SESSION 11 – ASSIGNMENT 11.1

Date: 11th Feb 2019

Use the link given below and locate the bank marketing dataset.
https://archive.ics.uci.edu/ml/machine-learning-databases/00222/

```
# Load the Data
library(readr)
bankdata <- read_delim("G:/DATA ANALYTICS/DATA/bank-additional/bank-additional/bankdata.csv", ";", escape_double = FALSE, trim_ws = TRUE)
str(bankdata)

if(length(which(is.na(bankdata)==TRUE)>0)){print("Missing Value found in the specified column") } else print("All okay: No Missing Value found in the specified column")

summary(bankdata)
dim(bankdata)
```

```
Console ~/
> library(readr)
> bankdata <- read_delim("G:/DATA ANALYTICS/DATA/bank-additional/bank-additional/bankdata.csv",
+                        ";", escape_double = FALSE, trim_ws = TRUE)
Parsed with column specification:
cols(
  .default = col_character(),
  age = col_double(),
  duration = col_double(),
  campaign = col_double(),
  pdays = col_double(),
  previous = col_double(),
  emp.var.rate = col_double(),
  cons.price.idx = col_double(),
  cons.conf.idx = col_double(),
  euribor3m = col_double(),
  nr.employed = col_double()
)
See spec(...) for full column specifications.
```

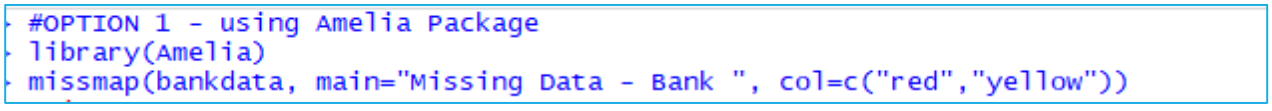
```
See spec(...) for full column specifications.
> str(bankdata)
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':        41188 obs. of  21 variables:
 $ age            : num  56 57 37 40 56 45 59 41 24 25 ...
 $ job            : chr  "housemaid" "services" "services" "admin." ...
 $ marital        : chr  "married" "married" "married" "married" ...
 $ education      : chr  "basic.4y" "high.school" "high.school" "basic.6y" ...
 $ default        : chr  "no" NA "no" "no" ...
 $ housing        : chr  "no" "no" "yes" "no" ...
 $ loan           : chr  "no" "no" "no" "no" ...
 $ contact        : chr  "telephone" "telephone" "telephone" "telephone" ...
 $ month          : chr  "may" "may" "may" "may" ...
 $ day_of_week    : chr  "mon" "mon" "mon" "mon" ...
 $ duration       : num  261 149 226 151 307 198 139 217 380 50 ...
 $ campaign       : num  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays          : num  999 999 999 999 999 999 999 999 999 999 ...
 $ previous       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome       : chr  "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...
 $ emp.var.rate   : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
 $ cons.price.idx: num  94 94 94 94 94 ...
 $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
 $ euribor3m      : num  4.86 4.86 4.86 4.86 4.86 ...
 $ nr.employed    : num  5191 5191 5191 5191 5191 ...
 $ y              : chr  "no" "no" "no" "no" ...
 - attr(*, "spec")=
  .. cols(
  ..   age = col_double(),
  ..   job = col_character(),
  ..   marital = col_character(),
  ..   education = col_character(),
  ..   default = col_character(),
  ..   housing = col_character(),
  ..   loan = col_character(),
  ..   contact = col_character(),
  ..   month = col_character(),
  ..   day_of_week = col_character(),
  ..   duration = col_double(),
  ..   campaign = col_double(),
  ..   pdays = col_double(),
  ..   previous = col_double(),
  ..   poutcome = col_character(),
  ..   emp.var.rate = col_double(),
  ..   cons.price.idx = col_double(),
  ..   cons.conf.idx = col_double(),
  ..   euribor3m = col_double(),
  ..   nr.employed = col_double(),
  ..   y = col_character()
  .. )
```

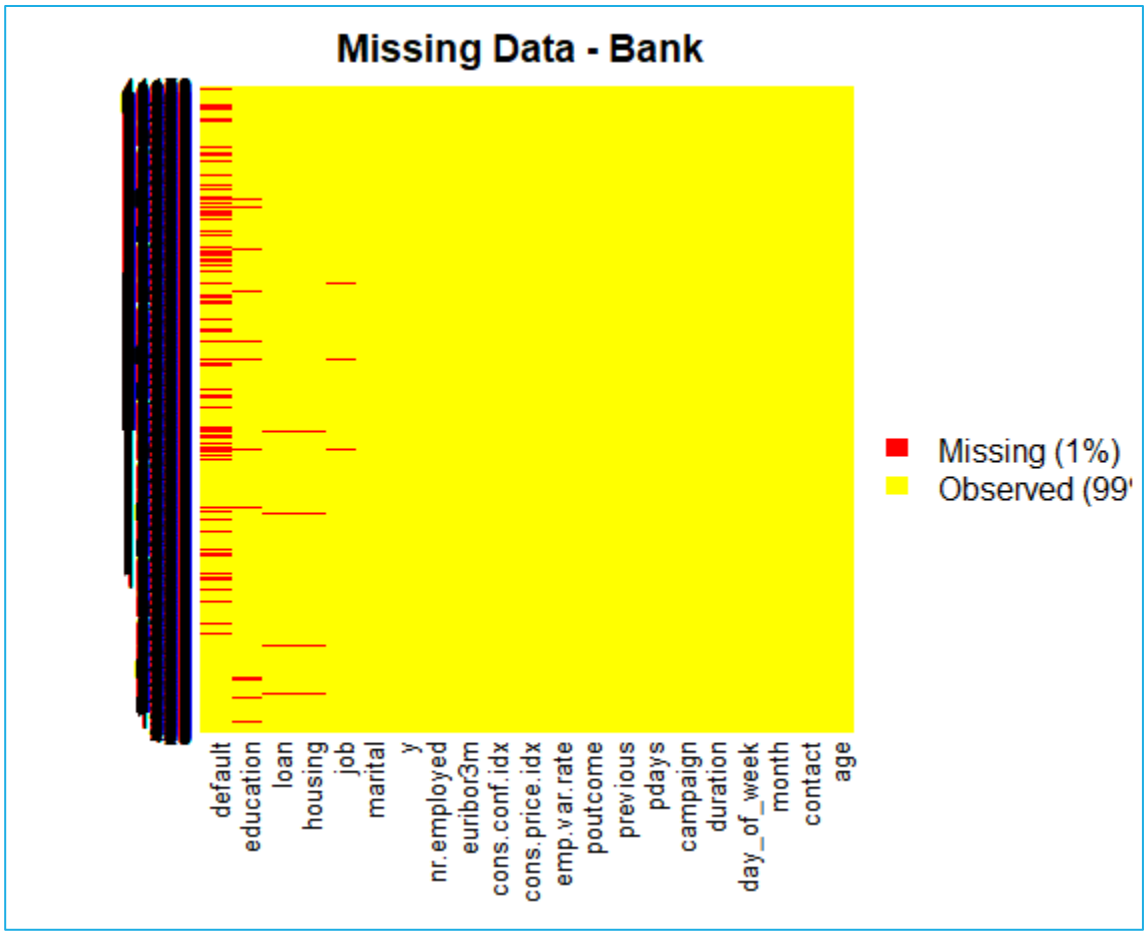
```
> if(length(which(is.na(bankdata)==TRUE)>0)){print("Missing Value found in the specified column")
+ } else print("All okay: No Missing Value found in the specified column")
[1] "Missing Value found in the specified column"
> summary(bankdata)
      age            job              marital           education          default            housing              loan              contact       
 Min.   :17.00   Length:41188       Length:41188       Length:41188       Length:41188       Length:41188       Length:41188       Length:41188      
 1st Qu.:32.00   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Median :38.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :40.02                                                                                                                                      
 3rd Qu.:47.00                                                                                                                                      
 Max.   :98.00                                                                                                                                      
    month           day_of_week          duration        campaign          pdays          previous       poutcome          emp.var.rate      
 Length:41188       Length:41188       Min.   :   0.0   Min.   : 1.000   Min.   :  0.0   Min.   :0.000   Length:41188       Min.   :-3.40000  
 Class :character   Class :character   1st Qu.: 102.0   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   Class :character   1st Qu.:-1.80000  
 Mode  :character   Mode  :character   Median : 180.0   Median : 2.000   Median :999.0   Median :0.000   Mode  :character   Median : 1.10000  
                                       Mean   : 258.3   Mean   : 2.568   Mean   :962.5   Mean   :0.173                      Mean   : 0.08189  
                                       3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000                      3rd Qu.: 1.40000  
                                       Max.   :4918.0   Max.   :56.000   Max.   :999.0   Max.   :7.000                      Max.   : 1.40000  
 cons.price.idx  cons.conf.idx     euribor3m      nr.employed       y            
 Min.   :92.20   Min.   :-50.8   Min.   :0.634   Min.   :4964   Length:41188      
 1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344   1st Qu.:5099   Class :character  
 Median :93.75   Median :-41.8   Median :4.857   Median :5191   Mode  :character  
 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                     
> dim(bankdata)
[1] 41188    21
```

Perform the below operations:

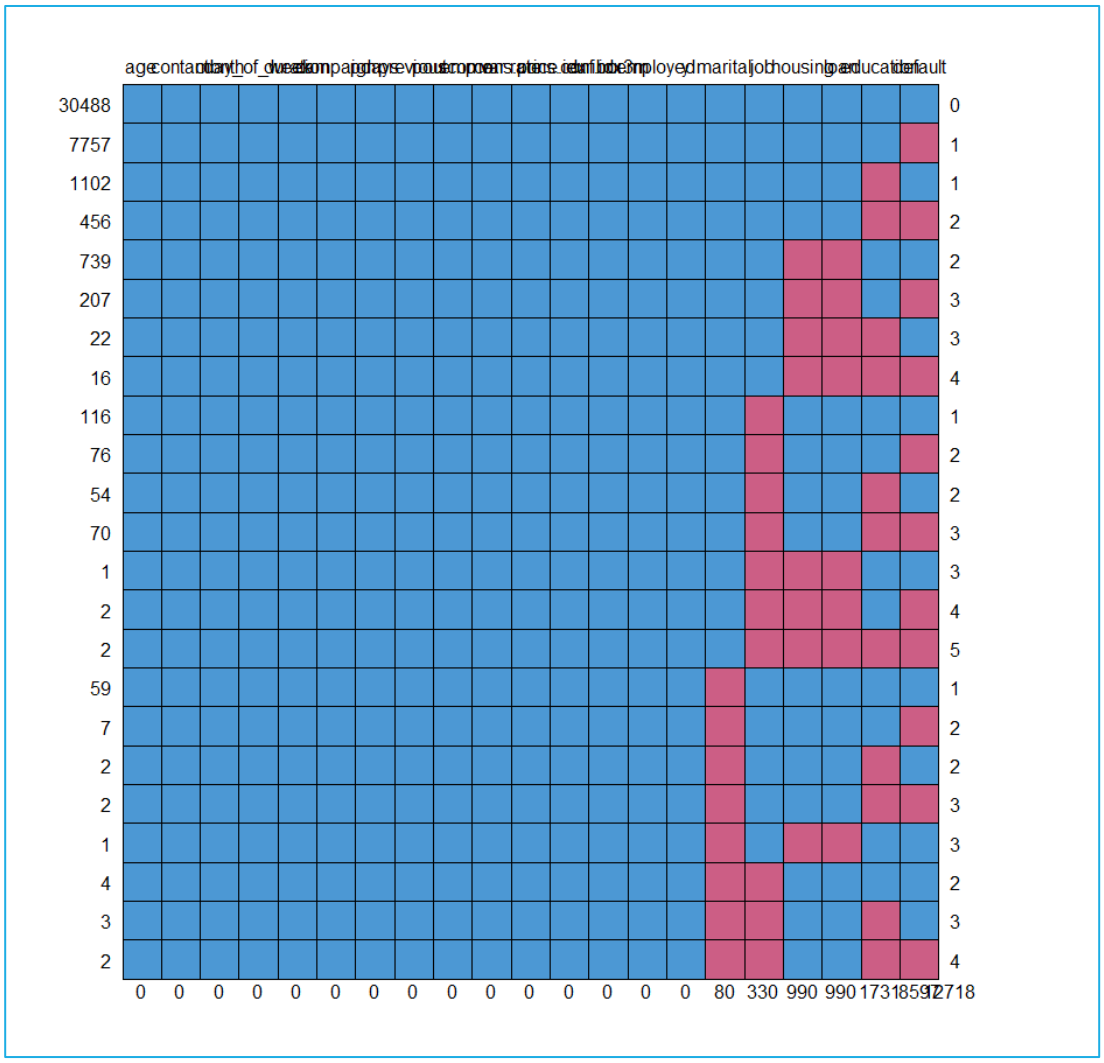
a. Create a visual for representing missing values in the dataset.

```
#OPTION 1 - using Amelia Package
library(Amelia)
missmap(bankdata, main="Missing Data - Bank ", col=c("red","yellow"))

#OPTION 2 - using Mice
library(mice)
md.pattern(bankdata)

#OPTION 3 - using the VIM package as follows
#we can visualize like this too

library(VIM)
missingvalue_plot <- aggr(bankdata, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE,
labels=names(bankdata), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))

#as there are lot many NA Values and for analysis we need complete cases, we can either impute
the data or take the complete cases only, so i am considering the complete cases only

bankdatanew<-bankdata[complete.cases(bankdata), ]
View(bankdatanew)
missmap(bankdatanew,col=c("yellow","red"))
```

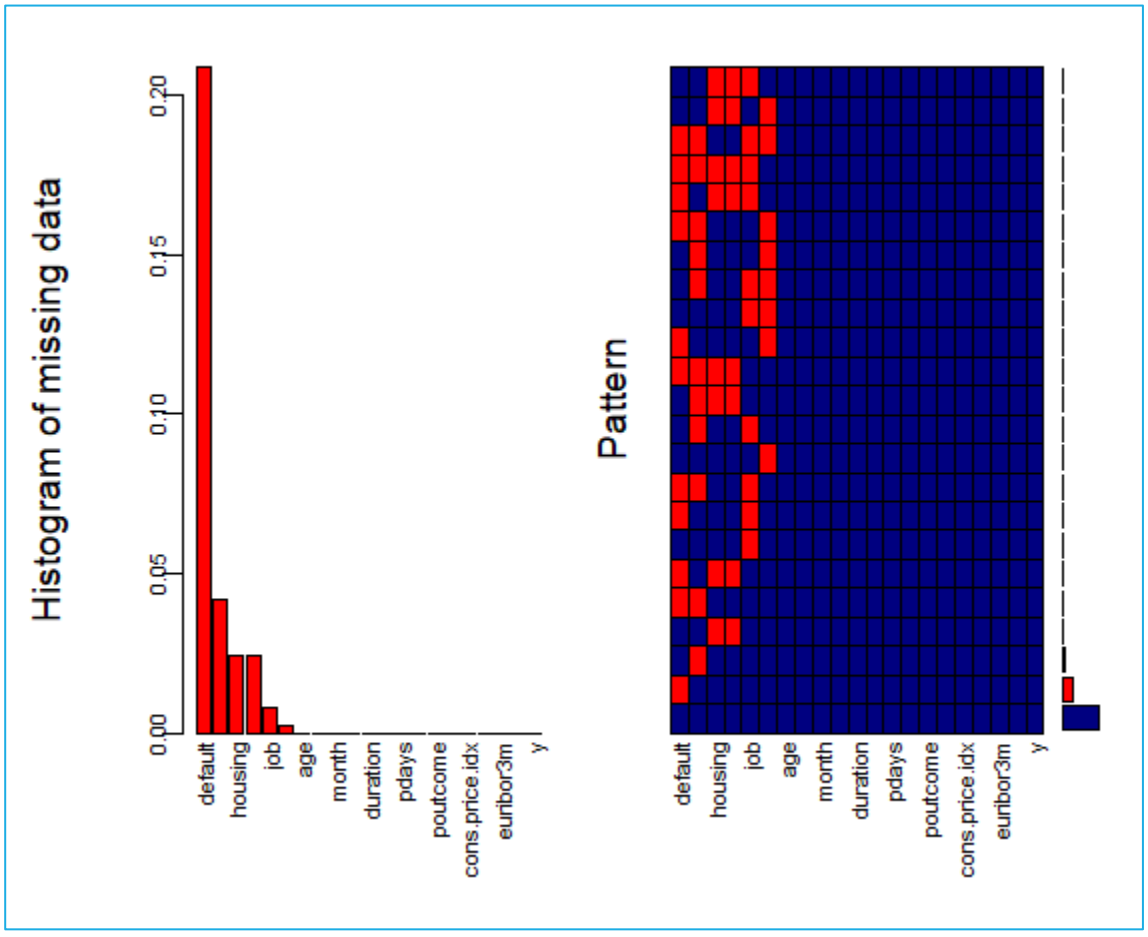





```
> #OPTION 2 - using Mice
> library(mice)
Loading required package: lattice

Attaching package: 'mice'

The following objects are masked from 'package:base':

    cbind, rbind

> md.pattern(bankdata)
      age contact month day_of_week duration campaign pdays previous poutcome emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed y marital job
30488   1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   1
7757    1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   1
1102    1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   1
456     1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   1
739     1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   1
207     1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   1
22      1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   1
16      1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   1
116     1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   0
76      1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   0
54      1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   0
70      1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   0
1       1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   0
2       1       1     1           1        1        1     1        1        1            1              1             1         1           1 1       1   0
```

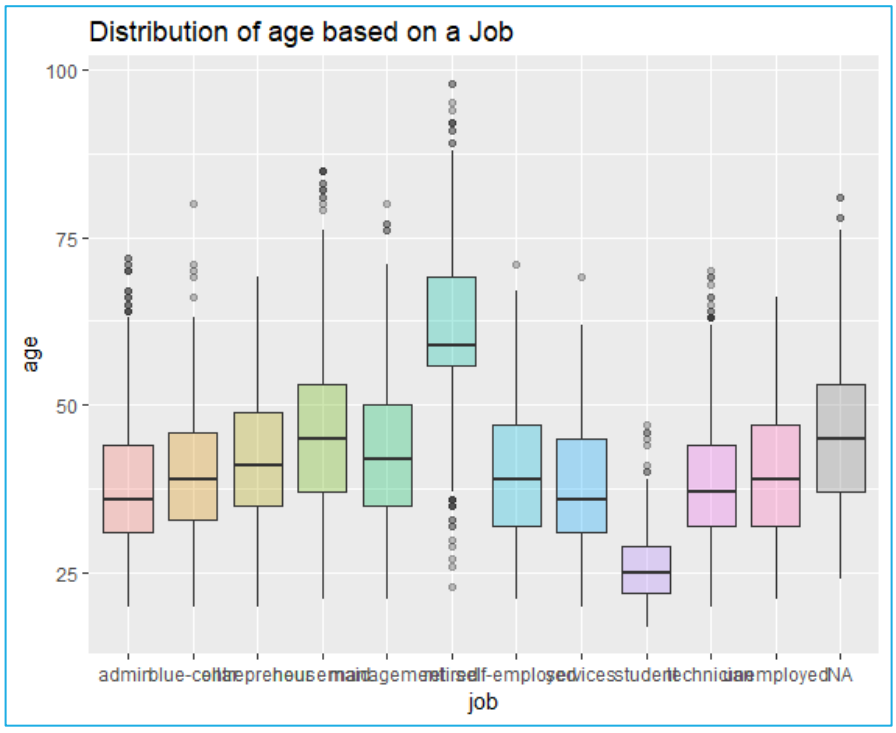


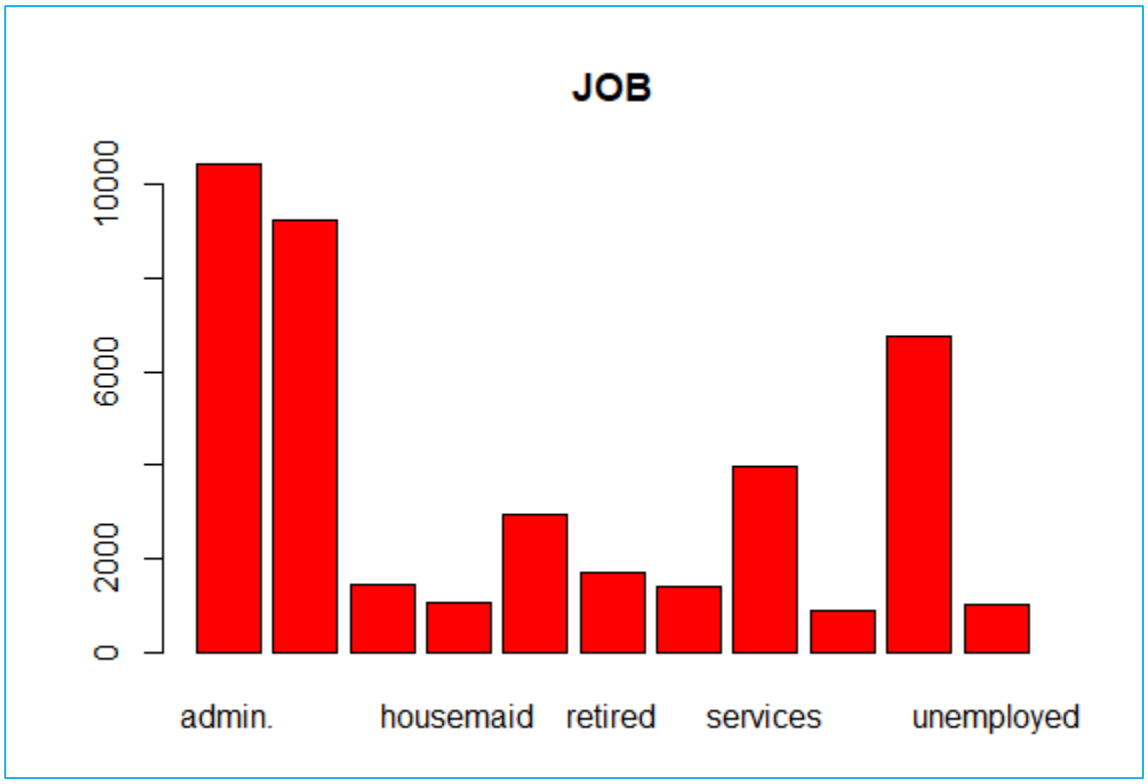

```
> library(VIM)
Loading required package: colorspace
Loading required package: grid
Loading required package: data.table
data.table 1.11.8  Latest news: r-datatable.com
VIM is ready to use.
 Since version 4.0.0 the GUI is in its own package VIMGUI.

          Please use the package to use the new (and old) GUI.

Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

Attaching package: 'VIM'

The following object is masked from 'package:datasets':

    sleep

> missingvalue_plot <- aggr(bankdata, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(bankdata), cex.axis=.7, gap=3, ylab=c("Histogram of miss
ng data","Pattern"))

 Variables sorted by number of missings:
       Variable       Count
        default 0.208725842
      education 0.042026804
        housing 0.024036127
           loan 0.024036127
            job 0.008012042
        marital 0.001942313
            age 0.000000000
        contact 0.000000000
          month 0.000000000
    day_of_week 0.000000000
       duration 0.000000000
       campaign 0.000000000
          pdays 0.000000000
       previous 0.000000000
       poutcome 0.000000000
   emp.var.rate 0.000000000
 cons.price.idx 0.000000000
  cons.conf.idx 0.000000000
      euribor3m 0.000000000
    nr.employed 0.000000000
              y 0.000000000
Warning message:
In plot.aggr(res, ...) : not enough horizontal space to display frequencies
>
```

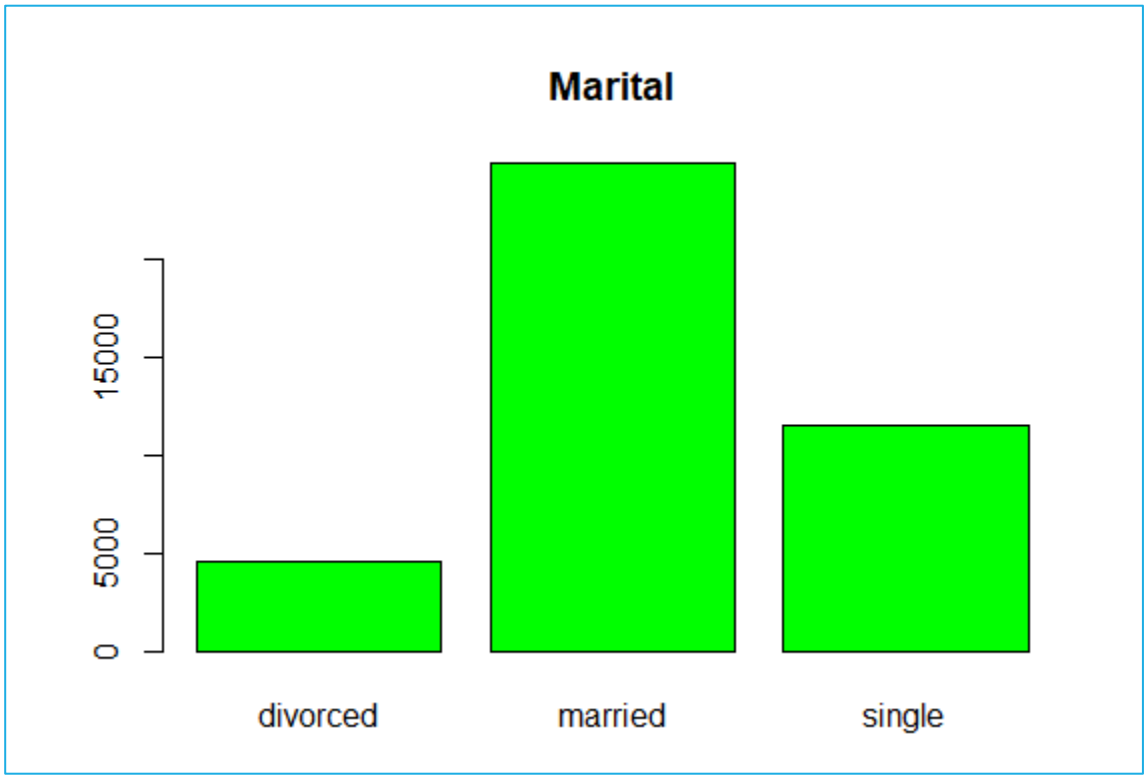


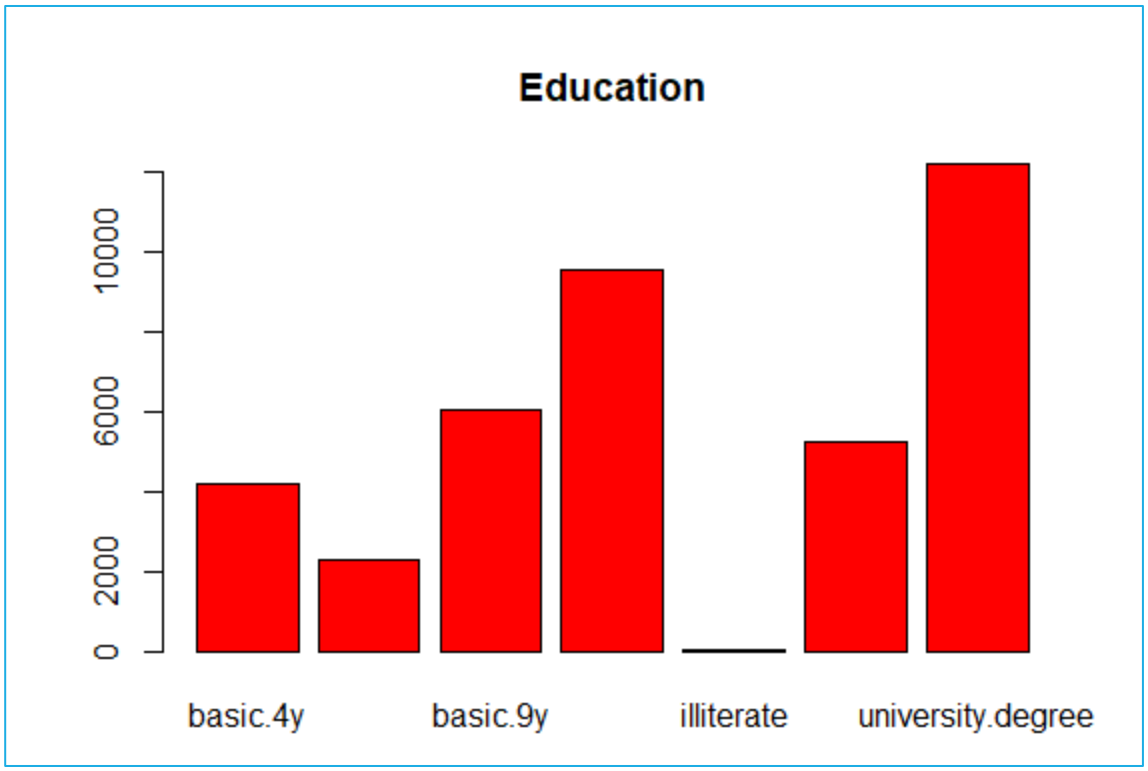
b. Show a distribution of clients based on a job.

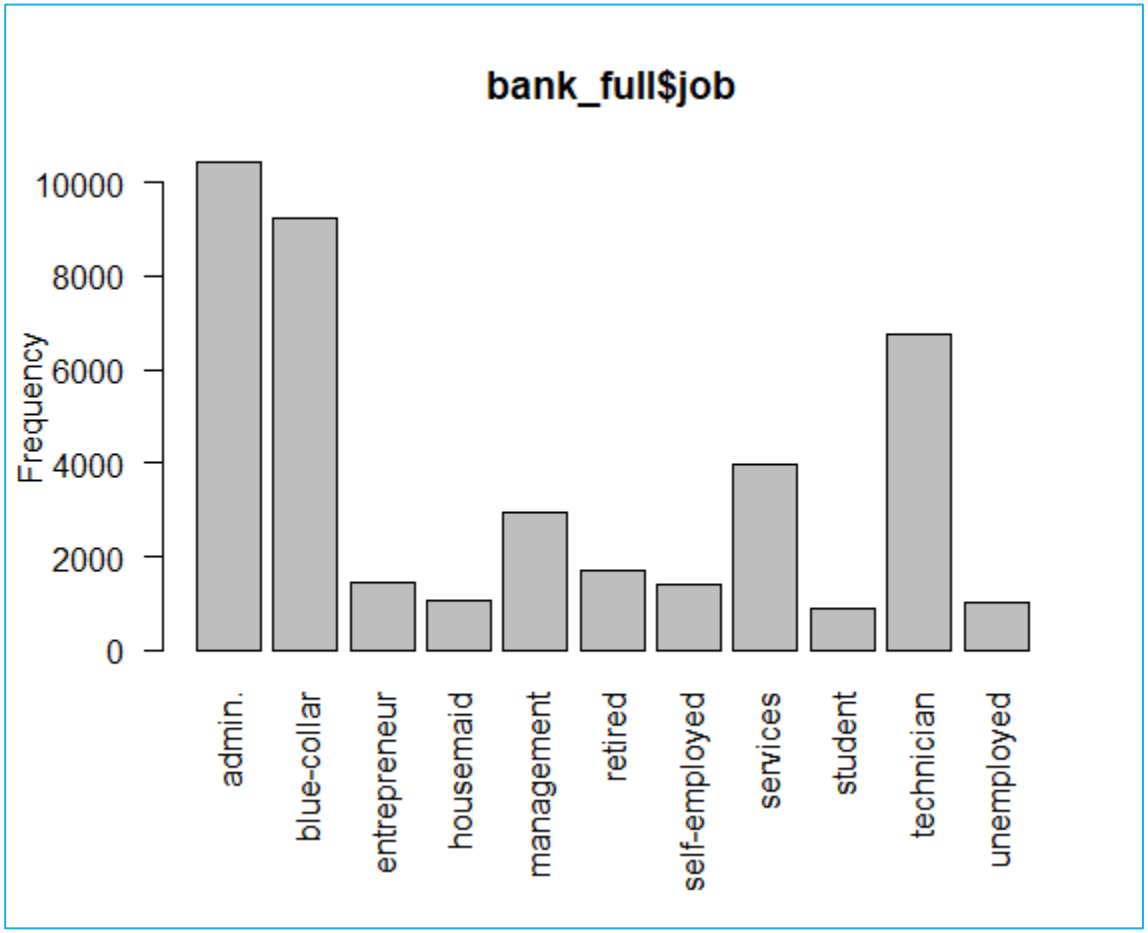
```
#b. Show a distribution of clients based on a Job.
#since in dataset I'm unable to find variable clients therefore i am using
#another variable say age for  showing you distribution of a age based on job
#Set a different color for each group

library(ggplot2)
ggplot(bankdata, aes(x=job, y=age, fill=job)) +  geom_boxplot(alpha=0.3) +
theme(legend.position="none")+ ggtitle("Distribution of age based on a Job")

## Barplotsfor Categorical Variables

par(oma=c(2,0,0,0)) #so labels are not cut off
barplot(table(bankdata$job),col="red",main="JOB")

par(oma=c(2,0,0,0)) #so labels are not cut off
barplot(table(bankdata$marital),col="green",main="Marital")

par(oma=c(2,0,0,0)) #so labels are not cut off
barplot(table(bankdata$education),col="red",main="Education")

par(oma=c(2,0,0,0)) #so labels are not cut off
barplot(table(bankdata$job),ylab = "Frequency", main = "bank_full$job",
      border="black", col="grey",las=2)
```




JOB
0
2000
6000
10000
admin.
housemaid
retired
services
unemployed


Marital
0
5000
15000
divorced
married
single


Education
0
2000
6000
10000
basic.4y
basic.9y
illiterate
university.degree


bank_full$job
Frequency
0
2000
4000
6000
8000
10000
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed

c. Check whether is there any relation between Job and Marital Status?

```
#c. Check whether is there any relation between Job and Marital Status?
#we are using Chi-Square Test for checking relation as both job and marital status are categorical variables so first defining the null hypothesis
#Ho: There is no relation between job and marital status
#Ha: There is relation between job and marital status

with(bankdata, chisq.test( job, marital))

#OR

chisq.test(bankdata$job ,bankdata$marital)
```

```
> #Set a different color for each group
> library(ggplot2)
> ## Barplotsfor Categorical Variables
> par(oma=c(2,0,0,0)) #so labels are not cut off
> barplot(table(bankdata$job),col="red",main="JOB")
> par(oma=c(2,0,0,0)) #so labels are not cut off
> barplot(table(bankdata$marital),col="green",main="Marital")
> par(oma=c(2,0,0,0)) #so labels are not cut off
> barplot(table(bankdata$education),col="red",main="Education")
> par(oma=c(2,0,0,0)) #so labels are not cut off
> barplot(table(bankdata$job),ylab = "Frequency", main = "bank_full$job",
+         border="black", col="grey",las=2)
> with(bankdata, chisq.test( job, marital))

        Pearson's Chi-squared test

data:  job and marital
X-squared = 4045.1, df = 20, p-value < 2.2e-16
```

```
> #OR
> chisq.test(bankdata$job ,bankdata$marital)

        Pearson's Chi-squared test

data:  bankdata$job and bankdata$marital
X-squared = 4045.1, df = 20, p-value < 2.2e-16
```

```
#now as we can see p value is nearly 0 or less which is henceforth less than 0.05
#p value<0.05 hence we will reject the null hypo and accept the alternative hypothesis
#which says that There is relation between job and marital status
```

d. Check whether is there any association between Job and Education?

```
#  d. Check whether is there any association between Job and Education?

#we are using Chi-Square Test for checking association as both job and education are categorical variables hence Chi-Square Test for checking association
#so first defining the null hypothesis
#Ho: There is no association between job and education
#Ha: There is association between job and education

with(bankdata, chisq.test( job, education))

with(bankdata, table(job, education))

with(bankdata, prop.table(table(job, education)))

#now as we can see p value is nearly 0 or less which is henceforth less than 0.05
#p value<0.05 hence we will reject the null hypo and accept the alternative hypothesis
#which says that There is association between job and education
```

```
> with(bankdata, chisq.test( job, education))

        Pearson's Chi-squared test

data:  job and education
X-squared = 35560, df = 60, p-value < 2.2e-16
```

```
> with(bankdata, table(job, education))
             education
job           basic.4y basic.6y basic.9y high.school illiterate professional.course university.degree
  admin.            77      151      499         3329          1                363               5753
  blue-collar     2318     1426     3623          878          8                453                 94
  entrepreneur     137       71      210          234          2                135                610
  housemaid        474       77       94          174          1                 59                139
  management       100       85      166          298          0                 89               2063
  retired          597       75      145          276          3                241                285
  self-employed     93       25      220          118          3                168                765
  services         132      226      388         2682          0                218                173
  student           26       13       99          357          0                 43                170
  technician        58       87      384          873          0               3320               1809
  unemployed       112       34      186          259          0                142                262
> with(bankdata, prop.table(table(job, education)))
             education
job                basic.4y     basic.6y     basic.9y  high.school   illiterate professional.course university.degree
  admin.       1.961384e-03 3.846350e-03 1.271079e-02 8.479800e-02 2.547252e-05        9.246523e-03      1.465434e-01
  blue-collar  5.904529e-02 3.632381e-02 9.228692e-02 2.236487e-02 2.037801e-04        1.153905e-02      2.394416e-03
  entrepreneur 3.489735e-03 1.808549e-03 5.349228e-03 5.960569e-03 5.094503e-05        3.438790e-03      1.553823e-02
  housemaid    1.207397e-02 1.961384e-03 2.394416e-03 4.432218e-03 2.547252e-05        1.502878e-03      3.540680e-03
  management   2.547252e-03 2.165164e-03 4.228438e-03 7.590810e-03 0.000000e+00        2.267054e-03      5.254980e-02
  retired      1.520709e-02 1.910439e-03 3.693515e-03 7.030414e-03 7.641755e-05        6.138876e-03      7.259667e-03
  self-employed 2.368944e-03 6.368129e-04 5.603953e-03 3.005757e-03 7.641755e-05        4.279383e-03      1.948647e-02
  services     3.362372e-03 5.756788e-03 9.883336e-03 6.831729e-02 0.000000e+00        5.553008e-03      4.406745e-03
  student      6.622854e-04 3.311427e-04 2.521779e-03 9.093688e-03 0.000000e+00        1.095318e-03      4.330328e-03
  technician   1.477406e-03 2.216109e-03 9.781446e-03 2.223751e-02 0.000000e+00        8.456875e-02      4.607978e-02
  unemployed   2.852922e-03 8.660655e-04 4.737888e-03 6.597381e-03 0.000000e+00        3.617097e-03      6.673799e-03
>
```