Preprocessing Bank Marketing Data

Crupanshu Ashishbhai Udani, Karan Jeeten Thakkar

09/05/20

## Importing libraries

library(dplyr)# Used for easier data manipulation  
library(caret)# Used for finding correlated attributes  
library(rpart)# Used for decision tree  
print("All Libraries loaded")

## [1] "All Libraries loaded"

## Reading the dataset

dataset <- read.csv('../data/bank-additional-full.csv',header = TRUE,sep = ';',na.strings = c("","NA","unknown"))  
summary(dataset)

## age job marital education   
## Min. :17.00 Length:41188 Length:41188 Length:41188   
## 1st Qu.:32.00 Class :character Class :character Class :character   
## Median :38.00 Mode :character Mode :character Mode :character   
## Mean :40.02   
## 3rd Qu.:47.00   
## Max. :98.00   
## default housing loan contact   
## Length:41188 Length:41188 Length:41188 Length:41188   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## month day\_of\_week duration campaign   
## Length:41188 Length:41188 Min. : 0.0 Min. : 1.000   
## Class :character Class :character 1st Qu.: 102.0 1st Qu.: 1.000   
## Mode :character Mode :character Median : 180.0 Median : 2.000   
## Mean : 258.3 Mean : 2.568   
## 3rd Qu.: 319.0 3rd Qu.: 3.000   
## Max. :4918.0 Max. :56.000   
## pdays previous poutcome emp.var.rate   
## Min. : 0.0 Min. :0.000 Length:41188 Min. :-3.40000   
## 1st Qu.:999.0 1st Qu.:0.000 Class :character 1st Qu.:-1.80000   
## Median :999.0 Median :0.000 Mode :character Median : 1.10000   
## Mean :962.5 Mean :0.173 Mean : 0.08189   
## 3rd Qu.:999.0 3rd Qu.:0.000 3rd Qu.: 1.40000   
## Max. :999.0 Max. :7.000 Max. : 1.40000   
## cons.price.idx cons.conf.idx euribor3m nr.employed   
## Min. :92.20 Min. :-50.8 Min. :0.634 Min. :4964   
## 1st Qu.:93.08 1st Qu.:-42.7 1st Qu.:1.344 1st Qu.:5099   
## Median :93.75 Median :-41.8 Median :4.857 Median :5191   
## Mean :93.58 Mean :-40.5 Mean :3.621 Mean :5167   
## 3rd Qu.:93.99 3rd Qu.:-36.4 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :94.77 Max. :-26.9 Max. :5.045 Max. :5228   
## y   
## Length:41188   
## Class :character   
## Mode :character   
##   
##   
##

# 1. Preprocessing data

## 1.1 Cleaning Data

As a part of this step we’ll be removing all the rows that contain NA values ( i.e. unknown in original dataset).

data <- na.omit(dataset)  
dim(data)

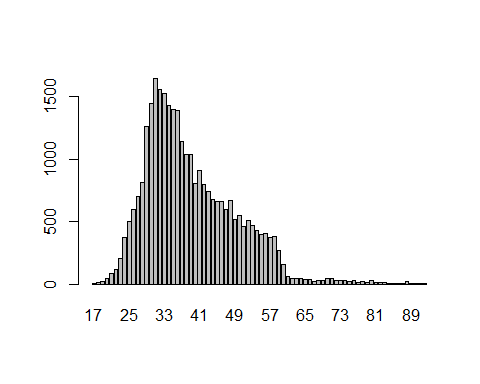
## [1] 30488 21

## 1.2 Exploring Data

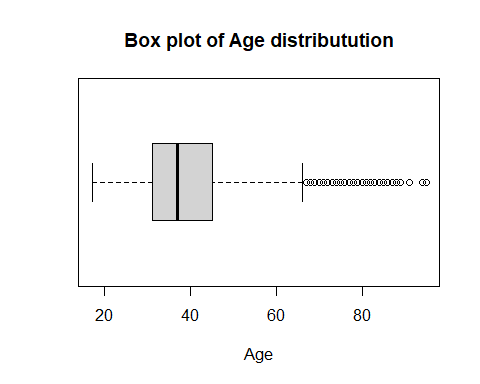
As a part of this step we would be exploring each data field and checking the values.

#### Age:

data$age<-as.integer(data$age)  
barplot(table(data$age))



boxplot(data$age,horizontal = T,xlab = "Age",main="Box plot of Age distributution")

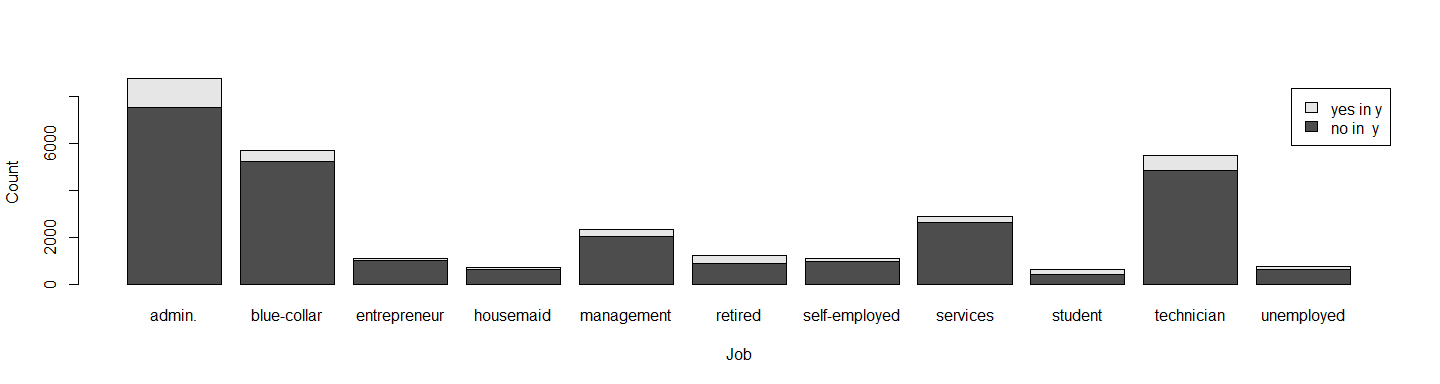


#abline(v=37.00,col="lightgrey")  
#text(37, 10, "37.00", col = "darkgrey")

Summary: Age data is mostly skewed towards lower,middle age groups thus older age group could cause issue in the model.

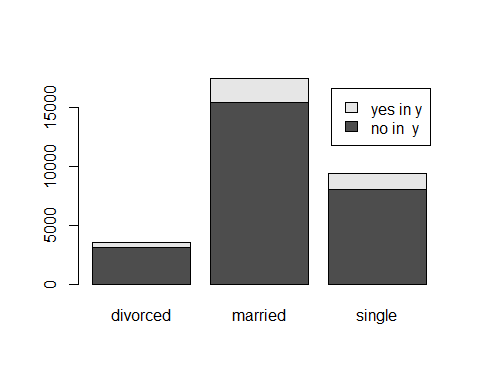
#### Job

job <- table(data$y,data$job)  
barplot(job,names.arg = names(job),legend.text = c("no in y","yes in y"),xlab = "Job",ylab = "Count")



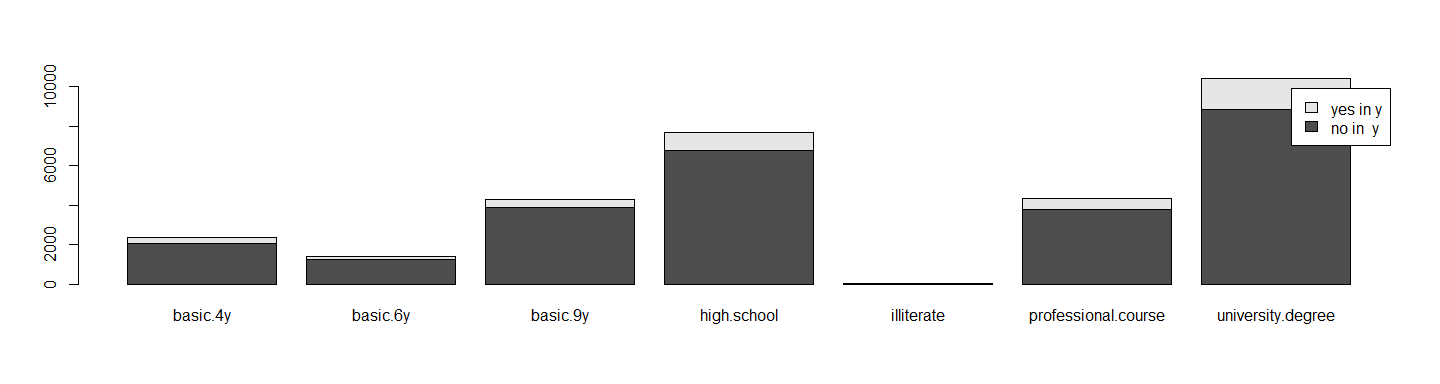
#### Marital

marital <- table(data$y,data$marital)  
barplot(marital,legend.text = c("no in y","yes in y"))



#### Education

education <- table(data$y,data$education)  
barplot(education, legend.text = c("no in y","yes in y"))



As per graph and data clearly for education *illiterate* has very low no of row of data (only **11**).

#### Default

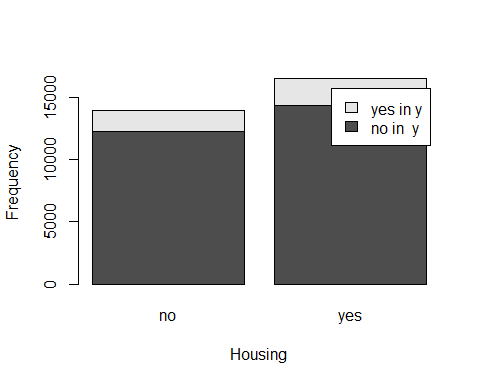
table(data$default)

##   
## no yes   
## 30485 3

Clearly for attribute *default* has only 3 rows with yes field. So the field would not be useful. Therefore the field can be dropped.

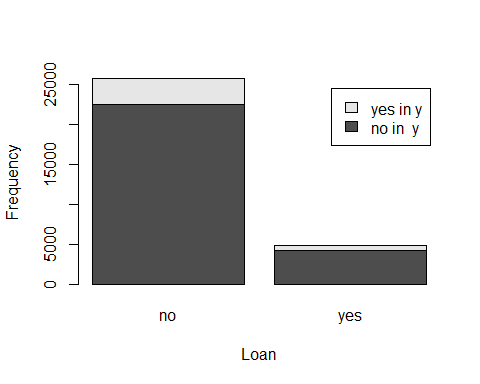
#### Housing

housing <- table(data$y,data$housing)  
barplot(housing,legend.text = c("no in y","yes in y"),xlab = "Housing",ylab = "Frequency")



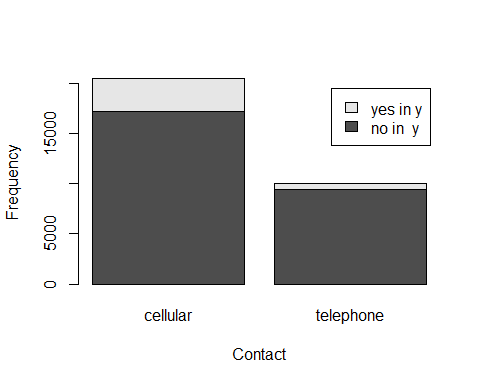
#### Loan

loan <- table(data$y,data$loan)  
barplot(loan,xlab = "Loan",ylab = "Frequency",legend.text = c("no in y","yes in y"))



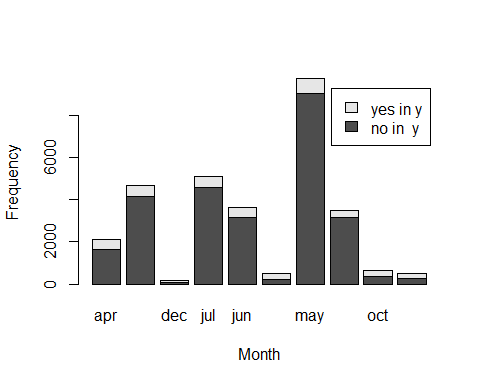
#### Contact

contact <- table(data$y,data$contact)  
barplot(contact,xlab = "Contact",ylab = "Frequency",legend.text = c("no in y","yes in y"))



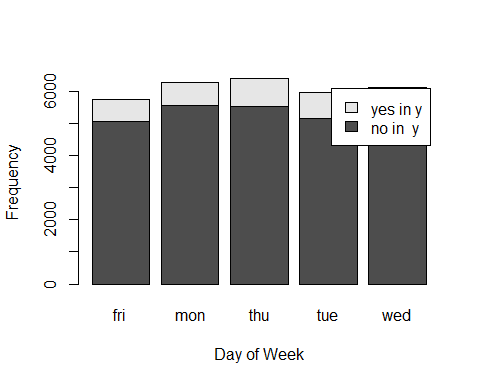
#### Month

month <- table(data$y,data$month)  
barplot(month,xlab = "Month",ylab = "Frequency",legend.text = c("no in y","yes in y"))



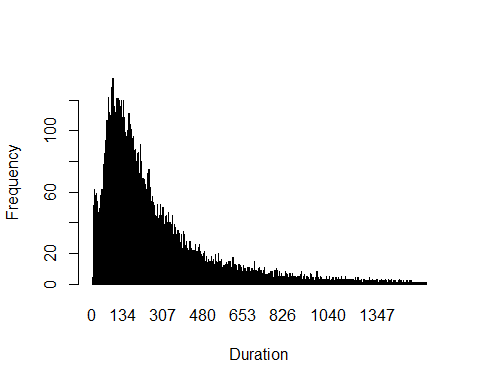
#### Day of Week

day\_of\_week <- table(data$y,data$day\_of\_week)  
barplot(day\_of\_week,xlab = "Day of Week",ylab = "Frequency",legend.text = c("no in y","yes in y"))

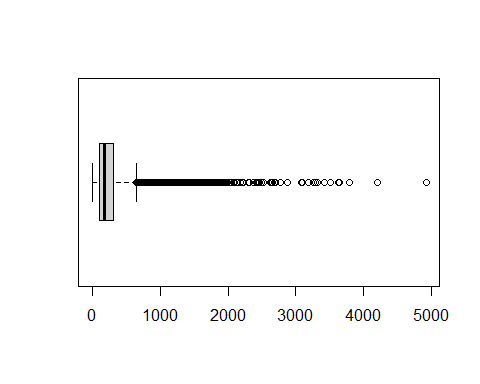


#### Duration

duration <- table(data$duration)  
barplot(duration,xlab = "Duration",ylab = "Frequency")



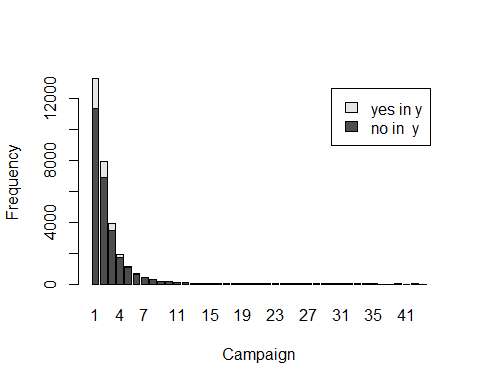
boxplot(data$duration,horizontal = T)



The above plots clearly shows that there are more number of calls that has happened for shorted duration then larger duration.

#### Campaign

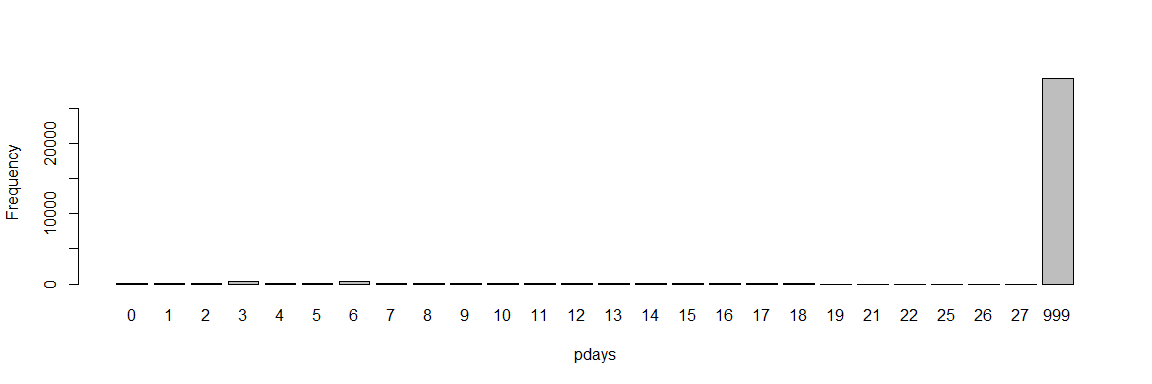
data$campaign <- as.integer(data$campaign)  
campaign <- table(data$y,data$campaign)  
barplot(campaign,xlab = "Campaign",ylab = "Frequency",legend.text = c("no in y","yes in y"))



The above plot clearly shows there is more data for campaign with less contact and very less data with more contacts done.

#### pdays

data$pdays<-as.integer(data$pdays)  
pdaysCount <- table(data$pdays)  
barplot(table(data$pdays),xlab = "pdays",ylab = "Frequency")

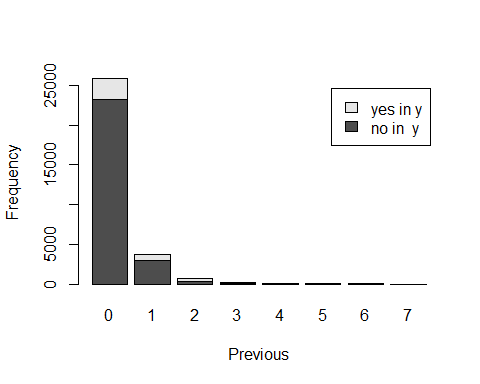


data %>% select(pdays) %>% filter(pdays == 999) %>% count()

## n  
## 1 29178

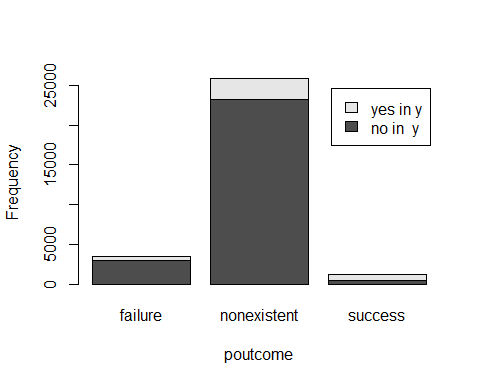
#### Previous

data$previous <- as.integer(data$previous)  
previous <- table(data$y,data$previous)  
barplot(previous,xlab = "Previous",ylab = "Frequency",legend.text = c("no in y","yes in y"))

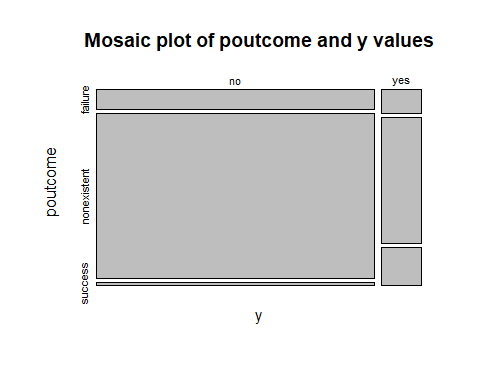


#### pOutcome

poutcome <- table(data$y,data$poutcome)  
barplot(poutcome,xlab = "poutcome",ylab = "Frequency",legend.text = c("no in y","yes in y"))

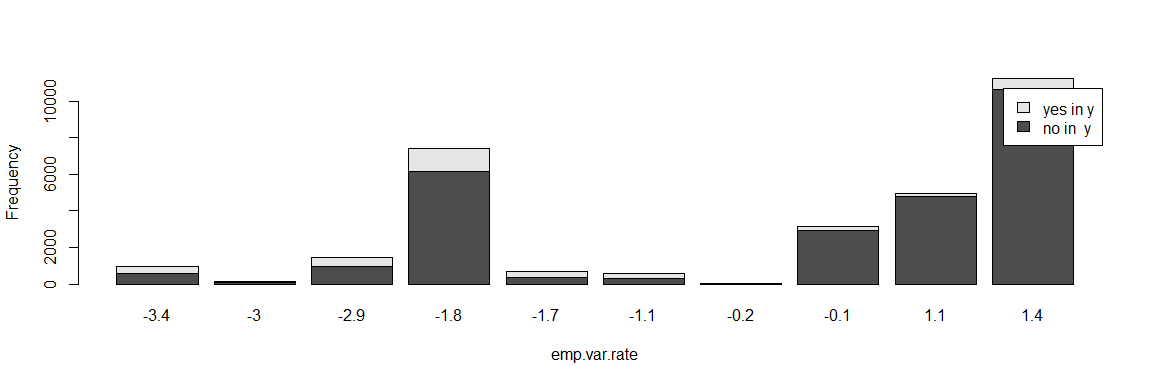


plot(poutcome,main = "Mosaic plot of poutcome and y values",xlab = "y",ylab = "poutcome")

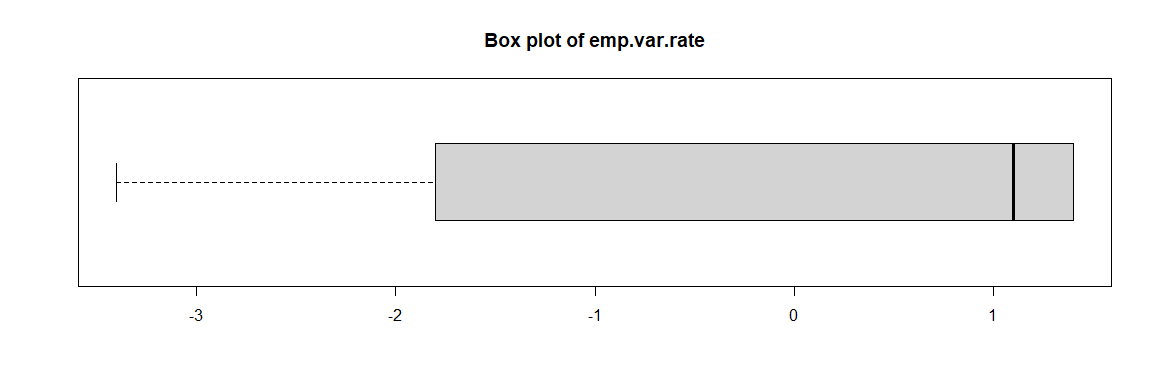


#### Social and Economical Context

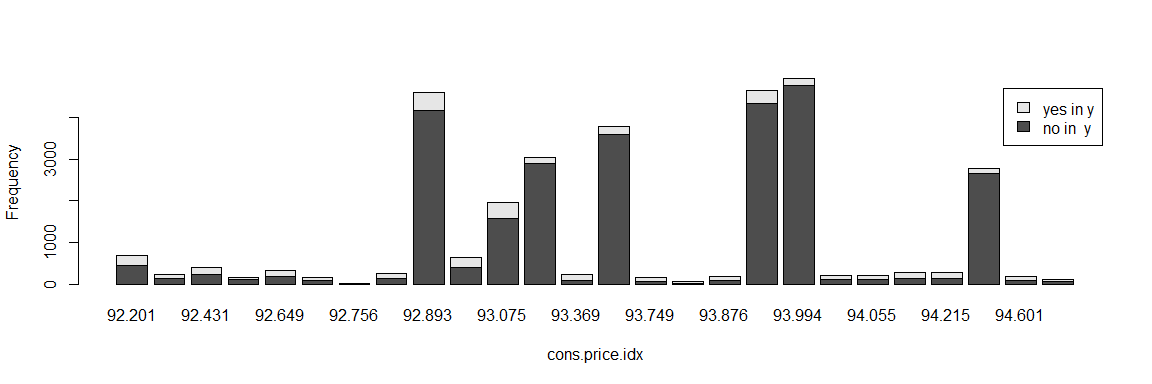
## emp.var.rate  
emp.var.rate <- table(data$y,data$emp.var.rate)  
barplot(emp.var.rate,xlab = "emp.var.rate",ylab = "Frequency",legend.text = c("no in y","yes in y"))



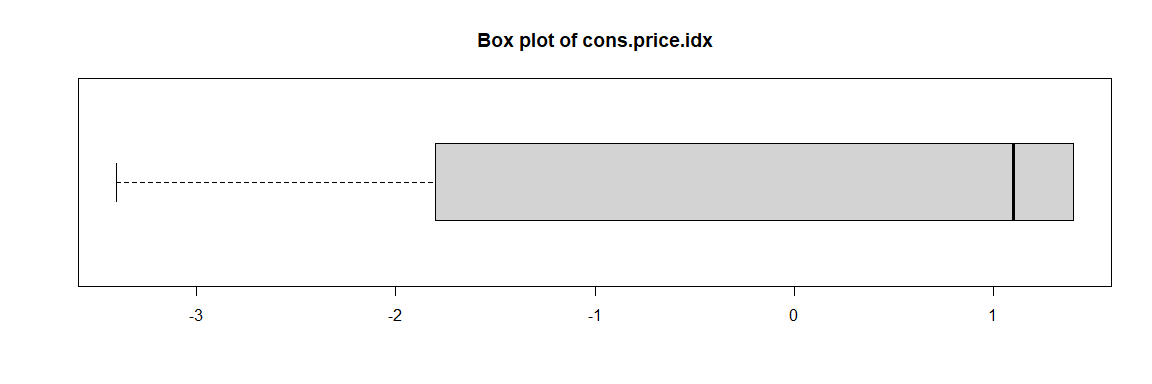
boxplot(data$emp.var.rate,horizontal = T,main ="Box plot of emp.var.rate")



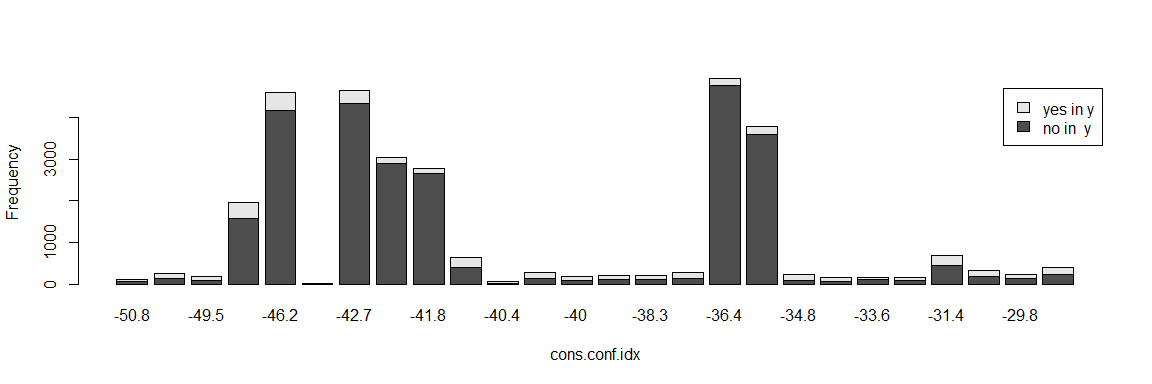
## cons.price.idx  
cons.price.idx <- table(data$y,data$cons.price.idx)  
barplot(cons.price.idx,xlab = "cons.price.idx",ylab = "Frequency",legend.text = c("no in y","yes in y"))



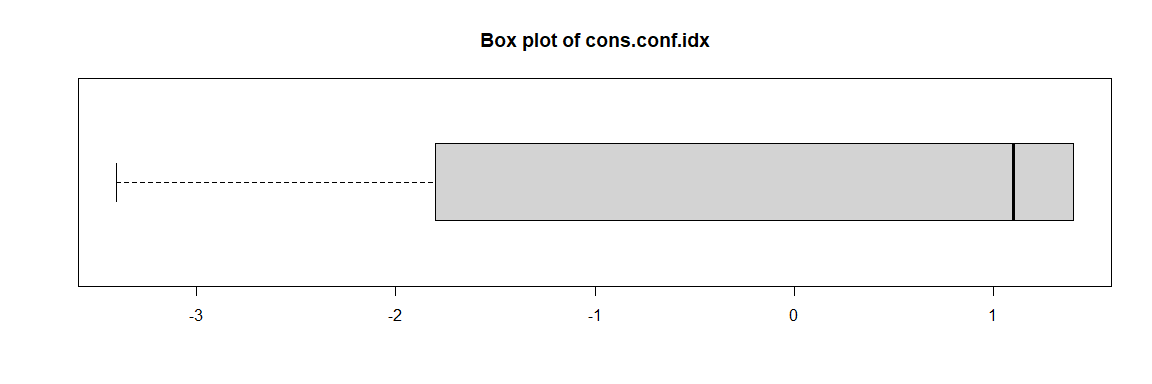
boxplot(data$emp.var.rate,horizontal = T,main ="Box plot of cons.price.idx")



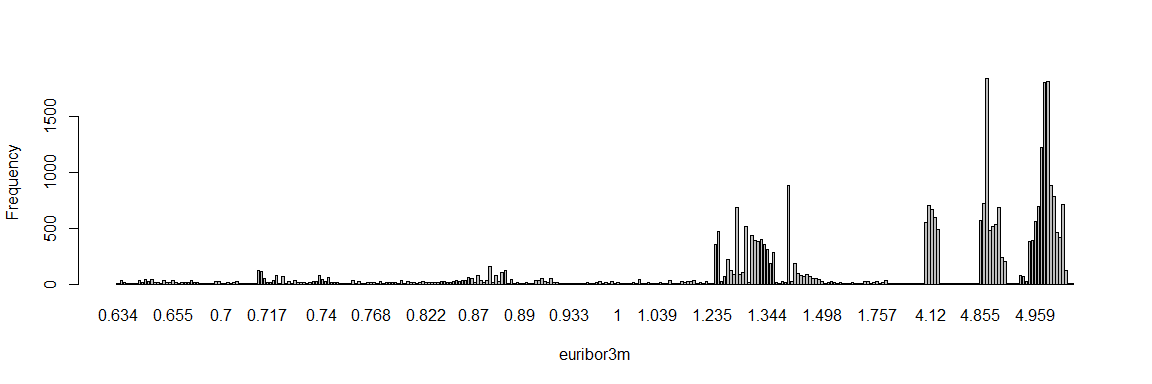
## cons.conf.idx  
cons.conf.idx <- table(data$y,data$cons.conf.idx)  
barplot(cons.conf.idx,xlab = "cons.conf.idx",ylab = "Frequency",legend.text = c("no in y","yes in y"))



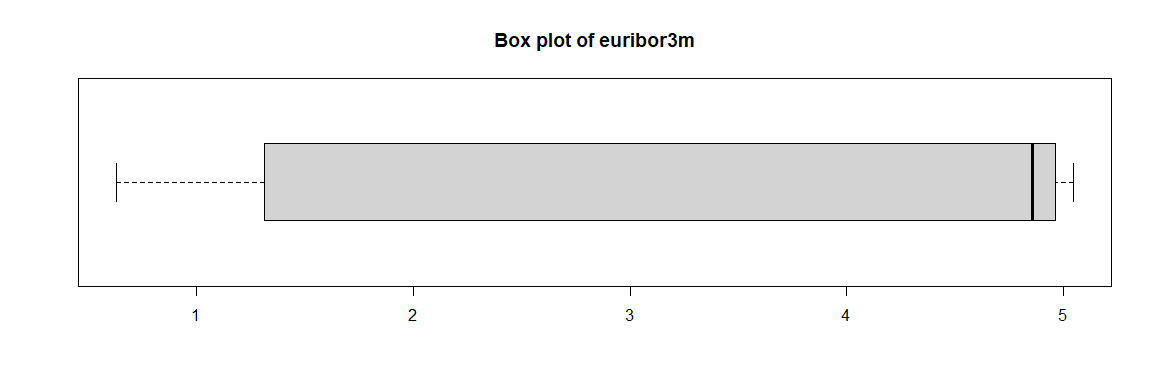
boxplot(data$emp.var.rate,horizontal = T,main ="Box plot of cons.conf.idx")



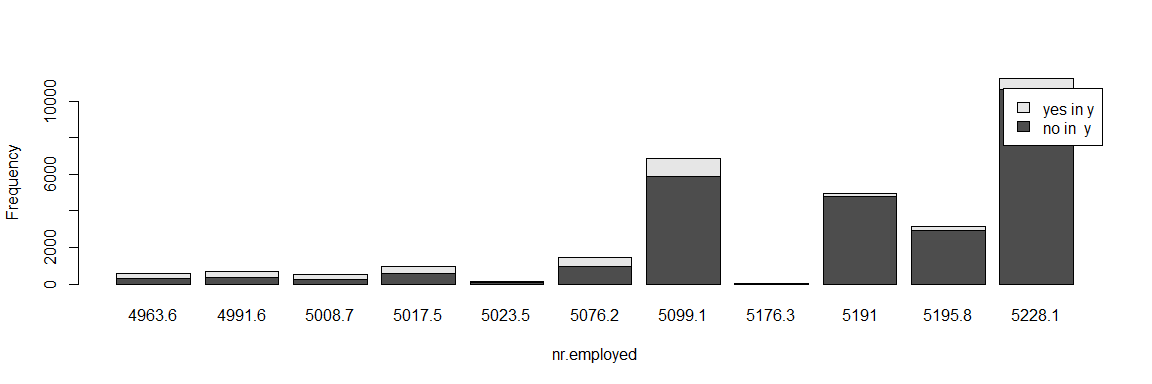
## euribor3m  
euribor3m <- table(data$euribor3m)  
barplot(euribor3m,xlab = "euribor3m",ylab = "Frequency")



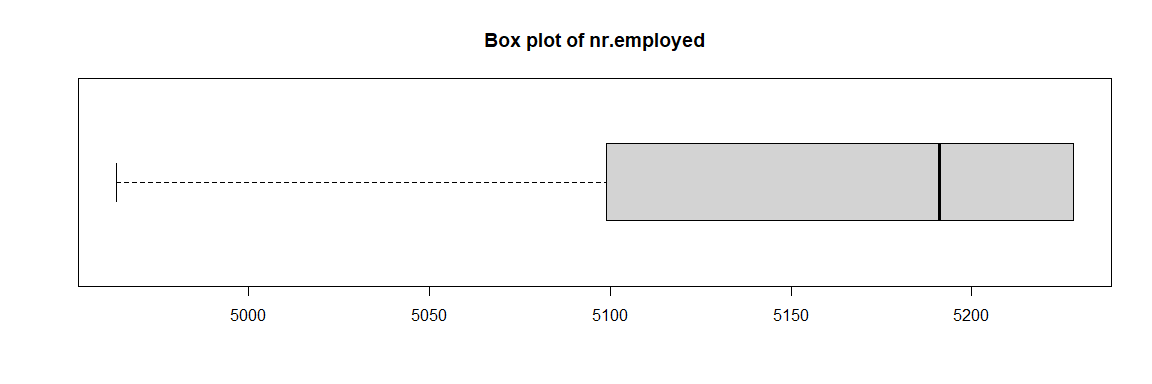
boxplot(data$euribor3m,horizontal = T,main ="Box plot of euribor3m")



## nr.employed  
nr.employed <- table(data$y,data$nr.employed)  
barplot(nr.employed,xlab = "nr.employed",ylab = "Frequency",legend.text = c("no in y","yes in y"))



boxplot(data$nr.employed,horizontal = T,main ="Box plot of nr.employed")



## 1.3 Changing the categorical/logical data to factors

data$job<-as.factor(data$job)  
data$marital<-as.factor(data$marital)  
data$education<-as.factor(data$education)  
data$default<-as.factor(data$default)  
data$housing<-as.factor(data$housing)  
data$loan<-as.factor(data$loan)  
data$contact<-as.factor(data$contact)  
data$month<-as.factor(data$month)  
data$day\_of\_week<-as.factor(data$day\_of\_week)  
data$poutcome<-as.factor(data$poutcome)  
data$y<-as.logical(data$y=="yes")

## 1.4 Checking Similarity of attributes using variance and correlation

X <- subset(data,select = age:nr.employed) %>% data.matrix()  
Y <- data$y  
var\_X <- var(X)  
cor\_X <- cor(X,method = "pearson")  
print(var\_X)

## age job marital education  
## age 1.067818e+02 -2.032600e-01 -2.584897e+00 -2.664026e+00  
## job -2.032600e-01 1.305200e+01 4.308684e-02 7.498994e-01  
## marital -2.584897e+00 4.308684e-02 3.889562e-01 1.399601e-01  
## education -2.664026e+00 7.498994e-01 1.399601e-01 3.955768e+00  
## default 3.250554e-04 5.526905e-04 -1.901047e-05 4.827068e-05  
## housing 2.398317e-02 1.007851e-02 2.764962e-03 2.008254e-02  
## loan -4.103704e-02 -2.490454e-02 8.156407e-04 6.710217e-03  
## contact -5.807290e-02 -5.029468e-02 -1.737138e-02 -8.944244e-02  
## month -2.537882e-01 -3.117252e-01 -3.284074e-02 -4.291148e-01  
## day\_of\_week -1.940744e-01 -1.079976e-02 5.166747e-03 -4.159144e-02  
## duration 2.139279e+01 -8.641877e+00 3.589956e-01 -9.738798e+00  
## campaign -6.645874e-02 -6.915168e-02 -9.458023e-03 4.593154e-02  
## pdays -1.058997e+02 -1.479499e+01 -3.641887e+00 -1.141455e+01  
## previous 2.659573e-01 2.440179e-02 9.339427e-03 1.683181e-02  
## poutcome 4.254688e-02 1.406920e-02 5.098250e-04 1.761848e-02  
## emp.var.rate -8.388581e-01 -2.102971e-02 -7.162743e-02 -2.329675e-02  
## cons.price.idx -2.163248e-01 -3.703746e-02 -1.832794e-02 -7.401408e-02  
## cons.conf.idx 6.187092e+00 7.251174e-01 -1.116204e-01 8.011694e-01  
## euribor3m -6.699703e-01 -2.741066e-02 -9.028306e-02 6.612979e-04  
## nr.employed -5.016035e+01 -3.294145e+00 -3.406438e+00 -4.523152e-01  
## default housing loan contact  
## age 3.250554e-04 2.398317e-02 -4.103704e-02 -5.807290e-02  
## job 5.526905e-04 1.007851e-02 -2.490454e-02 -5.029468e-02  
## marital -1.901047e-05 2.764962e-03 8.156407e-04 -1.737138e-02  
## education 4.827068e-05 2.008254e-02 6.710217e-03 -8.944244e-02  
## default 9.839292e-05 -2.052206e-05 -1.538912e-05 -3.242109e-05  
## housing -2.052206e-05 2.482538e-01 8.504961e-03 -1.886831e-02  
## loan -1.538912e-05 8.504961e-03 1.319361e-01 -1.539394e-03  
## contact -3.242109e-05 -1.886831e-02 -1.539394e-03 2.209281e-01  
## month -1.229484e-04 -1.581126e-02 -2.930475e-03 2.684866e-01  
## day\_of\_week 9.692436e-05 2.891894e-04 -4.875073e-03 -8.852922e-03  
## duration -1.536564e-02 -1.017866e+00 6.008085e-01 -3.561738e+00  
## campaign -1.169139e-04 -1.518329e-02 6.518232e-03 1.021427e-01  
## pdays 4.198613e-03 -7.567435e-01 2.235582e-01 1.089930e+01  
## previous 1.368388e-05 5.491297e-03 -6.656743e-04 -4.957765e-02  
## poutcome -2.547425e-05 -2.490251e-03 -5.246848e-04 1.921822e-02  
## emp.var.rate 9.559912e-05 -4.747209e-02 1.181382e-05 2.883334e-01  
## cons.price.idx -1.580784e-05 -2.258407e-02 -4.744915e-04 1.553159e-01  
## cons.conf.idx 2.495093e-04 -8.598124e-02 -3.284053e-02 5.075453e-01  
## euribor3m 1.213365e-04 -5.184736e-02 -1.589548e-03 3.207449e-01  
## nr.employed 5.561707e-03 -1.687759e+00 6.679791e-02 9.169184e+00  
## month day\_of\_week duration campaign  
## age -2.537882e-01 -1.940744e-01 2.139279e+01 -6.645874e-02  
## job -3.117252e-01 -1.079976e-02 -8.641877e+00 -6.915168e-02  
## marital -3.284074e-02 5.166747e-03 3.589956e-01 -9.458023e-03  
## education -4.291148e-01 -4.159144e-02 -9.738798e+00 4.593154e-02  
## default -1.229484e-04 9.692436e-05 -1.536564e-02 -1.169139e-04  
## housing -1.581126e-02 2.891894e-04 -1.017866e+00 -1.518329e-02  
## loan -2.930475e-03 -4.875073e-03 6.008085e-01 6.518232e-03  
## contact 2.684866e-01 -8.852922e-03 -3.561738e+00 1.021427e-01  
## month 5.682225e+00 7.963248e-02 -4.507666e-01 -4.381748e-01  
## day\_of\_week 7.963248e-02 1.957002e+00 8.943100e+00 -1.515998e-01  
## duration -4.507666e-01 8.943100e+00 6.849436e+04 -4.869843e+01  
## campaign -4.381748e-01 -1.515998e-01 -4.869843e+01 7.399218e+00  
## pdays -2.222390e+01 -2.100246e+00 -2.436904e+03 2.975006e+01  
## previous 1.305984e-01 -3.049730e-03 2.568463e+00 -1.148540e-01  
## poutcome -6.313398e-02 8.957937e-03 3.575422e+00 3.210201e-02  
## emp.var.rate -7.031314e-01 7.770366e-02 -9.851372e+00 6.909787e-01  
## cons.price.idx -2.770239e-02 4.526633e-03 1.956049e+00 2.026366e-01  
## cons.conf.idx -8.723352e-02 2.430163e-01 -1.193876e+01 -1.519539e-01  
## euribor3m -5.150042e-01 1.032787e-01 -1.345220e+01 6.808477e-01  
## nr.employed -3.886006e+01 3.432009e+00 -7.940118e+02 3.027132e+01  
## pdays previous poutcome emp.var.rate  
## age -1.058997e+02 2.659573e-01 4.254688e-02 -8.388581e-01  
## job -1.479499e+01 2.440179e-02 1.406920e-02 -2.102971e-02  
## marital -3.641887e+00 9.339427e-03 5.098250e-04 -7.162743e-02  
## education -1.141455e+01 1.683181e-02 1.761848e-02 -2.329675e-02  
## default 4.198613e-03 1.368388e-05 -2.547425e-05 9.559912e-05  
## housing -7.567435e-01 5.491297e-03 -2.490251e-03 -4.747209e-02  
## loan 2.235582e-01 -6.656743e-04 -5.246848e-04 1.181382e-05  
## contact 1.089930e+01 -4.957765e-02 1.921822e-02 2.883334e-01  
## month -2.222390e+01 1.305984e-01 -6.313398e-02 -7.031314e-01  
## day\_of\_week -2.100246e+00 -3.049730e-03 8.957937e-03 7.770366e-02  
## duration -2.436904e+03 2.568463e+00 3.575422e+00 -9.851372e+00  
## campaign 2.975006e+01 -1.148540e-01 3.210201e-02 6.909787e-01  
## pdays 4.055120e+04 -6.213869e+01 -3.812335e+01 8.715747e+01  
## previous -6.213869e+01 2.733073e-01 -5.779512e-02 -3.397069e-01  
## poutcome -3.812335e+01 -5.779512e-02 1.470458e-01 1.038860e-01  
## emp.var.rate 8.715747e+01 -3.397069e-01 1.038860e-01 2.593386e+00  
## cons.price.idx 8.016902e+00 -5.409792e-02 4.391139e-02 7.221495e-01  
## cons.conf.idx -9.872606e+01 -6.993129e-02 3.032756e-01 1.215453e+00  
## euribor3m 1.056440e+02 -4.077546e-01 1.079400e-01 2.774508e+00  
## nr.employed 5.612674e+03 -1.918872e+01 2.707381e+00 1.089783e+02  
## cons.price.idx cons.conf.idx euribor3m nr.employed  
## age -2.163248e-01 6.187092e+00 -6.699703e-01 -5.016035e+01  
## job -3.703746e-02 7.251174e-01 -2.741066e-02 -3.294145e+00  
## marital -1.832794e-02 -1.116204e-01 -9.028306e-02 -3.406438e+00  
## education -7.401408e-02 8.011694e-01 6.612979e-04 -4.523152e-01  
## default -1.580784e-05 2.495093e-04 1.213365e-04 5.561707e-03  
## housing -2.258407e-02 -8.598124e-02 -5.184736e-02 -1.687759e+00  
## loan -4.744915e-04 -3.284053e-02 -1.589548e-03 6.679791e-02  
## contact 1.553159e-01 5.075453e-01 3.207449e-01 9.169184e+00  
## month -2.770239e-02 -8.723352e-02 -5.150042e-01 -3.886006e+01  
## day\_of\_week 4.526633e-03 2.430163e-01 1.032787e-01 3.432009e+00  
## duration 1.956049e+00 -1.193876e+01 -1.345220e+01 -7.940118e+02  
## campaign 2.026366e-01 -1.519539e-01 6.808477e-01 3.027132e+01  
## pdays 8.016902e+00 -9.872606e+01 1.056440e+02 5.612674e+03  
## previous -5.409792e-02 -6.993129e-02 -4.077546e-01 -1.918872e+01  
## poutcome 4.391139e-02 3.032756e-01 1.079400e-01 2.707381e+00  
## emp.var.rate 7.221495e-01 1.215453e+00 2.774508e+00 1.089783e+02  
## cons.price.idx 3.426629e-01 7.630402e-02 6.942141e-01 2.150816e+01  
## cons.conf.idx 7.630402e-02 2.293691e+01 2.073739e+00 2.709806e+01  
## euribor3m 6.942141e-01 2.073739e+00 3.158550e+00 1.262095e+02  
## nr.employed 2.150816e+01 2.709806e+01 1.262095e+02 5.648735e+03

print(cor\_X)

## age job marital education  
## age 1.000000000 -0.005444585 -0.401092301 -0.1296206957  
## job -0.005444585 1.000000000 0.019122994 0.1043636179  
## marital -0.401092301 0.019122994 1.000000000 0.1128336128  
## education -0.129620696 0.104363618 0.112833613 1.0000000000  
## default 0.003171223 0.015422743 -0.003072987 0.0024467303  
## housing 0.004658113 0.005598995 0.008897970 0.0202654066  
## loan -0.010933157 -0.018978343 0.003600533 0.0092883673  
## contact -0.011956371 -0.029618164 -0.059259569 -0.0956759454  
## month -0.010302995 -0.036197137 -0.022090393 -0.0905104809  
## day\_of\_week -0.013425296 -0.002136880 0.005922039 -0.0149483369  
## duration 0.007910272 -0.009139916 0.002199436 -0.0187095174  
## campaign -0.002364343 -0.007036737 -0.005575157 0.0084899026  
## pdays -0.050891389 -0.020336417 -0.028998417 -0.0284997751  
## previous 0.049230881 0.012919851 0.028644700 0.0161878684  
## poutcome 0.010737242 0.010155575 0.002131790 0.0231007859  
## emp.var.rate -0.050408785 -0.003614608 -0.071317388 -0.0072735472  
## cons.price.idx -0.035762179 -0.017513349 -0.050203018 -0.0635719314  
## cons.conf.idx 0.125017394 0.041908525 -0.037370218 0.0841088162  
## euribor3m -0.036480684 -0.004269108 -0.081453928 0.0001870845  
## nr.employed -0.064585684 -0.012131894 -0.072673249 -0.0030258693  
## default housing loan contact  
## age 0.003171223 0.0046581128 -1.093316e-02 -0.011956371  
## job 0.015422743 0.0055989946 -1.897834e-02 -0.029618164  
## marital -0.003072987 0.0088979695 3.600533e-03 -0.059259569  
## education 0.002446730 0.0202654066 9.288367e-03 -0.095675945  
## default 1.000000000 -0.0041523226 -4.271203e-03 -0.006953768  
## housing -0.004152323 1.0000000000 4.699402e-02 -0.080567448  
## loan -0.004271203 0.0469940204 1.000000e+00 -0.009016596  
## contact -0.006953768 -0.0805674478 -9.016596e-03 1.000000000  
## month -0.005199741 -0.0133124936 -3.384519e-03 0.239628207  
## day\_of\_week 0.006984823 0.0004148956 -9.594088e-03 -0.013463735  
## duration -0.005918905 -0.0078057631 6.320143e-03 -0.028954035  
## campaign -0.004333025 -0.0112027614 6.597131e-03 0.079889349  
## pdays 0.002101948 -0.0075422154 3.056377e-03 0.115151990  
## previous 0.002638772 0.0210814950 -3.505535e-03 -0.201759898  
## poutcome -0.006697198 -0.0130337367 -3.766954e-03 0.106625539  
## emp.var.rate 0.005984644 -0.0591639021 2.019645e-05 0.380921768  
## cons.price.idx -0.002722432 -0.0774320293 -2.231583e-03 0.564490709  
## cons.conf.idx 0.005252153 -0.0360319950 -1.887821e-02 0.225466431  
## euribor3m 0.006882810 -0.0585510647 -2.462340e-03 0.383963893  
## nr.employed 0.007460203 -0.0450699466 2.446840e-03 0.259555135  
## month day\_of\_week duration campaign  
## age -0.0103029949 -0.0134252965 0.0079102719 -0.002364343  
## job -0.0361971372 -0.0021368799 -0.0091399161 -0.007036737  
## marital -0.0220903934 0.0059220392 0.0021994361 -0.005575157  
## education -0.0905104809 -0.0149483369 -0.0187095174 0.008489903  
## default -0.0051997406 0.0069848231 -0.0059189048 -0.004333025  
## housing -0.0133124936 0.0004148956 -0.0078057631 -0.011202761  
## loan -0.0033845191 -0.0095940876 0.0063201428 0.006597131  
## contact 0.2396282072 -0.0134637348 -0.0289540355 0.079889349  
## month 1.0000000000 0.0238800587 -0.0007225453 -0.067576434  
## day\_of\_week 0.0238800587 1.0000000000 0.0244267145 -0.039839163  
## duration -0.0007225453 0.0244267145 1.0000000000 -0.068406074  
## campaign -0.0675764342 -0.0398391635 -0.0684060738 1.000000000  
## pdays -0.0462976806 -0.0074554284 -0.0462390812 0.054311663  
## previous 0.1047979856 -0.0041700386 0.0187724193 -0.080765831  
## poutcome -0.0690681202 0.0166988296 0.0356265335 0.030776054  
## emp.var.rate -0.1831654829 0.0344915131 -0.0233741439 0.157738648  
## cons.price.idx -0.0198529220 0.0055277209 0.0127678824 0.127259878  
## cons.conf.idx -0.0076411154 0.0362720613 -0.0095249899 -0.011664110  
## euribor3m -0.1215647627 0.0415404513 -0.0289215828 0.140835836  
## nr.employed -0.2169045573 0.0326420328 -0.0403667686 0.148068542  
## pdays previous poutcome emp.var.rate  
## age -0.050891389 0.049230881 0.010737242 -5.040879e-02  
## job -0.020336417 0.012919851 0.010155575 -3.614608e-03  
## marital -0.028998417 0.028644700 0.002131790 -7.131739e-02  
## education -0.028499775 0.016187868 0.023100786 -7.273547e-03  
## default 0.002101948 0.002638772 -0.006697198 5.984644e-03  
## housing -0.007542215 0.021081495 -0.013033737 -5.916390e-02  
## loan 0.003056377 -0.003505535 -0.003766954 2.019645e-05  
## contact 0.115151990 -0.201759898 0.106625539 3.809218e-01  
## month -0.046297681 0.104797986 -0.069068120 -1.831655e-01  
## day\_of\_week -0.007455428 -0.004170039 0.016698830 3.449151e-02  
## duration -0.046239081 0.018772419 0.035626534 -2.337414e-02  
## campaign 0.054311663 -0.080765831 0.030776054 1.577386e-01  
## pdays 1.000000000 -0.590248079 -0.493699727 2.687628e-01  
## previous -0.590248079 1.000000000 -0.288296404 -4.035015e-01  
## poutcome -0.493699727 -0.288296404 1.000000000 1.682275e-01  
## emp.var.rate 0.268762788 -0.403501523 0.168227504 1.000000e+00  
## cons.price.idx 0.068009745 -0.176775212 0.195621794 7.660549e-01  
## cons.conf.idx -0.102367587 -0.027930483 0.165136716 1.575930e-01  
## euribor3m 0.295188353 -0.438863417 0.158384249 9.694121e-01  
## nr.employed 0.370844946 -0.488365300 0.093939342 9.003902e-01  
## cons.price.idx cons.conf.idx euribor3m nr.employed  
## age -0.035762179 0.125017394 -0.0364806837 -0.064585684  
## job -0.017513349 0.041908525 -0.0042691080 -0.012131894  
## marital -0.050203018 -0.037370218 -0.0814539280 -0.072673249  
## education -0.063571931 0.084108816 0.0001870845 -0.003025869  
## default -0.002722432 0.005252153 0.0068828104 0.007460203  
## housing -0.077432029 -0.036031995 -0.0585510647 -0.045069947  
## loan -0.002231583 -0.018878212 -0.0024623395 0.002446840  
## contact 0.564490709 0.225466431 0.3839638934 0.259555135  
## month -0.019852922 -0.007641115 -0.1215647627 -0.216904557  
## day\_of\_week 0.005527721 0.036272061 0.0415404513 0.032642033  
## duration 0.012767882 -0.009524990 -0.0289215828 -0.040366769  
## campaign 0.127259878 -0.011664110 0.1408358359 0.148068542  
## pdays 0.068009745 -0.102367587 0.2951883532 0.370844946  
## previous -0.176775212 -0.027930483 -0.4388634172 -0.488365300  
## poutcome 0.195621794 0.165136716 0.1583842488 0.093939342  
## emp.var.rate 0.766054915 0.157593021 0.9694121449 0.900390247  
## cons.price.idx 1.000000000 0.027217385 0.6672920675 0.488870874  
## cons.conf.idx 0.027217385 1.000000000 0.2436367707 0.075282707  
## euribor3m 0.667292067 0.243636771 1.0000000000 0.944871306  
## nr.employed 0.488870874 0.075282707 0.9448713058 1.000000000

print(findCorrelation(cor\_X,cutoff=0.9,names = T))

## [1] "euribor3m" "emp.var.rate"

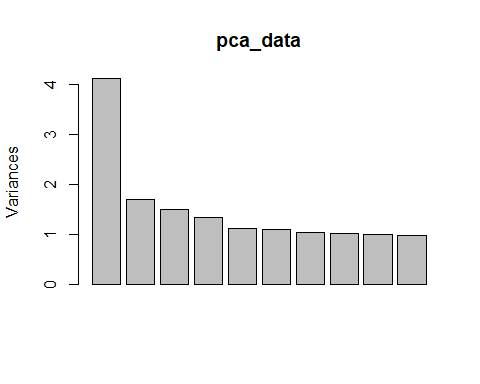
Based on similarity matrix we observe that field **emp.var.rate** and **euribor3m** are highly correlated and seemed redundant.

## 1.5 Dimensionality Reduction using PCA (Principle Component Analysis)

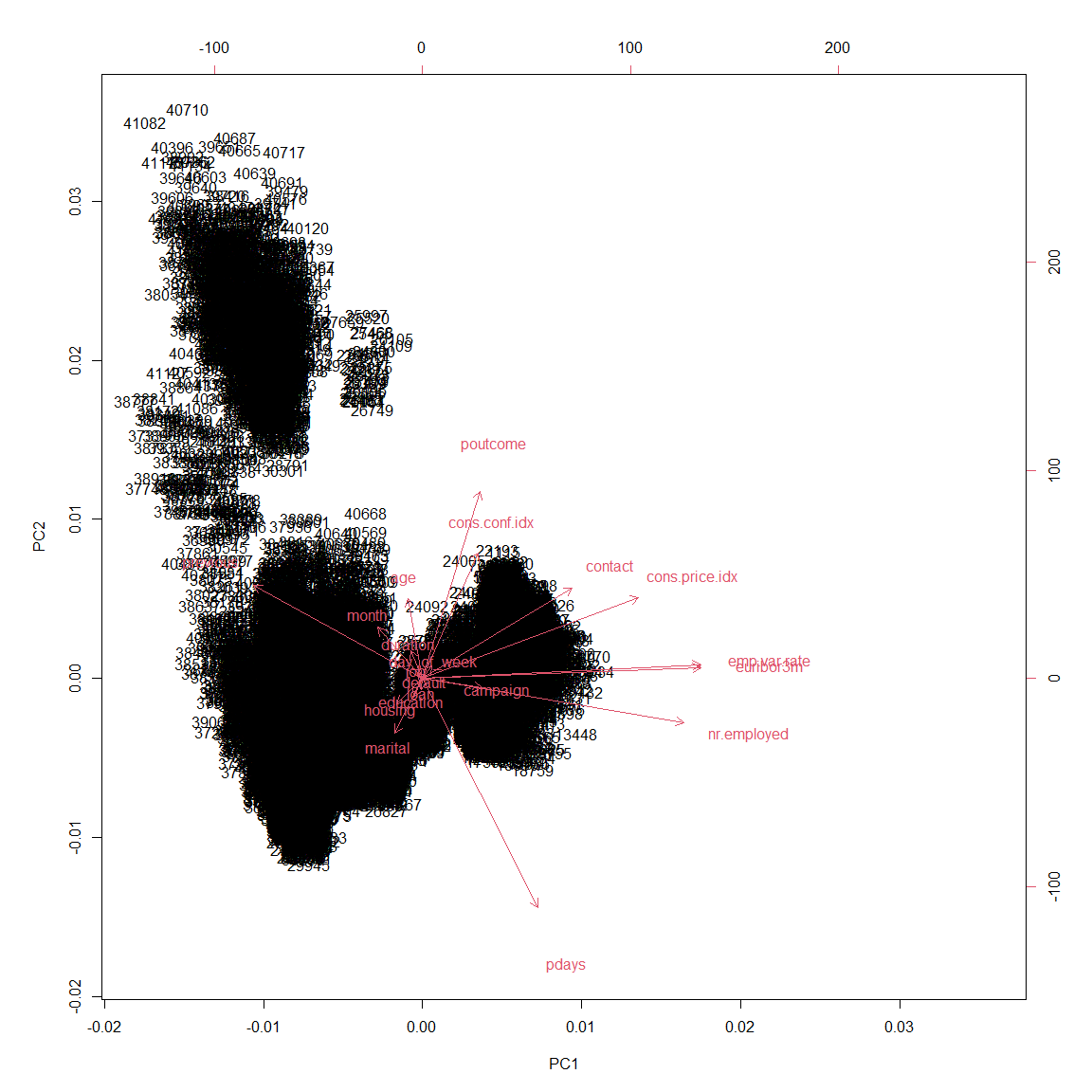
pca\_data<-prcomp(X,center = TRUE,scale. = TRUE)  
data\_pc<-pca\_data$x  
pca\_rotation<-pca\_data$rotation  
summary(pca\_data)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 2.0301 1.30293 1.22237 1.15338 1.05463 1.04333 1.01728  
## Proportion of Variance 0.2061 0.08488 0.07471 0.06651 0.05561 0.05443 0.05174  
## Cumulative Proportion 0.2061 0.29094 0.36565 0.43217 0.48778 0.54220 0.59395  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 1.00972 0.99726 0.9909 0.97426 0.95946 0.94455 0.89525  
## Proportion of Variance 0.05098 0.04973 0.0491 0.04746 0.04603 0.04461 0.04007  
## Cumulative Proportion 0.64492 0.69465 0.7438 0.79121 0.83724 0.88184 0.92192  
## PC15 PC16 PC17 PC18 PC19 PC20  
## Standard deviation 0.78252 0.75066 0.53583 0.26702 0.14233 0.08449  
## Proportion of Variance 0.03062 0.02817 0.01436 0.00356 0.00101 0.00036  
## Cumulative Proportion 0.95253 0.98071 0.99507 0.99863 0.99964 1.00000

plot(pca\_data)



biplot(pca\_data)



sort(abs(pca\_rotation[,1]))

## loan default job day\_of\_week education   
## 0.001610039 0.002041442 0.008390442 0.015272148 0.015510744   
## duration age housing marital month   
## 0.020236734 0.024407969 0.044024857 0.046850085 0.075486292   
## cons.conf.idx poutcome campaign pdays contact   
## 0.093771025 0.097331070 0.101106063 0.195276173 0.254072945   
## previous cons.price.idx nr.employed emp.var.rate euribor3m   
## 0.287018151 0.366310506 0.443237831 0.471316255 0.471461802

Based on the summary of pca components, we observe that at about 6 components determines 50% of result and about 15 components describes 95% of result (based on Cumulative Proportion). Further more as per the biplot we can clearly observe that our previous observation that field *emp.var.rate* and *euribor3m* are highly correlated. Based on PC1 values of PCA analysis *loan* and *default* are the least contributing attributes.

## 1.6 Sampling Dataset

set.seed(110) # Team 11  
dataSampleindex <- sample(1:nrow(data),10000)  
sampled\_data <- data[dataSampleindex, ]  
dim(sampled\_data)

## [1] 10000 21

sampled\_data$y <- as.factor(sampled\_data$y)

# 2. Classification

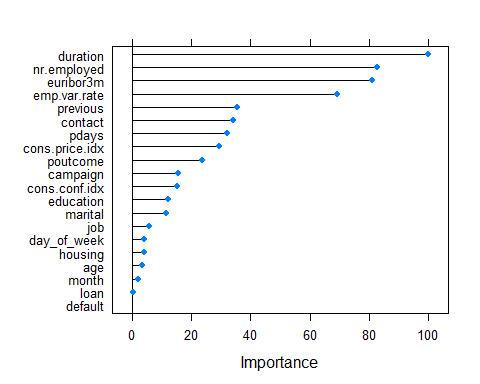
## 2.1 Splitting into training and test sets

### 2.1.1 Full sampled set benchmark 70:30

indxTrain70 <- createDataPartition(y = sampled\_data$y,p = 0.70,list = FALSE)  
training\_all\_col70 <- sampled\_data[indxTrain70,]  
testing\_all\_col70 <- sampled\_data[-indxTrain70,]  
x <- training\_all\_col70[, -21]  
y <- training\_all\_col70$y  
model <- train(x, y, 'nb', trControl=trainControl(method='cv', number=10))  
Predict <- predict(model, newdata = testing\_all\_col70)  
confusionMatrix(Predict, testing\_all\_col70$y,positive = "TRUE",mode = "prec\_recall")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction FALSE TRUE  
## FALSE 2423 201  
## TRUE 198 177  
##   
## Accuracy : 0.867   
## 95% CI : (0.8543, 0.8789)  
## No Information Rate : 0.874   
## P-Value [Acc > NIR] : 0.8811   
##   
## Kappa : 0.394   
##   
## Mcnemar's Test P-Value : 0.9202   
##   
## Precision : 0.47200   
## Recall : 0.46825   
## F1 : 0.47012   
## Prevalence : 0.12604   
## Detection Rate : 0.05902   
## Detection Prevalence : 0.12504   
## Balanced Accuracy : 0.69636   
##   
## 'Positive' Class : TRUE   
##

attributeImp <- varImp(model)  
plot(attributeImp)



### 2.1.2 Keep duration - drop pdays & loan & default - 70:30

sample\_data\_with\_dur\_3\_dropped\_70 <- sampled\_data  
sample\_data\_with\_dur\_3\_dropped\_70[, c('pdays', 'loan', 'default')] <- list(NULL)  
training\_with\_dur\_3dropped\_col70 <- sample\_data\_with\_dur\_3\_dropped\_70[indxTrain70,]  
testing\_with\_dur\_3dropped\_col70 <- sample\_data\_with\_dur\_3\_dropped\_70[-indxTrain70,]

### 2.1.3 Drop duration - drop pdays & loan & default - 70:30

sample\_data\_without\_dur\_3\_dropped\_70 <- sampled\_data  
sample\_data\_without\_dur\_3\_dropped\_70[, c('pdays', 'loan', 'default', 'duration')] <- list(NULL)  
training\_without\_dur\_3dropped\_col70 <- sample\_data\_without\_dur\_3\_dropped\_70[indxTrain70,]  
testing\_without\_dur\_3dropped\_col70 <- sample\_data\_without\_dur\_3\_dropped\_70[-indxTrain70,]

### 2.1.4 Keep duration - drop pdays & loan & default & emp.var.rate - 70:30

sample\_data\_with\_dur\_4\_dropped\_70 <- sampled\_data  
sample\_data\_with\_dur\_4\_dropped\_70[, c('pdays', 'loan', 'default', 'emp.var.rate')] <- list(NULL)  
training\_with\_dur\_4dropped\_col70 <- sample\_data\_with\_dur\_4\_dropped\_70[indxTrain70,]  
testing\_with\_dur\_4dropped\_col70 <- sample\_data\_with\_dur\_4\_dropped\_70[-indxTrain70,]

### 2.1.5 Drop duration - drop pdays & loan & default & emp.var.rate - 70:30

sample\_data\_without\_dur\_4\_dropped\_70 <- sampled\_data  
sample\_data\_without\_dur\_4\_dropped\_70[, c('pdays', 'loan', 'default', 'emp.var.rate', 'duration')] <- list(NULL)  
training\_without\_dur\_4dropped\_col70 <- sample\_data\_without\_dur\_4\_dropped\_70[indxTrain70,]  
testing\_without\_dur\_4dropped\_col70 <- sample\_data\_without\_dur\_4\_dropped\_70[-indxTrain70,]

### 2.1.6 Keep duration - drop pdays & loan & default - 50:50

indxTrain50 <- createDataPartition(y = sampled\_data$y,p = 0.50,list = FALSE)  
sample\_data\_with\_dur\_3\_dropped\_50 <- sampled\_data  
sample\_data\_with\_dur\_3\_dropped\_50[, c('pdays', 'loan', 'default')] <- list(NULL)  
training\_with\_dur\_3dropped\_col50 <- sample\_data\_with\_dur\_3\_dropped\_50[indxTrain50,]  
testing\_with\_dur\_3dropped\_col50 <- sample\_data\_with\_dur\_3\_dropped\_50[-indxTrain50,]

### 2.1.7 Drop duration - drop pdays & loan & default - 50:50

sample\_data\_without\_dur\_3\_dropped\_50 <- sampled\_data  
sample\_data\_without\_dur\_3\_dropped\_50[, c('pdays', 'loan', 'default', 'duration')] <- list(NULL)  
training\_without\_dur\_3dropped\_col50 <- sample\_data\_without\_dur\_3\_dropped\_50[indxTrain50,]  
testing\_without\_dur\_3dropped\_col50 <- sample\_data\_without\_dur\_3\_dropped\_50[-indxTrain50,]

### 2.1.8 Keep duration - drop pdays & loan & default & emp.var.rate - 50:50

sample\_data\_with\_dur\_4\_dropped\_50 <- sampled\_data  
sample\_data\_with\_dur\_4\_dropped\_50[, c('pdays', 'loan', 'default', 'emp.var.rate')] <- list(NULL)  
training\_with\_dur\_4dropped\_col50 <- sample\_data\_with\_dur\_4\_dropped\_50[indxTrain50,]  
testing\_with\_dur\_4dropped\_col50 <- sample\_data\_with\_dur\_4\_dropped\_50[-indxTrain50,]

### 2.1.9 Drop duration - drop pdays & loan & default & emp.var.rate - 50:50

sample\_data\_without\_dur\_4\_dropped\_50 <- sampled\_data  
sample\_data\_without\_dur\_4\_dropped\_50[, c('pdays', 'loan', 'default', 'emp.var.rate', 'duration')] <- list(NULL)  
training\_without\_dur\_4dropped\_col50 <- sample\_data\_without\_dur\_4\_dropped\_50[indxTrain50,]  
testing\_without\_dur\_4dropped\_col50 <- sample\_data\_without\_dur\_4\_dropped\_50[-indxTrain50,]

### 2.1.10 Keep duration - drop pdays & loan & default - 80:20

indxTrain80 <- createDataPartition(y = sampled\_data$y,p = 0.80,list = FALSE)  
sample\_data\_with\_dur\_3\_dropped\_80 <- sampled\_data  
sample\_data\_with\_dur\_3\_dropped\_80[, c('pdays', 'loan', 'default')] <- list(NULL)  
training\_with\_dur\_3dropped\_col80 <- sample\_data\_with\_dur\_3\_dropped\_80[indxTrain80,]  
testing\_with\_dur\_3dropped\_col80 <- sample\_data\_with\_dur\_3\_dropped\_80[-indxTrain80,]

### 2.1.11 Drop duration - drop pdays & loan & default - 80:20

sample\_data\_without\_dur\_3\_dropped\_80 <- sampled\_data  
sample\_data\_without\_dur\_3\_dropped\_80[, c('pdays', 'loan', 'default', 'duration')] <- list(NULL)  
training\_without\_dur\_3dropped\_col80 <- sample\_data\_without\_dur\_3\_dropped\_80[indxTrain80,]  
testing\_without\_dur\_3dropped\_col80 <- sample\_data\_without\_dur\_3\_dropped\_80[-indxTrain80,]

### 2.1.12 Keep duration - drop pdays & loan & default & emp.var.rate - 80:20

sample\_data\_with\_dur\_4\_dropped\_80 <- sampled\_data  
sample\_data\_with\_dur\_4\_dropped\_80[, c('pdays', 'loan', 'default', 'emp.var.rate')] <- list(NULL)  
training\_with\_dur\_4dropped\_col80 <- sample\_data\_with\_dur\_4\_dropped\_80[indxTrain80,]  
testing\_with\_dur\_4dropped\_col80 <- sample\_data\_with\_dur\_4\_dropped\_80[-indxTrain80,]

### 2.1.13 Drop duration - drop pdays & loan & default & emp.var.rate - 80:20

sample\_data\_without\_dur\_4\_dropped\_80 <- sampled\_data  
sample\_data\_without\_dur\_4\_dropped\_80[, c('pdays', 'loan', 'default', 'emp.var.rate', 'duration')] <- list(NULL)  
training\_without\_dur\_4dropped\_col80 <- sample\_data\_without\_dur\_4\_dropped\_80[indxTrain80,]  
testing\_without\_dur\_4dropped\_col80 <- sample\_data\_without\_dur\_4\_dropped\_80[-indxTrain80,]

## 2.2 Training models and making predictions for all splits

### 2.2.1 Classify - Keep duration - drop pdays & loan & default - 70:30

x <- training\_with\_dur\_3dropped\_col70[, -18]  
y <- training\_with\_dur\_3dropped\_col70$y  
model <- train(x, y, 'nb', trControl=trainControl(method='cv', number=10))  
Predict <- predict(model, newdata = testing\_with\_dur\_3dropped\_col70)  
confMatrix <- confusionMatrix(Predict, testing\_with\_dur\_3dropped\_col70$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.54232804 0.90614269 0.45454545   
## Neg Pred Value Precision Recall   
## 0.93210361 0.45454545 0.54232804   
## F1 Prevalence Detection Rate   
## 0.49457177 0.12604201 0.06835612   
## Detection Prevalence Balanced Accuracy   
## 0.15038346 0.72423537

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 2375 173  
## TRUE 246 205

### 2.2.2 Classify - Drop duration - drop pdays & loan & default - 70:30

x <- training\_without\_dur\_3dropped\_col70[, -17]  
y <- training\_without\_dur\_3dropped\_col70$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_without\_dur\_3dropped\_col70)  
confMatrix <- confusionMatrix(Predict, testing\_without\_dur\_3dropped\_col70$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.51851852 0.89622282 0.41880342   
## Neg Pred Value Precision Recall   
## 0.92809166 0.41880342 0.51851852   
## F1 Prevalence Detection Rate   
## 0.46335697 0.12604201 0.06535512   
## Detection Prevalence Balanced Accuracy   
## 0.15605202 0.70737067

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 2349 182  
## TRUE 272 196

### 2.2.3 Classify - Keep duration - drop pdays & loan & default & emp.var.rate - 70:30

x <- training\_with\_dur\_4dropped\_col70[, -17]  
y <- training\_with\_dur\_4dropped\_col70$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_with\_dur\_4dropped\_col70)  
confMatrix <- confusionMatrix(Predict, testing\_with\_dur\_4dropped\_col70$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.5052910 0.9164441 0.4658537   
## Neg Pred Value Precision Recall   
## 0.9277713 0.4658537 0.5052910   
## F1 Prevalence Detection Rate   
## 0.4847716 0.1260420 0.0636879   
## Detection Prevalence Balanced Accuracy   
## 0.1367122 0.7108676

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 2402 187  
## TRUE 219 191

### 2.2.4 Classify - Drop duration - drop pdays & loan & default & emp.var.rate - 70:30

x <- training\_without\_dur\_4dropped\_col70[, -16]  
y <- training\_without\_dur\_4dropped\_col70$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_without\_dur\_4dropped\_col70)  
confMatrix <- confusionMatrix(Predict, testing\_without\_dur\_4dropped\_col70$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.50000000 0.90537963 0.43249428   
## Neg Pred Value Precision Recall   
## 0.92622951 0.43249428 0.50000000   
## F1 Prevalence Detection Rate   
## 0.46380368 0.12604201 0.06302101   
## Detection Prevalence Balanced Accuracy   
## 0.14571524 0.70268981

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 2373 189  
## TRUE 248 189

### 2.2.5 Classify - Keep duration - drop pdays & loan & default - 50:50

x <- training\_with\_dur\_3dropped\_col50[, -18]  
y <- training\_with\_dur\_3dropped\_col50$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_with\_dur\_3dropped\_col50)  
confMatrix <- confusionMatrix(Predict, testing\_with\_dur\_3dropped\_col50$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.49047619 0.92835889 0.49678457   
## Neg Pred Value Precision Recall   
## 0.92666210 0.49678457 0.49047619   
## F1 Prevalence Detection Rate   
## 0.49361022 0.12602521 0.06181236   
## Detection Prevalence Balanced Accuracy   
## 0.12442488 0.70941754

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 4056 321  
## TRUE 313 309

### 2.2.6 Classify - Drop duration - drop pdays & loan & default - 50:50

x <- training\_without\_dur\_3dropped\_col50[, -17]  
y <- training\_without\_dur\_3dropped\_col50$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_without\_dur\_3dropped\_col50)  
confMatrix <- confusionMatrix(Predict, testing\_without\_dur\_3dropped\_col50$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.47460317 0.92217899 0.46791862   
## Neg Pred Value Precision Recall   
## 0.92408257 0.46791862 0.47460317   
## F1 Prevalence Detection Rate   
## 0.47123719 0.12602521 0.05981196   
## Detection Prevalence Balanced Accuracy   
## 0.12782557 0.69839108

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 4029 331  
## TRUE 340 299

### 2.2.7 Classify - Keep duration - drop pdays & loan & default & emp.var.rate - 50:50

x <- training\_with\_dur\_4dropped\_col50[, -17]  
y <- training\_with\_dur\_4dropped\_col50$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_with\_dur\_4dropped\_col50)  
confMatrix <- confusionMatrix(Predict, testing\_with\_dur\_4dropped\_col50$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.45396825 0.93980316 0.52094718   
## Neg Pred Value Precision Recall   
## 0.92269663 0.52094718 0.45396825   
## F1 Prevalence Detection Rate   
## 0.48515691 0.12602521 0.05721144   
## Detection Prevalence Balanced Accuracy   
## 0.10982196 0.69688571

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 4106 344  
## TRUE 263 286

### 2.2.8 Classify - Drop duration - drop pdays & loan & default & emp.var.rate - 50:50

x <- training\_without\_dur\_4dropped\_col50[, -16]  
y <- training\_without\_dur\_4dropped\_col50$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_without\_dur\_4dropped\_col50)  
confMatrix <- confusionMatrix(Predict, testing\_without\_dur\_4dropped\_col50$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.44761905 0.92813001 0.47315436   
## Neg Pred Value Precision Recall   
## 0.92096298 0.47315436 0.44761905   
## F1 Prevalence Detection Rate   
## 0.46003263 0.12602521 0.05641128   
## Detection Prevalence Balanced Accuracy   
## 0.11922384 0.68787453

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 4055 348  
## TRUE 314 282

### 2.2.9 Classify - Keep duration - drop pdays & loan & default - 80:20

x <- training\_with\_dur\_3dropped\_col80[, -18]  
y <- training\_with\_dur\_3dropped\_col80$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_with\_dur\_3dropped\_col80)  
confMatrix <- confusionMatrix(Predict, testing\_with\_dur\_3dropped\_col80$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.53968254 0.90955924 0.46258503   
## Neg Pred Value Precision Recall   
## 0.93196481 0.46258503 0.53968254   
## F1 Prevalence Detection Rate   
## 0.49816850 0.12606303 0.06803402   
## Detection Prevalence Balanced Accuracy   
## 0.14707354 0.72462089

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 1589 116  
## TRUE 158 136

### 2.2.10 Classify - Drop duration - drop pdays & loan & default - 80:20

x <- training\_without\_dur\_3dropped\_col80[, -17]  
y <- training\_without\_dur\_3dropped\_col80$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_without\_dur\_3dropped\_col80)  
confMatrix <- confusionMatrix(Predict, testing\_without\_dur\_3dropped\_col80$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.52777778 0.89868346 0.42903226   
## Neg Pred Value Precision Recall   
## 0.92954411 0.42903226 0.52777778   
## F1 Prevalence Detection Rate   
## 0.47330961 0.12606303 0.06653327   
## Detection Prevalence Balanced Accuracy   
## 0.15507754 0.71323062

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 1570 119  
## TRUE 177 133

### 2.2.11 Classify - Keep duration - drop pdays & loan & default & emp.var.rate - 80:20

x <- training\_with\_dur\_4dropped\_col80[, -17]  
y <- training\_with\_dur\_4dropped\_col80$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_with\_dur\_4dropped\_col80)  
confMatrix <- confusionMatrix(Predict, testing\_with\_dur\_4dropped\_col80$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.50793651 0.91528334 0.46376812   
## Neg Pred Value Precision Recall   
## 0.92803250 0.46376812 0.50793651   
## F1 Prevalence Detection Rate   
## 0.48484848 0.12606303 0.06403202   
## Detection Prevalence Balanced Accuracy   
## 0.13806903 0.71160993

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 1599 124  
## TRUE 148 128

### 2.2.12 Classify - Drop duration - drop pdays & loan & default & emp.var.rate - 80:20

x <- training\_without\_dur\_4dropped\_col80[, -16]  
y <- training\_without\_dur\_4dropped\_col80$y  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
Predict <- predict(model, newdata = testing\_without\_dur\_4dropped\_col80)  
confMatrix <- confusionMatrix(Predict, testing\_without\_dur\_4dropped\_col80$y,positive = "TRUE",mode = "prec\_recall")  
confMatrix$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.50396825 0.90726961 0.43944637   
## Neg Pred Value Precision Recall   
## 0.92690058 0.43944637 0.50396825   
## F1 Prevalence Detection Rate   
## 0.46950092 0.12606303 0.06353177   
## Detection Prevalence Balanced Accuracy   
## 0.14457229 0.70561893

confMatrix$table

## Reference  
## Prediction FALSE TRUE  
## FALSE 1585 125  
## TRUE 162 127