

Course Practical Assignment - 1st Delivery (17 de març del 2019)

Josep Clotet Ginovart

Eric Martin Obispo

Bank client data

Description of input 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', 'unknown')
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')# related with the last contact of the current campaign:
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 packages:

Loading data:

```
#dirwd<-"D:/Users/Usuari/Documents/ADEIpractica"
dirwd<-"D:/Documents/GitHub/ADEI"
setwd(dirwd)

df<-read.table( paste0(dirwd, "/bank-additional/bank-additional-full.csv"), header=TRUE, sep=";")

# General description of the bank data
```

```

#head(df)
nrow(df)

## [1] 41188

ncol(df)

## [1] 21

dim(df)

## [1] 41188    21

# Selection of our 5000 samples with a specific seed value
set.seed(17041998)
llista<-sample(size=5000, x=1:nrow(df), replace=FALSE)
llista<-sort(llista)

# Overwrite the dataframe with our chosen sample and save the RData
df<-df[llista,]
save.image( paste0(dirwd, "/bank-additional/Bank5000_raw.RData") )

```

Our chosen sample:

```

#load( paste0(dirwd, "/bank-additional/Bank5000_raw.RData") )
summary(df)

##      age                job                marital
##  Min.   :18.00   admin.       :1234   divorced: 556
##  1st Qu.:32.00   blue-collar:1154   married  :3053
##  Median :38.00   technician : 794   single   :1381
##  Mean   :40.07   services    : 500   unknown  : 10
##  3rd Qu.:47.00   management : 413
##  Max.   :87.00   retired     : 205
##                (Other)      : 700
##      education          default          housing          loan
##  university.degree :1472   no       :3966   no       :2219   no       :4091
##  high.school         :1171   unknown:1034   unknown: 137   unknown: 137
##  basic.9y            : 716   yes      : 0    yes      :2644   yes      : 772
##  professional.course: 602
##  basic.4y            : 513
##  basic.6y            : 291
##  (Other)             : 235
##      contact          month          day_of_week          duration
##  cellular :3130   may       :1743   fri: 924   Min.   : 1.0
##  telephone:1870   jul       : 831   mon:1018   1st Qu.: 101.0
##                aug       : 699   thu:1039   Median : 178.0
##                jun       : 653   tue:1045   Mean    : 254.8
##                nov       : 509   wed: 974   3rd Qu.: 317.0
##                apr       : 310   Max.    :3785.0
##                (Other): 255
##      campaign          pdays          previous          poutcome
##  Min.   : 1.000   Min.   : 0.0   Min.   :0.0000   failure   : 478
##  1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000   nonexistent:4363
##  Median : 2.000   Median :999.0   Median :0.0000   success    : 159

```

```
## Mean : 2.583 Mean :963.2 Mean :0.1606
## 3rd Qu.: 3.000 3rd Qu.:999.0 3rd Qu.:0.0000
## Max. :33.000 Max. :999.0 Max. :4.0000
##
## emp.var.rate cons.price.idx cons.conf.idx euribor3m
## Min. : -3.40000 Min. :92.20 Min. : -50.80 Min. :0.635
## 1st Qu.: -1.80000 1st Qu.:93.08 1st Qu.: -42.70 1st Qu.:1.334
## Median : 1.10000 Median :93.77 Median : -41.80 Median :4.857
## Mean : 0.06326 Mean :93.57 Mean : -40.43 Mean :3.613
## 3rd Qu.: 1.40000 3rd Qu.:93.99 3rd Qu.: -36.40 3rd Qu.:4.961
## Max. : 1.40000 Max. :94.77 Max. : -26.90 Max. :5.000
##
## nr.employed y
## Min. :4964 no :4435
## 1st Qu.:5099 yes: 565
## Median :5191
## Mean :5166
## 3rd Qu.:5228
## Max. :5228
##
```

Inicialitzacio del control d'errors, missings i outliers:

```
columns <- names(df) #list of column names

# creem 3 dataframes inicialitzats a 0 d'una fila amb les columnes de la nostra mostra;
# en ells hi posarem el nombre d'errors, missings i outliers per a cada variable
errors <- data.frame(matrix(0, ncol = length(columns), nrow = 1))
colnames(errors)<-columns

missings <- data.frame(matrix(0, ncol = length(columns), nrow = 1))
colnames(missings)<-columns

outliers <- data.frame(matrix(0, ncol = length(columns), nrow = 1))
colnames(outliers)<-columns

# columnes que portaran el control per individu:
df$num_missings <- 0
df$num_outliers <- 0
df$num_errors <- 0
```

UNIVARIATE DESCRIPTIVE ANALYSIS (to be included for each variable):

Aqui estudiem cada variable buscant missing values, outliers i possibles errors. En el cas que en trobem, els transformem en NAs i procedim a una imputacio manual o els eliminem, o una imputacio automatica (en un chunk posterior d'Imputation).

VARIABLES QUALITATIVE:

Tambe factoritzem aqui les categories (levels) de les variables qualitatives (discretes). Les etiquetes adicionales als factors s'afegeixen posteriorment als grafics per una qestio estetica, es redueix la mida de les etiquetes i

es poden veure amb mes claredad cada una de les variables.

Job

Els “unknowns” seran imputats mes endavant automaticament.

```
# Jobs "unknown" will be a missing value (set to NA):
```

```
sel<-which(df$job=="unknown");length(sel)
```

```
## [1] 39
```

```
df$job[sel]<-NA
```

```
# Missings:
```

```
miss<-which(is.na(df$job));
```

```
missings$job<-length(miss); length(miss)
```

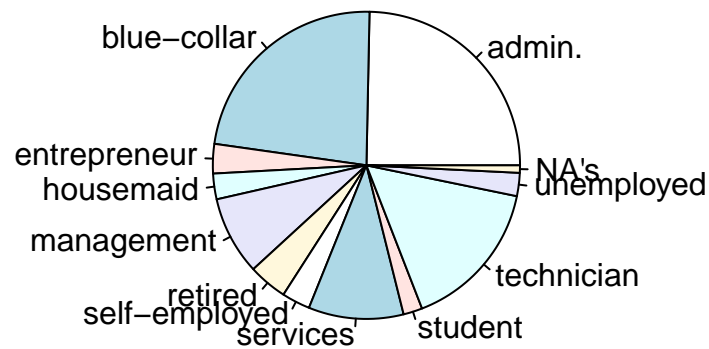
```
## [1] 39
```

```
df[miss, "num_missings"]<- df[miss, "num_missings"]+1
```

```
# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "job-":
```

```
df$job<-factor(df$job)
```

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



```
levels(df$job)<-paste0("job-",levels(df$job))
```

Marital

Els “unknowns” seran imputats mes endavant automaticament.

```

# Marital "unknown" will be a missing value (set to NA):
sel<-which(df$marital=="unknown");length(sel)

## [1] 10

df$marital[sel]<-NA

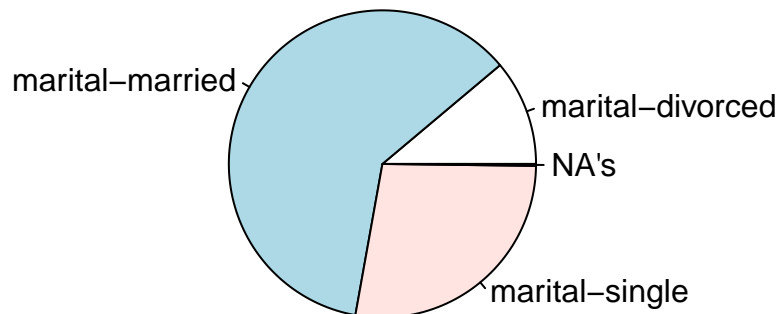
# Missings:
miss<-which(is.na(df$marital));
missings$marital<-length(miss); length(miss)

## [1] 10

df[miss, "num_missings"]<- df[miss, "num_missings"]+1

# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "marital-":
df$marital<-factor(df$marital)
levels(df$marital)<-paste0("marital-",levels(df$marital))
pie(summary(df$marital))

```



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

```
## marital-divorced  marital-married  marital-single      NA's
##                556             3053             1381          10
```

Education

Els “unknowns” seran imputats mes endavant automaticament. La categoria “illiterate” es ajuntada manualment a “basic.4y”.

```

# Education "unknown" will be a missing value (set to NA):
sel<-which(df$education=="unknown");length(sel)

## [1] 232

df$education[sel]<-NA

# Illiterates are consired as basic.4y.educated:
sel<-which(df$education=="illiterate");length(sel)

## [1] 3

df[sel, "education"]<-"basic.4y"

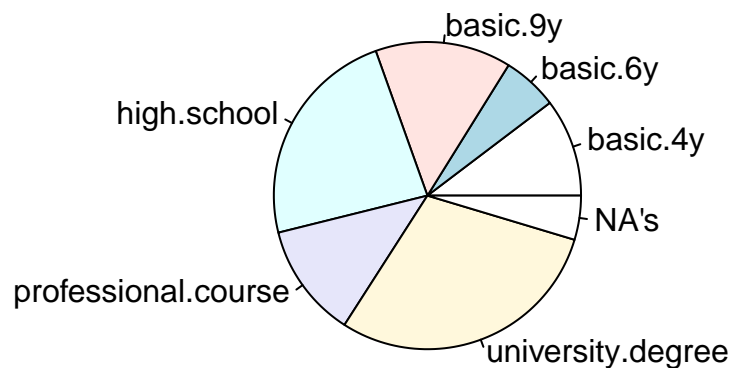
# Missings:
miss<-which(is.na(df$education));
missings$education<-length(miss); length(miss)

## [1] 232

df[miss, "num_missings"]<- df[miss, "num_missings"]+1

# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "education-":
df$education<-factor(df$education)
pie(summary(df$education))

```



```

levels(df$education)<-paste0("education-",levels(df$education))

```

Default (has credit in default?)

Default “unknown” sera considerada com a una categoria, no com a missing value.

```

table(df$default, useNA="always")

##
##      no unknown      yes      <NA>
##    3966    1034         0         0
# Missings:
miss<-which(is.na(df$default));
missings$default<-length(miss); length(miss)

## [1] 0
df[miss, "num_missings"]<- df[miss, "num_missings"]+1

# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "default-":
df$default<-factor(df$default)
summary(df$default)

##      no unknown
##    3966    1034
levels(df$default)<-paste0("default-",levels(df$default))

```

Housing

Els “unknowns” seran imputats mes endavant automaticament.

```

sel<-which(df$housing=="unknown");length(sel)

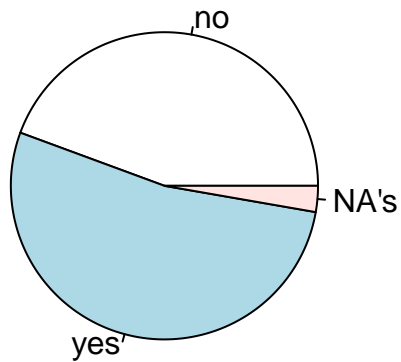
## [1] 137
df$housing[sel]<-NA

# Missings:
miss<-which(is.na(df$housing));
missings$housing<-length(miss); length(miss)

## [1] 137
df[miss, "num_missings"]<- df[miss, "num_missings"]+1

# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "housing-":
df$housing<-factor(df$housing)
pie(summary(df$housing))

```



```
levels(df$housing)<-paste0("housing-",levels(df$housing))
```

Loan (has personal loan?)

Loan “unknown” sera considerat com a missing value (NA), sera imputat mes endavant automaticament.

```
sel<-which(df$loan=="unknown");length(sel)
```

```
## [1] 137
```

```
df$loan[sel]<-NA
```

```
# Missings:
```

```
miss<-which(is.na(df$loan));
```

```
missings$loan<-length(miss); length(miss)
```

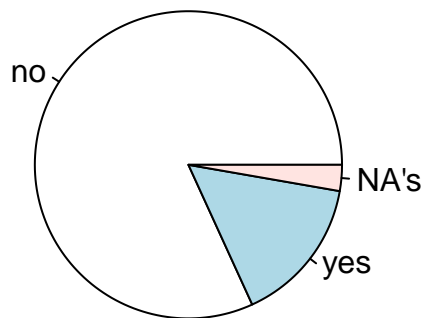
```
## [1] 137
```

```
df[miss, "num_missings"]<- df[miss, "num_missings"]+1
```

```
# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "loan-":
```

```
df$loan<-factor(df$loan)
```

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

```
levels(df$loan)<-paste0("loan-",levels(df$loan))
```

Contact

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

```
## cellular telephone
##      3130      1870
```

```
# Missings:
```

```
miss<-which(is.na(df$contact));
missings$contact<-length(miss); length(miss)
```

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

```
df[miss, "num_missings"]<- df[miss, "num_missings"]+1
```

```
# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "contact-":
```

```
df$contact<-factor(df$contact)
summary(df$contact)
```

```
## cellular telephone
##      3130      1870
```

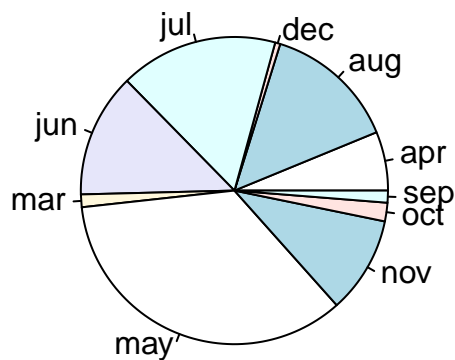
```
levels(df$contact)<-paste0("contact-",levels(df$contact))
```

Month

```
miss<-which(is.na(df$month));
missings$month<-length(miss); length(miss)
```

```
## [1] 0
df[miss, "num_missings"]<- df[miss, "num_missings"]+1

# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "month-":
df$month<-factor(df$month)
pie(summary(df$month))
```



```
levels(df$month)<-paste0("month-",levels(df$month))
```

Month -> definim noves factor categories per Season.

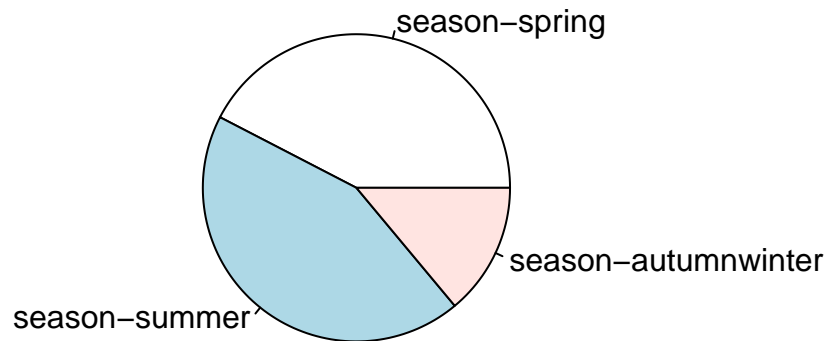
```
# Define new factor categories: 1-Spring 2-Summer 3-AutumnWinter
df$f.season <- 3
# 1 level - spring
sel<-which(df$month %in% c("month-mar","month-apr","month-may"))
df$f.season[sel] <-1

# 2 level - summer
sel<-which(df$month %in% c("month-jun","month-jul","month-aug"))
df$f.season[sel] <-2

df$f.season<-factor(df$f.season, levels=1:3, labels=c("season-spring","season-summer",
"season-autumnwinter"))

summary(df$f.season);pie(summary(df$f.season))
```

```
##          season-spring      season-summer season-autumnwinter
##                2120                2183                697
```



Day_of_week

```
miss<-which(is.na(df$day_of_week));
missings$day_of_week<-length(miss); length(miss)

## [1] 0

df[miss, "num_missings"]<- df[miss, "num_missings"]+1

# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "day_of_week-":
levels(df$day_of_week)<-c(levels(df$day_of_week), "1mon", "2tue", "3wed", "4thu", "5fri")
sel<-which(df$day_of_week=="mon"); df$day_of_week[sel]<-"1mon"
sel<-which(df$day_of_week=="tue"); df$day_of_week[sel]<-"2tue"
sel<-which(df$day_of_week=="wed"); df$day_of_week[sel]<-"3wed"
sel<-which(df$day_of_week=="thu"); df$day_of_week[sel]<-"4thu"
sel<-which(df$day_of_week=="fri"); df$day_of_week[sel]<-"5fri"

df$day_of_week<-factor(df$day_of_week)
summary(df$day_of_week)

## 1mon 2tue 3wed 4thu 5fri
## 1018 1045 974 1039 924

levels(df$day_of_week)<-paste0("day_of_week-",levels(df$day_of_week))
```

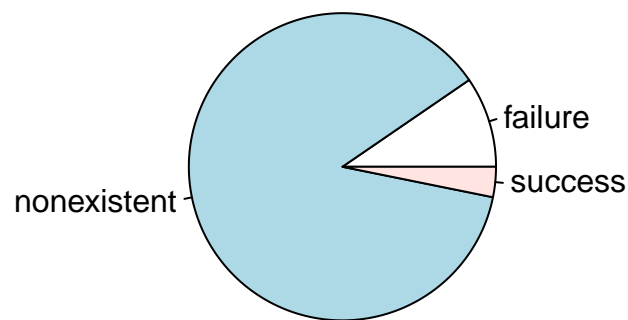
Poutcome (outcome of previous marketing campaign)

```
# Poutcome "nonexistent" will be considered a category, not a missing value.
table(df$poutcome, useNA="always")
```

```
##
##      failure nonexistent      success      <NA>
##      478      4363      159      0
# All missing data indicated as NA:
miss<-which(is.na(df$poutcome));
missings$poutcome<-length(miss); length(miss)

## [1] 0
df[miss, "num_missings"]<- df[miss, "num_missings"]+1

# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "poutcome-":
df$poutcome<-factor(df$poutcome)
pie(summary(df$poutcome))
```



```
levels(df$poutcome)<-paste0("poutcome-",levels(df$poutcome))
```

y (has the client subscribed a term deposit?)

```
miss<-which(is.na(df$y));
missings$y<-length(miss); length(miss)

## [1] 0
df[miss, "num_missings"]<- df[miss, "num_missings"]+1

# Factoritzem les categories (levels) de la columna i afegim l'etiqueta "y-":
df$y<-factor(df$y)
summary(df$y)
```

```
## no yes
## 4435 565
levels(df$y)<-paste0("y-",levels(df$y))
```

VARIABLES QUANTITATIVES:

Funcio de gran utilitat per a la deteccio d'outliers:

```
calcQ <- function(x){
  s.x <- summary(x)

  iqr <- s.x[5]-s.x[2] # IQR = Q3([5]) - Q1([2])

  list(souti=s.x[2]-3*iqr, mouti=s.x[2]-1.5*iqr, min=s.x[1], q1=s.x[2],
       q2=s.x[3], q3=s.x[5], max=s.x[6], mouts=s.x[5]+1.5*iqr, souts=s.x[5]+3*iqr)
}
```

Age

```
summary(df$age)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  18.00   32.00   38.00   40.07   47.00   87.00
```

```
# No tenim cap missing NA!
```

```
miss<-which(is.na(df$age))
missings$age<-length(miss); length(miss)
```

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

```
df[miss, "num_missings"]<- df[miss, "num_missings"]+1
```

```
par(mfrow=c(1,2))
hist(df$age, breaks=10, main="age - histogram")
Boxplot(df$age)
```

```
## [1] 4570 4634 3623 3628 3631 4755 4612 4734 4740 4512
```

```
# Errors are under aged people:
```

```
err<-which(df$age < 18)
errors$age<-length(err); length(err)
```

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

```
if(length(err)>0) df[err]<-NA
```

```
# Outliers:
```

```
out.var <- calcQ(df$age)
abline(h=out.var[["mouts"]], col="magenta", lwd=2); out.var[["mouts"]]
```

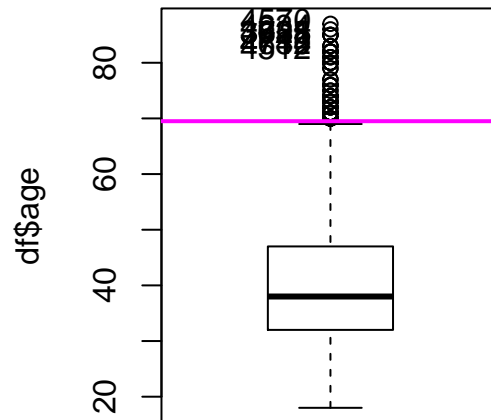
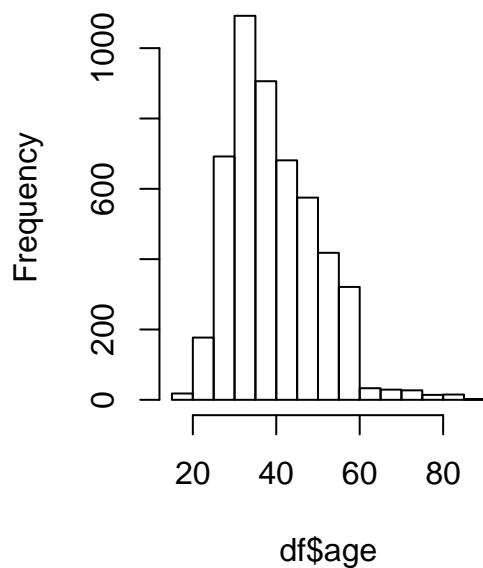
```
## 3rd Qu.
```

```
## 69.5
```

```
# But our outliers will be the ones above 100 years (there is none):
```

```
abline(h=100, col="red", lwd=2)
```

age – histogram



```
out<-which(df$age > 100)
outliers$age<-length(out); length(out)
```

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

```
if(length(out)>0) df[out]<-NA
```

Duration

Els outliers en la variable duracio han estat eliminats. Corresponen a duracions per sota els 5 segons (trucada massa curta a un client que potser no podia parlar en aquell moment o penja per error) i per sobre dels 1600 segons (26 minuts).

```
summary(df$duration)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.0   101.0   178.0   254.8   317.0  3785.0
```

```
# No tenim cap missing NA!
```

```
miss<-which(is.na(df$duration));
missings$duration<-length(miss); length(miss)
```

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

```
df[miss, "num_missings"]<- df[miss, "num_missings"]+1
```

```
par(mfrow=c(1,2))
hist(df$duration, breaks=20, main="duration - histogram")
Boxplot(df$duration)
```

```
## [1] 4929 3368 2817 4759 1285 2907 2033 3815 4998 3280
```

```

# Outliers:
out.var <- calcQ(df$duration)
abline(h=out.var[["mouts"]], col="magenta", lwd=2); out.var[["mouts"]]

## 3rd Qu.
##      641

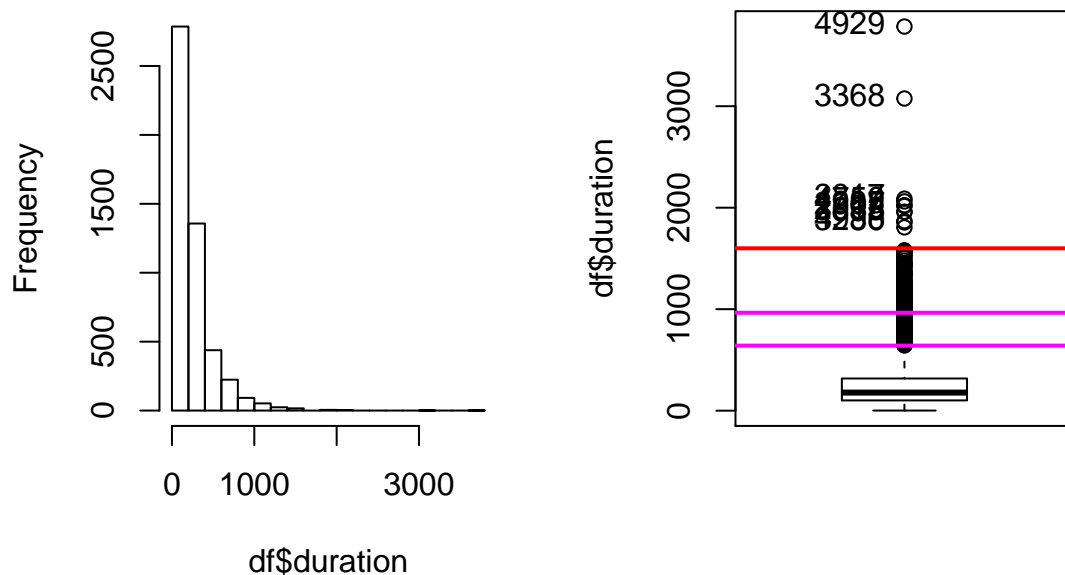
abline(h=out.var[["souts"]], col="magenta", lwd=2); out.var[["souts"]]

## 3rd Qu.
##      965

# But our outliers will be the ones above 1600 and below 5 seconds:
abline(h=1600, col="red", lwd=2)

```

duration – histogram



```

out<-which( (df$duration < 5) | (df$duration > 1600) )
outliers$duration=length(out); length(out)

## [1] 14

df[out, "num_outliers"]<- df[out, "num_outliers"]+1
df[out, "duration"]<-NA

# Eliminem els outliers:
if(length(out)>0) df<-df[-out,]

# Final summary of duration variable:
# par(mfrow=c(1,1))
# summary(df$duration)
# Boxplot(df$duration)

```

Duration -> creem una columna de duracio en minuts:

```
df$minutes<-df$duration/60
summary(df$minutes)
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.     Max.
## 0.08333  1.68333  2.95000  4.17703  5.26667 26.33333
```

Campaign

```
# summary(df$campaign)
# No tenim cap missing NA!
miss<-which(is.na(df$campaign));
missings$campaign<-length(miss); length(miss)

## [1] 0

df[miss, "num_missings"]<- df[miss, "num_missings"]+1

par(mfrow=c(1,2))
hist(df$campaign, breaks=10, main="campaign - histogram")
Boxplot(df$campaign)
```

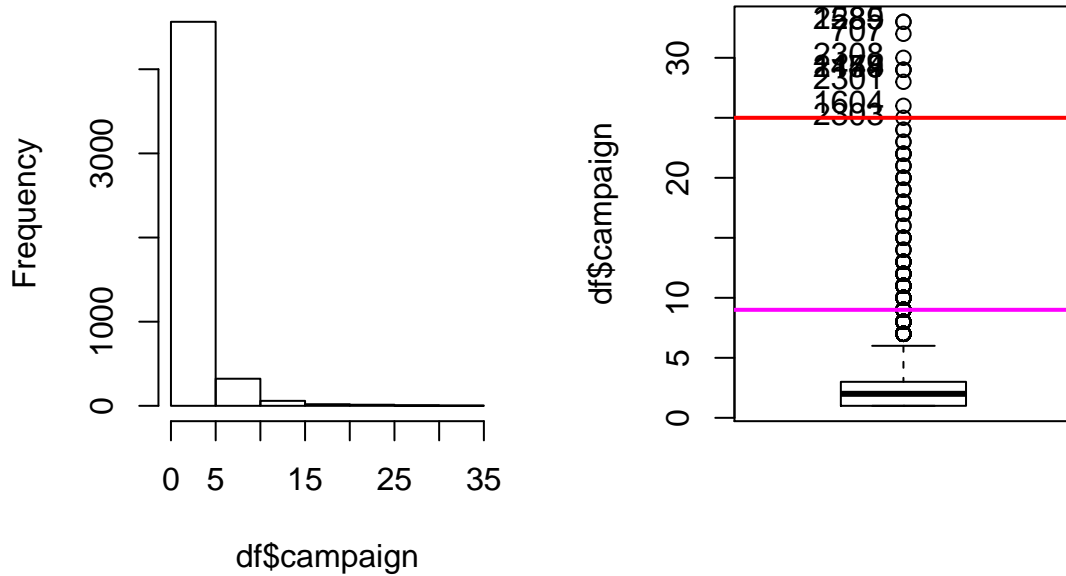
```
## [1] 1589 2285 707 2308 1158 1474 2149 2301 1604 2303
```

```
# Outliers:
out.var <- calcQ(df$campaign)
abline(h=out.var[["souts"]], col="magenta", lwd=2); out.var[["souts"]]
```

```
## 3rd Qu.
##      9
```

```
# But our outliers will be the ones contacted more than 25 times:
abline(h=25, col="red", lwd=2)
```


campaign – histogram



```
out<-which(df$campaign > 25)
df[out, "num_outliers"]<- df[out, "num_outliers"]+1
outliers$campaign=length(out); length(out)
```

```
## [1] 9
```

```
df[out, "campaign"]<-NA
```

```
# Final summary of campaign variable:
# par(mfrow=c(1,1))
# summary(df$campaign)
# Boxplot(df$campaign)
```

Pdays

Els valors 999 corresponen a mai contactats, son NA i s'imputen manualment al maxim(tret dels 999)+1.

```
# No tenim cap missing NA!
miss<-which(is.na(df$pdays));
missings$pdays<-length(miss); length(miss)
```

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

```
df[miss, "num_missings"]<- df[miss, "num_missings"]+1
```

```
# Values that are 999 mean never contacted before:
never<-which(df$pdays==999)
df$pdays[never]<-19 #imputacio manual al maxim+1
# Son outliers
df[out, "num_outliers"]<- df[out, "num_outliers"]+1
outliers$pdays=length(never); length(never)
```

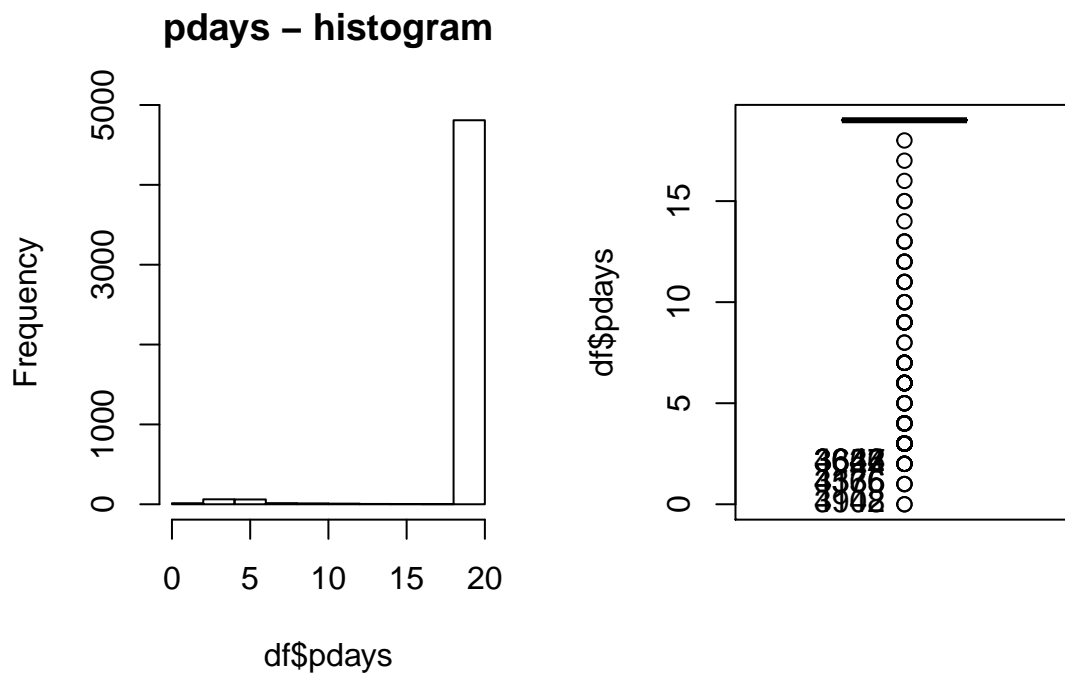
```
## [1] 4809
# They correspond to this percentage of rows:
length(never)/5000*100

## [1] 96.18
# No outliers!

# Final summary of pdays variable:
summary(df$pdays)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   19.00   19.00   18.53   19.00   19.00

par(mfrow=c(1,2))
hist(df$pdays, breaks=10, main="pdays - histogram")
Boxplot(df$pdays)
```



```
## [1] 3148 4902 3576 4135 4366 3627 3642 3644 3646 4352
```

Previous

```
# No tenim cap missing NA!
miss<-which(is.na(df$previous));
missings$previous<-length(miss); length(miss)

## [1] 0

df[miss, "num_missings"]<- df[miss, "num_missings"]+1

par(mfrow=c(1,2))
```

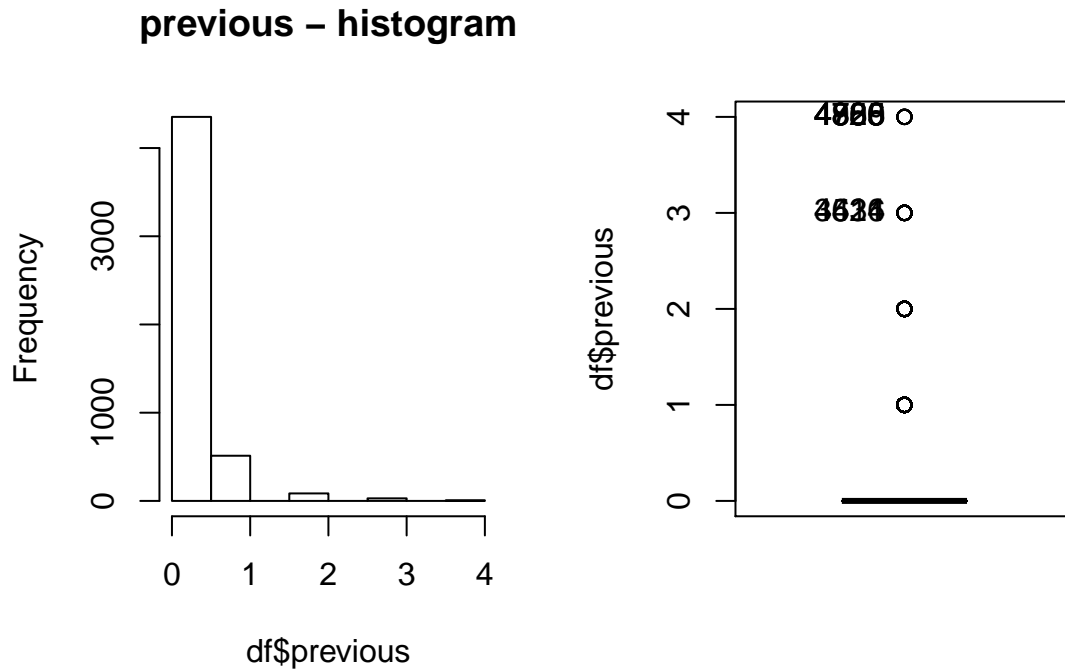
```
hist(df$previous, main="previous - histogram")
```

```
# Final summary of previous variable:
```

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000 0.0000 0.0000 0.1598 0.0000 4.0000
```

```
Boxplot(df$previous)
```



```
## [1] 4769 4786 4805 4826 4850 4888 4925 3431 4516 4624
```

emp.var.rate

```
# Neither missing, outliers nor error values.
```

```
par(mfrow=c(1,2))
```

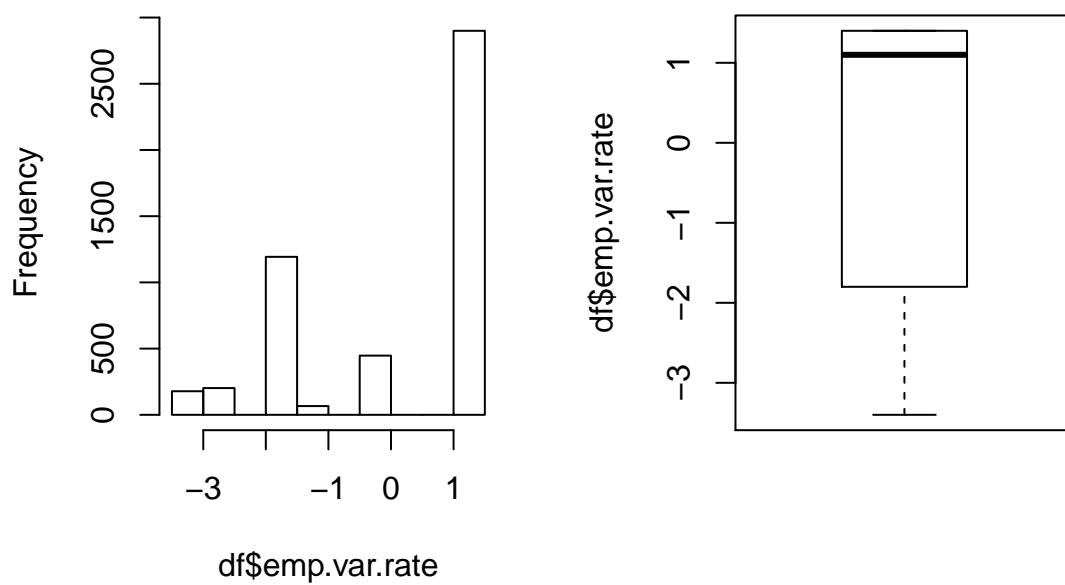
```
hist(df$emp.var.rate, main="emp.var.rate - histogram")
```

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.40000 -1.80000  1.10000  0.06446  1.40000  1.40000
```

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

emp.var.rate – histogram



cons.price.idx

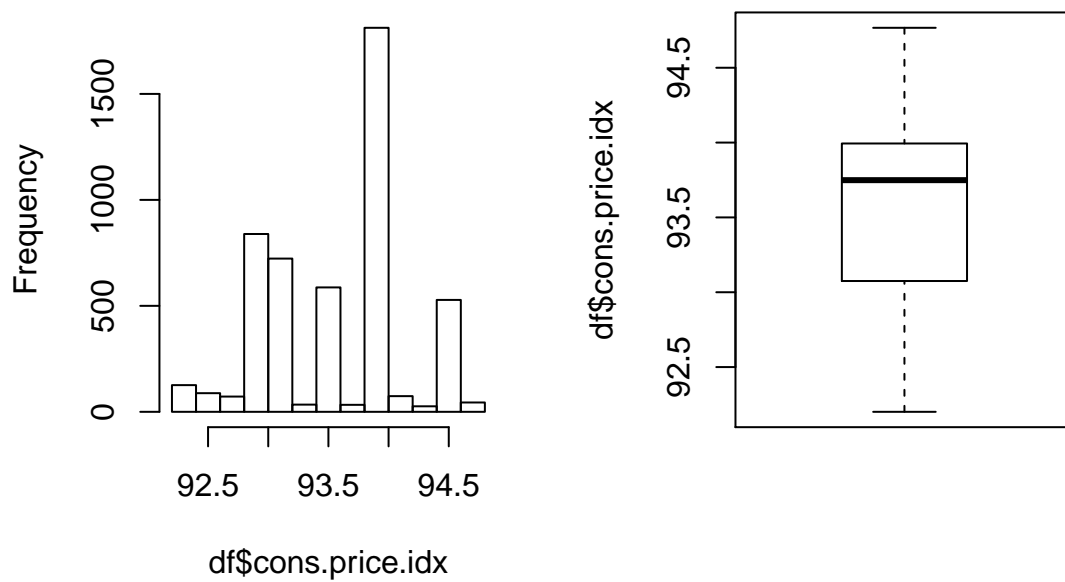
```
# Neither missing, outliers nor error values.
par(mfrow=c(1,2))

hist(df$cons.price.idx, main="cons.price.idx - histogram")
summary(df$cons.price.idx)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  92.20  93.08   93.75   93.57   93.99   94.77
```

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

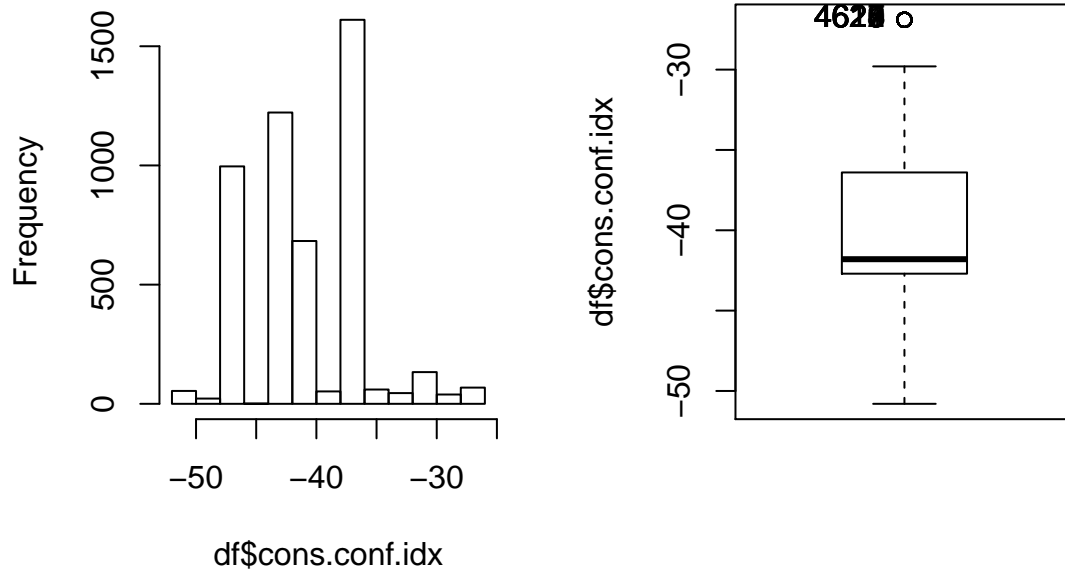
cons.price.idx – histogram



cons.conf.idx

```
# Neither missing, outliers nor error values.  
par(mfrow=c(1,2))  
  
hist(df$cons.conf.idx, main="cons.conf.idx - histogram")  
summary(df$cons.conf.idx)  
  
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
## -50.80 -42.70  -41.80  -40.43  -36.40  -26.90  
  
Boxplot(df$cons.conf.idx)
```

cons.conf.idx – histogram



```
## [1] 4617 4618 4619 4620 4621 4622 4623 4624 4625 4626
```

euribor3m

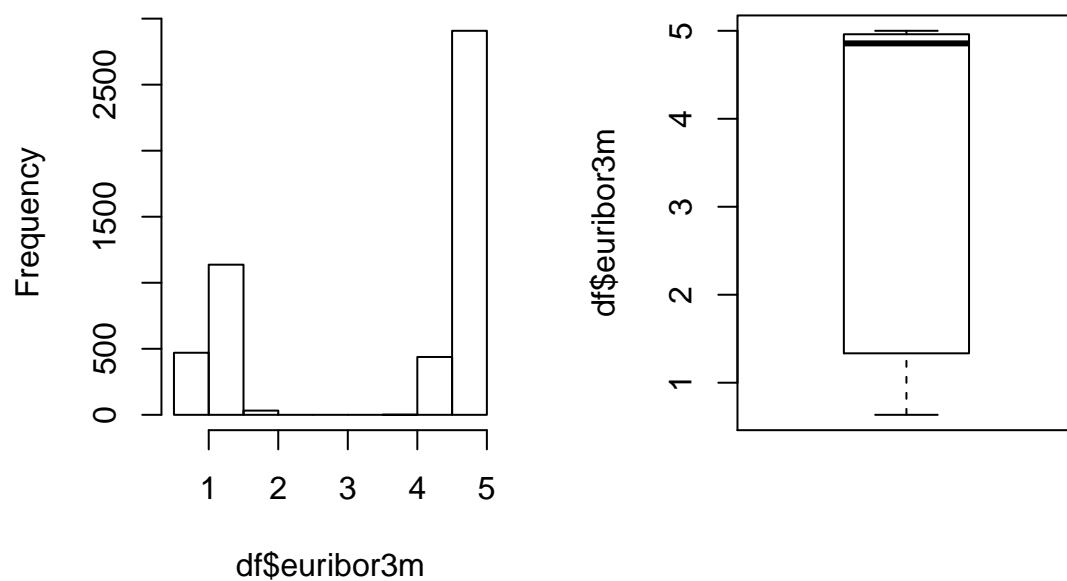
```
# Neither missing, outliers nor error values.
par(mfrow=c(1,2))

hist(df$euribor3m, main="euribor3m - histogram")
summary(df$euribor3m)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.635   1.334   4.857   3.614   4.961   5.000
```

```
Boxplot(df$euribor3m)
```

euribor3m – histogram



nr.employed

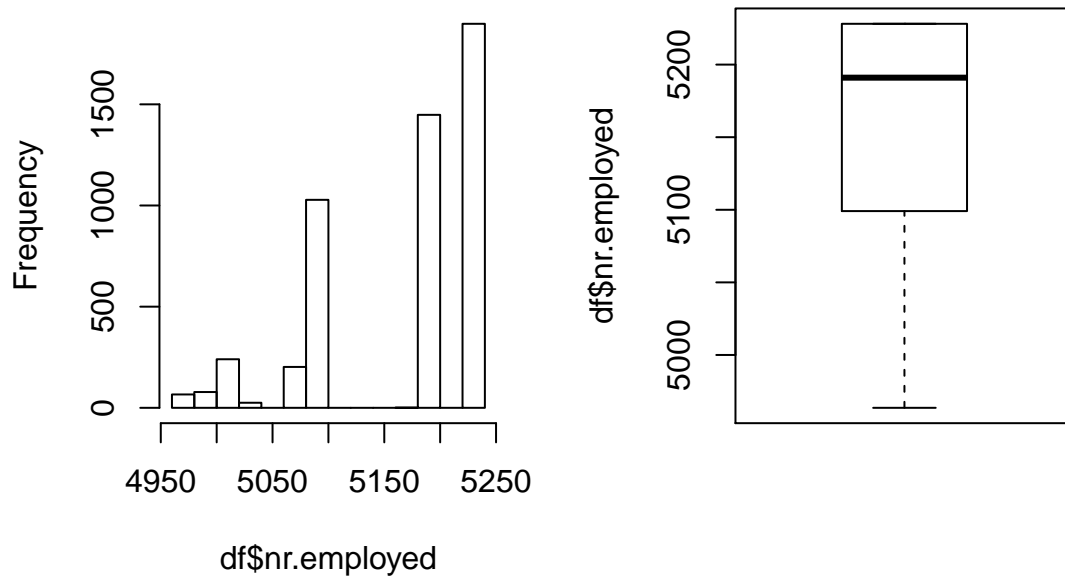
```
# Neither missing, outliers nor error values.
par(mfrow=c(1,2))

hist(df$nr.employed, main="nr.employed - histogram")
summary(df$nr.employed)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      4964   5099   5191   5166   5228   5228
```

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

nr.employed – histogram



DISCRETITZACIO DE VARIABLES NUMERIQUES:

Imputacio de variables numeriques abans de discretitzar-les:

La variable numerica campaign te certs individus que han estat considerats outliers previament. Aqui els imputem mitjançant la imputacio automatica imputePCA().

```
vars_con<-names(df)[c(1, 11:14, 16:20)]
res.imp<-imputePCA(df[,vars_con], ncp=8)
```

```
# Original:
summary(df$campaign)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##      1.000   1.000   2.000   2.535   3.000   25.000         9
```

```
# Amb dades imputades:
summary(res.imp$completeObs[, "campaign"])
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000   1.000   2.000   2.535   3.000   25.000
```

```
# Acceptem la imputacio
df$campaign<-res.imp$completeObs[, "campaign"]
#summary(df[,vars_con])
```

Les variables numeriques originals que corresponen a conceptes quantitativs reals es mantenen com a numeriques, pero tambe s'han de crear factors addicionals com a discretitzacio de cada variable numerica. Les etiquetes addicionals als factors s'afegeixen posterior als grafics per una qüestio estetica, es redueix la mida de les etiquetes i es poden veure amb mes claredat cada una de les variables.


```

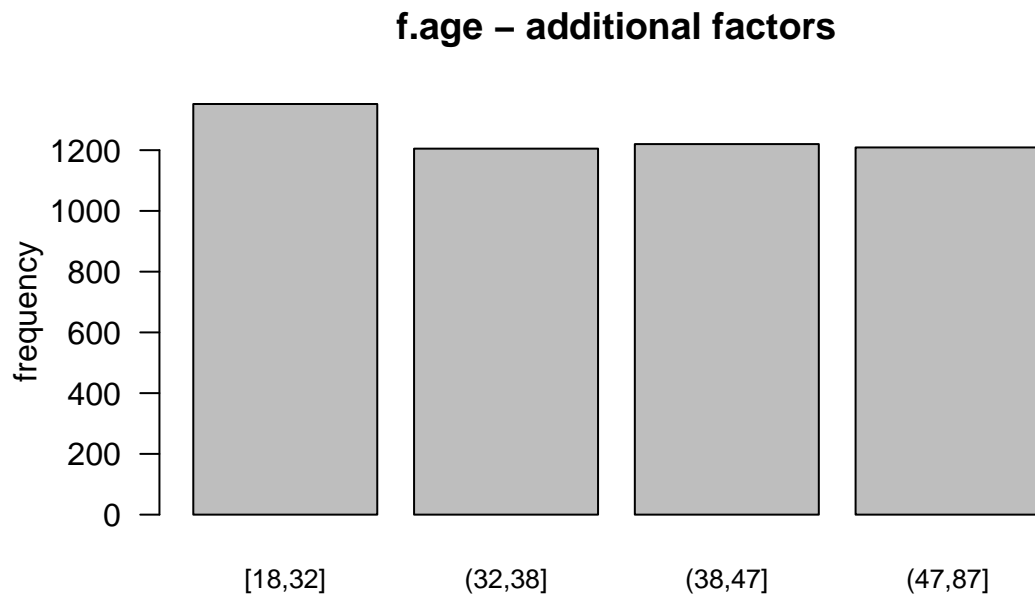
par(mfrow=c(1,1))

# AGE
qulist<-quantile(df$age, seq(0,1,0.25), na.rm=TRUE)

df$f.age<-factor( cut(df$age, breaks=qulist, include.lowest=T) )

# Es mostra una distribucio d'edats equitativa amb aquesta factoritzacio:
barplot(table(df$f.age), main="f.age - additional factors", ylab="frequency", las=1, cex.names=0.8)

```



```

summary(df$f.age)

## [18,32] (32,38] (38,47] (47,87]
##      1352      1205      1220      1209

levels(df$f.age)<-paste0("f.age-", levels(df$f.age) )

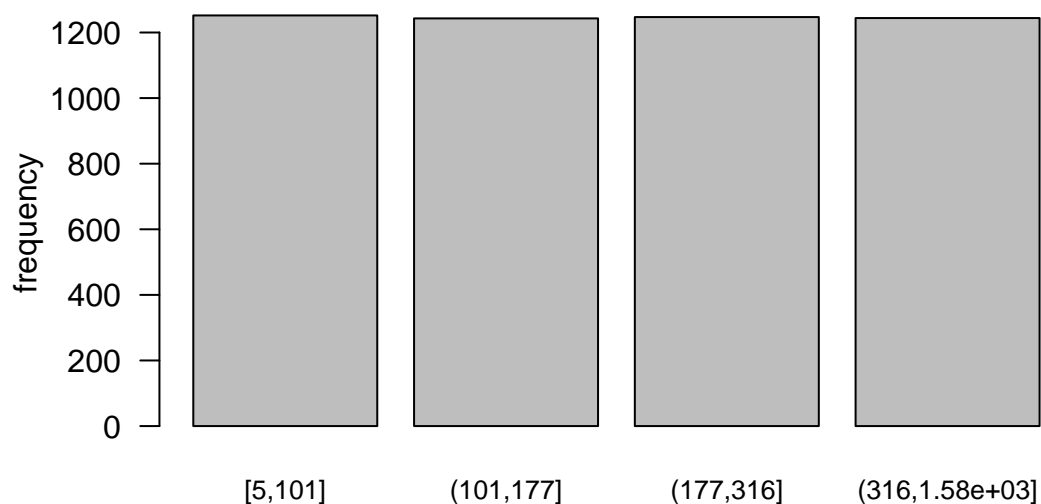
# DURATION
qulist<-quantile(df$duration, seq(0,1,0.25), na.rm=TRUE)

df$f.duration<-factor( cut(df$duration, breaks=qulist, include.lowest=T) )

# Es mostra una distribucio de duracions de la trucada equitativa amb aquesta factoritzacio:
barplot(table(df$f.duration), main="f.duration - additional factors", ylab="frequency", las=1, cex.names=0.8)

```

f.duration – additional factors



```
levels(df$f.duration)<-paste0("f.duration-", levels(df$f.duration) )
summary(df$f.duration)
```

```
##          f.duration-[5,101]      f.duration-(101,177]
##                1252                1243
##          f.duration-(177,316] f.duration-(316,1.58e+03]
##                1247                1244
```

```
# CAMPAIGN
```

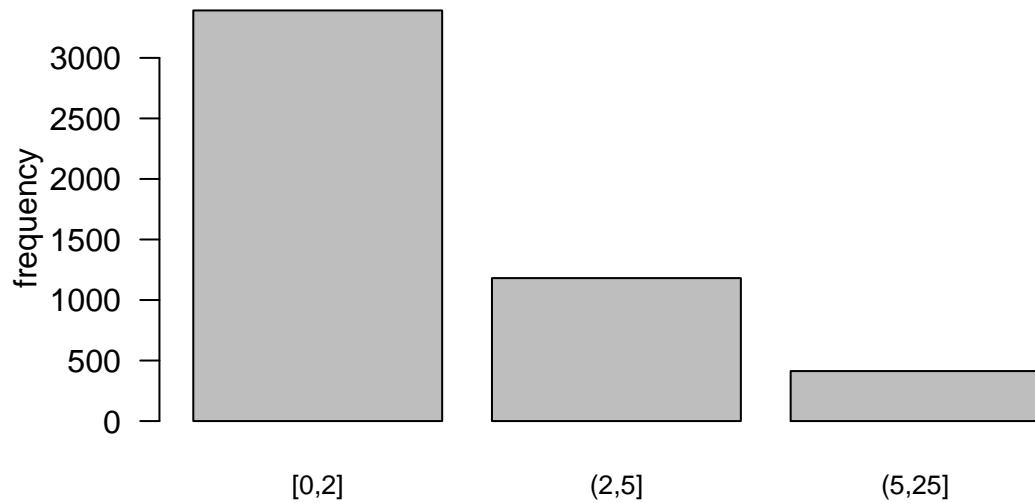
```
qulist<-quantile(df$campaign, seq(0,1,0.5), na.rm=TRUE)
```

```
df$f.campaign<-factor( cut(df$campaign, breaks=c(0,2,5,25), include.lowest=T) )
```

```
# Resultat de la factoritzacio de cops que s'ha contactat al client en la campanya actual:
```

```
barplot(table(df$f.campaign), main="f.campaign - additional factors", ylab="frequency", las=1, cex.names=1.5)
```

f.campaign – additional factors



```
levels(df$f.campaign)<-paste0("f.campaign-", levels(df$f.campaign) )  
summary(df$f.campaign)
```

```
## f.campaign-[0,2] f.campaign-(2,5] f.campaign-(5,25]  
##           3392           1181           413
```

```
# P DAYS
```

```
df$f.pdays<-factor( cut(df$pdays, breaks=c(0, 18, 19), include.lowest=T) )
```

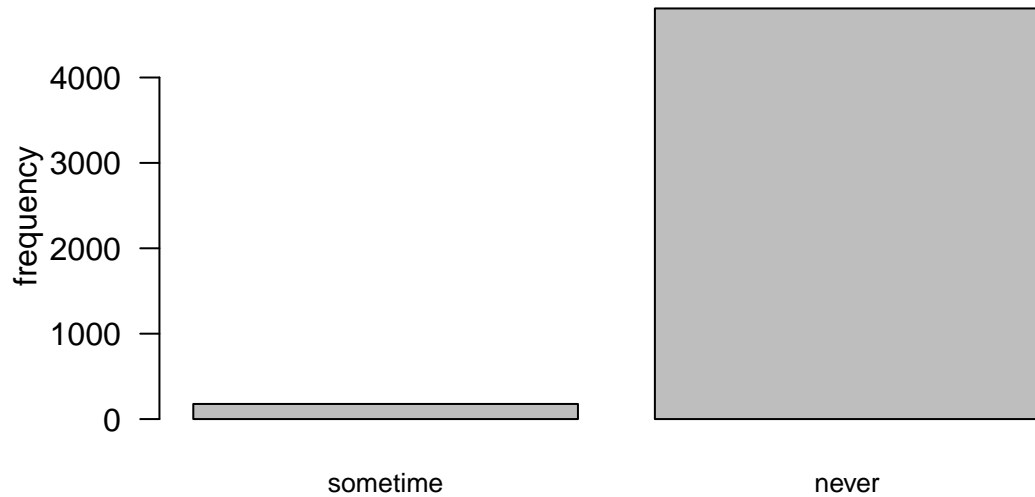
```
# Resultat de la factoritzacio dels dies que fa
```

```
# que s'ha contactat al client en una altra campanya:
```

```
levels(df$f.pdays)<-c("sometime", "never")
```

```
barplot(table(df$f.pdays), main="f.pdays - additional factors", ylab="frequency", las=1, cex.names=0.8)
```

f.pdays – additional factors



```
levels(df$f.pdays)<-paste0("f.pdays-", levels(df$f.pdays) )  
summary(df$f.pdays)
```

```
## f.pdays-sometime    f.pdays-never  
##                177                4809
```

```
# PREVIOUS
```

```
df$f.previous<-factor( cut(df$previous, breaks=c(-Inf, 0, +Inf), include.lowest=T) )
```

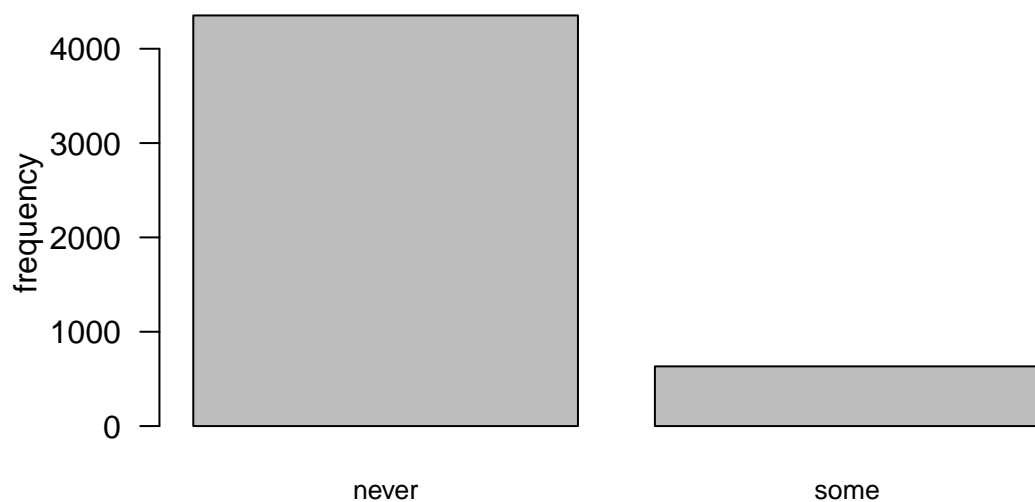
```
levels(df$f.previous)<-c("never", "some")
```

```
# Resultat de la factoritzacio de number of contacts performed
```

```
# before this campaign and for this client:
```

```
barplot(table(df$f.previous), main="f.previous - additional factors", ylab="frequency", las=1, cex.names=1.5)
```

f.previous – additional factors



```
levels(df$f.previous)<-paste0("f.previous-", levels(df$f.previous) )  
summary(df$f.previous)
```

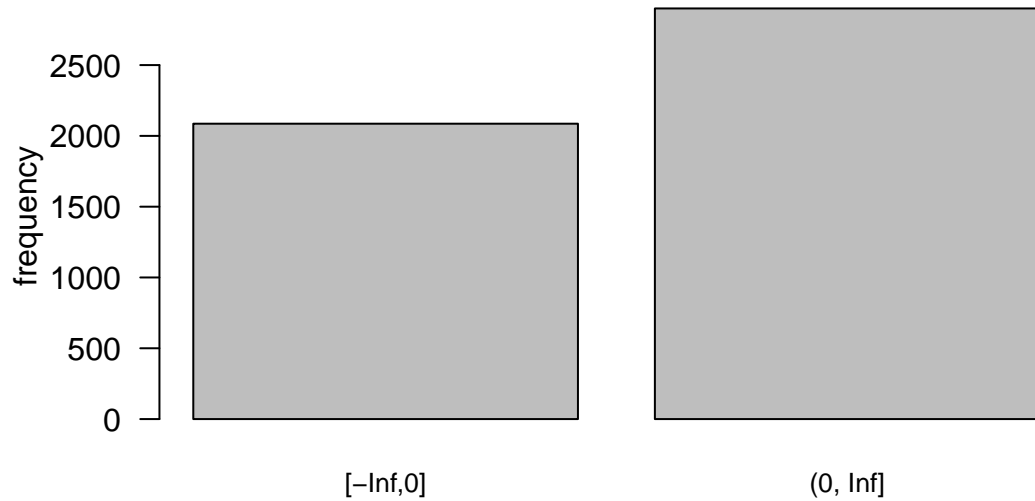
```
## f.previous-never  f.previous-some  
##           4353           633
```

```
# EMP.VAR.RATE
```

```
df$f.emp.var.rate <-factor( cut(df$emp.var.rate, breaks=c(-Inf, 0, +Inf), include.lowest=T) )
```

```
barplot(table(df$f.emp.var.rate), main="f.emp.var.rate - additional factors", ylab="frequency", las=1, col="gray")
```

f.emp.var.rate – additional factors



```
levels(df$f.emp.var.rate) <- paste0("f.emp.var.rate-", levels(df$f.emp.var.rate) )
summary(df$f.emp.var.rate)
```

```
## f.emp.var.rate-[-Inf,0] f.emp.var.rate-(0, Inf]
##                2086                2900
```

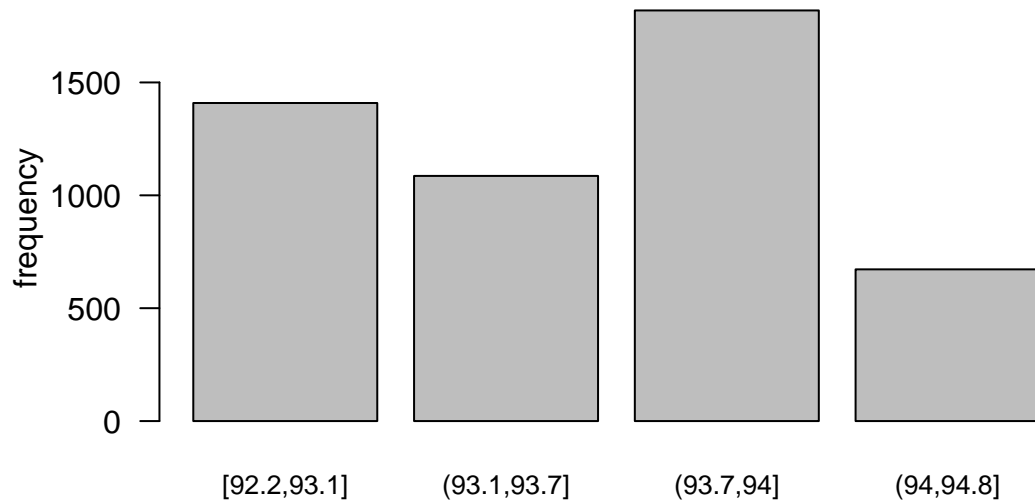
```
# CONS.PRICE.IDX
```

```
qulist <- quantile(df$cons.price.idx, seq(0,1,0.25), na.rm=TRUE)
```

```
df$f.cons.price.idx <- factor( cut(df$cons.price.idx , breaks=unique(qulist), include.lowest=T) )
```

```
barplot(table(df$f.cons.price.idx), main="f.cons.price.idx - additional factors", ylab="frequency", las=
```

f.cons.price.idx – additional factors



```
levels(df$f.cons.price.idx)<-paste0("f.cons.price.idx-", levels(df$f.cons.price.idx) )
summary(df$f.cons.price.idx)
```

```
## f.cons.price.idx-[92.2,93.1] f.cons.price.idx-(93.1,93.7]
##                1409                1086
## f.cons.price.idx-(93.7,94]  f.cons.price.idx-(94,94.8]
##                1819                672
```

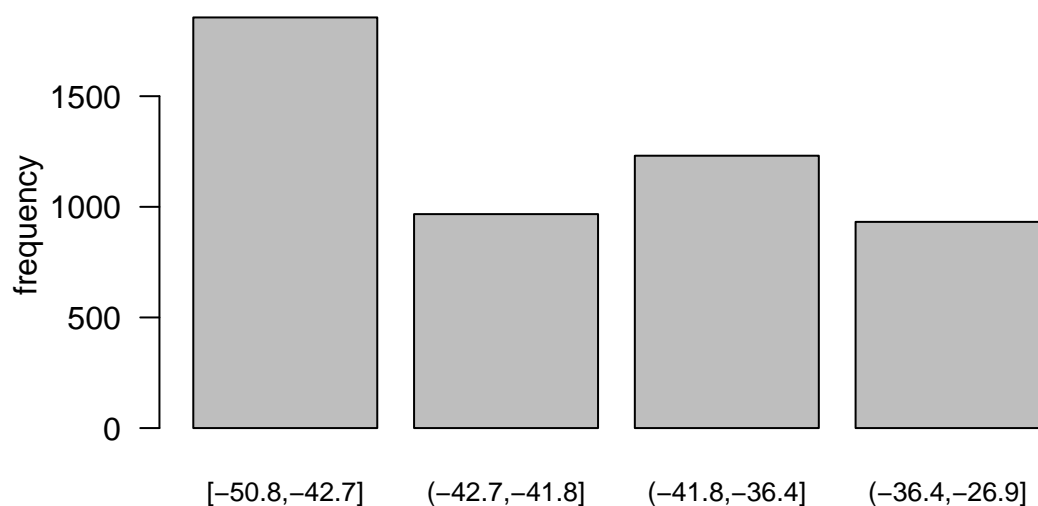
```
# CONS.CONF.IDX
```

```
qulist<-quantile(df$cons.conf.idx, seq(0,1,0.25), na.rm=TRUE)
```

```
df$f.cons.conf.idx <-factor( cut(df$cons.conf.idx , breaks=unique(qulist), include.lowest=T) )
```

```
barplot(table(df$f.cons.conf.idx), main="f.cons.conf.idx - additional factors", ylab="frequency", las=1)
```

f.cons.conf.idx – additional factors



```
levels(df$f.cons.conf.idx)<-paste0("f.cons.conf.idx-", levels(df$f.cons.conf.idx) )
summary(df$f.cons.conf.idx)
```

```
## f.cons.conf.idx-[-50.8,-42.7] f.cons.conf.idx-(-42.7,-41.8]
##                               1856                               967
## f.cons.conf.idx-(-41.8,-36.4] f.cons.conf.idx-(-36.4,-26.9]
##                               1231                               932
```

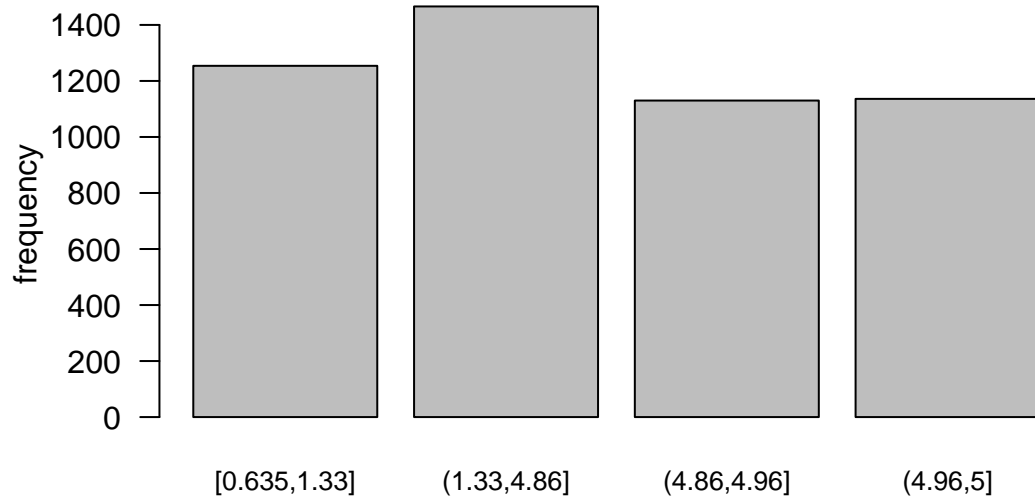
```
# EURIBOR3M
```

```
qulist<-quantile(df$euribor3m, seq(0,1,0.25), na.rm=TRUE)
```

```
df$f.euribor3m <-factor( cut(df$euribor3m , breaks=unique(qulist), include.lowest=T) )
```

```
barplot(table(df$f.euribor3m), main="f.euribor3m - additional factors", ylab="frequency", las=1, cex.na=
```


f.euribor3m – additional factors



```
levels(df$f.euribor3m)<-paste0("f.euribor3m-", levels(df$f.euribor3m) )
summary(df$f.euribor3m)
```

```
## f.euribor3m-[0.635,1.33]  f.euribor3m-(1.33,4.86]  f.euribor3m-(4.86,4.96]
##                        1254                        1466                        1130
##      f.euribor3m-(4.96,5]
##                        1136
```

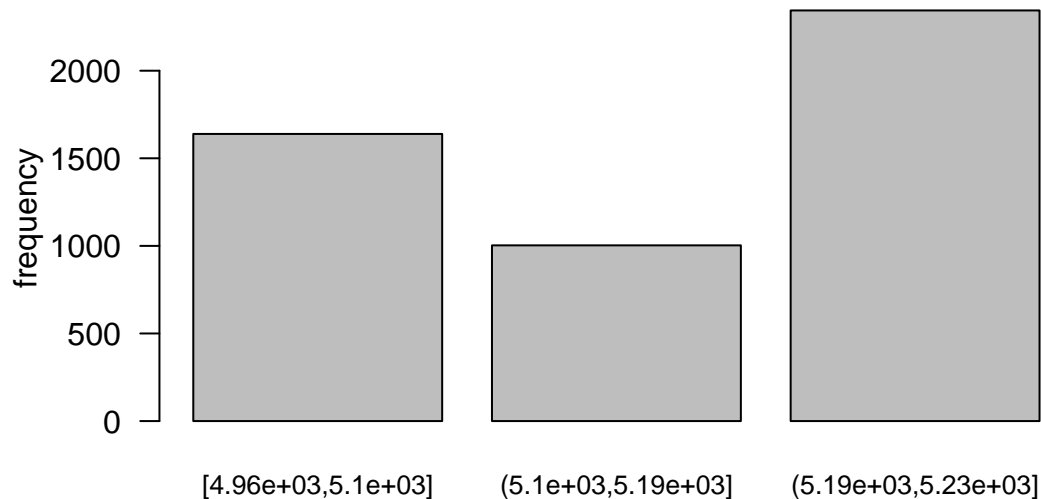
```
# NR.EMPLOYED
```

```
qulist<-quantile(df$nr.employed, seq(0,1,0.25), na.rm=TRUE)
```

```
df$f.nr.employed <-factor( cut(df$nr.employed , breaks=unique(qulist), include.lowest=T) )
```

```
barplot(table(df$f.nr.employed), main="f.nr.employed - additional factors", ylab="frequency", las=1, ce
```

f.nr.employed – additional factors



```
levels(df$f.nr.employed)<-paste0("f.nr.employed-", levels(df$f.nr.employed) )
summary(df$f.nr.employed)
```

```
## f.nr.employed-[4.96e+03,5.1e+03] f.nr.employed-(5.1e+03,5.19e+03]
##                                     1639                             1003
## f.nr.employed-(5.19e+03,5.23e+03]
##                                     2344
```

Llistat de variables continues i discretes:

```
vars<-names(df); vars
```

```
## [1] "age"           "job"           "marital"
## [4] "education"     "default"       "housing"
## [7] "loan"          "contact"       "month"
## [10] "day_of_week"   "duration"      "campaign"
## [13] "pdays"        "previous"      "poutcome"
## [16] "emp.var.rate"  "cons.price.idx" "cons.conf.idx"
## [19] "euribor3m"     "nr.employed"   "y"
## [22] "num_missings"  "num_outliers"  "num_errors"
## [25] "f.season"      "minutes"       "f.age"
## [28] "f.duration"    "f.campaign"    "f.pdays"
## [31] "f.previous"    "f.emp.var.rate" "f.cons.price.idx"
## [34] "f.cons.conf.idx" "f.euribor3m"   "f.nr.employed"
```

```
# Variables continues
```

```
vars_con<-names(df)[c(1, 11:14, 16:20)]; vars_con
```

```
## [1] "age"           "duration"      "campaign"      "pdays"
## [5] "previous"      "emp.var.rate"  "cons.price.idx" "cons.conf.idx"
```

```
## [9] "euribor3m"      "nr.employed"
# Variables discrete
vars_dis<-names(df)[c(2:10, 15, 21, 25, 27:36)]; vars_dis

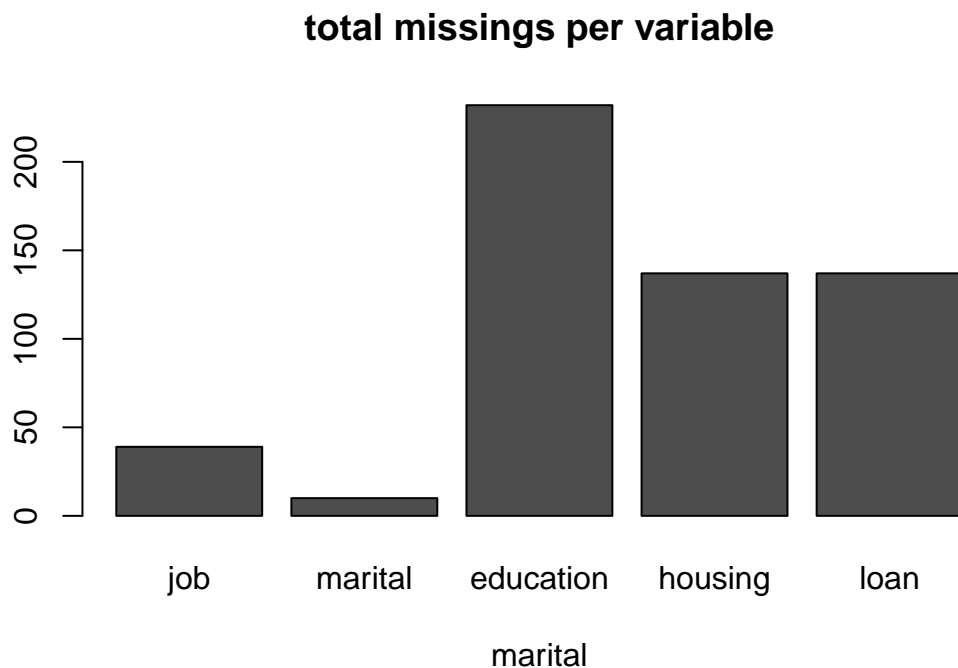
## [1] "job"      "marital"   "education"
## [4] "default"  "housing"   "loan"
## [7] "contact"  "month"     "day_of_week"
## [10] "poutcome" "y"         "f.season"
## [13] "f.age"     "f.duration" "f.campaign"
## [16] "f.pdays"  "f.previous" "f.emp.var.rate"
## [19] "f.cons.price.idx" "f.cons.conf.idx" "f.euribor3m"
## [22] "f.nr.employed"
```

DATA QUALITY REPORT:

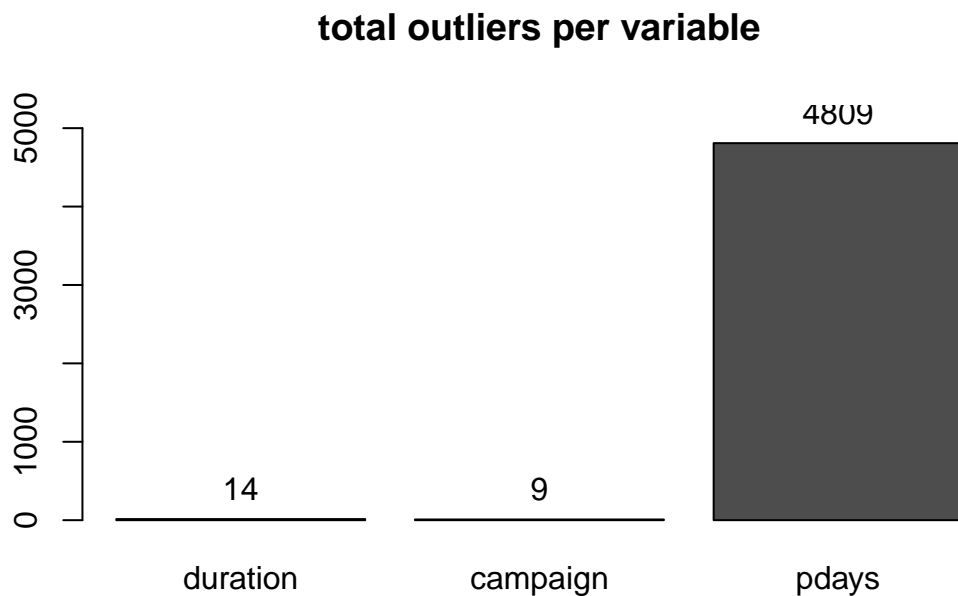
Per variable:

Nomes es mostren aquelles variables que tenen un valor diferent a 0 en el camp que expresa la grafica en concret.

```
barplot( t(c(missings[, c(2,3,4,6,7)])), main="total missings per variable", xlab="marital")
```



```
## Find a range of y's that'll leave sufficient space above the tallest bar
ylim <- c(0, 1.1*max( outliers[, c(11,12,13)] ))
## Plot, and store x-coordinates of bars in xx
data = t( c(outliers[, c(11,12,13)]) )
ylabls <- barplot(data, ylim = ylim, main="total outliers per variable")
## Add text at top of bars
text(x = ylabls, y = data, label = data, pos = 3)
```



```
#barplot( t(c(errors[, 13])), main="total errors per variable")
```

Per individu:

Cap individu en te mes d'un. Es mostra en format taula el numero d'individus que tenen 0 i/o 1 (o mes) missings, errors i outliers. Per ultim, es mostren alguns dels individus que han tingut algun outlier i que aquest ha estat imputat.

```
table(df$num_missings)
```

```
##
##    0    1    2    3
## 4591 241 149    5
```

```
table(df$num_errors)
```

```
##
##    0
## 4986
```

```
table(df$num_outliers)
```

```
##
##    0    2
## 4977    9
```

```
head(df[which(df$num_outliers>0), ], 2) #individus amb algun outlier
```

```
##      age      job      marital      education
## 5565  39  job-admin. marital-married education-university.degree
## 9014  30 job-blue-collar marital-married      education-basic.9y
##      default      housing      loan      contact      month
```

```

## 5565 default-no housing-yes loan-no contact-telephone month-may
## 9014 default-no housing-no loan-no contact-telephone month-jun
##          day_of_week duration campaign pdays previous
## 5565 day_of_week-1mon          14 2.556241    19      0
## 9014 day_of_week-4thu          53 2.585502    19      0
##          poutcome emp.var.rate cons.price.idx cons.conf.idx
## 5565 poutcome-nonexistent          1.1          93.994      -36.4
## 9014 poutcome-nonexistent          1.4          94.465      -41.8
##          euribor3m nr.employed    y num_missings num_outliers num_errors
## 5565    4.857        5191.0 y-no              0            2        0
## 9014    4.866        5228.1 y-no              0            2        0
##          f.season  minutes          f.age          f.duration
## 5565 season-spring 0.2333333 f.age-(38,47] f.duration-[5,101]
## 9014 season-summer 0.8833333 f.age-[18,32] f.duration-[5,101]
##          f.campaign    f.pdays    f.previous
## 5565 f.campaign-(2,5] f.pdays-never f.previous-never
## 9014 f.campaign-(2,5] f.pdays-never f.previous-never
##          f.emp.var.rate          f.cons.price.idx
## 5565 f.emp.var.rate-(0, Inf] f.cons.price.idx-(93.7,94]
## 9014 f.emp.var.rate-(0, Inf] f.cons.price.idx-(94,94.8]
##          f.cons.conf.idx          f.euribor3m
## 5565 f.cons.conf.idx-(-41.8,-36.4] f.euribor3m-(1.33,4.86]
## 9014 f.cons.conf.idx-(-42.7,-41.8] f.euribor3m-(4.86,4.96]
##          f.nr.employed
## 5565 f.nr.employed-(5.1e+03,5.19e+03]
## 9014 f.nr.employed-(5.19e+03,5.23e+03]

```

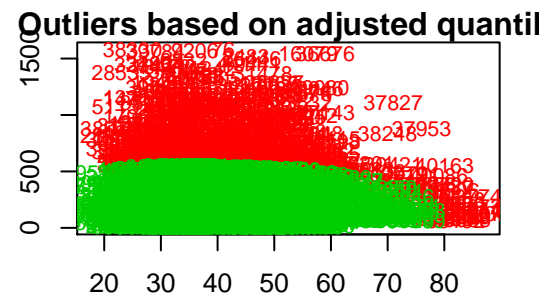
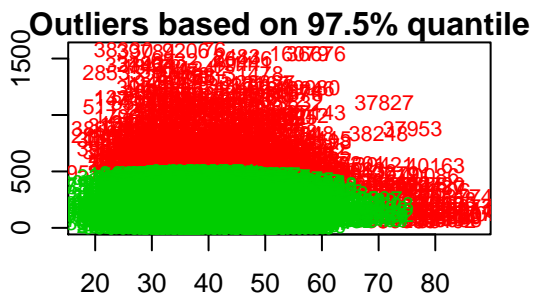
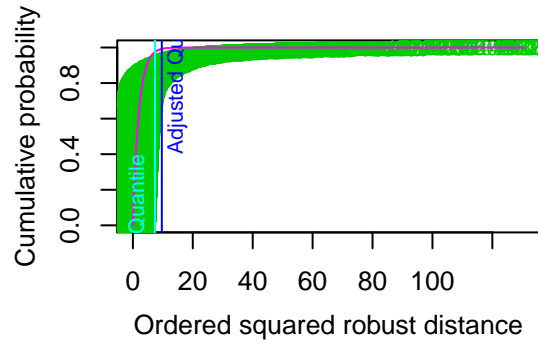
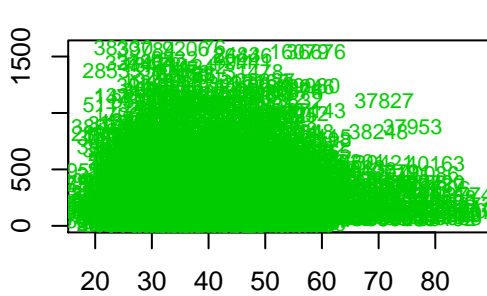
Outliers Multivariants:

No hem aconseguit trobar una configuració del `aq.plot` que ens doni una bona gràfica per a veure les distàncies de Mahalanobis i detectar outliers multivariants.

```

# Consider subset of numeric variables:
# summary(df[,vars_con])
vars_con_sub<-vars_con[c(1:2)]
x<-df[,vars_con_sub]
# aq.plot(x, delta=qchisq(0.995, df=ncol(x)) )
index <- data.frame(aq.plot(x, delta=qchisq(0.975, df=ncol(x)), quan=0.5, alpha=0.05))

```



```
table(index$outliers)
```

```
##
## FALSE TRUE
## 4429 557
```

IMPUTATION:

Factors:

De totes les variables discretes que hem analitzat, hem vist que algunes es podrien imputar fàcilment amb `imputeMCA()`, ja que els unknown (passats prèviament a NA) corresponen només una petita part de la mostra.

```
res.impf<-imputeMCA(df[,vars_dis], ncp=10)
```

```
# Original:
summary(df$marital)
```

```
## marital-divorced marital-married marital-single NA's
##           554           3046           1376           10
```

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

```
## loan-no loan-yes NA's
##    4080    769    137
```

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

```
## job-admin. job-blue-collar job-entrepreneur job-housemaid
##         1231         1151         154         135
```

```
##      job-management      job-retired job-self-employed      job-services
##           411           204           148           498
##      job-student      job-technician      job-unemployed      NA's
##           100           793           122           39
```

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

```
##      education-basic.4y      education-basic.6y
##           516           289
##      education-basic.9y      education-high.school
##           715           1168
## education-professional.course      education-university.degree
##           599           1468
##           NA's
##           231
```

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

```
##      housing-no housing-yes      NA's
##           2212           2637           137
```

```
# Amb dades imputades:
```

```
summary(res.impf$completeObs$marital)
```

```
##      marital-divorced      marital-married      marital-single
##           554           3055           1377
```

```
summary(res.impf$completeObs$loan)
```

```
##      loan-no loan-yes
##           4217           769
```

```
summary(res.impf$completeObs$job)
```

```
##      job-admin.      job-blue-collar      job-entrepreneur      job-housemaid
##           1246           1171           154           135
##      job-management      job-retired      job-self-employed      job-services
##           411           205           148           498
##      job-student      job-technician      job-unemployed
##           100           796           122
```

```
summary(res.impf$completeObs$education)
```

```
##      education-basic.4y      education-basic.6y
##           533           289
##      education-basic.9y      education-high.school
##           767           1218
## education-professional.course      education-university.degree
##           615           1564
```

```
summary(res.impf$completeObs$housing)
```

```
##      housing-no housing-yes
##           2261           2725
```

```
# Acceptem la imputacio:
```

```
df$marital<-res.impf$completeObs[, "marital"]
df$loan<-res.impf$completeObs[, "loan"]
df$job<-res.impf$completeObs[, "job"]
df$education<-res.impf$completeObs[, "education"]
```

```
df$housing<-res.impf$completeObs[, "housing"]
#summary(df[,vars_dis])
```

PROFILING:

CONTINUOUS DESCRIPTION - Numeric Target (Duration):

La funció d'R "condes" ens descriu la variable continua "duration" a partir d'altres variables quantitatives o de les variables categòriques. Això ho fa mitjançant els tres outputs diferenciats més avall; etiquetats com a "quant", "qual" i "category".

El primer dels quals (\$quant) ens mostra la correlació de la variable estudiada "duration" amb altres variables numèriques, mostrant només les correlacions que tenen un p-valor per sota del llindar o nivell de significació del 5% (en aquest cas). Com més petit és el p-valor, menys evidència hi ha de que la hipòtesi nul·la sigui certa i més segurs estem del rebuig de la hipòtesi nul·la. Aquesta hipòtesi nul·la H_0 afirma que la correlació o resultat obtingut és fruit d'una aleatorietat de les dades i no pot ser atribuïble a una causa específica. Per tant, a partir d'ara, direm que quan el p-valor està per sota del nivell de significació establert, els resultats són significatius.

Comentar que ens apareix el valor NA però no tenim cap valor en la nostra mostra (ho vam estar mirant a classe), tot i així no afecta al resultat obtingut, simplement l'obviem. De la mateixa manera obviem la correlació d'1 entre la duració de la trucada en segons i en minuts, ja que és una correlació perfecta deguda a una conversió d'unitats. Dit això, observem lleugeres correlacions negatives significatives (ordenades de més correlació positiva a no correlació i després a més correlació negativa) entre la duració de la trucada i la variable pdays, euribor3m, nr.employed i campaign. Es pot veure com la duració de la trucada augmenta com menys cops s'ha contactat al client en aquesta campanya (campaign), el qual és lògic perquè un client molt contactat estarà cansat ja de rebre trucades. També es pot veure com la duració de la trucada augmenta com menys dies fa que s'ha contactat a un client en relació a una campanya anterior (pdays), el que pot estar relacionat amb l'interès del client per les diferents campanyes actuals que se li han exposat. Finalment tenim dos indicadors socioeconòmics que tenen una lleugera correlació negativa amb la duració de la trucada.

El segon output (\$qual) ens mostra els factors (variables categòriques) que estan més relacionades amb la variable target "duration". Ens mostra els resultats significatius ordenats per factors de més a menys relacionats la duració. Obviant la discretització de la duració (f.duration) que obviament està molt relacionada, observem com la decisió final (y) del client a contractar un servei està força relacionada amb la duració d'una trucada. Molt menys relacionades (però lleugerament) ho estan les variables "f.campaign", "month", així com altres indicadors socioeconòmics.

El tercer output (\$category) ens indica una estimació de les unitats que la durada de la trucada està per sobre (+) o per sota (-) de la mitja global quan el registre pertany a la categoria en qüestió; ordenades per p-valor. Deixant de banda les categories de f.duration que són fruit de la discretització, pot veure com quan el producte es contracta (y=yes), la duració de la trucada està 148 segons per sobre, com era d'esperar en una contractació per telèfon. Altres resultats obtinguts interessants són que la duració de la trucada està 72 segons per sobre quan s'ha contactat amb el client en aquesta campanya 1 o 2 cops (f.campaign-[0,2]) i que també augmenta en 38 segons quan el resultat de la campanya anterior va ser positiu pel mateix client (poutcome=success). També podem destacar el mes d'abril (month-apr), en el qual les duracions de les trucades estan 28 segons per sobre de la mitja, o la primavera (season=spring) amb 18 segons per sobre de la mitja. D'altra banda podem veure com en el mes d'agost (month=aug) la duració de les trucades està 28 segons per sota la mitja, en el novembre (month=nov) 20 segons per sota, i que els clients que mai han estat contactats abans (f.pdays=never) estan 28 segons menys al telèfon que la mitja.

El oneway.test d'R ens compara si dues o més mostres de variables amb distribució normal tenen o no la mateixa mitjana (no cal assumir igualtat de variàncies pels grups implicats que es comparen). En aquest cas ens permet concloure que la mitjana de la durada de la trucada en els casos que s'ha contractat el servei és significativament diferent a la dels casos en els quals no s'ha contractat el servei. L'estadístic de contrast

segueix una distribució F de Fisher i pren el valor 447.7, que és molt significatiu ($p\text{-value} < 1e-16$).

```
pos_duration<-which(names(df)=="duration"); pos_duration
```

```
## [1] 11
```

```
condes(df, num.var=pos_duration, proba = 0.05)
```

```
## $quanti
```

	correlation	p.value
<NA>	NA	NA
minutes	1.00000000	0.000000e+00
pdays	-0.03190702	2.425821e-02
euribor3m	-0.03512962	1.311237e-02
num_outliers	-0.04065979	4.085021e-03
nr.employed	-0.04831097	6.438109e-04
campaign	-0.07479199	1.241586e-07

```
##
```

```
## $quali
```

	R2	p.value
f.duration	0.694658017	0.000000e+00
y	0.164777620	3.759496e-197
f.campaign	0.004516830	1.263332e-05
f.cons.conf.idx	0.004067507	1.465565e-04
f.nr.employed	0.002912867	6.975062e-04
f.cons.price.idx	0.003246051	1.031905e-03
f.season	0.002391413	2.566167e-03
month	0.005064462	2.674014e-03
f.euribor3m	0.002462249	6.473152e-03
poutcome	0.001851161	9.887924e-03
f.pdays	0.001211656	1.396985e-02
day_of_week	0.002352912	1.942616e-02

```
##
```

```
## $category
```

	Estimate	p.value
f.duration-(316,1.58e+03]	314.511430	0.000000e+00
y=yes	148.441504	3.759496e-197
season-spring	15.716194	5.877554e-04
poutcome-success	38.359032	5.480212e-03
f.campaign-[0,2]	20.816748	7.136472e-03
f.nr.employed-[4.96e+03,5.1e+03]	9.017147	8.355482e-03
f.cons.conf.idx-[-50.8,-42.7]	14.076002	1.238528e-02
f.pdays-sometime	21.670172	1.396985e-02
month-may	9.867780	1.599295e-02
f.cons.price.idx-(93.7,94]	11.621760	2.081111e-02
f.cons.conf.idx-(-41.8,-36.4]	16.349262	2.392080e-02
month-apr	27.731238	2.403940e-02
day_of_week-3wed	13.376659	4.495212e-02
education-high.school	11.950501	4.670195e-02
month-nov	-20.376410	4.421467e-02
f.duration-(177,316]	-12.420618	3.175920e-02
day_of_week-1mon	-15.133836	1.838350e-02
season-summer	-6.135532	1.752241e-02
f.pdays-never	-21.670172	1.396985e-02
f.cons.conf.idx-(-36.4,-26.9]	-14.862166	7.024095e-03

```
## f.cons.conf.idx-(-42.7,-41.8]      -15.563098  4.192506e-03
## f.euribor3m-(4.96,5]              -19.423787  1.079935e-03
## month-aug                         -28.383026  6.707022e-04
## f.nr.employed-(5.19e+03,5.23e+03] -16.466612  1.395228e-04
## f.cons.price.idx-(93.1,93.7]       -22.699701  8.027710e-05
## f.campaign-(5,25]                 -36.153827  2.638343e-06
## f.duration-(101,177]              -113.196196  4.416490e-92
## y-no                              -148.441504  3.759496e-197
## f.duration-[5,101]                -188.894616  8.444088e-278
```

```
# mitjana de la duracio per categoria de la duracio
# tapply(df$duration, df$f.duration, mean)
```

```
# duracio global
summary(df$duration)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      5.0   101.0   177.0   250.6   316.0  1580.0
```

```
# mitjana de la duracio per categoria de la y
tapply(df$duration, df$y, mean)
```

```
##      y-no    y-yes
## 217.4563 514.3393
```

```
oneway.test(df$duration~df$y)
```

```
##
## One-way analysis of means (not assuming equal variances)
##
## data: df$duration and df$y
## F = 447.7, num df = 1.00, denom df = 605.83, p-value < 2.2e-16
```

CATEGORICAL DESCRIPTION - Factor (Y, Final Decision):

La funció d'R "catdes" ens descriu la variable categòrica "y" a partir d'altres variables categòriques o de les variables quantitatives. Això ho fa mitjançant outputs diferenciats més avall. Notem que el nostre llindar de significació en aquest cas és del 0.025 per tal de limitar una mica la gran quantitat de resultats mostrats.

L'apartat "Link between the cluster variable and the categorical variables (chi-square test)" ens mostra les variables categòriques que han caracteritzat al factor "y" ordenades de més a menys caracterització del factor (de menys a més p-value). La columna "df" mostra els Degrees of Freedom, que corresponen amb el nombre de categories del factor menys 1. Les variables categòriques que han influenciat més en la decisió final (y) són la f.duration (però és una dada que s'obté a posteriori de la trucada, no ens serveix per a generar un perfil de client), f.pdays (nombre de dies des de l'últim contacte), poutcome (si la última campanya va ser acceptada per aquest client o no), el mes (month), previous (si havia estat contactat o no abans d'aquesta campanya), diferents indicadors socioeconòmics, contact (via de contacte), el job (feina), etc.

L'apartat "Description of each cluster by the categories" ens mostra per a cada categoria de la "y" (y-yes, y-no), una descripció de les variables categòriques per tal de poder estudiar-ne el seu enllaç. La primera columna Cla/Mod ens mostra el tant per cent de la categoria de la fila indicada que pertany a la resposta (y) corresponent. D'altra banda, per a una resposta (y-yes, y-no) fixada, la segona columna Mod/Cla ens mostra el tant per cent de valors de la fila corresponent que pertanyen a la resposta fixada. Aquesta columna pot ésser comparada amb la columna Global i d'aquesta manera trobar sobrerrepresentacions en algunes categories, ja que la tercera columna ens indica el tant per cent de valors que representa la categoria sense tenir en compte la resposta (y) fixada. Per acabar, v.test ens indica si la categoria de la fila es troba sobrerrepresentada (v.test>0) o infrarepresentada (v.test<0) dins una resposta (y) fixada. Al cluster "y-no", podem veure com el

fet de no haver contactat mai al client abans o fer-ho a través del telèfon fixe, estan sobrerrepresentats en la resposta (y) negativa, pel que no són bones caracteritzacions d'individu a l'hora d'acceptar el producte. Al cluster “y-yes”, podem veure una lleguera sobrerrepresentació dels individus que van ser contactats fa menys de 7 dies en altres campanyes i d'aquells que una campanya anterior va resultar exitosa, el que es pot interpretar com que en aquests casos el client es més propens a donar un si com a resposta. Així com el fet de realitzar la trucada al telèfon mòbil o altres categories, que estan sobrerrepresentades i poden ser observades en la llista donada per R. També hi ha certs valors socioeconòmics que estan més o menys representats en la resposta positiva que en la negativa del client, i viceversa.

L'apartat “Link between the cluster variable and the quantitative variables” ens mostra les variables quantitatives que han caracteritzat al factor “y” ordenades de més a menys caracterització del factor (de menys a més correlació). Les variables quantitatives que han influenciat més en la decisió final (y) són la duration i minutes (però són dades que s'obtenen a posteriori de la trucada, no ens serveixen per a generar un perfil de client), pdays (nombre de dies des de l'últim contacte), previous (si havia estat contactat o no abans d'aquesta campanya), diferents indicadors socioeconòmics, etc.

L'apartat “Description of each cluster by quantitative variables”. D'aquesta part de l'anàlisi no en podem extreure informació dels individus que conformen el cluster “y-no”, donat que els valors que es presenten de les categories dins el cluster i de manera general no presenten una diversificació notable. Per altra banda del cluster “y-yes” si que en podem extreure informació, podem veure que la mitjana de la duració de les trucades dels individus del cluster duplica la mitjana global (donat que la duració és una conseqüència del desenvolupament de la trucada). Altres factors com l'euríbor o la taxa de variació de la ocupació també tenen un impacte en la decisió final.

```
pos_y<-which(names(df)=="y"); pos_y
```

```
## [1] 21
```

```
catdes(df, num.var=pos_y, proba = 0.001)
```

```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## =====
##                p.value df
## f.duration      1.461354e-121  3
## f.pdays         1.117730e-99  1
## poutcome        3.053387e-95  2
## f.nr.employed   1.703080e-89  2
## f.euribor3m     5.470503e-79  3
## month           1.690776e-65  9
## f.emp.var.rate  4.141030e-61  1
## f.cons.price.idx 5.572278e-38  3
## f.previous      6.896103e-38  1
## f.cons.conf.idx  4.786677e-23  3
## contact         2.110136e-21  1
## job             4.816597e-17 10
## default         9.768051e-13  1
## f.season        7.457294e-10  2
## f.age           7.936723e-09  3
## education       8.419496e-08  5
## marital         6.658281e-05  2
## f.campaign      5.052372e-04  2
##
## Description of each cluster by the categories
## =====
## $`y-no`
```

	Cla/Mod	Mod/Cla
##		
## f.emp.var.rate=f.emp.var.rate-(0, Inf]	95.06897	62.2488146
## f.pdays=f.pdays-never	90.64255	98.4195078
## f.duration=f.duration-[5,101]	98.96166	27.9747121
## f.nr.employed=f.nr.employed-(5.19e+03,5.23e+03]	94.70990	50.1241815
## f.previous=f.previous-never	91.01769	89.4558591
## poutcome=poutcome-nonexistent	91.01769	89.4558591
## contact=contact-telephone	94.31330	39.6929329
## f.cons.price.idx=f.cons.price.idx-(93.7,94]	94.11765	38.6543238
## f.duration=f.duration-(101,177]	95.41432	26.7780537
## f.nr.employed=f.nr.employed-(5.1e+03,5.19e+03]	96.11167	21.7656356
## f.cons.conf.idx=f.cons.conf.idx-(-42.7,-41.8]	96.07032	20.9753895
## default=default-unknown	95.05814	22.1494694
## month=month-may	93.33716	36.6899977
## f.euribor3m=f.euribor3m-(4.86,4.96]	94.51327	24.1137954
## f.euribor3m=f.euribor3m-(4.96,5]	94.36620	24.2041093
## job=job-blue-collar	93.85141	24.8137277
## f.euribor3m=f.euribor3m-(1.33,4.86]	92.70123	30.6841273
## f.cons.price.idx=f.cons.price.idx-(93.1,93.7]	92.90976	22.7816663
## f.age=f.age-(38,47]	92.54098	25.4910815
## education=education-basic.9y	93.08996	16.1210205
## f.campaign=f.campaign-(5,25]	94.18886	8.7830210
## marital=marital-single	85.69354	26.6425830
## job=job-retired	78.53659	3.6351321
## f.age=f.age-[18,32]	85.35503	26.0555430
## job=job-student	70.00000	1.5804922
## month=month-apr	78.70968	5.5091443
## education=education-university.degree	85.23018	30.0970874
## f.season=season-autumnwinter	81.62084	12.7342515
## month=month-sep	57.37705	0.7902461
## month=month-mar	57.57576	0.8579815
## f.cons.conf.idx=f.cons.conf.idx-(-36.4,-26.9]	81.22318	17.0918943
## default=default-no	87.20283	77.8505306
## month=month-oct	54.63918	1.1966584
## contact=contact-cellular	85.55413	60.3070671
## f.cons.price.idx=f.cons.price.idx-[92.2,93.1]	80.48261	25.6039738
## f.previous=f.previous-some	73.77567	10.5441409
## poutcome=poutcome-success	37.82051	1.3321291
## f.pdays=f.pdays-sometime	39.54802	1.5804922
## f.emp.var.rate=f.emp.var.rate-[-Inf,0]	80.15340	37.7511854
## f.euribor3m=f.euribor3m-[0.635,1.33]	74.16268	20.9979679
## f.nr.employed=f.nr.employed-[4.96e+03,5.1e+03]	75.96095	28.1101829
## f.duration=f.duration-(316,1.58e+03]	71.46302	20.0722511
##	Global	p.value
## f.emp.var.rate=f.emp.var.rate-(0, Inf]	58.162856	3.963838e-61
## f.pdays=f.pdays-never	96.450060	2.410684e-59
## f.duration=f.duration-[5,101]	25.110309	4.912295e-55
## f.nr.employed=f.nr.employed-(5.19e+03,5.23e+03]	47.011633	2.158488e-37
## f.previous=f.previous-never	87.304452	1.438650e-30
## poutcome=poutcome-nonexistent	87.304452	1.438650e-30
## contact=contact-telephone	37.384677	3.447929e-23
## f.cons.price.idx=f.cons.price.idx-(93.7,94]	36.482150	7.057265e-21
## f.duration=f.duration-(101,177]	24.929803	3.376202e-20
## f.nr.employed=f.nr.employed-(5.1e+03,5.19e+03]	20.116326	1.424235e-19

## f.cons.conf.idx=f.cons.conf.idx-(-42.7,-41.8]	19.394304	1.401017e-18
## default=default-unknown	20.697954	1.230324e-14
## month=month-may	34.917770	1.726364e-14
## f.euribor3m=f.euribor3m-(4.86,4.96]	22.663458	1.693548e-13
## f.euribor3m=f.euribor3m-(4.96,5]	22.783795	6.639818e-13
## job=job-blue-collar	23.485760	4.385552e-11
## f.euribor3m=f.euribor3m-(1.33,4.86]	29.402327	6.796806e-09
## f.cons.price.idx=f.cons.price.idx-(93.1,93.7]	21.780987	4.701642e-07
## f.age=f.age-(38,47]	24.468512	9.135370e-07
## education=education-basic.9y	15.383073	1.890871e-05
## f.campaign=f.campaign-(5,25]	8.283193	1.084374e-04
## marital=marital-single	27.617329	2.164341e-05
## job=job-retired	4.111512	1.457985e-05
## f.age=f.age-[18,32]	27.115925	3.567657e-06
## job=job-student	2.005616	2.508620e-07
## month=month-apr	6.217409	1.047741e-07
## education=education-university.degree	31.367830	9.372698e-08
## f.season=season-autumnwinter	13.858805	1.173628e-09
## month=month-sep	1.223426	3.276634e-10
## month=month-mar	1.323706	7.597160e-11
## f.cons.conf.idx=f.cons.conf.idx-(-36.4,-26.9]	18.692339	1.352020e-14
## default=default-no	79.302046	1.230324e-14
## month=month-oct	1.945447	8.959508e-18
## contact=contact-cellular	62.615323	3.447929e-23
## f.cons.price.idx=f.cons.price.idx-[92.2,93.1]	28.259126	3.335427e-29
## f.previous=f.previous-some	12.695548	1.438650e-30
## poutcome=poutcome-success	3.128761	2.946325e-55
## f.pdays=f.pdays-sometime	3.549940	2.410684e-59
## f.emp.var.rate=f.emp.var.rate-[-Inf,0]	41.837144	3.963838e-61
## f.euribor3m=f.euribor3m-[0.635,1.33]	25.150421	3.042037e-70
## f.nr.employed=f.nr.employed-[4.96e+03,5.1e+03]	32.872042	1.759629e-84
## f.duration=f.duration-(316,1.58e+03]	24.949860	1.316787e-95
##		v.test
## f.emp.var.rate=f.emp.var.rate-(0, Inf]	16.495331	
## f.pdays=f.pdays-never	16.245323	
## f.duration=f.duration-[5,101]	15.625090	
## f.nr.employed=f.nr.employed-(5.19e+03,5.23e+03]	12.778626	
## f.previous=f.previous-never	11.492513	
## poutcome=poutcome-nonexistent	11.492513	
## contact=contact-telephone	9.918824	
## f.cons.price.idx=f.cons.price.idx-(93.7,94]	9.372891	
## f.duration=f.duration-(101,177]	9.206281	
## f.nr.employed=f.nr.employed-(5.1e+03,5.19e+03]	9.050417	
## f.cons.conf.idx=f.cons.conf.idx-(-42.7,-41.8]	8.797336	
## default=default-unknown	7.712857	
## month=month-may	7.669524	
## f.euribor3m=f.euribor3m-(4.86,4.96]	7.370998	
## f.euribor3m=f.euribor3m-(4.96,5]	7.186654	
## job=job-blue-collar	6.590430	
## f.euribor3m=f.euribor3m-(1.33,4.86]	5.795870	
## f.cons.price.idx=f.cons.price.idx-(93.1,93.7]	5.038105	
## f.age=f.age-(38,47]	4.909404	
## education=education-basic.9y	4.277402	
## f.campaign=f.campaign-(5,25]	3.870893	

## marital=marital-single	-4.247225		
## job=job-retired	-4.334942		
## f.age=f.age-[18,32]	-4.635100		
## job=job-student	-5.157057		
## month=month-apr	-5.318243		
## education=education-university.degree	-5.338484		
## f.season=season-autumnwinter	-6.083806		
## month=month-sep	-6.285090		
## month=month-mar	-6.508368		
## f.cons.conf.idx=f.cons.conf.idx-(-36.4,-26.9]	-7.700814		
## default=default-no	-7.712857		
## month=month-oct	-8.586582		
## contact=contact-cellular	-9.918824		
## f.cons.price.idx=f.cons.price.idx-[92.2,93.1]	-11.217779		
## f.previous=f.previous-some	-11.492513		
## poutcome=poutcome-success	-15.657639		
## f.pdays=f.pdays-sometime	-16.245323		
## f.emp.var.rate=f.emp.var.rate-[-Inf,0]	-16.495331		
## f.euribor3m=f.euribor3m-[0.635,1.33]	-17.718064		
## f.nr.employed=f.nr.employed-[4.96e+03,5.1e+03]	-19.475855		
## f.duration=f.duration-(316,1.58e+03]	-20.746562		
##			
## \$`y=yes`			
##		Cla/Mod	Mod/Cla
## f.duration=f.duration-(316,1.58e+03]	28.536977	63.734291	
## f.nr.employed=f.nr.employed-[4.96e+03,5.1e+03]	24.039048	70.736086	
## f.euribor3m=f.euribor3m-[0.635,1.33]	25.837321	58.168761	
## f.emp.var.rate=f.emp.var.rate-[-Inf,0]	19.846596	74.326750	
## f.pdays=f.pdays-sometime	60.451977	19.210054	
## poutcome=poutcome-success	62.179487	17.414722	
## f.previous=f.previous-some	26.224329	29.802513	
## f.cons.price.idx=f.cons.price.idx-[92.2,93.1]	19.517388	49.371634	
## contact=contact-cellular	14.445868	80.969479	
## month=month-oct	45.360825	7.899461	
## default=default-no	12.797167	90.843806	
## f.cons.conf.idx=f.cons.conf.idx-(-36.4,-26.9]	18.776824	31.418312	
## month=month-mar	42.424242	5.026930	
## month=month-sep	42.622951	4.667864	
## f.season=season-autumnwinter	18.379161	22.800718	
## education=education-university.degree	14.769821	41.472172	
## month=month-apr	21.290323	11.849192	
## job=job-student	30.000000	5.385996	
## f.age=f.age-[18,32]	14.644970	35.547576	
## job=job-retired	21.463415	7.899461	
## marital=marital-single	14.306463	35.368043	
## f.campaign=f.campaign-(5,25]	5.811138	4.308797	
## education=education-basic.9y	6.910039	9.515260	
## f.age=f.age-(38,47]	7.459016	16.337522	
## f.cons.price.idx=f.cons.price.idx-(93.1,93.7]	7.090239	13.824057	
## f.euribor3m=f.euribor3m-(1.33,4.86]	7.298772	19.210054	
## job=job-blue-collar	6.148591	12.926391	
## f.euribor3m=f.euribor3m-(4.96,5]	5.633803	11.490126	
## f.euribor3m=f.euribor3m-(4.86,4.96]	5.486726	11.131059	
## month=month-may	6.662837	20.825853	

## default=default-unknown	4.941860	9.156194
## f.cons.conf.idx=f.cons.conf.idx-(-42.7,-41.8]	3.929679	6.822262
## f.nr.employed=f.nr.employed-(5.1e+03,5.19e+03]	3.888335	7.001795
## f.duration=f.duration-(101,177]	4.585680	10.233393
## f.cons.price.idx=f.cons.price.idx-(93.7,94]	5.882353	19.210054
## contact=contact-telephone	5.686695	19.030521
## f.previous=f.previous-never	8.982311	70.197487
## poutcome=poutcome-nonexistent	8.982311	70.197487
## f.nr.employed=f.nr.employed-(5.19e+03,5.23e+03]	5.290102	22.262118
## f.duration=f.duration-[5,101]	1.038339	2.333932
## f.pdays=f.pdays-never	9.357455	80.789946
## f.emp.var.rate=f.emp.var.rate-(0, Inf]	4.931034	25.673250
##	Global	p.value
## f.duration=f.duration-(316,1.58e+03]	24.949860	1.316787e-95
## f.nr.employed=f.nr.employed-[4.96e+03,5.1e+03]	32.872042	1.759629e-84
## f.euribor3m=f.euribor3m-[0.635,1.33]	25.150421	3.042037e-70
## f.emp.var.rate=f.emp.var.rate-[-Inf,0]	41.837144	3.963838e-61
## f.pdays=f.pdays-sometime	3.549940	2.410684e-59
## poutcome=poutcome-success	3.128761	2.946325e-55
## f.previous=f.previous-some	12.695548	1.438650e-30
## f.cons.price.idx=f.cons.price.idx-[92.2,93.1]	28.259126	3.335427e-29
## contact=contact-cellular	62.615323	3.447929e-23
## month=month-oct	1.945447	8.959508e-18
## default=default-no	79.302046	1.230324e-14
## f.cons.conf.idx=f.cons.conf.idx-(-36.4,-26.9]	18.692339	1.352020e-14
## month=month-mar	1.323706	7.597160e-11
## month=month-sep	1.223426	3.276634e-10
## f.season=season-autumnwinter	13.858805	1.173628e-09
## education=education-university.degree	31.367830	9.372698e-08
## month=month-apr	6.217409	1.047741e-07
## job=job-student	2.005616	2.508620e-07
## f.age=f.age-[18,32]	27.115925	3.567657e-06
## job=job-retired	4.111512	1.457985e-05
## marital=marital-single	27.617329	2.164341e-05
## f.campaign=f.campaign-(5,25]	8.283193	1.084374e-04
## education=education-basic.9y	15.383073	1.890871e-05
## f.age=f.age-(38,47]	24.468512	9.135370e-07
## f.cons.price.idx=f.cons.price.idx-(93.1,93.7]	21.780987	4.701642e-07
## f.euribor3m=f.euribor3m-(1.33,4.86]	29.402327	6.796806e-09
## job=job-blue-collar	23.485760	4.385552e-11
## f.euribor3m=f.euribor3m-(4.96,5]	22.783795	6.639818e-13
## f.euribor3m=f.euribor3m-(4.86,4.96]	22.663458	1.693548e-13
## month=month-may	34.917770	1.726364e-14
## default=default-unknown	20.697954	1.230324e-14
## f.cons.conf.idx=f.cons.conf.idx-(-42.7,-41.8]	19.394304	1.401017e-18
## f.nr.employed=f.nr.employed-(5.1e+03,5.19e+03]	20.116326	1.424235e-19
## f.duration=f.duration-(101,177]	24.929803	3.376202e-20
## f.cons.price.idx=f.cons.price.idx-(93.7,94]	36.482150	7.057265e-21
## contact=contact-telephone	37.384677	3.447929e-23
## f.previous=f.previous-never	87.304452	1.438650e-30
## poutcome=poutcome-nonexistent	87.304452	1.438650e-30
## f.nr.employed=f.nr.employed-(5.19e+03,5.23e+03]	47.011633	2.158488e-37
## f.duration=f.duration-[5,101]	25.110309	4.912295e-55
## f.pdays=f.pdays-never	96.450060	2.410684e-59

```

## f.emp.var.rate=f.emp.var.rate-(0, Inf] 58.162856 3.963838e-61
## v.test
## f.duration=f.duration-(316,1.58e+03] 20.746562
## f.nr.employed=f.nr.employed-[4.96e+03,5.1e+03] 19.475855
## f.euribor3m=f.euribor3m-[0.635,1.33] 17.718064
## f.emp.var.rate=f.emp.var.rate-[-Inf,0] 16.495331
## f.pdays=f.pdays-sometime 16.245323
## poutcome=poutcome-success 15.657639
## f.previous=f.previous-some 11.492513
## f.cons.price.idx=f.cons.price.idx-[92.2,93.1] 11.217779
## contact=contact-cellular 9.918824
## month=month-oct 8.586582
## default=default-no 7.712857
## f.cons.conf.idx=f.cons.conf.idx-(-36.4,-26.9] 7.700814
## month=month-mar 6.508368
## month=month-sep 6.285090
## f.season=season-autumnwinter 6.083806
## education=education-university.degree 5.338484
## month=month-apr 5.318243
## job=job-student 5.157057
## f.age=f.age-[18,32] 4.635100
## job=job-retired 4.334942
## marital=marital-single 4.247225
## f.campaign=f.campaign-(5,25] -3.870893
## education=education-basic.9y -4.277402
## f.age=f.age-(38,47] -4.909404
## f.cons.price.idx=f.cons.price.idx-(93.1,93.7] -5.038105
## f.euribor3m=f.euribor3m-(1.33,4.86] -5.795870
## job=job-blue-collar -6.590430
## f.euribor3m=f.euribor3m-(4.96,5] -7.186654
## f.euribor3m=f.euribor3m-(4.86,4.96] -7.370998
## month=month-may -7.669524
## default=default-unknown -7.712857
## f.cons.conf.idx=f.cons.conf.idx-(-42.7,-41.8] -8.797336
## f.nr.employed=f.nr.employed-(5.1e+03,5.19e+03] -9.050417
## f.duration=f.duration-(101,177] -9.206281
## f.cons.price.idx=f.cons.price.idx-(93.7,94] -9.372891
## contact=contact-telephone -9.918824
## f.previous=f.previous-never -11.492513
## poutcome=poutcome-nonexistent -11.492513
## f.nr.employed=f.nr.employed-(5.19e+03,5.23e+03] -12.778626
## f.duration=f.duration-[5,101] -15.625090
## f.pdays=f.pdays-never -16.245323
## f.emp.var.rate=f.emp.var.rate-(0, Inf] -16.495331
##
##
## Link between the cluster variable and the quantitative variables
## =====
## Eta2 P-value
## duration 0.164777620 3.759496e-197
## minutes 0.164777620 3.759496e-197
## nr.employed 0.121012601 8.238443e-142
## euribor3m 0.090010720 3.115343e-104
## pdays 0.086552345 4.048268e-100

```



```

## emp.var.rate    0.085417483  8.992557e-99
## previous        0.042523921  5.101307e-49
## cons.price.idx  0.018386453  6.794885e-22
## cons.conf.idx   0.004669195  1.369222e-06
## campaign        0.004489048  2.189058e-06
## <NA>            NA          NA
##
## Description of each cluster by quantitative variables
## =====
## $`y-no`
##           v.test Mean in category Overall mean sd in category
## nr.employed    24.561104    5175.3298261 5166.47621340    64.3842715
## euribor3m      21.182621      3.7992890   3.61448034    1.6425449
## pdays          20.771698     18.7918266   18.52647413    1.6986882
## emp.var.rate   20.635071      0.2287424   0.06446049    1.4946001
## cons.price.idx  9.573739     93.6004884   93.57245006    0.5619158
## campaign       4.730529      2.5940749   2.53512993    2.5654605
## cons.conf.idx  -4.824514    -40.5398961 -40.42591256    4.4454152
## previous      -14.559593      0.1255362   0.15984757    0.4004406
## duration      -28.660364     217.4563107 250.62194144   191.6321071
## minutes       -28.660364      3.6242718   4.17703236    3.1938685
##           Overall sd      p.value
## nr.employed    71.7679377 3.291367e-133
## euribor3m      1.7370025 1.381286e-99
## pdays          2.5433666 7.804981e-96
## emp.var.rate   1.5850448 1.329502e-94
## cons.price.idx  0.5830800 1.031083e-21
## campaign       2.4808187 2.239356e-06
## cons.conf.idx  4.7037753 1.403451e-06
## previous       0.4691873 5.075919e-48
## duration      230.3904064 1.190744e-180
## minutes        3.8398401 1.190744e-180
##
## $`y-yes`
##           v.test Mean in category Overall mean sd in category
## minutes       28.660364      8.572322    4.17703236    5.3967235
## duration      28.660364     514.339318 250.62194144   323.8034093
## previous      14.559593      0.432675    0.15984757    0.7821222
## cons.conf.idx  4.824514    -39.519569 -40.42591256    6.3242738
## campaign      -4.730529      2.066427   2.53512993    1.5845655
## cons.price.idx -9.573739     93.349503   93.57245006    0.6904449
## emp.var.rate  -20.635071     -1.241831   0.06446049    1.6751620
## pdays         -20.771698     16.416517   18.52647413    5.4725311
## euribor3m     -21.182621      2.144969    3.61448034    1.7676126
## nr.employed   -24.561104     5096.076481 5166.47621340   86.9764988
##           Overall sd      p.value
## minutes        3.8398401 1.190744e-180
## duration      230.3904064 1.190744e-180
## previous       0.4691873 5.075919e-48
## cons.conf.idx  4.7037753 1.403451e-06
## campaign       2.4808187 2.239356e-06
## cons.price.idx  0.5830800 1.031083e-21
## emp.var.rate   1.5850448 1.329502e-94
## pdays          2.5433666 7.804981e-96

```

```
## euribor3m      1.7370025  1.381286e-99
## nr.employed    71.7679377  3.291367e-133
```