Meghana_Nadig_Practicum3

Meghana_Nadig

April 12, 2018

Problem 1

```
bank_data <- read.csv("C:/Users/Meghana Nadig/Downloads/bank/bank-full.csv", sep = ";")
test_data <- read.csv("C:/Users/Meghana Nadig/Downloads/bank/bank.csv", sep = ";")
str(bank_data)
```

```
## 'data.frame':    45211 obs. of  17 variables:
##  $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
##  $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
##  $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
##  $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
##  $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Exploring Data

```
prop.table(table(bank_data$y))
```

```
##
##        no       yes
## 0.8830152 0.1169848
```

```
nrow(bank_data)
```

```
## [1] 45211
```

```
prop.table(table(test_data$y))
```

```
##
##      no     yes
## 0.88476 0.11524
```

```
nrow(test_data)
```

```
## [1] 4521
```

```
summary(bank_data)
```

```
##       age                 job           marital          education
##  Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851
##  1st Qu.:33.00   management :9458   married :27214   secondary:23202
##  Median :39.00   technician :7597   single  :12790   tertiary :13301
##  Mean   :40.94   admin.     :5171                    unknown  : 1857
##  3rd Qu.:48.00   services   :4154
##  Max.   :95.00   retired    :2264
##                  (Other)    :6835
##  default        balance       housing      loan            contact
##  no :44396   Min.   : -8019   no :20081   no :37967   cellular :29285
##  yes:  815   1st Qu.:    72   yes:25130   yes: 7244   telephone: 2906
##              Median :   448                           unknown  :13020
##              Mean   :  1362
##              3rd Qu.:  1428
##              Max.   :102127
##
##       day            month          duration         campaign
##  Min.   : 1.00   may    :13766   Min.   :   0.0   Min.   : 1.000
##  1st Qu.: 8.00   jul    : 6895   1st Qu.: 103.0   1st Qu.: 1.000
##  Median :16.00   aug    : 6247   Median : 180.0   Median : 2.000
##  Mean   :15.81   jun    : 5341   Mean   : 258.2   Mean   : 2.764
##  3rd Qu.:21.00   nov    : 3970   3rd Qu.: 319.0   3rd Qu.: 3.000
##  Max.   :31.00   apr    : 2932   Max.   :4918.0   Max.   :63.000
##                  (Other): 6060
##      pdays          previous           poutcome       y
##  Min.   : -1.0   Min.   :  0.0000   failure: 4901   no :39922
##  1st Qu.: -1.0   1st Qu.:  0.0000   other  : 1840   yes: 5289
##  Median : -1.0   Median :  0.0000   success: 1511
##  Mean   : 40.2   Mean   :  0.5803   unknown:36959
##  3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :871.0   Max.   :275.0000
##
```

```
str(bank_data)
```

```
## 'data.frame':    45211 obs. of  17 variables:
##  $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
##  $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
##  $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
##  $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
##  $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
##  $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Support Vector Machine

```
library(kernlab)
set.seed(132)

# Building the SVM Model
svm_model <- ksvm(y ~ ., data = bank_data, kernal ="vanilladot")

svm_model
```

```
## Support Vector Machine object of class "ksvm"
##
## SV type: C-svc  (classification)
##  parameter : cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.0772346432145536
##
## Number of Support Vectors : 10032
##
## Objective Function Value : -8743.326
## Training error : 0.086284
```

Evaluating model performance

```
# Making predictions on testing dataset
data_predictions <- predict(svm_model, test_data)

head(data_predictions)
```

```
## [1] no no no no no no
## Levels: no yes
```

```
# Comparing predicted
table(data_predictions, test_data$y)
```

```
##
## data_predictions   no  yes
##              no  3936  317
##              yes   64  204
```

```
# Calculating the overall accuracy
agreement <- data_predictions == test_data$y

table(agreement)
```

```
## agreement
## FALSE  TRUE
##   381  4140
```

```r
# Accuracy in terms of percentage
prop.table(table(agreement))
```

```
## agreement
##      FALSE       TRUE
## 0.08427339 0.91572661
```

Absolute Accuracy

```r
library("caret")
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:kernlab':
##
##     alpha
```

```r
# SVM

confusionMatrix(data_predictions,test_data$y)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  3936  317
##        yes   64  204
##
##                Accuracy : 0.9157
##                  95% CI : (0.9072, 0.9237)
##     No Information Rate : 0.8848
##     P-Value [Acc > NIR] : 6.089e-12
##
##                   Kappa : 0.4761
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9840
##             Specificity : 0.3916
##          Pos Pred Value : 0.9255
##          Neg Pred Value : 0.7612
##              Prevalence : 0.8848
##          Detection Rate : 0.8706
##    Detection Prevalence : 0.9407
##       Balanced Accuracy : 0.6878
##
##        'Positive' Class : no
##
```

Area Under Curve (AUC)

```
# SVM_AUC
#install.packages("ROCR")

library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.4.4
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.4.4
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
svm_a <- prediction(as.numeric(data_predictions),as.numeric(test_data$y))

eval <- performance(svm_a, "acc")

eval
```

```
## An object of class "performance"
## Slot "x.name":
## [1] "Cutoff"
##
## Slot "y.name":
## [1] "Accuracy"
##
## Slot "alpha.name":
## [1] "none"
##
## Slot "x.values":
## [[1]]
## [1] Inf   2   1
##
##
## Slot "y.values":
## [[1]]
## [1] 0.8847600 0.9157266 0.1152400
##
##
## Slot "alpha.values":
## list()
```

Neural Network

Reading the data

```
bank_data <- read.csv("C:/Users/Meghana Nadig/Downloads/bank/bank-full.csv", sep = ";")
bank_test <- read.csv("C:/Users/Meghana Nadig/Downloads/bank/bank.csv", sep = ";")
str(bank_data)
```

```
## 'data.frame':    45211 obs. of  17 variables:
##  $ age       : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job       : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
##  $ marital   : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
##  $ default   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ balance   : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing   : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
##  $ loan      : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
##  $ contact   : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ day       : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month     : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays     : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ y         : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Converting factors to dummy variables and creating a new dataframe

```
#str(bank_data)
#bank_data$job_DV <- as.numeric(factor(bank_data$job))
#bank_data$marital_DV <- as.numeric(factor(bank_data$marital))
#bank_data$education_DV <- as.numeric(factor(bank_data$education))
#bank_data$default_DV <- as.numeric(factor(bank_data$default))
#bank_data$housing_DV <- as.numeric(factor(bank_data$housing))
#bank_data$loan_DV <- as.numeric(factor(bank_data$loan))
#bank_data$contact_DV <- as.numeric(factor(bank_data$contact))
#bank_data$month_DV <- as.numeric(factor(bank_data$month))
#bank_data$poutcome_DV <- as.numeric(factor(bank_data$poutcome))
#bank_data$y_DV <- as.numeric(factor(bank_data$y))

#df <- data.frame(bank_data[,c(1,6,10,12:15,18:27)])


#str(test_data)
#test_data$job_DV <- as.numeric(factor(test_data$job))
#test_data$marital_DV <- as.numeric(factor(test_data$marital))
#test_data$education_DV <- as.numeric(factor(test_data$education))
#test_data$default_DV <- as.numeric(factor(test_data$default))
```

```
#test_data$housing_DV <- as.numeric(factor(test_data$housing))
#test_data$loan_DV <- as.numeric(factor(test_data$loan))
#test_data$contact_DV <- as.numeric(factor(test_data$contact))
#test_data$month_DV <- as.numeric(factor(test_data$month))
#test_data$poutcome_DV <- as.numeric(factor(test_data$poutcome))
#test_data$y_DV <- as.numeric(factor(test_data$y))

#df_test <- data.frame(test_data[,c(1,6,10,12:15,18:27)])
library("caret")

dummybank <-dummyVars("~.",data=bank_data, fullRank = T) #creating dummy variables from factors in bank
bankdummy <- data.frame(predict(dummybank, newdata = bank_data))

dummytest <-dummyVars("~.",data=bank_test, fullRank = T) #creating the dummy variables in test set.
testdummy <- data.frame(predict(dummytest, newdata = bank_test))
```

Step 2: Exploring and preparing the data

```
# Normalizing the data

normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

df_norm <- as.data.frame(lapply(bankdummy,normalize))
df_test_norm <- as.data.frame(lapply(testdummy,normalize))

summary(df_norm)
```

```
##       age          job.blue.collar  job.entrepreneur  job.housemaid
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.1948   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.2727   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.2979   Mean   :0.2153   Mean   :0.03289   Mean   :0.02743
##  3rd Qu.:0.3896   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##  job.management    job.retired      job.self.employed  job.services
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.2092   Mean   :0.05008   Mean   :0.03493   Mean   :0.09188
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##   job.student      job.technician  job.unemployed     job.unknown
##  Min.   :0.00000   Min.   :0.000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.000   Median :0.00000   Median :0.00000
##  Mean   :0.02075   Mean   :0.168   Mean   :0.02882   Mean   :0.00637
##  3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.000   Max.   :1.00000   Max.   :1.00000
##  marital.married  marital.single   education.secondary education.tertiary
```

```
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000      Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000      1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :1.0000      Median :0.0000
##  Mean   :0.6019   Mean   :0.2829   Mean   :0.5132      Mean   :0.2942
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000      3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000      Max.   :1.0000
##  education.unknown  default.yes         balance         housing.yes
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.07346   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.07687   Median :1.0000
##  Mean   :0.04107   Mean   :0.01803   Mean   :0.08517   Mean   :0.5558
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.08577   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##     loan.yes      contact.telephone contact.unknown      day
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.2333
##  Median :0.0000   Median :0.00000   Median :0.000   Median :0.5000
##  Mean   :0.1602   Mean   :0.06428   Mean   :0.288   Mean   :0.4935
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.000   3rd Qu.:0.6667
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.000   Max.   :1.0000
##    month.aug        month.dec          month.feb         month.jan
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.000000   Median :0.00000   Median :0.00000
##  Mean   :0.1382   Mean   :0.004733   Mean   :0.05859   Mean   :0.03103
##  3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.000000   Max.   :1.00000   Max.   :1.00000
##    month.jul        month.jun        month.mar         month.may
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.1525   Mean   :0.1181   Mean   :0.01055   Mean   :0.3045
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##    month.nov         month.oct         month.sep          duration
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.02094
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.03660
##  Mean   :0.08781   Mean   :0.01632   Mean   :0.01281   Mean   :0.05249
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.06486
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##     campaign           pdays            previous        poutcome.other
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.01613   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.02845   Mean   :0.04725   Mean   :0.00211   Mean   :0.0407
##  3rd Qu.:0.03226   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##  poutcome.success  poutcome.unknown     y.yes
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.000
##  Median :0.00000   Median :1.0000   Median :0.000
##  Mean   :0.03342   Mean   :0.8175   Mean   :0.117
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.000
```

```
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.000
```

```
summary(df_test_norm)
```

```
##       age          job.blue.collar  job.entrepreneur  job.housemaid
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.2059   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.2941   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.3260   Mean   :0.2092   Mean   :0.03716   Mean   :0.02477
##  3rd Qu.:0.4412   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##  job.management    job.retired      job.self.employed  job.services
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.2143   Mean   :0.05087   Mean   :0.04048   Mean   :0.09224
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##   job.student      job.technician   job.unemployed     job.unknown
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :0.000000
##  Mean   :0.01858   Mean   :0.1699   Mean   :0.02831   Mean   :0.008405
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.000000
##  marital.married  marital.single   education.secondary education.tertiary
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000      Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000      1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :1.0000      Median :0.0000
##  Mean   :0.6187   Mean   :0.2645   Mean   :0.5101      Mean   :0.2986
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000      3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000      Max.   :1.0000
##  education.unknown  default.yes        balance         housing.yes
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.04540   1st Qu.:0.000
##  Median :0.00000   Median :0.00000   Median :0.05043   Median :1.000
##  Mean   :0.04136   Mean   :0.01681   Mean   :0.06356   Mean   :0.566
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.06433   3rd Qu.:1.000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000
##     loan.yes      contact.telephone contact.unknown       day
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.2667
##  Median :0.0000   Median :0.00000   Median :0.0000   Median :0.5000
##  Mean   :0.1528   Mean   :0.06658   Mean   :0.2929   Mean   :0.4972
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.6667
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##    month.aug      month.dec          month.feb        month.jan
##  Min.   :0.00   Min.   :0.000000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.00   Median :0.000000   Median :0.0000   Median :0.00000
##  Mean   :0.14   Mean   :0.004424   Mean   :0.0491   Mean   :0.03274
##  3rd Qu.:0.00   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.00   Max.   :1.000000   Max.   :1.0000   Max.   :1.00000
##    month.jul        month.jun        month.mar         month.may
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
```

```
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.1562   Mean   :0.1175   Mean   :0.01084   Mean   :0.3092
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##    month.nov         month.oct         month.sep         duration
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.03310
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.05991
##  Mean   :0.08604   Mean   :0.0177   Mean   :0.0115   Mean   :0.08605
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.10758
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##     campaign           pdays            previous      poutcome.other
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.02041   Median :0.00000   Median :0.0000   Median :0.00000
##  Mean   :0.03660   Mean   :0.04675   Mean   :0.0217   Mean   :0.04357
##  3rd Qu.:0.04082   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##  poutcome.success  poutcome.unknown     y.yes
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.02853   Mean   :0.8195   Mean   :0.1152
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
```

Problem 2

Importing Data

```
library(arules)
```

```
## Warning: package 'arules' was built under R version 3.4.4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

```
## The following object is masked from 'package:kernlab':
##
##     size
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
plants <- read.transactions("https://archive.ics.uci.edu/ml/machine-learning-databases/plants/plants.da

#summary(plants)
```

```
# Forming sparse matrix of the transaction data
plants1 <- as.data.frame(as(plants, "matrix"))
```

```
plants1[,1:70] <- lapply(plants1[,1:70], as.integer)

# Making new variable for preserving the original data
k <- sample(34781,2000)
samp <- plants1[k,]
```

Method 2: Using kmeans() for the model by first finding Principle Compenent Analysis of the data

```
# Data

plants3 <- plants1

library(caret)
```

```
# PCA analysis using BoxCox

trans <- preProcess(plants3[,1:70], method = c("BoxCox","center","scale","pca"))

pca<- predict(trans,plants3[,1:70])
```

```
## Warning in is.na(lam): is.na() applied to non-(list or vector) of type
## 'NULL'
```

```
head(pca)
```

```
##                              PC1        PC2        PC3        PC4
## abelia                  -2.363320 -0.9964322 -0.6384921 -0.5616725
## abelia x grandiflora    -2.363320 -0.9964322 -0.6384921 -0.5616725
## abelmoschus              1.743409 -4.5668129 -0.2124728 -2.5304105
## abelmoschus esculentus   1.698308 -4.4875365 -0.2107628 -2.2816549
## abelmoschus moschatus   -2.857036 -0.4380101 -0.8550407 -0.2236621
## abies                   14.676337  9.8601767 -4.7096665 -2.0542527
##                                PC5         PC6       PC7         PC8
## abelia                  -0.08342611 -0.04771028  0.181208 -0.4734805
## abelia x grandiflora    -0.08342611 -0.04771028  0.181208 -0.4734805
## abelmoschus             -0.40474839  3.18979846  3.322861  1.7107869
## abelmoschus esculentus  -0.54982897  2.36926272  2.821413  1.7126086
## abelmoschus moschatus    0.94851735  1.57864517  1.448945  0.2482808
## abies                   -2.55819910  0.70253388 -2.169991  1.7946008
##                                PC9        PC10       PC11       PC12
## abelia                   1.13918093 -0.09890436 -0.2680605 0.3522824
## abelia x grandiflora     1.13918093 -0.09890436 -0.2680605 0.3522824
## abelmoschus             -0.40041331 -1.18117627  1.0131713 0.9954396
## abelmoschus esculentus   0.07336372 -0.87334174 -1.8333507 0.4020818
## abelmoschus moschatus   -0.54402146 -0.69422135  2.0232595 0.3557644
## abies                    0.41747867 -0.97368858 -1.1724034 0.4516632
##                               PC13        PC14       PC15       PC16
## abelia                   0.02850500  0.06003033 -0.3020565 -0.1700643
## abelia x grandiflora     0.02850500  0.06003033 -0.3020565 -0.1700643
## abelmoschus             -0.05945455  0.31224964  0.3719323 -0.4146235
## abelmoschus esculentus   0.32396122 -0.07181356  0.3223300 -0.3212091
## abelmoschus moschatus   -0.63448310  0.14025059  0.2306581  0.1632809
## abies                    0.41839712  2.56974297  1.5538317 -0.2583637
##                               PC17       PC18       PC19        PC20
## abelia                  -0.68075667  0.4779729  0.3361087 -0.19776469
```

```
## abelia x grandiflora   -0.68075667  0.4779729  0.3361087 -0.19776469
## abelmoschus             0.22494397  0.2555790  1.0074287  0.04464414
## abelmoschus esculentus  0.05485747  0.4225649  0.9762965 -0.11712980
## abelmoschus moschatus   0.09704899  0.0747225 -0.5843541 -0.21698805
## abies                   0.33672106 -1.3835669  0.4513056  1.76187249
##                               PC21       PC22       PC23        PC24
## abelia                 -0.52235718 -0.1030986  0.1641071 -1.10010970
## abelia x grandiflora   -0.52235718 -0.1030986  0.1641071 -1.10010970
## abelmoschus             0.64650891 -0.4688548 -0.6277806  0.13375265
## abelmoschus esculentus  0.61421690 -0.3466219 -0.7204163  0.04085884
## abelmoschus moschatus  -0.31044611  0.0954600  0.3711566  1.62622940
## abies                   0.02816119 -1.5623135  3.3700450 -0.44126816
##                               PC25       PC26         PC27       PC28
## abelia                 -0.12657473  0.7877641  0.249371960 -0.4453595
## abelia x grandiflora   -0.12657473  0.7877641  0.249371960 -0.4453595
## abelmoschus            -0.04975872 -0.5564032 -0.433855607  0.3887175
## abelmoschus esculentus -0.10876093 -0.6931403 -0.403504885  0.4126448
## abelmoschus moschatus   0.36947814  0.7533770  0.004712676 -0.2658289
## abies                   0.34573649  1.7789431 -1.970888497 -0.3691903
##                               PC29       PC30        PC31        PC32
## abelia                 -0.2485916 -0.74278410  0.56762444 -0.07532935
## abelia x grandiflora   -0.2485916 -0.74278410  0.56762444 -0.07532935
## abelmoschus            -0.2903802 -0.34701257 -0.16857337 -0.16666085
## abelmoschus esculentus -0.3265784 -0.33036507 -0.20299027 -0.20139812
## abelmoschus moschatus   0.1072472 -0.02452656 -0.02110746  0.14683606
## abies                   1.0641816 -1.13271428 -1.74918759 -2.31656475
##                               PC33        PC34        PC35        PC36
## abelia                  0.001966704  0.22896343 -0.12128306 -0.04551171
## abelia x grandiflora    0.001966704  0.22896343 -0.12128306 -0.04551171
## abelmoschus            -0.445767156 -0.34893933  0.67024836 -0.66256469
## abelmoschus esculentus -0.499151026 -0.35296027  0.69814048 -0.65340186
## abelmoschus moschatus   0.071449189 -0.02562031  0.04357449 -0.04797041
## abies                   1.391002618  0.19185114 -0.13461111 -0.32618172
##                               PC37        PC38        PC39        PC40
## abelia                 -0.4916308  0.36467011 -0.61716067 -0.36888365
## abelia x grandiflora   -0.4916308  0.36467011 -0.61716067 -0.36888365
## abelmoschus            -2.1999704  1.00612535 -2.02999095 -1.42821192
## abelmoschus esculentus -2.2761508  1.01632575 -2.09389419 -1.44486169
## abelmoschus moschatus   0.1063877 -0.02880338  0.13122237  0.06562209
## abies                  -0.3949787  0.29520835 -0.01449247  0.41270622
##                               PC41        PC42        PC43         PC44
## abelia                  0.28138590 -0.39642733  0.03729649  0.067540690
## abelia x grandiflora    0.28138590 -0.39642733  0.03729649  0.067540690
## abelmoschus             0.40494486 -1.45703050 -0.36854502 -0.003944928
## abelmoschus esculentus  0.41854963 -1.48088433 -0.39200829  0.006334633
## abelmoschus moschatus  -0.03137518  0.03153023  0.04290167  0.004198505
## abies                  -1.18964844 -0.05123395  0.40905653  0.235706087
##                               PC45        PC46        PC47         PC48
## abelia                 0.022912836 -0.17041368  0.19699276 -0.338307632
## abelia x grandiflora   0.022912836 -0.17041368  0.19699276 -0.338307632
## abelmoschus            2.386242067 -0.04659075  0.21883528 -0.111627154
## abelmoschus esculentus 2.368101378 -0.05391284  0.18835006 -0.099013171
## abelmoschus moschatus  0.007254685 -0.01286911  0.04189869 -0.002814384
## abies                  0.257905228  1.51377642 -0.49221350 -0.598851382
```

```
##                                PC49
## abelia                  -0.165944368
## abelia x grandiflora    -0.165944368
## abelmoschus             -1.172280475
## abelmoschus esculentus  -1.159586020
## abelmoschus moschatus    0.003102498
## abies                   -0.012177111
```

```
# Model
# Selecting number of clusters = 6
plants3_clusters <- kmeans(plants3[1:1000,2:5],6)
```

```
plants3_clusters$size
```

```
## [1] 113  21  24 771  47  24
```

```
# Examining the coordinates of clusters
plants3_clusters$centers
```

```
##          ak        al        ar        az
## 1 0.0000000 1.0000000 0.5840708 0.0000000
## 2 0.0952381 0.0000000 1.0000000 0.1904762
## 3 0.0000000 1.0000000 0.8750000 1.0000000
## 4 0.0000000 0.0000000 0.0000000 0.2334630
## 5 1.0000000 0.0000000 0.0000000 0.3191489
## 6 1.0000000 0.9166667 0.9583333 0.6250000
```

Visualization of cluster

```
# Importing the .csv file of the data
plant_csv <- read.transactions("https://archive.ics.uci.edu/ml/machine-learning-databases/plants/plants


library(klaR)
```

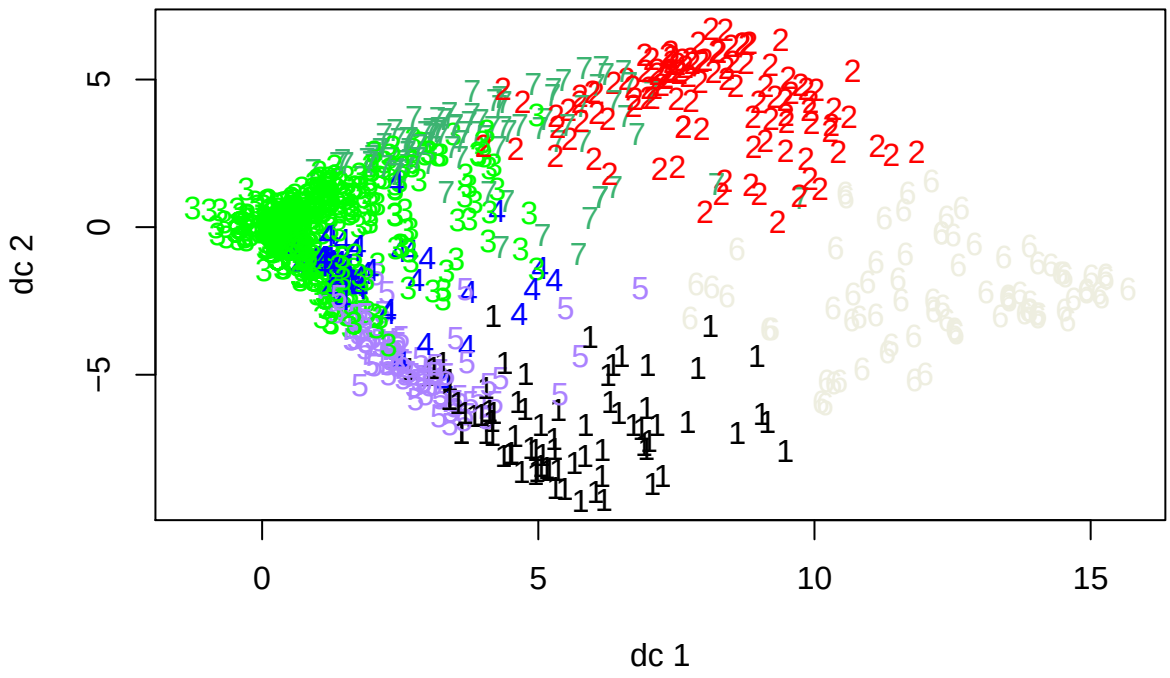
```
## Loading required package: MASS
```

```
library(MASS)
library("fpc")
```

```
## Warning: package 'fpc' was built under R version 3.4.4
```

```
# Model
# Selecting number of clusters = 6
#plants2_clusters <- kmodes(plants2[1:1000,2:5],6)

plant_cluster <- kmodes(samp,7,iter.max = 10, weighted = FALSE)
plotcluster(samp,plant_cluster$cluster)
```



Problem 2

Question 3

```
# Data
plant <- read.transactions("https://archive.ics.uci.edu/ml/machine-learning-databases/plants/plants.data

summary(plant)
```

```
## transactions as itemMatrix in sparse format with
##  34781 rows (elements/itemsets/transactions) and
##  70 columns (items) and a density of 0.1240883
##
## most frequent items:
##      ca      tx      or      az      fl (Other)
##   11676    8483    7028    6778    6621  261528
##
## element (itemset/transaction) length distribution:
## sizes
##     1     2     3     4     5     6     7     8     9    10    11    12
## 11566  4874  2954  2107  1366  1094   859   744   655   562   503   421
##    13    14    15    16    17    18    19    20    21    22    23    24
##   421   333   322   284   252   241   249   207   200   212   198   195
```

```
##    25    26    27    28    29    30    31    32    33    34    35    36
##   155   152   190   179   159   146   140   146   148   147   119   123
##    37    38    39    40    41    42    43    44    45    46    47    48
##   110   124   118   114    83    85   102    90    90    75    64    89
##    49    50    51    52    53    54    55    56    57    58    59    60
##    63    60    64    68    56    54    51    59    47    45    39    59
##    61    62    63    64    65    66    67    68    69
##    52    43    39    43    47    60    35    26     4
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   3.000   8.686   9.000  69.000
##
## includes extended item information - examples:
##   labels
## 1     ab
## 2     ak
## 3     al
##
## includes extended transaction information - examples:
##          transactionID
## 1               abelia
## 2 abelia x grandiflora
## 3          abelmoschus
```

```
# Examining transaction data
inspect(plant[1:3])
```

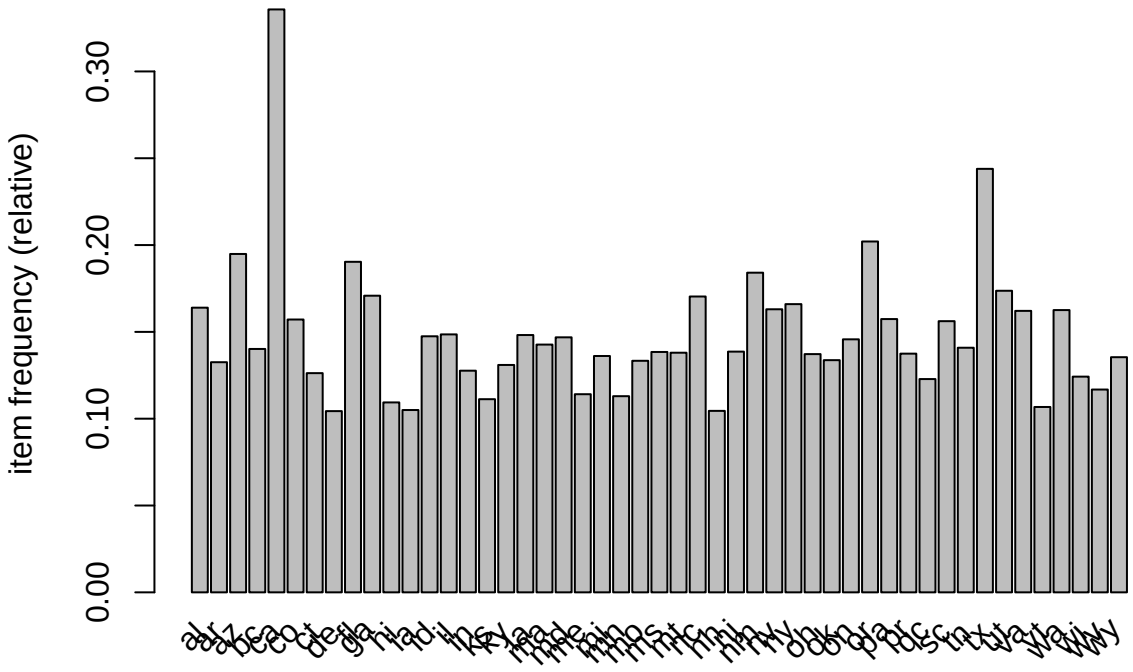
```
##     items                                          transactionID
## [1] {fl,nc}                                        abelia
## [2] {fl,nc}                                        abelia x grandiflora
## [3] {ct,dc,fl,hi,il,ky,la,md,mi,ms,nc,pr,sc,va,vi} abelmoschus
```

```
itemFrequency(plant[,1:3])
```

```
##         ab         ak         al
## 0.09798453 0.08536270 0.16394008
```

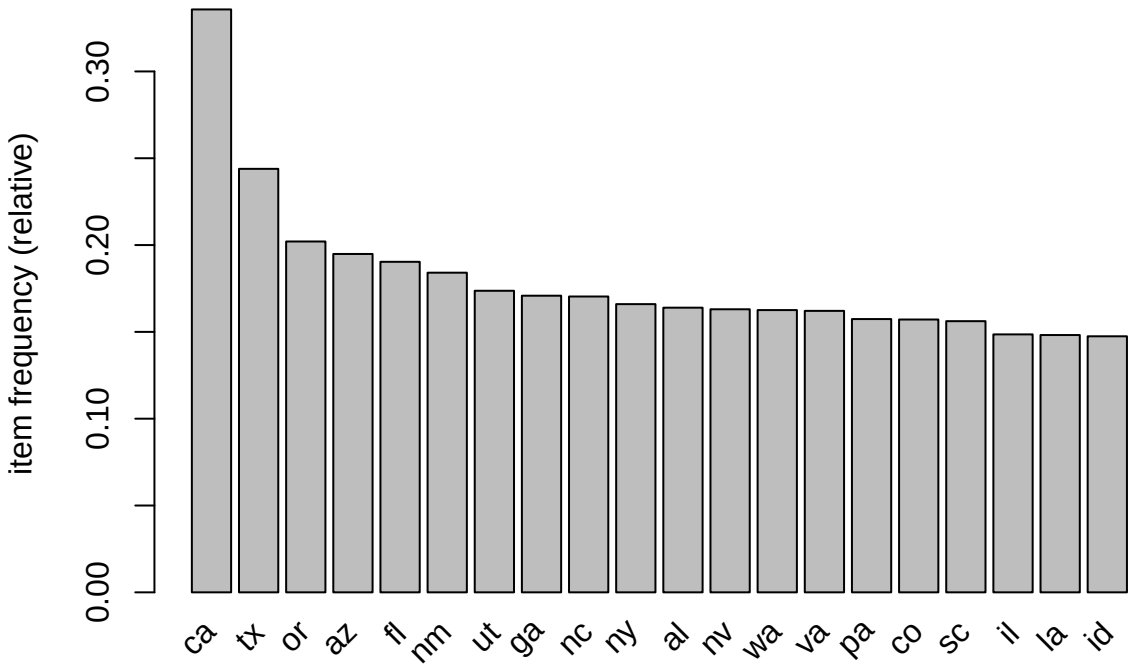
```
# Vizualizing item support - item frequency plot

itemFrequencyPlot(plant, support = 0.1)
```



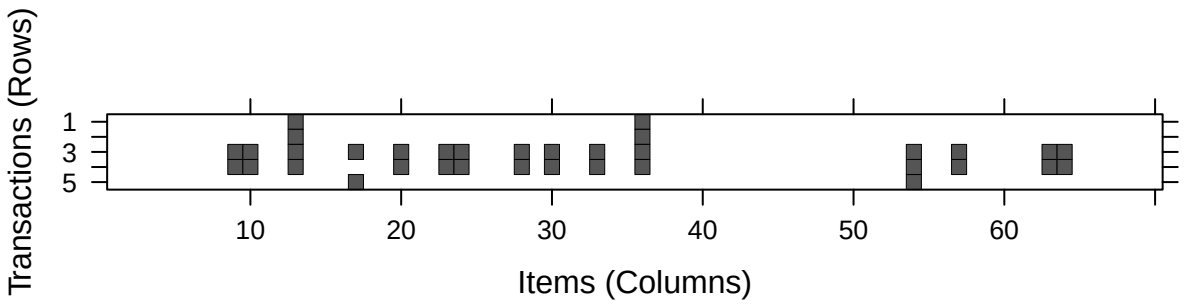

```
# Limiting the plot to a specific number

itemFrequencyPlot(plant, topN = 20)
```



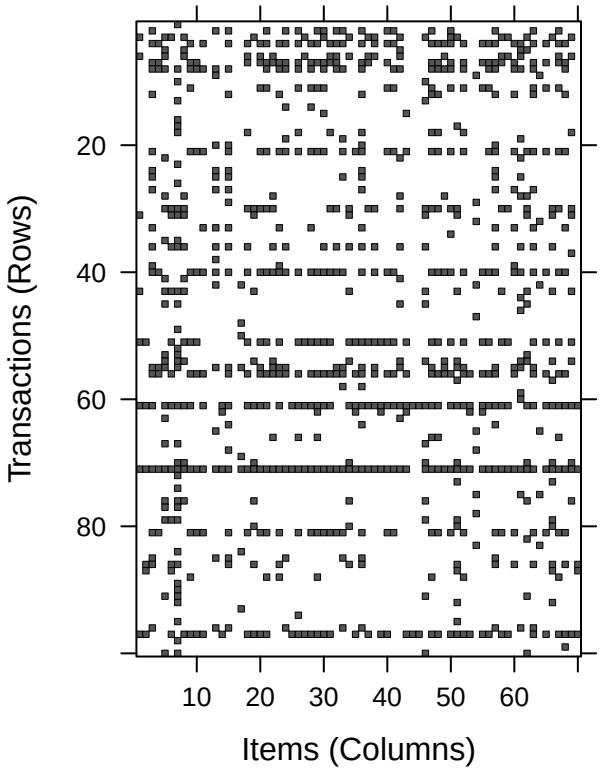

```
# Plotting the sparse matrix

image(plant[1:5])
```




```
# Selecting random transaction

image(sample(plant, 100))
```




```
# Finding associations

apriori(plant)
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.8    0.1    1 none FALSE            TRUE       5     0.1      1
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 3478
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[70 item(s), 34781 transaction(s)] done [0.04s].
## sorting and recoding items ... [49 item(s)] done [0.01s].
## creating transaction tree ... done [0.02s].
## checking subsets of size 1 2 3 4 5 done [0.07s].
## writing ... [506 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].

## set of 506 rules
```