

# Course Practical Assignment - Final Delivery (Part 1) - 11th June 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
##                      (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:

```

columnes <- 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(columnes), nrow = 1))
colnames(errors)<-columnes

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

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

# 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 chunck posterior d'Imputation).

## VARIABLES QUALITATIVE:

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

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

## Job

Els “unknowns” seran imputats mes endavant automaticament.

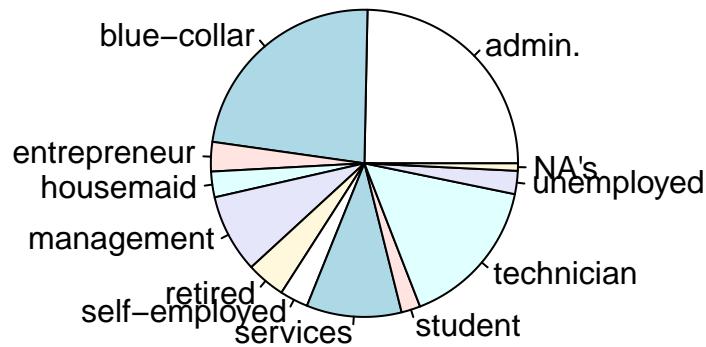
```
# Jobs "unknown" will be a missing value (set to NA):
sel<-which(df$job=="unknown");length(sel)

## [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

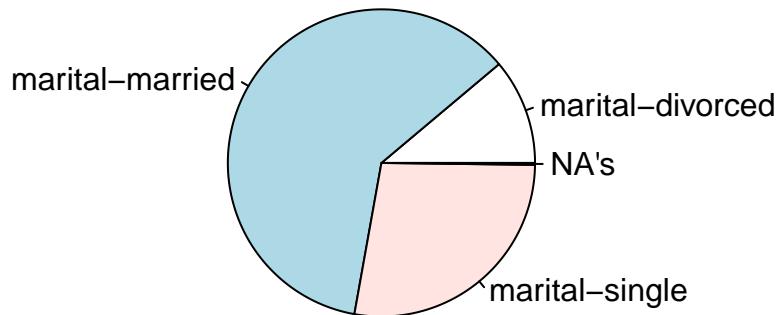
# Missings:
miss<-which(is.na(df$marital));
missings$marital<-length(miss); length(miss)

## [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)

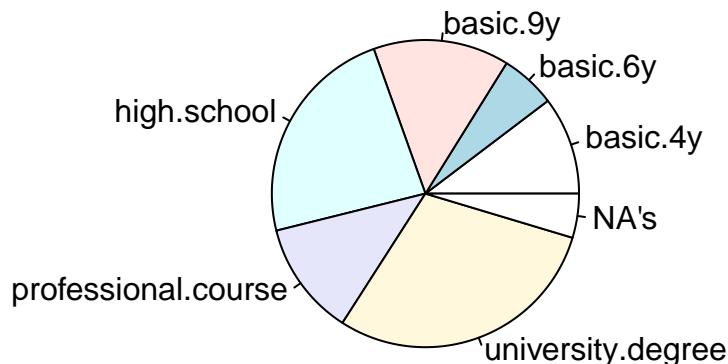
## [1] 3
df[sel, "education"]<-"basic.4y"

# 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.

```



```

## Housing

Els “unknowns” seran imputats mes endavant automaticament.

```

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

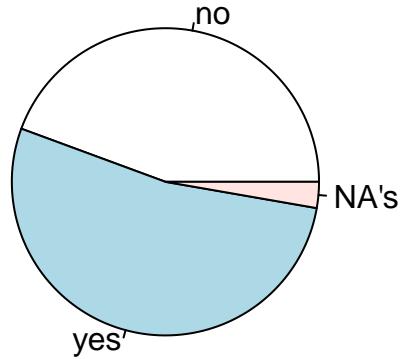
## [1] 137
df$housing[sel]<-NA

# 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
```

```
# Missing:
```

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

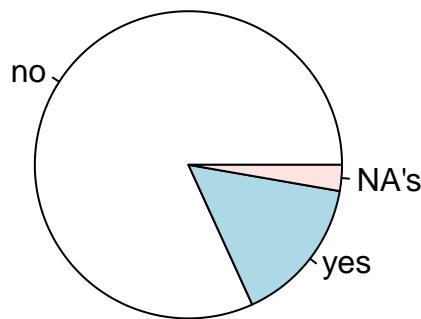
```
## [1] 137
```

```
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

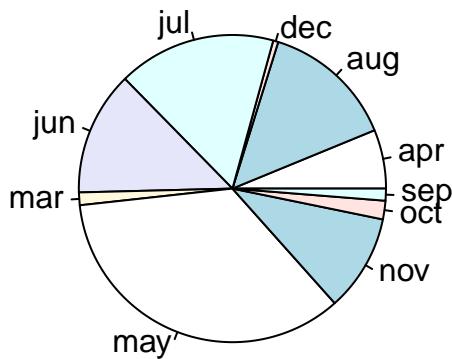
```
miss<-which(is.na(df$month));
missings$month<-length(miss); length(miss)
```

```

## [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

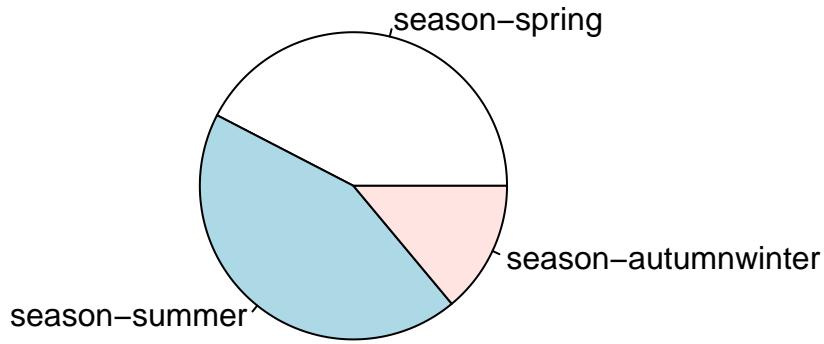
# 2 level - summer
sel<-which(df$month %in% c("month-jun","month-jul","month-aug"))
df$f.season[sel] <-2

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")

```

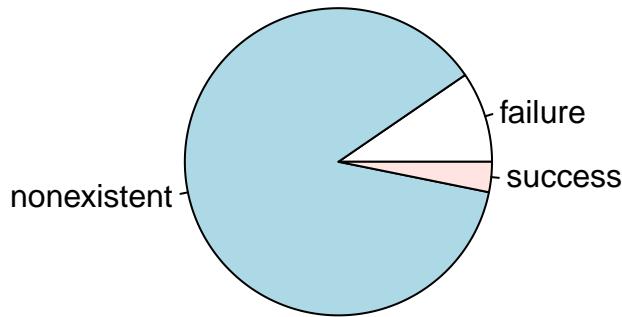
```

##           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:

Funció de gran utilitat per a la detecció 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)

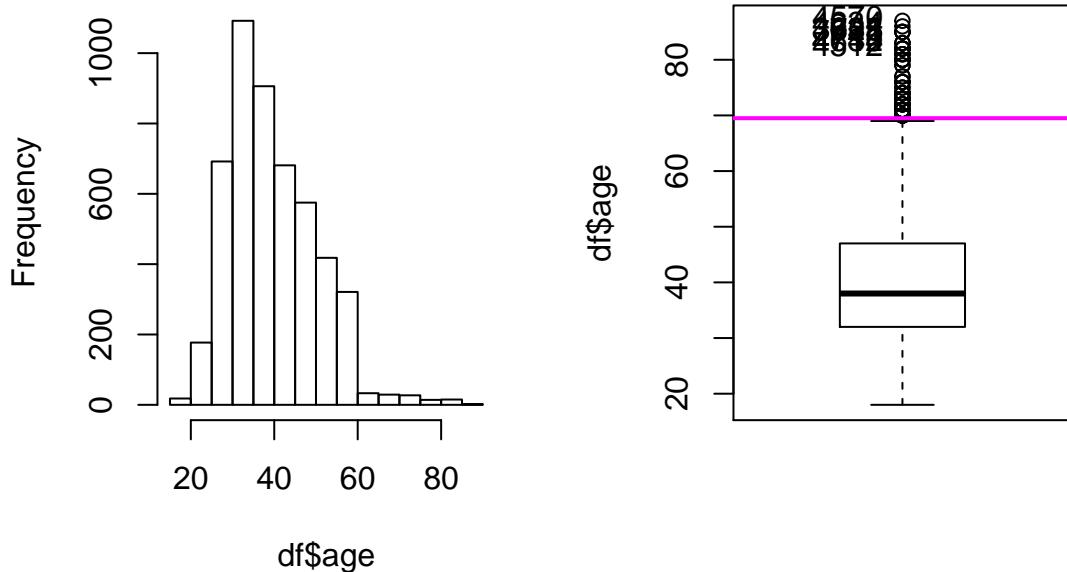
## [1] 0
if(length(err)>0) df[err]<-NA

# 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. Corresponden 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
```

```

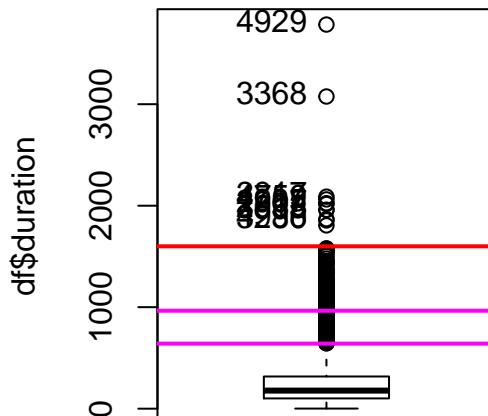
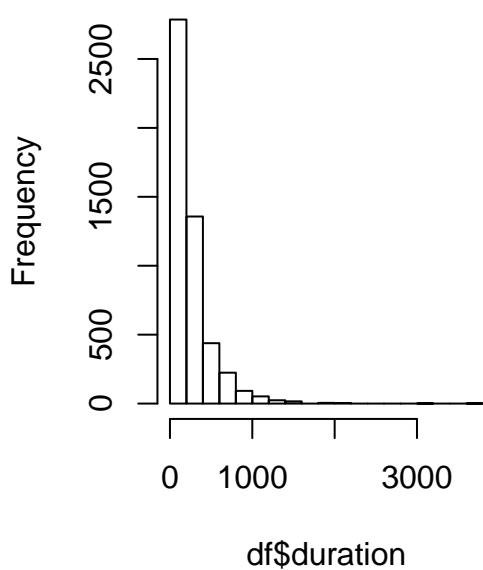
# Outliers:
out.var <- calcQ(df$duration)
abline(h=out.var[["mouts"]], col="magenta", lwd=2); out.var[["mouts"]]

## 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)

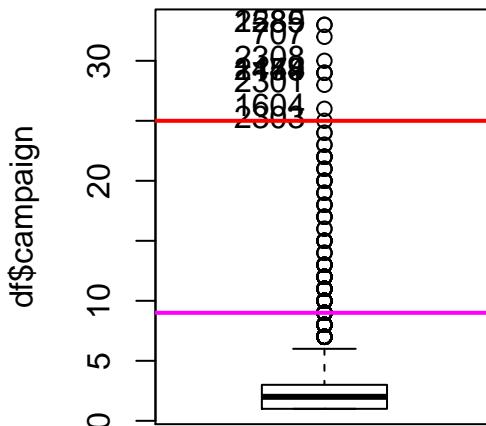
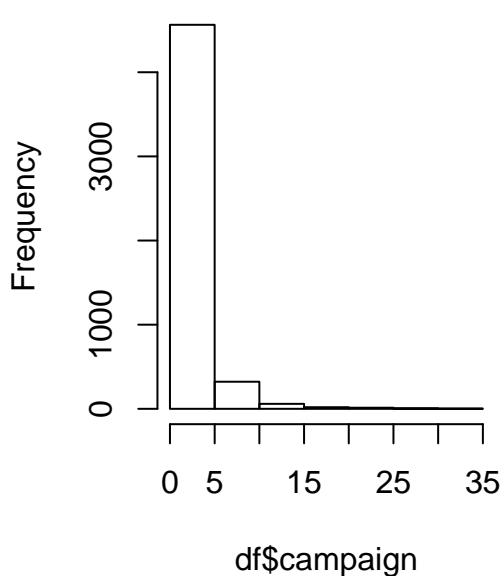
## [1] 0
df[miss, "num_missings"]<- df[miss, "num_missings"]+1

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

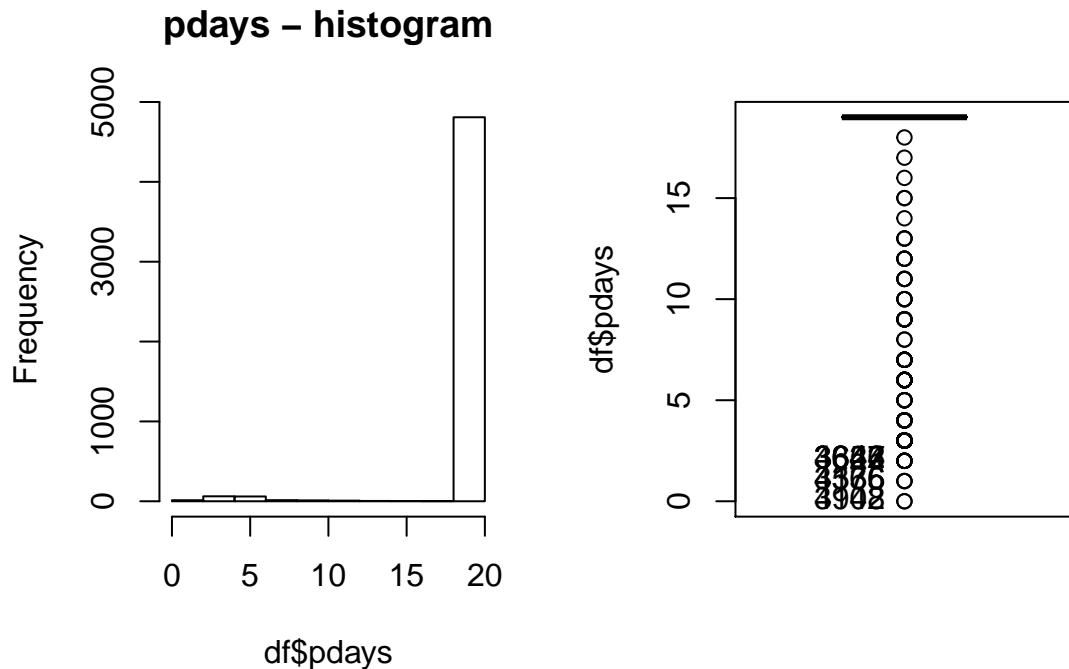
## [1] 96.18
# No outliers!

# 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 temim 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))

```

```

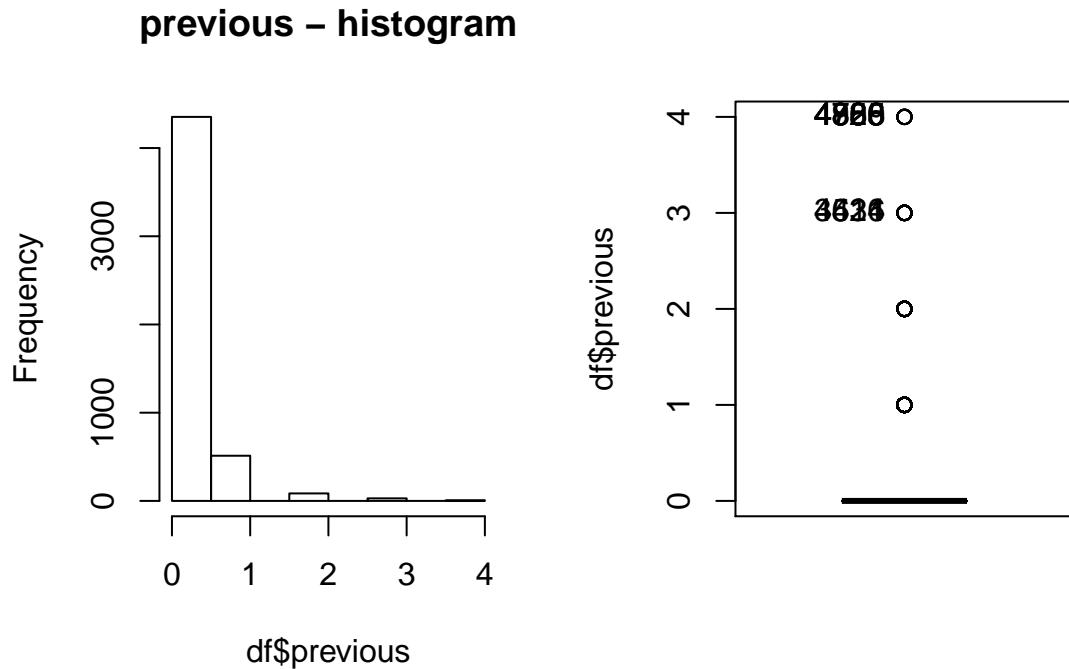
hist(df$previous, main="previous - histogram")

# 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))

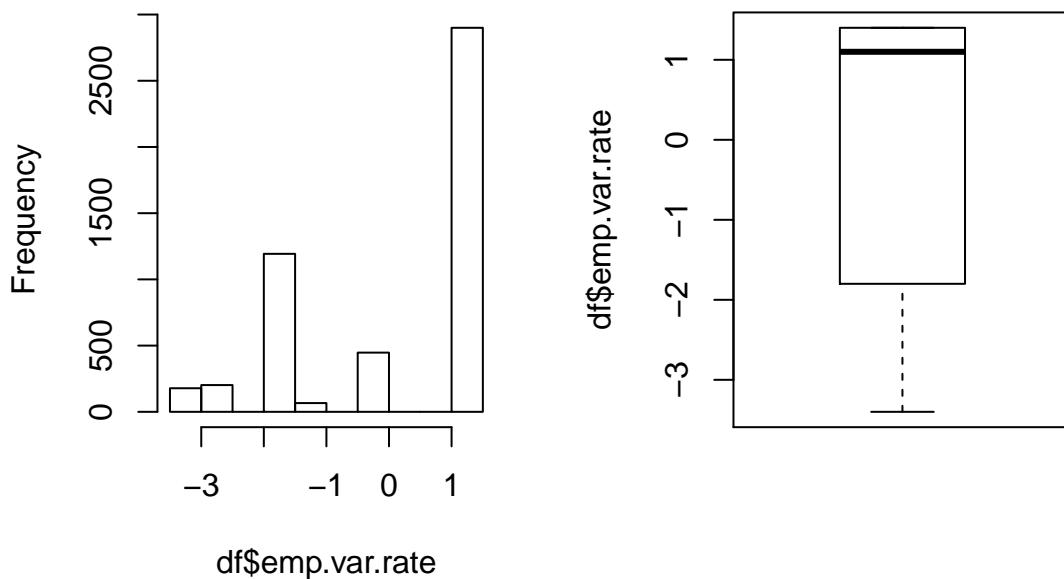
hist(df$emp.var.rate, main="emp.var.rate - histogram")
summary(df$emp.var.rate)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## -3.4000 -1.8000  1.1000  0.06446  1.40000  1.40000

Boxplot(df$emp.var.rate)

```

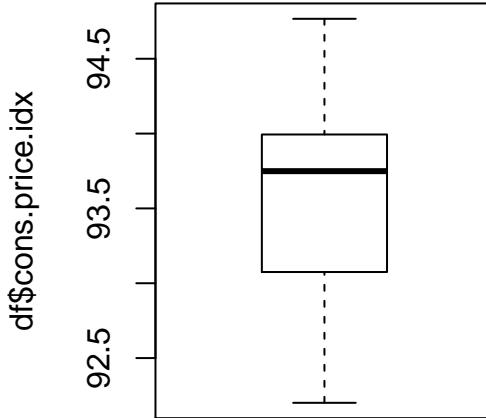
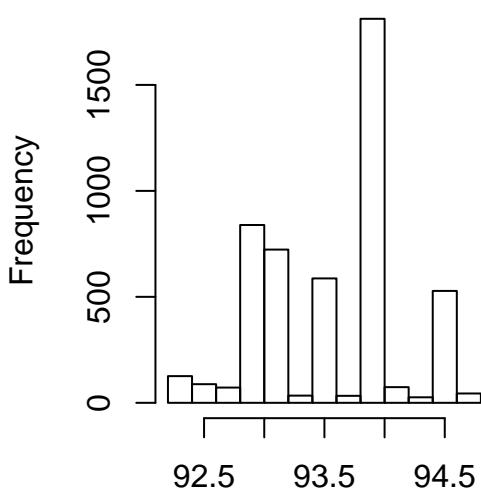
### emp.var.rate – histogram



### 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

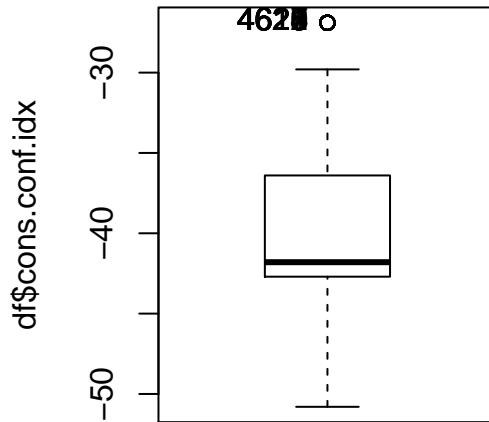
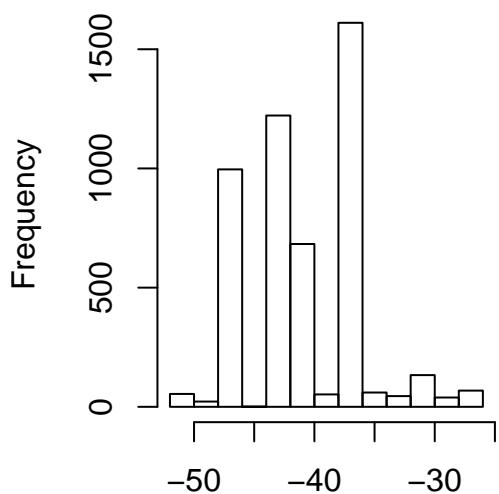


`df$cons.price.idx`

## 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

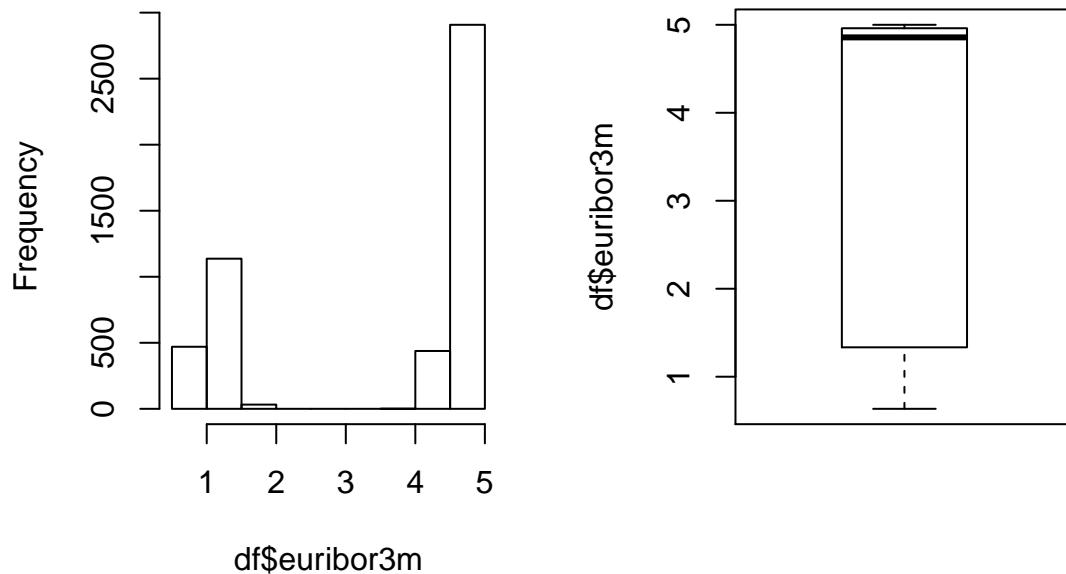


```
## [1] 4617 4618 4619 4620 4621 4622 4623 4624 4625 4626
```

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)
```

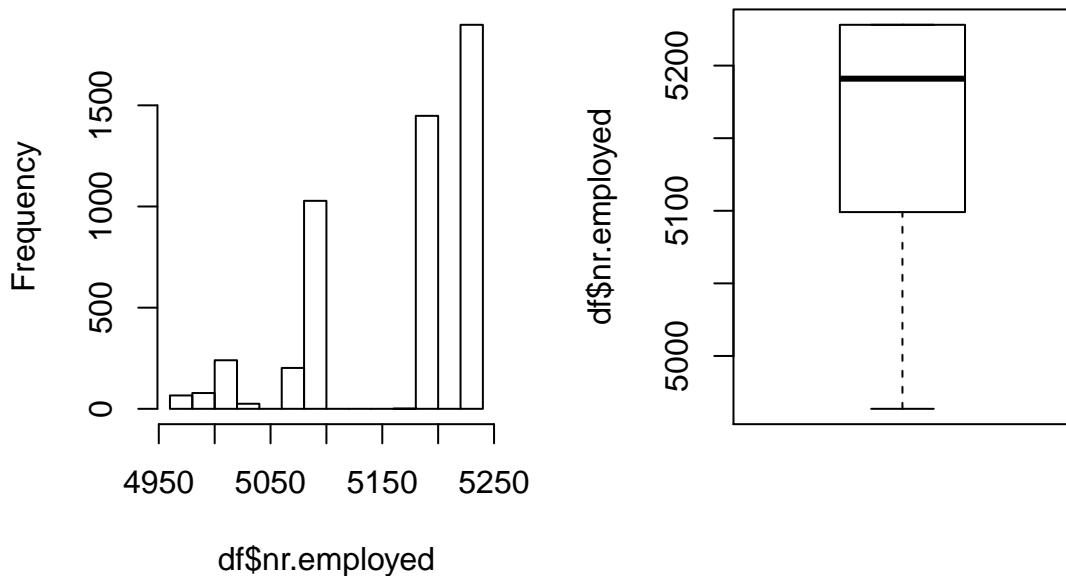
## euribor3m – histogram



## 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 té certs individus que han estat considerats outliers previament. Aquí els imputem mitjançant la imputació automàtica `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 imputació
df$campaign<-res.imp$completeObs[, "campaign"]
#summary(df[,vars_con])
```

Les variables numèriques originals que corresponen a conceptes quantitatius reals es mantenen com a numèriques, però també s'han de crear factors addicionals com a discretització de cada variable numèrica. Les etiquetes adicionals als factors s'afegeixen posterior als grafics per una qüestió estètica, es redueix la mida de les etiquetes i es poden veure amb més claredat cada una de les variables.

```

par(mfrow=c(1,1))

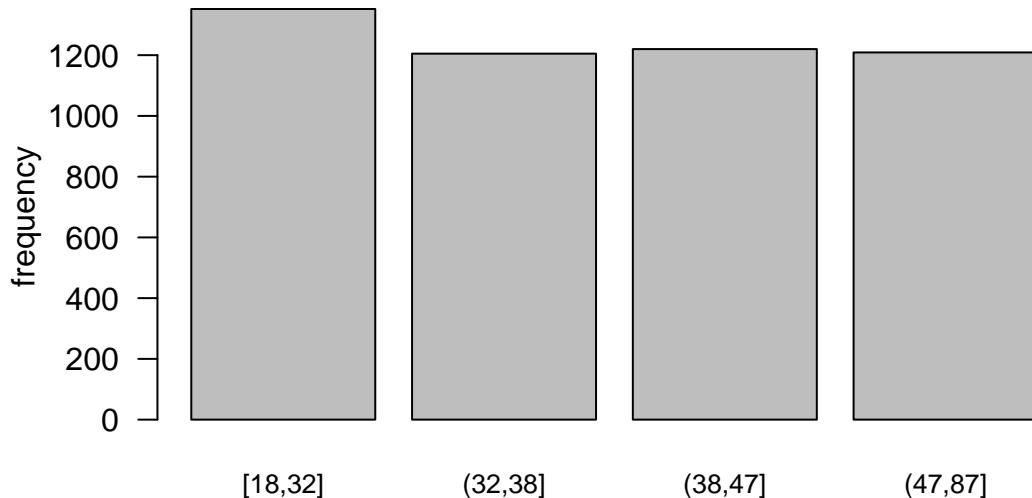
# AGE
qulist<-quantile(df$age, seq(0,1,0.25), na.rm=TRUE)

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)

```

## f.age – additional factors



```

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) )

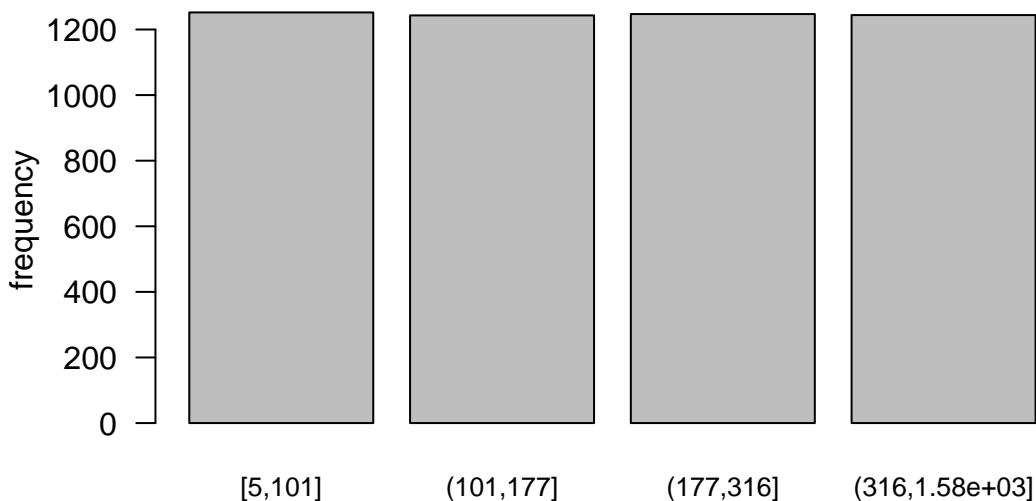
# DURATION
qulist<-quantile(df$duration, seq(0,1,0.25), na.rm=TRUE)

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)

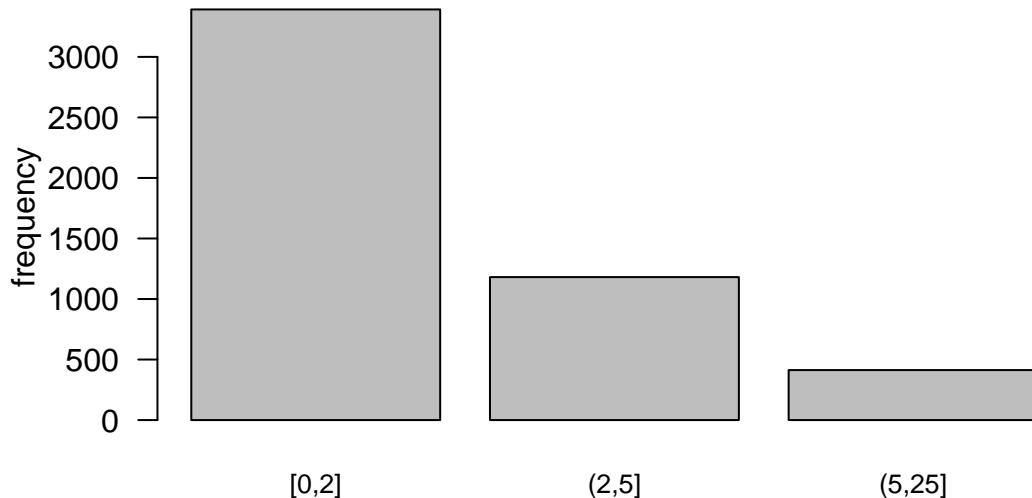
##          f.duration-[5,101]      f.duration-(101,177]
##                  1252                  1243
##          f.duration-(177,316] f.duration-(316,1.58e+03]
##                  1247                  1244

# 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)
```

## f.campaign – additional factors



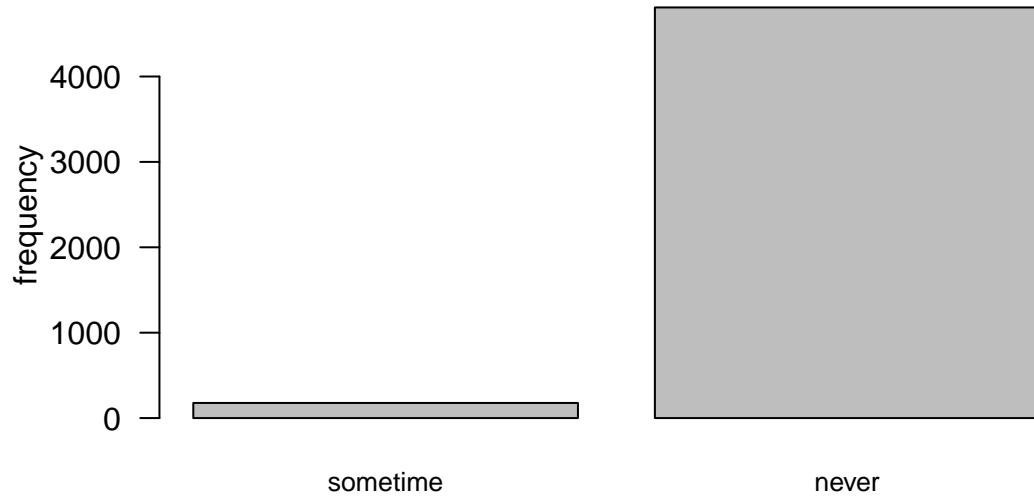
```
levels(df$f.campaign)<-paste0("f.campaign-", levels(df$f.campaign) )
summary(df$f.campaign)

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

# PDAYS
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 companya:
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



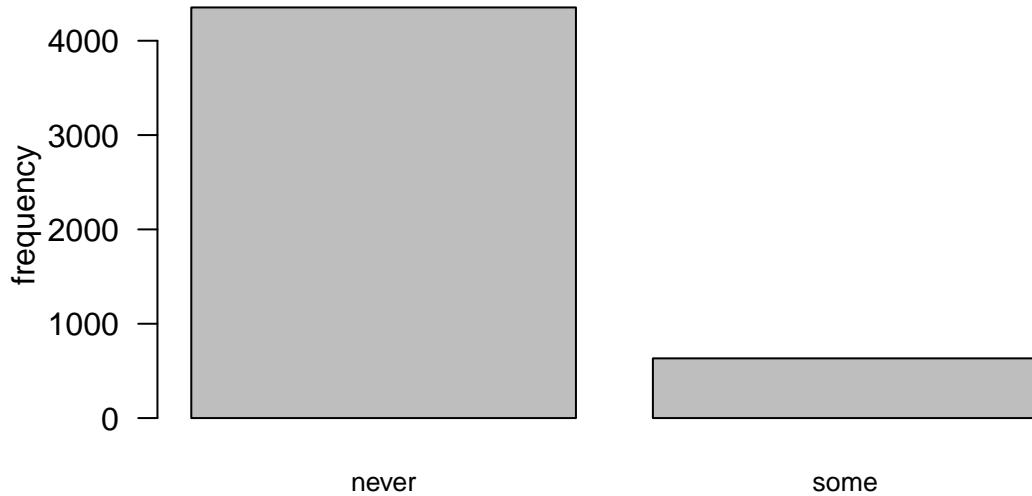
```
levels(df$f.pdays)<-paste0("f.pdays-", levels(df$f.pdays) )
summary(df$f.pdays)

## 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)
```

## f.previous – additional factors

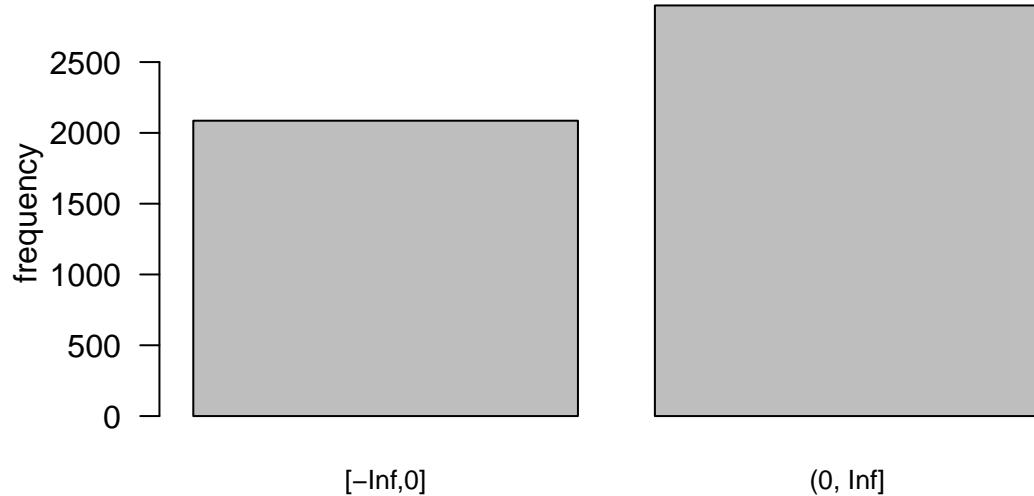


```
levels(df$f.previous)<-paste0("f.previous-", levels(df$f.previous) )
summary(df$f.previous)

## 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, cex.lab=0.8)
```

## f.emp.var.rate – additional factors

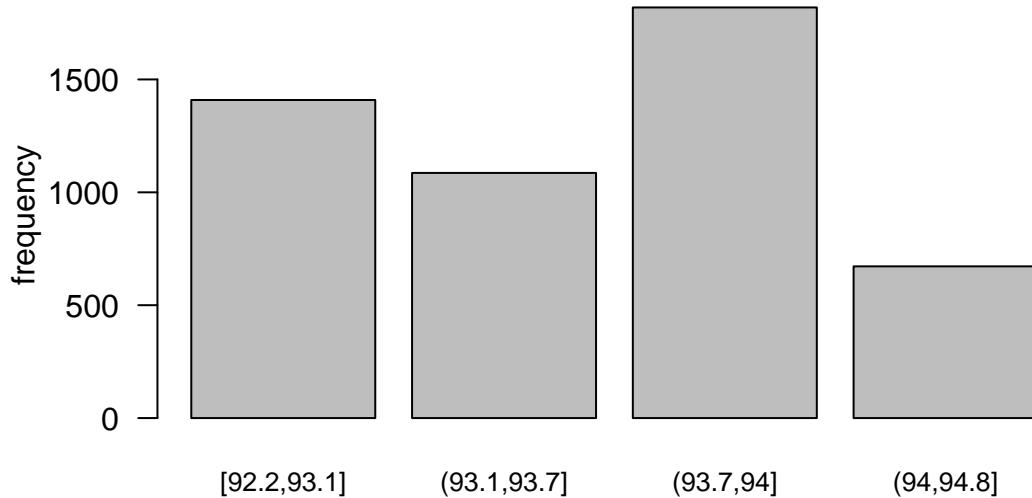


```
levels(df$f.emp.var.rate)<-paste0("f.emp.var.rate-", levels(df$f.emp.var.rate) )
summary(df$f.emp.var.rate)

## 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=1)
```

## 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)

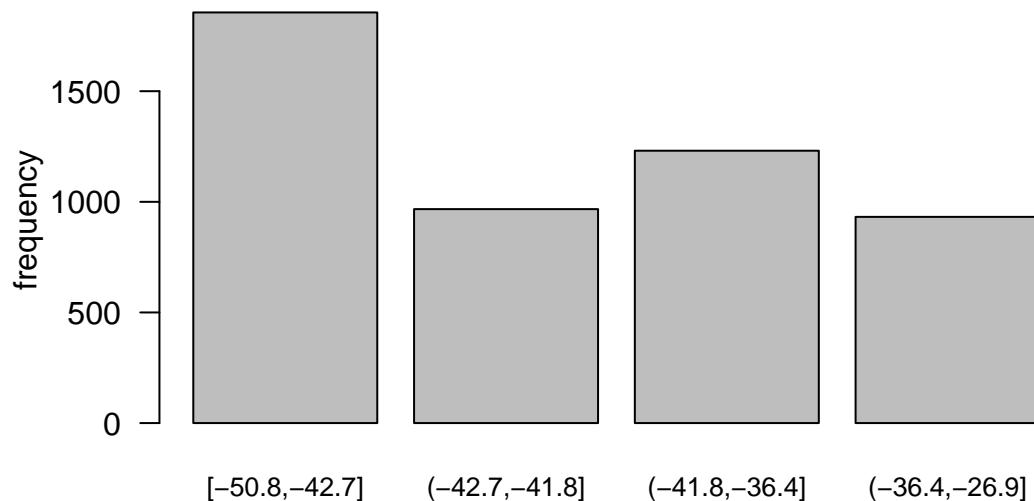
## f.cons.price.idx-[92.2,93.1] f.cons.price.idx-(93.1,93.7]
##                               1409                           1086
##   f.cons.price.idx-(93.7,94]   f.cons.price.idx-(94,94.8]
##                               1819                           672

# 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)

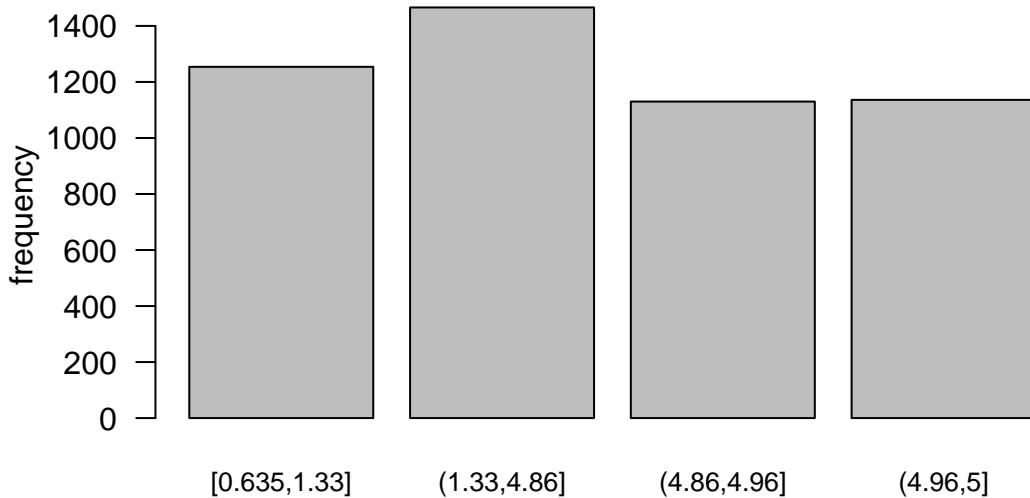
## f.cons.conf.idx-[-50.8,-42.7] f.cons.conf.idx-(-42.7,-41.8]
##                               1856                               967
## f.cons.conf.idx-(-41.8,-36.4] f.cons.conf.idx-(-36.4,-26.9]
##                               1231                               932

# 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.names=0.8)
```

## f.euribor3m – additional factors



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

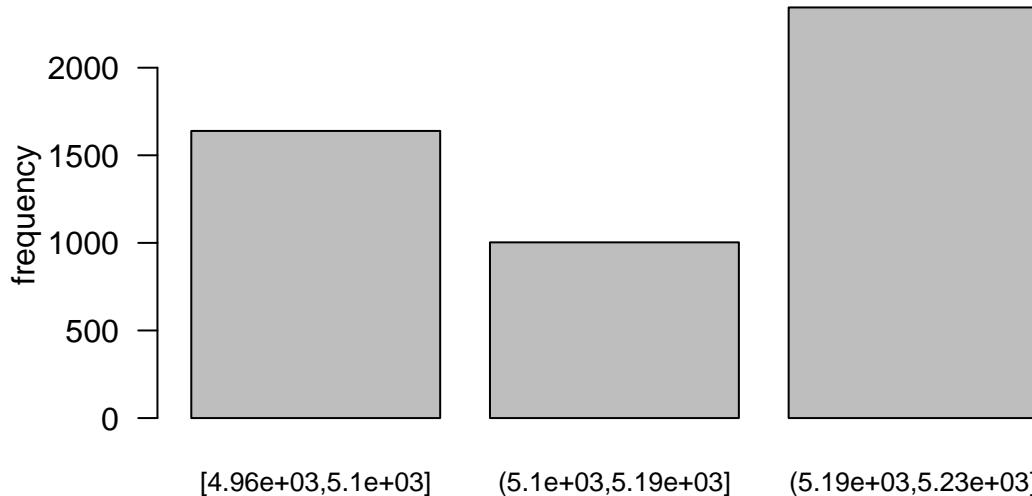
## f.euribor3m-[0.635,1.33]  f.euribor3m-(1.33,4.86]  f.euribor3m-(4.86,4.96]
##                      1254                  1466                  1130
##          f.euribor3m-(4.96,5]
##                      1136

# 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 discretes
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"

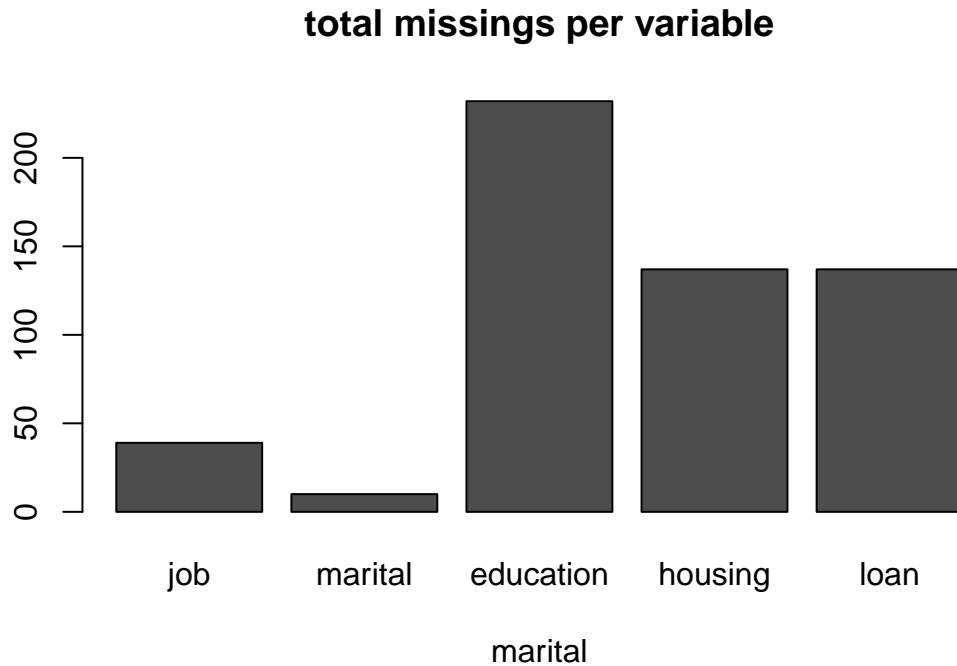
```

## 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.

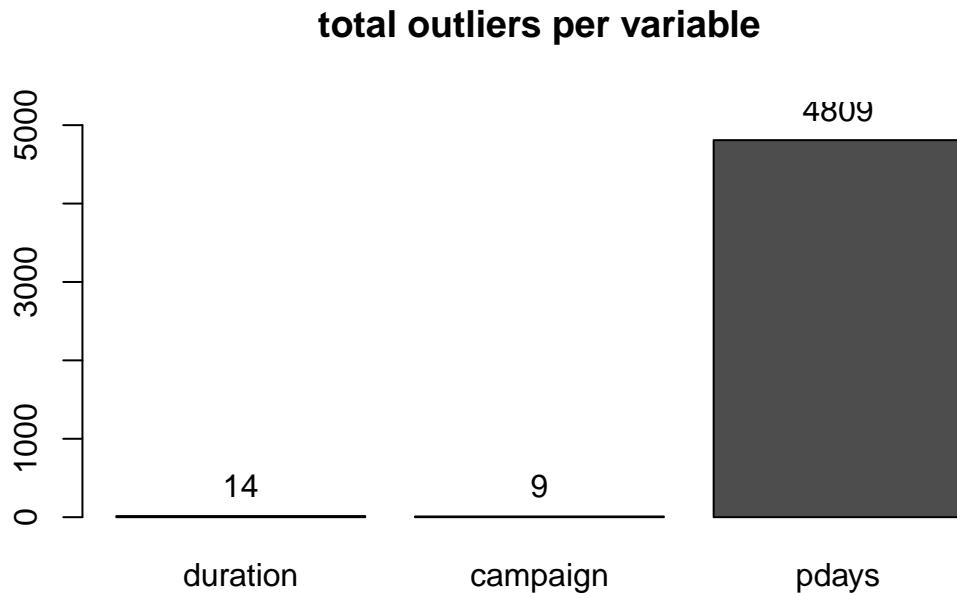
```
barplot( t(c(missings[, c(2,3,4,6,7)])), main="total missings per variable", xlab="marital")
```



```

## 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)]) )
ylabels <- barplot(data, ylim = ylim, main="total outliers per variable")
## Add text at top of bars
text(x = ylabels, 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]

```

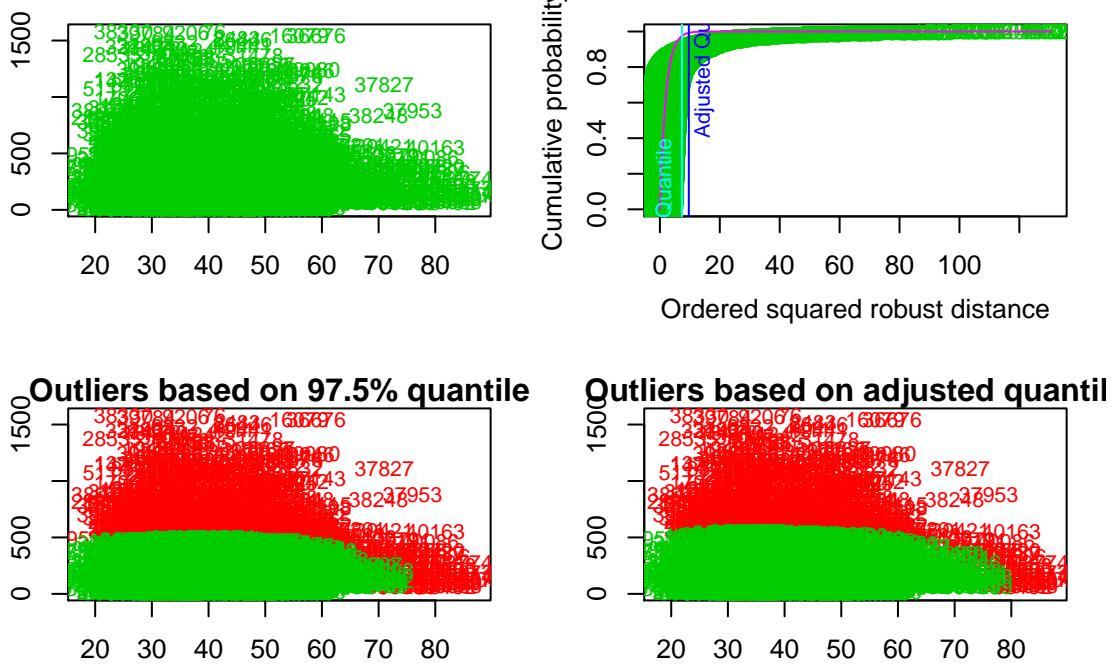
## Outliers Multivariants:

No hem aconseguit trobar una configuració del aq.plot que ens doni una bona grafica 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 facilment amb `imputeMCA()`, ja que els unknown (passats previament 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$complete0bs$marital)

## marital-divorced  marital-married  marital-single
##            554             3055            1377
summary(res.impf$complete0bs$loan)

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

summary(res.impf$complete0bs$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$complete0bs$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$complete0bs$housing)

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

# Acceptem la imputacio:
df$marital<-res.impf$complete0bs[, "marital"]
df$loan<-res.impf$complete0bs[, "loan"]
df$job<-res.impf$complete0bs[, "job"]
df$education<-res.impf$complete0bs[, "education"]

```

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

## PROFILING:

### CONTINOUS DESCRIPTION - Numeric Target (Duration):

La funcio d'R “condes” ens descriu la variable continua “duration” a partir d'altres variables quantitatives o de les variables categoriques. Això ho fa mitjançant els tres outputs diferenciatos mes avall; etiquetats com a “*quanti*”, “*quali*” i “*\$category*”.

El primer dels quals (*\$quanti*) ens mostra la correlació de la variable estudiada “duration” amb altres variables numeriques, mostrant nomes les correlacions que tenen un p-value per sota del llindar o nivell de significació del 5% (en aquest cas). Com mes petit es el p-valor, menys evidencia hi ha de que la hipotesi nul.la sigui certa i mes segurs estem del rebuig de la hipotesi nul.la. Aquesta hipotesi nul.la  $H_0$  afirma que la correlació o resultat obtingut es fruit d'una aleatorietat de les dades i no pot ser atribuible a una causa específica. Per tant, a partir d'ara, direm que quan el p-valor esta per sota del nivell de significació establert, els resultats son significatius.

Comentar que ens apareix el valor NA pero 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 entre la duració de la trucada en segons i en minuts, ja que es una correlació perfecta deguda a una conversió d'unitats. Dit això, observem lleugeres correlacions negatives significatives (ordenades de mes correlació positiva a no correlació i després a mes 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 es lògic perquè un client molt contactat està 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 socioeconomics que tenen una lleugera correlació negativa amb la duració de la trucada.

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

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 contractat (*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 variancias pels grups implicats que es comparen). En aquest cas ens permet concluir que la mitjana de la durada de la trucada en els casos que s'ha contractat el servei es 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 es molt significatiu (p-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       1.000000000 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$duration, df$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 funcio d'R “catdes” ens descriu la variable categorica “y” a partir d'altres variables categoriques o de les variables quantitatives. Aixo ho fa mitjançant outputs diferenciats mes avall. Notem que el nostre llindar de significacio en aquest cas es 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 categoriques que han caracteritzat al factor “y” ordenades de mes a menys caracteritzacio del factor (de menys a mes p-value). La columna “df” mostra els Degrees of Freedom, que corresponen amb el nombre de categories del factor menys 1. Les variables categoriques que han influenciat mes en la decisio final (y) son la f.duration (pero es una dada que s'obte a posteriori de la trucada, no ens serveix per a generar un perfil de client), f.pdays (nombre de dies des de l'ultim contacte), poutcome (si la ultima 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 socioeconomics, 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 descripcio de les variables categoriques per tal de poder estudiar-ne el seu enllac. La primera columna Cla/Mod esn 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 esser comparada amb la columna Global i d'aquesta manera trobar sobrerepresentacions 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 sobrerepresentada (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 sobrerepresentats 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 sobrerepresentació dels individus que van ser contactats fa 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 sobrerepresentades i poden ser observades en la llista donada per R. També hi ha certs valors socioeconomics 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 mes a menys caracterització del factor (de menys a mes correlació). Les variables quantitatives que han influenciat més en la decisió final (y) son 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 socioeconomics, 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 poden 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 un conseqüència del desenvolupament de la trucada). Altres factors com l'euribor 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
## f.outcome      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
```

# Course Practical Assignment - Final Delivery (Part 2)

*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:

### Load data from Deliverable 1:

```
#dirwd<- "D:/Users/Usuari/Documents/ADEIpractica"  
#dirwd<- "//pax/perfiles/1173408.CR/Downloads"  
dirwd<- "D:/Documents/GitHub/ADEI"  
setwd(dirwd)  
  
load( paste0(dirwd, "/bank-additional/Bank5000_validated.RData") )  
summary(df)
```

```

##      age          job          marital
##  Min.   :18.00   job-admin.    :1246   marital-divorced: 554
##  1st Qu.:32.00   job-blue-collar:1171  marital-married  :3055
##  Median :38.00   job-technician : 796   marital-single   :1377
##  Mean   :40.07   job-services   : 498
##  3rd Qu.:47.00   job-management: 411
##  Max.   :87.00   job-retired   : 205
##                  (Other)       : 659
##
##      education          default
##  education-basic.4y      : 533   default-no     :3954
##  education-basic.6y      : 289   default-unknown:1032
##  education-basic.9y      : 767
##  education-high.school   :1218
##  education-professional.course: 615
##  education-university.degree  :1564
##
##      housing          loan          contact
##  housing-no :2261   loan-no :4217   contact-cellular :3122
##  housing-yes:2725   loan-yes: 769   contact-telephone:1864
##
##      month          day_of_week          duration
##  month-may:1741   day_of_week-1mon:1016   Min.   : 5.0
##  month-jul: 829   day_of_week-2tue:1043   1st Qu.:101.0
##  month-aug: 697   day_of_week-3wed: 971   Median :177.0
##  month-jun: 652   day_of_week-4thu:1034   Mean   :250.6
##  month-nov: 507   day_of_week-5fri: 922   3rd Qu.:316.0
##  month-apr: 310                           Max.   :1580.0
##  (Other)   : 250
##
##      campaign          pdays          previous
##  Min.   : 1.000   Min.   : 0.00   Min.   :0.0000
##  1st Qu.: 1.000   1st Qu.:19.00   1st Qu.:0.0000
##  Median : 2.000   Median :19.00   Median :0.0000
##  Mean   : 2.535   Mean   :18.53   Mean   :0.1598
##  3rd Qu.: 3.000   3rd Qu.:19.00   3rd Qu.:0.0000
##  Max.   :25.000   Max.   :19.00   Max.   :4.0000
##
##      poutcome          emp.var.rate          cons.price.idx
##  poutcome-failure      : 477   Min.   :-3.40000   Min.   :92.20
##  poutcome-nonexistent:4353  1st Qu.:-1.80000  1st Qu.:93.08
##  poutcome-success     : 156   Median : 1.10000  Median :93.75
##                                Mean   : 0.06446  Mean   :93.57
##                                3rd Qu.: 1.40000  3rd Qu.:93.99
##                                Max.   : 1.40000  Max.   :94.77
##
##      cons.conf.idx          euribor3m          nr.employed          y
##  Min.   :-50.80   Min.   :0.635   Min.   :4964   y-no :4429
##  1st Qu.:-42.70   1st Qu.:1.334   1st Qu.:5099   y-yes: 557
##  Median :-41.80   Median :4.857   Median :5191
##  Mean   :-40.43   Mean   :3.614   Mean   :5166
##  3rd Qu.:-36.40   3rd Qu.:4.961   3rd Qu.:5228

```

```

##  Max.    :-26.90   Max.    :5.000   Max.    :5228
##
##  num_missings      num_outliers      num_errors
##  Min.    :0.0000   Min.    :0.00000   Min.    :0
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0
##  Median  :0.0000   Median  :0.00000   Median  :0
##  Mean    :0.1111   Mean    :0.00361   Mean    :0
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0
##  Max.    :3.0000   Max.    :2.00000   Max.    :0
##
##          f.season      minutes           f.age
##  season-spring     :2117   Min.    : 0.08333   f.age-[18,32]:1352
##  season-summer     :2178   1st Qu.: 1.68333   f.age-(32,38]:1205
##  season-autumnwinter: 691   Median  : 2.95000   f.age-(38,47]:1220
##                                         Mean    : 4.17703   f.age-(47,87]:1209
##                                         3rd Qu.: 5.26667
##                                         Max.    :26.33333
##
##          f.duration           f.campaign
##  f.duration-[5,101]    :1252   f.campaign-[0,2]  :3392
##  f.duration-(101,177]  :1243   f.campaign-(2,5]  :1181
##  f.duration-(177,316]  :1247   f.campaign-(5,25] : 413
##  f.duration-(316,1.58e+03]:1244
##
##
##
##          f.pdays           f.previous
##  f.pdays-sometime: 177   f.previous-never:4353
##  f.pdays-never    :4809   f.previous-some  : 633
##
##
##
##          f.emp.var.rate           f.cons.price.idx
##  f.emp.var.rate-[-Inf,0]  :2086   f.cons.price.idx-[92.2,93.1]:1409
##  f.emp.var.rate-(0, Inf]  :2900   f.cons.price.idx-(93.1,93.7]:1086
##                                         f.cons.price.idx-(93.7,94]  :1819
##                                         f.cons.price.idx-(94,94.8]  : 672
##
##
##
##          f.cons.conf.idx           f.euribor3m
##  f.cons.conf.idx-[-50.8,-42.7] :1856   f.euribor3m-[0.635,1.33]:1254
##  f.cons.conf.idx-(-42.7,-41.8]: 967   f.euribor3m-(1.33,4.86]  :1466
##  f.cons.conf.idx-(-41.8,-36.4] :1231   f.euribor3m-(4.86,4.96]  :1130
##  f.cons.conf.idx-(-36.4,-26.9]: 932   f.euribor3m-(4.96,5]    :1136
##
##
##
##          f.nr.employed
##  f.nr.employed-[4.96e+03,5.1e+03]  :1639
##  f.nr.employed-(5.1e+03,5.19e+03] :1003
##  f.nr.employed-(5.19e+03,5.23e+03]:2344

```

```
##  
##  
##  
##
```

## CORRESPONDENCE ANALYSIS (CA)

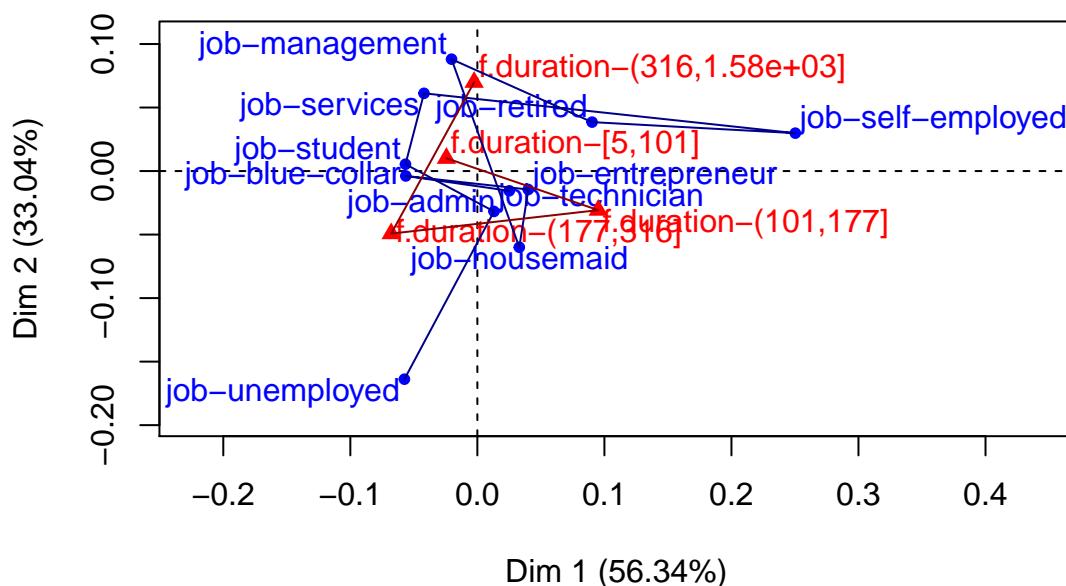
Realitzarem un analisi amb taules de correspondencia i mapes de factors entre la variable numerica target duracio discretitzada en 4 nivells (corregit de l'entrega 1) i diversos factors que hem trobat que tenen una correspondencia significativa.

Primer veiem com la duracio de la trucada no te cap relacio amb el job de l'individu, ja que no podem rebutjar la hipotesi  $H_0$ : *f.duration no te cap relacio amb la variable job* amb el valor p obtingut en el Chi Square test. Com mes apropi surten al grafic les categories d'ambdues variables analitzades, mes relacionades estan. En aquesta comparacio, com s'acaba de comentar, no es pot extreure res significatiu, mes enlla que potser les categories job-unemployed i job-self-employed van mes per lliure.

```
# H0: f.duration no te cap relacio amb variable job  
chisq.test( table( df$job, df$f.duration ) )
```

```
##  
## Pearson's Chi-squared test  
##  
## data: table(df$job, df$f.duration)  
## X-squared = 31.496, df = 30, p-value = 0.3913  
  
# CA - f.duration vs variable job  
res.ca<-CA( table( df$job, df$f.duration ) )  
lines(res.ca$row$coord[,1], res.ca$row$coord[,2], col="darkblue")  
lines(res.ca$col$coord[,1], res.ca$col$coord[,2], col="darkred")
```

CA factor map



Ara testearem la mateixa hipotesi i mostrarem el mateix mapa de factors pero amb altres categories factor que si que obtindrem que tenen una relacio significativa. El primer cas es l'epoca de l'any **f.season** en la qual es realitza la trucada (p valor = 2.506e-07). Comparant el profile de la taula de contingencia de proporcions per fila amb el profile marginal de la duracio veiem com hi ha un 28,5% de trucades amb duracions molt curtes a l'estiu respecte un 25% de trucades amb duracions curtes en tot l'any. D'altra banda, hi ha per sobre d'un 27% de trucades amb duracions llargues a la primavera, respecte un 25% de trucades en la mateixa duracio en tot l'any. Si comparem el profile de la taula de contingencia de proporcions per columna amb el profile marginal de la f.season veiem com el 42.5% de trucades es realitzen a la primavera, i en canvi mes d'un 46% de trucades corresponen a la primavera i a duracions llargues. A mes, el 43.7% de trucades es realitzen a l'estiu, i nomes prop d'un 40% de trucades corresponen a l'estiu i a duracions llargues. Aquesta mateixa informacio es pot veure representada en un mapa de factors de dues dimensions. Agafant nomes la primera dimensio ja seria suficient per a representar un 98.9% de la variancia del conjunt de les dades (Kaiser: take as many dimensions as eigenvalue > mean of eigenvalues).

```
chisq.test( table( df$f.season, df$f.duration) )

##
## Pearson's Chi-squared test
##
## data: table(df$f.season, df$f.duration)
## X-squared = 41.318, df = 6, p-value = 2.506e-07

#Row/Column profile
prop.table( table(df$f.season, df$f.duration), 1 ) #1->per files

##
##          f.duration-[5,101] f.duration-(101,177]
## season-spring           0.2106755           0.2442135
## season-summer            0.2855831           0.2534435
## season-autumnwinter     0.2662808           0.2518090
##
##          f.duration-(177,316] f.duration-(316,1.58e+03]
## season-spring            0.2735002           0.2716108
## season-summer             0.2277319           0.2332415
## season-autumnwinter      0.2489146           0.2329957

#Marginal Row/Column profile
prop.table( table(df$f.duration)) #1->per files

##
##          f.duration-[5,101]      f.duration-(101,177]
##                      0.2511031           0.2492980
##          f.duration-(177,316] f.duration-(316,1.58e+03]
##                      0.2501003           0.2494986

prop.table( table(df$f.season, df$f.duration), 2 ) #2->per columns

##
##          f.duration-[5,101] f.duration-(101,177]
## season-spring           0.3562300           0.4159292
## season-summer            0.4968051           0.4440869
## season-autumnwinter     0.1469649           0.1399839
##
##          f.duration-(177,316] f.duration-(316,1.58e+03]
## season-spring            0.4643144           0.4622186
## season-summer             0.3977546           0.4083601
## season-autumnwinter      0.1379310           0.1294212
```

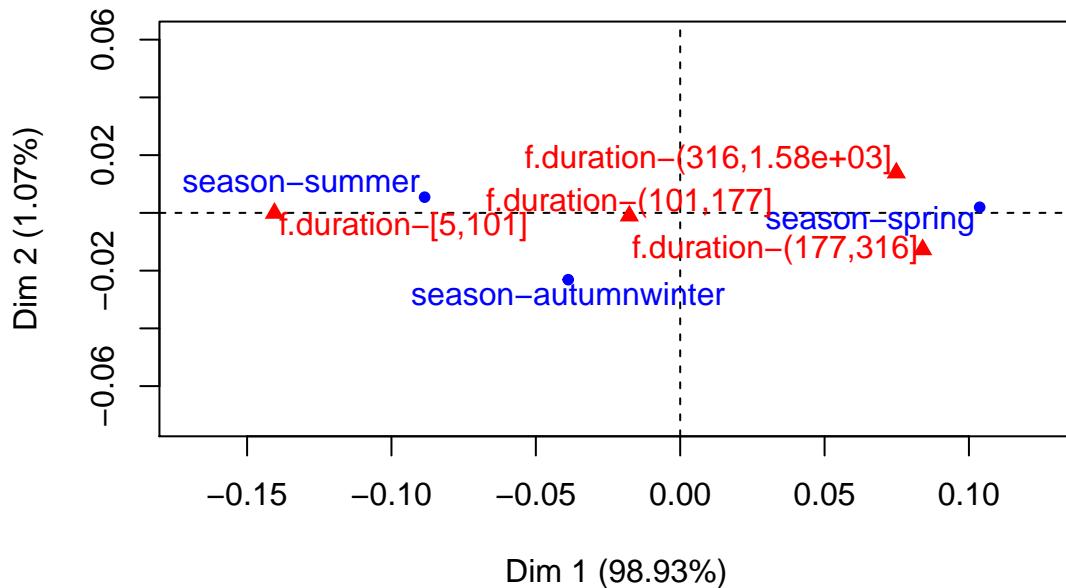
```

prop.table( table(df$f.season) ) #2->per columnes

##
##      season-spring      season-summer season-autumnwinter
##      0.4245888          0.4368231          0.1385880
res.ca<-CA( table( df$f.season, df$f.duration ) )

```

## CA factor map



```

attributes(res.ca); res.ca$eig #valors eig no normalitzats!

## $names
## [1] "eig"   "call"  "row"   "col"   "svd"
##
## $class
## [1] "CA"    "list"
##
##       eigenvalue percentage of variance
## dim 1 8.198140e-03      98.929073
## dim 2 8.874655e-05      1.070927
##
##       cumulative percentage of variance
## dim 1           98.92907
## dim 2          100.00000
mean(res.ca$eig[,1]) #Kaiser: take as many dimensions as eigenvalue > mean of eigenvalues

## [1] 0.004143443
#En una taula de correspondencies simples podem tenir maxim tantes
#dimensions com categories d'una variable menys 1!
#f.season te 3 categories -> -1 -> 2 dimensions!!

#La inercia total ens indica com de relacionades estan les dues variables,

```

```
#com mes proxim el valor a 0, menys relacionades estan!
sum(res.ca$eig[,1])
```

```
## [1] 0.008286887
#Coordenades:
#res.ca$row #files son la f.season!
#res.ca$col #columnes son la duration!
```

El segon cas en el qual obtenim una relació significativa (p valor = 1.203e-07) es el valor de l'euribor **f.euribor3m**, el qual es un indicador trimestral. Comparant el profile de la taula de contingència de proporcions per fila amb el profile marginal de la duració veiem com hi ha un 30.8% de trucades amb duracions molt curtes quan l'euribor té un valor alt, respecte un 25% de trucades amb duracions sense tenir en compte la fluctuació de l'indicador. De la mateixa manera, quan el valor de l'euribor es baix, hi ha major % de trucades que acaben amb duracions relativament altes. Si comparem el profile de la taula de contingència de proporcions per columna amb el profile marginal de f.euribor3m veiem com el 22.8% de trucades es realitzen amb un euribor molt alt, i en canvi un 28.0% de trucades corresponen a un euribor alt i a duracions molt curtes. A més, mirant el mapa de factors podem veure aquesta mateixa informació de manera gràfica. Si ens centrem en la 1a dimensió del grafic (que representa un 89% de la variancia del conjunt de les dades), podem observar com hi ha una tendència similar en ambdues variables (tot i que les duracions extremadament llargues trenquen una molt bona correlació del 1r eix). També veiem com les categories f.euribor3m-[4.96,5] i f.duration-[5,101] estan molt relacionades ja que es troben molt proximes en el mapa de factors; així com també les categories f.euribor3m-[0.635,1.33] i f.duration-(177,316]. (Kaiser: take as many dimensions as eigenvalue > mean of eigenvalues, el que equival a agafar 1 sola dimensió).

```
chisq.test( table( df$f.euribor3m, df$f.duration) )
```

```
##
## Pearson's Chi-squared test
##
## data: table(df$f.euribor3m, df$f.duration)
## X-squared = 49.745, df = 9, p-value = 1.203e-07
#Row/Column profile
prop.table( table(df$f.euribor3m, df$f.duration), 1 ) #1->per files
```

```
##
##          f.duration-[5,101] f.duration-(101,177]
##  f.euribor3m-[0.635,1.33]      0.2129187      0.2432217
##  f.euribor3m-(1.33,4.86]      0.2312415      0.2530696
##  f.euribor3m-(4.86,4.96]      0.2619469      0.2345133
##  f.euribor3m-(4.96,5]        0.3080986      0.2658451
##
##          f.duration-(177,316] f.duration-(316,1.58e+03]
##  f.euribor3m-[0.635,1.33]      0.2775120      0.2663477
##  f.euribor3m-(1.33,4.86]      0.2694407      0.2462483
##  f.euribor3m-(4.86,4.96]      0.2353982      0.2681416
##  f.euribor3m-(4.96,5]        0.2095070      0.2165493
```

```
#Marginal Row/Column profile
prop.table( table(df$f.duration)) #1->per files
```

```
##
##          f.duration-[5,101] f.duration-(101,177]
##          0.2511031      0.2492980
##          f.duration-(177,316] f.duration-(316,1.58e+03]
##          0.2501003      0.2494986
```

```

prop.table( table(df$f.euribor3m, df$f.duration), 2 ) #2->per columnes

##
##          f.duration-[5,101] f.duration-(101,177]
##  f.euribor3m-[0.635,1.33]      0.2132588      0.2453741
##  f.euribor3m-(1.33,4.86]      0.2707668      0.2984714
##  f.euribor3m-(4.86,4.96]      0.2364217      0.2131939
##  f.euribor3m-(4.96,5]        0.2795527      0.2429606
##
##          f.duration-(177,316] f.duration-(316,1.58e+03]
##  f.euribor3m-[0.635,1.33]      0.2790698      0.2684887
##  f.euribor3m-(1.33,4.86]      0.3167602      0.2901929
##  f.euribor3m-(4.86,4.96]      0.2133119      0.2435691
##  f.euribor3m-(4.96,5]        0.1908581      0.1977492

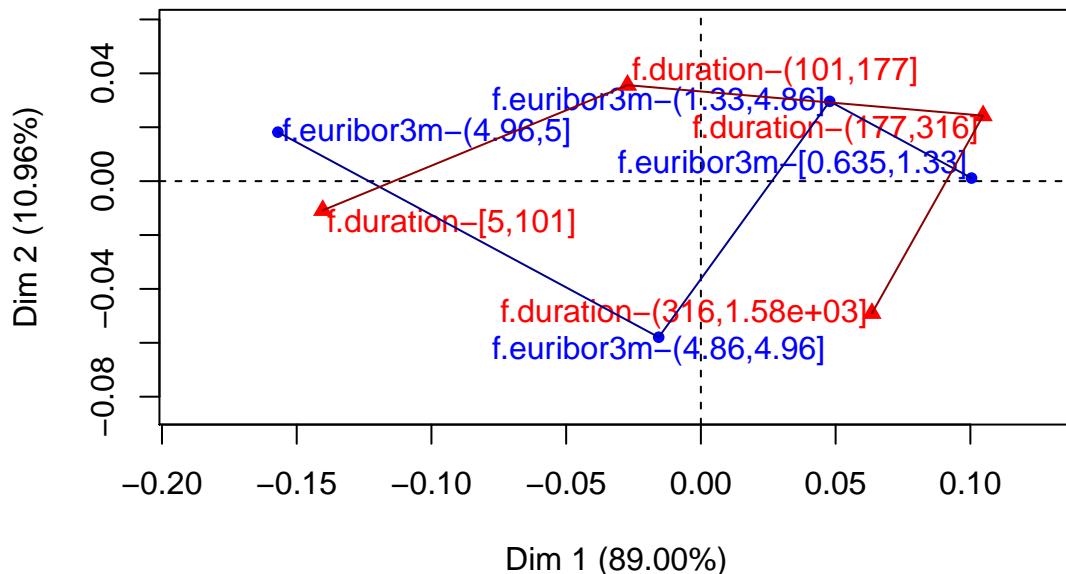
prop.table( table(df$f.euribor3m) ) #2->per columnes

##
##  f.euribor3m-[0.635,1.33]  f.euribor3m-(1.33,4.86]  f.euribor3m-(4.86,4.96]
##          0.2515042            0.2940233            0.2266346
##  f.euribor3m-(4.96,5]
##          0.2278379

res.ca<-CA( table( df$f.euribor3m, df$f.duration) )
lines(res.ca$row$coord[,1], res.ca$row$coord[,2], col="darkblue")
lines(res.ca$col$coord[,1], res.ca$col$coord[,2], col="darkred")

```

## CA factor map



```
attributes(res.ca); res.ca$eig #valors eig no normalitzats!
```

```
## $names
## [1] "eig"  "call" "row"  "col"  "svd"
```

```

## 
## $class
## [1] "CA"    "list"

##      eigenvalue percentage of variance
## dim 1 8.879478e-03      88.99940237
## dim 2 1.093876e-03      10.96396899
## dim 3 3.654443e-06      0.03662864
##      cumulative percentage of variance
## dim 1                  88.99940
## dim 2                  99.96337
## dim 3                 100.00000

mean(res.ca$eig[,1]) #Kaiser: take as many dimensions as eigenvalue > mean of eigenvalues

## [1] 0.00332567

#En una taula de correspondencies simples podem tenir maxim tantes
#dimensions com categories d'una variable menys 1!
#f.euribor3m te 4 categories -> -1 -> 3 dimensions!!

#La inercia total ens indica com de relacionades estan les dues variables,
#com mes proxim el valor a 0, menys relacionades estan!
sum(res.ca$eig[,1])

## [1] 0.009977009

#A vegades va be eliminar algunes categories amb pocs individus d'una variable
#per a poder veure millor les possibles relacions!

```

## PRINCIPAL COMPONENT ANALYSIS (PCA)

Primerament realitzem un PCA sobre les variables continues de la nostra mostra de dades, on la variables “duration” ha de ser suplementaria, ja que es tracta de la variable target!

### Eigenvalues and dominant axes

El summary ens permet veure els 9 diferents eigenvalues obtinguts amb aquest PCA, amb les seves dades corresponents de percentatges de variancia del conjunt de les dades que representen. Segons el criteri de Kaiser, que diu que s’han de descartar les dimensions amb valors eig normalitzats per sota d’1, hauríem d’agafar les 3 primeres dimensions per a una bona representació del conjunt de dades. Essent flexibles amb el criteri de Kaiser, podríem agafar també la quarta dimensió, la qual té una variancia del 0.9656, amb un valor molt proxim a 1. La incorporació d’aquesta nova dimensió ens donaria una variancia acumulada del 81%, obtenint d’aquesta manera una variancia acumulada per sobre el 80%. Per últim, si ens basem en la regla del colze (llegurament subjectiva), i l’apliquem sobre el grafic dels eigenvalues i %variancias obtingut amb les llibreries *ggplot*, hauríem d’agafar les 3 primeres dimensions.

```

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

## [1] "age"           "duration"       "campaign"       "pdays"
## [5] "previous"      "emp.var.rate"   "cons.price.idx" "cons.conf.idx"
## [9] "euribor3m"     "nr.employed"

vars_dis<-names(df)[c(2:10, 15, 21, 25, 27:36)] #variables discretas

# PCA:
res.pca<-PCA( df[, vars_con], quanti.sup=2, graph=FALSE) #"duration" com a suplementaria

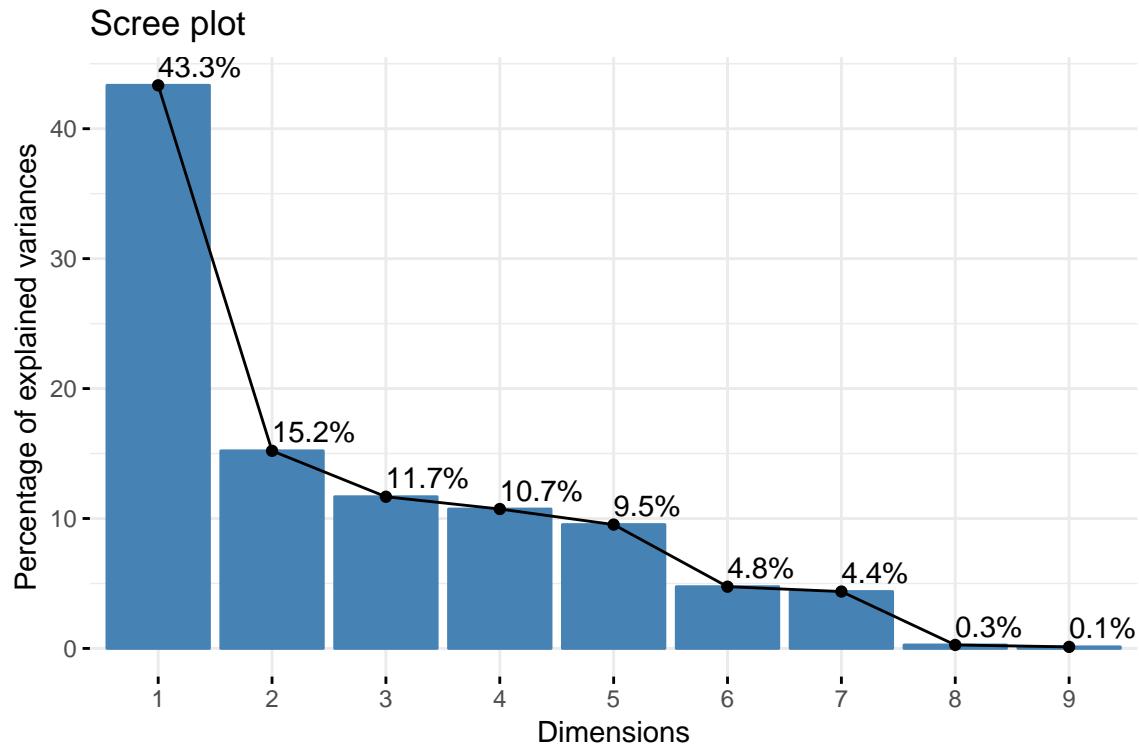
```

```

#nb.dec: number of decimal printed
#ncp: number of dimensions printed
summary(res.pca, nb.dec=2, ncp=5, nbind=0)

##
## Call:
## PCA(X = df[, vars_con], quanti.sup = 2, graph = FALSE)
##
##
## Eigenvalues
##          Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance     3.90    1.37    1.05    0.97    0.86    0.43    0.39
## % of var.  43.34   15.21   11.67   10.73   9.53    4.75    4.38
## Cumulative % of var. 43.34   58.55   70.22   80.95   90.48   95.24   99.61
##          Dim.8   Dim.9
## Variance     0.02    0.01
## % of var.   0.27    0.12
## Cumulative % of var. 99.88 100.00
##
## Variables
##          Dim.1   ctr   cos2   Dim.2   ctr   cos2   Dim.3   ctr   cos2
## age        -0.02  0.01  0.00 |  0.35  8.73  0.12 |  0.67 42.16  0.44
## campaign   0.20  1.03  0.04 |  0.01  0.01  0.00 | -0.35 11.85  0.12
## pdays      0.43  4.84  0.19 | -0.71 37.11  0.51 |  0.32  9.77  0.10
## previous   -0.59  9.05  0.35 |  0.55 21.75  0.30 | -0.32  9.71  0.10
## emp.var.rate 0.97 23.96  0.93 |  0.17  2.19  0.03 | -0.09  0.75  0.01
## cons.price.idx 0.75 14.49  0.57 |  0.25  4.49  0.06 | -0.25  5.88  0.06
## cons.conf.idx 0.16  0.68  0.03 |  0.56 23.16  0.32 |  0.46 19.73  0.21
## euribor3m   0.97 23.91  0.93 |  0.19  2.54  0.03 | -0.01  0.01  0.00
## nr.employed 0.93 22.02  0.86 |  0.01  0.02  0.00 | -0.04  0.14  0.00
##          Dim.4   ctr   cos2   Dim.5   ctr   cos2
## age        0.39 15.54  0.15 |  0.53 33.33  0.29 |
## campaign   0.89 81.66  0.79 | -0.22  5.44  0.05 |
## pdays      0.08  0.68  0.01 | -0.05  0.31  0.00 |
## previous   -0.05  0.29  0.00 |  0.14  2.42  0.02 |
## emp.var.rate -0.06  0.43  0.00 |  0.07  0.54  0.00 |
## cons.price.idx -0.07  0.52  0.00 |  0.26  7.62  0.07 |
## cons.conf.idx -0.03  0.09  0.00 | -0.66 50.14  0.43 |
## euribor3m   -0.07  0.54  0.01 | -0.02  0.05  0.00 |
## nr.employed -0.05  0.25  0.00 |  0.04  0.15  0.00 |
##
## Supplementary continuous variable
##          Dim.1   cos2   Dim.2   cos2   Dim.3   cos2   Dim.4   cos2
## duration   -0.03  0.00 |  0.00  0.00 |  0.00  0.00 | -0.07  0.00 |
##          Dim.5   cos2
## duration     0.04  0.00 |
##GGPLOT: Use modern ggplot facilities per la regla de l'ultim colze:
##at some point the marginal gain will drop, giving an angle in the graph
fviz_eig(res.pca, addlabels=TRUE)

```



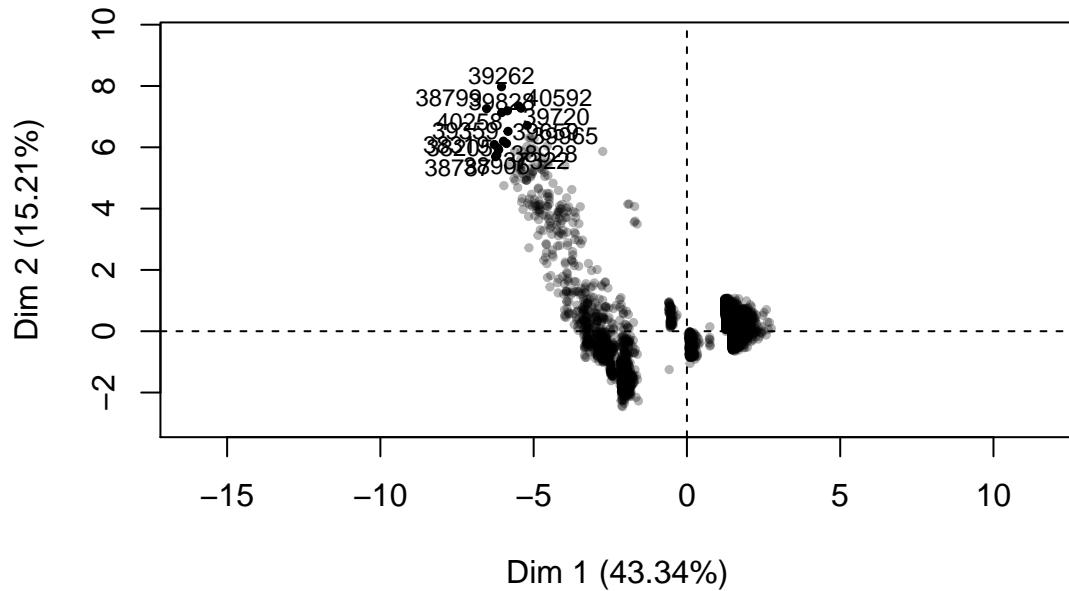
### Individuals point of view

Pintem primer el mapa de factors dels individus de les dues primeres dimensions (70% de la variancia del conjunt de dades) i etiquetem amb el numero d'individu els 15 mes contributius. A continuacio mostrem per a cadascun dels dos primers eixos, les coordenades i el registre complet d'aquests 3 individus mes contributius. Es fa exactament el mateix amb els individus mes ben representats ( $\cos^2$ ) en les dues primeres dimensions.

#nomes pinta les etiquetes dels 15 individus mes contributius!

```
plot(res.pca, choix="ind", cex=0.75, col.ind="black", select="contrib 15", title="Factor map - 15 individus")
```

## Factor map – 15 individus mes contributius



```
#2 individus mes contributius al 1r eix:
contrib<-sort(res.pca$ind$contrib[,1], decreasing=TRUE)[1:2]; contrib

##      38799      38319
## 0.2194923 0.2029850
df[c(names(contrib)), ]

##           age          job        marital             education
## 38799  62  job-housemaid marital-married education-university.degree
## 38319  37  job-blue-collar marital-married   education-basic.6y
##         default     housing     loan       contact      month
## 38799 default-no housing-yes loan-no contact-cellular month-nov
## 38319 default-no housing-no loan-no contact-cellular month-oct
##         day_of_week duration campaign pdays previous      poutcome
## 38799 day_of_week-4thu      237       1     3        3 poutcome-success
## 38319 day_of_week-4thu      128       2     6        3 poutcome-failure
##         emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed
## 38799        -3.4        92.649       -30.1      0.714      5017.5
## 38319        -3.4        92.431       -26.9      0.740      5017.5
##           y num_missings num_outliers num_errors      f.season
## 38799  y-no          0            0            0 season-autumnwinter
## 38319  y-yes          0            0            0 season-autumnwinter
##         minutes      f.age      f.duration      f.campaign
## 38799 3.950000 f.age-(47,87] f.duration-(177,316] f.campaign-[0,2]
## 38319 2.133333 f.age-(32,38] f.duration-(101,177] f.campaign-[0,2]
##         f.pdays      f.previous      f.emp.var.rate
## 38799 f.pdays-sometime f.previous-some f.emp.var.rate-[-Inf,0]
## 38319 f.pdays-sometime f.previous-some f.emp.var.rate-[-Inf,0]
##         f.cons.price.idx      f.cons.conf.idx
```

```

## 38799 f.cons.price.idx-[92.2,93.1] f.cons.conf.idx-(-36.4,-26.9]
## 38319 f.cons.price.idx-[92.2,93.1] f.cons.conf.idx-(-36.4,-26.9]
##                               f.euribor3m          f.nr.employed
## 38799 f.euribor3m-[0.635,1.33] f.nr.employed-[4.96e+03,5.1e+03]
## 38319 f.euribor3m-[0.635,1.33] f.nr.employed-[4.96e+03,5.1e+03]
#2 individus mes contributius al 2n eix:
contrib<-sort(res.pca$ind$contrib[,2], decreasing=TRUE)[1:2]; contrib

##      39262      39720
## 0.9322541 0.7912329

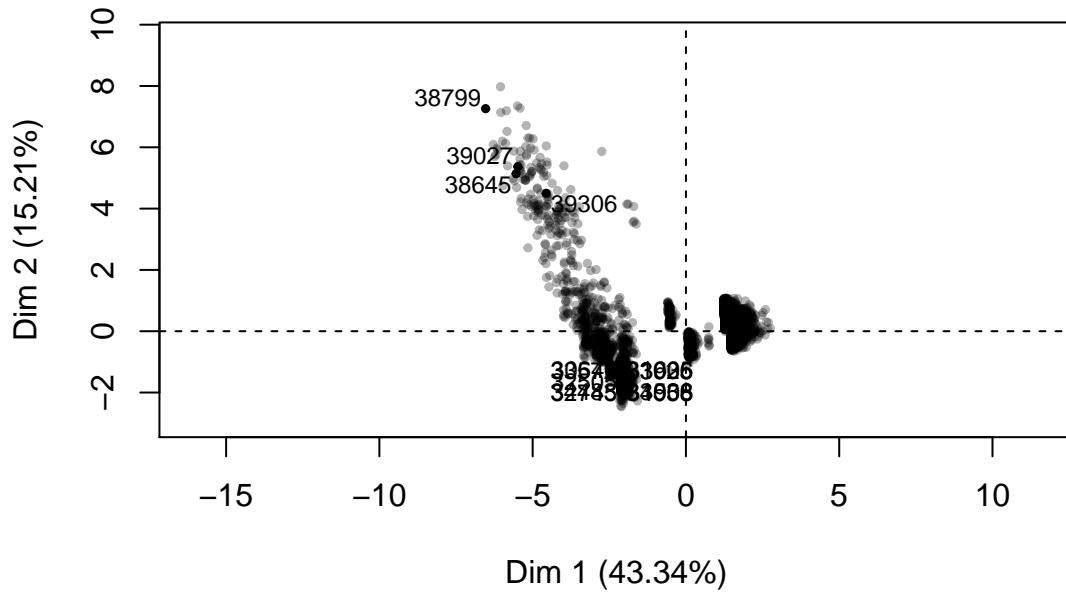
df[c(names(contrib)), ]

##           age         job       marital      education     default
## 39262 80 job-retired marital-married education-basic.4y default-no
## 39720 80 job-retired marital-married education-basic.4y default-no
##           housing    loan       contact      month   day_of_week
## 39262 housing-no loan-no contact-cellular month-mar day_of_week-5fri
## 39720 housing-no loan-no contact-cellular month-may day_of_week-1mon
##           duration campaign pdays previous      poutcome emp.var.rate
## 39262      213        3       6        4 poutcome-success      -1.8
## 39720      382        1       3        3 poutcome-success      -1.8
##           cons.price.idx cons.conf.idx euribor3m nr.employed      y
## 39262      93.369      -34.8      0.649      5008.7 y-yes
## 39720      93.876      -40.0      0.697      5008.7 y-yes
##           num_missings num_outliers num_errors      f.season minutes
## 39262          0            0          0 season-spring 3.550000
## 39720          0            0          0 season-spring 6.366667
##           f.age          f.duration      f.campaign
## 39262 f.age-(47,87] f.duration-(177,316] f.campaign-(2,5]
## 39720 f.age-(47,87] f.duration-(316,1.58e+03] f.campaign-[0,2]
##           f.pdays      f.previous      f.emp.var.rate
## 39262 f.pdays-sometime f.previous-some f.emp.var.rate-[-Inf,0]
## 39720 f.pdays-sometime f.previous-some f.emp.var.rate-[-Inf,0]
##           f.cons.price.idx          f.cons.conf.idx
## 39262 f.cons.price.idx-(93.1,93.7] f.cons.conf.idx-(-36.4,-26.9]
## 39720 f.cons.price.idx-(93.7,94] f.cons.conf.idx-(-41.8,-36.4]
##                               f.euribor3m          f.nr.employed
## 39262 f.euribor3m-[0.635,1.33] f.nr.employed-[4.96e+03,5.1e+03]
## 39720 f.euribor3m-[0.635,1.33] f.nr.employed-[4.96e+03,5.1e+03]

#nomes pinta les etiquetes dels 15 individus mes ben representats!
plot(res.pca, choix="ind", cex=0.75, col.ind="black", select="cos2 15", title="Factor map - 15 individus")

```

## Factor map – 15 individus mes ben representats



```
#2 individus mes ben representats al 1r eix:
repr<-sort(res.pca$ind$cos2[,1], decreasing=TRUE) [1:2]; repr

##      18385      13817
## 0.8444095 0.8442091
#df[c(names(repr)),]

#2 individus mes ben representats al 2n eix:
repr<-sort(res.pca$ind$cos2[,2], decreasing=TRUE) [1:2]; repr

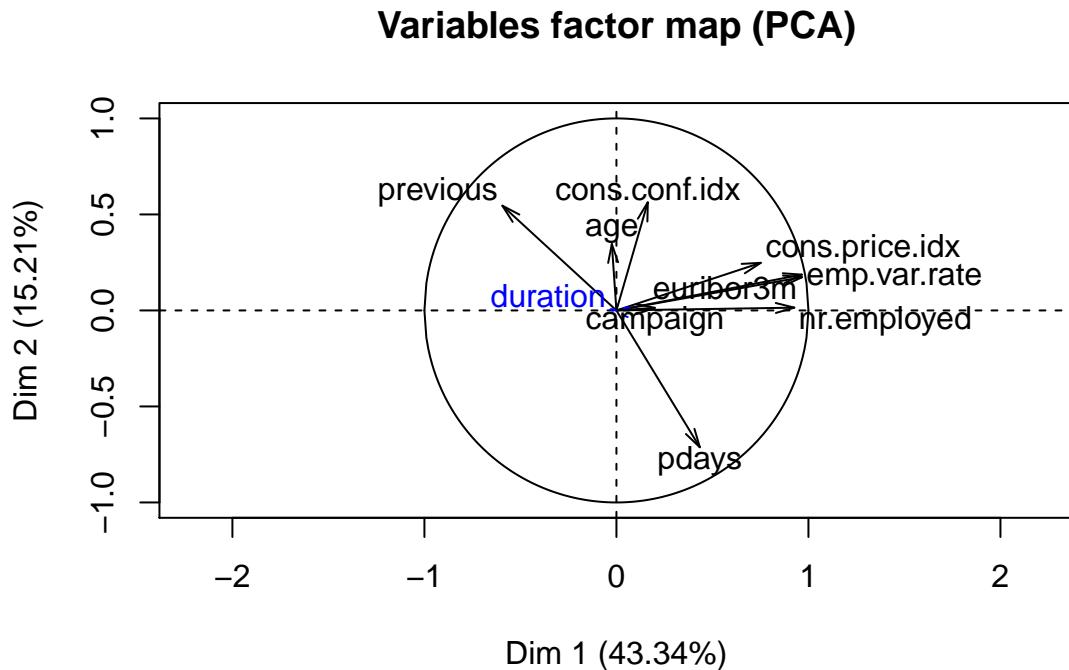
##      39431      39350
## 0.5706147 0.5679583
#df[c(names(repr)),]
```

### Interpretation of the PCA

En el mapa de factors de les variables (2D, primeres dues dimensions) es pot observar en blau la variable “duration” com a suplementaria, la qual surt quasi centrada, el que vol dir que les variables vars\_con utilitzades en el PCA no ens ajuden a dir res o predir els valors de la variable target. La variable previous (numero de contactes en campanyes antigues) està relacionada inversament amb pdays (dies que feia que no es trucava el client per altres campanyes), ja que que com es pot veure en el grafic, ambdues fletxes apunten oposadament. Sembla ser també que tots els indicadors socioeconomics (a excepció d'una mica de cons.conf.idx) apunten en la mateixa direcció, el que vol dir que estan relacionats entre ells i contribueixen d'una manera similar als eixos. Es pot veure el % de contribució de les variables a les tres primeres dimensions mitjançant una eina de la llibreria *ggplot*, on es veu com basicament els indicadors socioeconomics i les variables “previous”, “campaign” i “age” són les variables numèriques que mes han contribuït. El summary ens permet veure també de forma numèrica la contribució (ctr) de cadascuna de les variables en els 4 primers eixos, així com la qualitat de la representació (cos2) de les mateixes en cadascun dels eixos. Si se s'acaben de descriure les variables més contributives, les que estan millor representades (cos2 més proxim a 1) en el primer eix són euribor3m, emp.var.rate i nr.employed; i en el segon eix són pdays, previous i cons.conf.idx. Un l'últim mapa

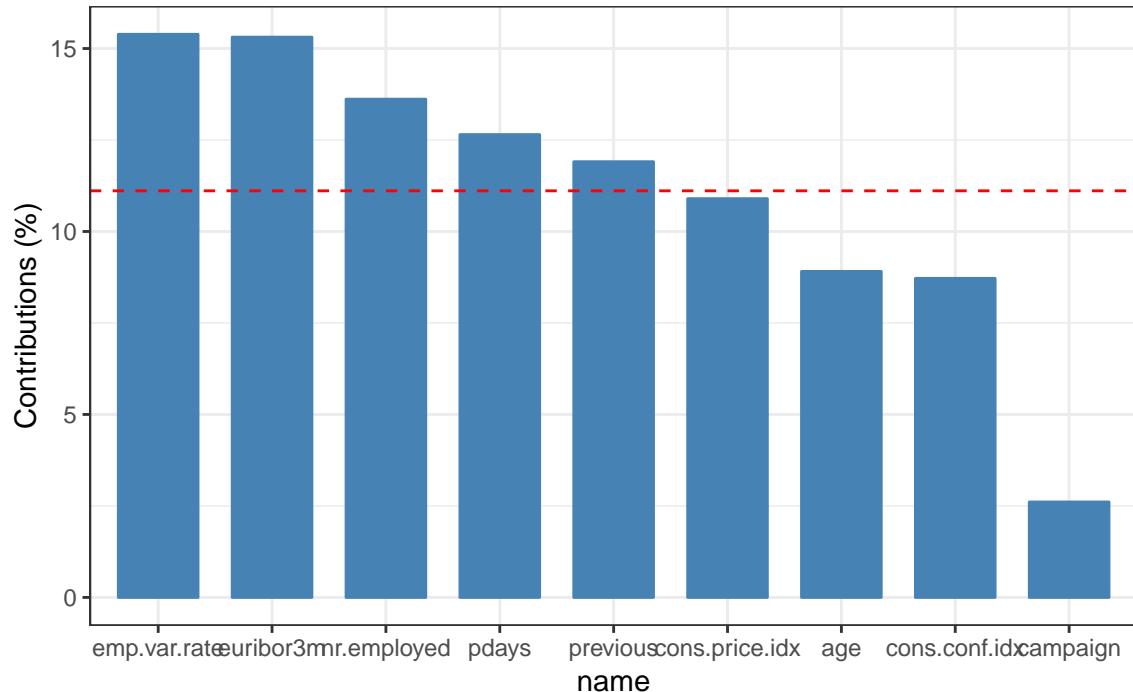
de factors de variables observem els eixos de les dimensions 3 i 4, les quals representen aproximadament un 11% de la variancia de les dades cadascuna.

```
# PCA:  
plot.PCA(res.pca, choix = c("var")) #variables factor map
```



```
#GGPLOT contribution of variables  
#fviz_pca_var(res.pca)  
fviz_contrib(res.pca, choice="var", axes=1:3)+theme_bw()
```

### Contribution of variables to Dim-1–2–3



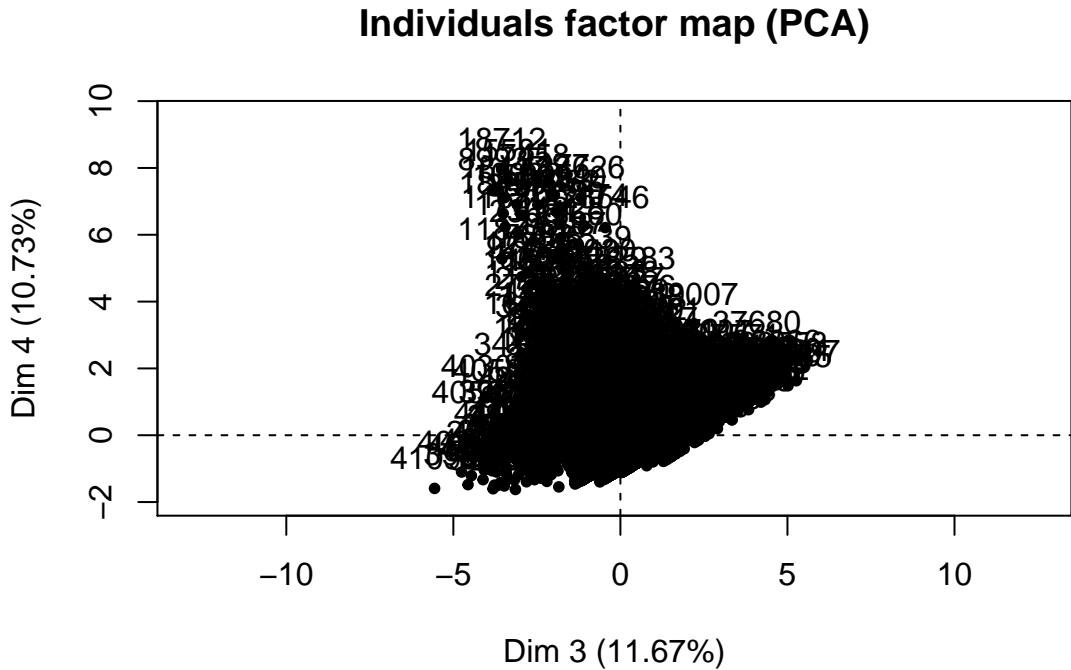
```
summary(res.pca, nb.dec=2, ncp=2, nbind=0)
```

```
##
## Call:
## PCA(X = df[, vars_con], quanti.sup = 2, graph = FALSE)
##
##
## Eigenvalues
##              Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance      3.90    1.37    1.05    0.97    0.86    0.43    0.39
## % of var.  43.34   15.21   11.67   10.73   9.53    4.75   4.38
## Cumulative % of var. 43.34  58.55  70.22  80.95  90.48  95.24  99.61
##                      Dim.8   Dim.9
## Variance      0.02    0.01
## % of var.   0.27    0.12
## Cumulative % of var. 99.88 100.00
##
## Variables
##           Dim.1   ctr  cos2   Dim.2   ctr  cos2
## age        -0.02  0.01  0.00   0.35  8.73  0.12 |
## campaign     0.20  1.03  0.04   0.01  0.01  0.00 |
## pdays       0.43  4.84  0.19  -0.71 37.11  0.51 |
## previous     -0.59  9.05  0.35   0.55 21.75  0.30 |
## emp.var.rate  0.97 23.96  0.93   0.17  2.19  0.03 |
## cons.price.idx  0.75 14.49  0.57   0.25  4.49  0.06 |
## cons.conf.idx  0.16  0.68  0.03   0.56 23.16  0.32 |
## euribor3m     0.97 23.91  0.93   0.19  2.54  0.03 |
## nr.employed    0.93 22.02  0.86   0.01  0.02  0.00 |
```

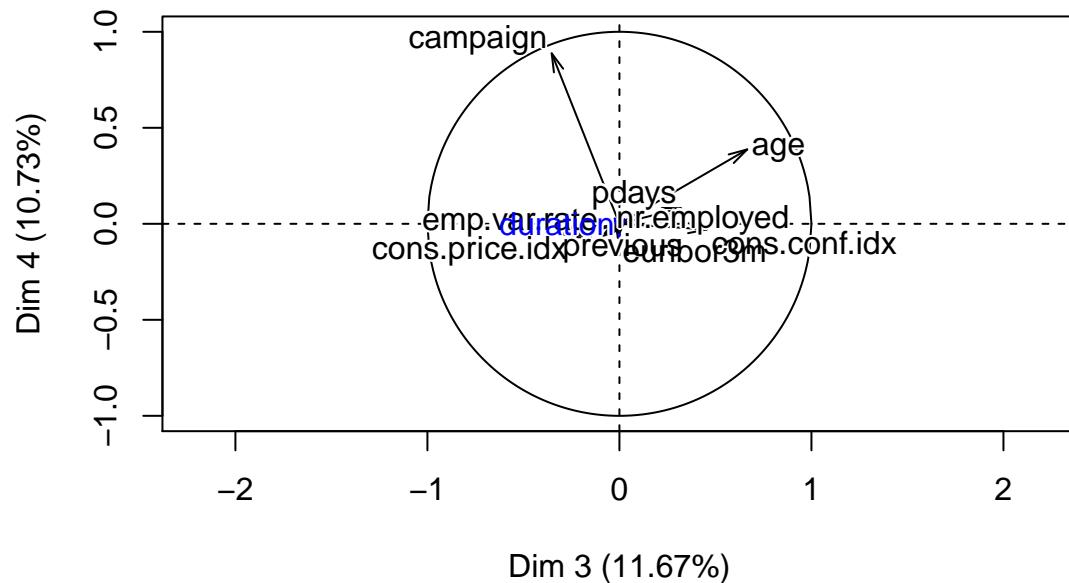
```

## Supplementary continuous variable
##                               Dim.1   cos2   Dim.2   cos2
## duration           | -0.03   0.00 |  0.00   0.00 |
#veure eixos 3 i 4:
res.pca<-PCA(df[, vars_con], quanti.sup=2, axes=3:4)

```



## Variables factor map (PCA)

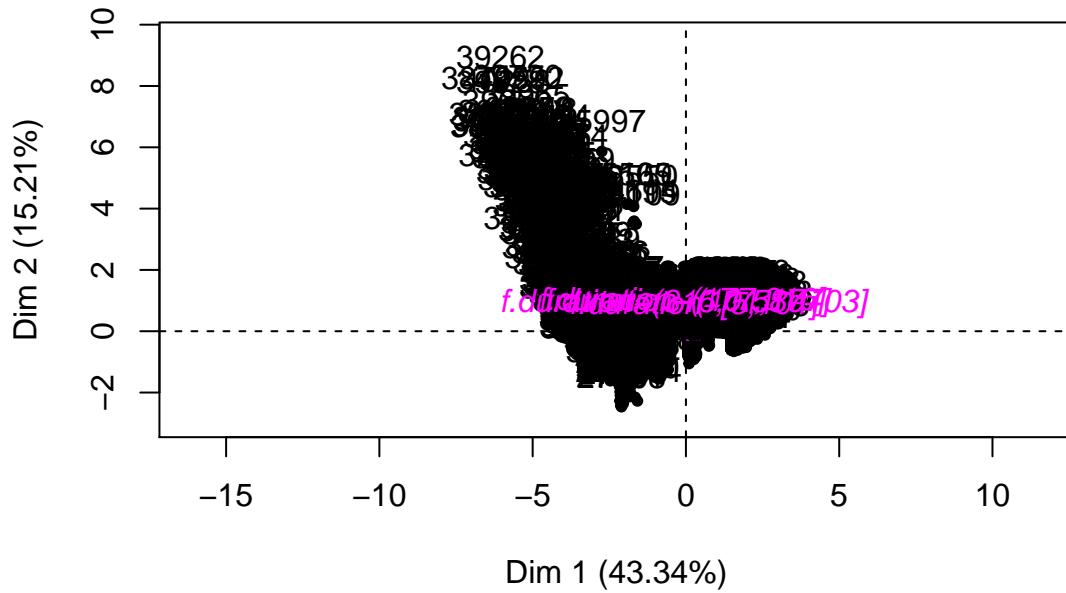


Com que es dificil extreure conclusions a partir nomes de les variables numeriques, en el grafics següents es pot observar el posicionament de les dues variables suplementaries “f.duration” i “f.euribor3m” en el mapa de factors dels individus. En el cas de f.duration, si es mira al mapa de factors ampliat on nomes es representa aquesta variable, sembla ser que té una tendència a creixer cap al 2n i 3r quadrant compost per les dues primeres dimensions. En el cas de f.euribor3m, de la mateixa manera, també es pot trobar un progrésio de les dades al llarg del primer eix, ja que el valor de l'euribor3m augmenta amb el valor de la primera dimensió de les dades.

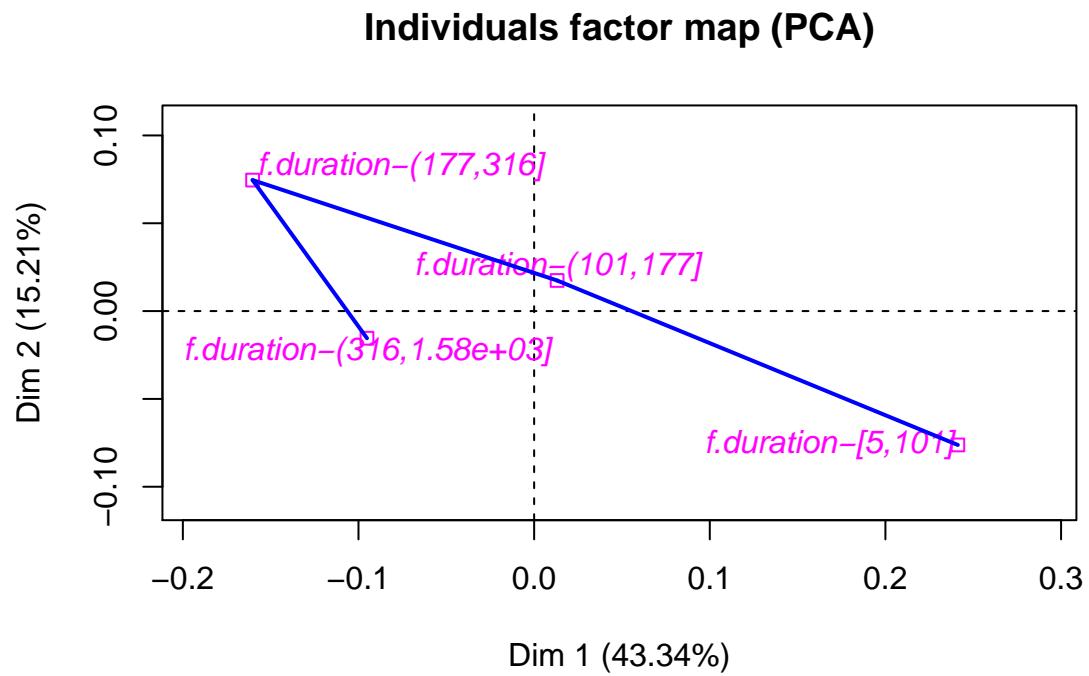
```
par(mfrow=c(1,1))

#ara afegim dues variables suplementaries!
res.pca<-PCA( df[, c("f.duration", vars_con) ], quanti.sup=3, quali.sup=1, graph = FALSE )
plot.PCA(res.pca, choix = c("ind") )
```

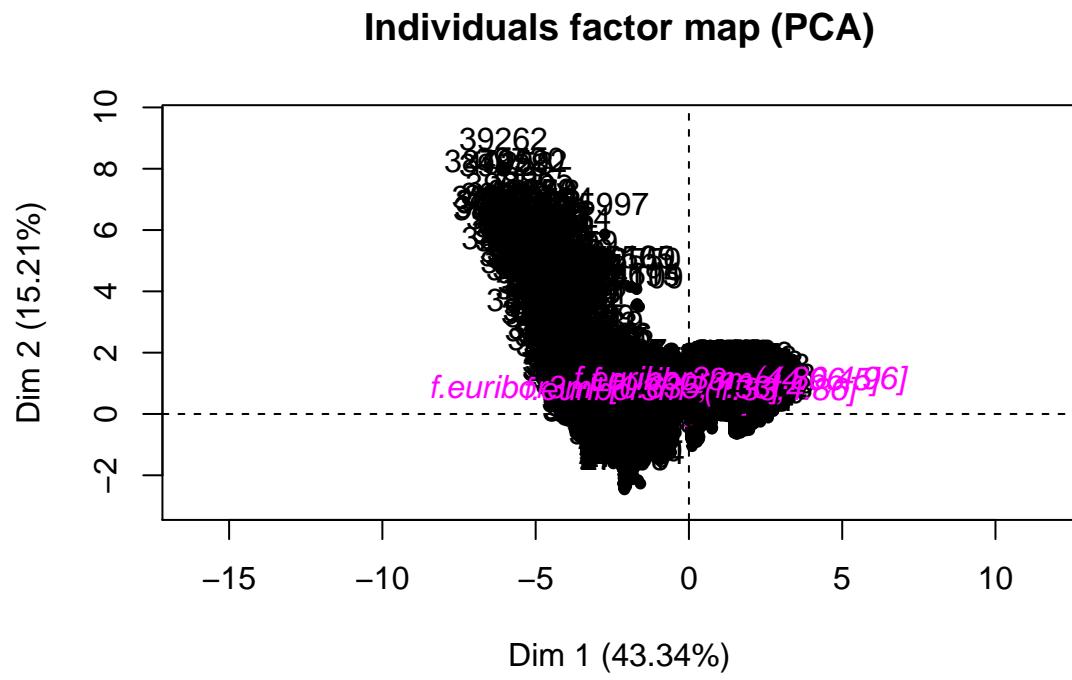
## Individuals factor map (PCA)



```
#Unir punts de variables suplementaries en el factor map per a una bona representació:  
plot.PCA(res.pca, choix="ind", invisible="ind")  
lines(res.pca$quali.sup$coord[1:2, 1:2], col="blue", lwd="2")  
lines(res.pca$quali.sup$coord[2:3, 1:2], col="blue", lwd="2")  
lines(res.pca$quali.sup$coord[3:4, 1:2], col="blue", lwd="2")
```



```
res.pca<-PCA( df[, c("f.euribor3m", vars_con) ], quanti.sup=3, quali.sup=1, graph = FALSE )
plot.PCA(res.pca, choix = c("ind") )
```



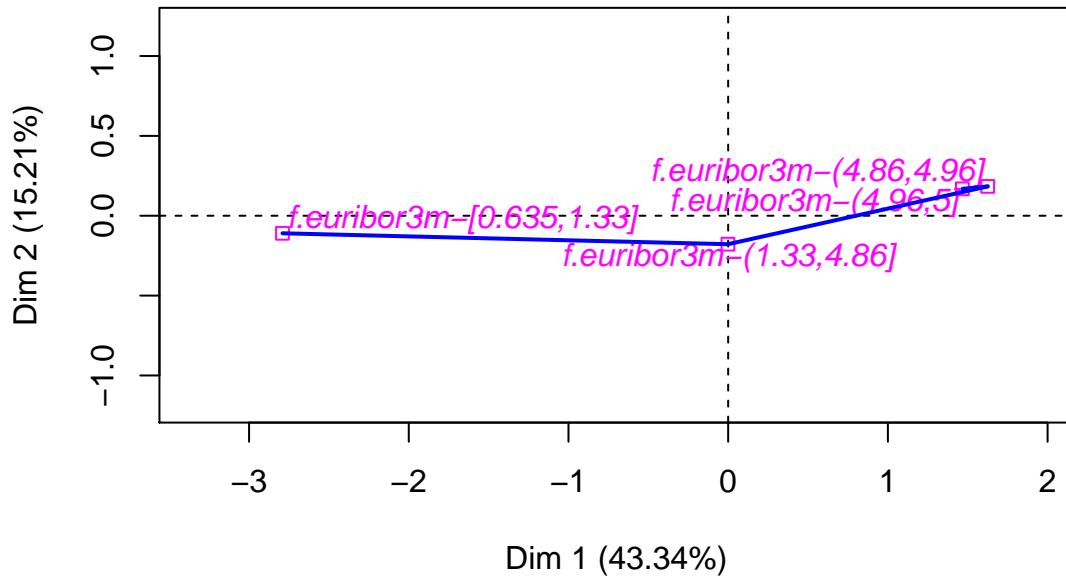
```
plot.PCA(res.pca, choix="ind", invisible="ind")
lines(res.pca$quali.sup$coord[1:2, 1:2], col="blue", lwd="2")
```

```

lines(res.pca$quali.sup$coord[2:3, 1:2], col="blue", lwd="2")
lines(res.pca$quali.sup$coord[3:4, 1:2], col="blue", lwd="2")

```

## Individuals factor map (PCA)



## K-Means Classification (Partitioning - Supervised learning)

K-means es un algoritme de clustering que té com a objectiu agrupar les observacions en un determinat nombre de grups o clusters els quals comparteixen característiques similars. Dit d'altra manera, agrupa els individus de manera que els que estan dins un mateix cluster tinguin unes distàncies euclidianes entre ells més petites que respecte els individus d'altres clusters. Cal tenir en compte que com que el paràmetre passat a la crida *kmeans* és el nombre de clusters que s'han d'obtenir (i no el conjunt inicial de centres), es seleccionen aleatoriament un conjunt inicial de  $k$  centres de cluster. Això vol dir que aquesta selecció aleatoria pot tenir gran influència en el resultat final, el qual serà diferent en cada execució de l'algoritme. Per tal de reduir soroll de les dades innecessari, treballarem només amb les 4 primeres dimensions obtingudes del PCA.

L'objecte retornat per la crida *kmeans* ens permet consultar diferents atributs. L'atribut *withinss* correspon a la suma del quadrat de les distàncies inter-cluster, és a dir, hi ha un valor per a cada cluster. L'atribut *betweenss* és un sol valor mig que correspon a la suma del quadrat de les distàncies inter-cluster. A partir d'aquest valor mig i el total de les distàncies tots podem obtenir el nivell de representació que obtenim només amb els centres de gravetat dels clusters sobre el conjunt de les dades.

```

#only 4 significant axes in order to avoid unnecessary noise
dclu<-res.pca$ind$coord[,1:4]

#fixed number of clusters (a random set of rows are chosen as the initial centers)
kcla<-kmeans(dclu, 4)
summary(kcla)

##          Length Class  Mode
##  cluster      4986   -none- numeric

```

```

## centers      16  -none- numeric
## totss        1   -none- numeric
## withinss     4   -none- numeric
## tot.withinss 1   -none- numeric
## betweenss    1   -none- numeric
## size         4   -none- numeric
## iter         1   -none- numeric
## ifault       1   -none- numeric

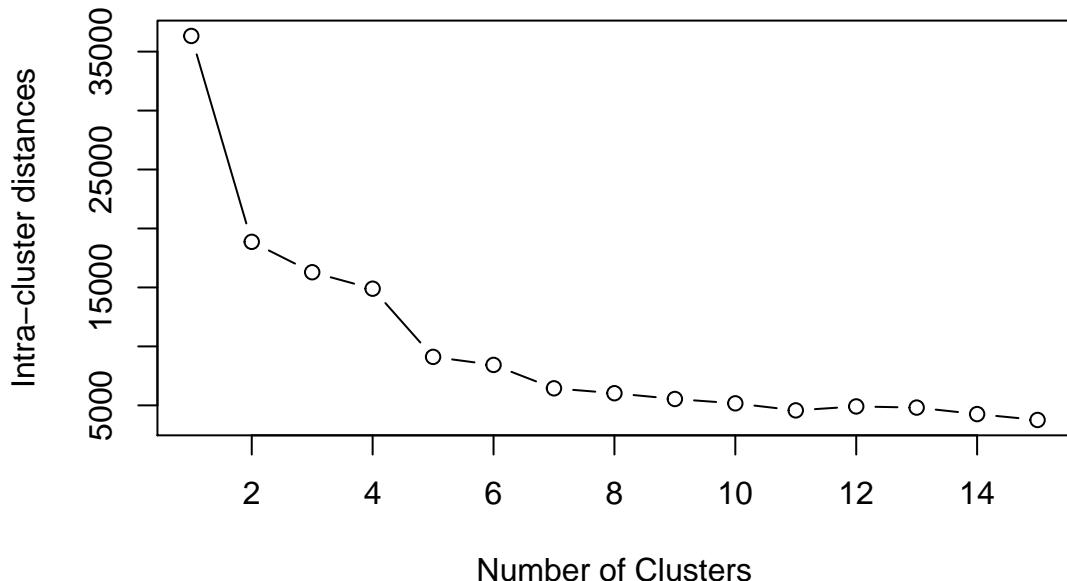


```

Si executem k-means per a diferents valors k podem observar com evolucionen les distancies intra-clusters o el % de representació obtingut en cada cas. L'execució d'unes quantes vegades del següent chunk de codi ens ha ajudat en la selecció del nombre de clusters, ja que per la regla de l'últim colze agafem una k=4.

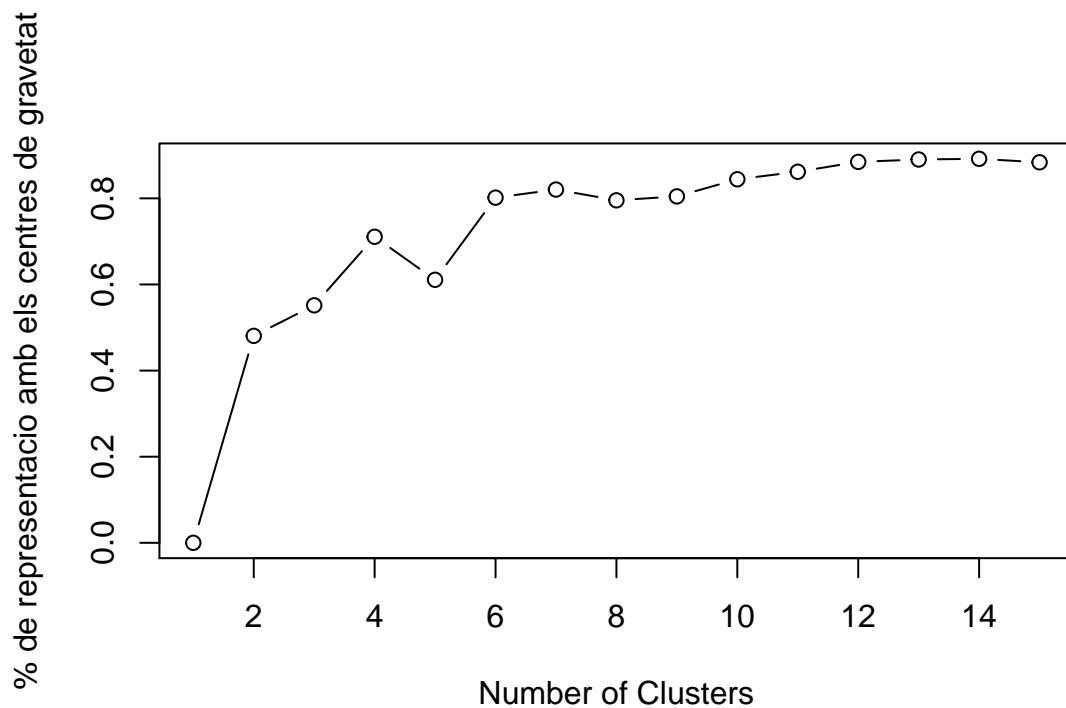
```
wss <- sum(kmeans(dclu,1)$withinss) #k=1
for (i in 2:15) wss[i] <- sum(kmeans(dclu,i)$withinss) #k=2 to 15

plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Intra-cluster distances")
```



```
km<-kmeans(dclu,1)
repr <- km$betweenss/km$totss #k=1
for (i in 2:15){
  km<-kmeans(dclu,i)
  repr[i] <- km$betweenss/km$totss
}

plot(1:15, repr, type="b", xlab="Number of Clusters", ylab="% de representacio amb els centres de gravitat")
```



## Descripcio dels clusters

En el **primer cluster** podem observar com les mostres que l'integren es trobaven en un situació econòmica poc favorable, els factors socio econòmics sobrepassen els valors mitjans en perjudici de la població. A més, l'edat es troba per sota de la mitjana global, la mitjana de contactes anteriors també es superior i el nombre de contactes en la present campanya es troba lleugerament per sobre. Per al **segon cluster**, podem observar com els contactes en la present campanya es troba més de 8 punts per sobre de la mitjana global. En relació als factors socio econòmics no en podem treure conclusions clares, l'índex de preus al consumidor i el nombre de empleats es troben per sobre de la mitjana mentre que l'euribor i la variança no són favorables. El nombre de dies transcorreguts des de l'últim contacte incrementa lleugerament i els contactes previs decauen de manera més dràstica. Com a característica principal del **tercer cluster** podem destacar que els contactes en la present campanya augmenten en 0.3 punts respecte a la mitjana global. Altres característiques amb les quals descriure el cluster poden ser el lleuger increment en la duració de les trucades i una lleugera baixada dels contactes en la present campanya. El valors oferts pels factors socio-econòmics no ens mostren una tendència clara de la situació socio-econòmica dels integrants del cluster. Per acabar, en el **\*quart cluster** podem observar com l'edat es troba 7 anys per sobre de la mitjana global i els valors socio-econòmics ofereixen condicions menys favorables que les que ofereix la mitjana global. El dies transcorreguts des del últim contacte augmenten lleugerament, mentre que els contactes en la present campanya i el nombre de contactes previs presenten valors inferiors respecte la mitjana global.

```
kclusters<-as.data.frame(kcla$cluster)
colnames(kclusters) = c("kcluster")
dades_cl<-merge(df[,vars_con], kclusters, by=0) #merge by row.names
dades_cl<-dades_cl[,-1] #eliminem columna row.names (correspon als num de fila anteriors)

dades_cl$kcluster<-as.factor(dades_cl$kcluster)
catdes(dades_cl, 11) #catdes per kcluster

## 
## Link between the cluster variable and the quantitative variables
```

```

## =====
##          Eta2      P-value
## campaign    0.550954418 0.000000e+00
## pdays       0.843397506 0.000000e+00
## previous    0.497123213 0.000000e+00
## emp.var.rate 0.885174767 0.000000e+00
## cons.price.idx 0.466478991 0.000000e+00
## euribor3m    0.968786556 0.000000e+00
## nr.employed   0.854396978 0.000000e+00
## cons.conf.idx 0.156579598 1.331688e-183
## age          0.007413433 4.436231e-08
## duration     0.004163600 1.164812e-04
##
## Description of each cluster by quantitative variables
## =====
## $`1`
##          v.test Mean in category Overall mean sd in category
## previous    47.006702      1.7040816 0.15984757 0.8476761
## cons.conf.idx 10.701121     -36.9015306 -40.42591256 6.6506826
## age          4.807612      43.5867347 40.06799037 15.2016254
## campaign    -3.491970      1.9285714 2.53512993 1.4160163
## cons.price.idx -8.941565     93.2074031 93.57245006 0.7438635
## emp.var.rate  -20.992479     -2.2653061 0.06446049 0.8827256
## euribor3m    -21.618653      0.9852041 3.61448034 0.6460893
## nr.employed  -26.738597    5032.1142857 5166.47621340 50.3600384
## pdays         -64.840841      6.9795918 18.52647413 5.0678651
## Overall sd           p.value
## previous      0.4691873 0.000000e+00
## cons.conf.idx 4.7037753 1.005546e-26
## age          10.4532458 1.527439e-06
## campaign     2.4808187 4.794719e-04
## cons.price.idx 0.5830800 3.836987e-19
## emp.var.rate  1.5850448 7.683224e-98
## euribor3m    1.7370025 1.199253e-103
## nr.employed  71.7679377 1.675774e-157
## pdays        2.5433666 0.000000e+00
##
## $`2`
##          v.test Mean in category Overall mean sd in category
## euribor3m    60.27191      4.80935925 3.61448034 0.2809108
## emp.var.rate 57.09961      1.09741978 0.06446049 0.5004926
## nr.employed  55.65945    5212.06708568 5166.47621340 17.7798692
## cons.price.idx 40.44479     93.84160304 93.57245006 0.3883514
## cons.conf.idx 20.00076     -39.35216672 -40.42591256 2.9906074
## pdays         16.31272      19.00000000 18.52647413 0.0000000
## campaign    -17.33235      2.04437903 2.53512993 1.1987427
## previous     -25.09387      0.02547139 0.15984757 0.1575519
## Overall sd           p.value
## euribor3m    1.7370025 0.000000e+00
## emp.var.rate 1.5850448 0.000000e+00
## nr.employed  71.7679377 0.000000e+00
## cons.price.idx 0.5830800 0.000000e+00
## cons.conf.idx 4.7037753 5.424298e-89
## pdays        2.5433666 8.014910e-60

```

```

## campaign      2.4808187 2.681214e-67
## previous     0.4691873 5.802097e-139
##
## $`3`
##           v.test Mean in category Overall mean sd in category
## previous      10.154358    0.2653766   0.15984757   0.4805084
## pdays         8.344120    18.9965446   18.52647413   0.1083352
## duration     1.976647   260.7090532  250.62194144  230.6955148
## age          -3.161983   39.3358673   40.06799037  12.3015208
## campaign     -8.141244   2.0877678   2.53512993  1.5373384
## cons.conf.idx -26.840556  -43.2223912  -40.42591256  6.1255475
## cons.price.idx -45.767655   92.9813511   93.57245006  0.4485074
## nr.employed   -55.561025  5078.1531444  5166.47621340 38.5615601
## emp.var.rate   -59.764486   -2.0337941   0.06446049  0.5599866
## euribor3m     -62.693709   1.2023663   3.61448034  0.2507120
##           Overall sd      p.value
## previous      0.4691873 3.168895e-24
## pdays        2.5433666 7.174593e-17
## duration     230.3904064 4.808160e-02
## age          10.4532458 1.566989e-03
## campaign     2.4808187 3.912382e-16
## cons.conf.idx 4.7037753 1.087242e-158
## cons.price.idx 0.5830800 0.000000e+00
## nr.employed   71.7679377 0.000000e+00
## emp.var.rate   1.5850448 0.000000e+00
## euribor3m     1.7370025 0.000000e+00
##
## $`4`
##           v.test Mean in category Overall mean sd in category
## campaign      52.399372    9.565625  2.535130e+00   4.1917411
## emp.var.rate  13.494227    1.221250  6.446049e-02   0.4830098
## nr.employed   13.137136   5217.467500 5.166476e+03  20.5272488
## euribor3m     13.091301    4.844319  3.614480e+00   0.4266389
## cons.price.idx 11.219672   93.926263 9.357245e+01   0.3911105
## pdays         3.442468    19.000000 1.852647e+01   0.0000000
## age          2.848041    41.678125 4.006799e+01   9.5009616
## duration     -3.911171   201.887500 2.506219e+02  228.3486777
## previous     -6.053037    0.006250 1.598476e-01   0.0788095
##           Overall sd      p.value
## campaign      2.4808187 0.000000e+00
## emp.var.rate   1.5850448 1.691208e-41
## nr.employed   71.7679377 2.017171e-39
## euribor3m     1.7370025 3.692280e-39
## cons.price.idx 0.5830800 3.264776e-29
## pdays        2.5433666 5.764313e-04
## age          10.4532458 4.398926e-03
## duration     230.3904064 9.184971e-05
## previous      0.4691873 1.421399e-09

```

## Hierarchical Clustering (Unsupervised learning)

Aquest punt compren la realització d'una clusterització aglomerativa jerarquizada dels individus. HCPC de la llibreria FactoMineR utilitza les distàncies entre clusters diferents, per tal de minimitzar la inercia

inter-cluster.

La primera de les comandes HCPC esta comentada ja que requereix l'interacio de l'usuari per a triar per on tallar l'arbre de clusters que es mostra a l'usuari. Despres d'estar interactuant amb diferents opcions de numero de clusters, s'ha decidit finalment agafar nb.clust = 7, tal i com es pot veure a continuacio. Es mostren els grafics corresponents a l'arbre el qual s'ha tallat a l'altura de 7 clusters, aixi com el mapa de factors de les dues primeres dimensions del PCA.

```
#PCA calculat amb 4 significant axes:
```

```
vars_con<-names(df)[c(1, 12:14, 16:20)]; vars_con
```

```
## [1] "age"           "campaign"        "pdays"          "previous"  
## [5] "emp.var.rate"  "cons.price.idx" "cons.conf.idx"  "euribor3m"  
## [9] "nr.employed"
```

```
res.pca<-PCA( df[, c("duration", "y", "loan", "month", "job", "poutcome", "education", "housing", vars_
```

```
#res.hcpc<-HCPC(res.pca, order=TRUE)
```

```
res.hcpc<-HCPC(res.pca, nb.clust=7, order=TRUE, graph=FALSE); res.hcpc
```

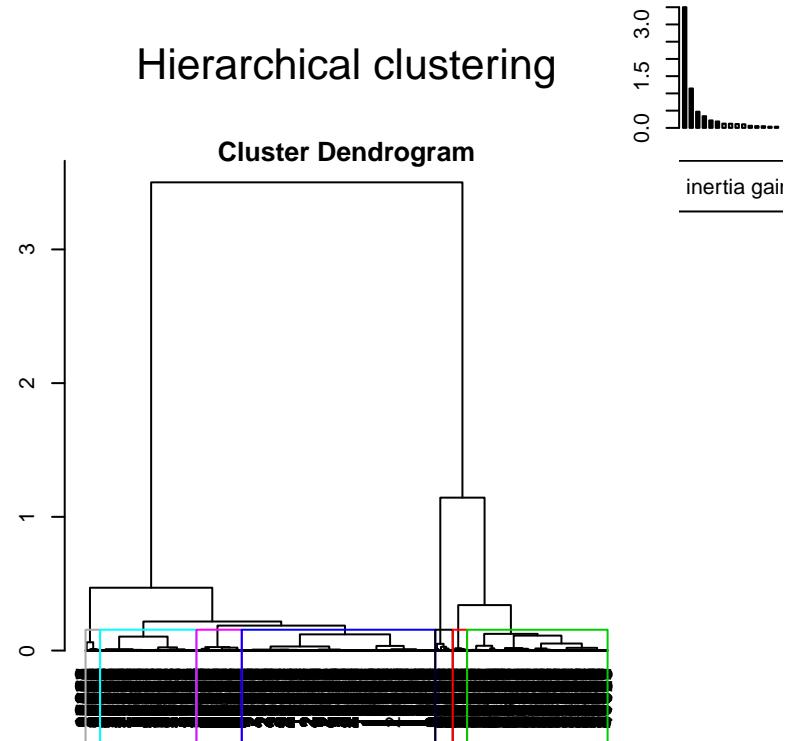
```
## **Results for the Hierarchical Clustering on Principal Components**
```

```
##   name  
## 1 "$data.clust"  
## 2 "$desc.var"  
## 3 "$desc.var$quanti.var"  
## 4 "$desc.var$quanti"  
## 5 "$desc.var$test.chi2"  
## 6 "$desc.axes$category"  
## 7 "$desc.axes"  
## 8 "$desc.axes$quanti.var"  
## 9 "$desc.axes$quanti"  
## 10 "$desc.ind"  
## 11 "$desc.ind$para"  
## 12 "$desc.ind$dist"  
## 13 "$call"  
## 14 "$call$t"  
##   description  
## 1 "dataset with the cluster of the individuals"  
## 2 "description of the clusters by the variables"  
## 3 "description of the cluster var. by the continuous var."  
## 4 "description of the clusters by the continuous var."  
## 5 "description of the cluster var. by the categorical var."  
## 6 "description of the clusters by the categories."  
## 7 "description of the clusters by the dimensions"  
## 8 "description of the cluster var. by the axes"  
## 9 "description of the clusters by the axes"  
## 10 "description of the clusters by the individuals"  
## 11 "parangons of each clusters"  
## 12 "specific individuals"  
## 13 "summary statistics"  
## 14 "description of the tree"
```

```
attributes(res.hcpc)
```

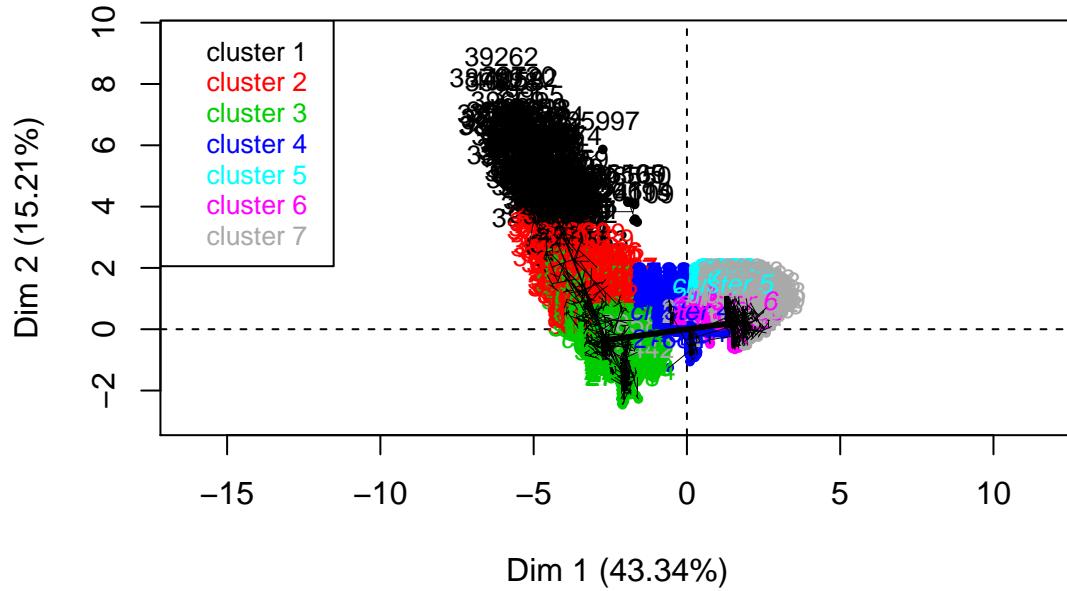
```
## $names  
## [1] "data.clust" "desc.var"    "desc.axes"   "call"       "desc.ind"
```

```
##  
## $class  
## [1] "HCPC"  
  
plot.HCPC(res.hcpc, choice="tree")
```



```
plot.HCPC(res.hcpc, choice="map")
```

## Factor map



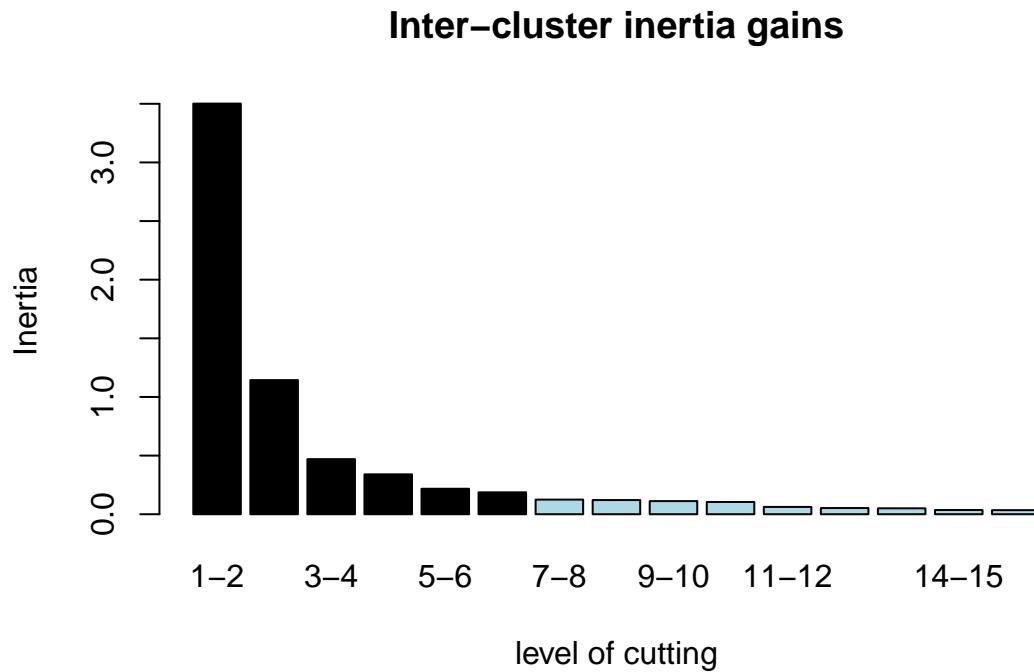
```
fviz_cluster(res.hcpc)
```



El guanys (d'1 a 2 clusters, de 2 a 3, de 3 a 4, etc.) d'inercia es poden observar en negre a la figura inferior, on tambe es pot veure la relacio entre altres nivells de tall i els seus guanys. Els valors *within* ens mostren la inercia dins de cada cluster, que si fos un sol cluster (1) seria la total del conjunt de dades, la qual equival

a la suma de tots els guanys d'inerçies *inert.gain*. Amb la selecció de 6 clusters que hem agafat estariem obtenint una representació de les dades del 80.4%, la qual es molt bona! Finalment, també podem veure la distribució en nombre d'individus en els diferents clusters.

```
plot.HCPC(res.hcpc, choice="bar")
```



```
res.hcpc$call$t$inert.gain[1:6]
## [1] 3.5016630 1.1436272 0.4699006 0.3400760 0.2171351 0.1867086
sum(res.hcpc$call$t$inert.gain)
## [1] 7.285793
res.hcpc$call$t$within[1:7]
## [1] 7.285793 3.784130 2.640503 2.170602 1.830526 1.613391 1.426683
#X clusters corresponen al %? d'inercia
sum(res.hcpc$call$t$inert.gain[1:6]) / sum(res.hcpc$call$t$inert.gain) * 100
## [1] 80.41829
table(res.hcpc$data.clust$clust) #nombre d'individus en cada cluster
##
##      1      2      3      4      5      6      7
## 176  313 1155  430 1268 1441  203
```

### Sobre les variables categoriques que han estat incloses com a supplementaries

Aquí analitzarem la representació i relació de les variables categòriques que s'han inclos com a supplementaries en el PCA destinat a la clusterització jeràrquica. Aquestes variables han estat les següents: “y”, “loan”,

“month”, “job”, “poutcome”, “education” i “housing”.

En un primer output basat en el *test chi2* veiem rapidament quines d'aquestes variables discretes supplementaries estan significativament ( $p\text{-value} < 0.05$ ) relacionades globalment amb la partició de clusters establecida. Les variables “month” i “poutcome” tenen un  $p$ -valor tan baix que R no el pot representar i ens el marca com a 0; de la mateixa manera, les variables “y(target)”, “job”, “education” i “housing” estan també relacionades amb el particionament seleccionat.

En un segon output desglosat per numero de cluster veiem la sobrerepresentació i infrarepresentació de diferents categories en els diferents clusters, ajudant així a caracteritzar els diferents grups formats. La presentació de les dades es la mateixa que la mostrada en el catdes però amb el numero de cluster com a variable fixada, per tant seran interpretades de la mateixa manera descrita en el primer deliverable. En una primera **descripció general**, podem veure com els clusters 1 i 2, i en menor mesura el 3, tenen una sobrerepresentació d'acceptació de producte financer. Pel que fa a la resta de clusters, tenen una infrarepresentació del mateix. El 71% d'acceptacions del producte financer es troben als 3 primers clusters, mentres que el 29% restant es troben als altres 4. Cada cluster a mes, té un lleuger esbiaix en certes categories professionals i d'educació. El **primer cluster** es troba caracteritzat per una immensa sobrerepresentació dels clients amb acceptació d'una campanya anterior (poutcome-success és un 86% dins el cluster 1 respecte un 3% global); això es tradueix també en que el 97% d'individus que havien acceptat anteriorment un producte estan situats en aquest cluster. El **segon cluster** es troba sobrerepresentat pels individus retirats (27% respecte un 4% global), el 41% dels quals pertanyen a aquest cluster. També trobem que mes de la meitat de trucades realitzades als mesos d'octubre, setembre i desembre son classificades dins del cluster número 2. A mes a mes, el segon i en especial el **tercer cluster** (59% dels poutcome-failure pertanyen a aquest) es troben força sobrerepresentats per individus que anteriorment no van acceptar una campanya (però aquesta sí). També inclouen una sobrerepresentació d'individus housing-yes. Aquest tercer cluster compren a mes una sobrerepresentació dels mesos d'abril, maig i juny. El 99.7% d'individus del **quart cluster** van ser contactats el mes de novembre, el que es tradueix en una immensa sobrerepresentació en aquest grup, que a mes no va acceptar el producte financer (97.0%). El **cinc cluster**, **sise cluster** i **sete cluster** estan formats integralment per individus dels quals no es té informació d'exit o fracas en contactes de campanyes anteriors. El cinquè cluster té una sobrerepresentació dels mesos agost i maig; mentres que el sise i sete clusters la tenen dels mesos juliol i juny. Aquests dos últims clusters tenen a mes una sobrerepresentació housing-no (i infra de housing-yes), essent el sete cluster lleugerament diferent del sise pel fet de contenir també una sobrerepresentació de l'agost i una major infrarepresentació d'acceptació del producte financer.

```
# Factors globally related to clustering partition:
```

```
res.hcpc$desc.var$test.chi2
```

```
##          p.value df
## month      0.000000e+00 54
## poutcome   0.000000e+00 12
## y          4.530266e-154  6
## job        1.045918e-146 60
## education  2.331835e-28 30
## housing    2.178875e-07  6
```

```
# Categories over/under represented in each cluster:
```

```
res.hcpc$desc.var$category
```

```
## $`1`
```

	Cla/Mod	Mod/Cla	Global
## poutcome=poutcome-success	96.794872	85.795455	3.1287605
## y=y-yes	18.850987	59.659091	11.1712796
## month=month-oct	19.587629	10.795455	1.9454473
## month=month-mar	22.727273	8.522727	1.3237064
## month=month-sep	18.032787	6.250000	1.2234256
## month=month-dec	26.923077	3.977273	0.5214601

```

## job=job-retired          9.268293 10.795455  4.1115122
## education=education-university.degree 5.051151 44.886364 31.3678299
## job=job-student         10.000000 5.681818  2.0056157
## month=month-nov          5.719921 16.477273 10.1684717
## job=job-housemaid        8.148148 6.250000  2.7075812
## job=job-admin.            4.494382 31.818182 24.9899719
## poutcome=poutcome-failure 5.241090 14.204545  9.5667870
## education=education-basic.9y 2.086050 9.090909 15.3830726
## job=job-services          1.405622 3.977273  9.9879663
## job=job-blue-collar       1.964133 13.068182 23.4857601
## month=month-jul           1.326900 6.250000  16.6265544
## month=month-may           1.378518 13.636364 34.9177698
## y=y-no                   1.603071 40.340909 88.8287204
## poutcome=poutcome-nonexistent 0.000000 0.000000 87.3044525
##                                     p.value      v.test
## poutcome=poutcome-success    3.881038e-254 34.051314
## y=y-yes                     2.249261e-57 15.964754
## month=month-oct             6.983202e-10 6.166468
## month=month-mar              5.595931e-09 5.828406
## month=month-sep              7.963183e-06 4.466171
## month=month-dec              2.444489e-05 4.219865
## job=job-retired             1.204719e-04 3.845164
## education=education-university.degree 1.343088e-04 3.818421
## job=job-student              3.281252e-03 2.940082
## month=month-nov              8.467352e-03 2.632843
## job=job-housemaid            1.043595e-02 2.561039
## job=job-admin.                3.744524e-02 2.080876
## poutcome=poutcome-failure    4.315053e-02 2.022250
## education=education-basic.9y 1.353542e-02 -2.469362
## job=job-services              3.081976e-03 -2.959441
## job=job-blue-collar            4.706686e-04 -3.496917
## month=month-jul               3.415301e-05 -4.143845
## month=month-may               1.098069e-10 -6.452794
## y=y-no                      2.249261e-57 -15.964754
## poutcome=poutcome-nonexistent 7.847689e-169 -27.695786
##                                     p.value      v.test
## $^2`                           Cla/Mod     Mod/Cla     Global
## job=job-retired             40.975610 26.8370607  4.1115122
## month=month-oct              58.762887 18.2108626  1.9454473
## y=y-yes                      18.312388 32.5878594 11.1712796
## month=month-sep              55.737705 10.8626198  1.2234256
## poutcome=poutcome-failure    18.867925 28.7539936  9.5667870
## month=month-dec              61.538462 5.1118211  0.5214601
## month=month-aug              10.329986 23.0031949 13.9791416
## education=education-basic.4y 10.318949 17.5718850 10.6899318
## education=education-university.degree 8.056266 40.2555911 31.3678299
## month=month-mar              16.666667 3.5143770  1.3237064
## housing=housing-yes           7.192661 62.6198083 54.6530285
## poutcome=poutcome-success     1.923077 0.9584665  3.1287605
## housing=housing-no             5.174701 37.3801917 45.3469715
## job=job-entrepreneur           1.298701 0.6389776  3.0886482
## education=education-high.school 4.515599 17.5718850 24.4283995
## month=month-jun               3.374233 7.0287540 13.0766145

```

```

## job=job-services          2.208835  3.5143770  9.9879663
## education=education-basic.9y 2.346806  5.7507987 15.3830726
## month=month-jul          2.412545  6.3897764 16.6265544
## poutcome=poutcome-nonexistent 5.053986 70.2875399 87.3044525
## job=job-blue-collar       1.707942  6.3897764 23.4857601
## y=y-no                   4.764055  67.4121406 88.8287204
## month=month-may          1.608271  8.9456869 34.9177698
##
##                                     p.value    v.test
## job=job-retired            3.206853e-50 14.901828
## month=month-oct           6.590944e-45 14.061041
## y=y-yes                  3.834595e-26 10.576387
## month=month-sep          8.371333e-26 10.502960
## poutcome=poutcome-failure 9.880353e-24 10.042825
## month=month-dec          1.246107e-13 7.411785
## month=month-aug          8.737157e-06 4.446276
## education=education-basic.4y 1.483350e-04 3.793840
## education=education-university.degree 6.031404e-04 3.430198
## month=month-mar           3.078719e-03 2.959767
## housing=housing-yes      3.315110e-03 2.936900
## poutcome=poutcome-success 1.178823e-02 -2.518421
## housing=housing-no       3.315110e-03 -2.936900
## job=job-entrepreneur     3.124217e-03 -2.955244
## education=education-high.school 2.715285e-03 -2.998257
## month=month-jun          4.715598e-04 -3.496412
## job=job-services         1.072126e-05 -4.402091
## education=education-basic.9y 6.730357e-08 -5.398225
## month=month-jul          3.048594e-08 -5.538623
## poutcome=poutcome-nonexistent 1.400973e-16 -8.264657
## job=job-blue-collar      1.157386e-16 -8.287413
## y=y-no                   3.834595e-26 -10.576387
## month=month-may          5.102772e-28 -10.973892
##
## $`3`
##                                     Cla/Mod   Mod/Cla   Global
## month=month-apr            87.741935 23.5497835 6.2174087
## month=month-may           39.402642 59.3939394 34.9177698
## poutcome=poutcome-failure 59.329140 24.5021645 9.5667870
## job=job-student          60.000000 5.1948052 2.0056157
## month=month-mar           60.606061 3.4632035 1.3237064
## y=y-yes                  33.572711 16.1904762 11.1712796
## housing=housing-yes      25.614679 60.4329004 54.6530285
## job=job-blue-collar      26.985482 27.3593074 23.4857601
## education=education-high.school 26.026273 27.4458874 24.4283995
## education=education-basic.9y 26.075619 17.3160173 15.3830726
## month=month-dec          7.692308 0.1731602 0.5214601
## month=month-oct          14.432990 1.2121212 1.9454473
## job=job-management       18.734793 6.6666667 8.2430806
## job=job-housemaid       12.592593 1.4718615 2.7075812
## job=job-technician       18.718593 12.9004329 15.9647012
## education=education-basic.4y 16.510319 7.6190476 10.6899318
## housing=housing-no       20.212295 39.5670996 45.3469715
## y=y-no                   21.855949 83.8095238 88.8287204
## job=job-retired          6.341463 1.1255411 4.1115122
## month=month-jun          12.269939 6.9264069 13.0766145

```

```

## poutcome=poutcome-success          1.282051  0.1731602  3.1287605
## month=month-nov                  3.353057  1.4718615 10.1684717
## poutcome=poutcome-nonexistent    19.986216 75.3246753 87.3044525
## month=month-aug                  1.865136  1.1255411 13.9791416
## month=month-jul                  1.809409  1.2987013 16.6265544
##
##                                     p.value      v.test
## month=month-apr                 8.449977e-140 25.170408
## month=month-may                 6.491023e-85 19.526860
## poutcome=poutcome-failure       2.400092e-72 17.988443
## job=job-student                1.919893e-15 7.946413
## month=month-mar                7.101962e-11 6.518488
## y=y-yes                         2.708602e-09 5.948349
## housing=housing-yes            6.372241e-06 4.513646
## job=job-blue-collar            4.638064e-04 3.500833
## education=education-high.school 6.941146e-03 2.699655
## education=education-basic.9y   3.965890e-02 2.057284
## month=month-dec                4.946472e-02 -1.964564
## month=month-oct                3.354840e-02 -2.125458
## job=job-management             2.394486e-02 -2.258013
## job=job-housemaid              1.872587e-03 -3.109730
## job=job-technician              9.690944e-04 -3.299349
## education=education-basic.4y   7.133660e-05 -3.971783
## housing=housing-no              6.372241e-06 -4.513646
## y=y-no                           2.708602e-09 -5.948349
## job=job-retired                5.601283e-11 -6.554008
## month=month-jun                6.684182e-14 -7.493929
## poutcome=poutcome-success       8.577891e-16 -8.045664
## month=month-nov                7.092964e-39 -13.041624
## poutcome=poutcome-nonexistent   4.123377e-39 -13.082912
## month=month-aug                3.750561e-66 -17.179978
## month=month-jul                1.381625e-80 -19.011049
##
## $^4
##
##                                     Cla/Mod     Mod/Cla     Global
## month=month-nov                 84.615385 99.7674419 10.168472
## y=y-no                          9.415218 96.9767442 88.828720
## job=job-management              17.761557 16.9767442 8.243081
## poutcome=poutcome-failure       16.561845 18.3720930 9.566787
## job=job-entrepreneur            23.376623 8.3720930 3.088648
## education=education-university.degree 11.700767 42.5581395 31.367830
## job=job-unemployed              16.393443 4.6511628 2.446851
## job=job-self-employed           14.189189 4.8837209 2.968311
## job=job-admin.                  7.223114 20.9302326 24.989972
## job=job-student                 3.000000 0.6976744 2.005616
## education=education-basic.9y   6.258149 11.1627907 15.383073
## month=month-sep                 0.000000 0.0000000 1.223426
## month=month-mar                 0.000000 0.0000000 1.323706
## job=job-blue-collar             6.233988 16.9767442 23.485760
## poutcome=poutcome-nonexistent   8.063405 81.6279070 87.304452
## month=month-oct                 0.000000 0.0000000 1.945447
## poutcome=poutcome-success       0.000000 0.0000000 3.128761
## y=y-yes                          2.333932 3.0232558 11.171280
## month=month-apr                 0.000000 0.0000000 6.217409
## month=month-jun                 0.000000 0.0000000 13.076615

```

## month=month-aug	0.000000	0.0000000	13.979142
## month=month-jul	0.000000	0.0000000	16.626554
## month=month-may	0.000000	0.0000000	34.917770
##	p.value	v.test	
## month=month-nov	0.000000e+00	Inf	
## y=y-no	1.445137e-10	6.411058	
## job=job-management	6.094995e-10	6.187954	
## poutcome=poutcome-failure	3.411201e-09	5.910476	
## job=job-entrepreneur	1.906432e-08	5.620285	
## education=education-university.degree	3.261080e-07	5.107698	
## job=job-unemployed	5.040211e-03	2.804453	
## job=job-self-employed	2.256934e-02	2.280647	
## job=job-admin.	3.956647e-02	-2.058247	
## job=job-student	2.833524e-02	-2.192613	
## education=education-basic.9y	8.935086e-03	-2.614528	
## month=month-sep	3.940443e-03	-2.882891	
## month=month-mar	2.495043e-03	-3.023942	
## job=job-blue-collar	6.095792e-04	-3.427316	
## poutcome=poutcome-nonexistent	4.218766e-04	-3.526007	
## month=month-oct	1.451605e-04	-3.799206	
## poutcome=poutcome-success	6.138981e-07	-4.986793	
## y=y-yes	1.445137e-10	-6.411058	
## month=month-apr	2.792982e-13	-7.304018	
## month=month-jun	3.466586e-28	-11.008782	
## month=month-aug	3.067865e-30	-11.426923	
## month=month-jul	2.141900e-36	-12.598870	
## month=month-may	1.367297e-85	-19.606258	
##			
## \$`5`			
##	Cla/Mod	Mod/Cla	Global
## poutcome=poutcome-nonexistent	29.129336	100.0000000	87.3044525
## month=month-aug	52.367288	28.78548896	13.9791416
## month=month-may	34.462952	47.31861199	34.9177698
## y=y-no	27.387672	95.66246057	88.8287204
## education=education-basic.4y	38.836773	16.32492114	10.6899318
## education=education-professional.course	30.406504	14.74763407	12.3345367
## loan=loan-no	26.179749	87.06624606	84.5768151
## job=job-retired	33.658537	5.44164038	4.1115122
## housing=housing-no	26.890756	47.94952681	45.3469715
## job=job-housemaid	33.333333	3.54889590	2.7075812
## job=job-blue-collar	27.754056	25.63091483	23.4857601
## housing=housing-yes	24.220183	52.05047319	54.6530285
## job=job-unemployed	17.213115	1.65615142	2.4468512
## job=job-admin.	22.953451	22.55520505	24.9899719
## education=education-high.school	22.495895	21.60883281	24.4283995
## loan=loan-yes	21.326398	12.93375394	15.4231849
## month=month-dec	0.000000	0.00000000	0.5214601
## education=education-university.degree	21.419437	26.41955836	31.3678299
## month=month-sep	0.000000	0.00000000	1.2234256
## month=month-mar	0.000000	0.00000000	1.3237064
## month=month-jun	15.644172	8.04416404	13.0766145
## job=job-student	2.000000	0.15772871	2.0056157
## month=month-oct	1.030928	0.07886435	1.9454473
## poutcome=poutcome-success	0.000000	0.00000000	3.1287605

```

## y=y-yes          9.874327  4.33753943 11.1712796
## month=month-apr 0.000000  0.00000000  6.2174087
## poutcome=poutcome-failure 0.000000  0.00000000  9.5667870
## month=month-nov 0.000000  0.00000000 10.1684717
##
##                                     p.value    v.test
## poutcome=poutcome-nonexistent 5.289007e-88 19.886855
## month=month-aug   1.191159e-61 16.567794
## month=month-may   3.493765e-26 10.585108
## y=y-no           2.522644e-22 9.718174
## education=education-basic.4y  5.108301e-13 7.222387
## education=education-professional.course 2.901818e-03 2.977952
## loan=loan-no      3.988941e-03 2.879035
## job=job-retired  7.243663e-03 2.685430
## housing=housing-no 3.131061e-02 2.153103
## job=job-housemaid 3.740547e-02 2.081311
## job=job-blue-collar 3.801325e-02 2.074712
## housing=housing-yes 3.131061e-02 -2.153103
## job=job-unemployed 3.030370e-02 -2.166098
## job=job-admin.    1.971720e-02 -2.331686
## education=education-high.school 6.416463e-03 -2.725703
## loan=loan-yes     3.988941e-03 -2.879035
## month=month-dec   4.751074e-04 -3.494412
## education=education-university.degree 9.020632e-06 -4.439409
## month=month-sep   1.482866e-08 -5.663534
## month=month-mar   3.344948e-09 -5.913706
## month=month-jun   1.357270e-10 -6.420613
## job=job-student  8.674956e-11 -6.488405
## month=month-oct   1.130837e-11 -6.788783
## poutcome=poutcome-success 5.633481e-21 -9.396637
## y=y-yes           2.522644e-22 -9.718174
## month=month-apr   9.931295e-42 -13.533406
## poutcome=poutcome-failure 3.651539e-65 -17.047444
## month=month-nov   1.753726e-69 -17.619231
##
## $`6`
##                                     Cla/Mod    Mod/Cla    Global
## month=month-jul       62.243667  35.8084663 16.6265544
## poutcome=poutcome-nonexistent 33.103607 100.0000000 87.3044525
## month=month-jun       57.055215  25.8154060 13.0766145
## y=y-no                 30.571235  93.9625260 88.8287204
## job=job-services      36.746988  12.6995142 9.9879663
## housing=housing-no    31.357806  49.2019431 45.3469715
## education=education-basic.9y 33.898305  18.0430257 15.3830726
## education=education-high.school 32.266010  27.2727273 24.4283995
## job=job-technician    32.035176  17.6960444 15.9647012
## job=job-blue-collar  31.255337  25.3990285 23.4857601
## job=job-management    22.627737  6.4538515 8.2430806
## housing=housing-yes   26.862385  50.7980569 54.6530285
## job=job-student      14.000000  0.9715475 2.0056157
## month=month-dec      0.000000  0.0000000  0.5214601
## month=month-oct      6.185567  0.4163775 1.9454473
## education=education-basic.4y 18.011257  6.6620402 10.6899318
## month=month-sep      0.000000  0.0000000  1.2234256
## month=month-mar      0.000000  0.0000000  1.3237064

```

```

## y=y-yes          15.619390  6.0374740 11.1712796
## month=month-may 20.964963  25.3296322 34.9177698
## poutcome=poutcome-success 0.000000  0.0000000 3.1287605
## job=job-retired 0.000000  0.0000000 4.1115122
## month=month-apr 0.000000  0.0000000 6.2174087
## poutcome=poutcome-failure 0.000000  0.0000000 9.5667870
## month=month-nov 0.000000  0.0000000 10.1684717
##
##                                p.value      v.test
## month=month-jul        2.193896e-108 22.116481
## poutcome=poutcome-nonexistent 2.146626e-102 21.485087
## month=month-jun        5.177232e-59 16.198381
## y=y-no                 1.177463e-14 7.718457
## job=job-services       6.653178e-05 3.988360
## housing=housing-no     4.994477e-04 3.481052
## education=education-basic.9y 1.044519e-03 3.278253
## education=education-high.school 3.082735e-03 2.959365
## job=job-technician     3.462568e-02 2.112709
## job=job-blue-collar   4.313066e-02 2.022442
## job=job-management    2.864908e-03 -2.981873
## housing=housing-yes   4.994477e-04 -3.481052
## job=job-student       4.497429e-04 -3.509031
## month=month-dec       1.370593e-04 -3.813416
## month=month-oct       2.045152e-08 -5.608138
## education=education-basic.4y 1.058277e-09 -6.100363
## month=month-sep       7.911717e-10 -6.146689
## month=month-mar       1.400443e-10 -6.415845
## y=y-yes                1.177463e-14 -7.718457
## month=month-may       3.650636e-20 -9.197885
## poutcome=poutcome-success 2.829400e-24 -10.165406
## job=job-retired      7.352963e-32 -11.746586
## month=month-apr       1.961207e-48 -14.624460
## poutcome=poutcome-failure 9.590135e-76 -18.417043
## month=month-nov       8.816429e-81 -19.034599
##
## $`7`
##                                Cla/Mod      Mod/Cla      Global      p.value
## poutcome=poutcome-nonexistent 4.6634505 100.0000000 87.304452 5.792869e-13
## month=month-jul             8.0820265 33.0049261 16.626554 5.404009e-09
## month=month-jun             8.2822086 26.6009852 13.076615 1.418693e-07
## y=y-no                      4.4027997 96.0591133 88.828720 2.087025e-04
## month=month-aug             5.7388809 19.7044335 13.979142 2.141058e-02
## housing=housing-no          4.6881911 52.2167488 45.346972 4.567393e-02
## housing=housing-yes         3.5596330 47.7832512 54.653028 4.567393e-02
## month=month-oct             0.0000000 0.0000000 1.945447 1.704213e-02
## poutcome=poutcome-success  0.0000000 0.0000000 3.128761 1.374795e-03
## y=y-yes                     1.4362657 3.9408867 11.171280 2.087025e-04
## month=month-apr             0.3225806 0.4926108 6.217409 2.660092e-05
## month=month-nov             0.5917160 1.4778325 10.168472 6.310181e-07
## month=month-may             2.1826536 18.7192118 34.917770 2.409376e-07
## poutcome=poutcome-failure 0.0000000 0.0000000 9.566787 8.711382e-10
##
##                                v.test
## poutcome=poutcome-nonexistent 7.205271
## month=month-jul             5.834228
## month=month-jun             5.262811

```

```

## y=y-no          3.708243
## month=month-aug 2.300665
## housing=housing-no 1.998394
## housing=housing-yes -1.998394
## month=month-oct -2.385798
## poutcome=poutcome-success -3.199891
## y=y-yes         -3.708243
## month=month-apr -4.200770
## month=month-nov -4.981474
## month=month-may -5.164612
## poutcome=poutcome-failure -6.131391

```

## Sobre les variables quantitatives (numeriques)

A continuacio es realitzara una altra descripcio dels diferents clusters formats pero basant-nos en les variables quantitatives (numeriques), les quals han estat utilitzades en el PCA. En el primer output es poden veure les que han estat “p-provades” com a globalment relacionades amb la clusteritzacio, mentres que el detall es pot veure en el segon output.

El **primer cluster** esta caracteritzat per individus (sempre respecte la mitja global) mes contactats en campanyes anteriors (previous) i menys en l'actual (campaign); aixi com individus que han estat contactats molt mes recentment (pdays). La duracio de les trucades en aquest cluster esta 35 segons per sobre la mitja global, aixi com tambe podem veure lleugers esbiaixos en els indicadors socioeconomics (valors baixos de euribor3m, emp.var.rate, nr.employed i cons.conf.idx). El **segon cluster** te una mitjana d'edat 12 anys per sobre la mitjana global, si be tambe una major desvacio estandard dins la categoria; a part d'un esbiaix similar a l'anterior amb els indicadors socioeconomics. El **tercer cluster** es similar al cluster 2 pero amb una mitja d'edat 4 anys per sota de la global; a mes en aquest grup la duracio de les trucades esta per sobre la mitja per 14 segons. Els quatre altres clusters contenen noms individus els quals no han estat contactats abans per cap altra campanya (mitja de pdays=19.0 amb un sd de 0) i els seus indicadors socioeconomics tenen un comportament similar en l'esbiaix. El **quart cluster** pero, a diferencia dels altres tres, no mostra una mitja de 0.0 en la variable “previous”, el que fa pensar que aquests individus han estat contactat abans pero no necessariament per a una campanya d'un producte. La duracio en aquest cluster es 30 segons per sota la mitja global. Entre els **cinque cluster** i **sise cluster** la diferencia principal esta en l'edat, que que el cinque agrupa individus per sobre la mitja i el sise per sota. Per ultim, el **sete cluster** destaca per a una mitja de duracio de les trucades extremadament curta (65 segons per sota la mitja global), aixi com per un numero de vegades que l'individu ha estat contactat en l'actual campanya molt per sobre de la mitja global (11.3 respecte 2.5).

```
# Numeric (quantitative) variables globally related to clustering partition:
res.hcpc$desc.var$quanti.var
```

```

##                  Eta2    P-value
## age            0.335988921 0.00000e+00
## campaign      0.540040103 0.00000e+00
## pdays          0.925018800 0.00000e+00
## previous       0.465947411 0.00000e+00
## emp.var.rate   0.960040097 0.00000e+00
## cons.price.idx 0.614485994 0.00000e+00
## cons.conf.idx  0.503365350 0.00000e+00
## euribor3m      0.988716657 0.00000e+00
## nr.employed    0.879442070 0.00000e+00
## duration       0.006588486 1.08361e-05
res.hcpc$desc.var$quanti

```

```
## $`1`
```

```

##          v.test Mean in category Overall mean sd in category
## previous      43.973874    1.687500   0.15984757  0.8848873
## cons.conf.idx  8.805587   -37.359091  -40.42591256  6.5879649
## age           3.296336    42.619318   40.06799037 14.8938468
## duration      2.067896    285.897727  250.62194144 213.7620565
## campaign     -3.562842    1.880682    2.53512993  1.3282215
## cons.price.idx -7.388255   93.253477   93.57245006  0.7521195
## emp.var.rate   -19.367388   -2.208523   0.06446049  0.8800737
## euribor3m     -20.279982   1.006216    3.61448034  0.6758827
## nr.employed   -25.213543   5032.493750  5166.47621340 52.1226287
## pdays          -67.905435   5.738636   18.52647413  3.5659192
##                         Overall sd p.value
## previous          0.4691873 0.000000e+00
## cons.conf.idx     4.7037753 1.301691e-18
## age              10.4532458 9.795477e-04
## duration         230.3904064 3.864981e-02
## campaign         2.4808187 3.668615e-04
## cons.price.idx   0.5830800 1.487681e-13
## emp.var.rate     1.5850448 1.454512e-83
## euribor3m        1.7370025 1.932036e-91
## nr.employed      71.7679377 2.845704e-140
## pdays            2.5433666 0.000000e+00
##
## $`2`
##          v.test Mean in category Overall mean sd in category
## cons.conf.idx   25.044818  -33.9789137 -40.42591256  6.3573009
## age             19.567738   51.2619808  40.06799037 16.7801531
## previous         7.959341   0.3642173   0.15984757  0.6047713
## pdays            3.149566   18.9648562  18.52647413  0.4099927
## campaign        -5.588859   1.7763578   2.53512993  1.2567327
## cons.price.idx  -26.189093   92.7367668  93.57245006  0.5664489
## euribor3m        -28.684949   0.8877157   3.61448034  0.2248025
## nr.employed     -31.489513   5042.7990415  5166.47621340 36.8870017
## emp.var.rate    -32.543552   -2.7584665   0.06446049  0.6816163
##                         Overall sd p.value
## cons.conf.idx   4.7037753 1.988210e-138
## age              10.4532458 2.913265e-85
## previous         0.4691873 1.729578e-15
## pdays            2.5433666 1.635132e-03
## campaign         2.4808187 2.285660e-08
## cons.price.idx   0.5830800 3.537697e-151
## euribor3m        1.7370025 5.878957e-181
## nr.employed      71.7679377 1.209022e-217
## emp.var.rate     1.5850448 2.583254e-232
##
## $`3`
##          v.test Mean in category Overall mean sd in category
## previous         8.754589   0.2658009   0.15984757  0.4829555
## pdays            7.006604   18.9861472  18.52647413  0.3326118
## duration        2.386888   264.8069264  250.62194144 234.8283557
## campaign        -5.291156   2.1965368   2.53512993  1.6640810
## age              -13.956513   36.3047619  40.06799037  8.5250191
## cons.price.idx  -35.103266   93.0444814  93.57245006  0.3872219
## cons.conf.idx   -42.294031  -45.5575758  -40.42591256  3.1552802

```

```

## nr.employed -43.023029 5086.8302165 5166.47621340 33.6303225
## emp.var.rate -46.855120 -1.8512554 0.06446049 0.3273476
## euribor3m -52.114578 1.2794519 3.61448034 0.1767441
## Overall sd p.value
## previous 0.4691873 2.048491e-18
## pdays 2.5433666 2.441716e-12
## duration 230.3904064 1.699168e-02
## campaign 2.4808187 1.215453e-07
## age 10.4532458 2.871432e-44
## cons.price.idx 0.5830800 6.010017e-270
## cons.conf.idx 4.7037753 0.000000e+00
## nr.employed 71.7679377 0.000000e+00
## emp.var.rate 1.5850448 0.000000e+00
## euribor3m 1.7370025 0.000000e+00
##
## $`4`
## v.test Mean in category Overall mean sd in category
## nr.employed 8.848963 5195.7546512 5166.47621340 9.392794e-01
## euribor3m 6.251387 4.1150930 3.61448034 7.151549e-02
## pdays 4.038403 19.0000000 18.52647413 0.000000e+00
## emp.var.rate -2.253765 -0.1002326 0.06446049 4.816817e-03
## duration -2.828692 220.5767442 250.62194144 2.118224e+02
## campaign -5.593856 1.8953488 2.53512993 1.308550e+00
## cons.conf.idx -7.300497 -42.0090698 -40.42591256 1.878559e-01
## cons.price.idx -13.893678 93.1989674 93.57245006 2.138667e-02
## Overall sd p.value
## nr.employed 71.767938 8.833617e-19
## euribor3m 1.737002 4.068245e-10
## pdays 2.543367 5.381623e-05
## emp.var.rate 1.585045 2.421094e-02
## duration 230.390406 4.673870e-03
## campaign 2.480819 2.220806e-08
## cons.conf.idx 4.703775 2.867066e-13
## cons.price.idx 0.583080 6.918945e-44
##
## $`5`
## v.test Mean in category Overall mean sd in category
## emp.var.rate 31.018314 1.256861 0.06446049 0.15450722
## euribor3m 30.778624 4.911099 3.61448034 0.05291853
## age 28.431726 47.276025 40.06799037 6.83638157
## nr.employed 25.303757 5210.519322 5166.47621340 18.52135643
## cons.conf.idx 23.501417 -37.744874 -40.42591256 2.61453060
## cons.price.idx 20.434823 93.861426 93.57245006 0.29713094
## pdays 7.676659 19.000000 18.52647413 0.00000000
## campaign -3.366631 2.332570 2.53512993 1.51977446
## previous -14.047446 0.000000 0.15984757 0.00000000
## Overall sd p.value
## emp.var.rate 1.5850448 3.053044e-211
## euribor3m 1.7370025 5.064155e-208
## age 10.4532458 8.199565e-178
## nr.employed 71.7679377 2.904135e-141
## cons.conf.idx 4.7037753 3.944955e-122
## cons.price.idx 0.5830800 8.198980e-93
## pdays 2.5433666 1.632909e-14

```

```

## campaign      2.4808187 7.609255e-04
## previous     0.4691873 7.986314e-45
##
## $`6`
##           v.test Mean in category Overall mean sd in category
## emp.var.rate 35.593579    1.317765   0.06446049   0.15938206
## cons.price.idx 34.404615    94.018094  93.57245006   0.31164559
## euribor3m    34.019646    4.927205   3.61448034   0.04863405
## nr.employed  32.673604   5218.568217 5166.47621340 16.19038972
## pdays        8.380911    19.000000  18.52647413   0.00000000
## cons.conf.idx 3.801019   -40.028730 -40.42591256  2.94116617
## campaign     -6.468091    2.178667   2.53512993  1.37338503
## previous     -15.336148    0.000000  0.15984757   0.00000000
## age          -27.777266   33.617627  40.06799037  5.41378149
##           Overall sd p.value
## emp.var.rate 1.5850448 1.761028e-277
## cons.price.idx 0.5830800 2.151036e-259
## euribor3m    1.7370025 1.141427e-253
## nr.employed  71.7679377 3.704005e-234
## pdays        2.5433666 5.251979e-17
## cons.conf.idx 4.7037753 1.441025e-04
## campaign     2.4808187 9.924849e-11
## previous     0.4691873 4.384517e-53
## age          10.4532458 8.165488e-170
##
## $`7`
##           v.test Mean in category Overall mean sd in category
## campaign      51.681982   11.349754 2.535130e+00   4.2854821
## emp.var.rate  11.261387    1.291626 6.446049e-02   0.3725545
## nr.employed  10.764761   5219.589655 5.166476e+03 18.9160146
## euribor3m    10.662441    4.887768 3.614480e+00   0.3732829
## cons.price.idx 9.771322   93.964148 9.357245e+01   0.3729365
## pdays        2.708103    19.000000 1.852647e+01   0.0000000
## duration     -4.128695   185.226601 2.506219e+02 230.3643674
## previous     -4.955532    0.000000 1.598476e-01   0.0000000
##           Overall sd p.value
## campaign     2.4808187 0.0000000e+00
## emp.var.rate 1.5850448 2.035309e-29
## nr.employed 71.7679377 5.049245e-27
## euribor3m   1.7370025 1.525396e-26
## cons.price.idx 0.5830800 1.494883e-22
## pdays        2.5433666 6.766903e-03
## duration    230.3904064 3.648276e-05
## previous    0.4691873 7.213279e-07

```

## Descripcio dels clusters mitjancant individus

Aqui descrivim els clusters mitjancant els individus, concretament ens centrarem en els individus que estan al centre de gravetat del cluster (para-parangons) i els que estan mes allunyats de la resta de clusters, es a dir, els mes propis del cluster en questio i menys dels altres (dist-especifics). En les següents grafiques es poden veure representats en blau els parangons i en taronja els individus específics per a cada cluster de la classificacio.

```
# Description of the clusters by individuals:  
names(res.hcpc$desc.ind)
```

```
## [1] "para" "dist"  
res.hcpc$desc.ind$para #parangons of each clusters
```

```
## Cluster: 1  
## 31670 30284 37727 36469 30158  
## 0.3892465 0.6099896 0.6405242 0.6517560 0.6658754
```

```
## -----  
## Cluster: 2  
## 38732 38503 37268 39000 37397  
## 0.4162434 0.5086237 0.5839632 0.5993900 0.6629911
```

```
## -----  
## Cluster: 3  
## 36703 36516 36745 36444 36765  
## 0.2333947 0.2626528 0.2918447 0.3074016 0.4287422
```

```
## -----  
## Cluster: 4  
## 27423 26065 25081 25324 26601  
## 0.2429945 0.2473604 0.2495809 0.2495809 0.2519679
```

```
## -----  
## Cluster: 5  
## 19520 19421 20018 23464 23491  
## 0.1425541 0.1427061 0.1428585 0.1431648 0.1434728
```

```
## -----  
## Cluster: 6  
## 12276 14034 16531 16228 18537  
## 0.2192499 0.2196483 0.2196483 0.2199155 0.2247631
```

```
## -----  
## Cluster: 7  
## 20528 15744 21650 23413 21920  
## 0.1715338 0.1947761 0.2990218 0.3655228 0.3672151
```

```
res.hcpc$desc.ind$dist #specific individuals
```

```
## Cluster: 1  
## 40258 39828 40592 39359 39659  
## 10.171491 10.130157 9.325015 9.035001 9.033797
```

```
## -----  
## Cluster: 2  
## 38207 38192 38185 38558 38253  
## 6.441959 6.210520 6.144412 6.115588 6.069564
```

```
## -----  
## Cluster: 3  
## 27734 34291 40877 27890 33321  
## 4.536960 4.490433 4.483955 4.461203 4.400579
```

```
## -----  
## Cluster: 4  
## 26974 26571 27451 25845 26975  
## 2.669911 2.636325 2.431859 2.426201 2.423020
```

```
## -----  
## Cluster: 5  
## 22846 743 22712 22178 20336
```

```

## 2.849191 2.722699 2.706876 2.651221 2.651213
## -----
## Cluster: 6
##    18411     11796      8740      8085     11253
## 2.757309 2.704960 2.704805 2.689915 2.651787
## -----
## Cluster: 7
##    18712     8993     15581     17458     18417
## 9.054553 8.685034 8.669008 8.303776 8.273290

#Characteristic individuals - as many as clusters
para1<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[1]]))
para2<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[2]]))
para3<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[3]]))
para4<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[4]]))
para5<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[5]]))
para6<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[6]]))
para7<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[7]]))

dist1<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[1]]))
dist2<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[2]]))
dist3<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[3]]))
dist4<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[4]]))
dist5<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[5]]))
dist6<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[6]]))
dist7<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[7]]))

df$clust<-factor(res.hcpc$data.clust$clust)

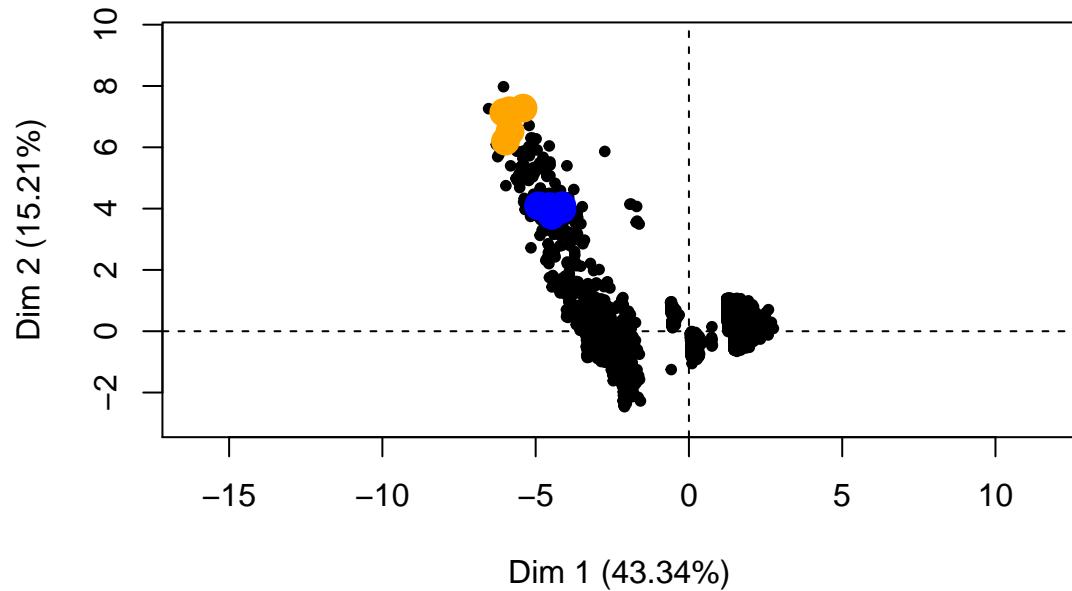
#cluster as variable suplementaria
res.pca<-PCA(df[,c("duration", "clust", vars_con)], quanti.sup=1, quali.sup=2, ncp=4, graph=FALSE)

#? habillage
#color the individuals among a categorical variable (give the number/name of the categorical variable)

plot.PCA(res.pca, label="none", invisible="quali", title="Characteristic individuals - Cluster 1")
#pintar "para" del cluster 1
points(res.pca$ind$coord[para1,1], res.pca$ind$coord[para1,2], col="blue", cex=2, pch=16)
#pintar "dist" del cluster 1
points(res.pca$ind$coord[dist1,1], res.pca$ind$coord[dist1,2], col="orange", cex=2, pch=16)

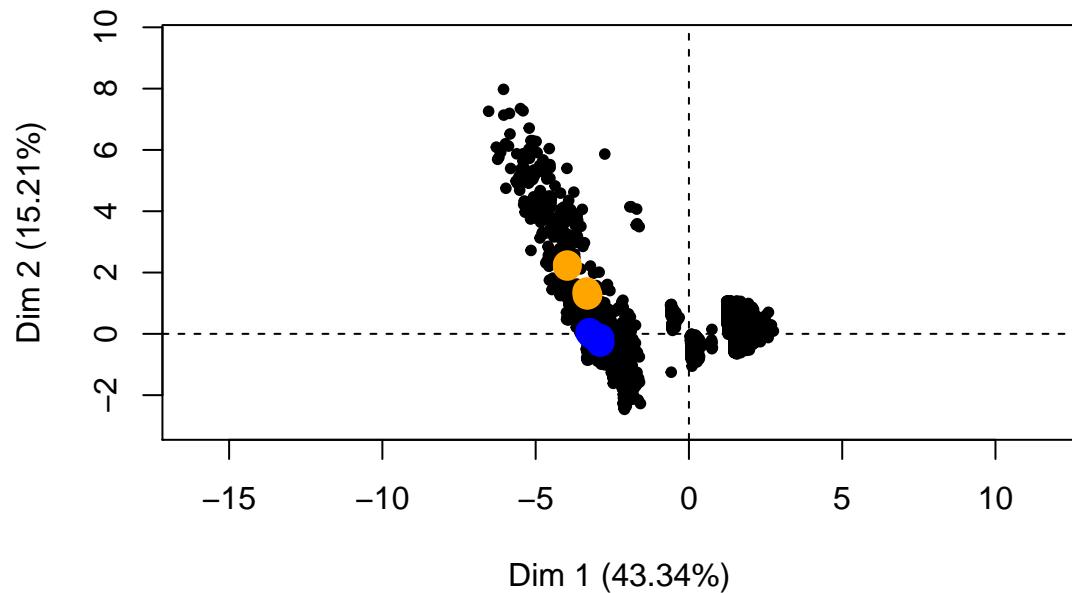
```

### Characteristic individuals – Cluster 1



```
plot.PCA(res.pca, label="none", invisible="quali", title="Characteristic individuals - Cluster 1")
points(res.pca$ind$coord[para2,1], res.pca$ind$coord[para2,2], col="blue", cex=2, pch=16)
points(res.pca$ind$coord[dist2,1], res.pca$ind$coord[dist2,2], col="orange", cex=2, pch=16)
```

### Characteristic individuals – Cluster 2

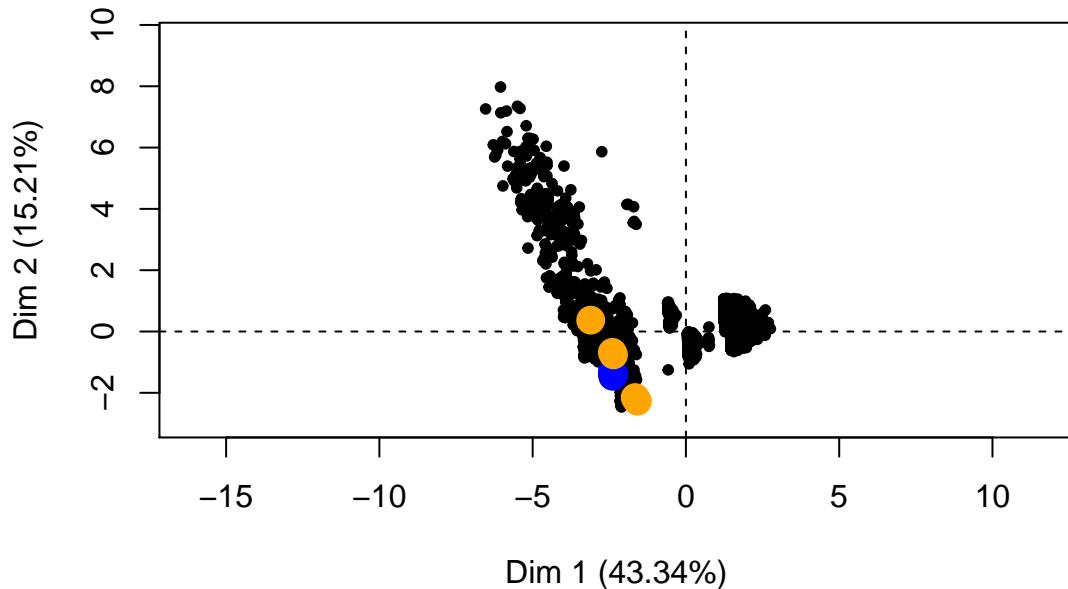


```

plot.PCA(res.pca, label="none", invisible="quali", title="Characteristic individuals - Cluster 3")
points(res.pca$ind$coord[para3,1], res.pca$ind$coord[para3,2], col="blue", cex=2, pch=16)
points(res.pca$ind$coord[dist3,1], res.pca$ind$coord[dist3,2], col="orange", cex=2, pch=16)

```

### Characteristic individuals – Cluster 3

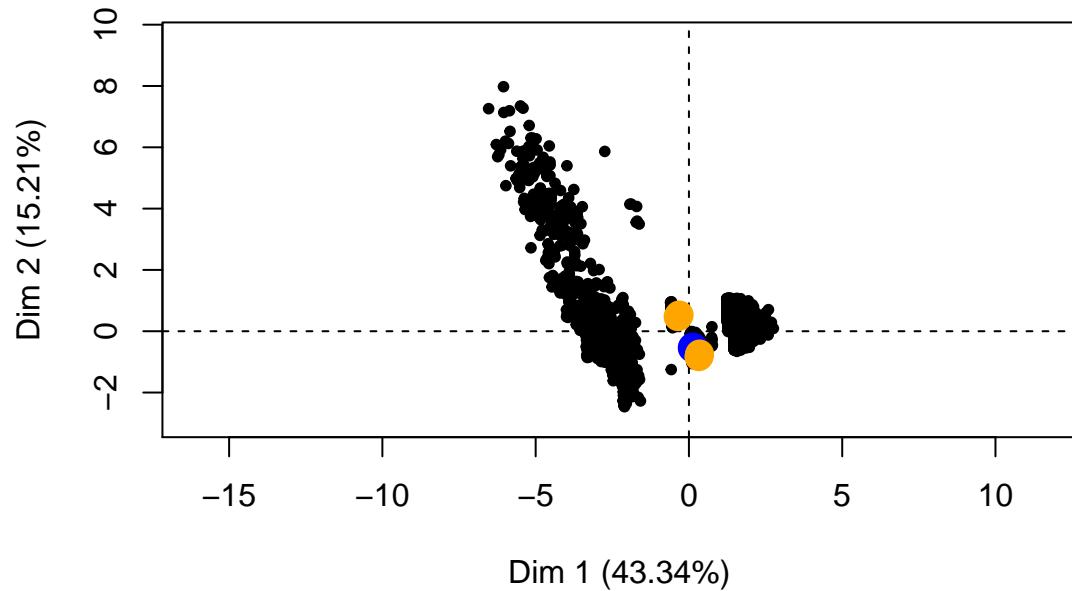


```

plot.PCA(res.pca, label="none", invisible="quali", title="Characteristic individuals - Cluster 4")
points(res.pca$ind$coord[para4,1], res.pca$ind$coord[para4,2], col="blue", cex=2, pch=16)
points(res.pca$ind$coord[dist4,1], res.pca$ind$coord[dist4,2], col="orange", cex=2, pch=16)

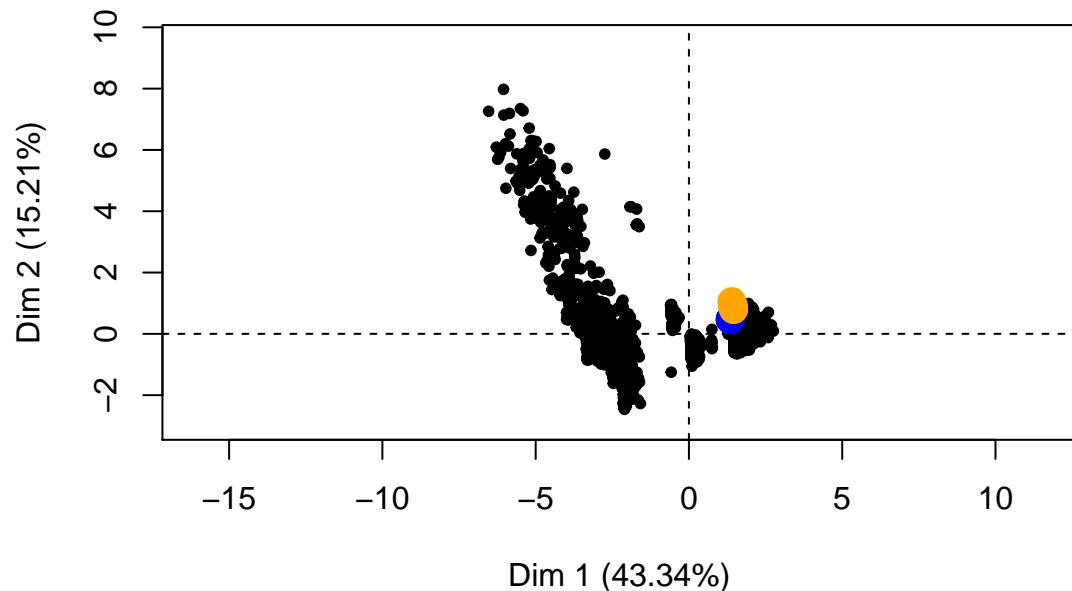
```

### Characteristic individuals – Cluster 4



```
plot.PCA(res.pca, label="none", invisible="quali", title="Characteristic individuals - Cluster 5")
points(res.pca$ind$coord[para5,1], res.pca$ind$coord[para5,2], col="blue", cex=2, pch=16)
points(res.pca$ind$coord[dist5,1], res.pca$ind$coord[dist5,2], col="orange", cex=2, pch=16)
```

### Characteristic individuals – Cluster 5

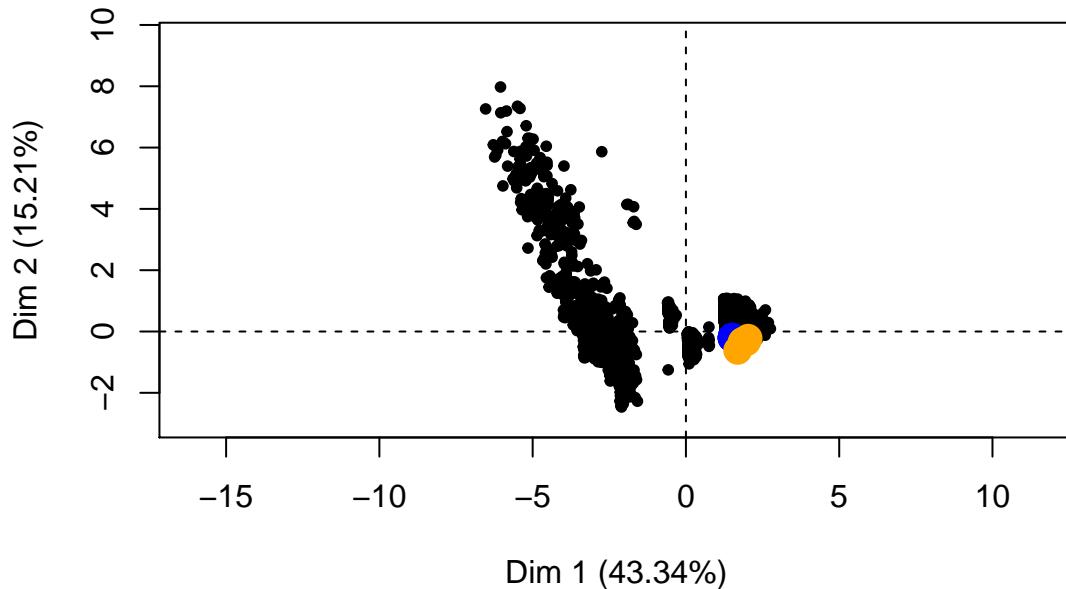


```

plot.PCA(res.pca, label="none", invisible="quali", title="Characteristic individuals - Cluster 6")
points(res.pca$ind$coord[para6,1], res.pca$ind$coord[para6,2], col="blue", cex=2, pch=16)
points(res.pca$ind$coord[dist6,1], res.pca$ind$coord[dist6,2], col="orange", cex=2, pch=16)

```

## Characteristic individuals – Cluster 6

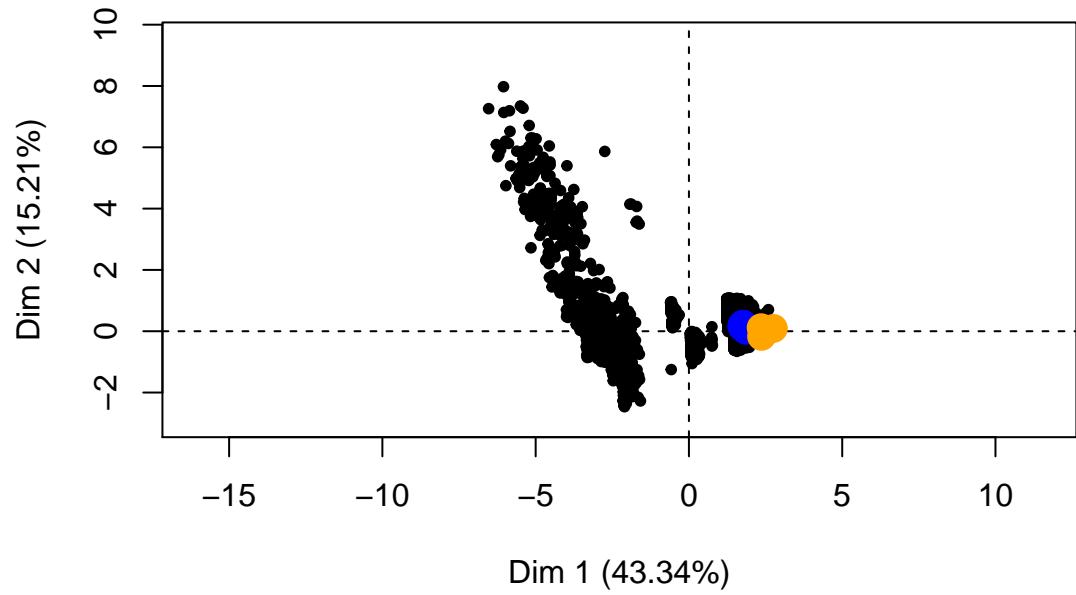


```

plot.PCA(res.pca, label="none", invisible="quali", title="Characteristic individuals - Cluster 7")
points(res.pca$ind$coord[para7,1], res.pca$ind$coord[para7,2], col="blue", cex=2, pch=16)
points(res.pca$ind$coord[dist7,1], res.pca$ind$coord[dist7,2], col="orange", cex=2, pch=16)

```

### Characteristic individuals – Cluster 7



# Course Practical Assignment - Final Delivery (Part 3)

*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:

### Load validated data from Deliverable 1:

```
#invisible() prevent hte output in console of the function
#dirwd<- "D:/Users/Usuari/Documents/ADEIpractica"
#dirwd<- "//pax/perfils/1173408.CR/Downloads/deliverable"
dirwd<- "D:/Documents/GitHub/ADEI"
setwd(dirwd)

load( paste0(dirwd, "/bank-additional/Bank5000_validated.RData") )
```

```
summary(df)
```

```
##      age          job          marital
## Min.   :18.00   job-admin.    :1246   marital-divorced: 554
## 1st Qu.:32.00   job-blue-collar:1171   marital-married  :3055
## Median :38.00   job-technician : 796   marital-single   :1377
## Mean   :40.07   job-services   : 498
## 3rd Qu.:47.00   job-management: 411
## Max.   :87.00   job-retired    : 205
## (Other)        : 659
##                  education          default
## education-basic.4y     : 533   default-no     :3954
## education-basic.6y     : 289   default-unknown:1032
## education-basic.9y     : 767
## education-high.school :1218
## education-professional.course: 615
## education-university.degree  :1564
##
##      housing          loan          contact
## housing-no :2261   loan-no :4217   contact-cellular :3122
## housing-yes:2725   loan-yes: 769   contact-telephone:1864
##
##      month          day_of_week          duration
## month-may:1741   day_of_week-1mon:1016   Min.   : 5.0
## month-jul: 829   day_of_week-2tue:1043   1st Qu.:101.0
## month-aug: 697   day_of_week-3wed: 971   Median :177.0
## month-jun: 652   day_of_week-4thu:1034   Mean   :250.6
## month-nov: 507   day_of_week-5fri: 922   3rd Qu.:316.0
## month-apr: 310
## (Other)   : 250   Max.   :1580.0
##      campaign          pdays          previous
## Min.   : 1.000   Min.   : 0.00   Min.   :0.0000
## 1st Qu.: 1.000   1st Qu.:19.00   1st Qu.:0.0000
## Median : 2.000   Median :19.00   Median :0.0000
## Mean   : 2.535   Mean   :18.53   Mean   :0.1598
## 3rd Qu.: 3.000   3rd Qu.:19.00   3rd Qu.:0.0000
## Max.   :25.000   Max.   :19.00   Max.   :4.0000
##
##      poutcome          emp.var.rate          cons.price.idx
## poutcome-failure     : 477   Min.   :-3.40000   Min.   :92.20
## poutcome-nonexistent:4353  1st Qu.:-1.80000   1st Qu.:93.08
## poutcome-success     : 156   Median : 1.10000   Median :93.75
##                               Mean   : 0.06446   Mean   :93.57
##                               3rd Qu.: 1.40000   3rd Qu.:93.99
##                               Max.   : 1.40000   Max.   :94.77
##
##      cons.conf.idx          euribor3m          nr.employed          y
## Min.   :-50.80   Min.   :0.635   Min.   :4964   y-no :4429
## 1st Qu.:-42.70   1st Qu.:1.334   1st Qu.:5099   y-yes: 557
## Median :-41.80   Median :4.857   Median :5191
```

```

##  Mean    :-40.43   Mean    :3.614   Mean    :5166
##  3rd Qu.:-36.40   3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :-26.90   Max.   :5.000   Max.   :5228
##
##      num_missings      num_outliers      num_errors
##  Min.   :0.0000   Min.   :0.00000   Min.   :0
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0
##  Median :0.0000   Median :0.00000   Median :0
##  Mean   :0.1111   Mean   :0.00361   Mean   :0
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0
##  Max.   :3.0000   Max.   :2.00000   Max.   :0
##
##          f.season      minutes           f.age
##  season-spring     :2117   Min.   : 0.08333   f.age-[18,32]:1352
##  season-summer     :2178   1st Qu.: 1.68333   f.age-(32,38]:1205
##  season-autumnwinter: 691   Median : 2.95000   f.age-(38,47]:1220
##                                         Mean   : 4.17703   f.age-(47,87]:1209
##                                         3rd Qu.: 5.26667
##                                         Max.   :26.33333
##
##          f.duration           f.campaign
##  f.duration-[5,101]   :1252   f.campaign-[0,2]  :3392
##  f.duration-(101,177]  :1243   f.campaign-(2,5]  :1181
##  f.duration-(177,316]  :1247   f.campaign-(5,25]: 413
##  f.duration-(316,1.58e+03] :1244
##
##      f.pdays           f.previous
##  f.pdays-sometime: 177   f.previous-never:4353
##  f.pdays-never   :4809   f.previous-some  : 633
##
##      f.emp.var.rate           f.cons.price.idx
##  f.emp.var.rate-[-Inf,0]  :2086   f.cons.price.idx-[92.2,93.1]:1409
##  f.emp.var.rate-(0, Inf]  :2900   f.cons.price.idx-(93.1,93.7]:1086
##                                         f.cons.price.idx-(93.7,94]  :1819
##                                         f.cons.price.idx-(94,94.8]  : 672
##
##      f.cons.conf.idx           f.euribor3m
##  f.cons.conf.idx-[-50.8,-42.7] :1856   f.euribor3m-[0.635,1.33]:1254
##  f.cons.conf.idx-(-42.7,-41.8]: 967   f.euribor3m-(1.33,4.86] :1466
##  f.cons.conf.idx-(-41.8,-36.4] :1231   f.euribor3m-(4.86,4.96] :1130
##  f.cons.conf.idx-(-36.4,-26.9]: 932   f.euribor3m-(4.96,5]    :1136
##
##      f.nr.employed
##  f.nr.employed-[4.96e+03,5.1e+03] :1639

```

```

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

## Linear Model Building - target numeric “duration” de la trucada

Per tal d'elaborar un model lineal que predigi el valor de la variable numerica target *duration*, primer hem de decidir quines son les variables que utilitzarem en la seva construccio. En altres paraules, trobar quines variables ens aporten informacio i precisio al model predictiu, pero sense sobreparametritzar-lo.

### Variables numeriques explicatives pel target numeric

#### Model inicial amb totes les variables numeriques

Una primera (i dolenta) aproximacio podria ser la d'usar un model lineal inicial que tingui en compte totes les variables numeriques aportades. Veiem com aquestes variables en un model lineal nomes ens expliquen l' 1.3% de la variabilitat de l'output *duration* (Multiple R-squared: 0.01309)! El que vol dir que gairebe un 99% d'questa variabilitat de la duracio queda sense explicar; aixi doncs aquesta primera aproximacio, a part d'estar sobreparametritzada, no prediu gens be.

```
vars_con
```

```

## [1] "age"           "duration"        "campaign"       "pdays"
## [5] "previous"      "emp.var.rate"    "cons.price.idx" "cons.conf.idx"
## [9] "euribor3m"     "nr.employed"

m1<-lm(duration~., data=df[,vars_con]) #es passa nomes el df amb les variables continues perque
#li hem posat un '.' al model que vol dir que les agafi totes, i ara nomes volem variables
#explicatives numeriques; si especificiquem les variables numeriques explicatives del model i volem
#que el calculi per tot el df incloent variables categoriques, "data" el passem com =df.
summary(m1)

##
## Call:
## lm(formula = duration ~ ., data = df[, vars_con])
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -363.78 -146.18 - 73.40  60.34 1336.54 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1509.2349  2363.9473  0.638   0.5232    
## age          -0.1649    0.3135 -0.526   0.5988    
## campaign     -6.9311    1.3298 -5.212 1.94e-07 *** 
## pdays         -3.2592    1.5930 -2.046   0.0408 *  
## previous     -20.6929   9.1633 -2.258   0.0240 *  
## emp.var.rate  25.7481   11.9837  2.149   0.0317 *  
## cons.price.idx 9.8281   14.1899  0.693   0.4886    
## cons.conf.idx -0.9407    1.1320 -0.831   0.4060    
## euribor3m    -13.2088   14.9069 -0.886   0.3756    
## nr.employed   -0.4030    0.2522 -1.598   0.1101    
## ---
```

```

## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.1 on 4976 degrees of freedom
## Multiple R-squared: 0.01309, Adjusted R-squared: 0.0113
## F-statistic: 7.333 on 9 and 4976 DF, p-value: 1.097e-10

```

### Model inicial amb noms les variables numeriques rellevants

Una altra opció mes adient seria la d'obtenir un model inicial utilitzant noms les variables numeriques que son rellevants, i a partir d'aquí mirar si es pot reduir la parametrització del model i seguir amb un bon ajust predictiu de la variabilitat del nostre output *duration*. Per a trobar les variables rellevants, podem realitzar tests de Fisher mitjançant la comanda Anova d'R, o bé utilitzar la comanda condes vista en anteriors entregues.

Inferential criteria o Bayesian info criteria Utilitzem la comanda Anova per a realitzar tests de Fisher i detectar i eliminar variables poc explicatives en els models. El test Anova ens diu línia a línia si cada variable es significativa a l'hora d'aportar informació en el model. Cada fila es refereix a un test de models encaixats del model m1 amb el model m1 sense la variable expressada en la fila. Per tant, si el p-valor es <0.05 podem refutar la H0 que deia que els models eren iguals. La podem refutar amb les variables age i indicadors socioeconomics, que vol dir que no ens aporten informació extra al model. En canvi, per les variables campaign, pdays, previous i emp.var.rate, no podem refutar la H0, el que vol dir que si que ens estan aportant informació al model i no les podem eliminar. Podriem contemplar també quedar-nos amb nr.employed, ja que esta prop de la frontera del p-valor teòric vs la flexibilitat a la pràctica. El model m2 es el model obtingut amb aquestes variables rellevants.

**#METODE TESTS FISHER:**

```

#remove non significant variables, per a saber quines son fem tests de Fisher amb la comanda Anova d'R
Anova(m1)

```

```

## Anova Table (Type II tests)
##
## Response: duration
##          Sum Sq Df F value    Pr(>F)
## age        14529  1  0.2768  0.59883
## campaign   1425963  1 27.1663 1.942e-07 ***
## pdays      219732  1  4.1861  0.04081 *
## previous    267683  1  5.0997  0.02397 *
## emp.var.rate 242319  1  4.6165  0.03171 *
## cons.price.idx 25180  1  0.4797  0.48858
## cons.conf.idx 36248  1  0.6906  0.40601
## euribor3m   41212  1  0.7851  0.37562
## nr.employed 134032  1  2.5535  0.11012
## Residuals  261191266 4976
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
m2<-lm(duration~campaign+pdays+previous+emp.var.rate+nr.employed, data=df)
Anova(m2)

```

```

## Anova Table (Type II tests)
##
## Response: duration
##          Sum Sq Df F value    Pr(>F)
## campaign   1354099  1 25.7778 3.970e-07 ***
## pdays       159994  1  3.0458  0.08101 .
## previous    242160  1  4.6100  0.03183 *
## emp.var.rate 894268  1 17.0240 3.751e-05 ***

```

```

## nr.employed    1275709      1 24.2855 8.576e-07 ***
## Residuals     261597980 4980
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Un altre metode pas a pas es el metode Akaike. Va eliminant en models successius les variables que treient-les, obtindriem un AIC mes baix, ja que ens interessa un coeficient d'Akaike que quan no li treiem cap variable es mantingui igual. Un trade-off entre el fitting del model i la sobre parametritzacio. El model m3  $duration \sim campaign + pdays + previous + emp.var.rate + cons.conf.idx + nr.employed$  es el model obtingut amb aquest metode.

#### #AKAIKE:

```
m3<-step(m1)
```

Hi ha un altre metode anomenat Bayesian, el qual es millor per a mostres grans com la nostra, ja que tot i funcionar com l'anterior, solem sortir models mes simplificats. El model m4  $duration \sim campaign + emp.var.rate + nr.employed$  es el model obtingut amb aquest metode.

#### #BAYESIAN (BIC):

```
m4<-step(m1, k=log(nrow(df)))
```

```

## Start:  AIC=54264.89
## duration ~ age + campaign + pdays + previous + emp.var.rate +
##           cons.price.idx + cons.conf.idx + euribor3m + nr.employed
##
##              Df Sum of Sq      RSS      AIC
## - age             1   14529 261205795 54257
## - cons.price.idx  1   25180 261216446 54257
## - cons.conf.idx   1   36248 261227514 54257
## - euribor3m       1   41212 261232478 54257
## - nr.employed     1   134032 261325298 54259
## - pdays            1   219732 261410997 54261
## - emp.var.rate    1   242319 261433584 54261
## - previous         1   267683 261458948 54261
## <none>                  261191266 54265
## - campaign         1   1425963 262617229 54284
##
## Step:  AIC=54256.65
## duration ~ campaign + pdays + previous + emp.var.rate + cons.price.idx +
##           cons.conf.idx + euribor3m + nr.employed
##
##              Df Sum of Sq      RSS      AIC
## - cons.price.idx  1   25242 261231037 54249
## - cons.conf.idx   1   39406 261245201 54249
## - euribor3m       1   42216 261248011 54249
## - nr.employed     1   131927 261337722 54251
## - pdays            1   218978 261424773 54252
## - emp.var.rate    1   243151 261448946 54253
## - previous         1   270229 261476024 54253
## <none>                  261205795 54257
## - campaign         1   1426321 262632115 54275
##
## Step:  AIC=54248.62
## duration ~ campaign + pdays + previous + emp.var.rate + cons.conf.idx +
##           euribor3m + nr.employed
##

```

```

##          Df Sum of Sq      RSS      AIC
## - euribor3m   1    25603 261256640 54241
## - cons.conf.idx 1    103360 261334397 54242
## - pdays       1    230438 261461474 54244
## - previous     1    275181 261506218 54245
## - nr.employed  1    438597 261669634 54248
## <none>                   261231037 54249
## - emp.var.rate  1    447006 261678042 54249
## - campaign      1    1413437 262644474 54267
##
## Step: AIC=54240.59
## duration ~ campaign + pdays + previous + emp.var.rate + cons.conf.idx +
##           nr.employed
##
##          Df Sum of Sq      RSS      AIC
## - pdays       1    226621 261483260 54236
## - previous     1    271870 261528510 54237
## - cons.conf.idx 1    341340 261597980 54239
## <none>                   261256640 54241
## - emp.var.rate  1    1109140 262365780 54253
## - campaign      1    1392303 262648943 54259
## - nr.employed   1    1458307 262714947 54260
##
## Step: AIC=54236.4
## duration ~ campaign + previous + emp.var.rate + cons.conf.idx +
##           nr.employed
##
##          Df Sum of Sq      RSS      AIC
## - previous     1    106252 261589512 54230
## - cons.conf.idx 1    274714 261757974 54233
## <none>                   261483260 54236
## - emp.var.rate  1    1158906 262642166 54250
## - campaign      1    1389083 262872343 54254
## - nr.employed   1    1566560 263049820 54258
##
## Step: AIC=54229.91
## duration ~ campaign + emp.var.rate + cons.conf.idx + nr.employed
##
##          Df Sum of Sq      RSS      AIC
## - cons.conf.idx 1    281879 261871391 54227
## <none>                   261589512 54230
## - emp.var.rate  1    1125903 262715415 54243
## - campaign      1    1382960 262972472 54248
## - nr.employed   1    1460309 263049821 54249
##
## Step: AIC=54226.77
## duration ~ campaign + emp.var.rate + nr.employed
##
##          Df Sum of Sq      RSS      AIC
## <none>                   261871391 54227
## - emp.var.rate  1    927321 262798712 54236
## - nr.employed   1    1277524 263148915 54243
## - campaign      1    1348382 263219773 54244

```

```

summary(m4)

##
## Call:
## lm(formula = duration ~ campaign + emp.var.rate + nr.employed,
##      data = df[, vars_con])
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -342.28 -147.72  -73.63   62.68 1318.28
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3091.9562   572.8442   5.398 7.07e-08 ***
## campaign     -6.7107    1.3250  -5.065 4.24e-07 ***
## emp.var.rate   21.1243    5.0293   4.200 2.71e-05 ***
## nr.employed   -0.5469    0.1109  -4.930 8.49e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.3 on 4982 degrees of freedom
## Multiple R-squared:  0.01052, Adjusted R-squared:  0.009924
## F-statistic: 17.66 on 3 and 4982 DF, p-value: 2.131e-11

```

Conedes per a obtenir les variables numeriques rellevants Utilitzant la comanda condes vista en anteriors entregues tambe podem trobar les variables que son rellevants pel nostre model. El model que obtenim així es el m5, i la comanda Anova ens diu que pdays no esta aportant res de nou (p-value=0.38). La treiem i ens queda el model m6, que correspon a *duration ~ euribor3m + nr.employed + campaign*. Si apliquem un metode BIC en aquest model obtenim un model m7 amb nomes un sol parametre aportant informacio que es campaign. Aquest model es massa simple i no ens va be per a treballar, així que ens quedem amb el model m4 obtingut anteriorment tambe pel metode Bayesian.

```

condes(df, 11)
#variable target: 11 (duration)

#Agafem com a variables explicatives les $quanti del condes:
m5<-lm(duration~pdays+euribor3m+nr.employed+campaign, data=df); Anova(m5)
m6<-lm(duration~euribor3m+nr.employed+campaign, data=df); Anova(m6)
#totes les variables ens aporten informacio nova

#BIC
m7<-step(m6, k=log(nrow(df)))

```

La comanda vif d'R ens diu les variables utilitzades en el model tenen redundancies. Si el seu valor esta per sota de 3 es valid; i si dos valors son iguals vol dir que d'aquelles dues variables ens n'hem de quedar nomes una! En el cas del model m4 ens quedarem amb la variable campaign i nr.employed (triada d'entre les dues amb valor similar), i obtindrem així el model m44.

```

vif(m4)

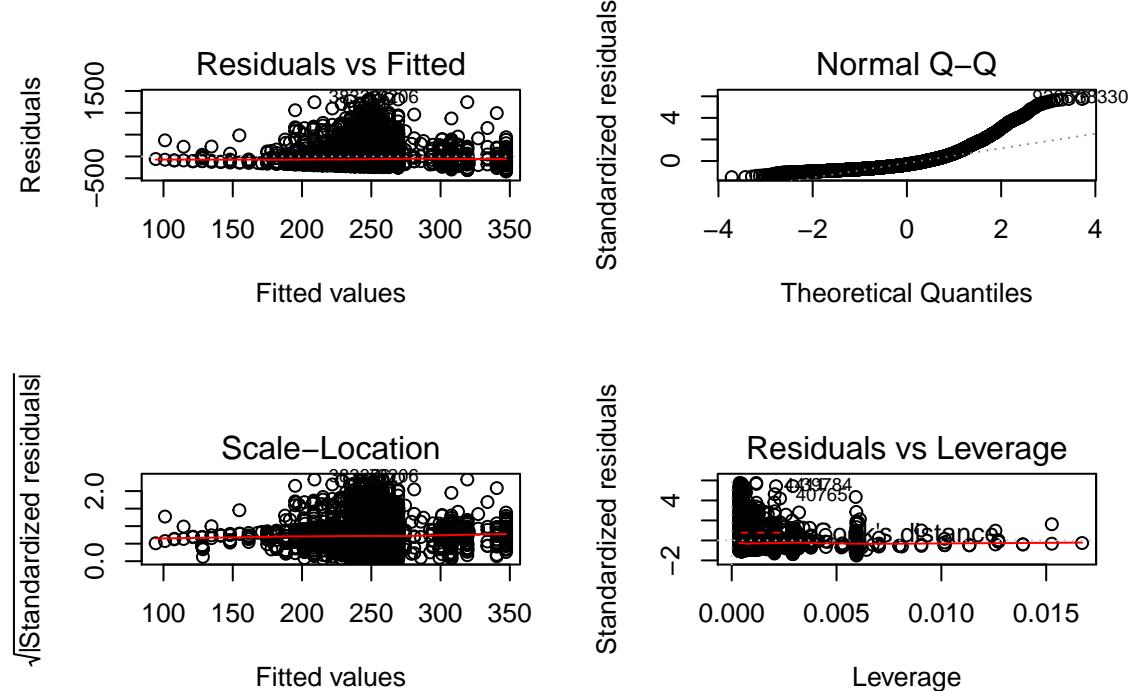
##      campaign emp.var.rate  nr.employed
##      1.024848     6.027929     6.013152
m44<-lm(duration~campaign+nr.employed, data=df)

```

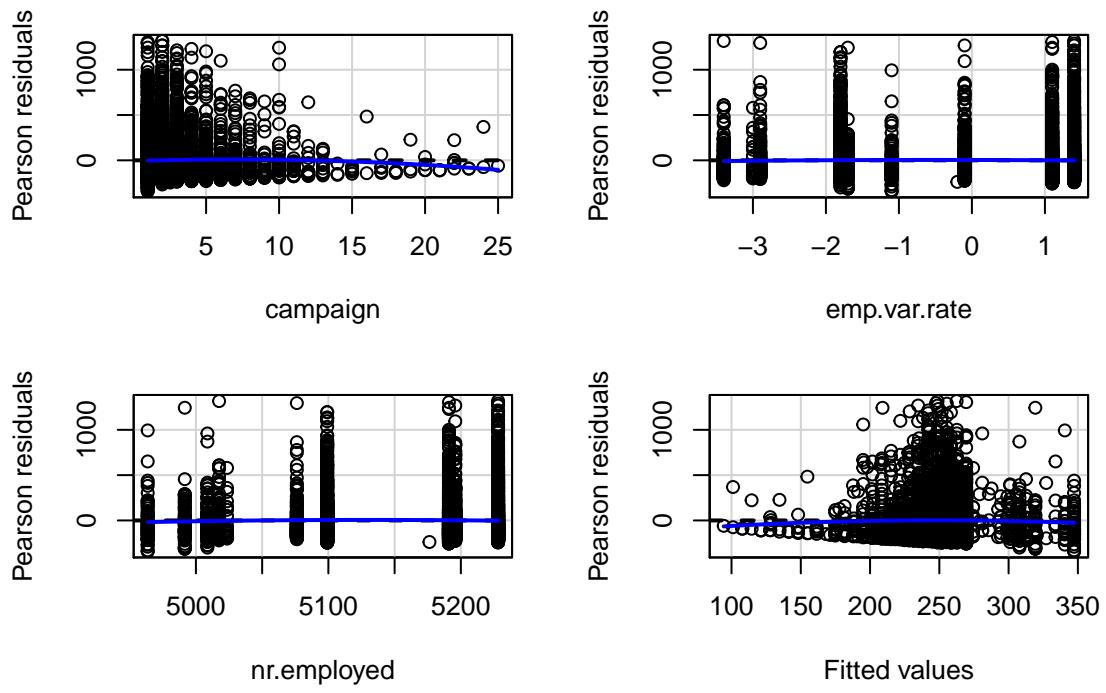
## Plots del model m4

Podem veure en el diagnostic del model que no es gens bo, a continuació doncs, farem un seguit de transformacions i hi afegirem variables factor explicatives. D'aquesta manera arribarem al nostre model definitiu, el qual diagnosticarem mes en profunditat.

```
par(mfrow=c(2,2))
plot(m4) #models forca dolents
```



```
par(mfrow=c(1,1))
residualPlots(m4)
```

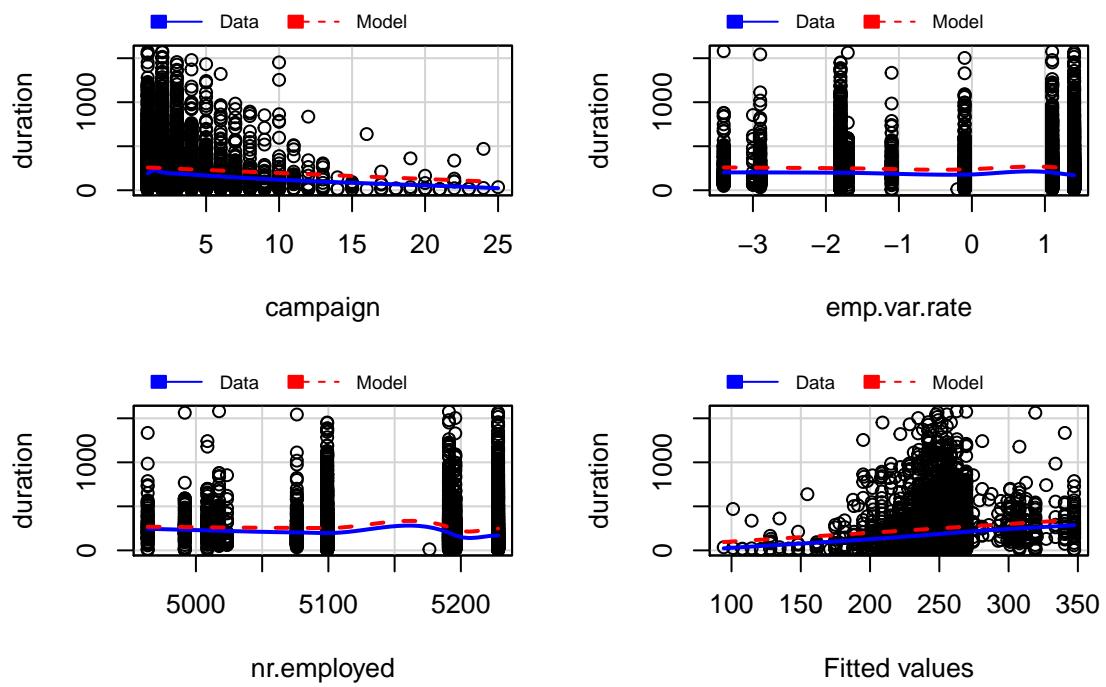


```

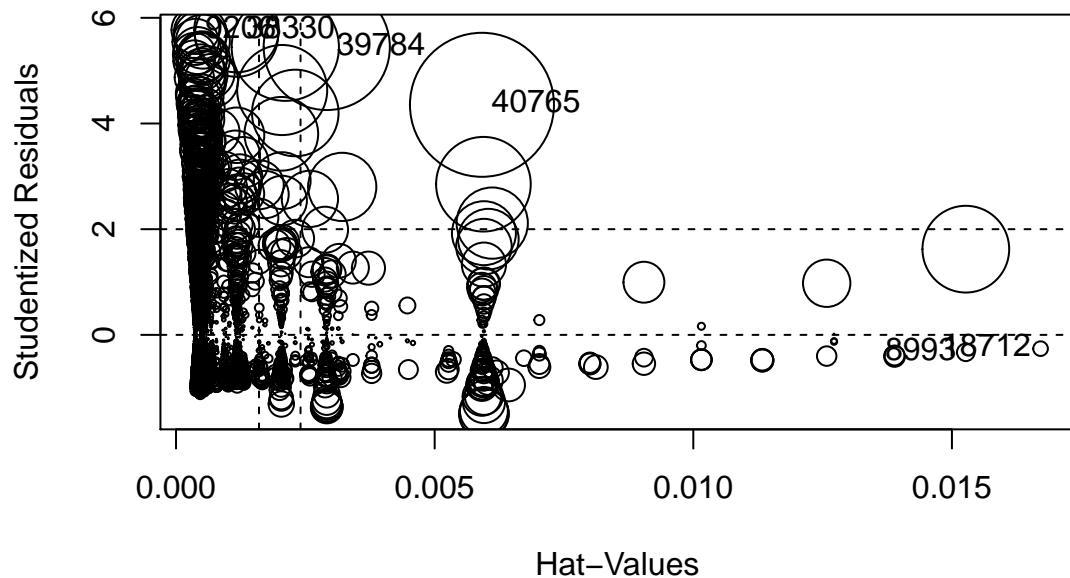
##           Test stat Pr(>|Test stat|)
## campaign      -1.8909    0.05870 .
## emp.var.rate   -0.7423    0.45794
## nr.employed    -1.6329    0.10255
## Tukey test     -2.4163    0.01568 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
marginalModelPlots(m4)

```

## Marginal Model Plots



```
influencePlot(m4)
```

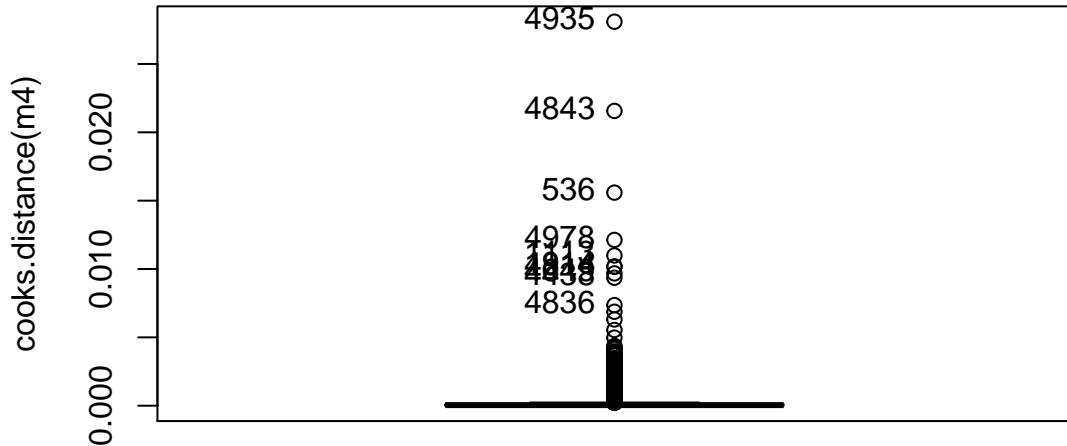


```
##          StudRes      Hat      CookD
## 8993   -0.3300087 0.0152684634 0.0004222271
## 9206    5.7696545 0.0003776632 0.0031239403
```

```

## 18712 -0.2611508 0.0167134662 0.0002898615
## 38330  5.7684911 0.0011697218 0.0096794548
## 39784  5.4482303 0.0029140821 0.0215638797
## 40765  4.3537452 0.0059141513 0.0280913224
Boxplot(cooks.distance(m4))

```



```

## [1] 4935 4843 536 4978 1113 1914 4814 4649 4438 4836

```

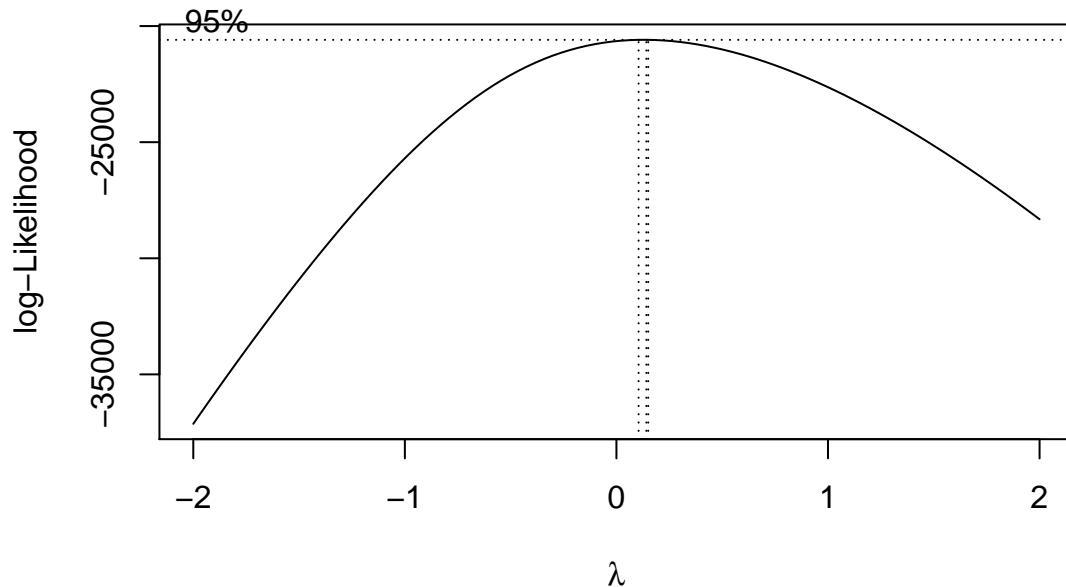
### Transformacio de la variable target numerica

A vegades una transformacio de la variable target numerica pot millorar el model. La comanda Bo-Cox ens mostra com el valor de lambda estimat es proper a 0, vol dir que hem d'elevar a 0 el target, com que això no es pot fer i amb les grafiques anteriors no em pogut indentificar cap patró que ens induís a elevar alguna de les variables, la transformacio estadistica del nostre target sera el logaritme, i la farem a partir del model m4 obtingut anteriorment (així donarem marge a possibles reduccions del mateix). A través de la grafica Normal Q-Q, podem observar com el nou model s'ajusta molt mes que l'anterior a una distribució normal i en podem identificar unes cues amb tendència inferior respecte a la línia de la normal. Per altra banda, també podem observar com en "Residuals vs Leverage" la majoria dels punts es concentra a la part esquerra de la grafica, ens indica que no tenim valors influents, cap dels valors es troba més enllaçat als marges del leverage (les línies no es dibuixen).

```

#Box-Cox
boxcox(m4, data=df)

```



```

#TRANSFORMACIO LOGARITMICA Y(m4) -> logY
m8<-lm(log(duration)~campaign+emp.var.rate+nr.employed, data=df); Anova(m8)

## Anova Table (Type II tests)
##
## Response: log(duration)
##             Sum Sq   Df F value    Pr(>F)
## campaign      139.3     1 175.543 < 2.2e-16 ***
## emp.var.rate   21.8     1  27.484 1.650e-07 ***
## nr.employed    31.0     1  39.109 4.343e-10 ***
## Residuals   3953.3 4982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#BIC
m10<-step(m8, k=log(nrow(df)))

## Start:  AIC=-1123.1
## log(duration) ~ campaign + emp.var.rate + nr.employed
##
##             Df Sum of Sq   RSS      AIC
## <none>            3953.3 -1123.10
## - emp.var.rate  1     21.809 3975.1 -1104.18
## - nr.employed   1     31.034 3984.3 -1092.62
## - campaign       1    139.296 4092.6  -958.95

vif(m10)

##      campaign emp.var.rate  nr.employed
## 1.024848     6.027929     6.013152

```

```

#emp.var.rate i nr.employed mostren molta colinearitat, ens quedem amb emp.var.rate
#per a ser una variable mes entenedora
m11<-lm(log(duration)~emp.var.rate+campaign, data=df)
vif(m11)

## emp.var.rate      campaign
##       1.024654    1.024654

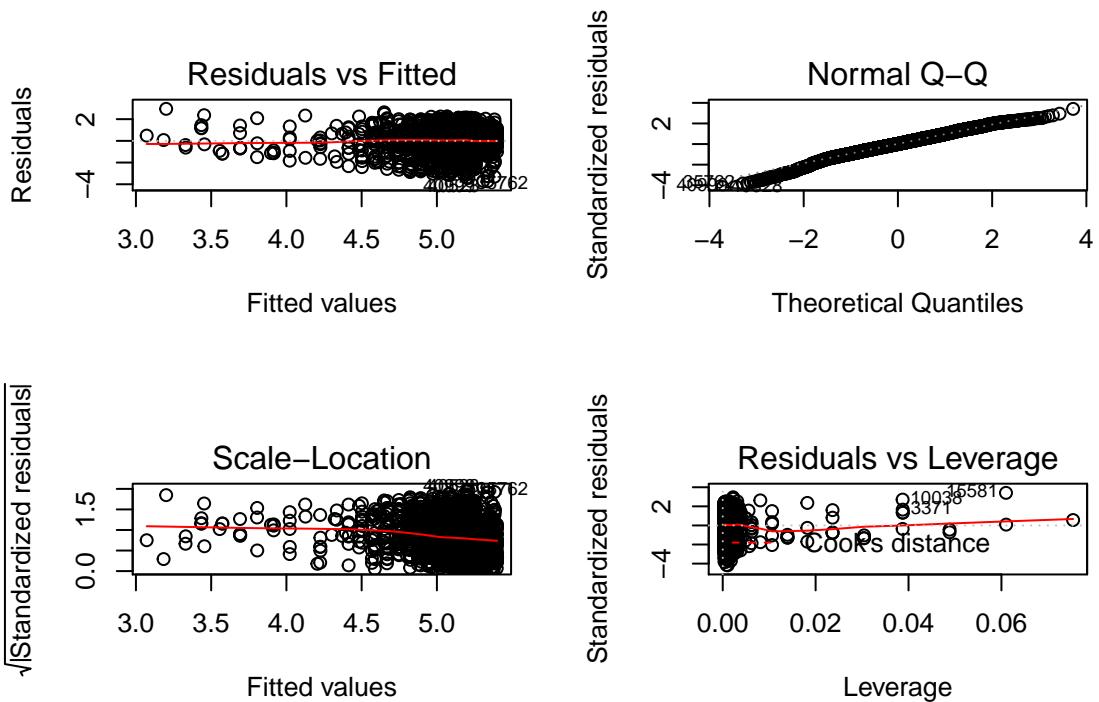
summary(m11)

##
## Call:
## lm(formula = log(duration) ~ emp.var.rate + campaign, data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -3.6643 -0.5519  0.0113  0.5914  2.6497 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 5.327250  0.018169 293.205 <2e-16 ***
## emp.var.rate -0.008891  0.008087 -1.099   0.272    
## campaign     -0.068651  0.005167 -13.286 <2e-16 ***  
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8942 on 4983 degrees of freedom
## Multiple R-squared:  0.03612,    Adjusted R-squared:  0.03574 
## F-statistic: 93.38 on 2 and 4983 DF,  p-value: < 2.2e-16

#POLINOMIC REGRESSION
#com hi ha poques variables provem de transformar-les totes amb regressio polinomica
m20<-lm(log(duration)~poly(euribor3m, 2)+poly(campaign, 2), data=df)
summary(m20)
Anova(m20)
#veiem com amb ambdos termes quadratics(2) tenen un p-value <0.05.
#Es millor que el terme lineal(1) en el cas d l'euribor3m, podriem fer per tant aquesta
#transformacio quadratica.

par(mfrow=c(2,2))
plot(m20)

```



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

### Variables discretes explicatives pel target numeric

Mitjançant la comanda condes intentarem trobar variables discretes que estiguin relacionades amb la variable target numèrica duration. D'aquesta manera sabrem quines variables discretes podem utilitzar en el model predictiu per a que ens aportin informació. A partir del millor model m11 anterior de variables continues, hem d'obtenir un nou model afegint les variables discretes i factoritzades (que no estiguin ja en el model de forma numèrica). En el nostre cas agafem campaign i nr.employed com a variables continues, i afegim f.cons.conf.idx+f.cons.price.idx+month+f.euribor3m+poutcome com a variables discretes. Com que el condes anterior ens ha donat com a variables factor significatives algunes que ja tenim en el model com a continues, hem de triar o una o altra versió. Per a saber si agafar una variable com a continua o factoritzada, hem de fer el següent i veiem com en ambdues variables obtenim que es millor usar la seva versió numèrica (no factoritzada).

```
condes(df[, c("duration", vars_dis)], 1, proba=0.01)
```

```
#a partir del millor model anterior (m11) amb variables continues afegim factors
m60<-lm(log(duration)~campaign+nr.employed+f.cons.conf.idx+f.cons.price.idx+month
+f.euribor3m+poutcome, data=df)
summary(m60)
```

```
##
## Call:
## lm(formula = log(duration) ~ campaign + nr.employed + f.cons.conf.idx +
##     f.cons.price.idx + month + f.euribor3m + poutcome, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -3.7849 -0.5447  0.0007  0.5763  2.5903
```

```

##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)           14.1263162  4.8821924  2.893
## campaign            -0.0685567  0.0051782 -13.240
## nr.employed         -0.0017450  0.0009734 -1.793
## f.cons.conf.idxf.cons.conf.idx-(-42.7,-41.8] -0.2215250  0.1385032 -1.599
## f.cons.conf.idxf.cons.conf.idx-(-41.8,-36.4] -0.1054326  0.0924315 -1.141
## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9]  0.0841657  0.0844295  0.997
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.1778539  0.0932539  1.907
## f.cons.price.idxf.cons.price.idx-(93.7,94]   0.2395880  0.1201263  1.994
## f.cons.price.idxf.cons.price.idx-(94,94.8]   0.1626503  0.0996614  1.632
## monthmonth-aug      -0.1611314  0.1410615 -1.142
## monthmonth-dec      -0.2665829  0.2010504 -1.326
## monthmonth-jul      -0.0104232  0.1309917 -0.080
## monthmonth-jun      0.0601159  0.1626995  0.369
## monthmonth-mar      -0.1319787  0.1316415 -1.003
## monthmonth-may      0.0062834  0.1106291  0.057
## monthmonth-nov      -0.0381862  0.1182949 -0.323
## monthmonth-oct      -0.2155702  0.1281603 -1.682
## monthmonth-sep      -0.1174620  0.1507313 -0.779
## f.euribor3mf.euribor3m-(1.33,4.86]  0.1413905  0.1056155  1.339
## f.euribor3mf.euribor3m-(4.86,4.96]  0.1738018  0.1201916  1.446
## f.euribor3mf.euribor3m-(4.96,5]    0.0598234  0.1297290  0.461
## poutcomepoutcome-nonexistent  0.0512680  0.0476555  1.076
## poutcomepoutcome-success   0.2738577  0.0862378  3.176
## Pr(>|t|)
## (Intercept)          0.00383 **
## campaign             < 2e-16 ***
## nr.employed          0.07308 .
## f.cons.conf.idxf.cons.conf.idx-(-42.7,-41.8] 0.10979
## f.cons.conf.idxf.cons.conf.idx-(-41.8,-36.4]  0.25407
## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9]  0.31887
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.05655 .
## f.cons.price.idxf.cons.price.idx-(93.7,94]   0.04616 *
## f.cons.price.idxf.cons.price.idx-(94,94.8]   0.10274
## monthmonth-aug       0.25339
## monthmonth-dec      0.18492
## monthmonth-jul      0.93658
## monthmonth-jun      0.71178
## monthmonth-mar      0.31612
## monthmonth-may      0.95471
## monthmonth-nov      0.74686
## monthmonth-oct      0.09262 .
## monthmonth-sep      0.43585
## f.euribor3mf.euribor3m-(1.33,4.86]  0.18072
## f.euribor3mf.euribor3m-(4.86,4.96]  0.14823
## f.euribor3mf.euribor3m-(4.96,5]    0.64472
## poutcomepoutcome-nonexistent  0.28207
## poutcomepoutcome-success   0.00150 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8878 on 4963 degrees of freedom

```

```

## Multiple R-squared:  0.05375,   Adjusted R-squared:  0.04955
## F-statistic: 12.81 on 22 and 4963 DF,  p-value: < 2.2e-16
Anova(m60)

## Anova Table (Type II tests)
##
## Response: log(duration)
##             Sum Sq Df  F value    Pr(>F)
## campaign        138.1  1 175.2846 < 2.2e-16 ***
## nr.employed     2.5  1   3.2138  0.073079 .
## f.cons.conf.idx 4.2  3   1.7649  0.151593
## f.cons.price.idx 5.7  3   2.4208  0.064119 .
## month           5.8  9   0.8134  0.603791
## f.euribor3m      4.9  3   2.0732  0.101547
## poutcome         8.0  2   5.0541  0.006416 **
## Residuals      3911.5 4963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#per campaign, mateix model pero amb f.campaign en lloc de campaign:
maux<-lm(log(duration)~f.campaign+nr.employed+f.cons.conf.idx+f.cons.price.idx+month
+f.euribor3m+poutcome, data=df)

#com que m60 i maux no son anidats, no els podem comparar amb un test de Fisher --> BIC
BIC(m60, maux)
#choose option with minimum BIC -> better
#      df      BIC
# m60 24 13143.82 --> millor model amb campaign com a numerica
# maux 25 13199.11

maux<-lm(log(duration)~campaign+f.nr.employed+f.cons.conf.idx+f.cons.price.idx+month
+f.euribor3m+poutcome, data=df)
BIC(m60, maux)
#      df      BIC
# m60 24 13143.82 --> millor model amb nr.employed com a numerica
# maux 25 13151.61

#si haguessim hagut de fer el mateix amb pdays, ojo!!!! pq la continua ha estat majoritariament
#imputada, per tant en aquest cas, tot i el test, hauriem d'agafar la factoritzada!

```

Mirem si podem simplificar el model eliminant variables poc significatives mitjancant la comanda step i veiem com ens podem quedar amb les variables numeriques nr.employed i campaign, i la variable factor f.cons.price.idx. Si ho fem mitjancant la comanda Anova, veiem com ens surt el mateix resultat pero agafant també la variable discreta poutcome. Així doncs també li farem cas i aquest model sera el m62, que explica el 5% ( $R^2=0.04959$ ) de la variabilitat.

```

m61<-step(m60, k=log(nrow(df)))

m62<-lm(log(duration)~campaign+nr.employed+f.cons.price.idx+poutcome, data=df); summary(m62)

```

## Interaccions

Partint del model anterior, li afegim interaccions 2 a 2 entre totes les seves variables, simplifiquem i veiem com hi ha dues interaccions significatives: campaign:nr.employed i campaign:f.cons.price.idx. En el nostre model nomes podem tenir en compte interaccions entre dos factors o entre un factor i una variable numerica,

així que amb els tests d'Anova mirarem manualment quines interaccions ens queden, el que ens porta a un model m73, que explica el 5.5% ( $R^2=0.05534$ ) de la variabilitat de l'output del logaritme de *duration*.

```
#interacció entre 2 variables:
m70<-lm(log(duration)~(campaign+nr.employed+f.cons.price.idx+poutcome)^2, data=df)
summary(m70)
#coef(m70)
invisible(
  m71<-step(m70, k=log(nrow(df)))
)#el criteri Anova(Fisher) reafirma el step(BIC) en aquest cas!
# log(duration) ~ campaign+nr.employed+f.cons.price.idx+campaign:nr.employed+campaign:f.cons.price.idx
#
#                               Df Sum of Sq    RSS      AIC
# <none>                      3907.1 -1130.7
# - campaign:nr.employed      1     13.192 3920.3 -1122.4
# - campaign:f.cons.price.idx 3     30.057 3937.1 -1118.0
invisible(
  Anova(m71)
)
anova(m71, m70) #Pr(>F) = 0.03967 * --> els models no són equivalents

Anova(m70)
#                                Pr(>F)
# campaign                         < 2.2e-16 ***
# nr.employed                      0.012590 *
# f.cons.price.idx                 1.162e-08 ***
# poutcome                          0.003345 **
# campaign:nr.employed             0.001721 ** --> entre dos numériques
# campaign:f.cons.price.idx       4.635e-07 *** --> entre numérica i factor --> AGAFEM
# campaign:poutcome                0.873389   --> entre numérica i factor
# nr.employed:f.cons.price.idx    0.058763 . --> entre numérica i factor
# nr.employed:poutcome             0.309191   --> entre numérica i factor
# f.cons.price.idx:poutcome       0.121019   --> entre factors --> AGAFEM aquest per l'entrega

m73<-lm(log(duration)~campaign+nr.employed+f.cons.price.idx+
          poutcome+campaign:f.cons.price.idx+f.cons.price.idx:poutcome, data=df)
anova(m73, m70) #p-value 0.003286 ** -> models no són equivalents -> H0 rejected -> m73 es millor

## Analysis of Variance Table
##
## Model 1: log(duration) ~ campaign + nr.employed + f.cons.price.idx + poutcome +
##           campaign:f.cons.price.idx + f.cons.price.idx:poutcome
## Model 2: log(duration) ~ (campaign + nr.employed + f.cons.price.idx +
##           poutcome)^2
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1    4969 3904.9
## 2    4961 3886.8  8     18.099 2.8877 0.003286 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
summary(m73)

##
## Call:
## lm(formula = log(duration) ~ campaign + nr.employed + f.cons.price.idx +
##       poutcome + campaign:f.cons.price.idx + f.cons.price.idx:poutcome,
```

```

##      data = df)
##
## Residuals:
##      Min    1Q Median    3Q   Max
## -3.8187 -0.5570  0.0047  0.5735  2.6169
##
## Coefficients:
##                               Estimate
## (Intercept)                9.4669014
## campaign            -0.1222066
## nr.employed        -0.0008058
## f.cons.price.idxf.cons.price.idx-(93.1,93.7] -0.1204673
## f.cons.price.idxf.cons.price.idx-(93.7,94]  0.5608898
## f.cons.price.idxf.cons.price.idx-(94,94.8]  0.0523764
## poutcomepoutcome-nonexistent  0.0875966
## poutcomepoutcome-success    0.1580875
## campaign:f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.0414029
## campaign:f.cons.price.idxf.cons.price.idx-(93.7,94]  0.0737345
## campaign:f.cons.price.idxf.cons.price.idx-(94,94.8]  0.0531081
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent  0.0103004
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent -0.5762804
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent -0.1861021
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success  0.2510572
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success -0.3243955
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success  0.2715234
##
##                               Std. Error
## (Intercept)                1.6422347
## campaign            0.0145243
## nr.employed        0.0003230
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.1100012
## f.cons.price.idxf.cons.price.idx-(93.7,94]  0.3416381
## f.cons.price.idxf.cons.price.idx-(94,94.8]  0.1619483
## poutcomepoutcome-nonexistent  0.0561413
## poutcomepoutcome-success    0.1096234
## campaign:f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.0184993
## campaign:f.cons.price.idxf.cons.price.idx-(93.7,94]  0.0163059
## campaign:f.cons.price.idxf.cons.price.idx-(94,94.8]  0.0184387
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent  0.1106143
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent  0.3461481
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent  0.1753390
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success  0.2305015
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success  0.4506457
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success  0.2301328
##
##                               t value
## (Intercept)                5.765
## campaign            -8.414
## nr.employed        -2.495
## f.cons.price.idxf.cons.price.idx-(93.1,93.7] -1.095
## f.cons.price.idxf.cons.price.idx-(93.7,94]  1.642
## f.cons.price.idxf.cons.price.idx-(94,94.8]  0.323
## poutcomepoutcome-nonexistent  1.560
## poutcomepoutcome-success    1.442
## campaign:f.cons.price.idxf.cons.price.idx-(93.1,93.7]  2.238
## campaign:f.cons.price.idxf.cons.price.idx-(93.7,94]  4.522

```

```

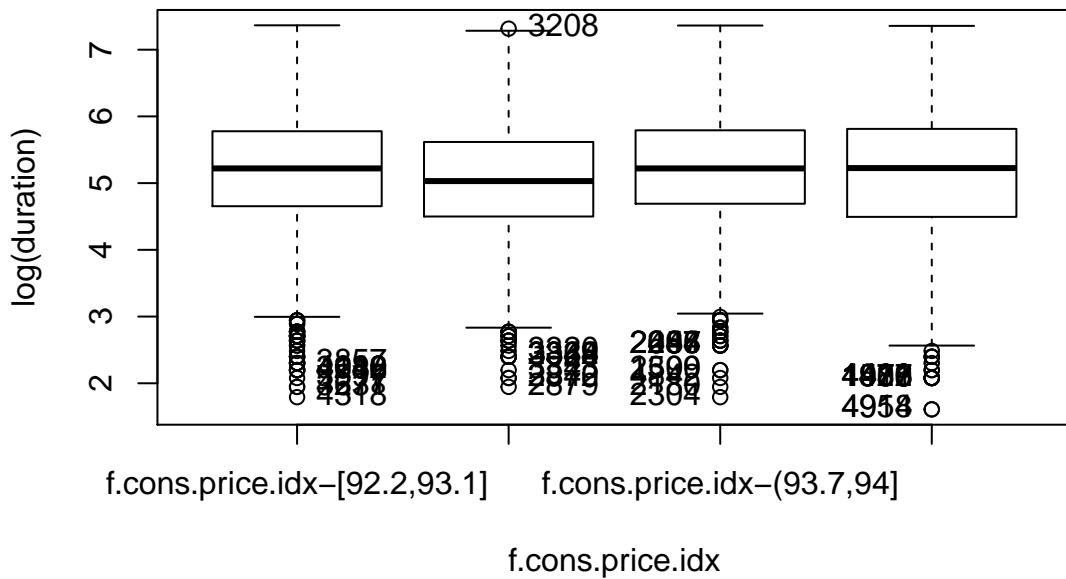
## campaign:f.cons.price.idxf.cons.price.idx-(94,94.8]                      2.880
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent   0.093
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent    -1.665
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent    -1.061
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success      1.089
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success       -0.720
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success        1.180
##
## Pr(>|t|)
## (Intercept)                                8.68e-09
## campaign                                     < 2e-16
## nr.employed                                  0.01263
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]          0.27351
## f.cons.price.idxf.cons.price.idx-(93.7,94]          0.10070
## f.cons.price.idxf.cons.price.idx-(94,94.8]          0.74640
## poutcomepoutcome-nonexistent                 0.11876
## poutcomepoutcome-success                     0.14934
## campaign:f.cons.price.idxf.cons.price.idx-(93.1,93.7]          0.02526
## campaign:f.cons.price.idxf.cons.price.idx-(93.7,94]          6.27e-06
## campaign:f.cons.price.idxf.cons.price.idx-(94,94.8]          0.00399
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent 0.92581
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent  0.09601
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent  0.28857
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success   0.27613
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success   0.47165
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success   0.23811
##
## (Intercept)                                ***
## campaign                                     ***
## nr.employed                                  *
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]          *
## f.cons.price.idxf.cons.price.idx-(93.7,94]          ***
## f.cons.price.idxf.cons.price.idx-(94,94.8]          **
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent .
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent .
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent .
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8865 on 4969 degrees of freedom
## Multiple R-squared:  0.05534,   Adjusted R-squared:  0.0523
## F-statistic: 18.19 on 16 and 4969 DF,  p-value: < 2.2e-16

```

### Interaction between a couple of factors in our model m73

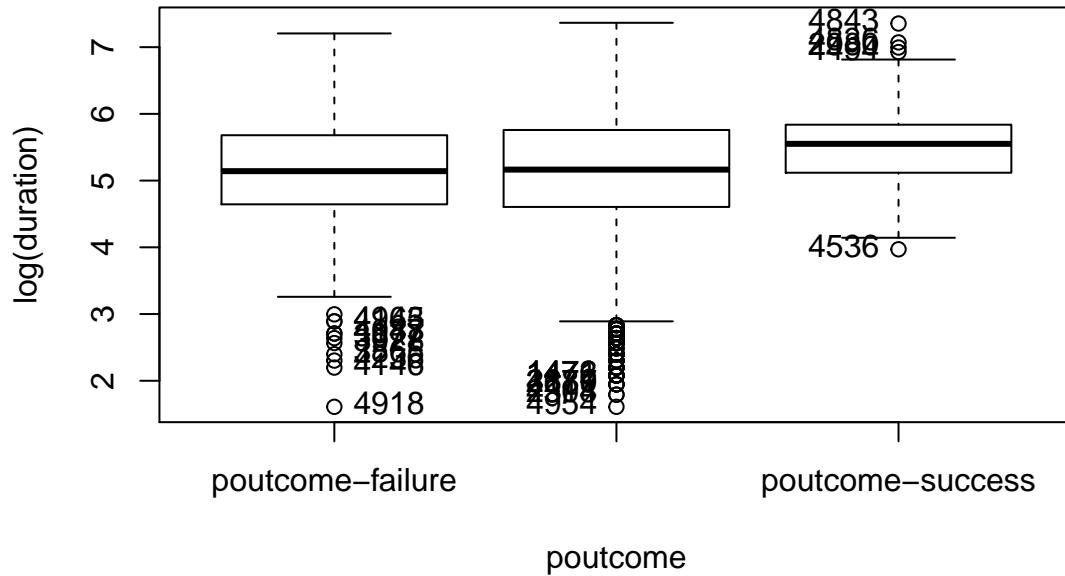
El model escollit m73 considera una interacció entre dos factors *f.cons.price.idx:poutcome*.

```
scatterplot(log(duration)~f.cons.price.idx, data=df)
```



```
## [1] "4318" "3671" "4237" "3597" "3682" "4088" "4146" "4210" "4280" "3857"  
## [11] "2879" "2870" "3342" "3343" "2868" "3329" "3332" "3344" "3345" "3320"  
## [21] "3208" "2304" "2180" "4842" "1599" "2300" "487" "666" "2048" "2254"  
## [31] "2297" "4918" "4954" "1471" "1472" "1476" "1486" "1489" "1507" "4916"  
## [41] "4928"
```

```
scatterplot(log(duration)~poutcome, data=df)
```

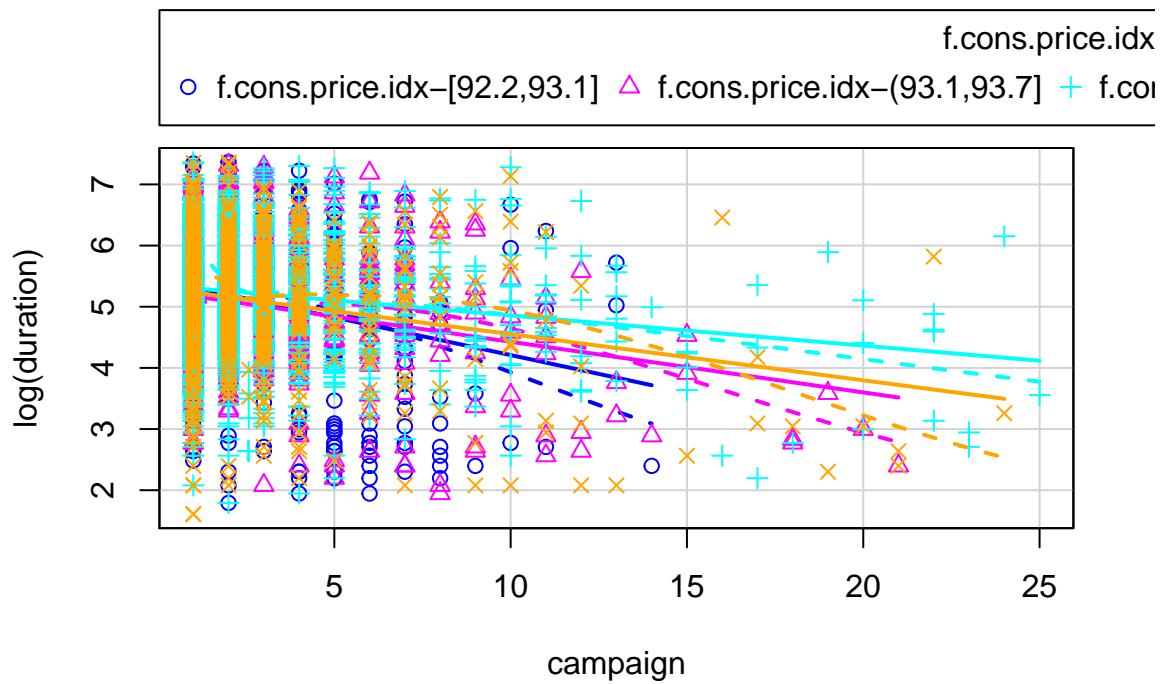


```
## [1] "4918" "4146" "4236" "3865" "3928" "3877" "4083" "4147" "4143" "4965"
## [11] "4954" "2304" "4318" "2180" "2879" "3671" "4237" "1471" "1472" "1476"
## [21] "4536" "2980" "4494" "4836" "4843"
```

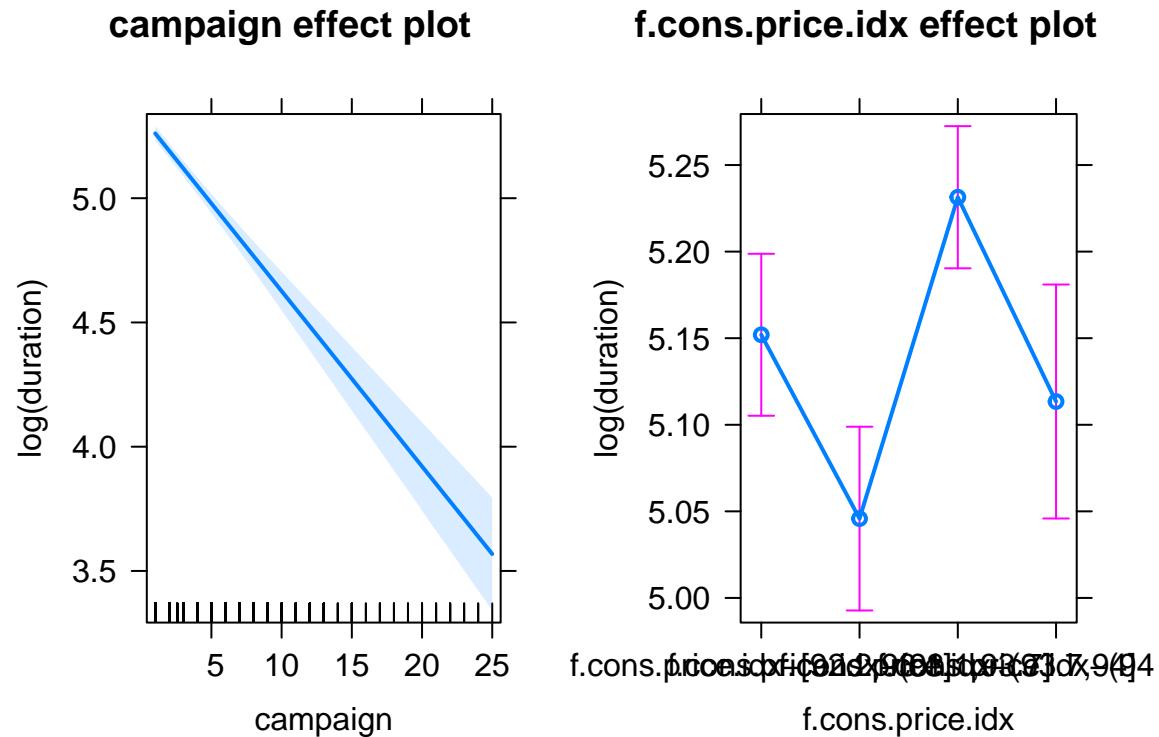
### Interaction between a factor and a covariate in our model m73

El model escollit m73 també considera una interacció entre un factor i una variable numèrica *campaign:f.cons.price*.

```
#model petit sense interaccions
m85<-lm(log(duration)~campaign+f.cons.price.idx, data=df)
scatterplot(log(duration)~campaign|f.cons.price.idx, data=df) #Suport visual
```



```
plot(allEffects(m85)) #effects library
```



```
#model gran amb interaccions: 3 parametres-> campaign, f.cons.price.idx, campaign:f.cons.price.idx
m855<-lm(log(duration)~campaign*f.cons.price.idx, data=df)
#are interactions significant?
```

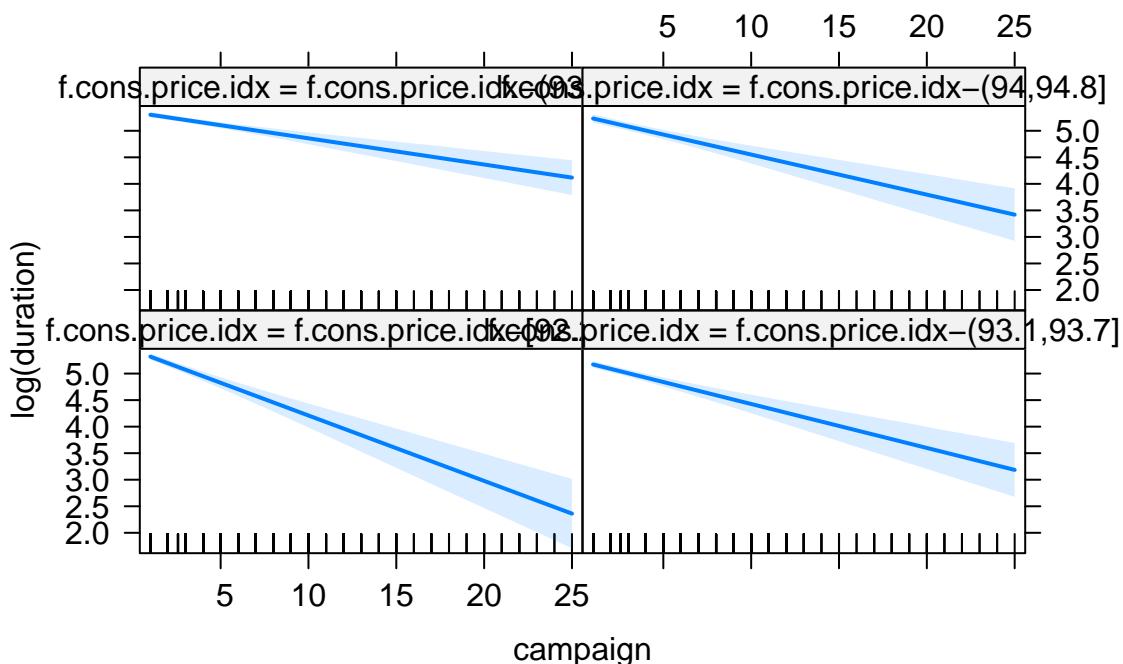
```

anova(m855, m855) #Pr(>F) 5.152e-05 *** --> H0 rejected --> m855 amb la interaccio es millor

## Analysis of Variance Table
##
## Model 1: log(duration) ~ campaign + f.cons.price.idx
## Model 2: log(duration) ~ campaign * f.cons.price.idx
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1    4981 3960.6
## 2    4978 3942.8  3     17.853 7.5136 5.152e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
plot(allEffects(m855))

```

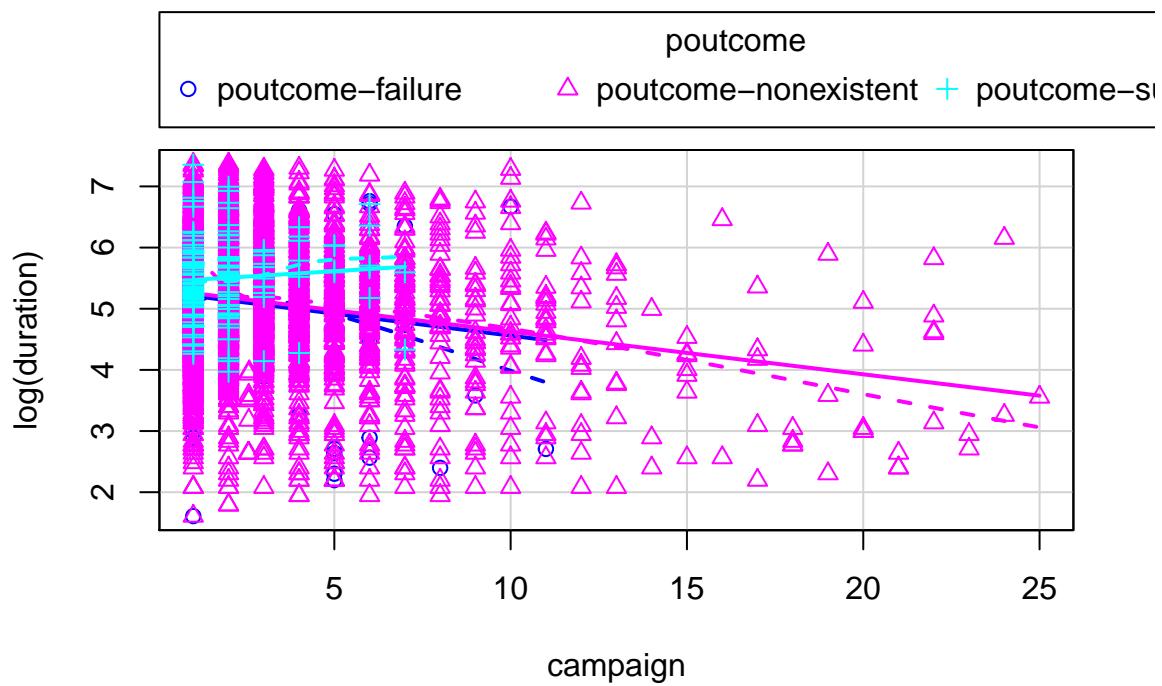
### campaign\*f.cons.price.idx effect plot



```

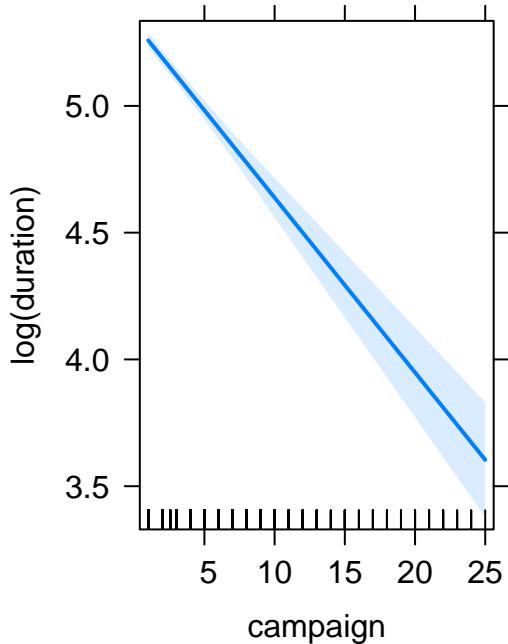
#campaign:poutcome --> segons el test anterior, la interaccio no ha sortit gaire
#significativa i no s'inclou en el model definitiu, pero s'interpreta millor visualment.
m86<-lm(log(duration)~campaign+poutcome, data=df)
scatterplot(log(duration)~campaign|poutcome, data=df)

```

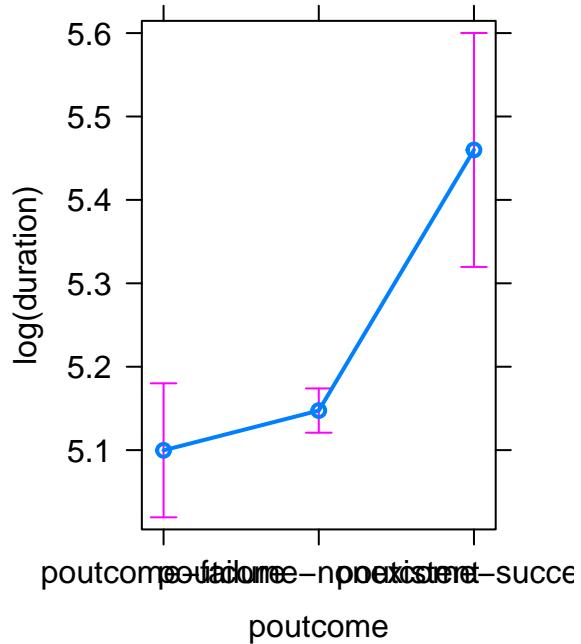


```
plot(allEffects(m86))
```

**campaign effect plot**



**poutcome effect plot**



```
m866<-lm(log(duration)~campaign*poutcome, data=df)
```

```
anova(m86, m866) #Pr(>F) 0.1435--> H0 accepted --> els models son iguals, per tant no cal
```

```

## Analysis of Variance Table
##
## Model 1: log(duration) ~ campaign + poutcome
## Model 2: log(duration) ~ campaign * poutcome
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     4982 3969.1
## 2     4980 3966.1  2      3.0931 1.9419 0.1435
#el model gran amb interaccions

```

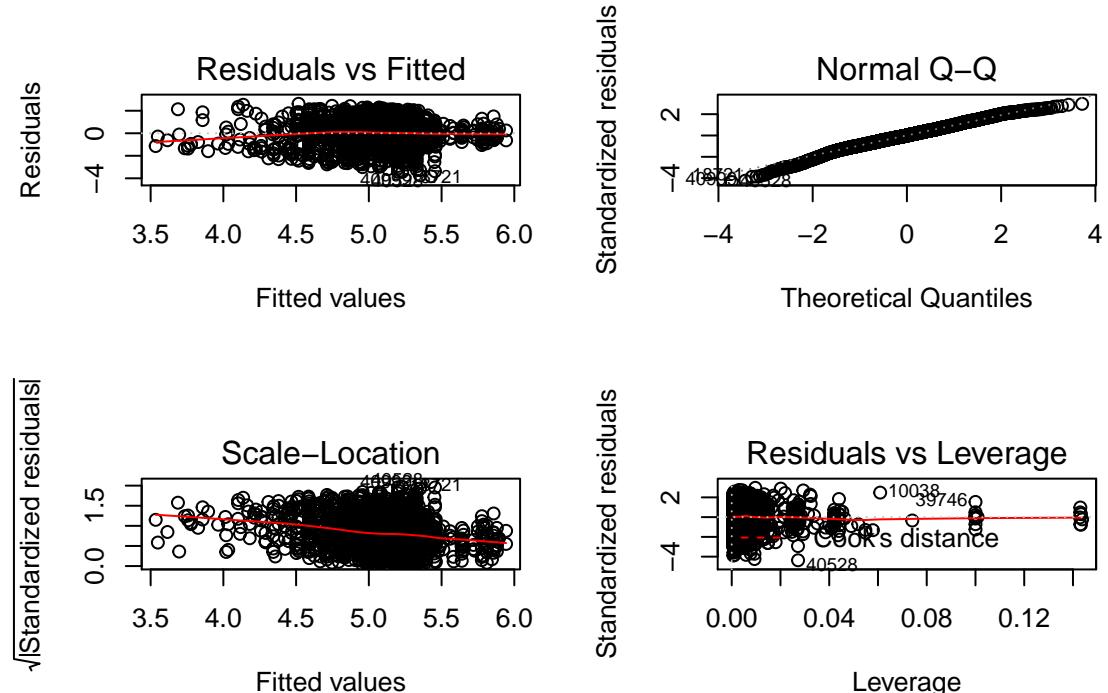
## Diagnostics del model definitiu m73

Residual vs Fitted En aquesta grafica podem veure els valors dels residus en l'eix Y i els valors de duracio en l'eix X. El nuvol de punts s'hauria de trobar encaixat dins un marge delimitat per dues línies paral·leles reconeixibles a la vista, fet que no podem veure en el nostre grafic i ens indica la incorrectesa del model. A més, també podem apreciar com els residus no es troben distribuïts de manera aleatoria, sino que es concentren en la part central de la grafica de manera molt significativa. De totes maneres, ha millorat respecte els models inicials. Normal Q-Q Aquesta grafica ens mostra els residus envers als seus valors esperats en el suposat d'una distribució normal. Podem dir que els residus no segueixen ben una distribució normal donat que s'allunyen significativament de la recta de punts en els seus extrems. Scale-Location L'objectiu d'aquesta grafica es visualitzar la primera grafica però amb valors de residus estandarditzats per tal de poder demostrar-ne homoscedasticitat. Es pot apreciar que no n'hi ha, es a dir, que la variància dels errors en les diferents observacions no és constant. Residual vs Leverage Aquesta grafica ens ajuda a identificar "influent data" en els nostres residus basant-se en la distància de Cook. En la nostra grafica no s'hi aprecia cap observació realment influent en el model amb gran distància de Cook, però si que s'observen forces observacions amb gran residu. Tot i així, no podem extreure grans conclusions mes enllà de dir que el model no es gaire bo.

```

par(mfrow=c(2,2))
plot(m73)

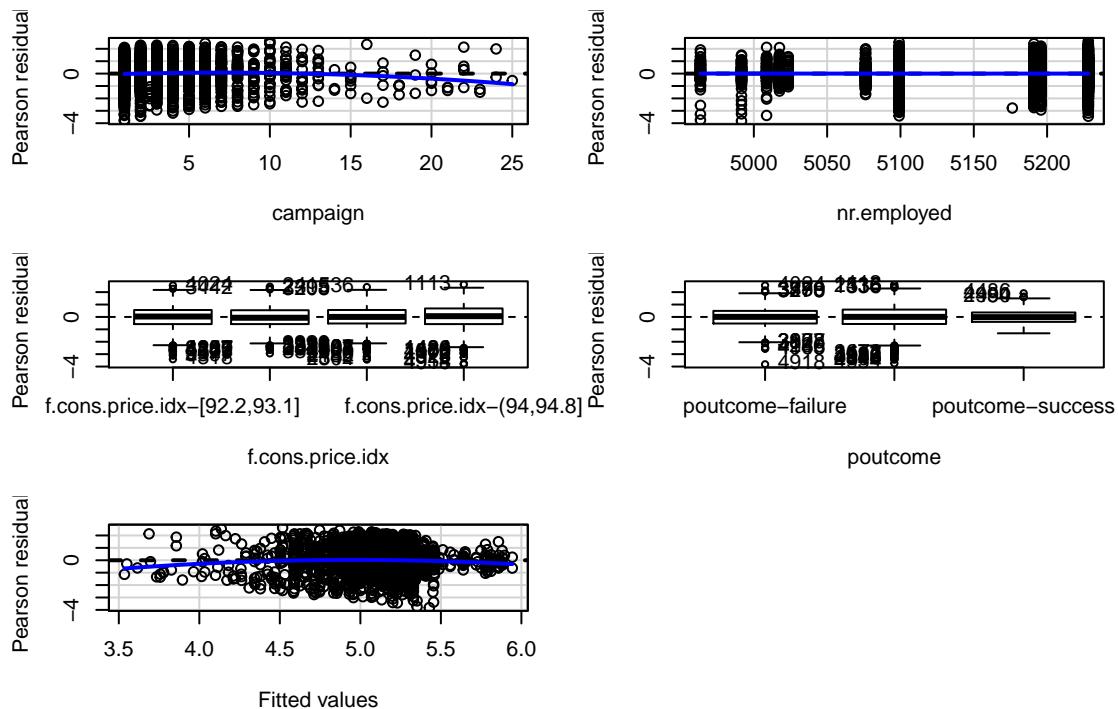
```



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

ResidualPlots Aquesta comanda ens mostra els residus de Pearson per a cada variable explicativa del model. En cap d'elles podem indentificar un patro que ens suggereixi un canvi o transformacio de la variable en el nostre model. marginalModelPlots Mitjançant un plot dels valors de log(duration) de les nostres dades, als qual hi afegeix un smoother blau, podem veure com el nostre model, que esta representat amb un de vermell, no esta predint gaire acuradament (tret del cas nr.employed). Això es veu perque no es sobreposen ambdues línies de suavitzacio, aixi doncs, el model obtingut no es gaire apropiat, sobretot en valors de la part esquerra de la grafica. influencePlot Aquesta comanda crea un grafic de bombolles amb el valor dels residus "Studentized", on el diametre de les bombolles mante una relacio proporcional a la seva distancia de Cook de l'individu (es veu tambe en el boxplot). Amb l'ajuda de la taula que s'ens mostra podem identificar com les mostres 10038 i 40528 son les mes influents, les que tenen una distancia de Cook mes elevada, es a dir, que treient-les el model canviaria més. Les observacions 39584 i 39699 son les que tenen mes leverage (hat value), es a dir, que cauen mes fora de la majoria de valors predicts del model. Tot i que les observacions amb mes leverage poden tenir mes habilitat per a afectar al model, en aquest cas no ho fan, ja que han han caigut en el patro del model i tenen una distancia de Cook baixa, es a dir, no son influents. Per ultim, també tenim algunes observacions amb un residu força elevat, com son les 10038, 40528 i 40999.

```
residualPlots(m73)
```

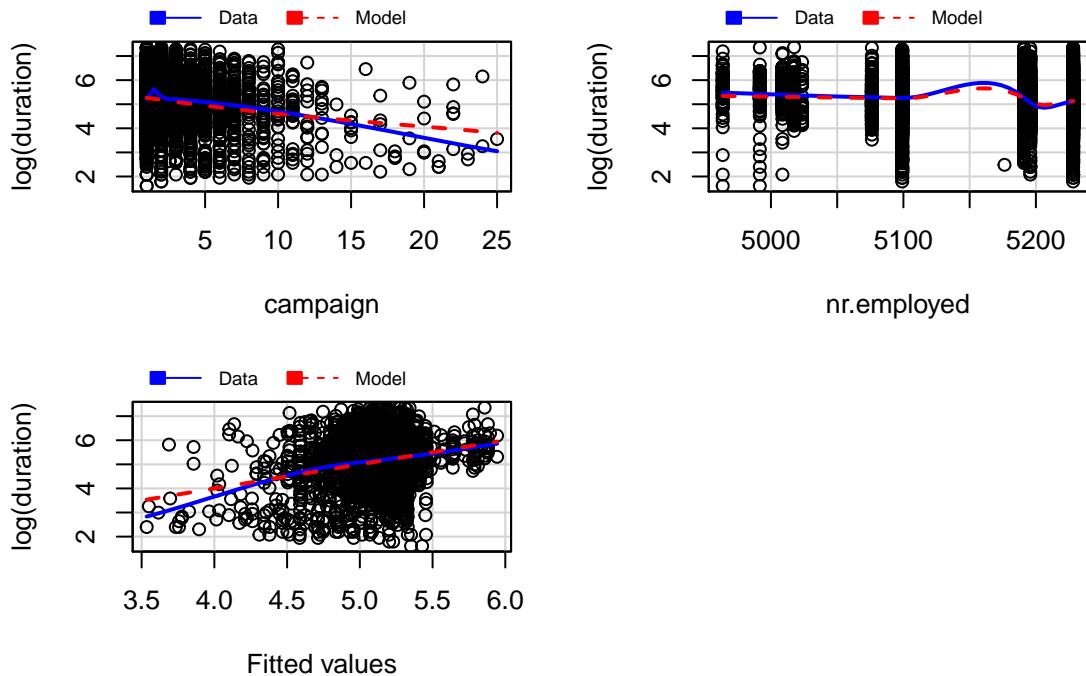


```
##           Test stat Pr(>|Test stat|)  
## campaign      -4.2245    2.438e-05 ***  
## nr.employed   0.1803     0.8569  
## f.cons.price.idx  
## poutcome  
## Tukey test     -5.1074    3.266e-07 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

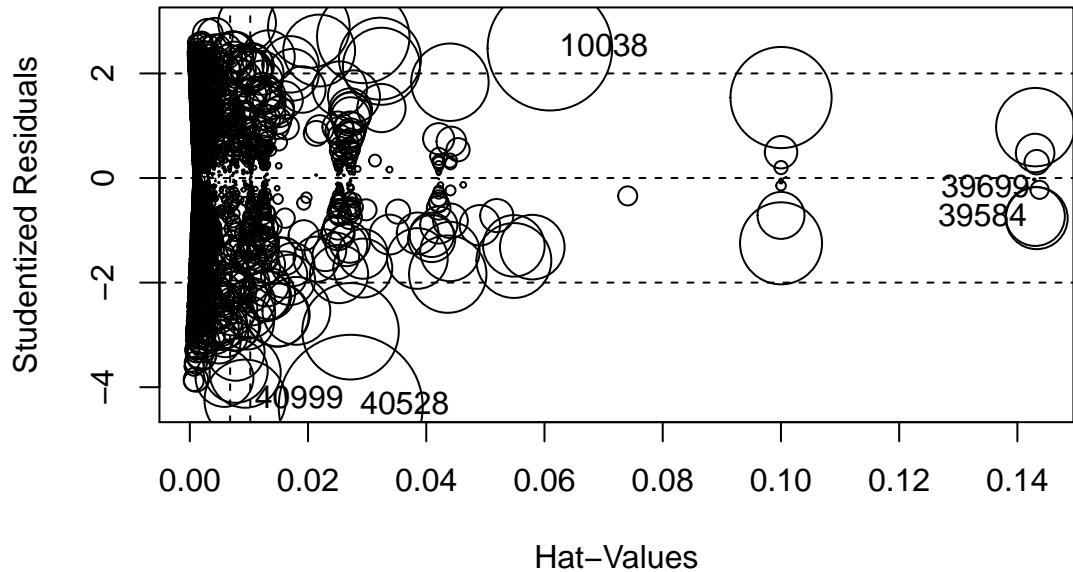
```
marginalModelPlots(m73)
```

```
## Warning in mmpls(...): Interactions and/or factors skipped
```

### Marginal Model Plots

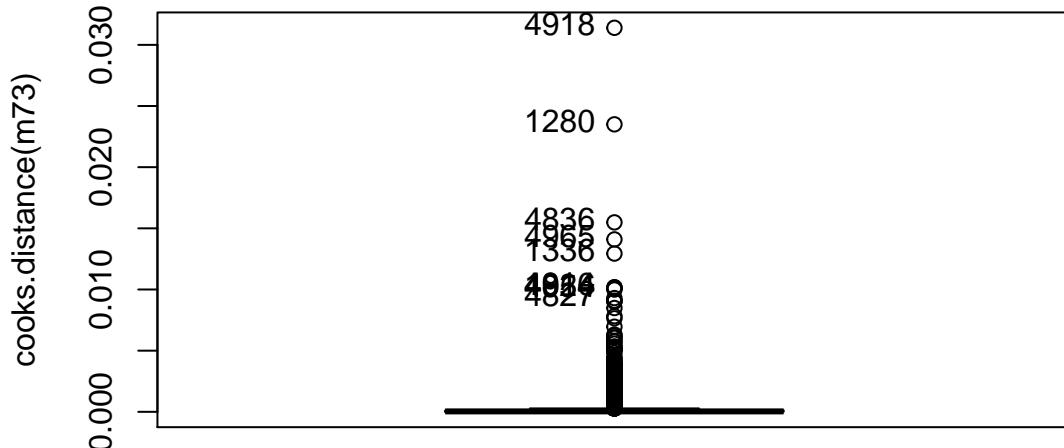


```
influencePlot(m73)
```



```
##          StudRes      Hat      CookD
## 10038  2.4832006 0.060909838 0.0235019298
## 39584 -0.7680869 0.143271206 0.0058039438
## 39699 -0.2249883 0.143752609 0.0005000004
## 40528 -4.3754701 0.027217651 0.0313944221
## 40999 -4.2491549 0.009357579 0.0099980359
```

```
Boxplot(cooks.distance(m73))
```



```
## [1] 4918 1280 4836 4965 1336 4816 1914 4024 4954 4827
```

## Binary Regression Models - target binari d'acceptacio del producte financer “y”

Per tal d'elaborar un model lineal que predigui el valor de la variable binaria target  $y$ , primer hem de decidir quines son les variables (columnes) que utilitzarem en la seva construccio. En altres paraules, trobar quines variables ens aporten informacio i precisio al model predictiu, pero sense sobreparametritzar-lo.

### Work and test samples division

Dividim la nostra mostra en dues submostres: el 75% de la mostra inicial serà per a treballar amb les dades (dataframe work -  $dfw$ ), i el 25% restant serà per a testejar-les (dataframe test -  $dft$ ).

```
set.seed(69)
sam<-sample(1:nrow(df), 0.75*nrow(df)) #random sample without replacement

dfw<-df[sam,] #work75%
dft<-df[-sam,] #test25%
```

### Variables numeriques explicatives pel target binari

A partir de la informacio del catdes, que ens diu les variables numeriques mes explicatives, generem un model inicial del tipus binomial, i a partir d'aqui el simplificarem per a que no quedi sobreparametritzat (model gm2).

```
#numeric variables
catdes( dfw[,c("y", vars_con)], 1)
```

```

#glm amb les variables continues significatives
gm1<-glm(y~nr.employed+euribor3m+emp.var.rate+pdays+previous+cons.price.idx+
           cons.conf.idx+campaign, family=binomial, data=dfw); summary(gm1)

## 
## Call:
## glm(formula = y ~ nr.employed + euribor3m + emp.var.rate + pdays +
##       previous + cons.price.idx + cons.conf.idx + campaign, family = binomial,
##       data = dfw)
## 
## Deviance Residuals:
##    Min      1Q   Median      3Q      Max
## -1.6348 -0.3897 -0.3574 -0.2725  2.7048
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -86.895389  41.451889 -2.096 0.036056 *
## nr.employed -0.000228  0.003897 -0.058 0.953354
## euribor3m   -0.343097  0.238396 -1.439 0.150096
## emp.var.rate -0.413371  0.192678 -2.145 0.031921 *
## pdays        -0.106215  0.018245 -5.822 5.82e-09 ***
## previous     -0.362876  0.124945 -2.904 0.003681 **
## cons.price.idx  0.974082  0.265630  3.667 0.000245 ***
## cons.conf.idx   0.051647  0.016582  3.115 0.001842 **
## campaign      -0.044702  0.030582 -1.462 0.143822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
## Null deviance: 2619.2 on 3738 degrees of freedom
## Residual deviance: 2196.2 on 3730 degrees of freedom
## AIC: 2214.2
## 
## Number of Fisher Scoring iterations: 5
Anova(gm1)

```

```

## Analysis of Deviance Table (Type II tests)
## 
## Response: y
##             LR Chisq Df Pr(>Chisq)
## nr.employed    0.003  1  0.9533446
## euribor3m      2.088  1  0.1484697
## emp.var.rate    4.572  1  0.0324990 *
## pdays          35.065  1  3.189e-09 ***
## previous        8.879  1  0.0028853 **
## cons.price.idx 12.730  1  0.0003599 ***
## cons.conf.idx   9.757  1  0.0017864 **
## campaign        2.323  1  0.1274845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
vif(gm1) #check colinear variables

```

```

##      nr.employed      euribor3m    emp.var.rate      pdays      previous
##      32.326246       60.227711     33.905107      1.750267     1.944268
## cons.price.idx  cons.conf.idx      campaign
##      10.091478       2.980404     1.025755

#our strategy: remove 2 colinear variables and campagin
gm2<-glm(y~emp.var.rate+pdays+previous+cons.price.idx+cons.conf.idx,
           family=binomial, data=dfw); summary(gm2)

##
## Call:
## glm(formula = y ~ emp.var.rate + pdays + previous + cons.price.idx +
##       cons.conf.idx, family = binomial, data = dfw)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q        Max
## -1.7093  -0.3918  -0.3725  -0.2786   2.6749
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.127e+02  1.312e+01 -8.590 < 2e-16 ***
## emp.var.rate -8.590e-01  5.815e-02 -14.771 < 2e-16 ***
## pdays        -1.045e-01  1.820e-02 -5.744 9.26e-09 ***
## previous     -3.158e-01  1.227e-01 -2.573  0.0101 *
## cons.price.idx 1.218e+00  1.407e-01  8.653 < 2e-16 ***
## cons.conf.idx  3.905e-02  9.953e-03  3.924 8.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2619.2 on 3738 degrees of freedom
## Residual deviance: 2204.1 on 3733 degrees of freedom
## AIC: 2216.1
##
## Number of Fisher Scoring iterations: 5

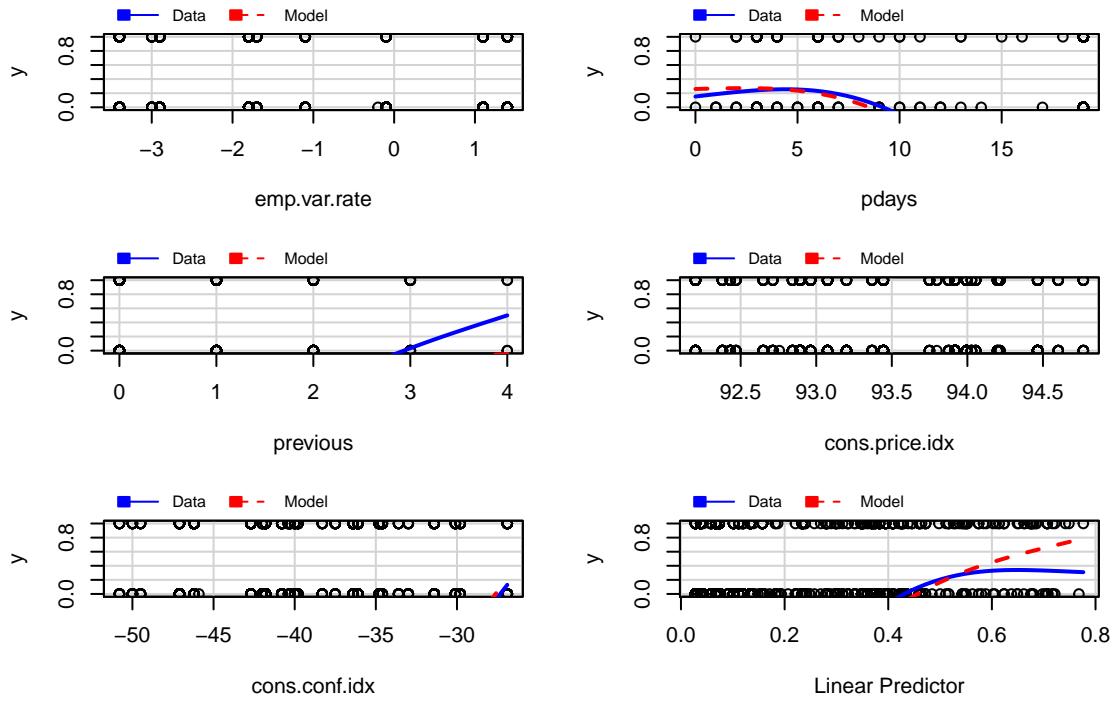
vif(gm2)

##      emp.var.rate      pdays      previous cons.price.idx  cons.conf.idx
##      3.120624       1.738325     1.878031     2.877660     1.060744

marginalModelPlots(gm2) #some missfit data vs model

```

## Marginal Model Plots



### Transforming variables

No veiem cap patró en els marginal plots que ens pugui ajudar a l'hora de seleccionar una transformació de variable en el model. Alguns que hem provat no milloraven el model.

```
#gm3<-glm(y~emp.var.rate+pdays+poly(previous, 2)+cons.price.idx+cons.conf.idx,
#           family=binomial, data=dfw); summary(gm3)
#Anova(gm3)
#marginalModelPlots(gm3)
```

### Variables discretes explicatives pel target binari

A partir del millor model en aquest punt (gm2), comprovem per a cada variable numèrica si es millor la seva utilització com a factor o com a numèrica. Ens quedem doncs amb un model gm2f, el qual té les variables pdays i f.cons.conf.idx com a factors i la resta com a numèriques:  $y \sim emp.var.rate + f.pdays + previous + cons.price.idx + cons.conf.idx$ .

```
#amb pdays, ojo!!!! pq la continua ha estat majoritàriament imputada, per tant en aquest cas,
#no fem cas del test, agafem la factoritzada!
gm2<-glm(y~emp.var.rate+f.pdays+previous+cons.price.idx+cons.conf.idx, family=binomial, data=dfw)

#f.emp.var.rate or emp.var.rate?
gm2f<-glm(y~f.emp.var.rate+f.pdays+previous+cons.price.idx+cons.conf.idx, family=binomial, data=dfw)
BIC(gm2, gm2f)
#f.emp.var.rate dona pitjor (mes baix) BIC, ens quedem amb la variable numèrica

gm2f<-glm(y~emp.var.rate+f.pdays+f.previous+cons.price.idx+cons.conf.idx, family=binomial, data=dfw)
BIC(gm2, gm2f)
#previous com a numèrica
```

```

gm2f<-glm(y~emp.var.rate+f.pdays+previous+f.cons.price.idx+cons.conf.idx, family=binomial, data=dfw)
BIC(gm2, gm2f)
#cons.price.idx com a numerica

gm2f<-glm(y~emp.var.rate+f.pdays+previous+cons.price.idx+f.cons.conf.idx, family=binomial, data=dfw)
BIC(gm2, gm2f)
#f.cons.conf.idx com a factor

```

Ara afegirem les variables discretes explicatives que siguin significatives, ho farem a partir de les que ens indiqui el catdes, sempre sense repetir una variable si ja esta representada com a numerica. Mirem les colinearitats i eliminem les variables que en tinguin, per a continuaciO simplifcar el model amb la comanda step. Veiem com finalment el model gm44 es un model sense colinearitats, on tots els efectes de les variables explicatives son significatius.

```

#discrete variables
catdes( dfw[,c("y", vars_dis)], 1)

#assumim gm2f el millor model en aquest punt i li afegim les significatives
#(entre month i f.season triem la primera):
gm4<-glm(y~emp.var.rate+f.pdays+previous+cons.price.idx+f.cons.conf.idx+f.nr.employed+poutcome
          +f.euribor3m+contact+default+f.age+education+month+marital+f.campaign, family=binomial, data=dfw)
summary(gm4)

## 
## Call:
## glm(formula = y ~ emp.var.rate + f.pdays + previous + cons.price.idx +
##       f.cons.conf.idx + f.nr.employed + poutcome + f.euribor3m +
##       contact + default + f.age + education + month + marital +
##       f.campaign, family = binomial, data = dfw)
## 
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.0209 -0.4060 -0.3271 -0.2619  2.8284
## 
## Coefficients:
##                               Estimate Std. Error
## (Intercept)                -1.172e+02  4.731e+01
## emp.var.rate                 -8.825e-01  4.864e-01
## f.pdaysf.pdays-never        -3.195e-01  6.806e-01
## previous                      -1.861e-01  2.263e-01
## cons.price.idx                  1.226e+00  4.978e-01
## f.cons.conf.idxf.cons.conf.idx-(-42.7,-41.8]   -1.839e-01  5.606e-01
## f.cons.conf.idxf.cons.conf.idx-(-41.8,-36.4]    8.917e-01  4.222e-01
## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9]    6.494e-01  3.654e-01
## f.nr.employedf.nr.employed-(5.1e+03,5.19e+03]   -5.090e-01  1.319e+00
## f.nr.employedf.nr.employed-(5.19e+03,5.23e+03]   3.442e-02  1.460e+00
## poutcomepoutcome-nonexistent      2.584e-01  3.287e-01
## poutcomepoutcome-success         1.107e+00  6.593e-01
## f.euribor3mf.euribor3m-(1.33,4.86]      1.395e-01  3.972e-01
## f.euribor3mf.euribor3m-(4.86,4.96]      8.233e-01  4.758e-01
## f.euribor3mf.euribor3m-(4.96,5]        5.810e-01  5.496e-01
## contactcontact-telephone        -4.058e-01  2.195e-01
## defaultdefault-unknown         -1.550e-01  1.864e-01
## f.agef.age-(32,38]              -3.159e-01  1.692e-01

```

```

## f.agef.age-(38,47]          -2.719e-01  1.817e-01
## f.agef.age-(47,87]          -1.099e-01  1.777e-01
## educationeducation-basic.6y 1.258e-03  3.308e-01
## educationeducation-basic.9y -9.690e-02  2.532e-01
## educationeducation-high.school -1.759e-02  2.276e-01
## educationeducation-professional.course 2.463e-01  2.474e-01
## educationeducation-university.degree 2.076e-01  2.153e-01
## monthmonth-aug             -8.638e-02  4.435e-01
## monthmonth-dec             -7.972e-01  6.361e-01
## monthmonth-jul             -6.037e-02  4.813e-01
## monthmonth-jun             -6.067e-01  4.825e-01
## monthmonth-mar             5.768e-01  4.024e-01
## monthmonth-may             -3.841e-01  4.299e-01
## monthmonth-nov             -5.517e-01  5.415e-01
## monthmonth-oct             -5.562e-02  4.900e-01
## monthmonth-sep             -2.901e-01  4.559e-01
## maritalmarital-married    7.928e-02  1.972e-01
## maritalmarital-single     7.256e-02  2.238e-01
## f.campaignf.campaign-(2,5] 8.189e-02  1.405e-01
## f.campaignf.campaign-(5,25] -2.359e-01  2.540e-01
##
## (Intercept)                  -2.477   0.0133 *
## emp.var.rate                 -1.814   0.0696 .
## f.pdaysf.pdays-never        -0.469   0.6387
## previous                      -0.822   0.4110
## cons.price.idx                2.464   0.0137 *
## f.cons.conf.idxf.cons.conf.idx(-42.7,-41.8] -0.328   0.7428
## f.cons.conf.idxf.cons.conf.idx(-41.8,-36.4]  2.112   0.0347 *
## f.cons.conf.idxf.cons.conf.idx(-36.4,-26.9]  1.777   0.0755 .
## f.nr.employedf.nr.employed-(5.1e+03,5.19e+03] -0.386   0.6996
## f.nr.employedf.nr.employed-(5.19e+03,5.23e+03]  0.024   0.9812
## poutcomepoutcome-nonexistent 0.786   0.4319
## poutcomepoutcome-success     1.680   0.0930 .
## f.euribor3mf.euribor3m-(1.33,4.86]  0.351   0.7254
## f.euribor3mf.euribor3m-(4.86,4.96]  1.730   0.0836 .
## f.euribor3mf.euribor3m-(4.96,5]      1.057   0.2904
## contactcontact-telephone     -1.849   0.0645 .
## defaultdefault-unknown       -0.831   0.4057
## f.agef.age-(32,38]           -1.867   0.0619 .
## f.agef.age-(38,47]           -1.496   0.1346
## f.agef.age-(47,87]           -0.619   0.5362
## educationeducation-basic.6y  0.004   0.9970
## educationeducation-basic.9y  -0.383   0.7020
## educationeducation-high.school -0.077   0.9384
## educationeducation-professional.course 0.995   0.3195
## educationeducation-university.degree 0.964   0.3348
## monthmonth-aug              -0.195   0.8456
## monthmonth-dec              -1.253   0.2101
## monthmonth-jul              -0.125   0.9002
## monthmonth-jun              -1.257   0.2086
## monthmonth-mar              1.433   0.1518
## monthmonth-may              -0.893   0.3717
## monthmonth-nov              -1.019   0.3082
## monthmonth-oct              -0.114   0.9096

```

```

## monthmonth-sep -0.636 0.5246
## maritalmarital-married 0.402 0.6877
## maritalmarital-single 0.324 0.7457
## f.campaignf.campaign-(2,5] 0.583 0.5600
## f.campaignf.campaign-(5,25] -0.929 0.3528
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2619.2 on 3738 degrees of freedom
## Residual deviance: 2116.9 on 3701 degrees of freedom
## AIC: 2192.9
##
## Number of Fisher Scoring iterations: 6
vif(gm4) #valor de colinealitat en la 2a columna (eliminem fins a obtenir tots els valors <3)

##                               GVIF Df GVIF^(1/(2*Df))
## emp.var.rate      221.469476 1     14.881851
## f.pdays          11.753022 1     3.428268
## previous         6.098233 1     2.469460
## cons.price.idx   31.890396 1     5.647158
## f.cons.conf.idx  219.801351 3     2.456622
## f.nr.employed    1148.647706 2     5.821658
## poutcome        30.415086 2     2.348401
## f.euribor3m     104.117842 3     2.168973
## contact         2.676215 1     1.635914
## default         1.156703 1     1.075501
## f.age            1.574669 3     1.078611
## education       1.320436 5     1.028186
## month           1483.148770 9     1.500296
## marital          1.422431 2     1.092088
## f.campaign      1.078035 2     1.018962

gm4<-glm(y~previous+cons.price.idx+f.cons.conf.idx+poutcome+f.euribor3m+contact+default+f.age
          +education+month+marital+f.campaign, family=binomial, data=dfw); summary(gm4); vif(gm4)

##
## Call:
## glm(formula = y ~ previous + cons.price.idx + f.cons.conf.idx +
##       poutcome + f.euribor3m + contact + default + f.age + education +
##       month + marital + f.campaign, family = binomial, data = dfw)
##
## Deviance Residuals:
##      Min        1Q        Median        3Q        Max 
## -1.8595   -0.4196   -0.3331   -0.2717    2.8330 
##
## Coefficients:
##                               Estimate Std. Error z value
## (Intercept)                 -8.025e+00  1.504e+01 -0.534
## previous                   -9.403e-02  2.023e-01 -0.465
## cons.price.idx                8.483e-02  1.609e-01  0.527
## f.cons.conf.idxf.cons.conf.idxx(-42.7,-41.8] -5.047e-01  4.076e-01 -1.238
## f.cons.conf.idxf.cons.conf.idxx(-41.8,-36.4]  9.311e-01  2.464e-01  3.778

```

```

## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9] 1.214e-01 2.709e-01 0.448
## poutcomepoutcome-nonexistent 2.804e-01 3.186e-01 0.880
## poutcomepoutcome-success 1.447e+00 2.680e-01 5.397
## f.euribor3mf.euribor3m-(1.33,4.86] -1.606e+00 2.688e-01 -5.973
## f.euribor3mf.euribor3m-(4.86,4.96] -1.212e+00 2.632e-01 -4.606
## f.euribor3mf.euribor3m-(4.96,5] -1.688e+00 2.763e-01 -6.109
## contactcontact-telephone -3.760e-01 2.046e-01 -1.838
## defaultdefault-unknown -2.144e-01 1.852e-01 -1.158
## f.agef.age-(32,38] -3.252e-01 1.677e-01 -1.939
## f.agef.age-(38,47] -2.509e-01 1.801e-01 -1.394
## f.agef.age-(47,87] -7.597e-02 1.764e-01 -0.431
## educationeducation-basic.6y -3.253e-02 3.277e-01 -0.099
## educationeducation-basic.9y -1.432e-01 2.513e-01 -0.570
## educationeducation-high.school -6.174e-02 2.253e-01 -0.274
## educationeducation-professional.course 1.846e-01 2.454e-01 0.752
## educationeducation-university.degree 1.908e-01 2.127e-01 0.897
## monthmonth-aug -1.283e+00 3.746e-01 -3.424
## monthmonth-dec -1.380e+00 6.047e-01 -2.282
## monthmonth-jul -1.297e+00 3.379e-01 -3.839
## monthmonth-jun -1.321e+00 4.690e-01 -2.817
## monthmonth-mar -9.427e-05 4.013e-01 0.000
## monthmonth-may -2.156e+00 3.174e-01 -6.792
## monthmonth-nov -1.203e+00 3.909e-01 -3.078
## monthmonth-oct -5.709e-01 3.857e-01 -1.480
## monthmonth-sep -1.081e+00 4.439e-01 -2.435
## maritalmarital-married 1.148e-01 1.974e-01 0.582
## maritalmarital-single 1.178e-01 2.233e-01 0.527
## f.campaignf.campaign-(2,5] 1.108e-01 1.390e-01 0.797
## f.campaignf.campaign-(5,25] -2.182e-01 2.521e-01 -0.866
## Pr(>|z|)
## (Intercept) 0.593560
## previous 0.642085
## cons.price.idx 0.597954
## f.cons.conf.idxf.cons.conf.idx-(-42.7,-41.8] 0.215647
## f.cons.conf.idxf.cons.conf.idx-(-41.8,-36.4] 0.000158 ***
## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9] 0.653985
## poutcomepoutcome-nonexistent 0.378757
## poutcomepoutcome-success 6.77e-08 ***
## f.euribor3mf.euribor3m-(1.33,4.86] 2.33e-09 ***
## f.euribor3mf.euribor3m-(4.86,4.96] 4.10e-06 ***
## f.euribor3mf.euribor3m-(4.96,5] 1.00e-09 ***
## contactcontact-telephone 0.066101 .
## defaultdefault-unknown 0.246929
## f.agef.age-(32,38] 0.052509 .
## f.agef.age-(38,47] 0.163427
## f.agef.age-(47,87] 0.666642
## educationeducation-basic.6y 0.920933
## educationeducation-basic.9y 0.568902
## educationeducation-high.school 0.784055
## educationeducation-professional.course 0.451778
## educationeducation-university.degree 0.369771
## monthmonth-aug 0.000617 ***
## monthmonth-dec 0.022475 *
## monthmonth-jul 0.000123 ***

```

```

## monthmonth-jun          0.004853 **
## monthmonth-mar          0.999813
## monthmonth-may          1.10e-11 ***
## monthmonth-nov          0.002087 **
## monthmonth-oct          0.138834
## monthmonth-sep          0.014906 *
## maritalmarital-married 0.560792
## maritalmarital-single   0.597989
## f.campaignf.campaign-(2,5] 0.425166
## f.campaignf.campaign-(5,25] 0.386737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2619.2  on 3738  degrees of freedom
## Residual deviance: 2140.3  on 3705  degrees of freedom
## AIC: 2208.3
##
## Number of Fisher Scoring iterations: 6

##                      GVIF Df GVIF^(1/(2*Df))
## previous           4.846187 1    2.201406
## cons.price.idx    3.254165 1    1.803930
## f.cons.conf.idx   28.988886 3    1.752691
## poutcome          5.029068 2    1.497517
## f.euribor3m      16.970692 3    1.603061
## contact           2.340668 1    1.529924
## default            1.147276 1    1.071110
## f.age              1.545620 3    1.075269
## education          1.300784 5    1.026646
## month              64.251018 9    1.260195
## marital            1.411570 2    1.089998
## f.campaign         1.070570 2    1.017194

```

#### Anova(gm4)

```

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                      LR Chisq Df Pr(>Chisq)
## previous           0.216   1   0.64181
## cons.price.idx    0.278   1   0.59809
## f.cons.conf.idx   35.173   3   1.120e-07 ***
## poutcome          30.291   2   2.645e-07 ***
## f.euribor3m      69.337   3   5.918e-15 ***
## contact            3.516   1   0.06078 .
## default            1.383   1   0.23951
## f.age              4.815   3   0.18588
## education          4.885   5   0.43004
## month              81.877   9   6.843e-14 ***
## marital            0.361   2   0.83497
## f.campaign         1.687   2   0.43025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

gm44<-step(gm4, k=log(nrow(dfw)))
vif(gm44)
summary(gm44)

#provem si f.season va millor que la variable month
gm44season<-glm(y~f.cons.conf.idx+poutcome+f.euribor3m+f.season, family=binomial, data=dfw)
#equivalent al test de Fisher pero per al target factor (binari):
anova(gm44, gm44season, test="Chisq") #Pr(>Chi) < 2.2e-1 -> no son equivalents, per tant el model

## Analysis of Deviance Table
##
## Model 1: y ~ f.cons.conf.idx + poutcome + f.euribor3m + month
## Model 2: y ~ f.cons.conf.idx + poutcome + f.euribor3m + f.season
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3721     2158.4
## 2      3728    2259.4 -7   -100.97 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#mes gran (amb month que te mes categories) hauria de ser millor

#podriem crear una nova variable a partir de months que sigui months bons i dolents,
#podria simplificar el model!!

```

## Interaccions

En aquest apartat intentarem millorar el nostre model actual (gm44) a traves de les interaccions. Establirem un model inicial (gmint) que anira acumulant les millores en el cas de que l'aportacio de les interaccions sigui positiva per al model. En el primer cas provarem la interaccio f.euribor3m amb f.cons.conf.idx, la qual veiem que ens aporta informacio al model.

```

gmint<-glm(y~f.cons.conf.idx+poutcome+f.euribor3m+month, family=binomial, data=dfw)
gmint1<-glm(y~f.cons.conf.idx*f.euribor3m+poutcome+month, family=binomial, data=dfw)
```

```
Anova(gmint1); anova(gmint, gmint1, test="Chisq") #H0 rej -> models no equivalents
```

```

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                         LR Chisq Df Pr(>Chisq)
## f.cons.conf.idx          43.105  3  2.337e-09 ***
## f.euribor3m              93.292  3  < 2.2e-16 ***
## poutcome                  32.667  2  8.062e-08 ***
## month                      25.674  9  0.002309 **
## f.cons.conf.idx:f.euribor3m 30.062  6  3.825e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##
## Model 1: y ~ f.cons.conf.idx + poutcome + f.euribor3m + month
## Model 2: y ~ f.cons.conf.idx * f.euribor3m + poutcome + month
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3721     2158.4
## 2      3715     2128.3  6   30.062 3.825e-05 ***
## ---
```

```
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

A continuacio, tot i que segons tests Anova anteriors cap variable numerica ens aportava informacio significativa, provarem la interaccion entre una variable factor i una numerica, tal i com se'sns diu a la practica. Sera el cas de la interaccio entre cons.price.idx i f.euribor3m, la qual l'affegirem al model gmint1. Veiem doncs com si que aquesta variable numerica en forma d'interaccio ens aporta infomacio significativa i ens quedem amb el model gmint2. Si mitjancant la comanda step simplifiquem el model, obtenim el model gmint3, que trobem no ser equivalent al model gmint2, pel qual no acabarem de fer-li cas a la comanda step i ens quedarem amb el model anterior gmint2.
```

```
gmint2<-glm(y~cons.price.idx:f.euribor3m+poutcome+month+f.cons.conf.idx*f.euribor3m,
              family=binomial, data=dfw)
Anova(gmint2); anova(gmint1, gmint2, test="Chisq") #H0 rej -> models no equivalents
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                         LR Chisq Df Pr(>Chisq)
## poutcome                  30.767  2  2.084e-07 ***
## month                      12.929  9   0.165863
## f.cons.conf.idx            47.082  3  3.339e-10 ***
## f.euribor3m                 93.292  3 < 2.2e-16 ***
## cons.price.idx:f.euribor3m    9.542  3   0.022893 *
## f.euribor3m:f.cons.conf.idx 18.694  5   0.002191 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ f.cons.conf.idx * f.euribor3m + poutcome + month
## Model 2: y ~ cons.price.idx:f.euribor3m + poutcome + month + f.cons.conf.idx *
##             f.euribor3m
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3715     2128.3
## 2      3712     2118.8  3    9.5416  0.02289 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gmint3<-step(gmint2, k=log(nrow(dfw)))
```

```
## Start: AIC=2340.91
## y ~ cons.price.idx:f.euribor3m + poutcome + month + f.cons.conf.idx *
##             f.euribor3m
##
##                         Df Deviance    AIC
## - month                      9   2131.7 2279.8
## - f.euribor3m:f.cons.conf.idx 5   2137.5 2318.5
## - cons.price.idx:f.euribor3m   3   2128.3 2325.8
## <none>                      2118.8 2340.9
## - poutcome                     2   2149.6 2355.2
##
## Step: AIC=2279.8
## y ~ poutcome + f.cons.conf.idx + f.euribor3m + cons.price.idx:f.euribor3m +
##             f.euribor3m:f.cons.conf.idx
##
##                         Df Deviance    AIC
```

```

## - f.cons.conf.idx:f.euribor3m 5 2166.0 2273.0
## - f.euribor3m:cons.price.idx 3 2154.0 2277.4
## <none> 2131.7 2279.8
## - poutcome 2 2160.9 2292.5
##
## Step: AIC=2272.98
## y ~ poutcome + f.cons.conf.idx + f.euribor3m + f.euribor3m:cons.price.idx
##
##                                     Df Deviance    AIC
## <none>                         2166.0 2273.0
## - poutcome                      2 2198.2 2288.7
## - f.cons.conf.idx                3 2279.5 2361.8
## - f.euribor3m:cons.price.idx   4 2288.1 2362.1
anova(gmint2, gmint3, test="Chisq")

## Analysis of Deviance Table
##
## Model 1: y ~ cons.price.idx:f.euribor3m + poutcome + month + f.cons.conf.idx *
##           f.euribor3m
## Model 2: y ~ poutcome + f.cons.conf.idx + f.euribor3m + f.euribor3m:cons.price.idx
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     3712    2118.8
## 2     3726    2166.0 -14  -47.234 1.759e-05 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#H0 rej -> models no equivalents

```

### Interaction between a couple of factors in our model mgint2

El model escollit mgint2 considera una interacció entre dos factors *f.cons.conf.idx:f.euribor3m*.

```
table(dfw$f.euribor3m, dfw$f.cons.conf.idx)
```

```

##
##                                     f.cons.conf.idx-[-50.8,-42.7]
## f.euribor3m-[0.635,1.33]          507
## f.euribor3m-(1.33,4.86]          281
## f.euribor3m-(4.86,4.96]          241
## f.euribor3m-(4.96,5]             351
##
##                                     f.cons.conf.idx-(-42.7,-41.8]
## f.euribor3m-[0.635,1.33]           0
## f.euribor3m-(1.33,4.86]          328
## f.euribor3m-(4.86,4.96]          347
## f.euribor3m-(4.96,5]              53
##
##                                     f.cons.conf.idx-(-41.8,-36.4]
## f.euribor3m-[0.635,1.33]          164
## f.euribor3m-(1.33,4.86]          499
## f.euribor3m-(4.86,4.96]          266
## f.euribor3m-(4.96,5]              3
##
##                                     f.cons.conf.idx-(-36.4,-26.9]
## f.euribor3m-[0.635,1.33]          260

```

```

##   f.euribor3m-(1.33,4.86]          0
##   f.euribor3m-(4.86,4.96]          0
##   f.euribor3m-(4.96,5]           439

```

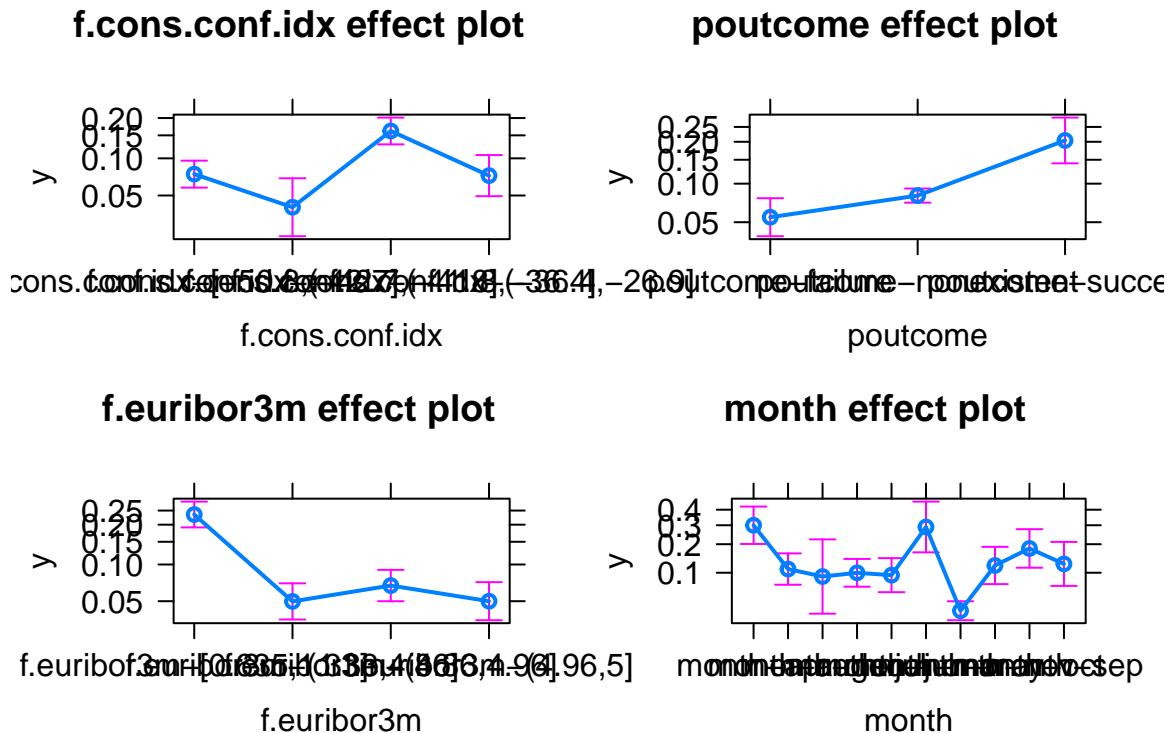
### Interaction between a factor and a covariate in our model mgint2

El model escollit mgint2 també considera una interacció entre un factor i una variable numèrica *cons.price.idx:f.euribor3m*.

```

#model sense interaccions
plot(allEffects(gmint))

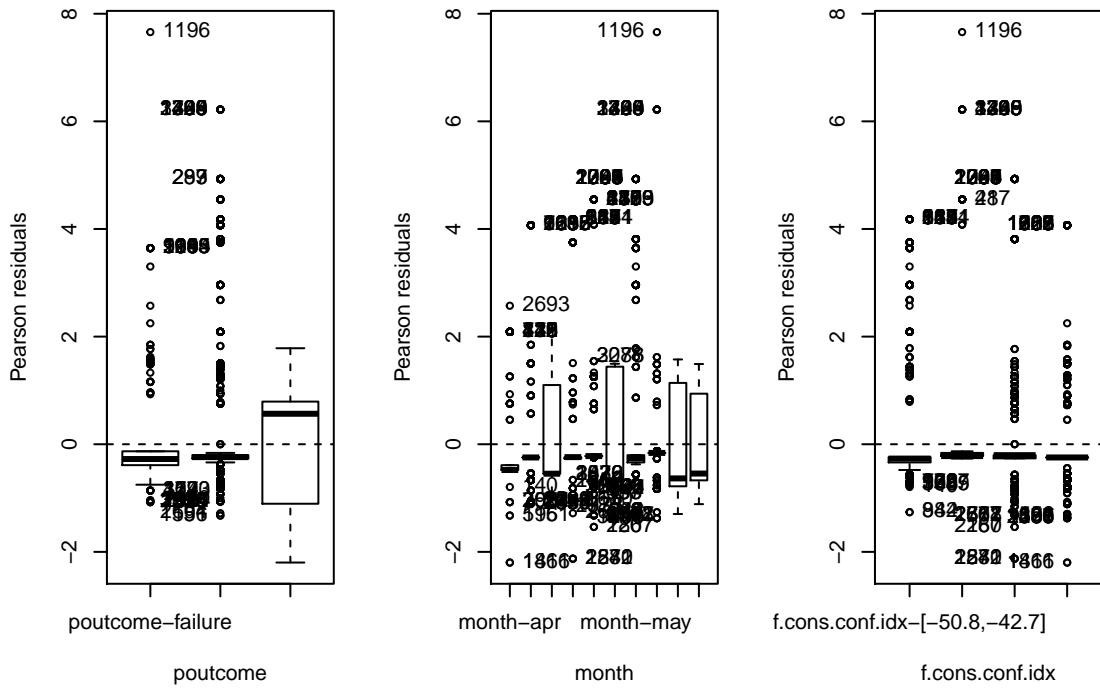
```



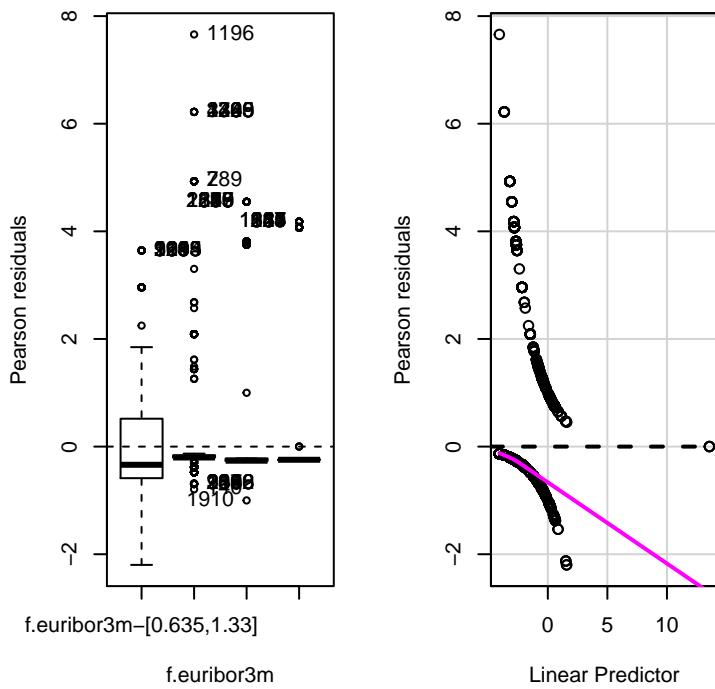
### Diagnostics del model definitiu gmint2

En els 5 primers grafics de les dues primeres figures que hi ha a continuació es poden veure els residus de les prediccions del model per a cada variable explicativa. El 6e grafic mostra un test de curvatura. En les següents figures i taula veiem les distàncies de Cook i els valors barret per a cadascuna de les observacions. Les observacions 24043 i 24031 son observacions molt influents en el model.

```
residualPlots(gmint2, layout=c(1, 3))
```

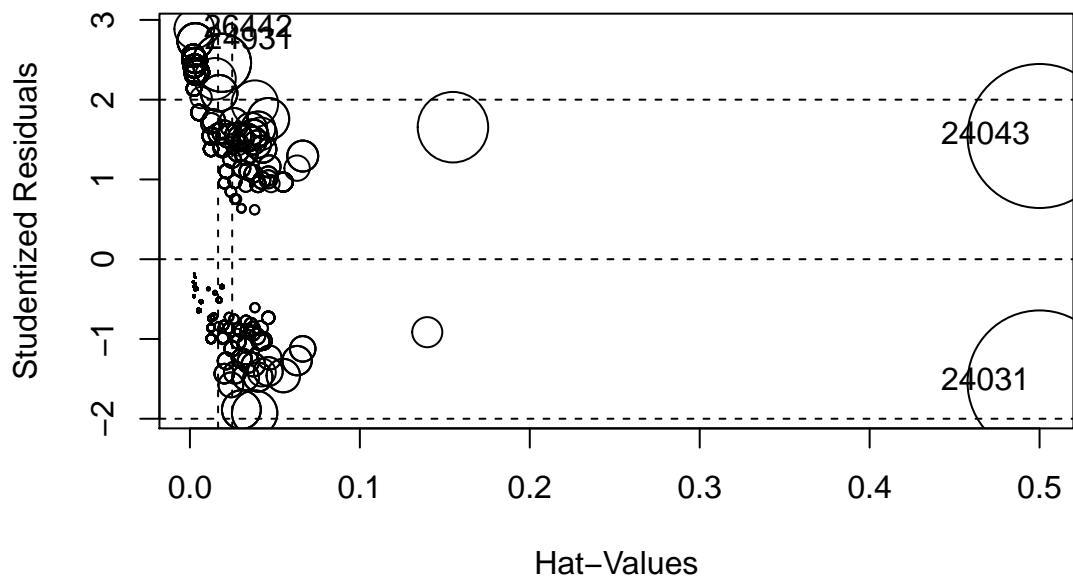
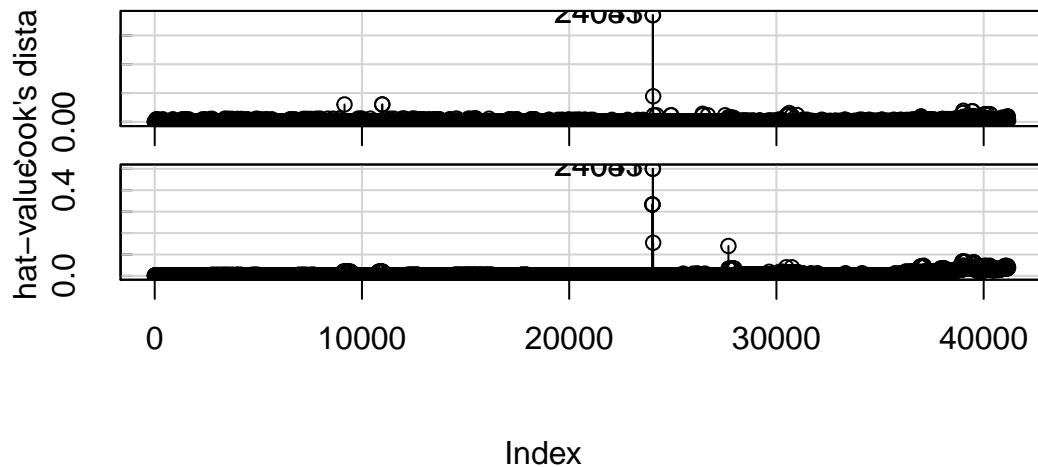


```
## Warning in residualPlots.default(model, ...): No possible lack-of-fit tests
```



```
influenceIndexPlot(gmint2, vars=c("Cook", "hat")); influencePlot(gmint2)
```

Diagnostic Plots



```
##          StudRes        Hat       CookD
## 24931  2.735807 0.003169808 0.004569850
## 26442  2.885971 0.002580695 0.005635131
```

```
## 24043 1.544764 0.500000000 0.074074074
## 24031 -1.544764 0.500000000 0.074074074
```

## Predictions i matriu de confusió de l'acceptació del producte financer $y$

Primer utilitzarem el model obtingut per a predir la resposta  $y$  de les dades de treball  $dfw$ , que son les mateixes amb les quals s'ha construït el model. Després farem el mateix però amb les dades de test  $dft$ , les quals no s'han usat per a la construcció del model. Si obtenim resultats similars en ambdós tests voldrà dir que el nostre model està correctament parametritzat i no ens hem adaptat a les dades de treball a l'hora de construir-lo. Veiem com per ambdós jocs de dades obtenim els mateixos resultats, amb una precisió del 90% d'ençert a l'hora de determinar l'acceptació o no del producte financer. Les prediccions correctes d'acceptació del producte financer o *sensitivity* són del 60-70%, mentre que les prediccions correctes de rebutj del producte financer són del 90%.

```
#work data
predw<-predict(gmint2, type="response")
predictionw<-prediction(predw, dfw$y)
predw.y <- factor(ifelse(as.numeric(predw)<0.5, 0, 1), labels=c("predw.y-no", "predw.y-yes"))
tablew<-addmargins(table(predw.y, dfw$y)); tablew

##
## predw.y      y-no y-yes Sum
##   predw.y-no 3265  334 3599
##   predw.y-yes  56   84  140
##   Sum         3321  418 3739

#test data
predt<-predict(gmint2, type="response", newdata=dft)

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
predictiont<-prediction(predt, dft$y)
predt.y <- factor(ifelse(as.numeric(predt)<0.5, 0, 1), labels=c("predt.y-no", "predt.y-yes"))
tablet<-addmargins(table(predt.y, dft$y)); tablet

##
## predt.y      y-no y-yes Sum
##   predt.y-no 1095  108 1203
##   predt.y-yes  13   31  44
##   Sum         1108  139 1247

#confusion matrix values
predicions_correctes_w<-sum(diag(tablew[1:2, 1:2]))/sum(tablew[1:2, 1:2])*100; predicions_correctes_w
## [1] 89.5694
predicions_correctes_t<-sum(diag(tablet[1:2, 1:2]))/sum(tablet[1:2, 1:2])*100; predicions_correctes_t
## [1] 90.29671
predicions_incorrectes_w<-(100-predicions_correctes_w); predicions_incorrectes_w
## [1] 10.4306
predicions_incorrectes_t<-(100-predicions_correctes_t); predicions_incorrectes_t
## [1] 9.703288
```

```

sensibility_w<-tablew[2,2]/sum(tablew[2, 1:2])*100; sensibility_w
## [1] 60
sensibility_t<-tablet[2,2]/sum(tablet[2, 1:2])*100; sensibility_t
## [1] 70.45455
specificity_w<-tablew[1,1]/sum(tablew[1, 1:2])*100; specificity_w
## [1] 90.71964
specificity_t<-tablet[1,1]/sum(tablet[1, 1:2])*100; specificity_t
## [1] 91.02244

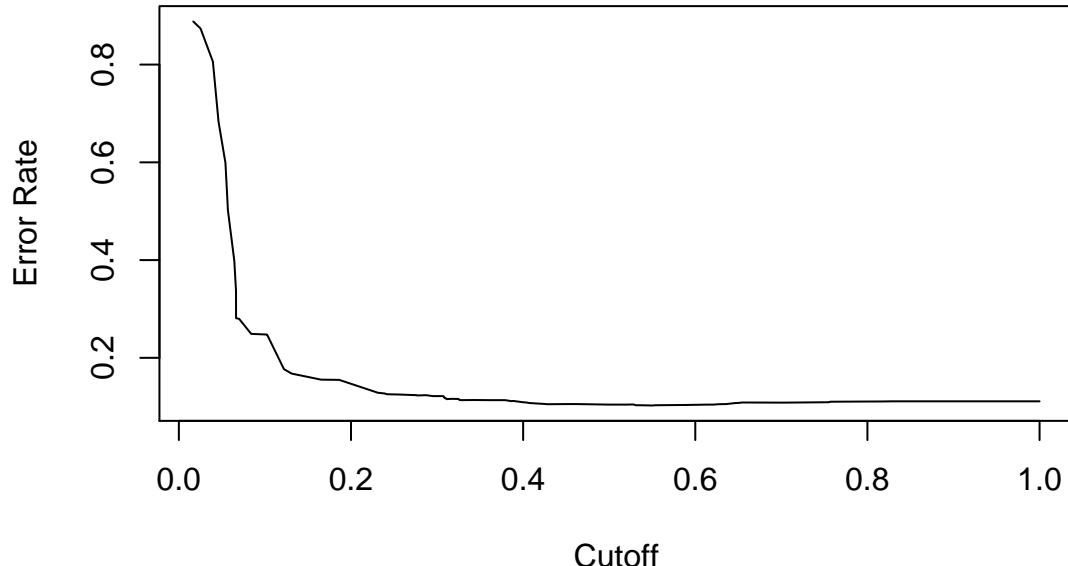
```

El llindar seleccionat anteriorment per a decidir si una probabilitat es una acceptació o un rebuig del producte ha estat l'standard de 0.5; a continuació també es buscarà si es poden millorar els resultats considerant un altre valor threshold, el qual valorarem a partir de les curbes ROC. En el primer grafic veiem com el llindar actual ens garantitza el maxim % d'encert global del model; en el segon grafic però, veiem també com la sensibilitat del model té encara marge de millora, a canvi d'incrementar els falsos positius, es a dir, predir una acceptació del producte financer i que finalment aquest sigui rebutjat.

```

dadesroc<-prediction(predict(gmint2, type="response"), dfw$y)
plot(performance(dadesroc, "err"));

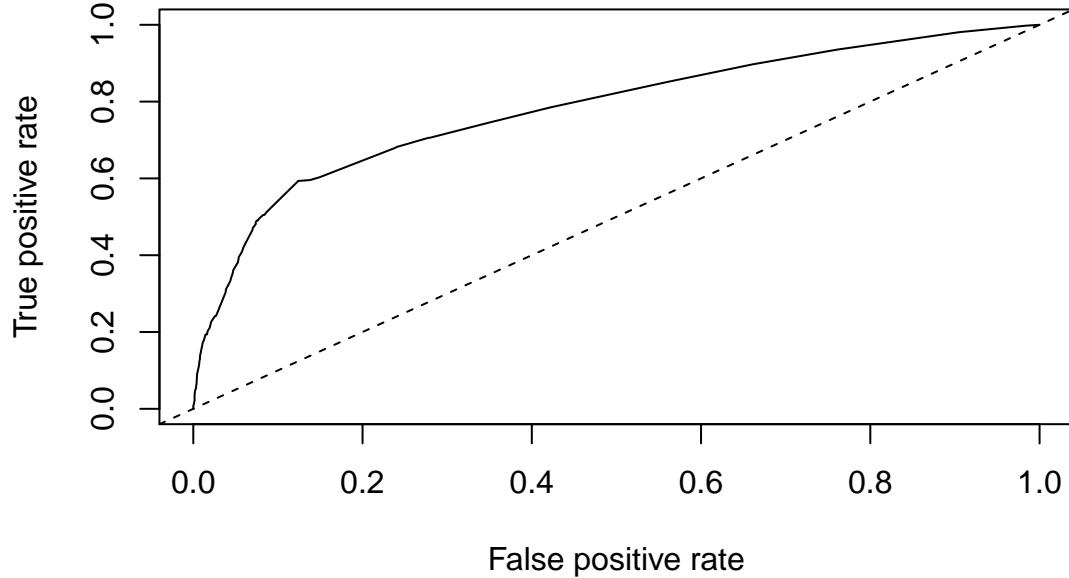
```



```

plot(performance(dadesroc, "tpr", "fpr"))
abline(0,1,lty=2)

```



Si canviem el valor llindar *cut off* a 0.7, el que vol dir que serem “mes pesimistes” a l’hora de dir que el client acceptara el producte financer a partir de la probabilitat donada pel model, obtenim uns millors resultats de sensibilitat (per sobre el 70% en ambdues mostres), mentres que la resta de valors segueixen similars. Tot i això, aquest valor es podria ajustar segons l’aplicació física real del model predictiu. Si ens interessa estar molt segurs que si predim un outcome positiu així sigui, hauríem de pujar el llindar, fet que comportaria que ens estiguessim predint algun valor positiu com a negatiu. Si el que destijem es el contrari, i simplement volem descartar nomes casos que l’outcome seria negatiu amb molta seguretat, hauríem de baixar el llindar.

```
#work data
predw<-predict(gmint2, type="response")
predictionw<-prediction(predw, dfw$y)
predw.y <- factor(ifelse(as.numeric(predw)<0.7, 0, 1), labels=c("predw.y-no", "predw.y-yes"))
tablew<-addmargins(table(predw.y, dfw$y)); tablew

##
## predw.y      y-no y-yes Sum
##   predw.y-no 3314  398 3712
##   predw.y-yes    7   20  27
##   Sum        3321  418 3739

#test data
predt<-predict(gmint2, type="response", newdata=dft)

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
predictiont<-prediction(predt, dft$y)
predt.y <- factor(ifelse(as.numeric(predt)<0.7, 0, 1), labels=c("predt.y-no", "predt.y-yes"))
tablet<-addmargins(table(predt.y, dft$y)); tablet

##
```

```

## predt.y      y-no y-yes Sum
##   predt.y-no 1105   131 1236
##   predt.y-yes    3     8  11
##   Sum          1108   139 1247
#confussion matrix values
predicions_correctes_w<-sum(diag(tablew[1:2, 1:2]))/sum(tablew[1:2, 1:2])*100; predicions_correctes_w

## [1] 89.16823
predicions_correctes_t<-sum(diag(tablet[1:2, 1:2]))/sum(tablet[1:2, 1:2])*100; predicions_correctes_t

## [1] 89.25421
predicions_incorrectes_w<-(100-predicions_correctes_w); predicions_incorrectes_w

## [1] 10.83177
predicions_incorrectes_t<-(100-predicions_correctes_t); predicions_incorrectes_t

## [1] 10.74579
sensibility_w<-tablew[2,2]/sum(tablew[2, 1:2])*100; sensibility_w

## [1] 74.07407
sensibility_t<-tablet[2,2]/sum(tablet[2, 1:2])*100; sensibility_t

## [1] 72.72727
specificity_w<-tablew[1,1]/sum(tablew[1, 1:2])*100; specificity_w

## [1] 89.27802
specificity_t<-tablet[1,1]/sum(tablet[1, 1:2])*100; specificity_t

## [1] 89.40129

```