ADEI

Data Processing, Description, Validation and Profiling

Fujie Mei Sergio Delgado Mario Wang

May 16, 2023

Contents

1 Introduction

The course project is concerned with Multivariant Data Analysis and model building for response variables for recollected data of the outcome of a marketing campaign performed by a bank: Y (Binary Target) and numeric variable Duration (Numeric Target) are the targets Aim is to predict how much probability you have to be successful given some socioeconomics characteristics. It involves a binary outcome. As a secondary goal, is to predict the duration of the calls. The first part will consist on data prepropcessing(dealing with missings, outliers. . .), univariate descriptive analysis and profiling.

2 Data description

- Description https://archive.ics.uci.edu/ml/datasets/bank+marketing

2.1 Variables

- 1 - age (numeric)
- 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- 3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
- 4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','un
- 5 - default: has credit in default? (categorical: 'no','yes','unknown')
- 6 - housing: has housing loan? (categorical: 'no','yes','unknown')
- 7 - loan: has personal loan? (categorical: 'no','yes','unknown')
- 8 - contact: contact communication type (categorical: 'cellular','telephone')
- 9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', . . . , 'nov', 'dec')
- 10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- 11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
- 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

- 14 - previous: number of contacts performed before this campaign and for this client (numeric)
- 15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
-social and economic context attributes
- 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
- 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
- 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- 19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
- 20 - nr.employed: number of employees - quarterly indicator (numeric)
- 21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

#Loading the data

```
df<-read.csv2("bank-additional.csv")
head(df)
```

```
##   age       job marital   education default housing loan   contact month
## 1  56 housemaid married    basic.4y      no      no   no telephone   may
## 2  57  services married high.school unknown      no   no telephone   may
## 3  37  services married high.school      no     yes   no telephone   may
## 4  40    admin. married    basic.6y      no      no   no telephone   may
## 5  56  services married high.school      no      no  yes telephone   may
## 6  45  services married    basic.9y unknown      no   no telephone   may
##   day_of_week duration campaign pdays previous    poutcome emp.var.rate
## 1         mon      261        1   999        0 nonexistent          1.1
## 2         mon      149        1   999        0 nonexistent          1.1
## 3         mon      226        1   999        0 nonexistent          1.1
## 4         mon      151        1   999        0 nonexistent          1.1
## 5         mon      307        1   999        0 nonexistent          1.1
## 6         mon      198        1   999        0 nonexistent          1.1
##   cons.price.idx cons.conf.idx euribor3m nr.employed  y
## 1         93.994         -36.4     4.857        5191 no
## 2         93.994         -36.4     4.857        5191 no
## 3         93.994         -36.4     4.857        5191 no
## 4         93.994         -36.4     4.857        5191 no
## 5         93.994         -36.4     4.857        5191 no
## 6         93.994         -36.4     4.857        5191 no
```

2.2 Data sampling 5000 individuals and balancing positives and negatives

```
set.seed(1)
n <- 5000
number_of_trues = as.integer(runif(1, min = 2400, max=2600))

df_yes = df[df$y=="yes",]
df_yes = df_yes[sample(1:2600), ]

df_no = df[df$y=="no",]
df_no = df_no[sample(1:2400),]
df = rbind(df_yes, df_no)
summary(df)
```

```
##       age            job              marital           education
##  Min.   :18.00   Length:5000        Length:5000        Length:5000
##  1st Qu.:32.00   Class :character   Class :character   Class :character
##  Median :38.00   Mode  :character   Mode  :character   Mode  :character
```

```
##  Mean   :39.96
##  3rd Qu.:47.00
##  Max.   :88.00
##    default            housing              loan             contact
##  Length:5000        Length:5000        Length:5000        Length:5000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     month           day_of_week           duration         campaign
##  Length:5000        Length:5000        Min.   :   4.0   Min.   : 1.000
##  Class :character   Class :character   1st Qu.: 175.0   1st Qu.: 1.000
##  Mode  :character   Mode  :character   Median : 342.0   Median : 2.000
##                                        Mean   : 479.8   Mean   : 2.117
##                                        3rd Qu.: 686.0   3rd Qu.: 3.000
##                                        Max.   :4199.0   Max.   :23.000
##      pdays           previous        poutcome          emp.var.rate
##  Min.   :  0.0   Min.   :0.0000   Length:5000        Length:5000
##  1st Qu.:999.0   1st Qu.:0.0000   Class :character   Class :character
##  Median :999.0   Median :0.0000   Mode  :character   Mode  :character
##  Mean   :974.5   Mean   :0.0748
##  3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :999.0   Max.   :3.0000
##  cons.price.idx     cons.conf.idx       euribor3m         nr.employed
##  Length:5000        Length:5000        Length:5000        Length:5000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       y
##  Length:5000
##  Class :character
##  Mode  :character
##
##
##
```

3 Useful Functions

```
#Creates a dataframe of missing per variable
miss<-function(df) {
  missing <-c()
  names<-c()
  colsnames = colnames(df)
  for (variable in 1:length(colnames(df))){
    names<-append(names,colsnames[variable])
    missing<-append(missing,sum(df[,variable]=="unknown")+sum(is.na(df[,variable])))
  }
  df1<-cbind.data.frame(names,missing)
}
```

```r
#Calculates mild and extreme outliers
Outliers<-function(x) {
  sumlist <- summary(x)
  iqr <- sumlist[5]-sumlist[2]
  list(ext_inf_lim = sumlist[2]-3*iqr,ext_sup_lim = sumlist[5]+3*iqr,mild_inf_lim = sumlist[2]-3*iqr,mi
}
```

```r
indivOut<-rep(0,5000)
indivMiss<-rep(0,5000)
indivErrs<-rep(0,5000)
```

```r
colnames<-colnames(df)
outliers<-rep(0,21)
errors<-rep(0,21)
```

```r
plots<-function(df,vector,imputed) {
  if (imputed == TRUE)
    for (var in vector) {
      boxplot(df[var],main = paste("After Imputation: ",var),col=4,las=2)
    }
  else {
    for (var in vector) {
      boxplot(df[var],main = paste("Before Imputation: ",var),col=4,las=2)
    }
  }
}
plotscat<-function(df,vector,imputed) {
  if (imputed == TRUE)
    for (var in vector) {
      barplot(100*prop.table(table(df[var])),main = paste("After Imputation: ",var),col=4,las=2)
    }
  else {
    for (var in vector) {
      barplot(100*prop.table(table(df[var])),main = paste("Before Imputation: ",var),col=4,las=2)
    }
  }
}
```

```r
#Finds in which index is located a variable in the columnames vector
findIndex<-function(x,colnames) {
  i<-1
  while(i < 22) {
    if (colnames[i] == x) {
      break
    }
    i=i+1
  }
  i
}
```

```r
missings<-miss(df)
```

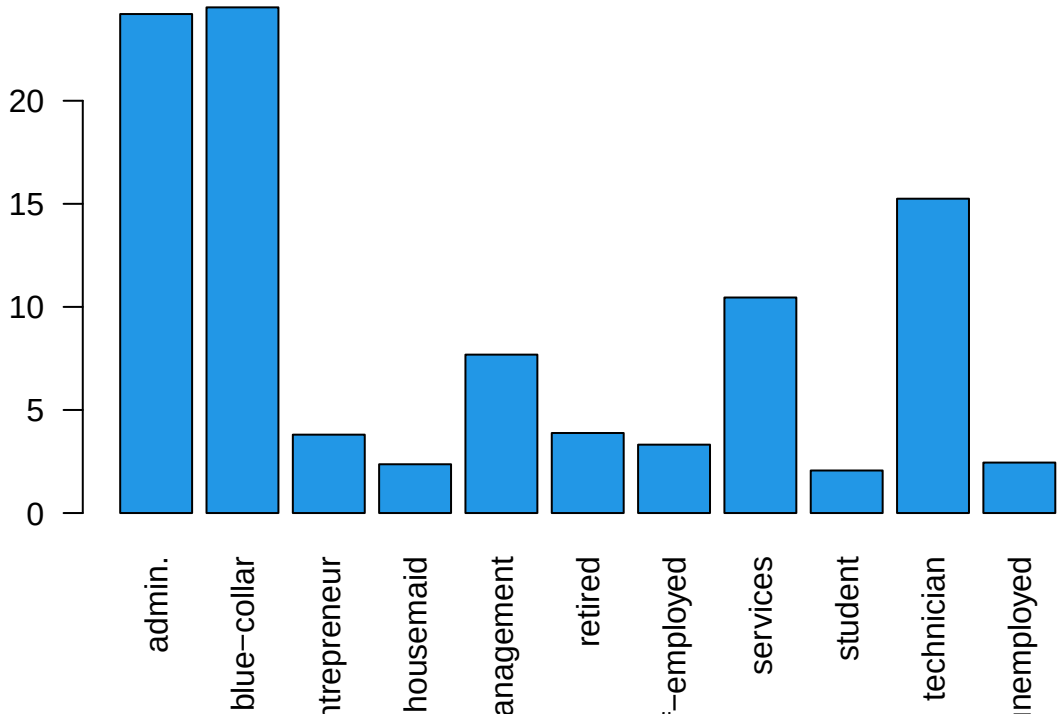
4 Univariate Descriptive Analysis of variables

4.1 Qualitative Variables (Factors) / Categorical

What we will do in this part is to explicitly assign as factors every categorical variable, assign missings as NA's(in case they have) and put it in our vectors for further calculation when needed, finally we will plot it the see each of their proportions and the structure of the graphs.

4.1.1 Job

```
df$job <- as.factor(df$job)
miss<-which(df$job=="unknown")
indivMiss[miss]<-indivMiss[miss]+1
levels(df$job) <- c("admin.","blue-collar","entrepreneur","housemaid","management","retired","self-empl
barplot(100*prop.table(table(df$job)),las=2,col=4)
```

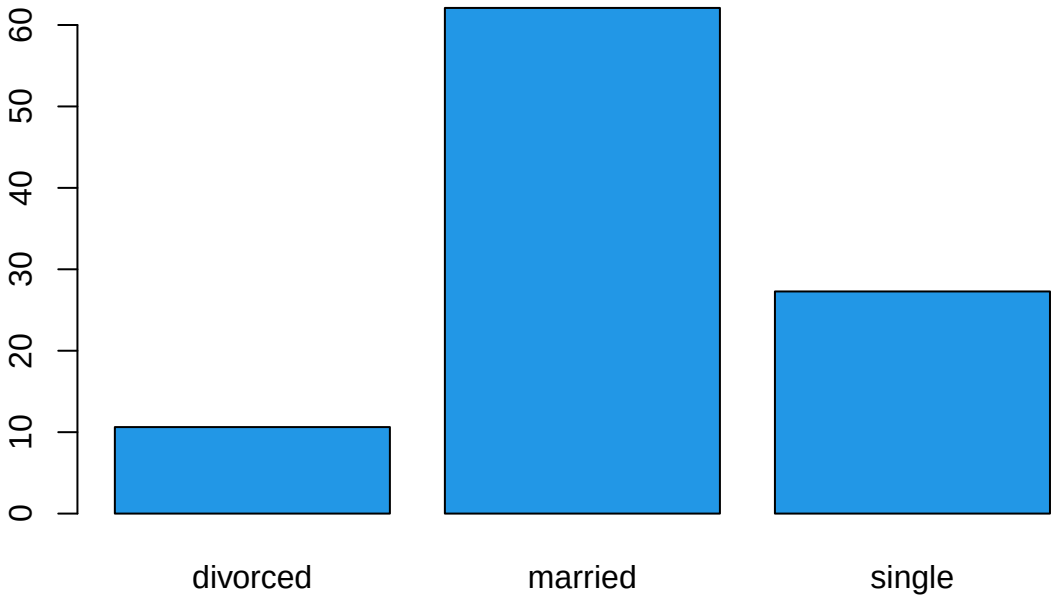



```
summary(df$job)
```

```
##        admin.   blue-collar  entrepreneur     housemaid    management
##          1197          1213           188           117           380
##       retired self-employed      services       student    technician
##           192           164           517           102           754
##    unemployed          NA's
##           121            55
```

4.1.2 Marital

```
df$marital <- as.factor(df$marital)
miss<-which(df$marital=="unknown")
indivMiss[miss]<-indivMiss[miss]+1
levels(df$marital) <- c("divorced","married","single",NA)
barplot(100*prop.table(table(df$marital)),col=4)
```

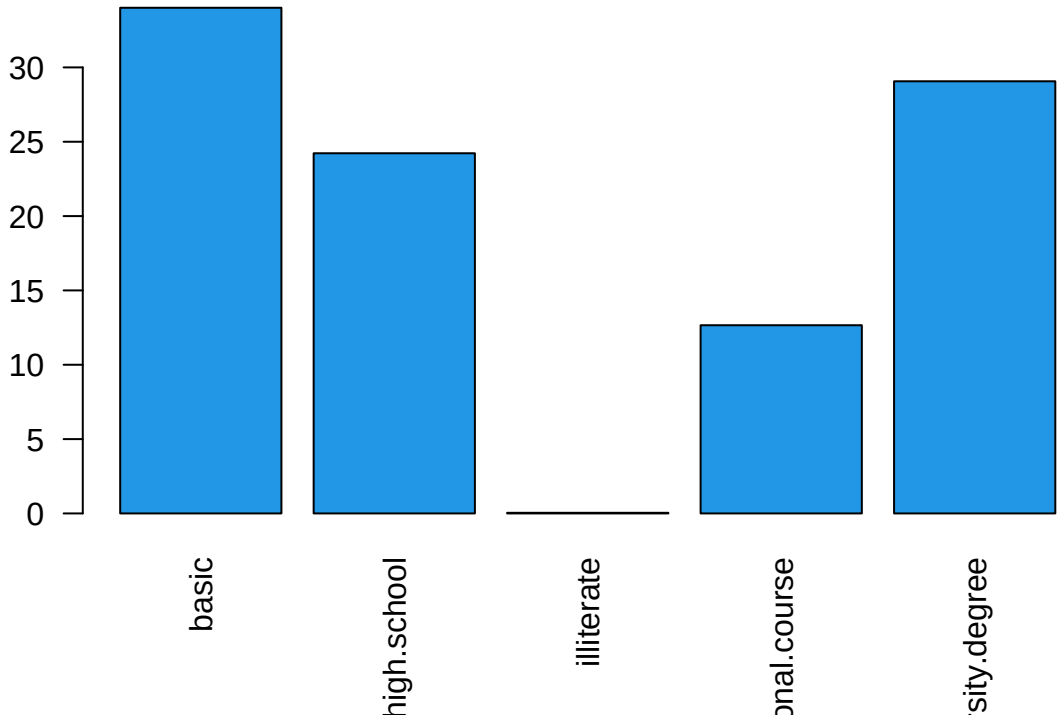



```
summary(df$marital)
```

```
## divorced  married   single     NA's
##      530     3098     1361       11
```

4.1.3 Education

```
df$education <- as.factor(df$education)
miss<-which(df$education=="unknown")
indivMiss[miss]<-indivMiss[miss]+1
levels(df$education) <- c("basic","basic","basic","high.school","illiterate",   "professional.course","u
barplot(100*prop.table(table(df$education)),col=4,las=2)
```



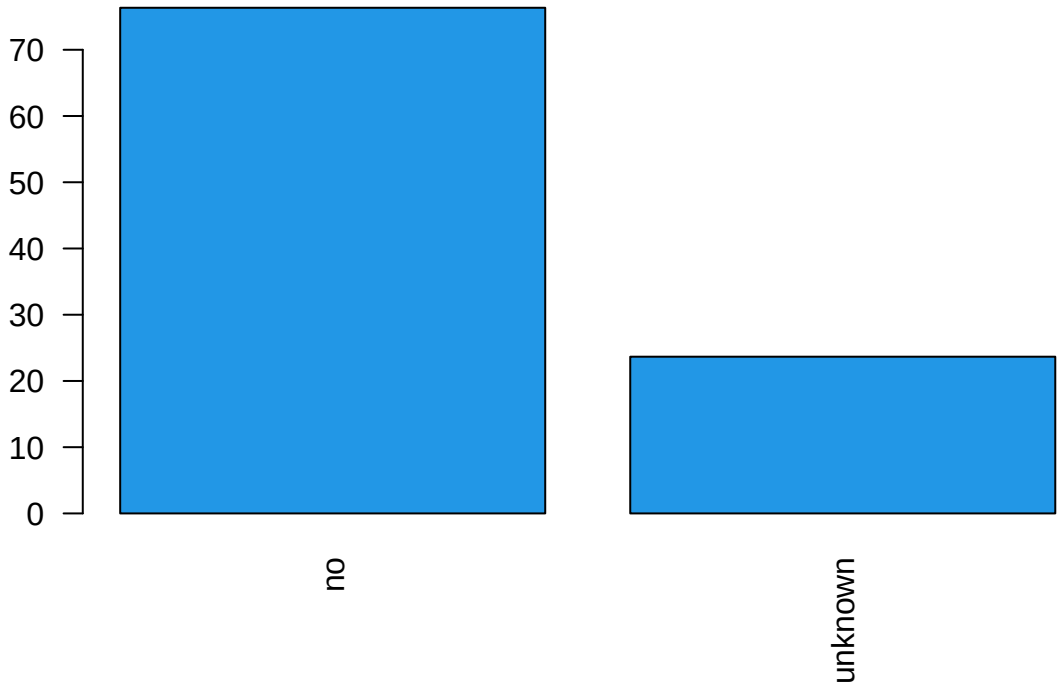

```
summary(df$education)
```

```
##                basic          high.school          illiterate professional.course
##                 1623                 1156                    2                  604
##    university.degree                 NA's
##                 1387                  228
```

4.1.4 Default

We only have two levels in here, no and unknown, so this variable is not explicative and won't contribute to anything so it will be deleted.

```
df$default <- as.factor(df$default)
miss<-which(df$default=="unknown")
indivMiss[miss]<-indivMiss[miss]+1
barplot(100*prop.table(table(df$default)),col=4,las=2)
```



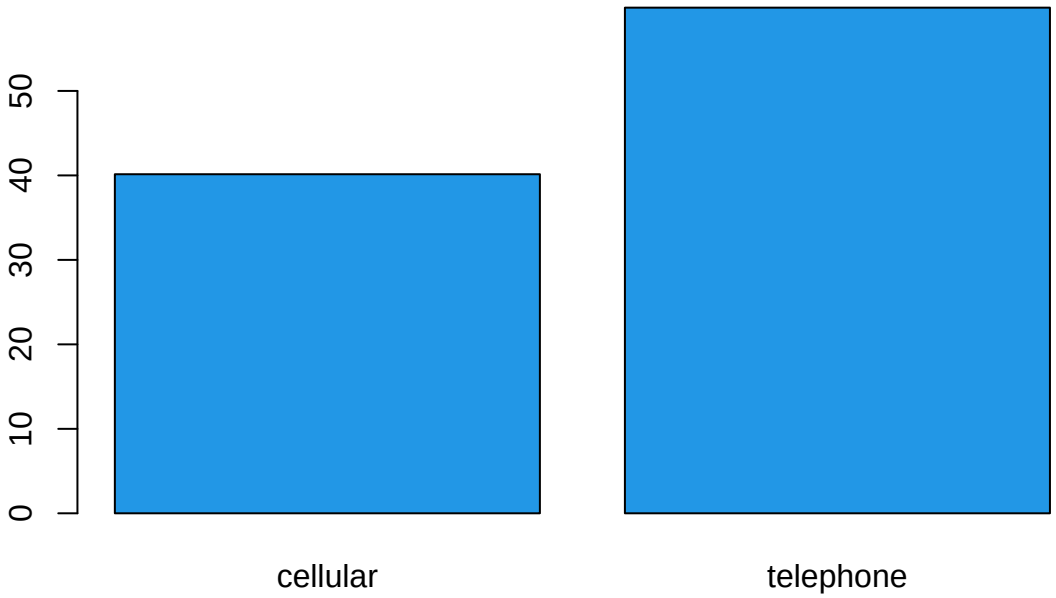

```
summary(df$default)
```

```
##      no unknown
##    3817    1183
```

```
df<-subset(df,select=-default)
```

4.1.5 Contact

```
df$contact <- as.factor(df$contact)
barplot(100*prop.table(table(df$contact)),col=4)
```

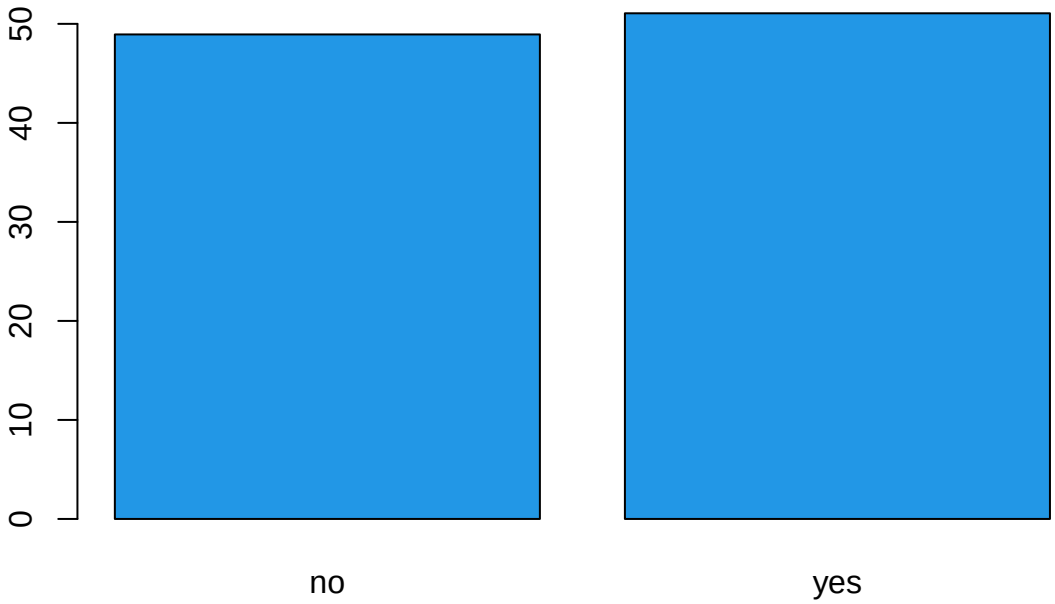


```
summary(df$contact)
```

```
##  cellular telephone
##      2007      2993
```

4.1.6 Housing

```
df$housing <- as.factor(df$housing)
miss<-which(df$housing=="unknown")
indivMiss[miss]<-indivMiss[miss]+1
levels(df$housing) <- c("no",NA,"yes")
barplot(100*prop.table(table(df$housing)),col=4)
```

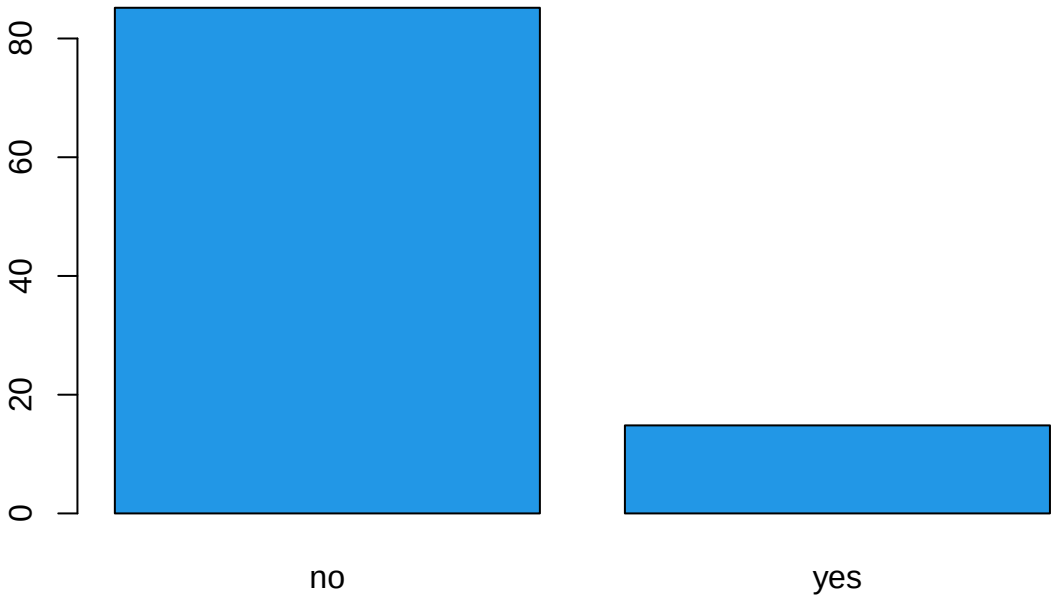



```
summary(df$housing)
```

```
##   no  yes NA's
## 2383 2487  130
```

4.1.7 Loan

```
df$loan <- as.factor(df$loan)
miss<-which(df$loan=="unknown")
indivMiss[miss]<-indivMiss[miss]+1
levels(df$loan) <- c("no",NA,"yes")
barplot(100*prop.table(table(df$loan)),col=4)
```

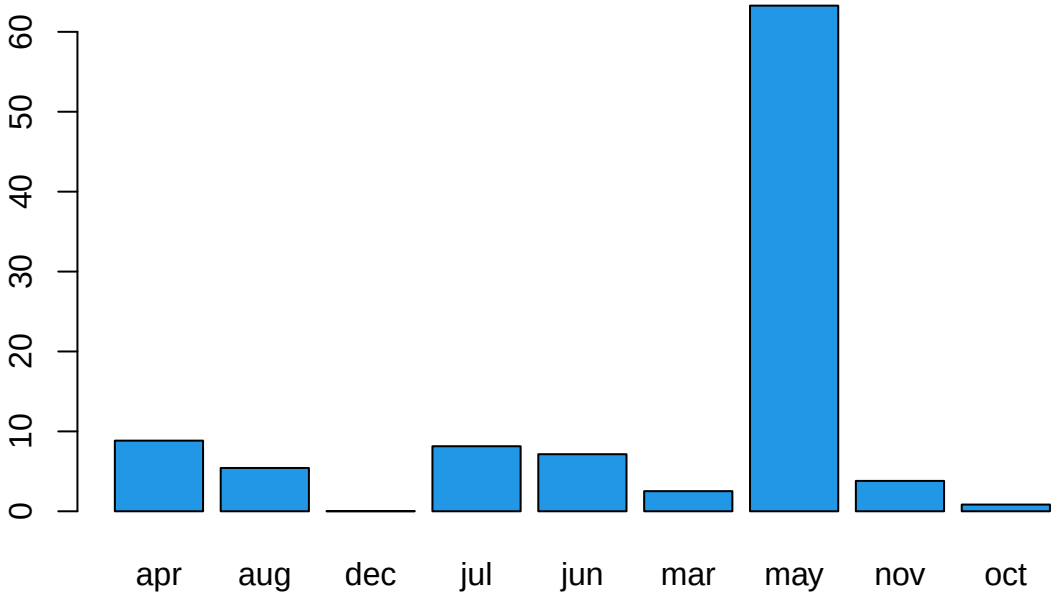



```
summary(df$loan)
```

```
##   no  yes NA's
## 4148  722  130
```

4.1.8 Month

```
df$month <- as.factor(df$month)
barplot(100*prop.table(table(df$month)),col=4)
```

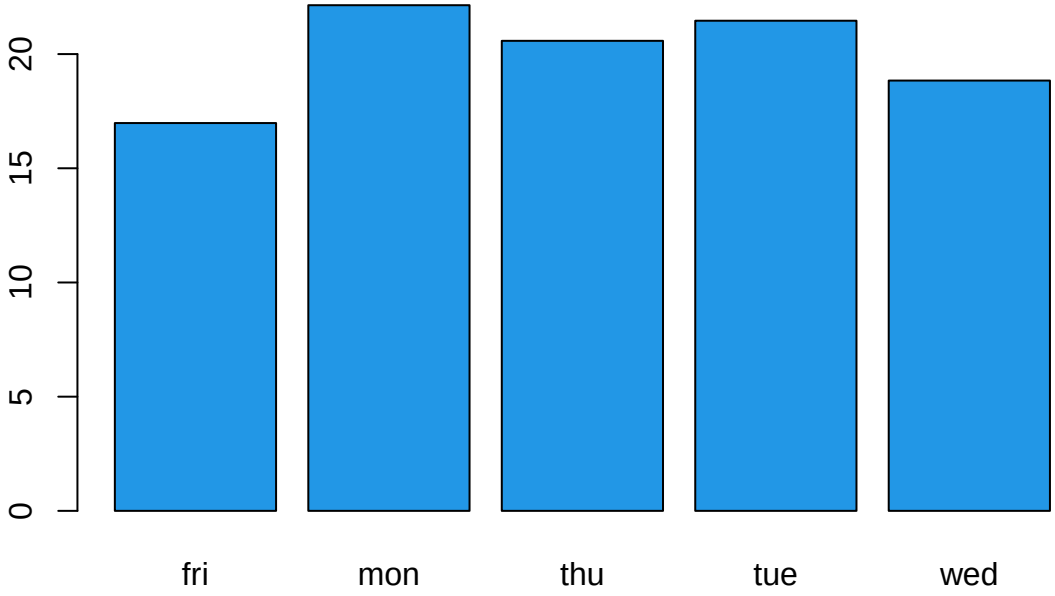



```
summary(df$month)
```

```
##  apr  aug  dec  jul  jun  mar  may  nov  oct
##  442  271    1  407  357  126 3164  190   42
```

4.1.9 Day of the week

```
df$day_of_week <- as.factor(df$day_of_week)
barplot(100*prop.table(table(df$day_of_week)),col=4)
```

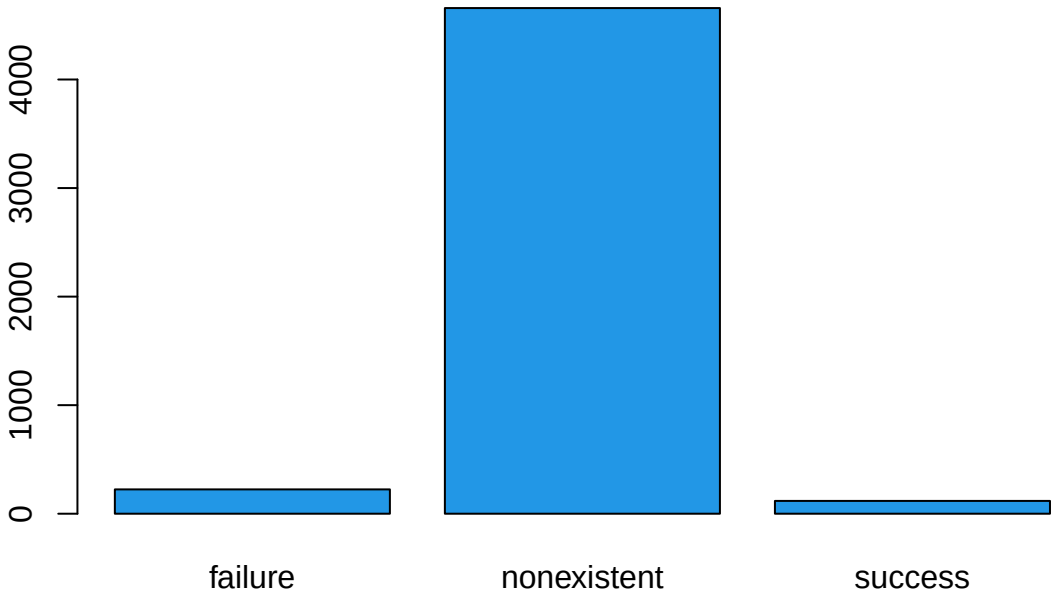



```
summary(df$day_of_week)
```

```
##  fri  mon  thu  tue  wed
##  849 1107 1029 1073  942
```

4.1.10 Poutcome

```
df$poutcome <- as.factor(df$poutcome)
barplot(table(df$poutcome),col=4)
```

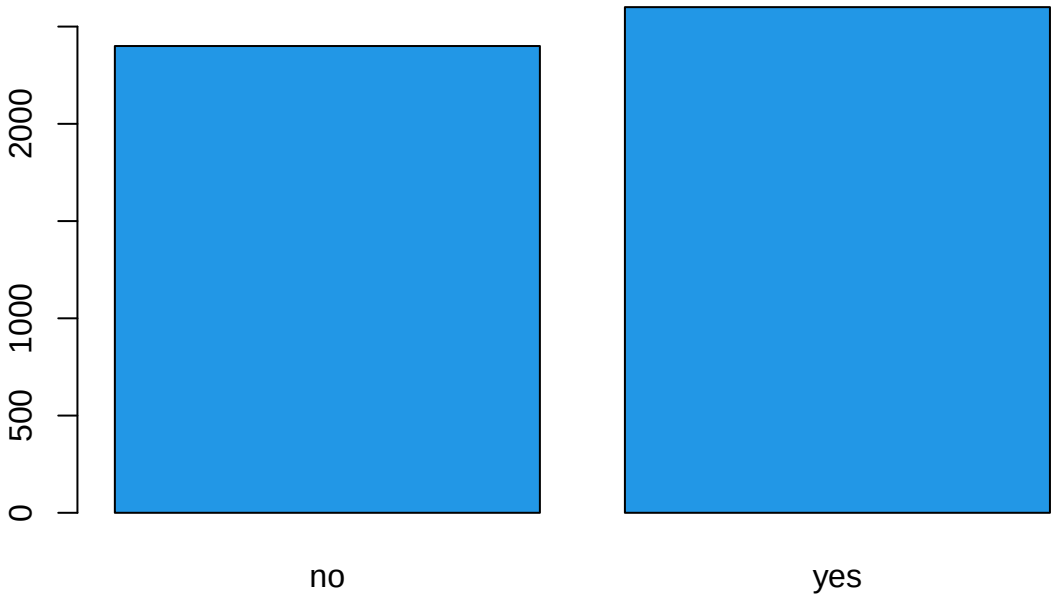



```
summary(df$poutcome)
```

```
##     failure nonexistent     success
##         224        4658         118
```

4.1.11 y

```
df$y <- as.factor(df$y)
barplot(table(df$y),col=4)
```




```
summary(df$y)
```

```
##   no  yes
## 2400 2600
```

4.2 Quantitative Variables / Numerical

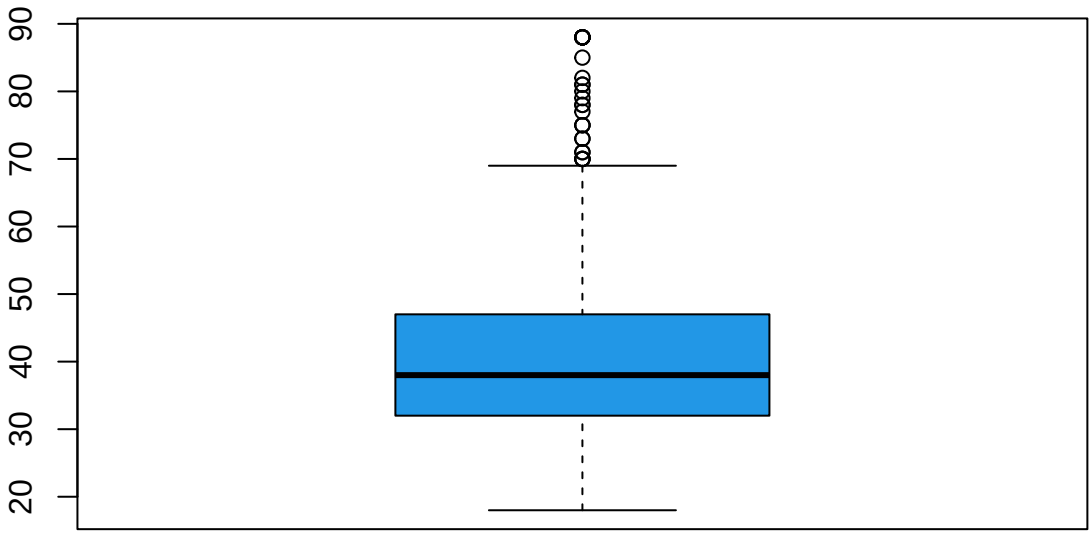
What we will do in this part is to explicitly assign as numericals every numerical variable, in this case we don't have any missings but some of the variables have errors and outliers so we will calculate the mild and extreme outliers and put the mild in our vector for further calculations and use the extremes to assign them as NA's for imputations afterwards, we will see the boxplot for a basic understanding of their structures and where are the extreme outliers located.

4.2.1 Age

```
df$age<-as.numeric(df$age)
boxplot(df$age,col=4)
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   32.00   38.00   39.96   47.00   88.00
```

```
out<-Outliers(df$age)
abline(h=out$ext_sup_lim,col="red")
```




```
i<-findIndex("age",colnames)
ext_nulls<-which(df$age>=out$ext_sup_lim)
mild_nulls<-which(df$age>=out$mild_sup_lim)
l<-(length(mild_nulls))
outliers[i]<-l
indivOut[mild_nulls]<-indivOut[mild_nulls]+1
df[ext_nulls,"age"]<-NA
```

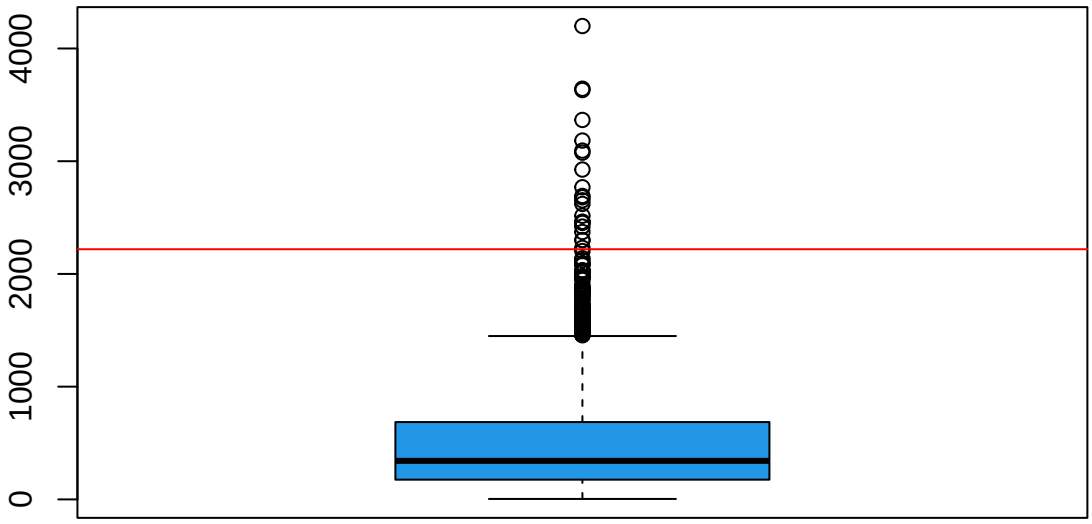
4.2.2 Duration

Since duration is a target variable we will not impute it and leave it as it is.

```
df$duration<-as.numeric(df$duration)
boxplot(df$duration,col=4)
summary(df$duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0   175.0   342.0   479.8   686.0  4199.0
```

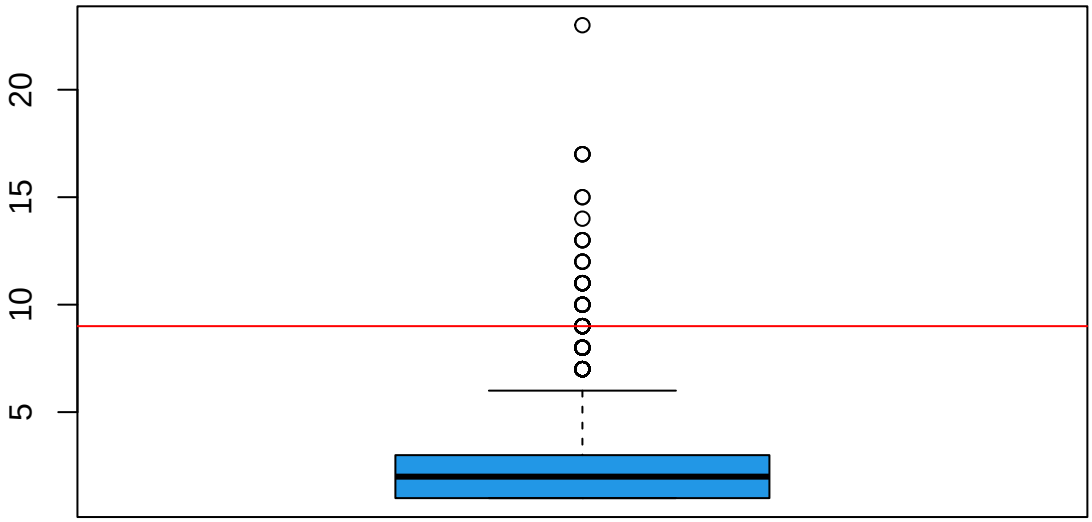
```
out<-Outliers(df$duration)
abline(h=out$ext_sup_lim,col="red")
```



```
i<-findIndex("duration",colnames)
ext_nulls<-which(df$duration>=out$ext_sup_lim)
mild_nulls<-which(df$duration>=out$mild_sup_lim)
l<-(length(mild_nulls))
outliers[i]<-l
indivOut[mild_nulls]<-indivOut[mild_nulls]+1
```

4.2.3 Campaign

```
df$campaign<-as.numeric(df$campaign)
boxplot(df$campaign,col=4)
out<-Outliers(df$campaign)
abline(h=out$mild_sup_lim, col = "red")
```



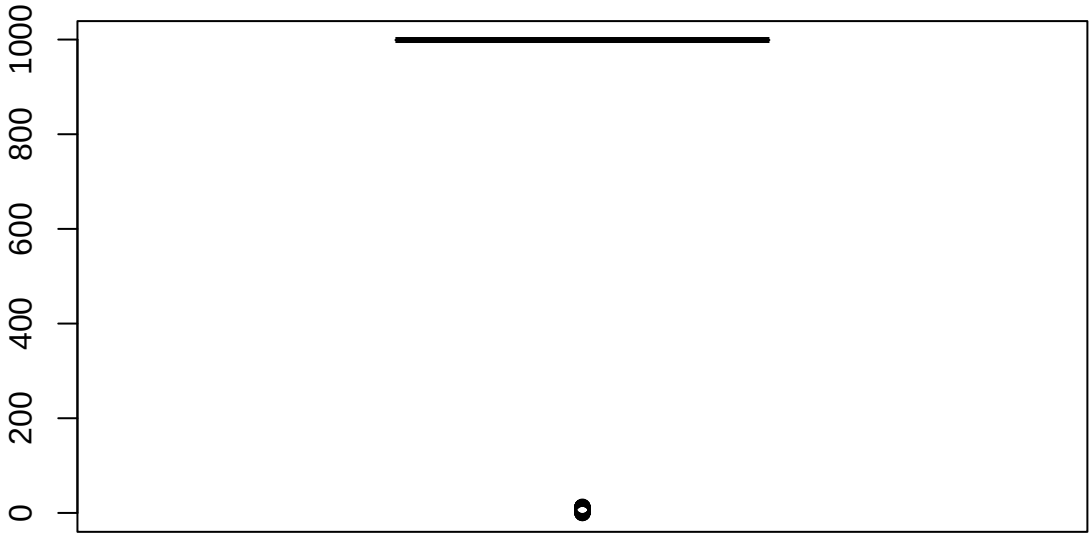

```
summary(df$campaign)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.117   3.000  23.000
```

```
i<-findIndex("campaign",colnames)
ext_nulls<-which(df$campaign>=out$ext_sup_lim)
mild_nulls<-which(df$campaign>=out$mild_sup_lim)
l<-(length(mild_nulls))
outliers[i]<-l
indivOut[mild_nulls]<-indivOut[mild_nulls]+1
df[ext_nulls,"campaign"]<-NA
```

4.2.4 Pay days

```
df$pdays<-as.numeric(df$pdays,col=4)
boxplot(df$pdays)
```



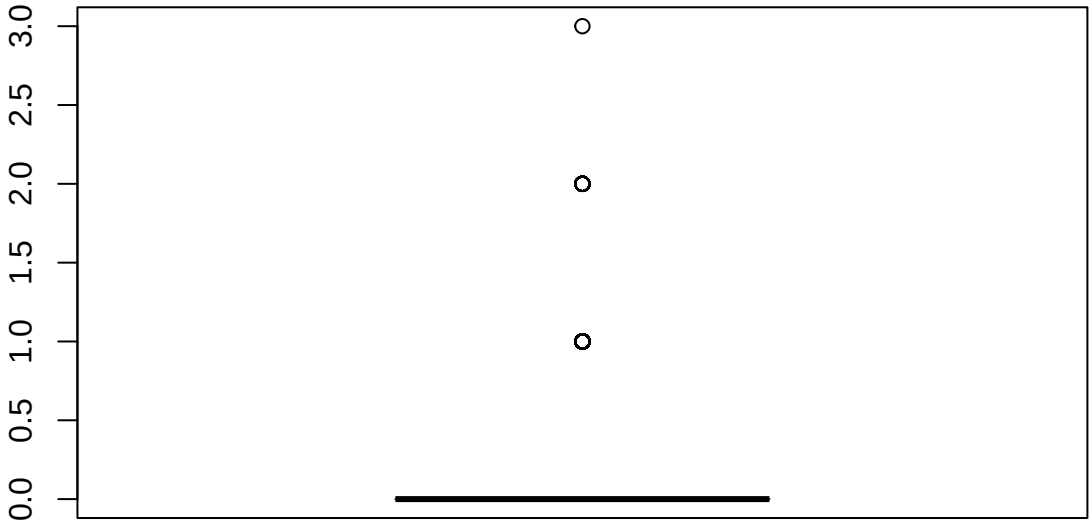

```
summary(df$pdays)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   999.0   999.0   974.5   999.0   999.0
```

4.2.5 Previous

In this case we have errors and inconsistencies, when pdays = 999 and we have previous>0, it's and impossible case because we have never contacted the client in the previous campaign but it appears in the database that the client had > 0 contacts, which is inconsistent.

```
df$previous<-as.numeric(df$previous)
boxplot(df$previous,col=4)
```




```
summary(df$previous)
```

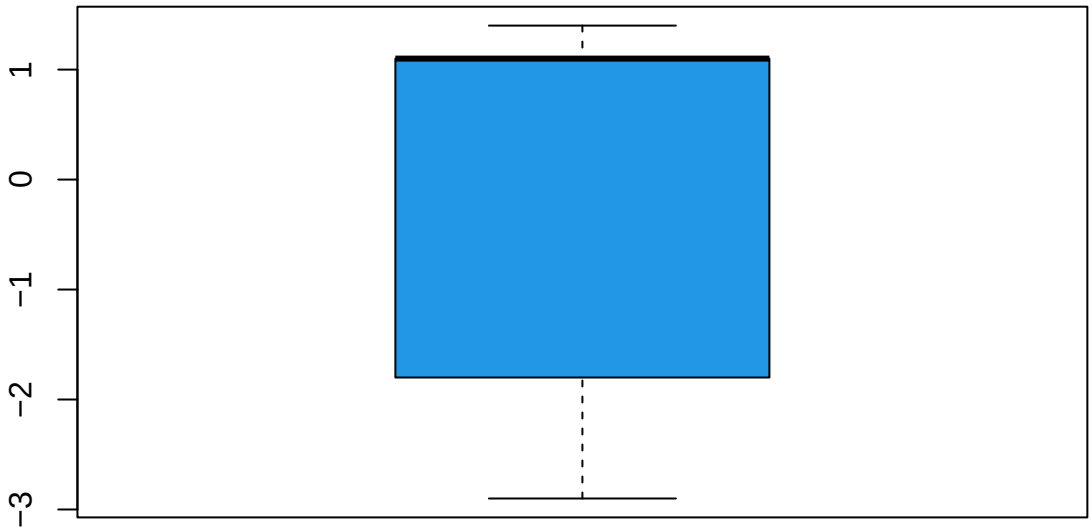
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.0748  0.0000  3.0000
```

```
i<-findIndex("pdays",colnames)
y<-findIndex("previous",colnames)
l<-which(df$pdays==999 & df$previous>0)
```

```
errors[i]<-length(l)
errors[y]<-length(l)
indivErrs[l]<-indivErrs[l]+1
```

4.2.6 Employment variation rate

```
df$emp.var.rate<-as.numeric(df$emp.var.rate)
boxplot(df$emp.var.rate,col=4)
```

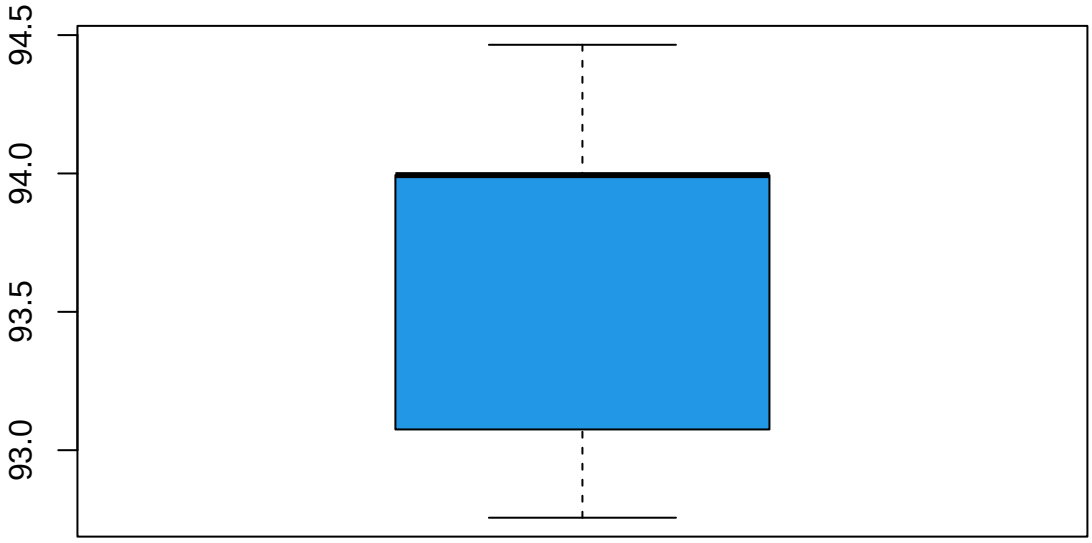



```
summary(df$emp.var.rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.9000 -1.8000  1.1000  0.3275  1.1000  1.4000
```

4.2.7 Consumer price index

```
df$cons.price.idx<-as.numeric(df$cons.price.idx)
boxplot(df$cons.price.idx,col=4)
```

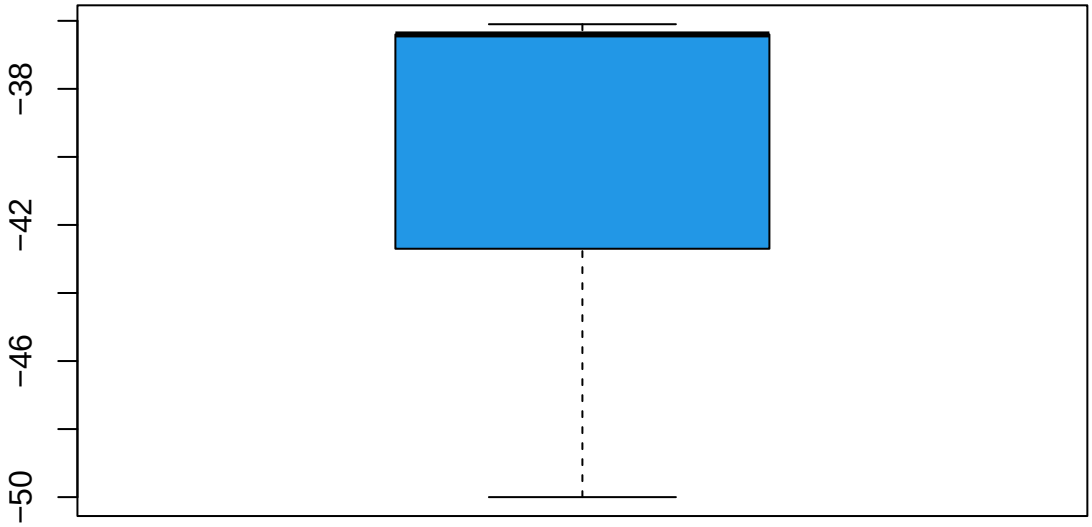


```
summary(df$cons.price.idx)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   92.76   93.08   93.99   93.68   93.99   94.47
```

4.2.8 Consumer confidence index

```
df$cons.conf.idx<-as.numeric(df$cons.conf.idx)
boxplot(df$cons.conf.idx,col=4)
```

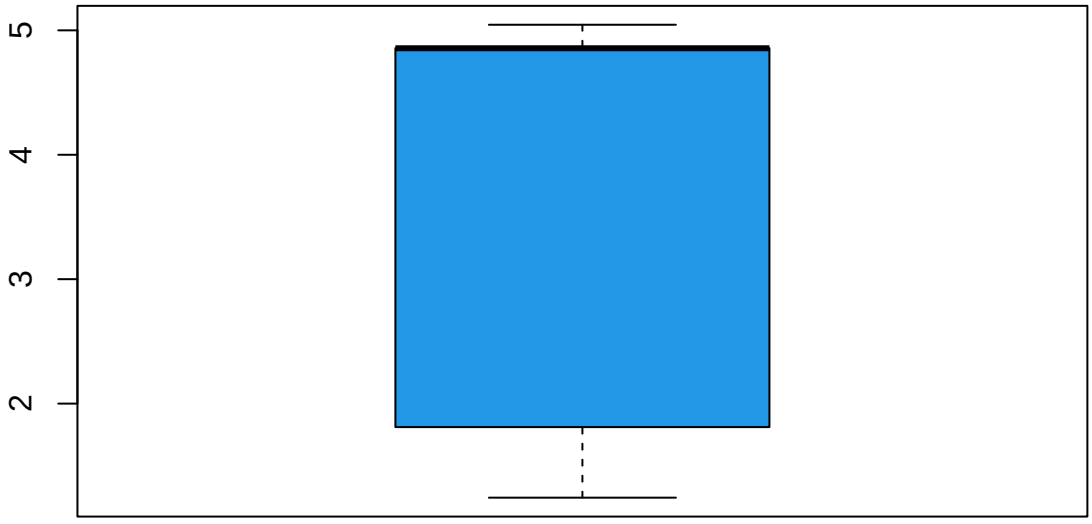



```
summary(df$cons.conf.idx)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -50.00  -42.70  -36.40  -39.81  -36.40  -36.10
```

4.2.9 Euribor 3 month rate

```
df$euribor3m<-as.numeric(df$euribor3m)
boxplot(df$euribor3m,col=4)
```

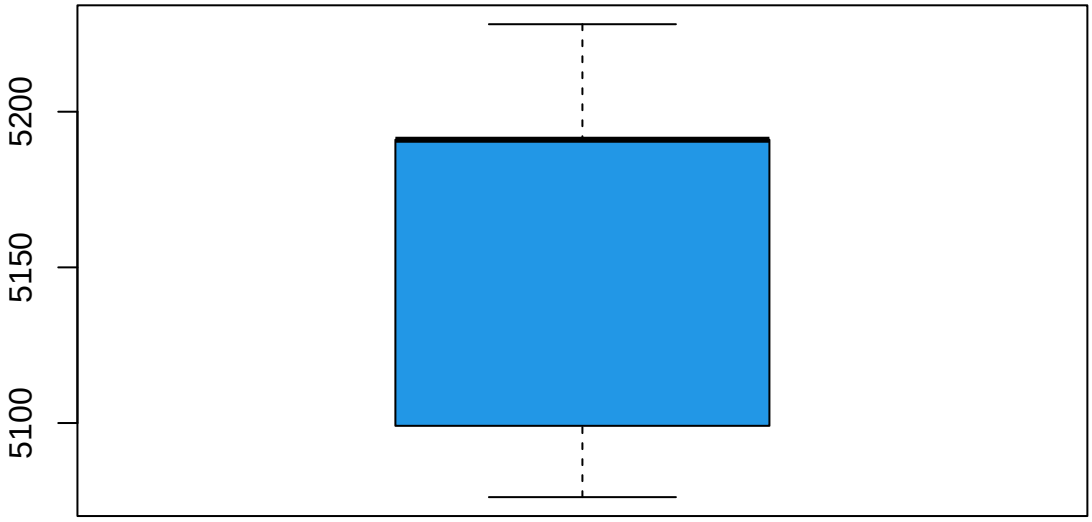


```
summary(df$euribor3m)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.244   1.811   4.856   3.965   4.857   5.045
```

4.2.10 Nr.employed

```
df$nr.employed<-as.numeric(df$nr.employed)
boxplot(df$nr.employed,col=4)
```




```
summary(df$nr.employed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5076    5099    5191    5174    5191    5228
```

5 Imputation

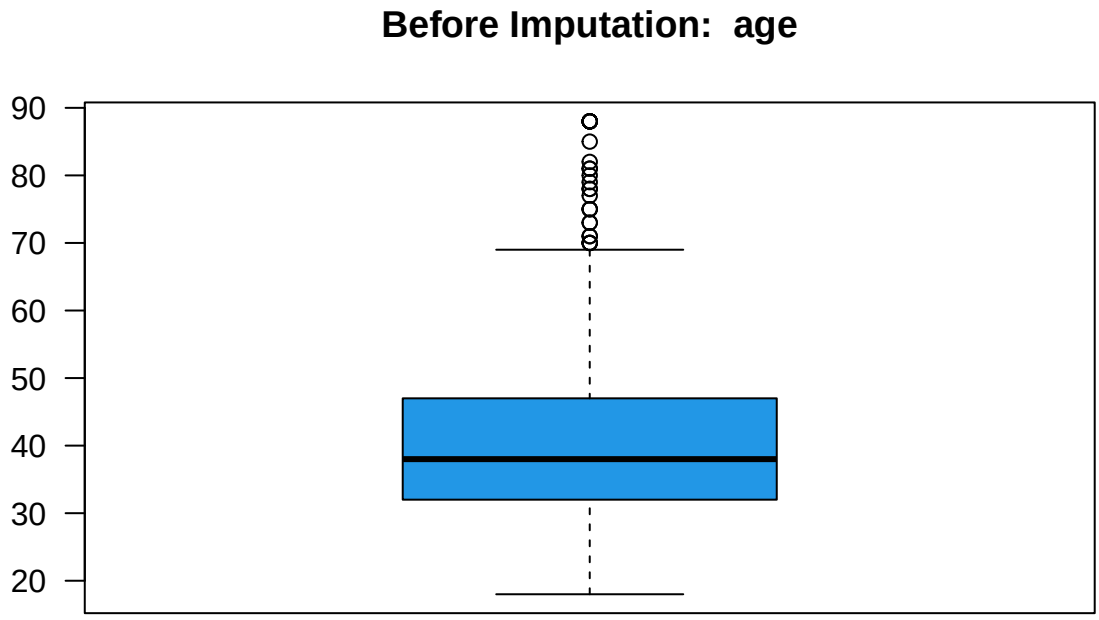
For each imputation we will see that the graphs structure stays almost exactly the same and with summary we see the NA's going away so we can validate the imputations as correct.

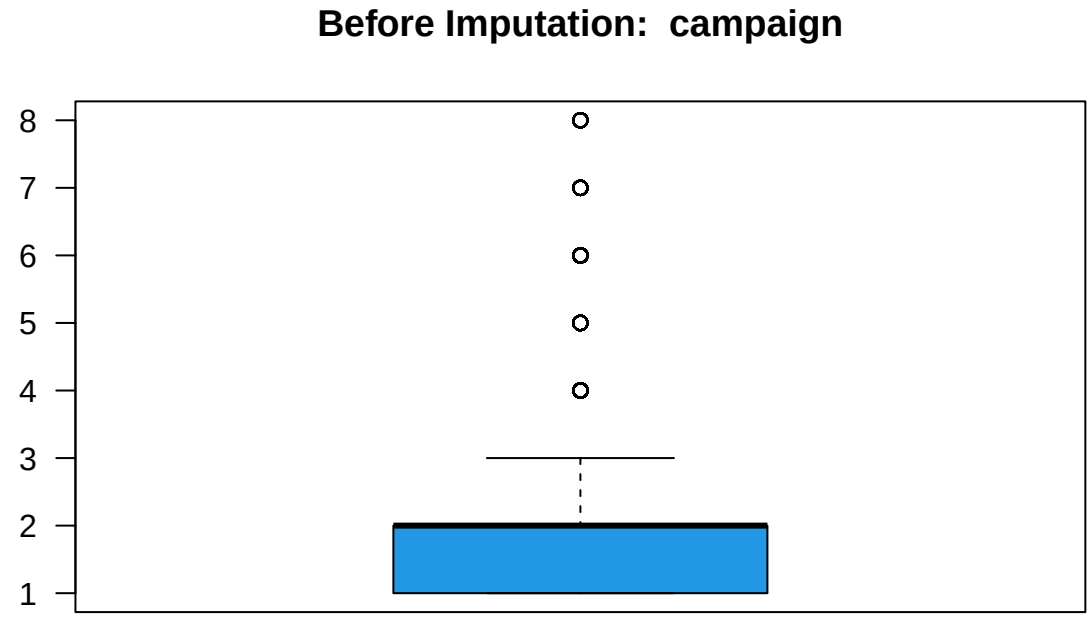
5.1 Imputation numerical

```
print("Abans d'imputació")
```

```
## [1] "Abans d'imputació"
```

```
numerical<-c("age","campaign")
plots(df,numerical,FALSE)
```






```
summary(df)
```

```
##       age                 job          marital                    education
##  Min.   :18.00   blue-collar:1213   divorced: 530   basic              :1623
##  1st Qu.:32.00   admin.      :1197   married :3098   high.school        :1156
##  Median :38.00   technician : 754   single  :1361   illiterate         :   2
##  Mean   :39.96   services   : 517   NA's    :  11   professional.course: 604
##  3rd Qu.:47.00   management : 380                   university.degree  :1387
##  Max.   :88.00   (Other)    : 884                   NA's               : 228
##                  NA's       :  55
##  housing       loan           contact         month      day_of_week
##  no  :2383   no  :4148   cellular :2007   may    :3164   fri: 849
##  yes :2487   yes : 722   telephone:2993   apr    : 442   mon:1107
##  NA's: 130   NA's: 130                    jul    : 407   thu:1029
##                                           jun    : 357   tue:1073
```

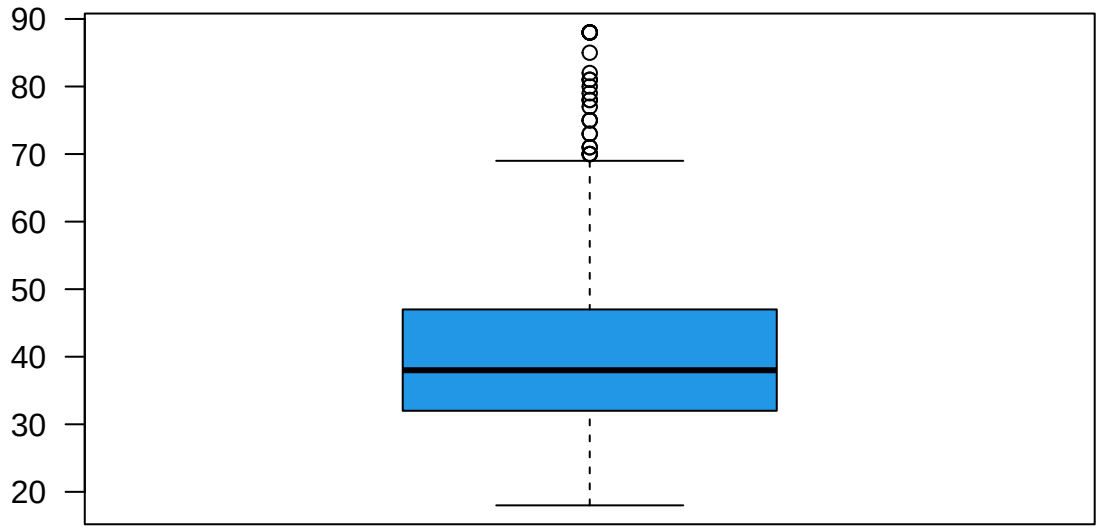
```
##                                          aug    : 271   wed: 942
##                                          nov    : 190
##                                          (Other): 169
##     duration         campaign         pdays          previous
##  Min.   :   4.0   Min.   :1.000   Min.   :  0.0   Min.   :0.0000
##  1st Qu.: 175.0   1st Qu.:1.000   1st Qu.:999.0   1st Qu.:0.0000
##  Median : 342.0   Median :2.000   Median :999.0   Median :0.0000
##  Mean   : 479.8   Mean   :2.006   Mean   :974.5   Mean   :0.0748
##  3rd Qu.: 686.0   3rd Qu.:2.000   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :4199.0   Max.   :8.000   Max.   :999.0   Max.   :3.0000
##                   NA's   :61
##         poutcome     emp.var.rate     cons.price.idx  cons.conf.idx
##  failure    : 224   Min.   :-2.9000   Min.   :92.76   Min.   :-50.00
##  nonexistent:4658   1st Qu.:-1.8000   1st Qu.:93.08   1st Qu.:-42.70
##  success    : 118   Median : 1.1000   Median :93.99   Median :-36.40
##                     Mean   : 0.3275   Mean   :93.68   Mean   :-39.81
##                     3rd Qu.: 1.1000   3rd Qu.:93.99   3rd Qu.:-36.40
##                     Max.   : 1.4000   Max.   :94.47   Max.   :-36.10
##
##    euribor3m      nr.employed     y
##  Min.   :1.244   Min.   :5076   no :2400
##  1st Qu.:1.811   1st Qu.:5099   yes:2600
##  Median :4.856   Median :5191
##  Mean   :3.965   Mean   :5174
##  3rd Qu.:4.857   3rd Qu.:5191
##  Max.   :5.045   Max.   :5228
##
```

```
print("Després d'imputació")
```

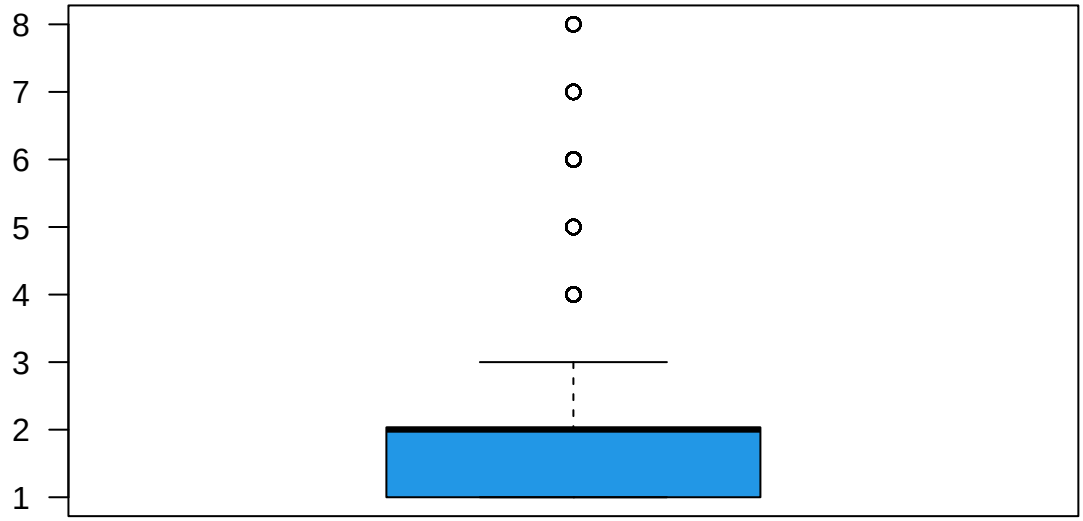
```
## [1] "Després d'imputació"
```

```
res.input<-imputePCA(df[,numerical],ncp=1)
df[,numerical]<-res.input$completeObs
plots(df,numerical,TRUE)
```

After Imputation: age



After Imputation: campaign




```
summary(df)
```

```
##       age                 job           marital                    education
##  Min.   :18.00   blue-collar:1213   divorced: 530   basic              :1623
##  1st Qu.:32.00   admin.      :1197   married :3098   high.school        :1156
##  Median :38.00   technician : 754   single  :1361   illiterate         :   2
##  Mean   :39.96   services   : 517   NA's    :  11   professional.course: 604
##  3rd Qu.:47.00   management : 380                   university.degree  :1387
##  Max.   :88.00   (Other)    : 884                   NA's               : 228
##                  NA's       :  55
##  housing       loan           contact         month      day_of_week
##  no  :2383   no  :4148   cellular :2007   may    :3164   fri: 849
##  yes :2487   yes : 722   telephone:2993   apr    : 442   mon:1107
##  NA's: 130   NA's: 130                    jul    : 407   thu:1029
##                                           jun    : 357   tue:1073
##                                           aug    : 271   wed: 942
##                                           nov    : 190
##                                           (Other): 169
##     duration         campaign         pdays          previous
##  Min.   :   4.0   Min.   :1.000   Min.   :  0.0   Min.   :0.0000
##  1st Qu.: 175.0   1st Qu.:1.000   1st Qu.:999.0   1st Qu.:0.0000
##  Median : 342.0   Median :2.000   Median :999.0   Median :0.0000
##  Mean   : 479.8   Mean   :2.006   Mean   :974.5   Mean   :0.0748
##  3rd Qu.: 686.0   3rd Qu.:2.033   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :4199.0   Max.   :8.000   Max.   :999.0   Max.   :3.0000
##
##        poutcome     emp.var.rate     cons.price.idx  cons.conf.idx
##  failure    : 224   Min.   :-2.9000   Min.   :92.76   Min.   :-50.00
##  nonexistent:4658   1st Qu.:-1.8000   1st Qu.:93.08   1st Qu.:-42.70
##  success    : 118   Median : 1.1000   Median :93.99   Median :-36.40
##                     Mean   : 0.3275   Mean   :93.68   Mean   :-39.81
##                     3rd Qu.: 1.1000   3rd Qu.:93.99   3rd Qu.:-36.40
##                     Max.   : 1.4000   Max.   :94.47   Max.   :-36.10
##
```

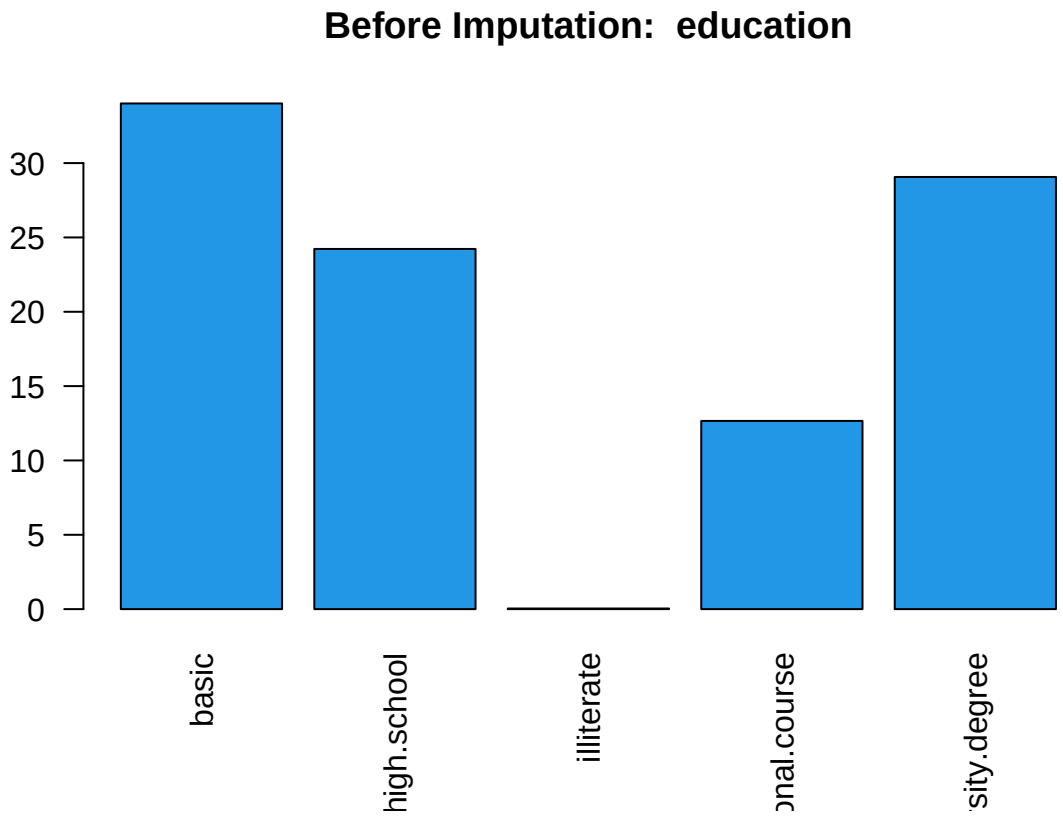
```
##    euribor3m      nr.employed     y
##  Min.   :1.244   Min.   :5076   no :2400
##  1st Qu.:1.811   1st Qu.:5099   yes:2600
##  Median :4.856   Median :5191
##  Mean   :3.965   Mean   :5174
##  3rd Qu.:4.857   3rd Qu.:5191
##  Max.   :5.045   Max.   :5228
##
```

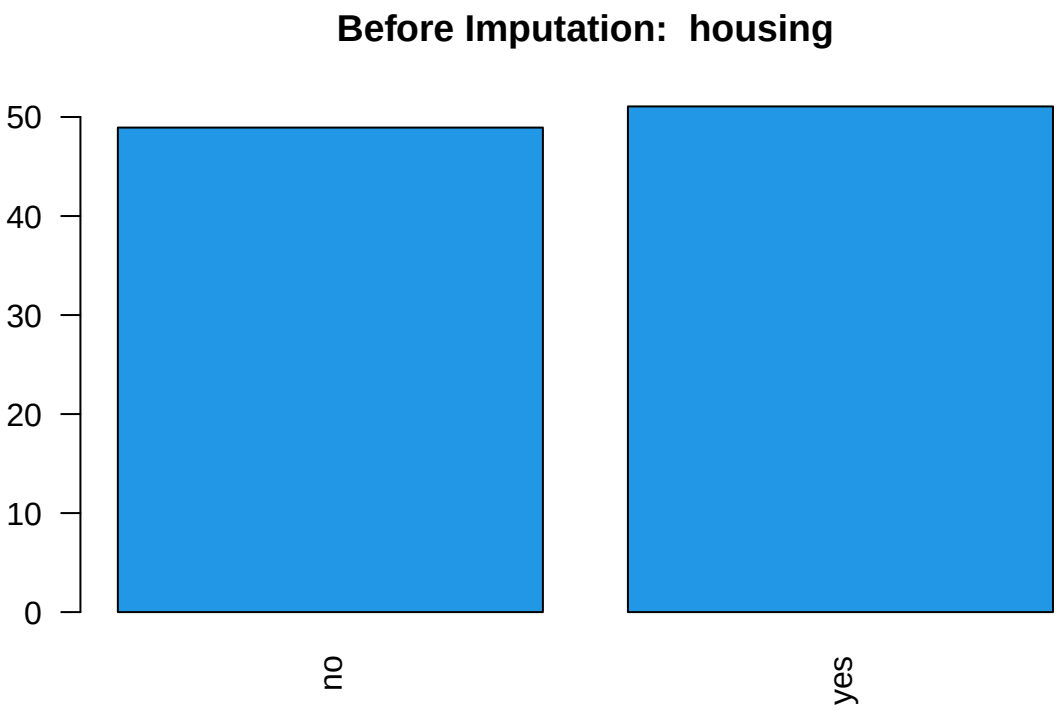
5.2 Imputation categorical

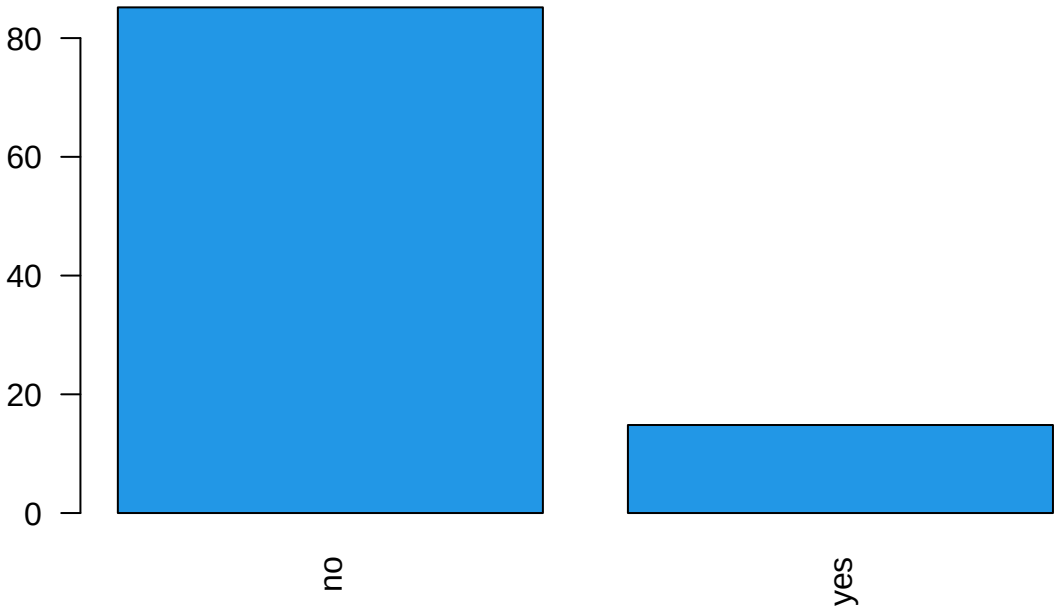
```
print("Abans d'imputació")
```

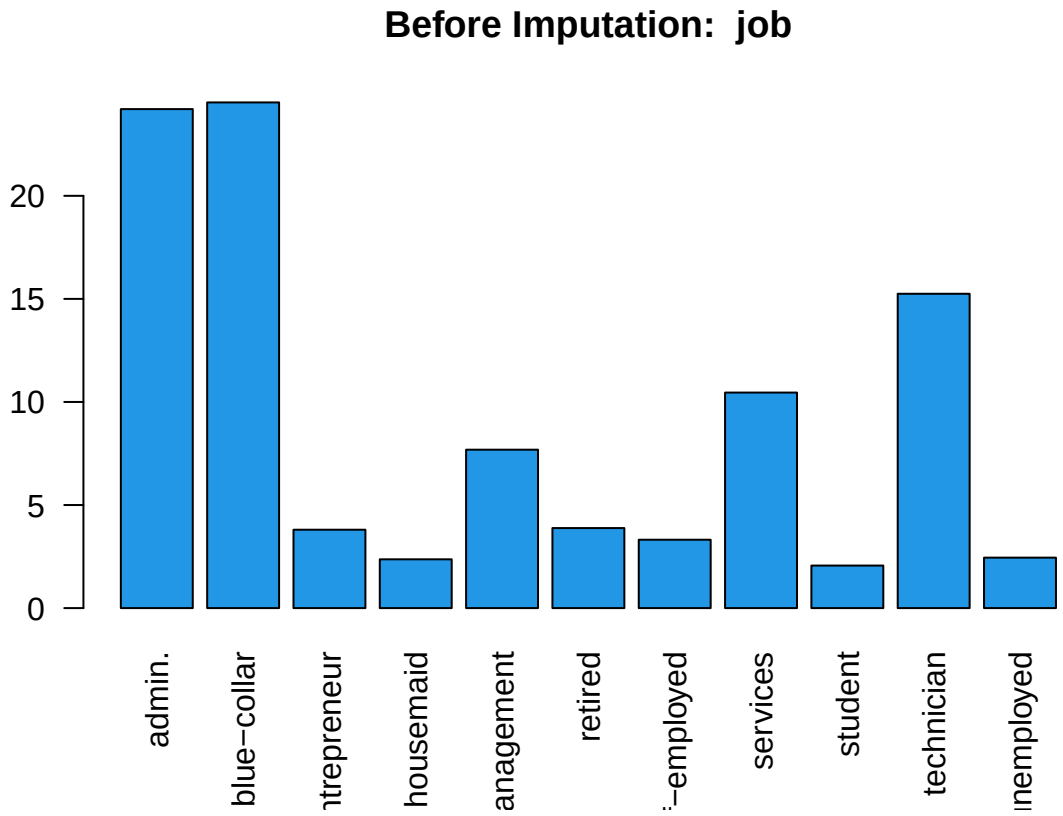
```
## [1] "Abans d'imputació"
```

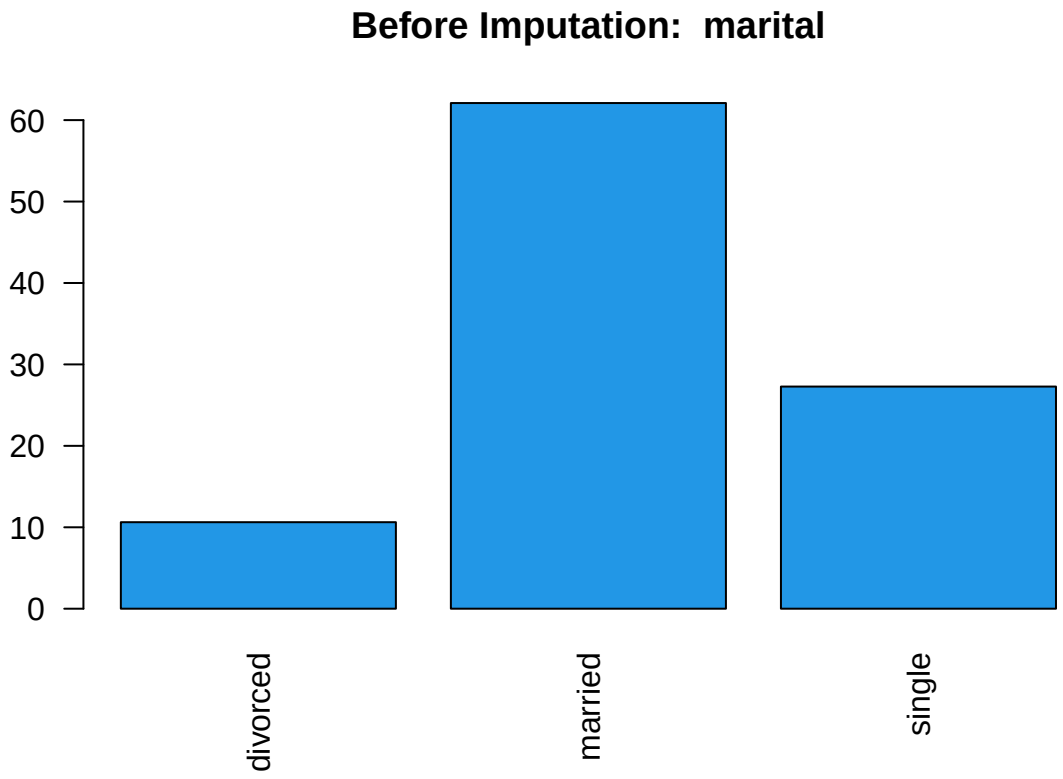
```
categorical<-c("education","housing","loan","job","marital")
plotscat(df,categorical,FALSE)
```




Before Imputation: housing
50
40
30
20
10
0
no
yes


Before Imputation: loan
80
60
40
20
0
no
yes






```
summary(df)
```

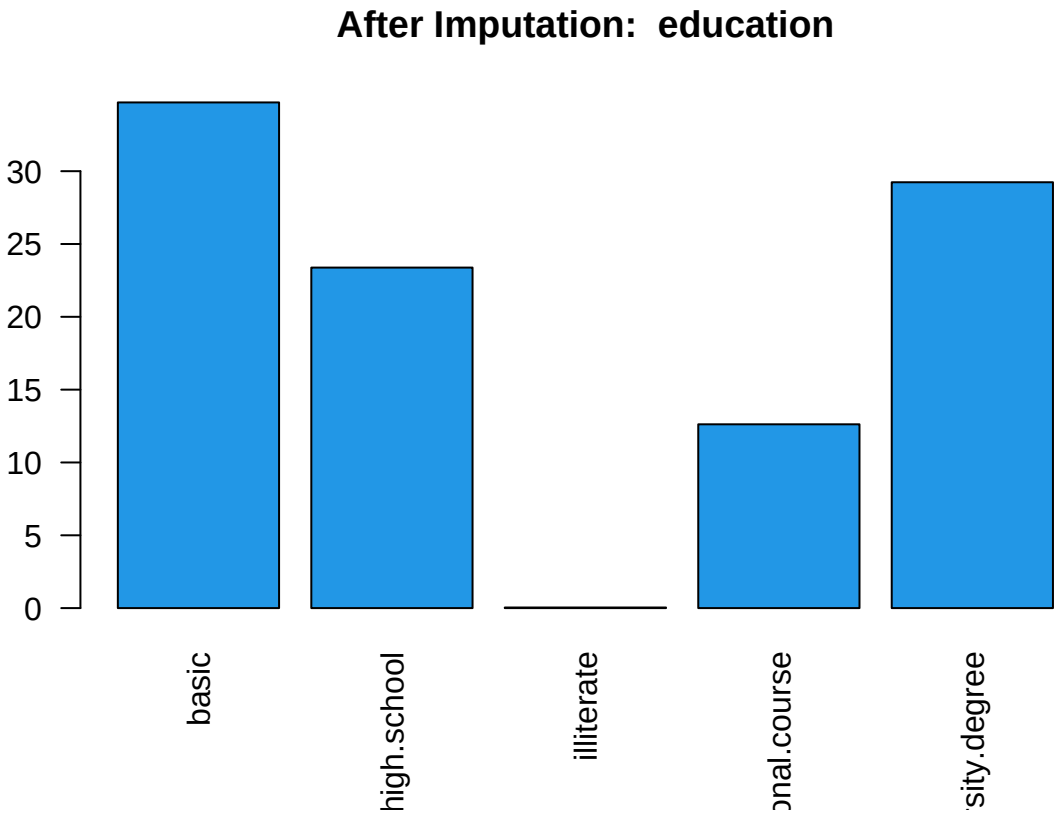
```
##       age                job          marital                    education
##  Min.   :18.00   blue-collar:1213   divorced: 530   basic              :1623
```

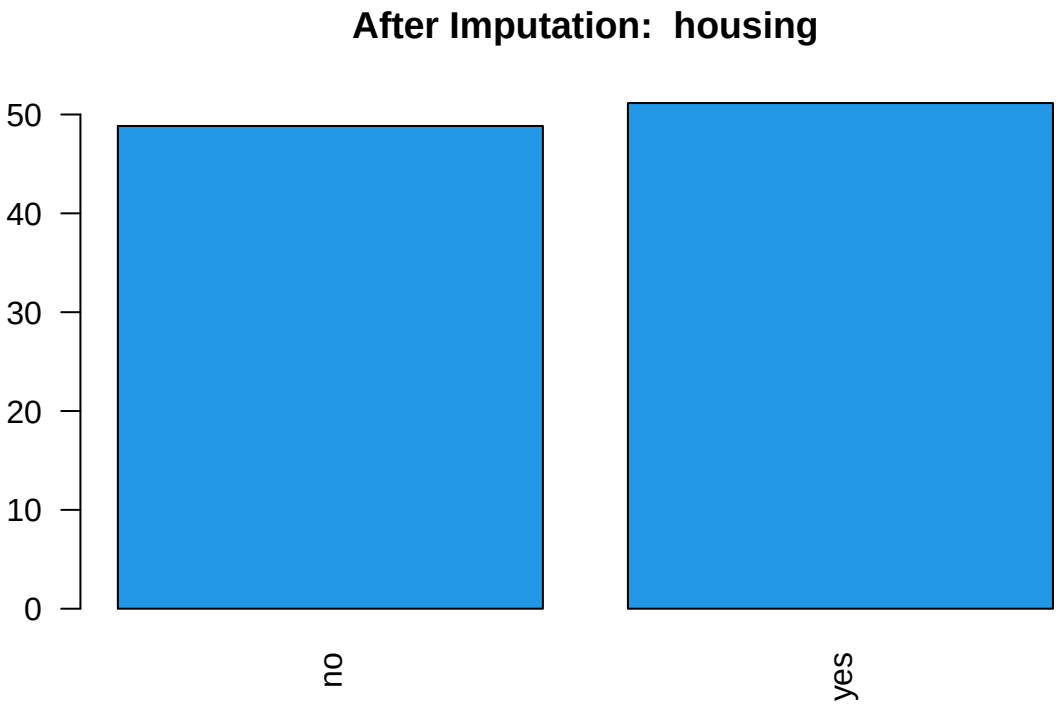
```
##  1st Qu.:32.00   admin.     :1197   married :3098   high.school        :1156
##  Median :38.00   technician : 754   single  :1361   illiterate         :   2
##  Mean   :39.96   services   : 517   NA's    :  11   professional.course: 604
##  3rd Qu.:47.00   management : 380                    university.degree  :1387
##  Max.   :88.00   (Other)    : 884                    NA's               : 228
##                  NA's       :  55
##  housing       loan           contact         month      day_of_week
##  no  :2383   no  :4148   cellular :2007   may    :3164   fri: 849
##  yes :2487   yes : 722   telephone:2993   apr    : 442   mon:1107
##  NA's: 130   NA's: 130                    jul    : 407   thu:1029
##                                           jun    : 357   tue:1073
##                                           aug    : 271   wed: 942
##                                           nov    : 190
##                                           (Other): 169
##     duration         campaign         pdays          previous
##  Min.   :   4.0   Min.   :1.000   Min.   :  0.0   Min.   :0.0000
##  1st Qu.: 175.0   1st Qu.:1.000   1st Qu.:999.0   1st Qu.:0.0000
##  Median : 342.0   Median :2.000   Median :999.0   Median :0.0000
##  Mean   : 479.8   Mean   :2.006   Mean   :974.5   Mean   :0.0748
##  3rd Qu.: 686.0   3rd Qu.:2.033   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :4199.0   Max.   :8.000   Max.   :999.0   Max.   :3.0000
##
##         poutcome     emp.var.rate     cons.price.idx  cons.conf.idx
##  failure    : 224   Min.   :-2.9000   Min.   :92.76   Min.   :-50.00
##  nonexistent:4658   1st Qu.:-1.8000   1st Qu.:93.08   1st Qu.:-42.70
##  success    : 118   Median : 1.1000   Median :93.99   Median :-36.40
##                     Mean   : 0.3275   Mean   :93.68   Mean   :-39.81
##                     3rd Qu.: 1.1000   3rd Qu.:93.99   3rd Qu.:-36.40
##                     Max.   : 1.4000   Max.   :94.47   Max.   :-36.10
##
##    euribor3m      nr.employed     y
##  Min.   :1.244   Min.   :5076   no :2400
##  1st Qu.:1.811   1st Qu.:5099   yes:2600
##  Median :4.856   Median :5191
##  Mean   :3.965   Mean   :5174
##  3rd Qu.:4.857   3rd Qu.:5191
##  Max.   :5.045   Max.   :5228
##
```

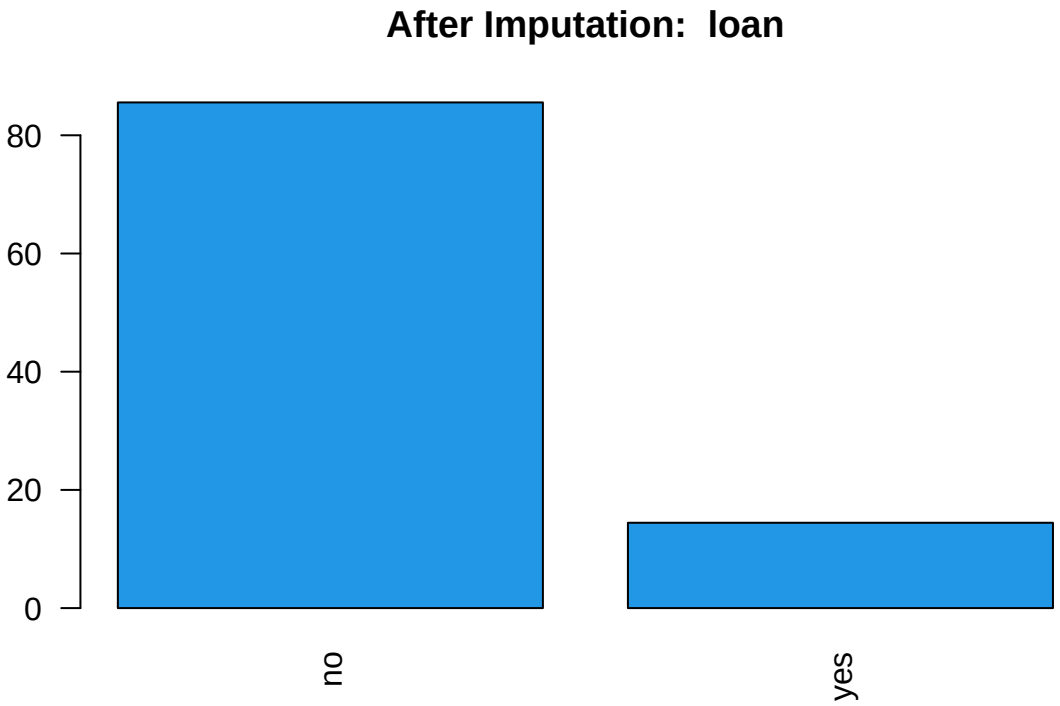
```
print("Després d'imputació")
```

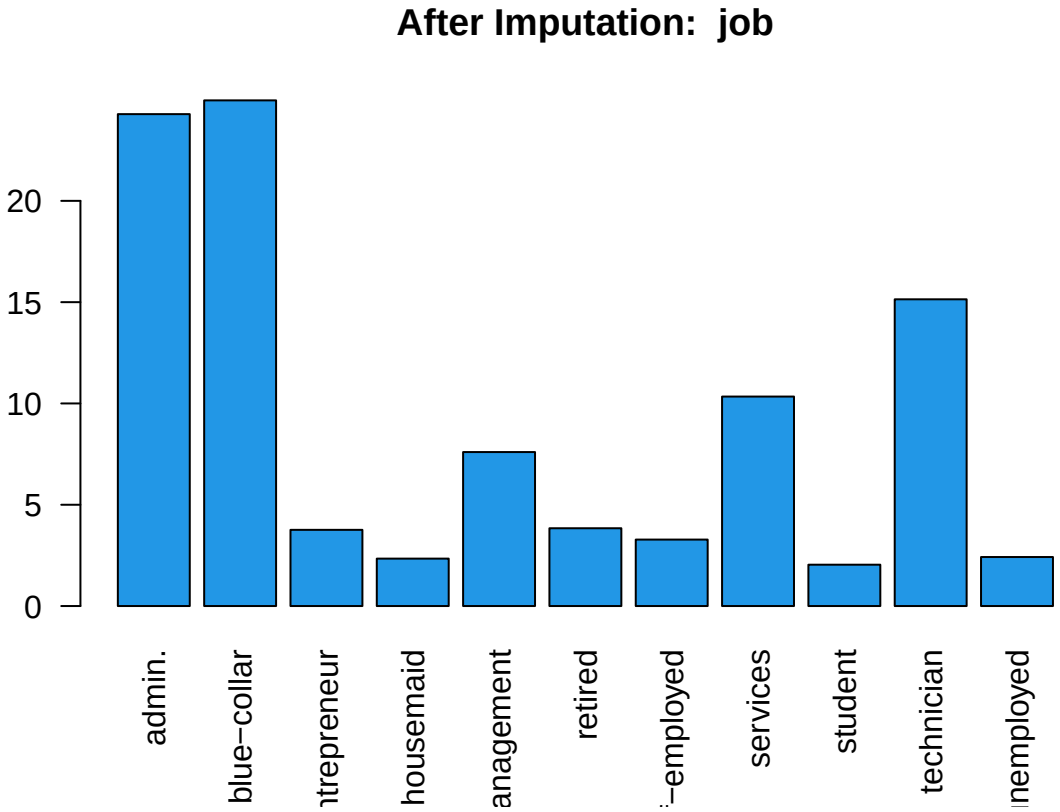
```
## [1] "Després d'imputació"
```

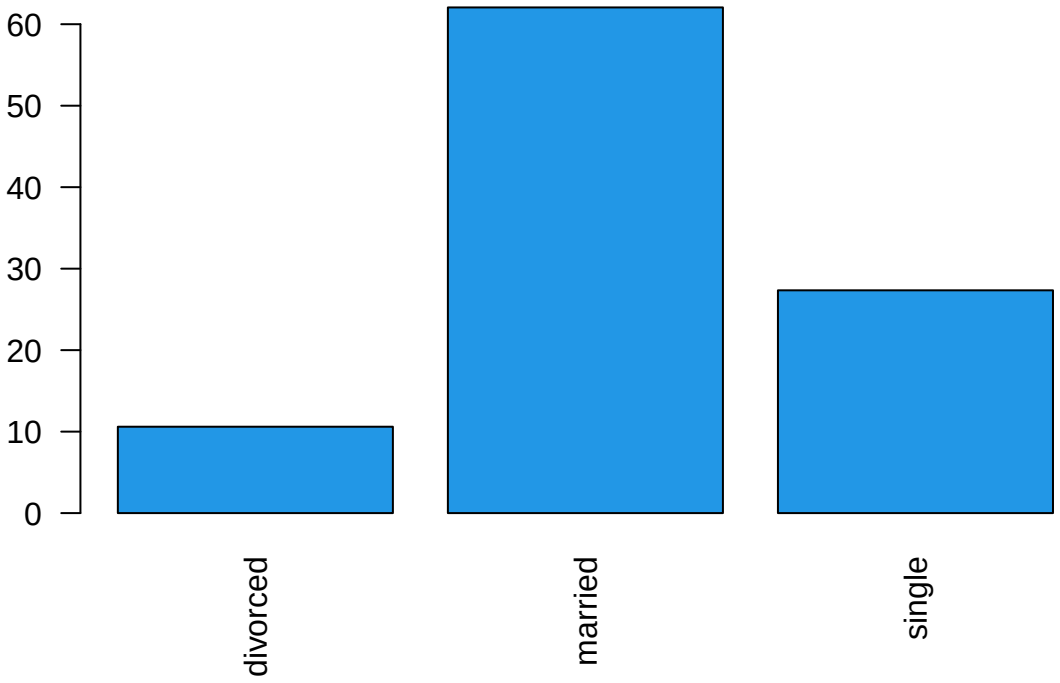
```
res.input<-imputeMCA(df[,categorical],method="EM")
df[,categorical]<-res.input$completeObs
plotscat(df,categorical,TRUE)
```


After Imputation: education
30
25
20
15
10
5
0
basic
high.school
illiterate
onal.course
sity.degree


After Imputation: housing
50
40
30
20
10
0
no
yes


After Imputation: loan
80
60
40
20
0
no
yes


After Imputation: job
20
15
10
5
0
admin.
blue−collar
housemaid
retired
services
student
technician




```
summary(df)
```

```
##       age                job          marital                    education
##  Min.   :18.00   blue-collar:1248   divorced: 530   basic              :1736
##  1st Qu.:32.00   admin.     :1214   married :3103   high.school        :1169
##  Median :38.00   technician : 757   single  :1367   illiterate         :   2
##  Mean   :39.96   services   : 517                   professional.course: 631
##  3rd Qu.:47.00   management : 380                   university.degree  :1462
##  Max.   :88.00   retired    : 192
##                  (Other)    : 692
##  housing     loan           contact         month      day_of_week
##  no :2442   no :4278   cellular :2007   may    :3164   fri: 849
##  yes:2558   yes: 722   telephone:2993   apr    : 442   mon:1107
##                                         jul    : 407   thu:1029
##                                         jun    : 357   tue:1073
##                                         aug    : 271   wed: 942
##                                         nov    : 190
##                                         (Other): 169
##     duration         campaign          pdays          previous
##  Min.   :   4.0   Min.   :1.000   Min.   :  0.0   Min.   :0.0000
##  1st Qu.: 175.0   1st Qu.:1.000   1st Qu.:999.0   1st Qu.:0.0000
##  Median : 342.0   Median :2.000   Median :999.0   Median :0.0000
##  Mean   : 479.8   Mean   :2.006   Mean   :974.5   Mean   :0.0748
##  3rd Qu.: 686.0   3rd Qu.:2.033   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :4199.0   Max.   :8.000   Max.   :999.0   Max.   :3.0000
##
##         poutcome     emp.var.rate      cons.price.idx  cons.conf.idx
##  failure    : 224   Min.   :-2.9000   Min.   :92.76   Min.   :-50.00
##  nonexistent:4658   1st Qu.:-1.8000   1st Qu.:93.08   1st Qu.:-42.70
##  success    : 118   Median : 1.1000   Median :93.99   Median :-36.40
```

```
##                     Mean   : 0.3275   Mean   :93.68   Mean   :-39.81
##                     3rd Qu.: 1.1000   3rd Qu.:93.99   3rd Qu.:-36.40
##                     Max.   : 1.4000   Max.   :94.47   Max.   :-36.10
##
##    euribor3m      nr.employed     y
##  Min.   :1.244   Min.   :5076   no :2400
##  1st Qu.:1.811   1st Qu.:5099   yes:2600
##  Median :4.856   Median :5191
##  Mean   :3.965   Mean   :5174
##  3rd Qu.:4.857   3rd Qu.:5191
##  Max.   :5.045   Max.   :5228
##
```

6 Discretizatio

6.1 Age discretization

We are going to discretize the age, the first section will be kids from ages 0-10, then teenager from 10-20, afterwards there will be the young adults that goes 20-30 and the adults from 30-50, finally we will have the elderly which will be 60 and above.

```
df$Age_group[df$age>=0 & df$age<10]<-"0-10"
df$Age_group[df$age>=10 & df$age<20]<-"10-20"
df$Age_group[df$age>=20 & df$age<30]<-"20-30"
df$Age_group[df$age>=30 & df$age<50]<-"30-50"
df$Age_group[df$age>=50 & df$age<60]<-"40-60"
df$Age_group[df$age>=60 & df$age<60]<-">=60"
df$Age_group<-as.factor(df$Age_group)

head(df)
```

```
##       age         job marital           education housing loan   contact month
## 15936  41      admin. married   university.degree      no   no  cellular   jul
## 34002  35 blue-collar married               basic      no   no telephone   may
## 20259  30  technician  single   university.degree     yes   no  cellular   aug
## 28435  29 blue-collar married               basic     yes   no  cellular   apr
## 12984  30 blue-collar married               basic      no   no  cellular   jul
## 35830  40  technician  single professional.course     yes   no  cellular   may
##       day_of_week duration campaign pdays previous    poutcome emp.var.rate
## 15936         mon     1360        3   999        0 nonexistent          1.4
## 34002         wed      622        3   999        0 nonexistent         -1.8
## 20259         mon      720        1   999        0 nonexistent          1.4
## 28435         thu     1042        2   999        0 nonexistent         -1.8
## 12984         tue      623        2   999        0 nonexistent          1.4
## 35830         fri      317        1   999        1     failure         -1.8
##       cons.price.idx cons.conf.idx euribor3m nr.employed   y Age_group
## 15936         93.918         -42.7     4.960      5228.1 yes     30-50
## 34002         92.893         -46.2     1.281      5099.1 yes     30-50
## 20259         93.444         -36.1     4.965      5228.1 yes     30-50
## 28435         93.075         -47.1     1.435      5099.1 yes     20-30
## 12984         93.918         -42.7     4.962      5228.1 yes     30-50
## 35830         92.893         -46.2     1.259      5099.1 yes     30-50
```

6.2 Caompaign discretization

We are going to discretize the campaing variable, for people who have been contacted for 0 to 5 times, they are considered contacted Infrequently, from 5 to 10 they are considered frequent and more than that very frequent.

```
  df$Campaign_contacts[df$campaign>=0 & df$campaign<5]<-"Infrequent"
df$Campaign_contacts[df$campaign>=5 & df$campaign<=10]<-"Frequent"
df$Campaign_contacts[df$campaign>=10]<-"Very frequent"

df$Campaign_contacts<-as.factor(df$Campaign_contacts)

head(df)
```

```
##       age         job marital           education housing loan   contact month
## 15936  41      admin. married   university.degree      no   no  cellular   jul
## 34002  35 blue-collar married               basic      no   no telephone   may
## 20259  30  technician  single   university.degree     yes   no  cellular   aug
## 28435  29 blue-collar married               basic     yes   no  cellular   apr
## 12984  30 blue-collar married               basic      no   no  cellular   jul
## 35830  40  technician  single professional.course     yes   no  cellular   may
##       day_of_week duration campaign pdays previous    poutcome emp.var.rate
## 15936         mon     1360        3   999        0 nonexistent          1.4
## 34002         wed      622        3   999        0 nonexistent         -1.8
## 20259         mon      720        1   999        0 nonexistent          1.4
## 28435         thu     1042        2   999        0 nonexistent         -1.8
## 12984         tue      623        2   999        0 nonexistent          1.4
## 35830         fri      317        1   999        1     failure         -1.8
##       cons.price.idx cons.conf.idx euribor3m nr.employed   y Age_group
## 15936         93.918         -42.7     4.960      5228.1 yes     30-50
## 34002         92.893         -46.2     1.281      5099.1 yes     30-50
## 20259         93.444         -36.1     4.965      5228.1 yes     30-50
## 28435         93.075         -47.1     1.435      5099.1 yes     20-30
## 12984         93.918         -42.7     4.962      5228.1 yes     30-50
## 35830         92.893         -46.2     1.259      5099.1 yes     30-50
##       Campaign_contacts
## 15936        Infrequent
## 34002        Infrequent
## 20259        Infrequent
## 28435        Infrequent
## 12984        Infrequent
## 35830        Infrequent
```

7 Per variable

7.1 Number of missing values,outliers and errors

7.1.1 Ranking per variable by missings and errors

```
miss_errs<-missings$missing+errors
missings<-cbind.data.frame(missings,miss_errs)
print(missings[order(missings$miss_errs, decreasing = T), ]   )
```

```
##             names missing miss_errs
## 5         default    1183      1183
```

```
## 4       education     228       228
## 13          pdays       0       219
## 14       previous       0       219
## 6         housing     130       130
## 7            loan     130       130
## 2             job      55        55
## 3         marital      11        11
## 1             age       0         0
## 8         contact       0         0
## 9           month       0         0
## 10    day_of_week       0         0
## 11       duration       0         0
## 12       campaign       0         0
## 15       poutcome       0         0
## 16   emp.var.rate       0         0
## 17 cons.price.idx       0         0
## 18  cons.conf.idx       0         0
## 19      euribor3m       0         0
## 20    nr.employed       0         0
## 21              y       0         0
```

8 Per individual

8.1 Missing

```
print(sum(indivMiss))
```

```
## [1] 1737
```

8.2 Outlier

```
print(sum(indivOut))
```

```
## [1] 84
```

8.3 Errors

```
print(sum(indivErrs))
```

```
## [1] 219
```

8.4 Create variable adding the total number missing values, outliers and errors

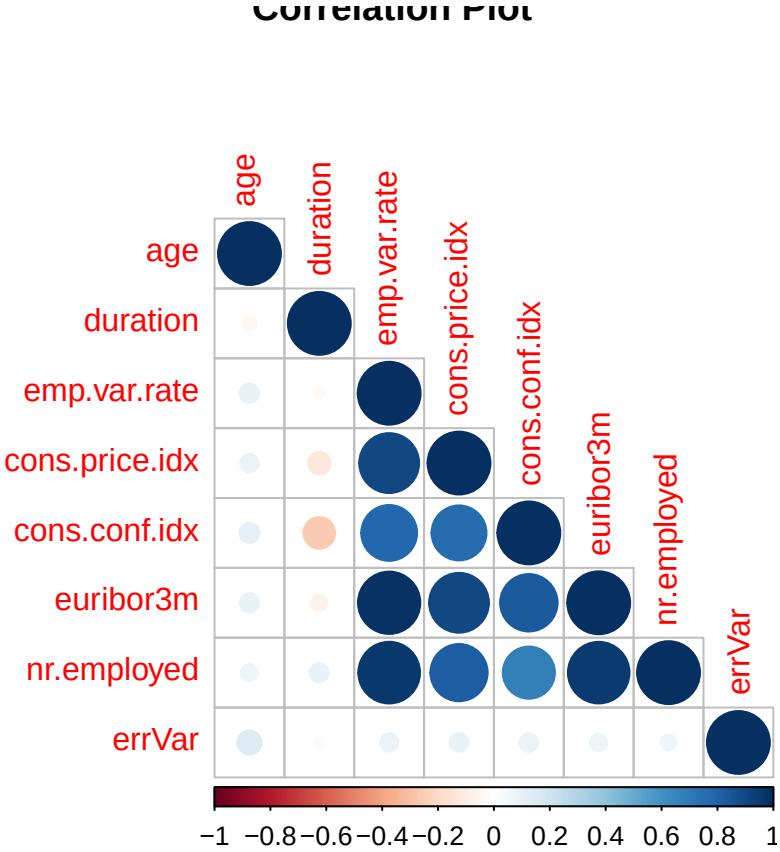
```
errVar<-indivMiss+indivOut+indivErrs
df<-cbind.data.frame(df,errVar)
head(df)
```

```
##       age         job marital           education housing loan   contact month
## 15936  41      admin. married   university.degree      no   no  cellular   jul
## 34002  35 blue-collar married               basic      no   no telephone   may
## 20259  30  technician  single   university.degree     yes   no  cellular   aug
## 28435  29 blue-collar married               basic     yes   no  cellular   apr
## 12984  30 blue-collar married               basic      no   no  cellular   jul
```

```
## 35830  40  technician  single professional.course     yes   no  cellular   may
##       day_of_week duration campaign pdays previous    poutcome emp.var.rate
## 15936         mon     1360        3   999        0 nonexistent          1.4
## 34002         wed      622        3   999        0 nonexistent         -1.8
## 20259         mon      720        1   999        0 nonexistent          1.4
## 28435         thu     1042        2   999        0 nonexistent         -1.8
## 12984         tue      623        2   999        0 nonexistent          1.4
## 35830         fri      317        1   999        1     failure         -1.8
##       cons.price.idx cons.conf.idx euribor3m nr.employed   y Age_group
## 15936         93.918         -42.7     4.960      5228.1 yes     30-50
## 34002         92.893         -46.2     1.281      5099.1 yes     30-50
## 20259         93.444         -36.1     4.965      5228.1 yes     30-50
## 28435         93.075         -47.1     1.435      5099.1 yes     20-30
## 12984         93.918         -42.7     4.962      5228.1 yes     30-50
## 35830         92.893         -46.2     1.259      5099.1 yes     30-50
##       Campaign_contacts errVar
## 15936        Infrequent      0
## 34002        Infrequent      0
## 20259        Infrequent      0
## 28435        Infrequent      0
## 12984        Infrequent      0
## 35830        Infrequent      1
```

8.5 Describe these variables, to which other variables exist higher associations.

```
corrplot(cor(df[,c("age","duration","emp.var.rate","cons.price.idx","cons.conf.idx","euribor3m",
                "nr.employed","errVar")]),title="Correlation Plot",type="lower")
```



* We can see interesting things here: * As the consumer confidence decreases, the duration of the call increases and otherwise. Seems coherent as the confidence is greater, the client will agree or disagree with the caller faster, and if the confidence is lower the client will be more cautious with the conditions and more questions will be asked. * The positive and strong correlations between number of employed and employment variation seems ovbious. * As the euribor increase, the employment variation also increases since a higher euribor rate implies job losses. * Another obvious strong and positive correlation is between IPC and euribor, the higher the the euribor rates the more expensive everthing will be.

8.6 Groups and its means

We will see 3 differents groups from the education variable and how it changes the mean on wether they are uneducated,educated or highly educated.

```
group1<-df[df$education=="illiterate","errVar"]
mean1<-sum(group1)/length(group1)
mean1
```

```
## [1] 0
```

```
group2<-df[df$education=="high.school","errVar"]
mean2<-sum(group2)/length(group2)
mean2
```

```
## [1] 0.2985458
```

```
group3<-df[df$education=="university.degree","errVar"]
mean3<-sum(group3)/length(group3)
mean3
```

```
## [1] 0.3194254
```

```
cols<-c("illitarate","high.school","university.degree")
means<-c(mean1,mean2,mean3)
groups<-cbind.data.frame(cols,means)
groups
```

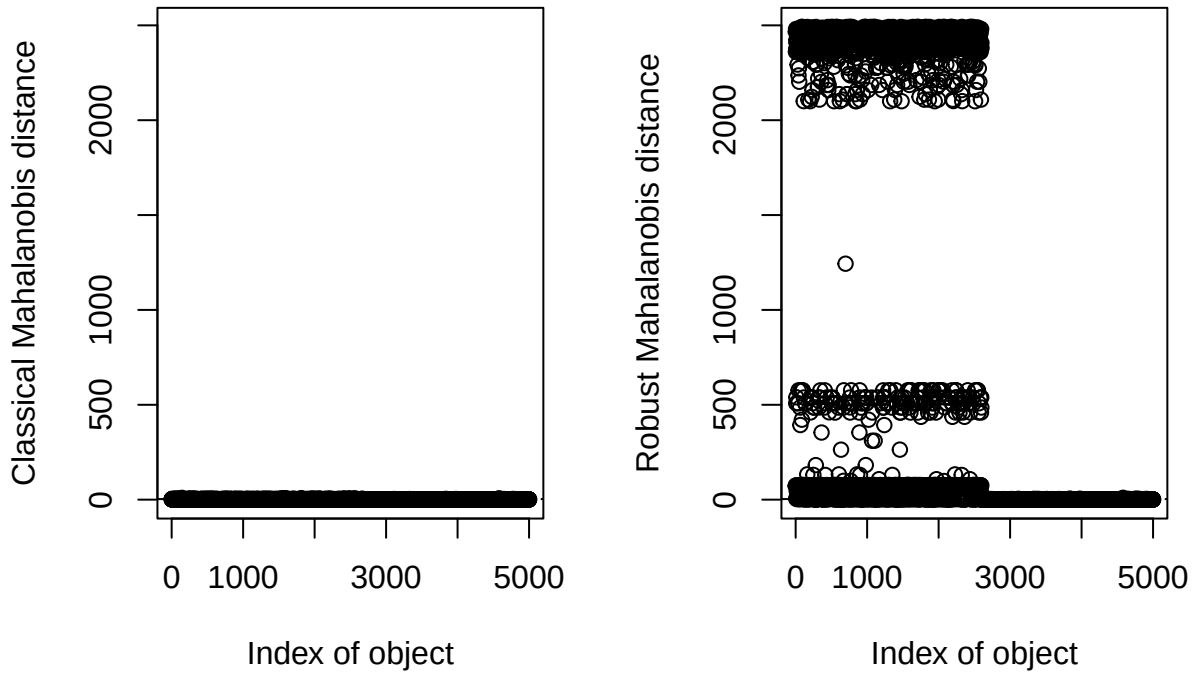
```
##                cols     means
## 1        illitarate 0.0000000
## 2       high.school 0.2985458
## 3 university.degree 0.3194254
```

- We see that for illitarate people, the are consistently inside the common data and no errors or incosistencies occurs, this might indicate as more uneducated someone is, the less power of negotation they have, so it's less prone that in the database might occur something weird.
- There isn't much difference between the educated and highly educated people, they fit the same way in the data structure and few errors or incosistencies occur in them.

8.7 Multivarite outliers

We only chose 3 variable because with the others the following error is produced: "Warning in covMcd(X) : The 2502-th order statistic of the absolute deviation of variable 3 is zero. There are 2640 observations (in the entire dataset of 5000 obs.) lying on the plane with equation 0 (x_i1-m_1) + 0 (x_i2-m_2) + 1 (x_i3-m_3) = 0 with (m_1,m_2) the mean of these observations. Warning in sqrt(mahalanobis(X, X.mcd$center$, $X.mcd$cov)). NaNs produced", and the execution is halted.

```
mout<-Moutlier(df[,c("age","duration","euribor3m")],quantile =0.975, plot = TRUE)
```



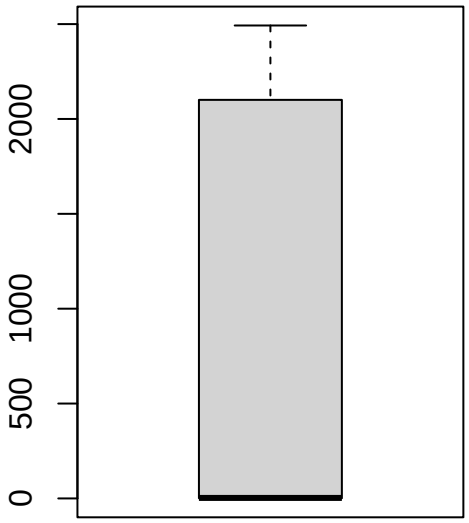

```
# Classical: Assumption of normality on the underlying generating mechanism
# Robust: Median and absolute median deviations -> Not normal generating mechanism

length(which(mout$rd>mout$cutoff))
```

```
## [1] 2446
```

```
ll<-which(mout$rd>mout$cutoff)
boxplot(mout$rd)
df$mout <- 0
df$mout[ ll ]<-1
df$mout <- factor( df$mout, labels=c( "NoMOut","YesMOut"))
table(df$mout)
```

```
##
##  NoMOut YesMOut
##    2554    2446
```



9 Profiling

9.1 Duration

```
i<-findIndex("duration",colnames)
res.condes<-condes(df,i, proba=0.05)
```

```
res.condes$quanti  # Global association to numeric variables
```

```
##                 correlation      p.value
## nr.employed      0.11290678 1.172220e-15
## emp.var.rate     0.10542465 7.801900e-14
## duration         0.09372579 3.128793e-11
## euribor3m        0.08023345 1.338285e-08
## cons.price.idx   0.07072555 5.555230e-07
## pdays            0.04108769 3.662701e-03
## age              0.02806418 4.721889e-02
## previous        -0.03463980 1.430442e-02
```

- We can see that all numerical variables have a relationship with the numerical target feature duration, since all p-values are lesser than 0.05. So we can conclude that all the quantitative variable are being affected by the duration. All the variables show except previous, the greater their value the greater the duration will be, on the other hand, the greater the duration the lower previous value will be.

```
res.condes$quali # Global association to factors
```

```
##                            R2      p.value
```

```
## Campaign_contacts 0.516762248 0.000000e+00
## month             0.033180385 2.641683e-32
## day_of_week       0.010767255 5.051624e-11
## poutcome          0.002085205 5.432666e-03
## y                 0.001188048 1.479430e-02
## marital           0.001678997 1.501811e-02
## mout              0.001061058 2.125875e-02
```

```
res.condes$category
```

```
##                                  Estimate      p.value
## Campaign_contacts=Frequent     2.01750577 0.000000e+00
## month=jul                      0.66265937 2.640994e-14
## month=aug                      0.67847145 2.573792e-10
## day_of_week=mon                0.19506466 1.550922e-08
## poutcome=nonexistent           0.15828242 8.060483e-03
## y=yes                          0.04525886 1.479430e-02
## mout=YesMOut                   0.04274741 2.125875e-02
## job=retired                    0.21412933 2.822288e-02
## marital=divorced               0.09168584 4.625455e-02
## mout=NoMOut                   -0.04274741 2.125875e-02
## y=no                          -0.04525886 1.479430e-02
## marital=single                -0.09062562 1.326356e-02
## day_of_week=thu               -0.09474721 1.048736e-02
## poutcome=success              -0.21688507 2.425109e-03
## month=nov                     -0.25880983 1.555199e-06
## month=oct                     -0.81664013 5.891712e-07
## day_of_week=tue               -0.18359656 2.800299e-07
## month=apr                     -0.12642666 1.129388e-07
## Campaign_contacts=Infrequent -2.01750577 0.000000e+00
```

- Since all p-values are lesser than 0.05, we can conclude that all leves of the categorical variables have and impact on its result, that is, depending on the categorical level of a variable, the duration of a call will be bigger or smaller depending on all the factors,

9.2 Yes

```
colnames<-colnames(df)
i<-findIndex("y",colnames)
res.catdes<-catdes(df, i , proba = 0.05)
```

```
res.catdes$quanti.var  # Global association to numeric variables
```

```
##                       Eta2       P-value
## duration       0.296307368  0.000000e+00
## emp.var.rate   0.295636408  0.000000e+00
## cons.price.idx 0.397806228  0.000000e+00
## cons.conf.idx  0.550214817  0.000000e+00
## euribor3m      0.316976151  0.000000e+00
## nr.employed    0.123968952 7.269081e-146
## previous       0.062674047  2.555064e-72
## pdays          0.023280131  1.998680e-27
## age            0.010754180  1.969403e-13
## errVar         0.008051012  2.071767e-10
## campaign       0.001188048  1.479430e-02
```

- We see that for all the numerical variables the target y has categorical effect since p-value for all of them are <0.05. So we can conclude that the

```
res.catdes$quanti # Partial association of numeric variables to levels of outcome factor
```

```
## $no
##                    v.test Mean in category Overall mean sd in category
## cons.conf.idx   52.445437       -36.400000   -39.812260   0.000000e+00
## cons.price.idx  44.594095        93.994000    93.683170   0.000000e+00
## euribor3m       39.806580         4.856075     3.965195   8.671073e-04
## emp.var.rate    38.443288         1.100000     0.327460   0.000000e+00
## nr.employed     24.894192      5191.000000  5173.694300   0.000000e+00
## pdays           10.787835       999.000000   974.545000   0.000000e+00
## age              7.332131        41.045833    39.960200   8.881558e+00
## errVar           6.344053         0.468750     0.408000   6.897029e-01
## campaign        -2.437017         1.958846     2.005916   1.234057e+00
## previous       -17.700496         0.000000     0.074800   0.000000e+00
## duration       -38.486888       245.359167   479.758800   2.149088e+02
##                   Overall sd       p.value
## cons.conf.idx    4.4197217  0.000000e+00
## cons.price.idx   0.4734852  0.000000e+00
## euribor3m        1.5202845  0.000000e+00
## emp.var.rate     1.3650890  0.000000e+00
## nr.employed     47.2227912 8.599724e-137
## pdays          153.9904918  3.929392e-27
## age             10.0580424  2.265212e-13
## errVar           0.6504890  2.237976e-10
## campaign         1.3120151  1.480898e-02
## previous         0.2870626  4.156304e-70
## duration       413.7182218  0.000000e+00
##
## $yes
##                    v.test Mean in category Overall mean sd in category
## duration        38.486888      696.1276923   479.758800    434.7342361
## previous        17.700496        0.1438462     0.074800      0.3854076
## campaign         2.437017        2.0493642     2.005916      1.3786481
## errVar          -6.344053        0.3519231     0.408000      0.6066904
## age             -7.332131       38.9580769    39.960200     10.9380147
## pdays          -10.787835      951.9711538   974.545000    211.0460584
## nr.employed    -24.894192     5157.7198077  5173.694300     61.2928380
## emp.var.rate   -38.443288       -0.3856538     0.327460      1.5887580
## euribor3m      -39.806580        3.1428442     3.965195      1.7423728
## cons.price.idx -44.594095       93.3962496    93.683170      0.5095336
## cons.conf.idx  -52.445437      -42.9620385   -39.812260      4.1105111
##                   Overall sd       p.value
## duration       413.7182218  0.000000e+00
## previous         0.2870626  4.156304e-70
## campaign         1.3120151  1.480898e-02
## errVar           0.6504890  2.237976e-10
## age             10.0580424  2.265212e-13
## pdays          153.9904918  3.929392e-27
## nr.employed     47.2227912 8.599724e-137
## emp.var.rate     1.3650890  0.000000e+00
## euribor3m        1.5202845  0.000000e+00
## cons.price.idx   0.4734852  0.000000e+00
```

```
## cons.conf.idx    4.4197217  0.000000e+00
```

- For both responses, yes and no, the means are not equal to the global mean.

```
res.catdes$test.chi2 # Global association to factors
```

```
##                         p.value df
## contact           0.000000e+00  1
## month             0.000000e+00  8
## mout              0.000000e+00  1
## poutcome          2.600625e-74  2
## Age_group         9.581755e-33  4
## marital           8.493818e-27  2
## education         1.493193e-26  4
## job               1.056644e-25 10
## day_of_week       4.420802e-17  4
## housing           2.178167e-08  1
## Campaign_contacts 4.960635e-03  1
```

```
res.catdes$category  # Partial association to significative levesls in factors
```

```
## $no
##                                    Cla/Mod     Mod/Cla Global       p.value
## mout=NoMOut                     93.2654659  99.2500000  51.08  0.000000e+00
## month=may                       75.8533502 100.0000000  63.28  0.000000e+00
## contact=telephone               80.1871032 100.0000000  59.86  0.000000e+00
## poutcome=nonexistent            51.5242593 100.0000000  93.16 7.252158e-103
## marital=married                 53.3032549  68.9166667  62.06  6.397639e-22
## education=basic                 56.9124424  41.1666667  34.72  3.468137e-20
## job=blue-collar                 57.2916667  29.7916667  24.96  3.325791e-14
## Age_group=30-50                 51.0720667  71.4583333  67.16  4.772182e-10
## housing=no                      52.0475020  52.9583333  48.84  2.173760e-08
## day_of_week=mon                 55.1038844  25.4166667  22.14  8.397675e-08
## day_of_week=tue                 54.7996272  24.5000000  21.46  4.977022e-07
## Age_group=40-60                 54.0792541  19.3333333  17.16  9.138512e-05
## job=services                    54.9323017  11.8333333  10.34  8.753892e-04
## job=housemaid                   61.5384615   3.0000000   2.34  3.080632e-03
## Campaign_contacts=Infrequent    48.4925690  95.1666667  94.20  4.887560e-03
## job=technician                  43.4610304  13.7083333  15.14  6.625221e-03
## Campaign_contacts=Frequent      40.0000000   4.8333333   5.80  4.887560e-03
## job=retired                     36.4583333   2.9166667   3.84  1.048274e-03
## Age_group=10-20                  0.0000000   0.0000000   0.28  1.039277e-04
## day_of_week=thu                 41.1078717  17.6250000  20.58  6.514524e-07
## day_of_week=wed                 40.6581741  15.9583333  18.84  5.228270e-07
## housing=yes                     44.1360438  47.0416667  51.16  2.173760e-08
## job=admin.                      40.8566722  20.6666667  24.28  9.652705e-09
## Age_group=NA                    21.3675214   1.0416667   2.34  1.981788e-09
## job=student                     18.6274510   0.7916667   2.04  5.320835e-10
## month=oct                        0.0000000   0.0000000   0.84  1.005875e-12
## education=university.degree     37.0725034  22.5833333  29.24  1.772546e-23
## Age_group=20-30                 30.0153139   8.1666667  13.06  1.576945e-23
## marital=single                  35.5523043  20.2500000  27.34  1.634630e-27
## poutcome=success                 0.0000000   0.0000000   2.36  8.354284e-35
## month=mar                        0.0000000   0.0000000   2.52  3.707909e-37
## month=nov                        0.0000000   0.0000000   3.80  3.511500e-56
## poutcome=failure                 0.0000000   0.0000000   4.48  1.951387e-66
```

```
## month=aug                      0.0000000   0.0000000   5.42  8.686216e-81
## month=jun                      0.0000000   0.0000000   7.14 1.357986e-107
## month=jul                      0.0000000   0.0000000   8.14 1.623552e-123
## month=apr                      0.0000000   0.0000000   8.84 8.154365e-135
## mout=YesMOut                   0.7358953   0.7500000  48.92  0.000000e+00
## contact=cellular               0.0000000   0.0000000  40.14  0.000000e+00
##                                   v.test
## mout=NoMOut                          Inf
## month=may                            Inf
## contact=telephone                    Inf
## poutcome=nonexistent           21.535428
## marital=married                 9.622939
## education=basic                 9.203395
## job=blue-collar                 7.584964
## Age_group=30-50                 6.226419
## housing=no                      5.597571
## day_of_week=mon                 5.358369
## day_of_week=tue                 5.027197
## Age_group=40-60                 3.912395
## job=services                    3.327787
## job=housemaid                   2.959576
## Campaign_contacts=Infrequent    2.814353
## job=technician                 -2.715118
## Campaign_contacts=Frequent     -2.814353
## job=retired                    -3.277240
## Age_group=10-20                -3.881234
## day_of_week=thu                -4.975304
## day_of_week=wed                -5.017741
## housing=yes                    -5.597571
## job=admin.                     -5.736721
## Age_group=NA                   -5.999293
## job=student                    -6.209338
## month=oct                      -7.129701
## education=university.degree    -9.985026
## Age_group=20-30                -9.996616
## marital=single                -10.868161
## poutcome=success              -12.306526
## month=mar                     -12.736471
## month=nov                     -15.792367
## poutcome=failure              -17.217839
## month=aug                     -19.035379
## month=jun                     -22.034072
## month=jul                     -23.636539
## month=apr                     -24.710960
## mout=YesMOut                        -Inf
## contact=cellular                    -Inf
##
## $yes
##                                  Cla/Mod    Mod/Cla Global       p.value
## mout=YesMOut                  99.264105 93.3846154  48.92  0.000000e+00
## contact=cellular             100.000000 77.1923077  40.14  0.000000e+00
## month=apr                    100.000000 17.0000000   8.84 8.154365e-135
## month=jul                    100.000000 15.6538462   8.14 1.623552e-123
## month=jun                    100.000000 13.7307692   7.14 1.357986e-107
```

```
## month=aug                     100.000000 10.4230769   5.42  8.686216e-81
## poutcome=failure              100.000000  8.6153846   4.48  1.951387e-66
## month=nov                     100.000000  7.3076923   3.80  3.511500e-56
## month=mar                     100.000000  4.8461538   2.52  3.707909e-37
## poutcome=success              100.000000  4.5384615   2.36  8.354284e-35
## marital=single                 64.447696 33.8846154  27.34  1.634630e-27
## Age_group=20-30                69.984686 17.5769231  13.06  1.576945e-23
## education=university.degree    62.927497 35.3846154  29.24  1.772546e-23
## month=oct                     100.000000  1.6153846   0.84  1.005875e-12
## job=student                    81.372549  3.1923077   2.04  5.320835e-10
## Age_group=NA                   78.632479  3.5384615   2.34  1.981788e-09
## job=admin.                     59.143328 27.6153846  24.28  9.652705e-09
## housing=yes                    55.863956 54.9615385  51.16  2.173760e-08
## day_of_week=wed                59.341826 21.5000000  18.84  5.228270e-07
## day_of_week=thu                58.892128 23.3076923  20.58  6.514524e-07
## Age_group=10-20               100.000000  0.5384615   0.28  1.039277e-04
## job=retired                    63.541667  4.6923077   3.84  1.048274e-03
## Campaign_contacts=Frequent     60.000000  6.6923077   5.80  4.887560e-03
## job=technician                 56.538970 16.4615385  15.14  6.625221e-03
## Campaign_contacts=Infrequent   51.507431 93.3076923  94.20  4.887560e-03
## job=housemaid                  38.461538  1.7307692   2.34  3.080632e-03
## job=services                   45.067698  8.9615385  10.34  8.753892e-04
## Age_group=40-60                45.920746 15.1538462  17.16  9.138512e-05
## day_of_week=tue                45.200373 18.6538462  21.46  4.977022e-07
## day_of_week=mon                44.896116 19.1153846  22.14  8.397675e-08
## housing=no                     47.952498 45.0384615  48.84  2.173760e-08
## Age_group=30-50                48.927933 63.1923077  67.16  4.772182e-10
## job=blue-collar                42.708333 20.5000000  24.96  3.325791e-14
## education=basic                43.087558 28.7692308  34.72  3.468137e-20
## marital=married                46.696745 55.7307692  62.06  6.397639e-22
## poutcome=nonexistent           48.475741 86.8461538  93.16 7.252158e-103
## mout=NoMOut                     6.734534  6.6153846  51.08  0.000000e+00
## month=may                      24.146650 29.3846154  63.28  0.000000e+00
## contact=telephone              19.812897 22.8076923  59.86  0.000000e+00
##                                   v.test
## mout=YesMOut                         Inf
## contact=cellular                     Inf
## month=apr                      24.710960
## month=jul                      23.636539
## month=jun                      22.034072
## month=aug                      19.035379
## poutcome=failure               17.217839
## month=nov                      15.792367
## month=mar                      12.736471
## poutcome=success               12.306526
## marital=single                 10.868161
## Age_group=20-30                 9.996616
## education=university.degree     9.985026
## month=oct                       7.129701
## job=student                     6.209338
## Age_group=NA                    5.999293
## job=admin.                      5.736721
## housing=yes                     5.597571
## day_of_week=wed                 5.017741
```

```
## day_of_week=thu                4.975304
## Age_group=10-20                3.881234
## job=retired                    3.277240
## Campaign_contacts=Frequent     2.814353
## job=technician                 2.715118
## Campaign_contacts=Infrequent  -2.814353
## job=housemaid                 -2.959576
## job=services                  -3.327787
## Age_group=40-60               -3.912395
## day_of_week=tue               -5.027197
## day_of_week=mon               -5.358369
## housing=no                    -5.597571
## Age_group=30-50               -6.226419
## job=blue-collar               -7.584964
## education=basic               -9.203395
## marital=married               -9.622939
## poutcome=nonexistent         -21.535428
## mout=NoMOut                        -Inf
## month=may                          -Inf
## contact=telephone                  -Inf
```

```
write.csv2(df,file="clean_data.csv")
```