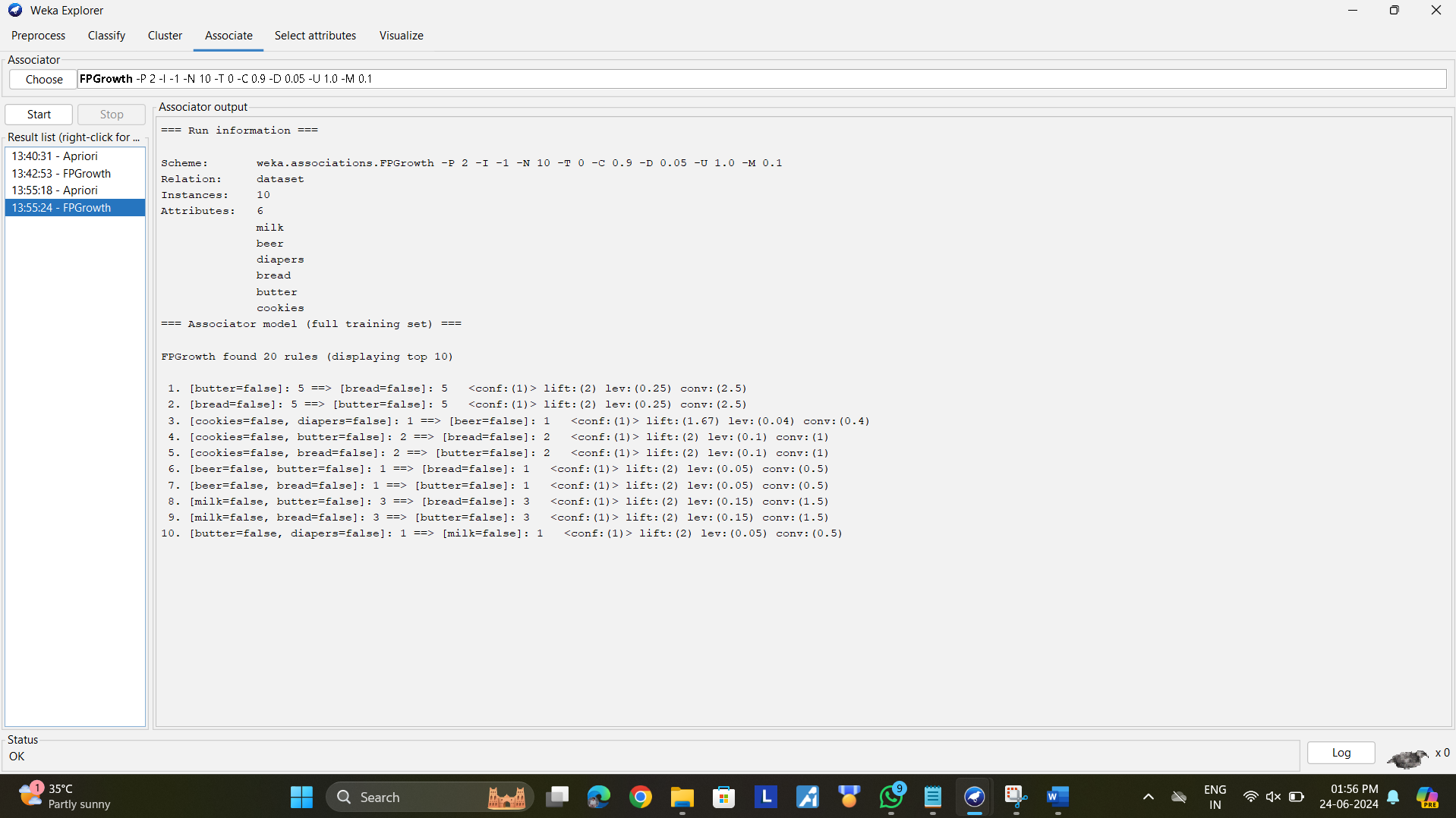
false true true false false false

true false true true true false

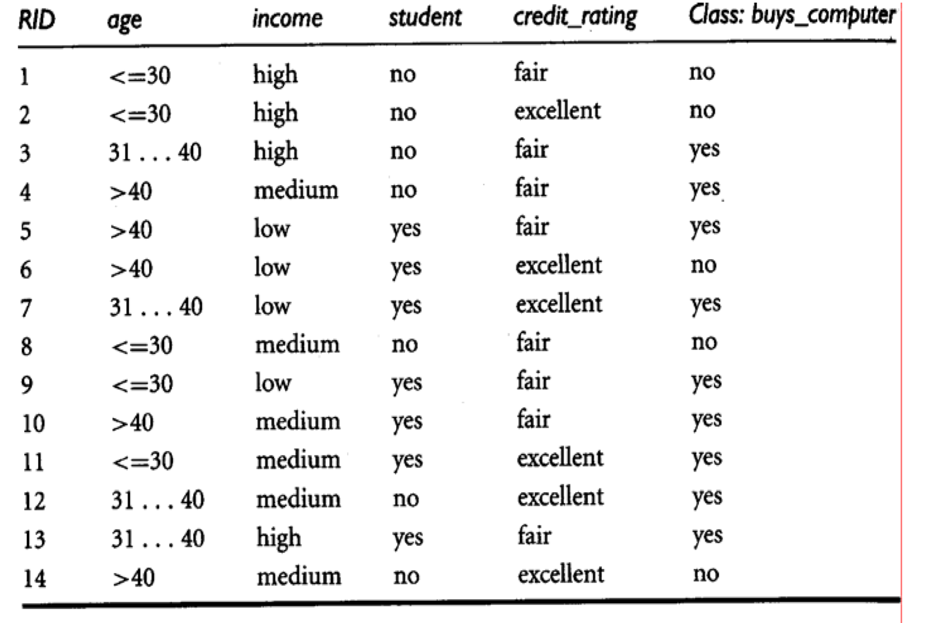
false true false false false true

**Output:**





3, Bayes classification and descion tree (using training and test data)



**Dataset:** (sample.arff)

@relation dataset

@attribute age {<=30,31-40,>40}

@attribute income {high,medium,low}

@attribute student {yes,no}

@attribute credit {fair,excellent}

@attribute class {yes,no}

@data

<=30,high,no,fair,no

<=30,high,no,excellent,no

31-40,high,no,fair,yes

>40,medium,no,fair,yes

>40,low,yes,fair,yes

>40,low,yes,excellent,no

31-40,low,yes,excellent,yes

<=30,medium,no,fair,no

<=30,low,yes,fair,yes

>40,medium,yes,fair,yes

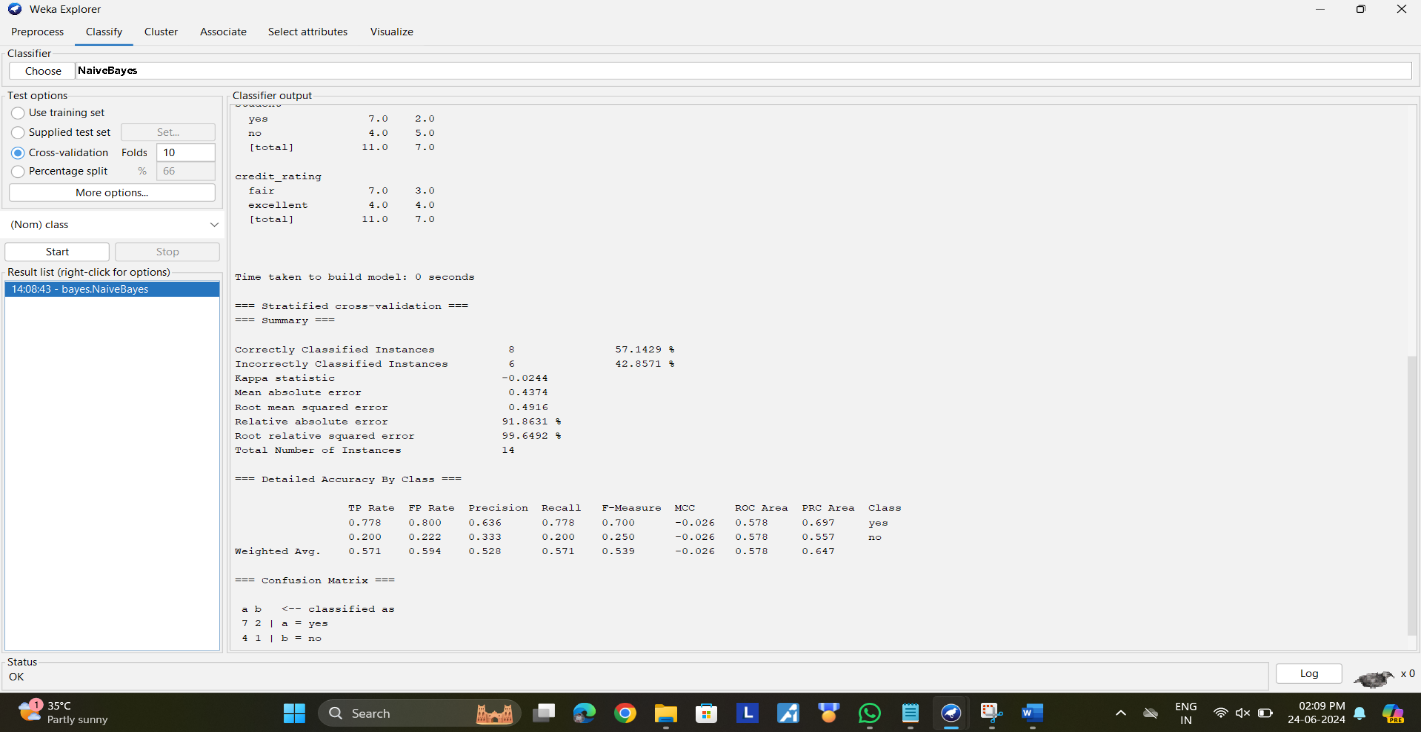
<=30,medium,yes,excellent,yes

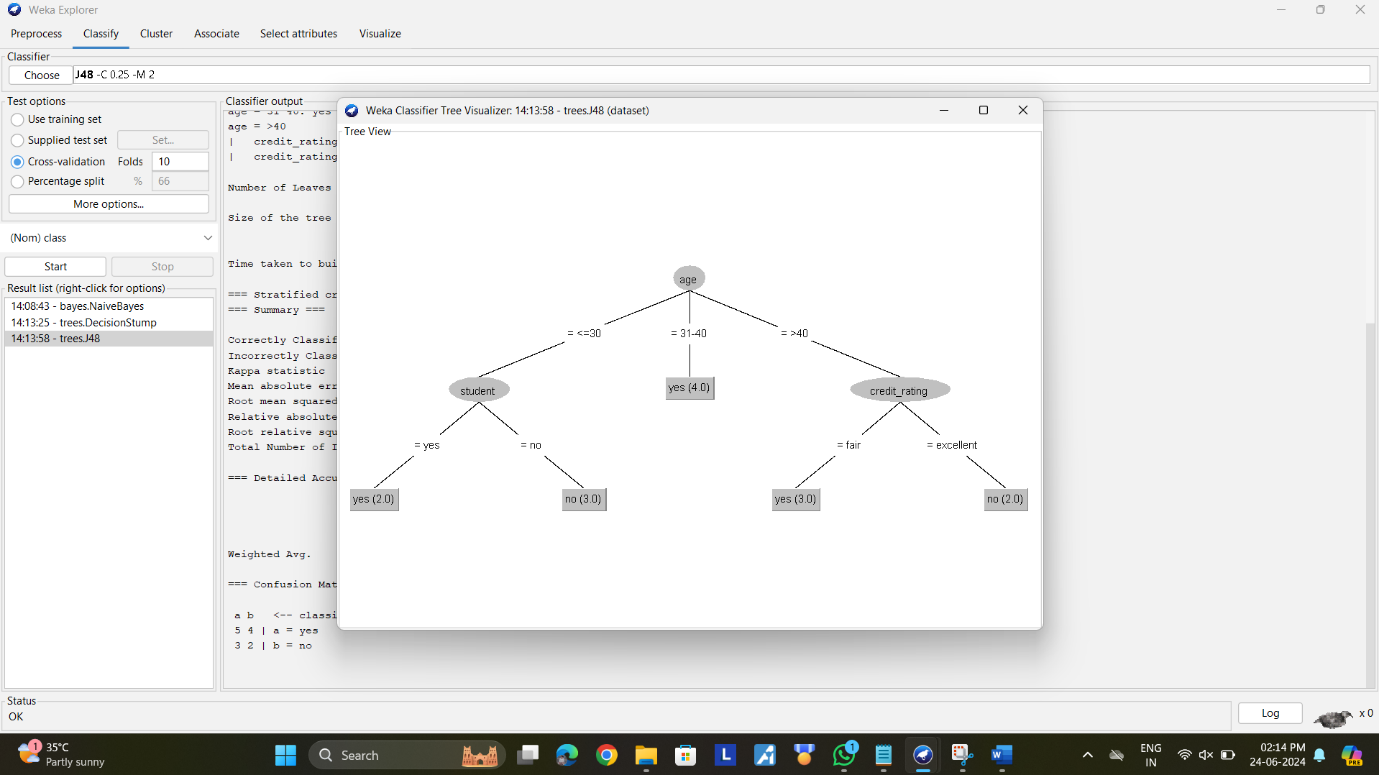
31-40,medium,no,excellent,yes

31-40,high,yes,fair,yes

>40,medium,no,excellent,no

**Output:**





4, 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

**Dataset: (sample.c)**

@relation dataset

@attribute M{yes,no}

@attribute O{yes,no}

@attribute N{yes,no}

@attribute K{yes,no}

@attribute E{yes,no}

@attribute Y{yes,no}

@attribute D{yes,no}

@attribute A{yes,no}

@attribute C{yes,no}

@attribute U{yes,no}

@attribute I{yes,no}

@data

yes yes yes yes yes yes no no no no no

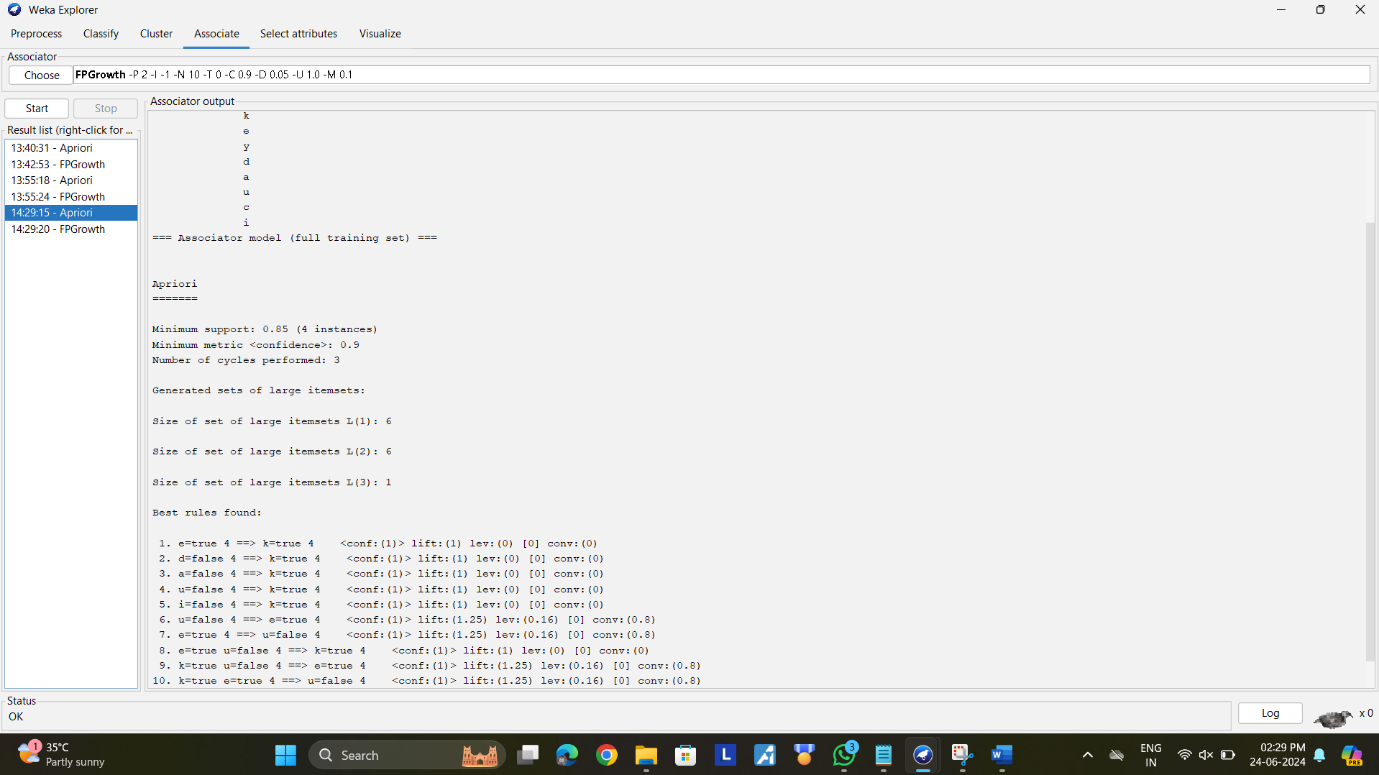
no yes yes yes yes yes yes no no no no

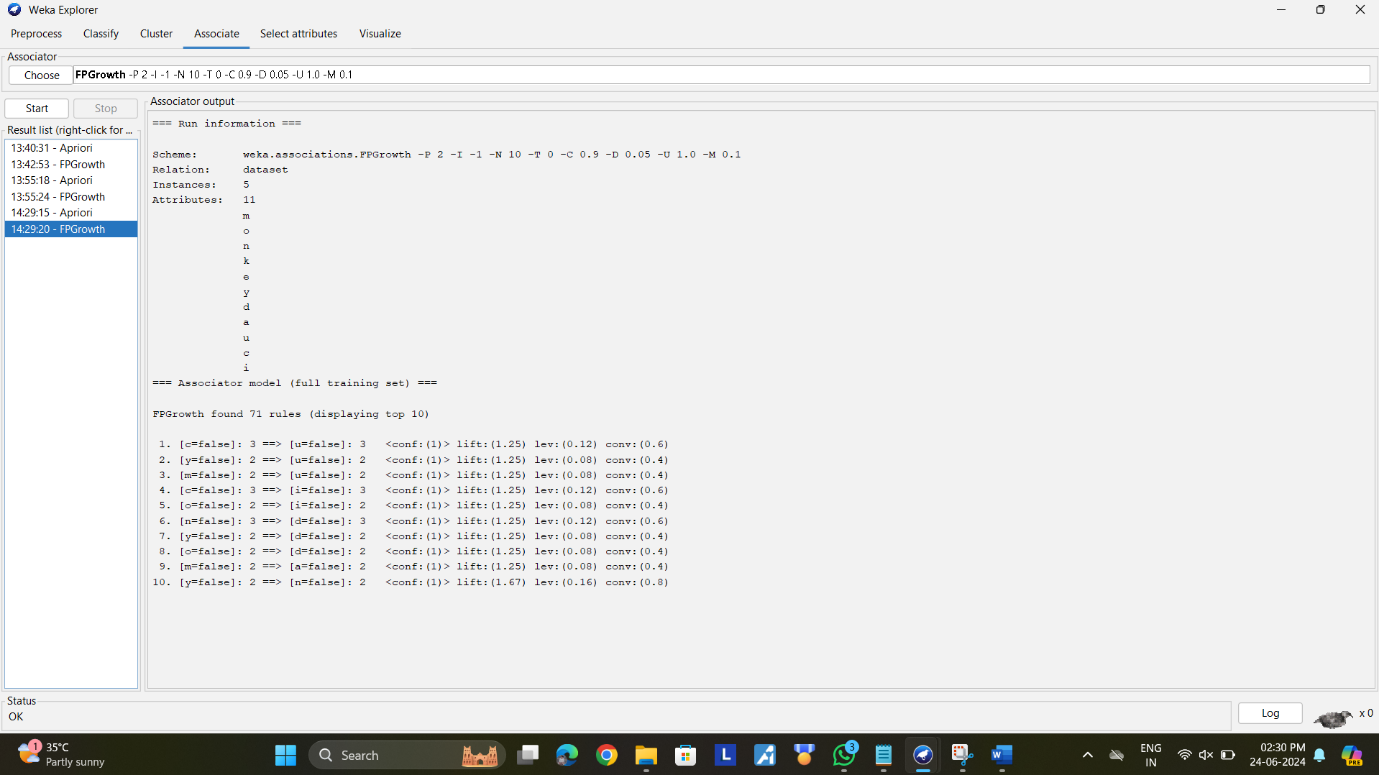
yes no no yes yes no no yes no no no

yes no no yes no yes no no yes yes no

no no no yes yes no no no yes no no

**Output:**





5, 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.

**Dataset: (Sample.arff)**

@relation dataset

@attribute hotdogs{yes,no}

@attribute buns{yes,no}

@attribute ketchup{yes,no}

@attribute coke{yes,no}

@attribute chips{yes,no}

@data

yes yes yes no no

yes yes no no no

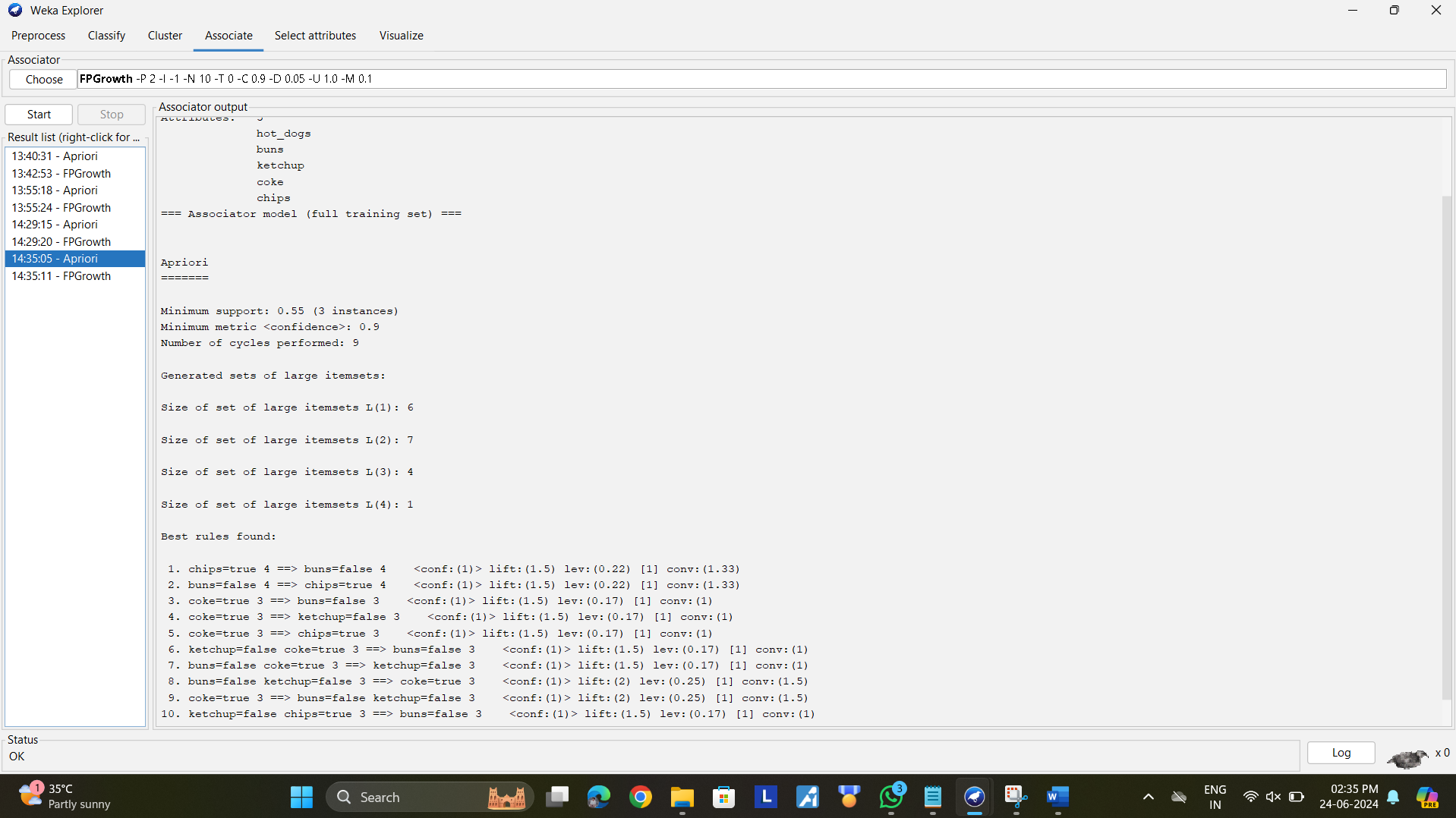
yes no no yes yes

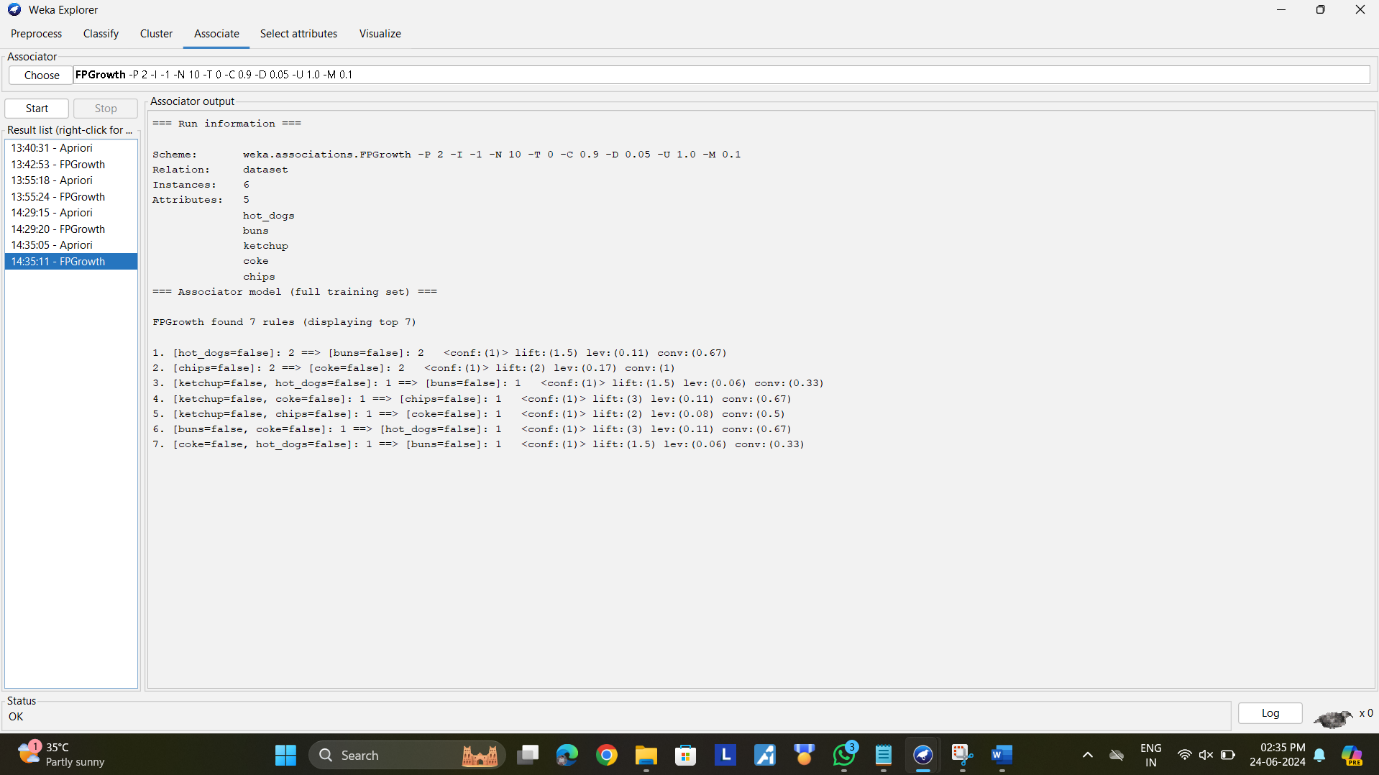
no no no yes yes

no no yes no yes

yes no no yes yes

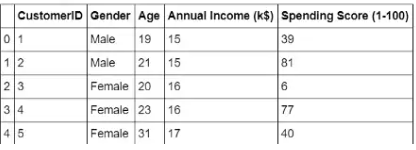
**Output:**





**Day 4**

1.Consider that you are owning a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. For the above scenario, the Problem Statement was You want to understand the customers who can easily converge [Target Customers] so that the data can be given to the marketing team and plan the strategy accordingly. For the above scenario prepare a dataset and perform **Clustering Analysis** to segment the customers in the Mall. There are clearly Five segments of Customers based on their Annual Income and Spending Score namely *Usual Customers, Priority Customers, Senior Citizen Target Customers, and Young Target Customers.*Sample data



**Dataset: (sample.arff)**

@relation dataset

@attribute gender{male,female}

@attribute age{19,21,20,23,31}

@attribute annual\_income{15,16,17}

@attribute Score{39,81,6,77,40}

@data

male 19 15 39

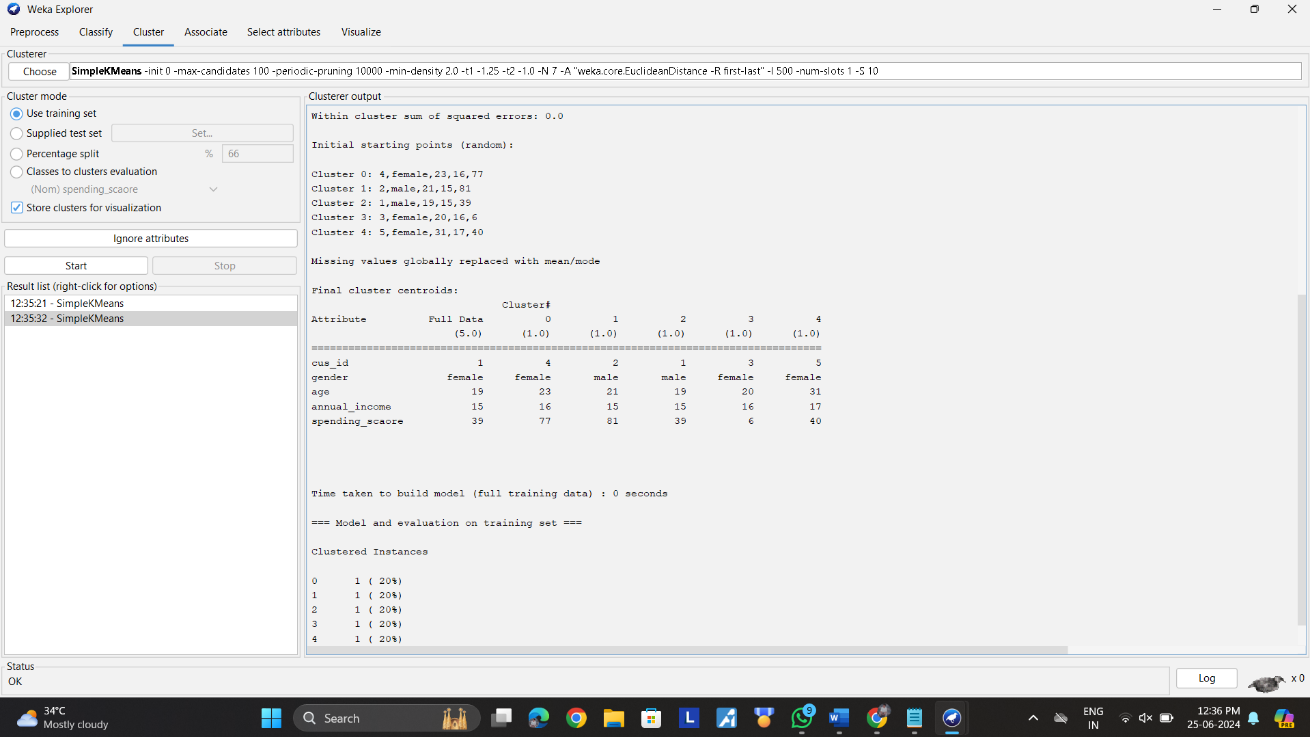
male 21 15 81

female 20 16 6

female 23 16 77

female 31 17 40

**Output:**



2.Create the following dataset using CSV file format. To perform cluster analysis using K- Means in WEKA. To change the cluster size and plot the graph and illustrate the visualization of cluster.

| 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 |

**Dataset: (sample.arff)**

@relation dataset

@attribute gender{male,female}

@attribute age{28,25,26,29,30}

@attribute salary{150000,160000,170000,200000}

@attribute Credit{39,27,42,40,64,72}

@data

male 28 150000 39

male 25 150000 27

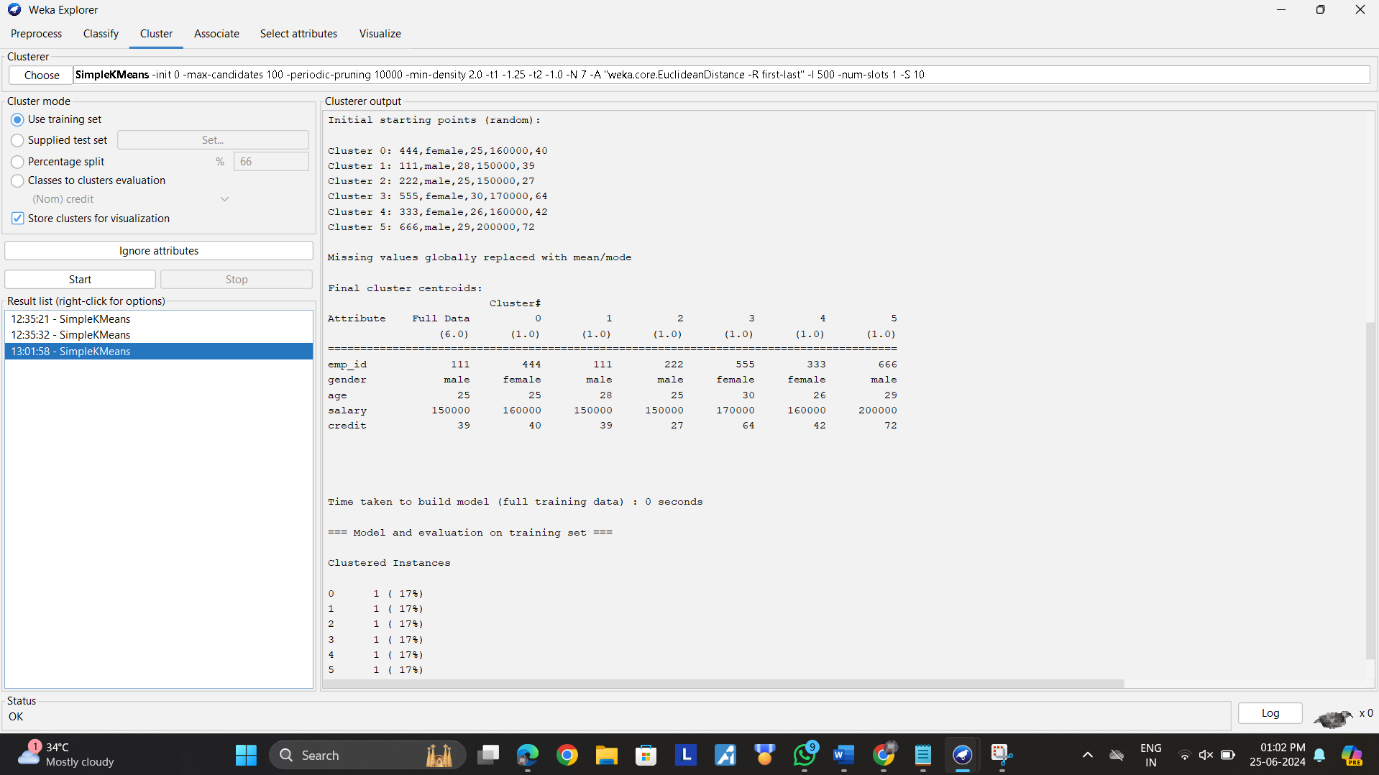
female 26 160000 42

female 25 160000 40

female 30 170000 64

male 29 200000 72

**Output:**



**4.T**he 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?

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Person** | Gopu | Babu | Baby | Gopal | Krishna | Jai | Dev | Malini | Hema | Anu |
| **Vegetarian** | yes | yes | yes | no | yes | no | no | yes | yes | yes |

**Dataset: (Sample.arff)**

@relation dataset

@attribute person {gopu,babu,baby,gopal,krishna,jai,dev,malini,hema,anu}

@attribute vegetarian {yes,no}

@data

gopu,yes

babu,yes

baby,yes

gopal,no

krishna,yes

jai,no

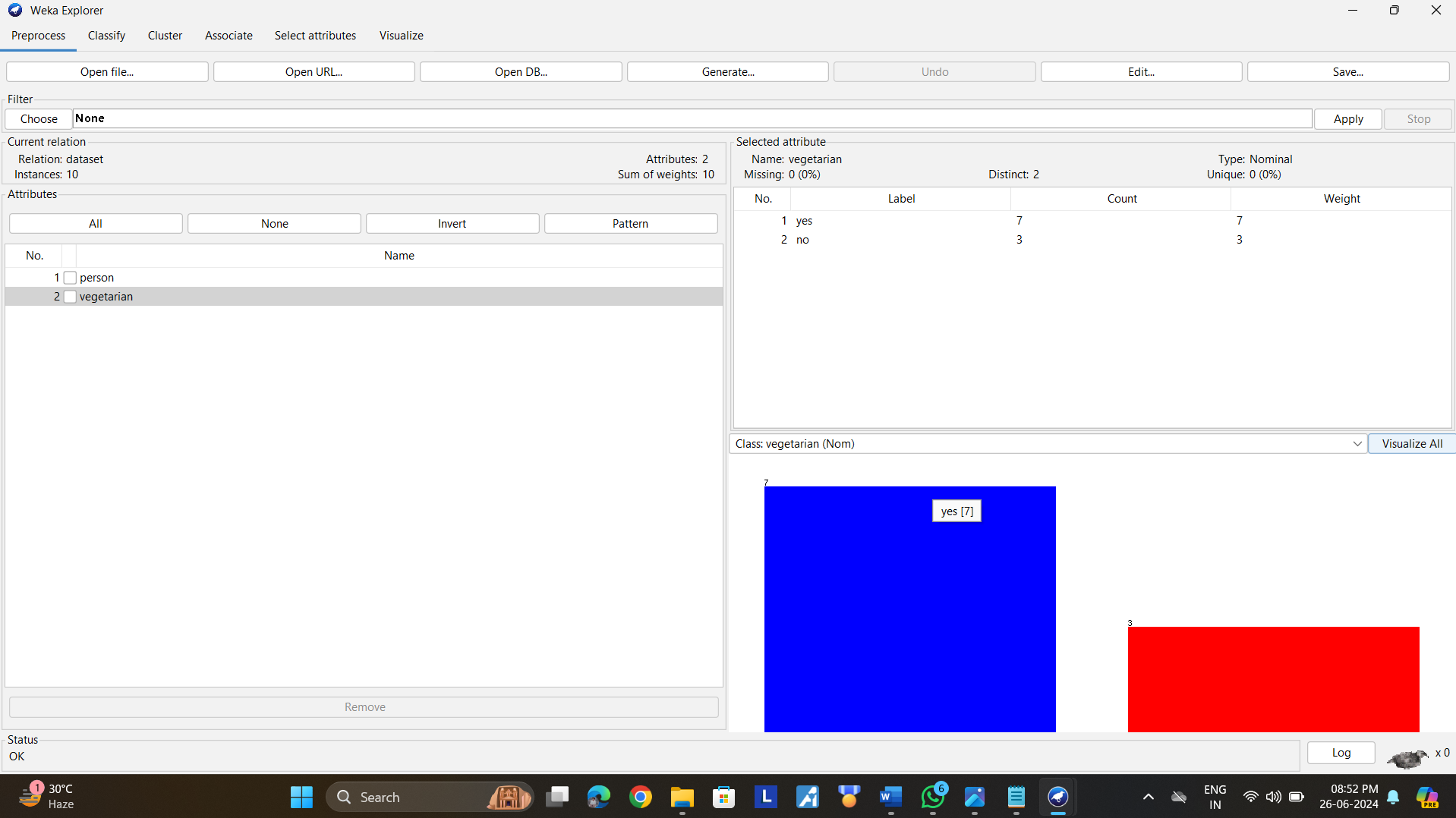
dev,no

malini,yes

hema,yes

anu,yes

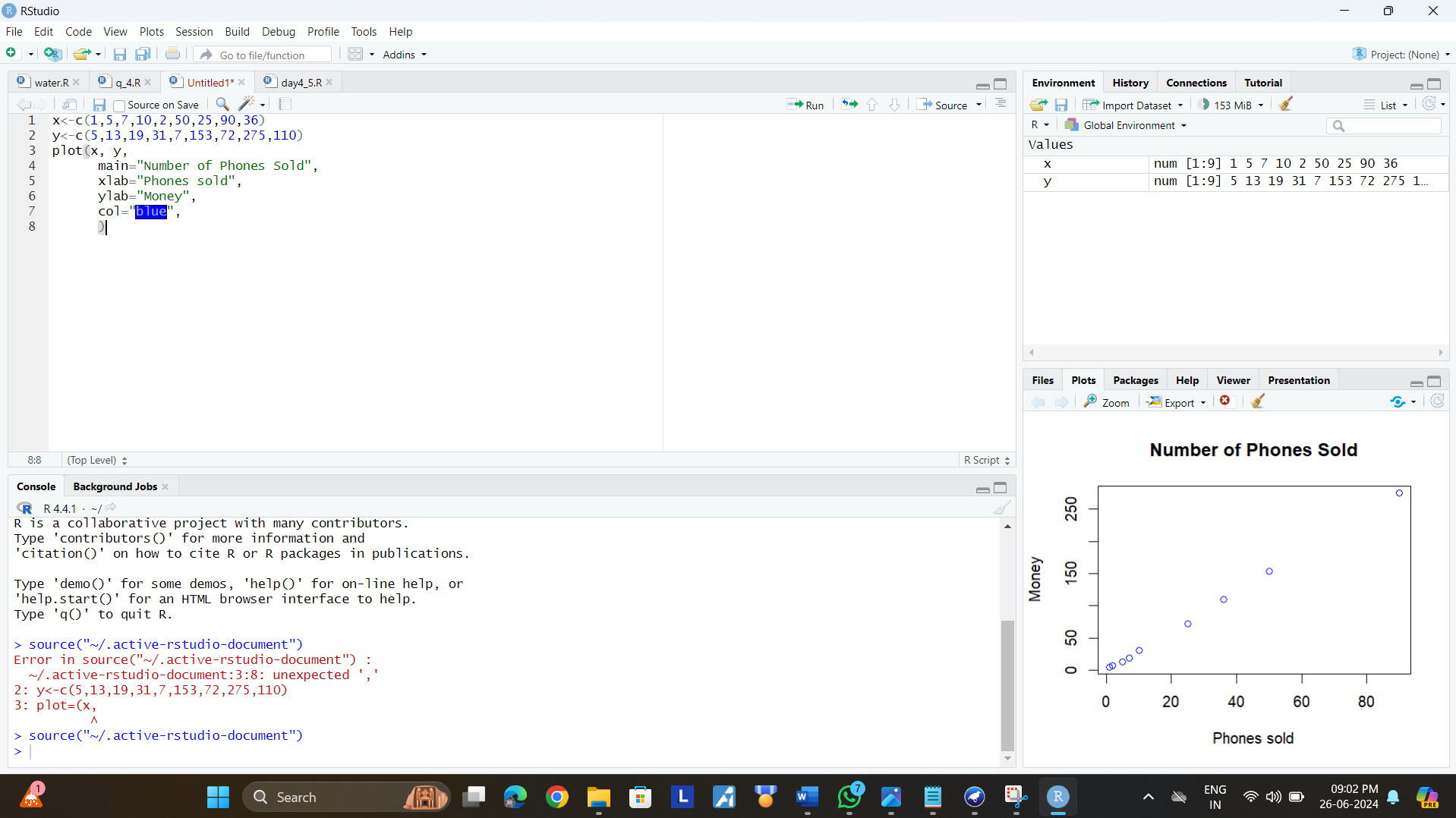
**Output:**



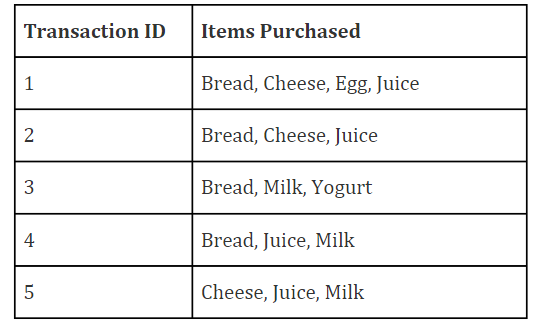
5.The following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones sold.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 4 | 1 | 5 | 7 | 10 | 2 | 50 | 25 | 90 | 36 |
| y | 12 | 5 | 13 | 19 | 31 | 7 | 153 | 72 | 275 | 110 |

**Output:**



6.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%.



**Dataset: (sample.arff)**

@relation dataset

@attribute bread{true,false}

@attribute cheese{true,false}

@attribute egg{ture,false}

@attribute juice{ture,false}

@attribute milk{ture,false}

@attribute yogurt{ture,false}

@data

true true true true false false

true true false true false fasle

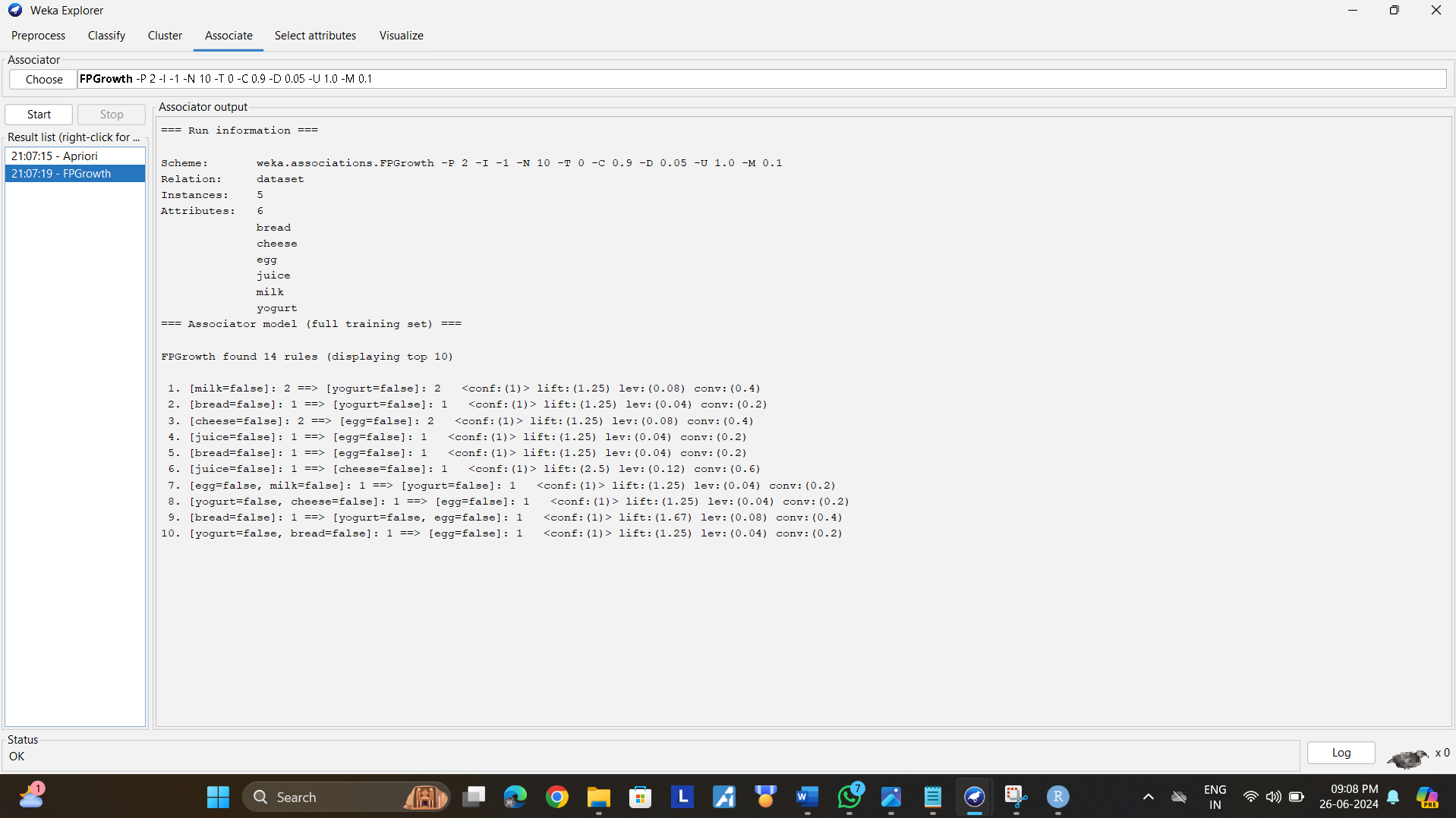
true false false false true true

true false false true true false

false true false true true false

**Output:**

****

****

10.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.

NOTE: Assume Min\_sup=2 and confidence= 50%

|  |  |
| --- | --- |
| T.ID | ITEMS |
| T1 | SONY, BPL, LG |
| T2 | BPL, SAMSUNG |
| T3 | BPL, ONIDA |
| T4 | SONY, BPL, SAMSUNG |
| T5 | SONY, ONIDA |
| T6 | BPL, ONIDA |
| T7 | SONY, ONIDA |
| T8 | SONY, BPL, ONIDA, LG |
| T9 | SONY, BPL, ONIDA |

**Dataset: (sample.arff)**

@relation dataset

@attribute sony{true,false}

@attribute bpl{true,false}

@attribute lg{true,false}

@attribute samsung{true,false}

@attribute onida{true,false}

@data

true true true false false

false true false true false

false true false false true

true true false true false

true false false false true

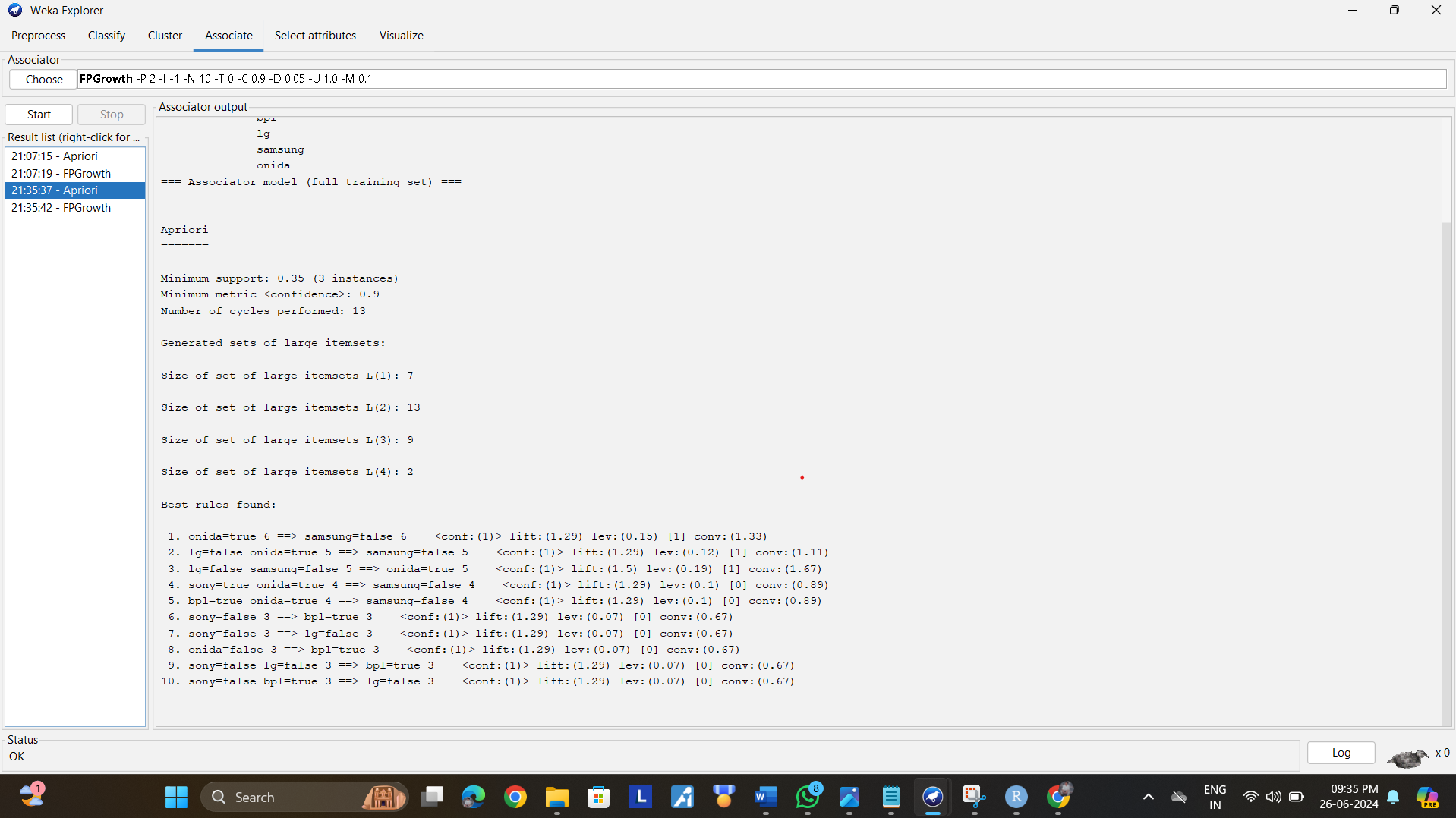
false true false false true

true false false false true

true true true false true

true true false false true

**Output:**

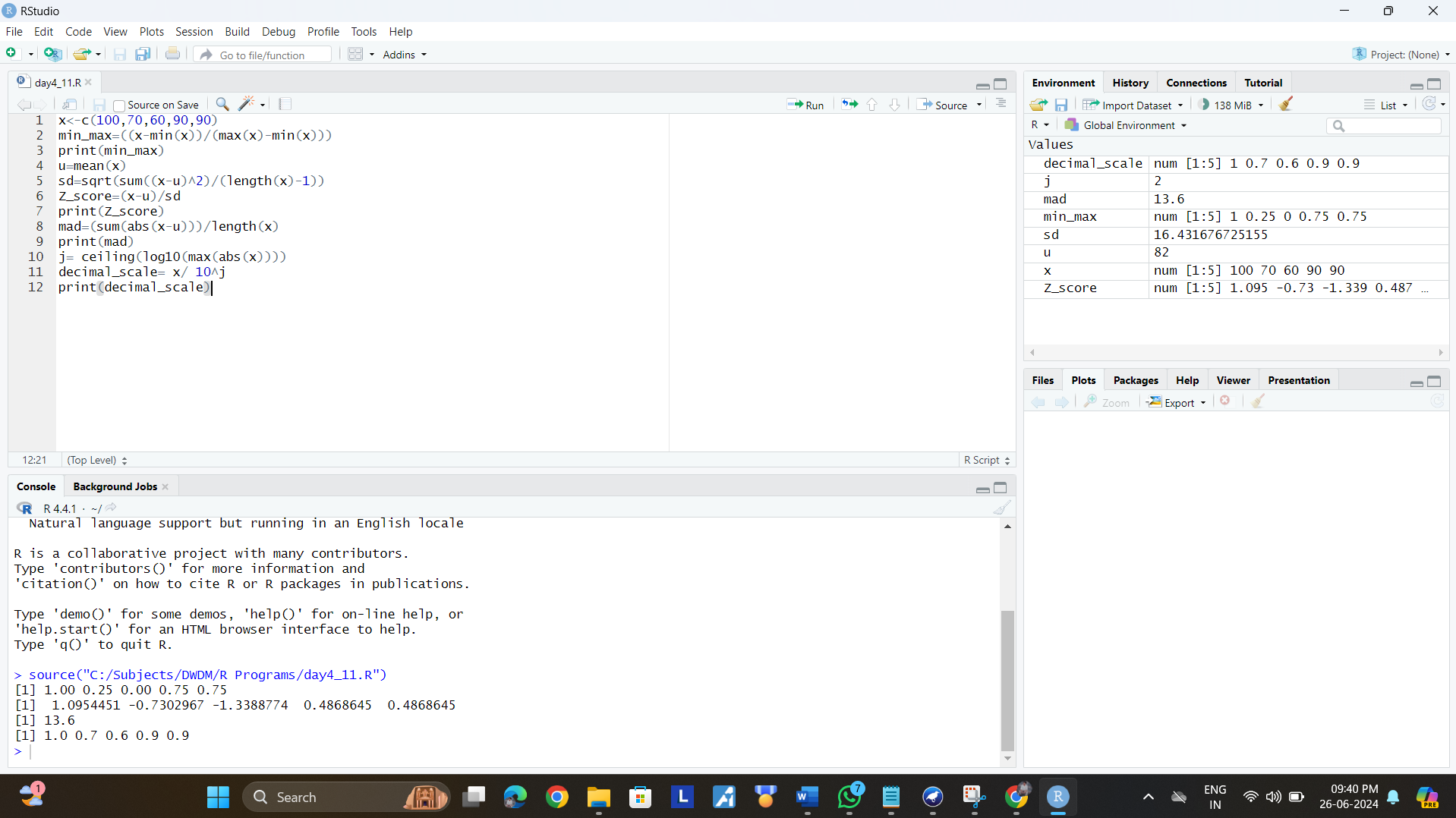
****

****

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

1. min-max normalization by setting min = 0 and max = 1
2. z-score normalization
3. z-score normalization using the mean absolute deviation instead of standard deviation
4. normalization by decimal scaling

**Output:**

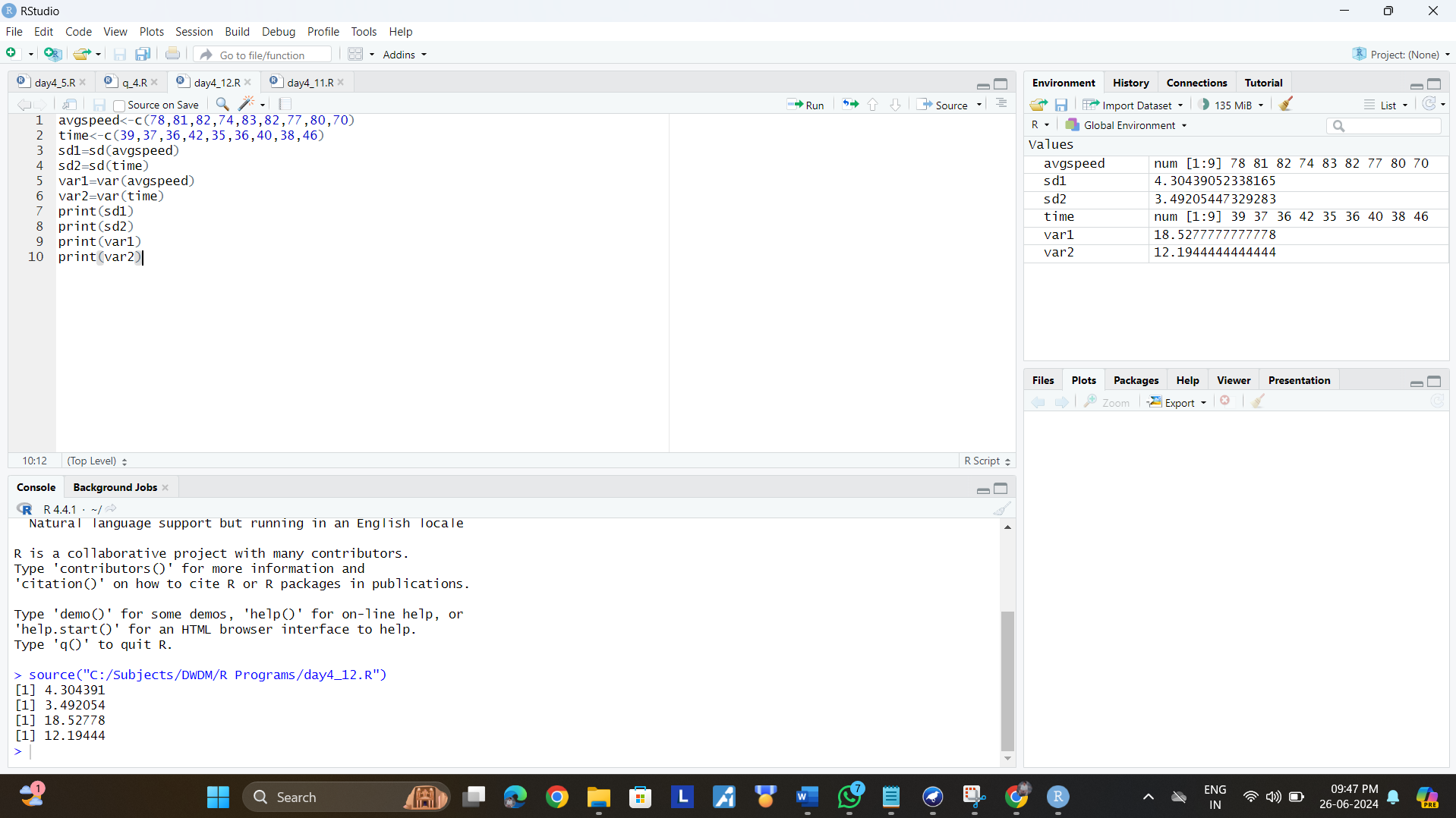


12,Suppose some car is tested for the AvgSpeed and TotalTime data for 9 randomly selected car with the following result

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

1. Calculate the standard deviation of AvgSpeed and TotalTime.
2. Calculate the Variance of AvgSpeed and TotalTime for the above dataset.

**Output:**



13.Consider this table

**TID items bought**

**T100 {M, O, N, K, E, Y}**

**T200 {D, O, N, K, E, Y }**

**T300 {M, A, K, E}**

**T400 {M, U, C, K, Y}**

**T500 {C, O, O, K, I ,E**}

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

(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.):

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

**Dataset: (sample.c)**

@relation dataset

@attribute M{yes,no}

@attribute O{yes,no}

@attribute N{yes,no}

@attribute K{yes,no}

@attribute E{yes,no}

@attribute Y{yes,no}

@attribute D{yes,no}

@attribute A{yes,no}

@attribute C{yes,no}

@attribute U{yes,no}

@attribute I{yes,no}

@data

yes yes yes yes yes yes no no no no no

no yes yes yes yes yes yes no no no no

yes no no yes yes no no yes no no no

yes no no yes no yes no no yes yes no

no no no yes yes no no no yes no no

**Output:**

