DA5030.P2.Parpattedar

Shruti Parpattedar

March 17, 2019

Note : Collaborated with Shantanu Bafna on Problem 2 # Problem 1

Question 1

Import the data and add headers to the data frame.

```
data <- read.csv("income.csv", header = FALSE, fileEncoding = "UTF-8-BOM")
names(data) <- c("Age", "WorkClass", "fnlwgt", "Education", "Edu-num", "Marital-status", "Occupation",
str(data)
```

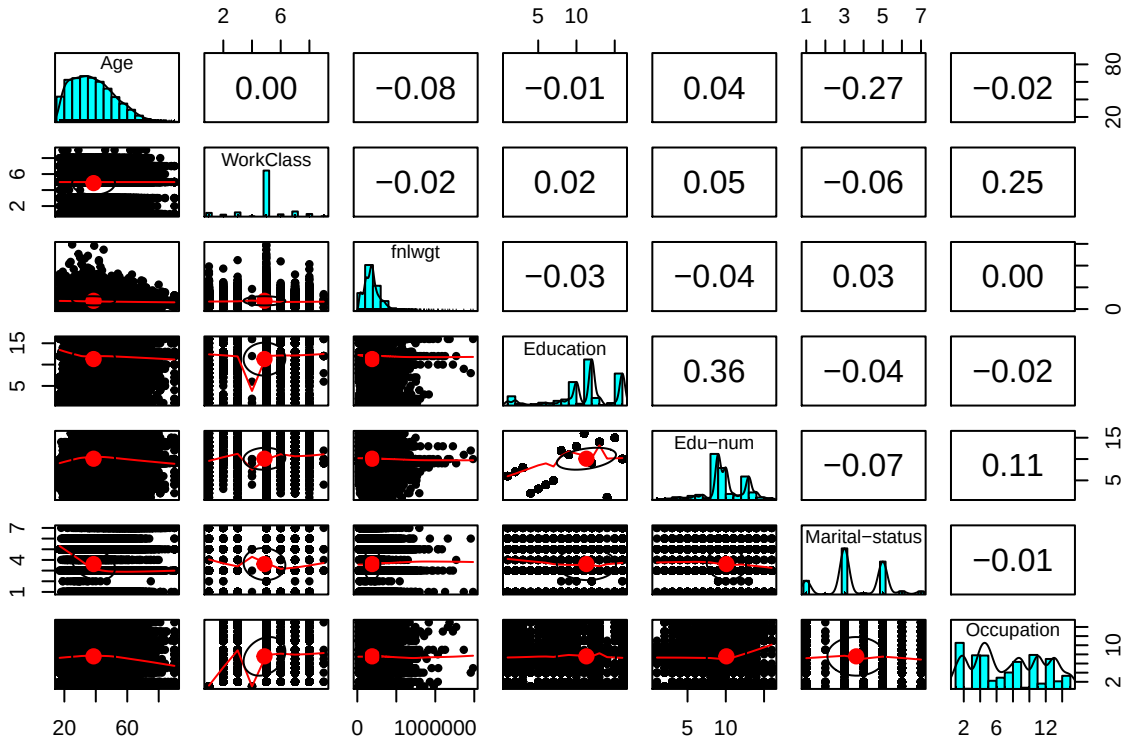
```
## 'data.frame':    32561 obs. of  15 variables:
##  $ Age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ WorkClass     : Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ Education     : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ Edu-num       : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ Marital-status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ Occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ Relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ Race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ Sex           : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ Cap-gain      : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ Cap-loss      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ HrsPerWeek    : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ NativeCountry : Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
##  $ Income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
```

Question 2

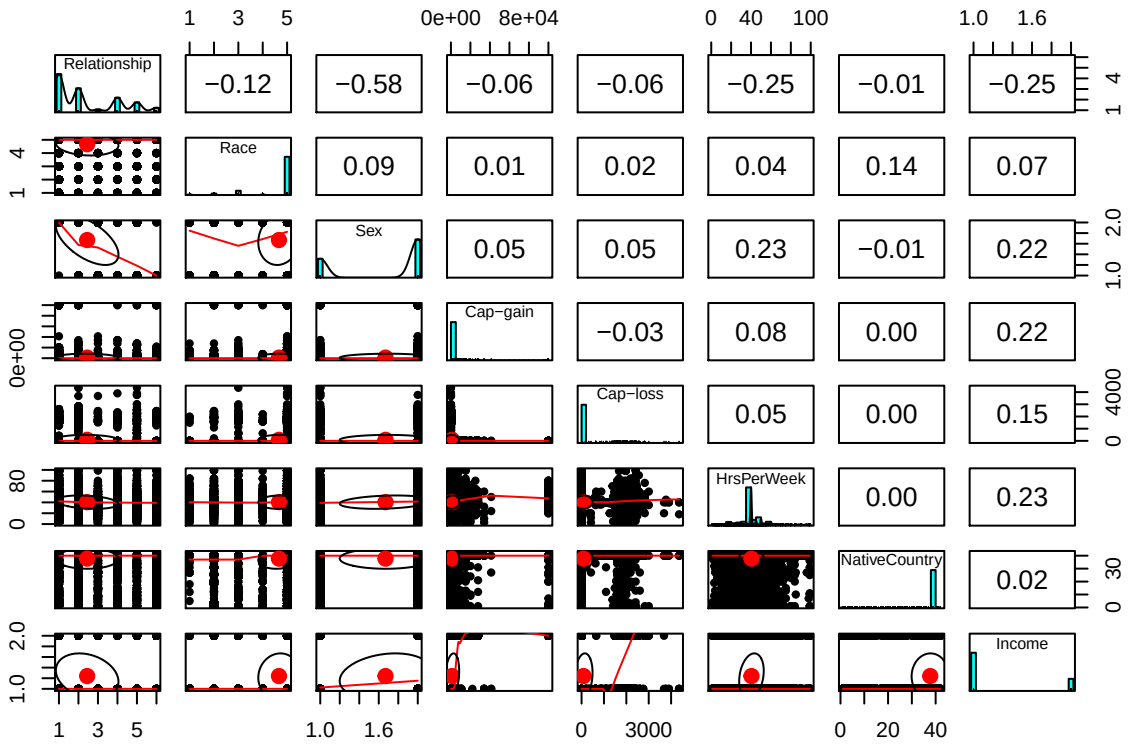
After exploring the data, I noticed that most of the variables are almost normally distributed.

ANother observation was that fnlwgt is right skewed, so we can apply the log transform to rectify the skew.

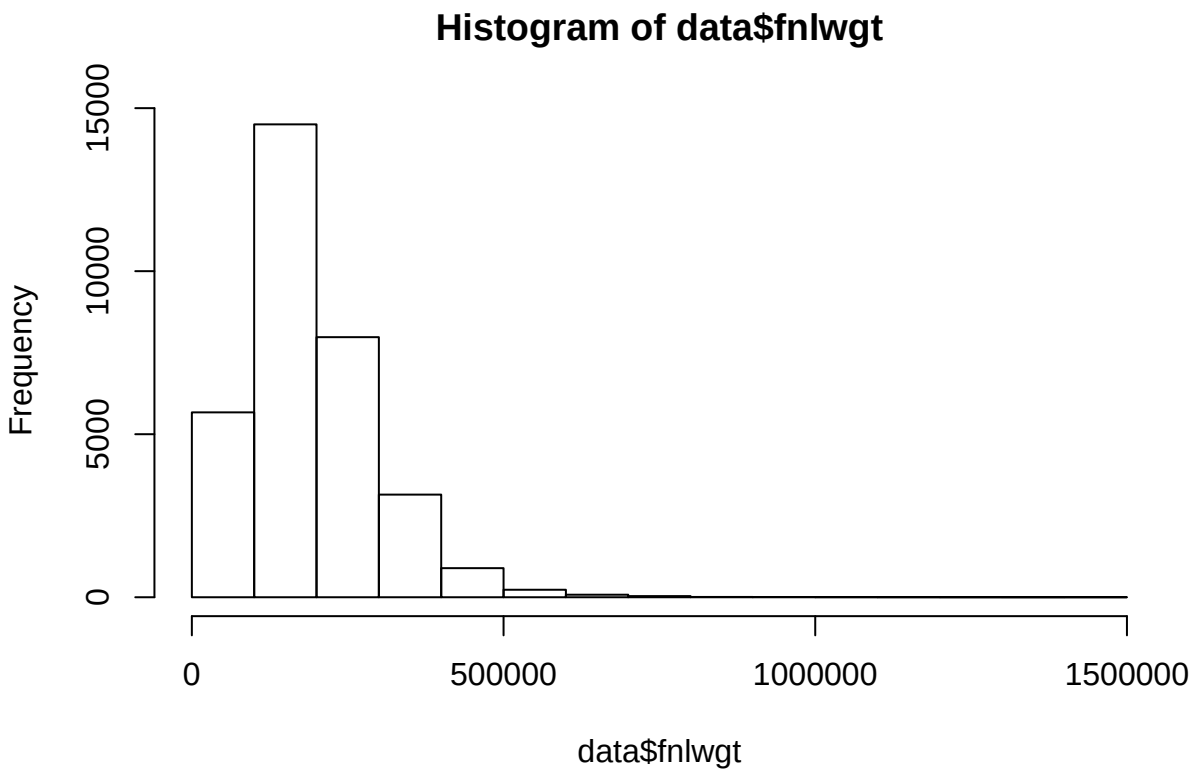
```
library(psych)
pairs.panels(data[,1:7])
```



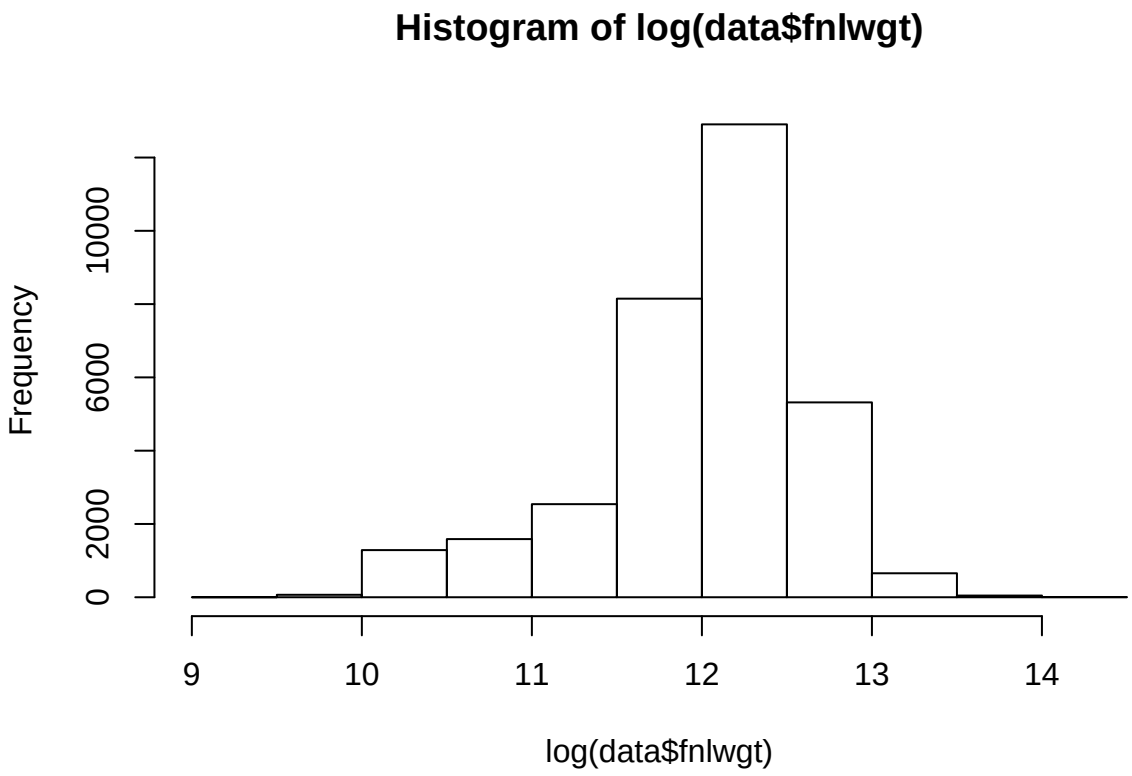

```
pairs.panels(data[,8:15])
```




```r
hist(data$fnlwgt)
```




```
hist(log(data$fnlwgt))
```



Question 3

Using the categorical variables to create the frequency and likelihood tables.

Next, implementing a Naive Bayes Classifer

```
library(gmodels)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Column numbers for the categorical variables
cat_variables <- c(2,4,6,7,8,9,10,14,15)

# Dataset with only categorical variables
cat_data <- data[, cat_variables]

createLikelihoodT <- function(dataset)
{
  n <- ncol(dataset) - 1 # excluding the Income column from the count
```

```r
  # Creating the initial frequency and likelihood tables using only the first categorical variable
  # with the income variable
  firstTable <- CrossTable(dataset$Income, dataset[, 1])
  freqTable <- firstTable$t
  likelihoodTable <- firstTable$prop.row

  # Cbinding the frequency and likelihood tables for the remaining categorical variables to the intial
  for(i in 2:n)
  {
    newTable <- CrossTable(dataset$Income, dataset[,i])
    freqTable <- cbind(freqTable, newTable$t)
    likelihoodTable <- cbind(likelihoodTable, newTable$prop.row)
  }
  return(likelihoodTable)
}

likelihoodTable_orig <- createLikelihoodT(cat_data)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  32561
##
##
##                | dataset[, 1]
## dataset$Income |                 ? |      Federal-gov |        Local-gov |     Never-worked |
## ---------------|-------------------|------------------|------------------|------------------|---------
##          <=50K |              1645 |              589 |             1476 |                7 |
##                |            45.244 |           26.825 |            8.034 |            0.535 |
##                |             0.067 |            0.024 |            0.060 |            0.000 |
##                |             0.896 |            0.614 |            0.705 |            1.000 |
##                |             0.051 |            0.018 |            0.045 |            0.000 |
## ---------------|-------------------|------------------|------------------|------------------|---------
##           >50K |               191 |              371 |              617 |                0 |
##                |           142.639 |           84.569 |           25.328 |            1.686 |
##                |             0.024 |            0.047 |            0.079 |            0.000 |
##                |             0.104 |            0.386 |            0.295 |            0.000 |
##                |             0.006 |            0.011 |            0.019 |            0.000 |
## ---------------|-------------------|------------------|------------------|------------------|---------
##   Column Total |              1836 |              960 |             2093 |                7 |
##                |             0.056 |            0.029 |            0.064 |            0.000 |
## ---------------|-------------------|------------------|------------------|------------------|---------
##
##
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  32561
##
##
##                | dataset[, i]
## dataset$Income |         10th |         11th |         12th |      1st-4th |      5th-6th |      7th-
## ---------------|--------------|--------------|--------------|--------------|--------------|----------
##          <=50K |          871 |         1115 |          400 |          162 |          317 |
##                |       37.360 |       55.723 |       15.452 |        9.308 |       16.298 |       27
##                |        0.035 |        0.045 |        0.016 |        0.007 |        0.013 |        0
##                |        0.934 |        0.949 |        0.924 |        0.964 |        0.952 |        0
##                |        0.027 |        0.034 |        0.012 |        0.005 |        0.010 |        0
## ---------------|--------------|--------------|--------------|--------------|--------------|----------
##           >50K |           62 |           60 |           33 |            6 |           16 |
##                |      117.784 |      175.674 |       48.715 |       29.346 |       51.382 |       85
##                |        0.008 |        0.008 |        0.004 |        0.001 |        0.002 |        0
##                |        0.066 |        0.051 |        0.076 |        0.036 |        0.048 |        0
##                |        0.002 |        0.002 |        0.001 |        0.000 |        0.000 |        0
## ---------------|--------------|--------------|--------------|--------------|--------------|----------
##   Column Total |          933 |         1175 |          433 |          168 |          333 |
##                |        0.029 |        0.036 |        0.013 |        0.005 |        0.010 |        0
## ---------------|--------------|--------------|--------------|--------------|--------------|----------
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  32561
##
##
##                | dataset[, i]
## dataset$Income |              Divorced |     Married-AF-spouse |    Married-civ-spouse | Married-spou
## ---------------|-----------------------|-----------------------|-----------------------|-------------
##          <=50K |                  3980 |                    13 |                  8284 |
```

```
##                |               109.202 |                 1.140 |               837.419 |
##                |                 0.161 |                 0.001 |                 0.335 |
##                |                 0.896 |                 0.565 |                 0.553 |
##                |                 0.122 |                 0.000 |                 0.254 |
## ---------------|-----------------------|-----------------------|-----------------------|------------
##           >50K |                   463 |                    10 |                  6692 |
##                |               344.277 |                 3.594 |              2640.097 |
##                |                 0.059 |                 0.001 |                 0.853 |
##                |                 0.104 |                 0.435 |                 0.447 |
##                |                 0.014 |                 0.000 |                 0.206 |
## ---------------|-----------------------|-----------------------|-----------------------|------------
##   Column Total |                  4443 |                    23 |                 14976 |
##                |                 0.136 |                 0.001 |                 0.460 |
## ---------------|-----------------------|-----------------------|-----------------------|------------
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  32561
##
##
##                | dataset[, i]
## dataset$Income |                 ? |      Adm-clerical |      Armed-Forces |      Craft-repair |   E
## ---------------|-------------------|-------------------|-------------------|-------------------|-----
##          <=50K |              1652 |              3263 |                 8 |              3170 |
##                |            45.679 |            56.140 |             0.199 |             1.084 |
##                |             0.067 |             0.132 |             0.000 |             0.128 |
##                |             0.896 |             0.866 |             0.889 |             0.773 |
##                |             0.051 |             0.100 |             0.000 |             0.097 |
## ---------------|-------------------|-------------------|-------------------|-------------------|-----
##           >50K |               191 |               507 |                 1 |               929 |
##                |           144.011 |           176.992 |             0.629 |             3.417 |
##                |             0.024 |             0.065 |             0.000 |             0.118 |
##                |             0.104 |             0.134 |             0.111 |             0.227 |
##                |             0.006 |             0.016 |             0.000 |             0.029 |
## ---------------|-------------------|-------------------|-------------------|-------------------|-----
##   Column Total |              1843 |              3770 |                 9 |              4099 |
##                |             0.057 |             0.116 |             0.000 |             0.126 |
## ---------------|-------------------|-------------------|-------------------|-------------------|-----
##
##
##
##
##    Cell Contents
```

```
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  32561
##
##
##                 | dataset[, i]
## dataset$Income |        Husband |  Not-in-family | Other-relative |      Own-child |      Unmarried
## ---------------|----------------|----------------|----------------|----------------|----------------
##          <=50K |           7275 |           7449 |            944 |           5001 |           3228
##                |        750.108 |        207.541 |         53.298 |        345.772 |        143.085
##                |          0.294 |          0.301 |          0.038 |          0.202 |          0.131
##                |          0.551 |          0.897 |          0.962 |          0.987 |          0.937
##                |          0.223 |          0.229 |          0.029 |          0.154 |          0.099
## ---------------|----------------|----------------|----------------|----------------|----------------
##           >50K |           5918 |            856 |             37 |             67 |            218
##                |       2364.834 |        654.305 |        168.029 |       1090.101 |        451.099
##                |          0.755 |          0.109 |          0.005 |          0.009 |          0.028
##                |          0.449 |          0.103 |          0.038 |          0.013 |          0.063
##                |          0.182 |          0.026 |          0.001 |          0.002 |          0.007
## ---------------|----------------|----------------|----------------|----------------|----------------
##   Column Total |          13193 |           8305 |            981 |           5068 |           3446
##                |          0.405 |          0.255 |          0.030 |          0.156 |          0.106
## ---------------|----------------|----------------|----------------|----------------|----------------
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  32561
##
##
##                 | dataset[, i]
## dataset$Income | Amer-Indian-Eskimo | Asian-Pac-Islander |              Black |              Other |
## ---------------|--------------------|--------------------|--------------------|--------------------|-
##          <=50K |                275 |                763 |               2737 |                246 |
##                |              6.406 |              0.844 |             56.262 |              7.878 |
##                |              0.011 |              0.031 |              0.111 |              0.010 |
##                |              0.884 |              0.734 |              0.876 |              0.908 |
```

```
##                |              0.008 |              0.023 |              0.084 |              0.008 |
## ---------------|--------------------|--------------------|--------------------|--------------------|-
##           >50K |                 36 |                276 |                387 |                 25 |
##                |             20.197 |              2.660 |            177.373 |             24.837 |
##                |              0.005 |              0.035 |              0.049 |              0.003 |
##                |              0.116 |              0.266 |              0.124 |              0.092 |
##                |              0.001 |              0.008 |              0.012 |              0.001 |
## ---------------|--------------------|--------------------|--------------------|--------------------|-
##   Column Total |                311 |               1039 |               3124 |                271 |
##                |              0.010 |              0.032 |              0.096 |              0.008 |
## ---------------|--------------------|--------------------|--------------------|--------------------|-
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  32561
##
##
##                | dataset[, i]
## dataset$Income |    Female |      Male | Row Total |
## ---------------|-----------|-----------|-----------|
##          <=50K |      9592 |     15128 |     24720 |
##                |   244.770 |   120.992 |           |
##                |     0.388 |     0.612 |     0.759 |
##                |     0.891 |     0.694 |           |
##                |     0.295 |     0.465 |           |
## ---------------|-----------|-----------|-----------|
##           >50K |      1179 |      6662 |      7841 |
##                |   771.677 |   381.447 |           |
##                |     0.150 |     0.850 |     0.241 |
##                |     0.109 |     0.306 |           |
##                |     0.036 |     0.205 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |     10771 |     21790 |     32561 |
##                |     0.331 |     0.669 |           |
## ---------------|-----------|-----------|-----------|
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
```

```
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  32561
##
##
##                | dataset[, i]
## dataset$Income |                       ? |                 Cambodia |                   Canada
## ---------------|-------------------------|-------------------------|-------------------------
##          <=50K |                     437 |                      12 |                      81
##                |                   0.071 |                   0.408 |                   1.059
##                |                   0.018 |                   0.000 |                   0.003
##                |                   0.750 |                   0.632 |                   0.678
##                |                   0.013 |                   0.000 |                   0.003
## ---------------|-------------------------|-------------------------|-------------------------
##           >50K |                     146 |                       7 |                      39
##                |                   0.224 |                   1.285 |                   3.338
##                |                   0.019 |                   0.001 |                   0.005
##                |                   0.250 |                   0.368 |                   0.321
##                |                   0.004 |                   0.000 |                   0.001
## ---------------|-------------------------|-------------------------|-------------------------
##   Column Total |                     583 |                      19 |                     121
##                |                   0.018 |                   0.001 |                   0.004
## ---------------|-------------------------|-------------------------|-------------------------
##
##
```

```r
# Implementing the Naive Bayes classifier that returns the probability of the income being less than 50
naiveBayes <- function(values, LH_Table)
{
  # P(<=50k)*product of P(<=50k and parameters)
  P_le50k <- table(data$Income)[1]*prod(LH_Table[1, values])

  # P(>50k)*product of P(>50k and parameters)
  P_gt50k <- table(data$Income)[2]*prod(LH_Table[2, values])

  result <- P_le50k / (P_le50k + P_gt50k)

  # Returning probability of income being <=50k using the values calculated above
  return(ifelse(result < 0.5, ">50k", "<=50k"))
}
```

Question 4

Using the classifier to predict the outcome for the new data

```r
vals <- c("White", "Female", "Federal-gov", "Bachelors", "India")
output <- naiveBayes(vals, likelihoodTable_orig)
paste("Result is", output)
```

```
## [1] "Result is >50k"
```

Question 5

10-fold cross validation on the classifier

Problem 2

Question 1

Finding and eliminating outliers

```
data_Q2 <- read.csv("uffidata.csv", stringsAsFactors = FALSE)
data_Q2 <- data_Q2[-100,]
data_Q2$Sale.Price <- as.numeric(gsub(",", "",data_Q2$Sale.Price))
str(data_Q2)
```

```
## 'data.frame':    99 obs. of  12 variables:
##  $ Observation   : int  37 79 75 32 69 4 28 30 18 63 ...
##  $ Year.Sold     : int  2009 2009 2011 2011 2010 2011 2010 2011 2010 2011 ...
##  $ Sale.Price    : num  76900 78000 79000 80000 82000 84000 84000 84000 85000 85000 ...
##  $ UFFI.IN       : int  1 1 0 0 1 1 0 0 0 1 ...
##  $ Brick.Ext     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ X45.Yrs.      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Bsmnt.Fin_SF  : int  0 154 400 0 157 398 0 0 0 0 ...
##  $ Lot.Area      : int  2772 4490 5840 5040 5441 4800 3300 5313 4125 2897 ...
##  $ Enc.Pk.Spaces : int  0 0 0 0 0 1 2 0 0 1 ...
##  $ Living.Area_SF: num  1018 536 721 512 672 ...
##  $ Central.Air   : int  0 1 1 0 0 0 0 0 1 0 ...
##  $ Pool          : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
bsmnt_mean   <- mean(data_Q2$Bsmnt.Fin_SF)
bsmnt_sd     <- sd(data_Q2$Bsmnt.Fin_SF)
bsmnt_zscore <- (bsmnt_mean - data_Q2$Bsmnt.Fin_SF) / bsmnt_sd

o <- which(bsmnt_zscore > 1)
data2 <- data_Q2[-o,]

Lot.Area_mean   <- mean(data2$Lot.Area)
Lot.Area_sd     <- sd(data2$Lot.Area)
Lot.Area_zscore <- (Lot.Area_mean - data2$Lot.Area) / Lot.Area_sd
o <- which(Lot.Area_zscore > 1)
data3 <- data2[-o,]

final_data <- data3
```

Question 2

Full correlation matrix

```
cor(final_data)
```

```
##                Observation    Year.Sold  Sale.Price     UFFI.IN
## Observation     1.000000000  0.005570781  0.12128684  0.01267699
## Year.Sold       0.005570781  1.000000000  0.61526298 -0.26950465
## Sale.Price      0.121286838  0.615262982  1.00000000 -0.21234278
## UFFI.IN         0.012676990 -0.269504652 -0.21234278  1.00000000
## Brick.Ext       0.431290609  0.341621849  0.28325307 -0.15190020
```

```
## X45.Yrs.       -0.705361838 -0.159994384 -0.22695205  0.02408006
## Bsmnt.Fin_SF    0.280411928  0.137878277  0.05090485  0.06000121
## Lot.Area        0.173395311  0.278858020  0.29979675  0.14415703
## Enc.Pk.Spaces  -0.014970809  0.243395146  0.44926531 -0.16878624
## Living.Area_SF  0.075920031  0.485110445  0.70499664 -0.08994836
## Central.Air     0.050684948 -0.057293946  0.16355806  0.06045008
## Pool            0.109092043  0.141406393  0.52216520 -0.11695218
##                   Brick.Ext    X45.Yrs. Bsmnt.Fin_SF    Lot.Area
## Observation     0.431290609 -0.70536184  0.280411928  0.17339531
## Year.Sold       0.341621849 -0.15999438  0.137878277  0.27885802
## Sale.Price      0.283253068 -0.22695205  0.050904855  0.29979675
## UFFI.IN        -0.151900199  0.02408006  0.060001209  0.14415703
## Brick.Ext       1.000000000 -0.42382302  0.174791333  0.11664286
## X45.Yrs.       -0.423823021  1.00000000 -0.479031374 -0.39920944
## Bsmnt.Fin_SF    0.174791333 -0.47903137  1.000000000  0.27241373
## Lot.Area        0.116642863 -0.39920944  0.272413730  1.00000000
## Enc.Pk.Spaces   0.002529482 -0.03667698 -0.039352840  0.22648183
## Living.Area_SF  0.096593967 -0.07869559  0.044501368  0.22493875
## Central.Air     0.119241192 -0.07297615  0.111425714  0.25574070
## Pool            0.014199962 -0.20589660  0.005395474  0.09909482
##                Enc.Pk.Spaces Living.Area_SF Central.Air         Pool
## Observation     -0.014970809     0.07592003  0.05068495  0.109092043
## Year.Sold        0.243395146     0.48511045 -0.05729395  0.141406393
## Sale.Price       0.449265312     0.70499664  0.16355806  0.522165201
## UFFI.IN         -0.168786244    -0.08994836  0.06045008 -0.116952176
## Brick.Ext        0.002529482     0.09659397  0.11924119  0.014199962
## X45.Yrs.        -0.036676985    -0.07869559 -0.07297615 -0.205896602
## Bsmnt.Fin_SF    -0.039352840     0.04450137  0.11142571  0.005395474
## Lot.Area         0.226481829     0.22493875  0.25574070  0.099094823
## Enc.Pk.Spaces    1.000000000     0.39504286  0.09696348  0.157280151
## Living.Area_SF   0.395042863     1.00000000  0.15955145  0.283149070
## Central.Air      0.096963484     0.15955145  1.00000000 -0.014199962
## Pool             0.157280151     0.28314907 -0.01419996  1.000000000
```

Question 3

Multiple regression model for Sales Price

```
model_Q2 <- lm(Sale.Price ~ ., final_data)
summary(model_Q2)
```

```
##
## Call:
## lm(formula = Sale.Price ~ ., data = final_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57584  -8352  -1810   8798  92814
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.093e+07  3.565e+06  -3.066 0.003595 **
## Observation    -2.896e+01  1.488e+02  -0.195 0.846499
## Year.Sold       5.467e+03  1.775e+03   3.081 0.003445 **
## UFFI.IN        -1.947e+03  8.200e+03  -0.237 0.813313
```

```
## Brick.Ext       1.098e+04  8.188e+03   1.341 0.186296
## X45.Yrs.       -4.387e+03  1.170e+04  -0.375 0.709432
## Bsmnt.Fin_SF   -1.334e+01  1.926e+01  -0.693 0.491967
## Lot.Area        7.744e-01  2.369e+00   0.327 0.745196
## Enc.Pk.Spaces   8.698e+03  4.582e+03   1.898 0.063812 .
## Living.Area_SF  4.874e+01  1.236e+01   3.944 0.000266 ***
## Central.Air     8.725e+03  7.126e+03   1.224 0.226879
## Pool            6.464e+04  1.483e+04   4.360 7.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23050 on 47 degrees of freedom
## Multiple R-squared:  0.7589, Adjusted R-squared:  0.7024
## F-statistic: 13.45 on 11 and 47 DF,  p-value: 4.314e-11
```

```
model2_Q2 <- lm(Sale.Price ~ Year.Sold + Enc.Pk.Spaces + Living.Area_SF + Pool, final_data)
summary(model2_Q2)
```

```
##
## Call:
## lm(formula = Sale.Price ~ Year.Sold + Enc.Pk.Spaces + Living.Area_SF +
##     Pool, data = final_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -70199 -11003    879   7343 102538
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.239e+07  2.933e+06  -4.225 9.26e-05 ***
## Year.Sold       6.195e+03  1.460e+03   4.244 8.68e-05 ***
## Enc.Pk.Spaces   9.222e+03  4.382e+03   2.105     0.04 *
## Living.Area_SF  5.018e+01  1.170e+01   4.290 7.44e-05 ***
## Pool            6.506e+04  1.411e+04   4.612 2.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22800 on 54 degrees of freedom
## Multiple R-squared:  0.729,  Adjusted R-squared:  0.7089
## F-statistic: 36.31 on 4 and 54 DF,  p-value: 1.014e-14
```

Question 4

UFFI and change in the value of a property

```
UFFI.glm <- glm(formula = UFFI.IN ~ Sale.Price,
                data = final_data,
                family = binomial)

newdata <- data.frame(Sale.Price = 128304.2)

predict(UFFI.glm, newdata, type = "response")
```

```
##         1
## 0.1687216
```

Question 5

Predicting Sales price for the given data along with its 95% CI.

```
model1 <- lm(Sale.Price ~ X45.Yrs. + Bsmnt.Fin_SF + Brick.Ext + Lot.Area + Enc.Pk.Spaces + Living.Area_S

sm1 <- summary(model1)

X45      <- 1
Basement <- 0
Brick    <- 0
LotArea  <- 7800
Parking  <- 1
Living   <- 1720
UFFI     <- 1
Central  <- 0
Pool     <- 0
salePrice <- sm1$coefficients[[1]] + sm1$coefficients[[2]]*X45 + sm1$coefficients[[3]]*Basement + sm1$co

paste("Predicted Value =", salePrice)
```

```
## [1] "Predicted Value = 165748.485603114"
```

```
paste("95% CI =", salePrice - 1.96*22230, salePrice + 1.96*22230)
```

```
## [1] "95% CI = 122177.685603114 209319.285603114"
```

Problem 3

Question 1

Importing data

```
data_Q3 <- read.csv("bank-full.csv", sep = ";")

validation_Q3 <- read.csv("bank.csv", sep = ";")
```

Question 2

Creating dummy codes.

Next, implementing a logistic regression model.

```
data_Q3$job1 <- ifelse(data_Q3$job == "admin.", 1, 0)
data_Q3$job2 <- ifelse(data_Q3$job == "unknown", 1, 0)
data_Q3$job3 <- ifelse(data_Q3$job == "unemployed", 1, 0)
data_Q3$job4 <- ifelse(data_Q3$job == "management", 1, 0)
data_Q3$job5 <- ifelse(data_Q3$job == "housemaid", 1, 0)
data_Q3$job6 <- ifelse(data_Q3$job == "entrepreneur", 1, 0)
data_Q3$job7 <- ifelse(data_Q3$job == "student", 1, 0)
data_Q3$job8 <- ifelse(data_Q3$job == "blue-collar", 1, 0)
data_Q3$job9 <- ifelse(data_Q3$job == "self-employed", 1, 0)
data_Q3$job10 <- ifelse(data_Q3$job == "retired", 1, 0)
data_Q3$job11 <- ifelse(data_Q3$job == "technician", 1, 0)

data_Q3$marital1 <- ifelse(data_Q3$marital == "married", 1, 0)
data_Q3$marital2 <- ifelse(data_Q3$marital == "divorced", 1, 0)
```

```r
data_Q3$edu1 <- ifelse(data_Q3$education == "secondary", 1, 0)
data_Q3$edu2 <- ifelse(data_Q3$education == "primary", 1, 0)
data_Q3$edu3 <- ifelse(data_Q3$education == "tertiary", 1, 0)

data_Q3$default10 <- ifelse(data_Q3$default == "yes", 1, 0)

data_Q3$housing10 <- ifelse(data_Q3$housing == "yes", 1, 0)

data_Q3$loan10 <- ifelse(data_Q3$loan == "yes", 1, 0)

data_Q3$contact1 <- ifelse(data_Q3$contact == "telephone", 1, 0)
data_Q3$contact2 <- ifelse(data_Q3$contact == "cellular", 1, 0)

data_Q3$month1 <- ifelse(data_Q3$month == "jan", 1, 0)
data_Q3$month2 <- ifelse(data_Q3$month == "feb", 2, 0)
data_Q3$month3 <- ifelse(data_Q3$month == "mar", 3, 0)
data_Q3$month4 <- ifelse(data_Q3$month == "apr", 4, 0)
data_Q3$month5 <- ifelse(data_Q3$month == "may", 5, 0)
data_Q3$month6 <- ifelse(data_Q3$month == "jun", 6, 0)
data_Q3$month7 <- ifelse(data_Q3$month == "jul", 7, 0)
data_Q3$month8 <- ifelse(data_Q3$month == "aug", 8, 0)
data_Q3$month9 <- ifelse(data_Q3$month == "sep", 9, 0)
data_Q3$month10 <- ifelse(data_Q3$month == "oct", 10, 0)
data_Q3$month11 <- ifelse(data_Q3$month == "nov", 11, 0)

data_Q3$poutcome1 <- ifelse(data_Q3$poutcome == "unknown", 1, 0)
data_Q3$poutcome2 <- ifelse(data_Q3$poutcome == "other", 1, 0)
data_Q3$poutcome3 <- ifelse(data_Q3$poutcome == "failure", 1, 0)

data_Q3$y10 <- ifelse(data_Q3$y == "yes", 1, 0)

model_logreg <- glm(y10 ~ age +
                      job1 + job2 + job3 + job4 + job5 +
                      job6 + job7 + job8 + job9 + job10 + job11 +
                      marital1 + marital2 +
                      edu1 + edu2 + edu3 +
                      default10 + housing10 + loan10 + balance +
                      contact1 + contact2 + day +
                      month1 + month2 + month3 + month4 + month5 + month6 +
                      month7 + month8 + month9 + month10 + month11 + duration +
                      campaign + pdays + previous +
                      poutcome1 + poutcome2 + poutcome3, data = data_Q3)

model2 <- step(model_logreg, direction = "backward")
```

```
## Start:  AIC=9482.46
## y10 ~ age + job1 + job2 + job3 + job4 + job5 + job6 + job7 +
##     job8 + job9 + job10 + job11 + marital1 + marital2 + edu1 +
##     edu2 + edu3 + default10 + housing10 + loan10 + balance +
##     contact1 + contact2 + day + month1 + month2 + month3 + month4 +
##     month5 + month6 + month7 + month8 + month9 + month10 + month11 +
##     duration + campaign + pdays + previous + poutcome1 + poutcome2 +
##     poutcome3
```

```
##
##             Df Deviance     AIC
## - default10  1   3258.5  9480.5
## - job11      1   3258.5  9480.7
## - job4       1   3258.5  9480.7
## - edu1       1   3258.5  9481.0
## - job3       1   3258.5  9481.1
## - job2       1   3258.5  9481.1
## - job9       1   3258.5  9481.4
## - job8       1   3258.5  9481.5
## - month10    1   3258.5  9481.5
## - age        1   3258.6  9481.8
## - month9     1   3258.6  9482.1
## - job6       1   3258.6  9482.2
## <none>           3258.5  9482.5
## - edu2       1   3258.6  9483.1
## - previous   1   3258.7  9483.1
## - edu3       1   3258.7  9483.6
## - marital2   1   3258.7  9483.9
## - job5       1   3258.8  9485.4
## - job1       1   3258.8  9485.7
## - pdays      1   3258.9  9485.9
## - balance    1   3259.1  9488.9
## - campaign   1   3259.4  9493.3
## - job10      1   3260.4  9507.9
## - day        1   3260.7  9511.4
## - month6     1   3261.4  9520.8
## - marital1   1   3261.4  9520.9
## - month3     1   3261.4  9521.9
## - loan10     1   3261.7  9525.8
## - month4     1   3262.1  9530.6
## - month2     1   3263.0  9543.4
## - job7       1   3263.4  9548.4
## - month5     1   3264.2  9560.5
## - month8     1   3267.3  9602.9
## - month7     1   3267.4  9603.9
## - month11    1   3267.5  9605.7
## - contact1   1   3268.3  9617.4
## - month1     1   3269.6  9635.4
## - housing10  1   3276.8  9734.0
## - contact2   1   3288.1  9890.3
## - poutcome2  1   3378.0 11109.6
## - poutcome1  1   3428.6 11781.4
## - poutcome3  1   3437.7 11901.3
## - duration   1   3913.3 17760.2
##
## Step:  AIC=9480.47
## y10 ~ age + job1 + job2 + job3 + job4 + job5 + job6 + job7 +
##     job8 + job9 + job10 + job11 + marital1 + marital2 + edu1 +
##     edu2 + edu3 + housing10 + loan10 + balance + contact1 + contact2 +
##     day + month1 + month2 + month3 + month4 + month5 + month6 +
##     month7 + month8 + month9 + month10 + month11 + duration +
##     campaign + pdays + previous + poutcome1 + poutcome2 + poutcome3
##
```

```
##             Df Deviance     AIC
## - job11      1   3258.5  9478.7
## - job4       1   3258.5  9478.7
## - edu1       1   3258.5  9479.0
## - job3       1   3258.5  9479.1
## - job2       1   3258.5  9479.1
## - job9       1   3258.5  9479.4
## - month10    1   3258.5  9479.5
## - job8       1   3258.5  9479.5
## - age        1   3258.6  9479.8
## - month9     1   3258.6  9480.2
## - job6       1   3258.6  9480.2
## <none>           3258.5  9480.5
## - edu2       1   3258.6  9481.1
## - previous   1   3258.7  9481.1
## - edu3       1   3258.7  9481.6
## - marital2   1   3258.7  9481.9
## - job5       1   3258.8  9483.4
## - job1       1   3258.8  9483.7
## - pdays      1   3258.9  9483.9
## - balance    1   3259.1  9487.0
## - campaign   1   3259.4  9491.3
## - job10      1   3260.4  9505.9
## - day        1   3260.7  9509.4
## - month6     1   3261.4  9518.8
## - marital1   1   3261.4  9518.9
## - month3     1   3261.4  9519.9
## - loan10     1   3261.8  9524.1
## - month4     1   3262.1  9528.6
## - month2     1   3263.0  9541.4
## - job7       1   3263.4  9546.4
## - month5     1   3264.2  9558.5
## - month8     1   3267.3  9600.9
## - month7     1   3267.4  9601.9
## - month11    1   3267.5  9603.8
## - contact1   1   3268.3  9615.5
## - month1     1   3269.7  9633.5
## - housing10  1   3276.8  9732.0
## - contact2   1   3288.1  9888.4
## - poutcome2  1   3378.0 11107.6
## - poutcome1  1   3428.6 11780.0
## - poutcome3  1   3437.7 11899.3
## - duration   1   3913.4 17758.7
##
## Step:  AIC=9478.67
## y10 ~ age + job1 + job2 + job3 + job4 + job5 + job6 + job7 +
##     job8 + job9 + job10 + marital1 + marital2 + edu1 + edu2 +
##     edu3 + housing10 + loan10 + balance + contact1 + contact2 +
##     day + month1 + month2 + month3 + month4 + month5 + month6 +
##     month7 + month8 + month9 + month10 + month11 + duration +
##     campaign + pdays + previous + poutcome1 + poutcome2 + poutcome3
##
##             Df Deviance     AIC
## - job4       1   3258.5  9476.7
```

```
## - job3       1   3258.5  9477.1
## - edu1       1   3258.5  9477.2
## - job2       1   3258.5  9477.5
## - month10    1   3258.5  9477.7
## - age        1   3258.6  9478.0
## - month9     1   3258.6  9478.4
## - job9       1   3258.6  9478.4
## <none>           3258.5  9478.7
## - edu2       1   3258.7  9479.3
## - previous   1   3258.7  9479.4
## - job8       1   3258.7  9479.4
## - job6       1   3258.7  9479.4
## - edu3       1   3258.7  9479.9
## - marital2   1   3258.7  9480.2
## - pdays      1   3258.9  9482.1
## - job1       1   3258.9  9483.0
## - job5       1   3259.0  9483.4
## - balance    1   3259.1  9485.2
## - campaign   1   3259.4  9489.5
## - day        1   3260.7  9507.7
## - job10      1   3260.8  9508.4
## - month6     1   3261.4  9517.1
## - marital1   1   3261.4  9517.3
## - month3     1   3261.5  9518.1
## - loan10     1   3261.8  9522.3
## - month4     1   3262.1  9526.9
## - month2     1   3263.0  9539.6
## - job7       1   3263.8  9550.8
## - month5     1   3264.3  9556.7
## - month8     1   3267.3  9599.0
## - month7     1   3267.4  9600.2
## - month11    1   3267.5  9601.9
## - contact1   1   3268.4  9613.6
## - month1     1   3269.7  9631.7
## - housing10  1   3276.8  9730.3
## - contact2   1   3288.2  9886.6
## - poutcome2  1   3378.0 11105.9
## - poutcome1  1   3428.7 11778.8
## - poutcome3  1   3437.7 11897.4
## - duration   1   3913.4 17756.7
##
## Step:  AIC=9476.75
## y10 ~ age + job1 + job2 + job3 + job5 + job6 + job7 + job8 +
##     job9 + job10 + marital1 + marital2 + edu1 + edu2 + edu3 +
##     housing10 + loan10 + balance + contact1 + contact2 + day +
##     month1 + month2 + month3 + month4 + month5 + month6 + month7 +
##     month8 + month9 + month10 + month11 + duration + campaign +
##     pdays + previous + poutcome1 + poutcome2 + poutcome3
##
##             Df Deviance     AIC
## - job3       1   3258.5  9475.1
## - edu1       1   3258.5  9475.3
## - job2       1   3258.5  9475.7
## - month10    1   3258.6  9475.8
```

```
## - age        1   3258.6  9476.1
## - month9     1   3258.6  9476.4
## <none>           3258.5  9476.7
## - job9       1   3258.6  9476.9
## - edu2       1   3258.7  9477.4
## - previous   1   3258.7  9477.4
## - job6       1   3258.7  9478.0
## - job8       1   3258.7  9478.0
## - marital2   1   3258.7  9478.2
## - edu3       1   3258.7  9478.4
## - pdays      1   3258.9  9480.2
## - job1       1   3258.9  9481.2
## - job5       1   3259.0  9482.0
## - balance    1   3259.1  9483.3
## - campaign   1   3259.4  9487.6
## - day        1   3260.7  9505.8
## - job10      1   3260.8  9507.6
## - month6     1   3261.4  9515.2
## - marital1   1   3261.4  9515.3
## - month3     1   3261.5  9516.2
## - loan10     1   3261.8  9520.5
## - month4     1   3262.1  9525.0
## - month2     1   3263.0  9537.7
## - job7       1   3263.9  9549.8
## - month5     1   3264.3  9554.8
## - month8     1   3267.3  9597.1
## - month7     1   3267.4  9598.3
## - month11    1   3267.5  9600.0
## - contact1   1   3268.4  9611.7
## - month1     1   3269.7  9629.8
## - housing10  1   3276.8  9728.6
## - contact2   1   3288.2  9884.8
## - poutcome2  1   3378.0 11103.9
## - poutcome1  1   3428.7 11776.8
## - poutcome3  1   3437.7 11895.4
## - duration   1   3913.4 17754.7
##
## Step:  AIC=9475.14
## y10 ~ age + job1 + job2 + job5 + job6 + job7 + job8 + job9 +
##     job10 + marital1 + marital2 + edu1 + edu2 + edu3 + housing10 +
##     loan10 + balance + contact1 + contact2 + day + month1 + month2 +
##     month3 + month4 + month5 + month6 + month7 + month8 + month9 +
##     month10 + month11 + duration + campaign + pdays + previous +
##     poutcome1 + poutcome2 + poutcome3
##
##             Df Deviance     AIC
## - edu1       1   3258.5  9473.7
## - job2       1   3258.6  9474.1
## - month10    1   3258.6  9474.2
## - age        1   3258.6  9474.5
## - month9     1   3258.6  9474.8
## <none>           3258.5  9475.1
## - job9       1   3258.7  9475.4
## - edu2       1   3258.7  9475.7
```

```
## - previous   1   3258.7  9475.8
## - job6       1   3258.8  9476.5
## - marital2   1   3258.8  9476.6
## - edu3       1   3258.8  9476.7
## - job8       1   3258.8  9476.9
## - pdays      1   3258.9  9478.6
## - job1       1   3258.9  9479.2
## - job5       1   3259.1  9480.7
## - balance    1   3259.1  9481.7
## - campaign   1   3259.4  9486.0
## - day        1   3260.7  9504.2
## - job10      1   3260.8  9505.6
## - month6     1   3261.4  9513.6
## - marital1   1   3261.4  9513.7
## - month3     1   3261.5  9514.6
## - loan10     1   3261.8  9519.3
## - month4     1   3262.1  9523.4
## - month2     1   3263.0  9535.9
## - job7       1   3263.9  9547.8
## - month5     1   3264.3  9553.3
## - month8     1   3267.4  9595.7
## - month7     1   3267.4  9596.7
## - month11    1   3267.5  9598.4
## - contact1   1   3268.4  9610.2
## - month1     1   3269.7  9628.0
## - housing10  1   3277.0  9728.5
## - contact2   1   3288.2  9883.3
## - poutcome2  1   3378.2 11103.9
## - poutcome1  1   3428.7 11775.1
## - poutcome3  1   3437.9 11895.7
## - duration   1   3913.8 17757.9
##
## Step:  AIC=9473.65
## y10 ~ age + job1 + job2 + job5 + job6 + job7 + job8 + job9 +
##     job10 + marital1 + marital2 + edu2 + edu3 + housing10 + loan10 +
##     balance + contact1 + contact2 + day + month1 + month2 + month3 +
##     month4 + month5 + month6 + month7 + month8 + month9 + month10 +
##     month11 + duration + campaign + pdays + previous + poutcome1 +
##     poutcome2 + poutcome3
##
##             Df Deviance     AIC
## - job2       1   3258.6  9472.4
## - month10    1   3258.6  9472.7
## - age        1   3258.7  9473.2
## - month9     1   3258.7  9473.4
## <none>           3258.5  9473.7
## - job9       1   3258.7  9473.9
## - previous   1   3258.7  9474.3
## - edu2       1   3258.8  9475.0
## - job6       1   3258.8  9475.0
## - marital2   1   3258.8  9475.2
## - job8       1   3258.8  9475.3
## - pdays      1   3258.9  9477.1
## - job1       1   3259.0  9477.7
```

```
## - job5       1   3259.1  9479.2
## - balance    1   3259.2  9480.2
## - campaign   1   3259.5  9484.5
## - edu3       1   3260.8  9502.6
## - day        1   3260.8  9502.7
## - job10      1   3260.9  9504.0
## - month6     1   3261.5  9512.1
## - marital1   1   3261.5  9512.3
## - month3     1   3261.5  9513.1
## - loan10     1   3261.9  9518.4
## - month4     1   3262.2  9521.9
## - month2     1   3263.1  9534.5
## - job7       1   3264.1  9548.3
## - month5     1   3264.3  9551.8
## - month8     1   3267.4  9594.4
## - month7     1   3267.5  9595.2
## - month11    1   3267.6  9597.0
## - contact1   1   3268.4  9608.6
## - month1     1   3269.7  9626.5
## - housing10  1   3277.1  9727.9
## - contact2   1   3288.2  9881.4
## - poutcome2  1   3378.3 11102.8
## - poutcome1  1   3428.8 11774.0
## - poutcome3  1   3437.9 11894.4
## - duration   1   3913.9 17756.4
##
## Step:  AIC=9472.42
## y10 ~ age + job1 + job5 + job6 + job7 + job8 + job9 + job10 +
##     marital1 + marital2 + edu2 + edu3 + housing10 + loan10 +
##     balance + contact1 + contact2 + day + month1 + month2 + month3 +
##     month4 + month5 + month6 + month7 + month8 + month9 + month10 +
##     month11 + duration + campaign + pdays + previous + poutcome1 +
##     poutcome2 + poutcome3
##
##             Df Deviance     AIC
## - month10    1   3258.7  9471.4
## - age        1   3258.7  9471.9
## - month9     1   3258.7  9472.1
## <none>           3258.6  9472.4
## - job9       1   3258.8  9472.6
## - previous   1   3258.8  9473.1
## - job6       1   3258.8  9473.6
## - edu2       1   3258.8  9473.8
## - job8       1   3258.9  9473.9
## - marital2   1   3258.9  9473.9
## - pdays      1   3259.0  9475.9
## - job1       1   3259.1  9476.7
## - job5       1   3259.1  9477.7
## - balance    1   3259.2  9479.0
## - campaign   1   3259.5  9483.3
## - day        1   3260.9  9501.6
## - edu3       1   3260.9  9502.1
## - job10      1   3261.0  9503.6
## - month6     1   3261.5  9511.0
```

```
## - marital1   1   3261.5  9511.0
## - month3     1   3261.6  9511.8
## - loan10     1   3261.9  9516.8
## - month4     1   3262.2  9520.7
## - month2     1   3263.1  9533.2
## - job7       1   3264.2  9547.6
## - month5     1   3264.4  9550.6
## - month8     1   3267.5  9593.2
## - month7     1   3267.5  9594.1
## - month11    1   3267.6  9595.8
## - contact1   1   3268.5  9607.2
## - month1     1   3269.8  9625.5
## - housing10  1   3277.1  9725.9
## - contact2   1   3288.3  9880.7
## - poutcome2  1   3378.3 11101.3
## - poutcome1  1   3428.9 11773.0
## - poutcome3  1   3438.0 11893.5
## - duration   1   3914.0 17755.5
##
## Step:  AIC=9471.41
## y10 ~ age + job1 + job5 + job6 + job7 + job8 + job9 + job10 +
##     marital1 + marital2 + edu2 + edu3 + housing10 + loan10 +
##     balance + contact1 + contact2 + day + month1 + month2 + month3 +
##     month4 + month5 + month6 + month7 + month8 + month9 + month11 +
##     duration + campaign + pdays + previous + poutcome1 + poutcome2 +
##     poutcome3
##
##             Df Deviance     AIC
## - month9     1   3258.7  9470.1
## - age        1   3258.8  9470.9
## <none>           3258.7  9471.4
## - job9       1   3258.8  9471.6
## - previous   1   3258.9  9472.1
## - job6       1   3258.9  9472.6
## - edu2       1   3258.9  9472.7
## - job8       1   3258.9  9472.8
## - marital2   1   3258.9  9472.9
## - pdays      1   3259.1  9474.9
## - job1       1   3259.1  9475.7
## - job5       1   3259.2  9476.7
## - balance    1   3259.3  9478.0
## - campaign   1   3259.6  9482.5
## - day        1   3260.9  9501.0
## - edu3       1   3261.0  9501.2
## - job10      1   3261.1  9502.6
## - marital1   1   3261.6  9510.1
## - loan10     1   3262.0  9515.8
## - month3     1   3263.7  9539.4
## - job7       1   3264.2  9546.4
## - contact1   1   3268.5  9606.2
## - month6     1   3271.9  9651.9
## - month4     1   3274.6  9690.0
## - housing10  1   3277.1  9724.7
## - month2     1   3277.5  9729.3
```

```
## - month5     1   3285.6  9840.9
## - contact2   1   3288.4  9879.7
## - month1     1   3296.6  9992.3
## - month11    1   3297.3 10002.5
## - month8     1   3298.0 10012.1
## - month7     1   3298.5 10018.4
## - poutcome2  1   3378.3 11099.7
## - poutcome1  1   3428.9 11771.1
## - poutcome3  1   3438.0 11891.5
## - duration   1   3914.0 17753.5
##
## Step:  AIC=9470.1
## y10 ~ age + job1 + job5 + job6 + job7 + job8 + job9 + job10 +
##     marital1 + marital2 + edu2 + edu3 + housing10 + loan10 +
##     balance + contact1 + contact2 + day + month1 + month2 + month3 +
##     month4 + month5 + month6 + month7 + month8 + month11 + duration +
##     campaign + pdays + previous + poutcome1 + poutcome2 + poutcome3
##
##             Df Deviance     AIC
## - age        1   3258.8  9469.6
## <none>           3258.7  9470.1
## - job9       1   3258.9  9470.3
## - previous   1   3258.9  9470.8
## - job6       1   3259.0  9471.3
## - edu2       1   3259.0  9471.4
## - job8       1   3259.0  9471.6
## - marital2   1   3259.0  9471.6
## - pdays      1   3259.1  9473.5
## - job1       1   3259.2  9474.4
## - job5       1   3259.3  9475.4
## - balance    1   3259.3  9476.6
## - campaign   1   3259.7  9481.1
## - day        1   3261.0  9499.2
## - edu3       1   3261.0  9499.9
## - job10      1   3261.1  9501.2
## - marital1   1   3261.7  9508.8
## - loan10     1   3262.1  9514.5
## - month3     1   3264.1  9542.6
## - job7       1   3264.3  9545.3
## - contact1   1   3268.6  9604.4
## - housing10  1   3277.2  9723.6
## - month6     1   3278.4  9740.8
## - month4     1   3281.9  9788.2
## - month2     1   3285.9  9844.2
## - contact2   1   3288.4  9877.8
## - month5     1   3300.1 10038.4
## - month1     1   3307.9 10144.6
## - month11    1   3314.9 10240.6
## - month8     1   3317.9 10281.5
## - month7     1   3318.2 10285.4
## - poutcome2  1   3378.5 11099.6
## - poutcome1  1   3428.9 11770.0
## - poutcome3  1   3438.3 11892.7
## - duration   1   3914.0 17751.6
```

```
##
## Step:  AIC=9469.58
## y10 ~ job1 + job5 + job6 + job7 + job8 + job9 + job10 + marital1 +
##     marital2 + edu2 + edu3 + housing10 + loan10 + balance + contact1 +
##     contact2 + day + month1 + month2 + month3 + month4 + month5 +
##     month6 + month7 + month8 + month11 + duration + campaign +
##     pdays + previous + poutcome1 + poutcome2 + poutcome3
##
##             Df Deviance     AIC
## <none>           3258.8  9469.6
## - job9       1   3259.0  9469.7
## - marital2   1   3259.0  9470.0
## - previous   1   3259.0  9470.3
## - edu2       1   3259.0  9470.5
## - job6       1   3259.1  9470.7
## - job8       1   3259.1  9471.3
## - pdays      1   3259.2  9473.0
## - job1       1   3259.3  9473.8
## - job5       1   3259.3  9474.6
## - balance    1   3259.5  9476.7
## - campaign   1   3259.8  9480.5
## - day        1   3261.1  9498.5
## - edu3       1   3261.1  9498.9
## - marital1   1   3261.8  9508.1
## - job10      1   3262.1  9513.1
## - loan10     1   3262.2  9514.4
## - month3     1   3264.2  9542.0
## - job7       1   3264.3  9543.4
## - contact1   1   3268.9  9607.3
## - housing10  1   3277.9  9731.3
## - month6     1   3278.6  9741.3
## - month4     1   3282.1  9789.0
## - month2     1   3286.1  9844.9
## - contact2   1   3288.4  9876.0
## - month5     1   3300.4 10041.2
## - month1     1   3308.1 10145.5
## - month11    1   3315.0 10240.6
## - month8     1   3318.0 10281.2
## - month7     1   3318.5 10287.7
## - poutcome2  1   3378.7 11100.9
## - poutcome1  1   3429.4 11773.6
## - poutcome3  1   3438.4 11892.7
## - duration   1   3914.0 17750.0
```

Question 3

Above model as an equation

```
model2$call[[2]]
```

```
## y10 ~ job1 + job5 + job6 + job7 + job8 + job9 + job10 + marital1 +
##     marital2 + edu2 + edu3 + housing10 + loan10 + balance + contact1 +
##     contact2 + day + month1 + month2 + month3 + month4 + month5 +
##     month6 + month7 + month8 + month11 + duration + campaign +
##     pdays + previous + poutcome1 + poutcome2 + poutcome3
```

Question 4

Prediction accuracy

```
validation_Q3$job1 <- ifelse(validation_Q3$job == "admin.", 1, 0)
validation_Q3$job2 <- ifelse(validation_Q3$job == "unknown", 1, 0)
validation_Q3$job3 <- ifelse(validation_Q3$job == "unemployed", 1, 0)
validation_Q3$job4 <- ifelse(validation_Q3$job == "management", 1, 0)
validation_Q3$job5 <- ifelse(validation_Q3$job == "housemaid", 1, 0)
validation_Q3$job6 <- ifelse(validation_Q3$job == "entrepreneur", 1, 0)
validation_Q3$job7 <- ifelse(validation_Q3$job == "student", 1, 0)
validation_Q3$job8 <- ifelse(validation_Q3$job == "blue-collar", 1, 0)
validation_Q3$job9 <- ifelse(validation_Q3$job == "self-employed", 1, 0)
validation_Q3$job10 <- ifelse(validation_Q3$job == "retired", 1, 0)
validation_Q3$job11 <- ifelse(validation_Q3$job == "technician", 1, 0)

validation_Q3$marital1 <- ifelse(validation_Q3$marital == "married", 1, 0)
validation_Q3$marital2 <- ifelse(validation_Q3$marital == "divorced", 1, 0)

validation_Q3$edu1 <- ifelse(validation_Q3$education == "secondary", 1, 0)
validation_Q3$edu2 <- ifelse(validation_Q3$education == "primary", 1, 0)
validation_Q3$edu3 <- ifelse(validation_Q3$education == "tertiary", 1, 0)

validation_Q3$default10 <- ifelse(validation_Q3$default == "yes", 1, 0)

validation_Q3$housing10 <- ifelse(validation_Q3$housing == "yes", 1, 0)

validation_Q3$loan10 <- ifelse(validation_Q3$loan == "yes", 1, 0)

validation_Q3$contact1 <- ifelse(validation_Q3$contact == "telephone", 1, 0)
validation_Q3$contact2 <- ifelse(validation_Q3$contact == "cellular", 1, 0)

validation_Q3$month1 <- ifelse(validation_Q3$month == "jan", 1, 0)
validation_Q3$month2 <- ifelse(validation_Q3$month == "feb", 2, 0)
validation_Q3$month3 <- ifelse(validation_Q3$month == "mar", 3, 0)
validation_Q3$month4 <- ifelse(validation_Q3$month == "apr", 4, 0)
validation_Q3$month5 <- ifelse(validation_Q3$month == "may", 5, 0)
validation_Q3$month6 <- ifelse(validation_Q3$month == "jun", 6, 0)
validation_Q3$month7 <- ifelse(validation_Q3$month == "jul", 7, 0)
validation_Q3$month8 <- ifelse(validation_Q3$month == "aug", 8, 0)
validation_Q3$month9 <- ifelse(validation_Q3$month == "sep", 9, 0)
validation_Q3$month10 <- ifelse(validation_Q3$month == "oct", 10, 0)
validation_Q3$month11 <- ifelse(validation_Q3$month == "nov", 11, 0)

validation_Q3$poutcome1 <- ifelse(validation_Q3$poutcome == "unknown", 1, 0)
validation_Q3$poutcome2 <- ifelse(validation_Q3$poutcome == "other", 1, 0)
validation_Q3$poutcome3 <- ifelse(validation_Q3$poutcome == "failure", 1, 0)

validation_Q3$y10 <- ifelse(validation_Q3$y == "yes", 1, 0)
final_logreg <- glm(model2$call[[2]], data = data_Q3)
pred <- predict(final_logreg, validation_Q3)
pred <- ifelse(pred < 0.5, 0, 1)

acc <- sum(pred == validation_Q3$y10) / length(pred) * 100
paste0("Prediction accuracy = ", acc, "%")
```

```
## [1] "Prediction accuracy = 89.758902897589%"
```