**ASSIGNMENT - 1**

**LARGE SCALE DATA PROCESSING**

**PRIYANSHU SHARMA**

**15BCE1282**

**“APRIORI ALGORITHM USING R”**

**CODE -**

#APRIORI METHOD BY PRIYANSHU SHARMA

#install.package(‘arules’)

#install.package(‘arulesViz’)

#install.package(‘dataset’)

library(arules)

dataset = read.csv('groceries.csv', header = FALSE)

dataset = read.transactions('groceries.csv', sep = ',', rm.duplicates = TRUE)

summary(dataset)

itemFrequencyPlot(dataset, topN = 20)

itemFrequencyPlot(dataset, topN = 10)

#Training Apriori on the dataset

rules = apriori(data = dataset, parameter = list(support = 0.001, confidence = 0.8))

#Visualization

options(digits=2)

inspect(rules[1:5])

rules<-sort(rules, by="confidence", decreasing=TRUE)

rules<-apriori(data=dataset, parameter=list(supp=0.001,conf = 0.08),

appearance = list(default="lhs",rhs="whole milk"),

control = list(verbose=F))

rules<-sort(rules, decreasing=TRUE,by="confidence")

inspect(rules[1:5])

rules<-apriori(data=dataset, parameter=list(supp=0.001,conf = 0.15,minlen=2),

appearance = list(default="rhs",lhs="whole milk"),

control = list(verbose=F))

rules<-sort(rules, decreasing=TRUE,by="confidence")

inspect(rules[1:5])

library(arulesViz)

plot(rules)

plot(rules[1:5],method = "graph",control = list(type = "items"))

arulesViz::plotly\_arules(rules)

**EXPLAINATION**

We are going to make the use of different libraries such as Arules, ArulesViz, etc. So in order to install these library we have to use the following commands : -

*#install.package(‘arules’)*

*#install.package(‘arulesViz’)*

*#install.package(‘dataset’)*

For importing the arules library use : -

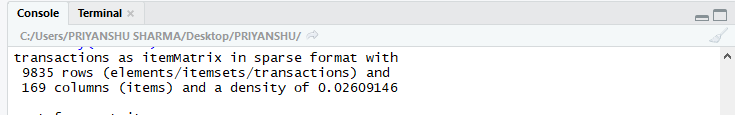
*library(arules)*

Now we have to read the dataset and we have to exclude the header of each column so use : -

*dataset = read.csv('groceries.csv', header = FALSE)*

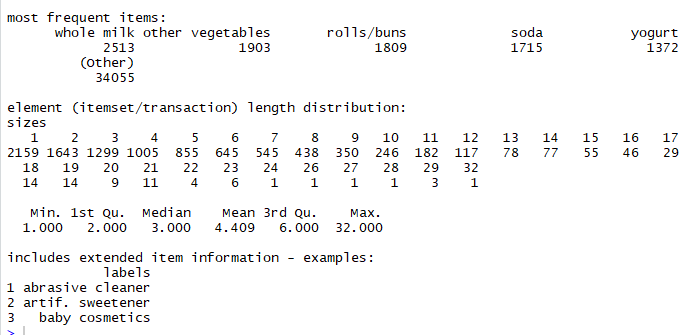
Since ‘arules’ library make the use of spare matrix in order to train the model and this dataset is comma separated and may be our dataset has many duplicate. So in order to take care of it we have added rm.duplicate :-

*dataset = read.transactions('groceries.csv', sep = ',', rm.duplicates = TRUE)*



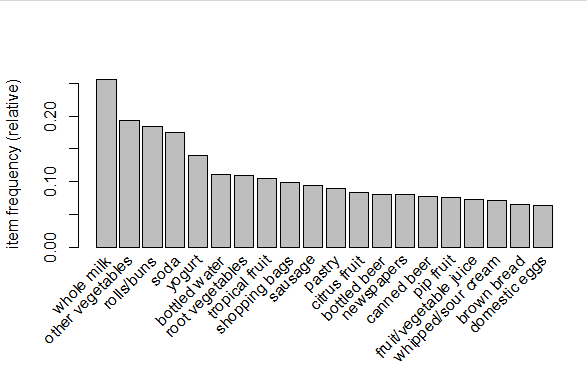
To get the summary of our complete dataset : -

*summary(dataset)*

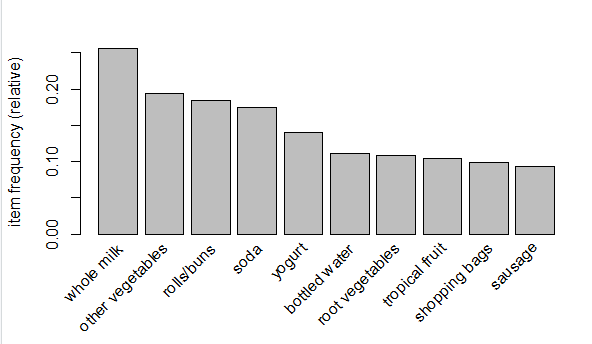


To visualize the top 20 and top 10 product :-

*itemFrequencyPlot(dataset, topN = 20)*



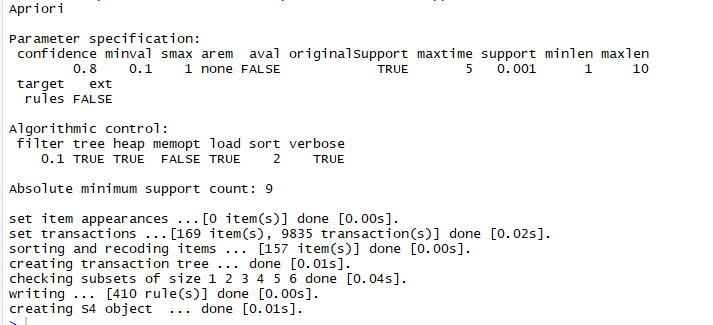
*itemFrequencyPlot(dataset, topN = 10)*



Now we are going to set the minimum support and confidence for training our model and after training we get 410 rules based on our assumption of minimum support and confidence value.

*#Training Apriori on the dataset*

*rules = apriori(data = dataset, parameter = list(support = 0.001, confidence = 0.8))*

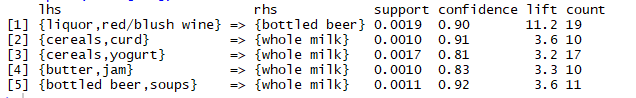


*#Visualization*

Looking for the first 5 rules which we get from our trained model based no relevance : -

*options(digits=2)*

*inspect(rules[1:5])*



Sorting the rules in decreasing order based on confidence values.

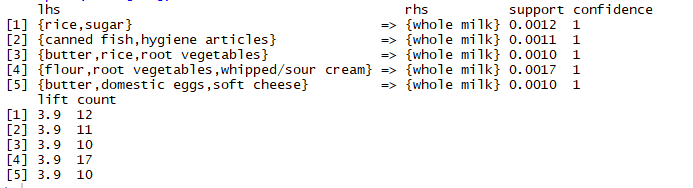
*rules<-sort(rules, by="confidence", decreasing=TRUE)*

Here we have considered those rules only which are having “*whole milk*” in rhs means all those rules in which consumer buy something before buying “*whole milk*” and then we are sorting those rules in decreasing order based on confidence values.

*rules<-apriori(data=dataset, parameter=list(supp=0.001,conf = 0.08), appearance = list(default="lhs",rhs="whole milk"), control = list(verbose=F))*

*rules<-sort(rules, decreasing=TRUE,by="confidence")*

*inspect(rules[1:5])*



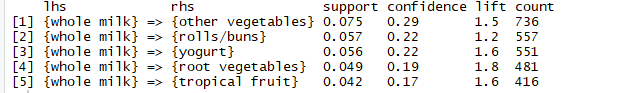
Here we are going to considered those rules only in which consumer are buying something after buying “*whole milk”* and then sorting them in decreasing order.

*rules<-apriori(data=dataset, parameter=list(supp=0.001,conf = 0.15,minlen=2), appearance = list(default="rhs",lhs="whole milk"),*

*control = list(verbose=F))*

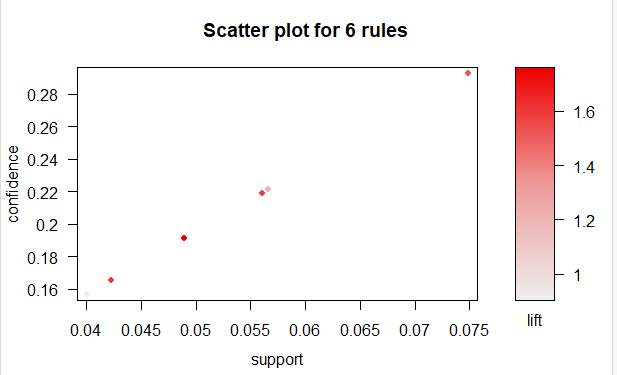
*rules<-sort(rules, decreasing=TRUE,by="confidence")*

*inspect(rules[1:5])*

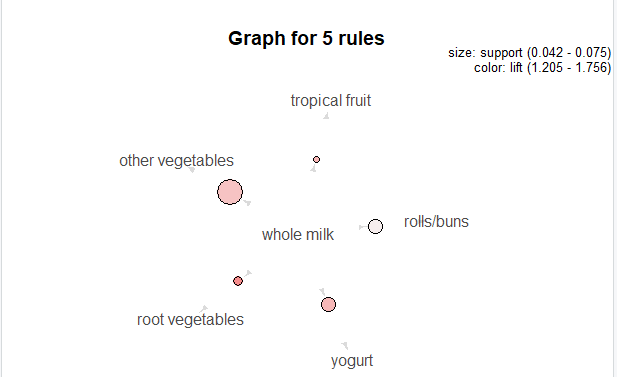


*library(arulesViz)*

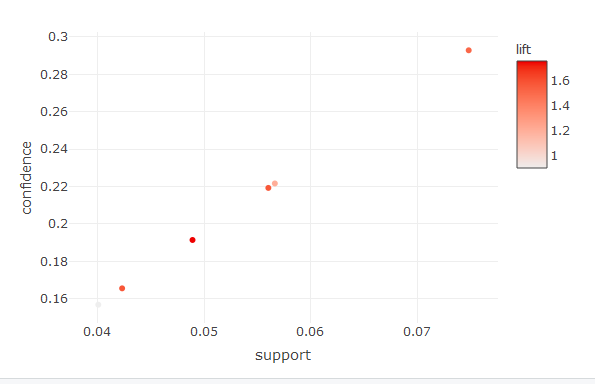
*plot(rules)*



*plot(rules[1:5],method = "graph",control = list(type = "items"))*



*arulesViz::plotly\_arules(rules)*



**DATASET**

**Groceries.csv**

