

First delivery - ADEI

Alex Rubio i Josep Bernad

1 de març de 2019

Contents

1	Introduction	2
1.1	Bank client data	2
1.2	Loading packages	2
1.3	Loading data	2
2	Univariate Descriptive Analysis	4
2.1	Transform missing and wrong data to NAs	4
2.2	Create new factors corresponding to qualitative concepts.	5
2.3	Create new factors corresponding to quantitative concepts.	9
2.4	Exploratory Data Analysis	10
3	Data Quality Report	30
3.1	Variables	30
3.2	Individual	40
4	Correlation	40
5	Imputation	45
5.1	Numeric Variables	45
5.2	Factors	46
6	Profiling	47
7	PCA analysis	54
7.1	Eigenvalues and dominant axes analysis	54
7.2	Individuals point of view	60
7.3	Interpreting the axes	71
8	K-Means Classification	72
8.1	Description of clusters	74
9	Hierarchical Clustering	83
9.1	Description of clusters	85
10	CA analysis	90
10.1	f.duration - f.job	90
10.2	f.duration - f.age	93

1 Introduction

We are going to work with dataset bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014].

Deliverables are concerned with Multivariant Data Analysis and model building for response variables: Y-Duration of the call and binary factor Y (Binary Target) if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

1.1 Bank client data

1.1.1 Description

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')

1.2 Loading packages

1.3 Loading data

```
#rm(list=ls())
# Load Required Packages: to be increased over the course

# Josep
#setwd("~/Developer/r-studio/laboratory-adei/data-directory")
#load("~/Developer/r-studio/laboratory-adei/data-directory/5000_samples.RData")
```

```
# Alex
setwd("D:/Google Drive/Uni/ADEI/data-directory")
load(path.expand("D:/Google Drive/Uni/ADEI/data-directory/5000_samples.RData"))

summary(df)
```

```
##      age                job                marital
## Min.   :17.00   admin.      :1288   divorced: 546
## 1st Qu.:32.00   blue-collar:1156   married  :3029
## Median :38.00   technician : 831   single   :1416
## Mean   :39.97   services   : 471   unknown  : 9
## 3rd Qu.:47.00   management : 345
## Max.   :92.00   retired    : 187
##                (Other)    : 722
##      education      default      housing      loan
## university.degree :1431   no      :3939   no      :2226   no      :4138
## high.school        :1169   unknown:1061   unknown: 112   unknown: 112
## basic.9y           : 758   yes      : 0   yes      :2662   yes      : 750
## professional.course: 668
## basic.4y           : 493
## basic.6y           : 272
## (Other)            : 209
##      contact      month      day_of_week      duration
## cellular :3182   may      :1679   fri: 948   Min.    : 4.0
## telephone:1818   jul      : 907   mon:1017   1st Qu.: 104.0
##                aug      : 699   thu:1031   Median  : 181.0
##                jun      : 660   tue:1005   Mean    : 263.7
##                nov      : 502   wed: 999   3rd Qu.: 328.0
##                apr      : 323   Max.    :3078.0
##                (Other): 230
##      campaign      pdays      previous      poutcome
## Min.    : 1.000   Min.    : 0.0   Min.    :0.0000   failure   : 493
## 1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000   nonexistent:4315
## Median : 2.000   Median :999.0   Median :0.0000   success   : 192
## Mean    : 2.647   Mean    :957.9   Mean    :0.1772
## 3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000
## Max.    :42.000   Max.    :999.0   Max.    :5.0000
##
##      emp.var.rate   cons.price.idx   cons.conf.idx   euribor3m
## Min.    :-3.4000   Min.    :92.20   Min.    :-50.80   Min.    :0.634
## 1st Qu.: -1.8000   1st Qu.:93.08   1st Qu.: -42.70   1st Qu.:1.344
## Median : 1.1000   Median :93.88   Median : -41.80   Median :4.857
## Mean    : 0.1029   Mean    :93.58   Mean    : -40.59   Mean    :3.641
## 3rd Qu.: 1.4000   3rd Qu.:93.99   3rd Qu.: -36.40   3rd Qu.:4.961
## Max.    : 1.4000   Max.    :94.77   Max.    : -26.90   Max.    :5.045
##
##      nr.employed      y
## Min.    :4964   no :4416
## 1st Qu.:5099   yes: 584
## Median :5191
## Mean    :5168
## 3rd Qu.:5228
## Max.    :5228
```

```
##
```

2 Univariate Descriptive Analysis

Creem factors per cada variable posant abans NA a aquells valors erronis o faltants.

2.1 Transform missing and wrong data to NAs

```
#Default  
sel<-which(df$default=="unknown");length(sel)
```

```
## [1] 1061
```

```
df$default[sel] <- NA  
df$default <- factor(df$default)  
summary(df$default)
```

```
## no NA's  
## 3939 1061
```

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

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

```
df$marital[sel] <- NA  
df$marital <- factor(df$marital)  
summary(df$marital)
```

```
## divorced married single NA's  
## 546 3029 1416 9
```

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

```
## [1] 112
```

```
df$housing[sel] <- NA  
df$housing <- factor(df$housing)  
summary(df$housing)
```

```
## no yes NA's  
## 2226 2662 112
```

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

```
## [1] 112
```

```
df$loan[sel] <- NA  
df$loan <- factor(df$loan)  
summary(df$loan)
```

```
## no yes NA's  
## 4138 750 112
```

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

```
## [1] 43
```

```
df$job[sel] <- NA
df$job <- factor(df$job)
summary(df$job)
```

##	admin.	blue-collar	entrepreneur	housemaid	management
##	1288	1156	181	132	345
##	retired	self-employed	services	student	technician
##	187	152	471	100	831
##	unemployed	NA's			
##	114	43			

```
#Education
sel<-which(df$education=="unknown");length(sel)
```

```
## [1] 207
```

```
df$education[sel] <- NA
df$education <- factor(df$education)
summary(df$education)
```

##	basic.4y	basic.6y	basic.9y
##	493	272	758
##	high.school	illiterate	professional.course
##	1169	2	668
##	university.degree	NA's	
##	1431	207	

```
#Pdays
sel<-which(df$pdays==999);length(sel)
```

```
## [1] 4793
```

```
df$pdays[sel] <- NA
summary(df$pdays)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.000	3.000	6.000	5.792	7.000	18.000	4793

```
#Poutcome
sel<-which(df$poutcome=="nonexistent");length(sel)
```

```
## [1] 4315
```

```
df$poutcome[sel] <- NA
df$poutcome <- factor(df$poutcome)
summary(df$poutcome)
```

##	failure	success	NA's
##	493	192	4315

2.2 Create new factors corresponding to qualitative concepts.

2.2.1 Month

```
#Modify factor levels label
df$f.month <- factor(df$month, labels=paste("Month", sep="-", levels(df$month)))
table(df$f.month)
```

```
##
```

```
## Month-apr Month-aug Month-dec Month-jul Month-jun Month-mar Month-may
##      323      699      19      907      660      66      1679
## Month-nov Month-oct Month-sep
##      502      79      66

# Define new factor categories: 1-Spring / 2-Summer / 3-Resta
df$f.season <- 3

# 1 level - spring
sel<-which(df$f.month %in% c("Month-mar","Month-apr","Month-may"))
df$f.season[sel] <-1

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

table(df$f.season);summary(df$f.season)

##
##      1      2      3
## 2068 2266  666

##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      1.00   1.00   2.00   1.72   2.00   3.00

df$f.season<-factor(df$f.season,levels=1:3,labels=c("Spring","Summer","Aut-Win"))
summary(df$f.season)

##      Spring      Summer      Aut-Win
##      2068      2266      666
```

2.2.2 Job

```
#Modify factor levels label
df$f.job <- factor(df$job, labels=paste("Job", sep="-", levels(df$job)))

table(df$f.job)

##
##      Job-admin.      Job-blue-collar      Job-entrepreneur      Job-housemaid
##      1288      1156      181      132
##      Job-management      Job-retired      Job-self-employed      Job-services
##      345      187      152      471
##      Job-student      Job-technician      Job-unemployed
##      100      831      114

# Define new factor categories: 1-selfemployed / 2-worker / 3-other
df$f.jobsituation<-3

# 1 level - self-employed
sel<-which(df$f.job %in% c("Job-entrepreneur","Job-housemaid","Job-self-employed"))
df$f.jobsituation[sel] <- 1

# 2 level - worker
sel<-which(df$f.job %in% c("Job-admin","Job-blue-collar","Job-management","Job-services","Job-technician"))
df$f.jobsituation[sel] <- 2
```

```
table(df$f.jobssituation);summary(df$f.jobssituation)

##
##      1      2      3
## 465 2803 1732

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000   2.000   2.000   2.253   3.000   3.000

df$f.jobssituation<-factor(df$f.jobssituation,levels=1:3,labels=c("Self-employed","Worker","Other"))
summary(df$f.jobssituation)

## Self-employed      Worker      Other
##           465           2803           1732
```

2.2.3 Pdays

```
table(df$pdays)

##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 15 16 17 18
##  1  5 12 62 17  5 48 13  5  9  7  2  4  8  3  1  4  1

# Define new factor categories: 1-contacted / 2-not contacted
df$f.prev_contacted<-2

# 1 level - contacted
sel<-which(df$pdays %in% c(1:20))
df$f.prev_contacted[sel] <- 1

# 2 level - not contacted
sel<-which(df$pdays %in% c(21:1000))
df$f.prev_contacted[sel] <- 2

table(df$f.prev_contacted);summary(df$f.prev_contacted)

##
##      1      2
## 206 4794

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000   2.000   2.000   1.959   2.000   2.000

df$f.prev_contacted<-factor(df$f.prev_contacted,levels=1:2,labels=c("Contacted","No-contacted"))
summary(df$pdays)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      0.000   3.000   6.000   5.792   7.000  18.000    4793
```

2.2.4 Education

```
#Modify factor levels label
df$education <- factor(df$education, labels=paste("Edu", sep="-", levels(df$education)))

table(df$education)

##
```

```
##           Edu-basic.4y           Edu-basic.6y           Edu-basic.9y
##           493           272           758
##           Edu-high.school           Edu-illiterate Edu-professional.course
##           1169           2           668
##           Edu-university.degree
##           1431

# Define new factor categories: 1-mandatory / 2-nonmandatory / 3-other
df$f.education<-3

# 1 level - mandatory
sel<-which(df$education %in% c("Edu-basic.4y","Edu-basic.6y", "Edu-basic.9y", "Edu-high.school"))
df$f.education[sel] <- 1

# 2 level - nonmandatory
sel<-which(df$education %in% c("Edu-professional.course","Edu-university.degree"))
df$f.education[sel] <- 2

table(df$f.education);summary(df$f.education)

##
##      1      2      3
## 2692 2099  209

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000  1.000   1.000   1.503   2.000   3.000

df$f.education<-factor(df$f.education,levels=1:3,labels=c("Mandatory","Non-Mandatory","Other"))
summary(df$f.education)

##      Mandatory Non-Mandatory      Other
##           2692           2099           209
```

2.2.5 Extra Factorization

```
#Housing

df$f.housing<-factor(df$housing,labels=paste("f",sep=".",levels(df$housing)))
table(df$f.housing);summary(df$f.housing);

##
## f.no f.yes
##  2226  2662

## f.no f.yes NA's
##  2226  2662   112

#Marital

df$f.marital<-factor(df$marital,labels=paste("f",sep=".",levels(df$marital)))
table(df$f.marital);summary(df$f.marital);

##
## f.divorced f.married f.single
##           546          3029          1416

## f.divorced f.married f.single      NA's
##           546          3029          1416          9
```



```
#Default
df$f.default<-factor(df$default, labels=paste("f",sep=".",levels(df$default)))
df$f.default <- factor(df$f.default , levels = c(levels(df$f.default), "f.si"))
table(df$f.default);
```

```
##
## f.no f.si
## 3939    0
```

```
#Loan
df$f.loan<-factor(df$loan,labels=paste("f",sep=".",levels(df$loan)))
table(df$f.loan);summary(df$f.loan)
```

```
##
## f.no f.yes
## 4138    750

## f.no f.yes NA's
## 4138    750   112
```

```
#Contact
df$f.contact<-factor(df$contact,labels=paste("f",sep=".",levels(df$contact)))
table(df$f.contact);summary(df$f.contact)
```

```
##
## f.cellular f.telephone
##          3182          1818

## f.cellular f.telephone
##          3182          1818
```

```
#Day of Week
df$f.day<-factor(df$day_of_week,labels=paste("f.day",sep=".",levels(df$day)))
table(df$f.day);summary(df$f.day)
```

```
##
## f.day.fri f.day.mon f.day.thu f.day.tue f.day.wed
##          948          1017          1031          1005          999

## f.day.fri f.day.mon f.day.thu f.day.tue f.day.wed
##          948          1017          1031          1005          999
```

2.3 Create new factors corresponding to quantitative concepts.

2.3.1 Age discretization

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      17.00  32.00   38.00   39.97  47.00   92.00
```

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

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

```
## varaux
## [17,32] (32,38] (38,47] (47,92]
##      1353      1248      1202      1197
```

```

tapply(df$age,varaux,median)

## [17,32] (32,38] (38,47] (47,92]
##      29      35      43      53

varaux<-factor(cut(df$age,breaks=c(17,30,40,50,95),include.lowest=T))
table(varaux)

## varaux
## [17,30] (30,40] (40,50] (50,95]
##      887      2003      1252      858

tapply(df$age,varaux,median)

## [17,30] (30,40] (40,50] (50,95]
##      28      35      45      55

df$f.age<-factor(cut(df$age,breaks=c(17,30,40,50,95),include.lowest=T))

summary(df$f.age)

## [17,30] (30,40] (40,50] (50,95]
##      887      2003      1252      858

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

```

2.4 Exploratory Data Analysis

2.4.1 Age

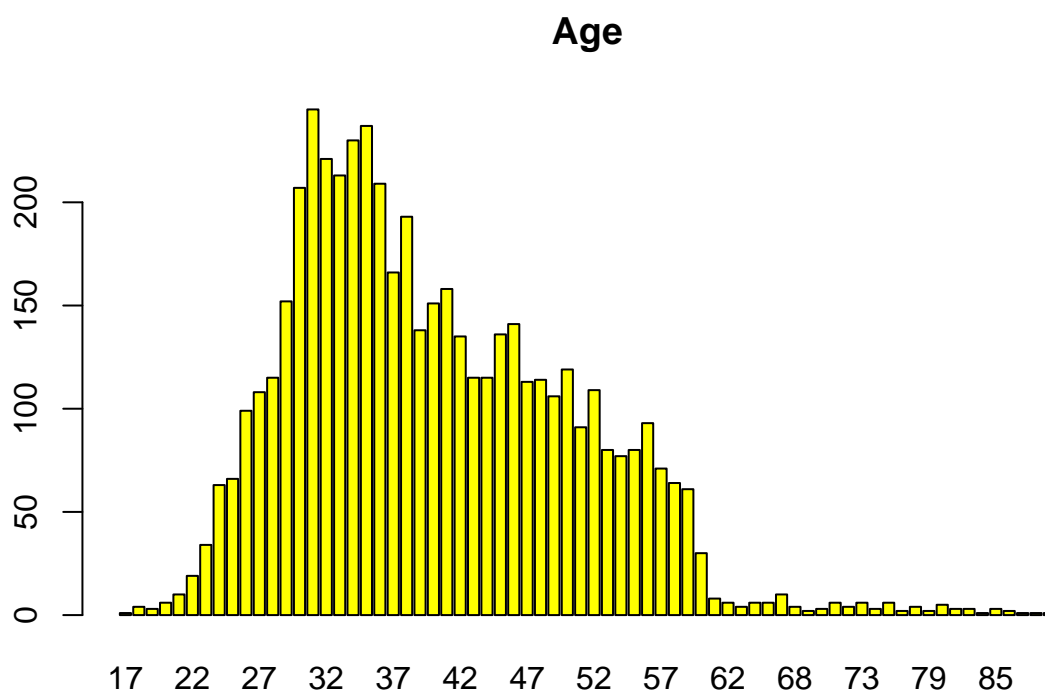
```

summary(df$age)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    17.00  32.00   38.00   39.97  47.00   92.00

barplot(table(df$age), main= "Age",col="yellow")

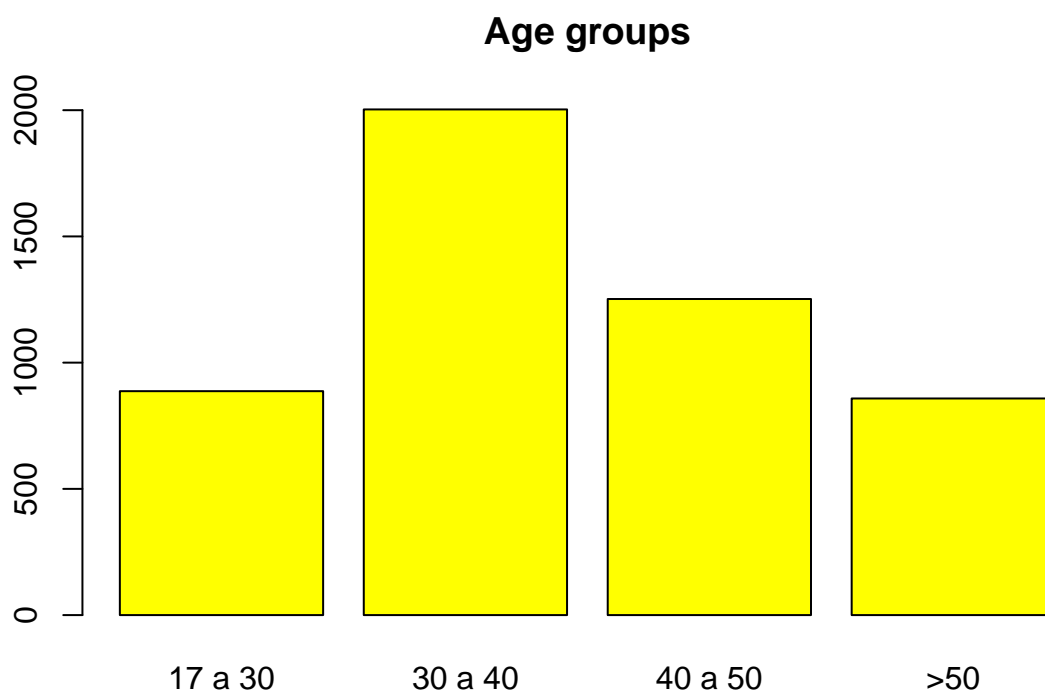
```



```
summary(df$f.age)
```

```
## f.age-[17,30] f.age-(30,40] f.age-(40,50] f.age-(50,95]
##          887          2003          1252          858
```

```
barplot(table(df$f.age), main="Age groups",names.arg=c("17 a 30","30 a 40","40 a 50", ">50"),col="yellow")
```



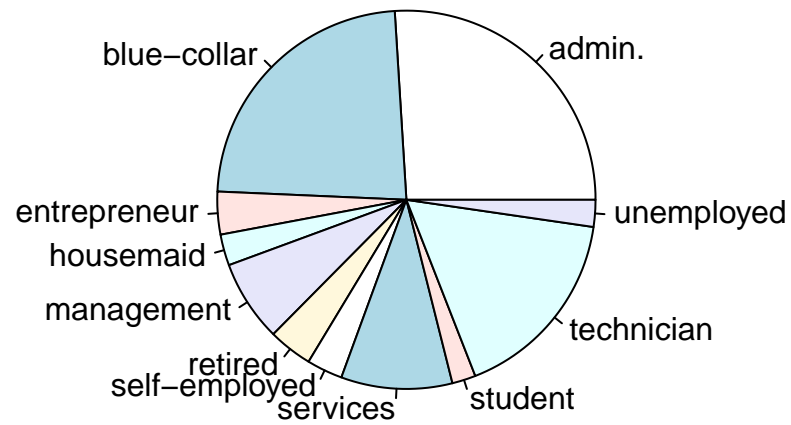
2.4.2 Job

```
table(df$job)
```

```
##
##      admin.  blue-collar  entrepreneur  housemaid  management
##      1288      1156      181      132      345
##      retired self-employed  services  student  technician
##      187      152      471      100      831
##      unemployed
##      114
```

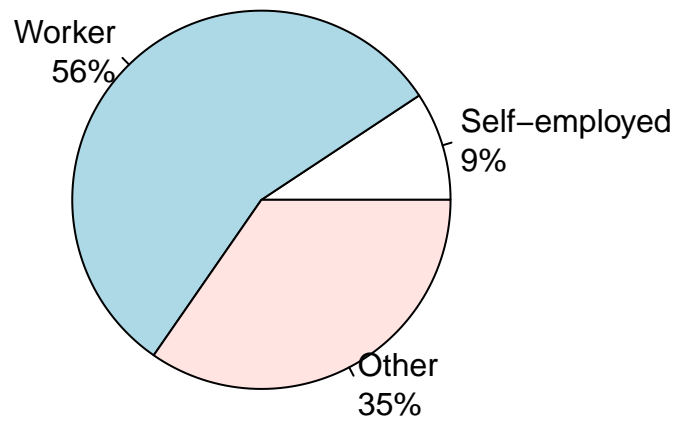
```
pie(table(df$job), main= "Job")
```

Job



```
aux <- table(df$f.jobssituation)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # add % to labels
pie(aux,labels = lbls, main="Job Situation")
```

Job Situation



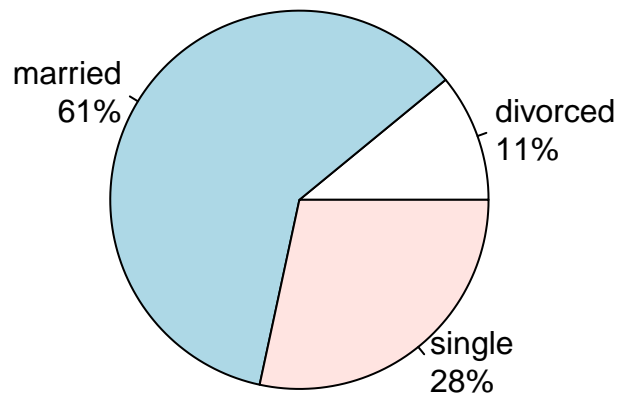
2.4.3 Marital

```
table(df$marital)
```

```
##  
## divorced married single  
##      546      3029      1416
```

```
aux <- table(df$marital)  
pct <- round(aux/sum(aux)*100)  
lbls <- paste(names(aux), "\n", pct, sep="")  
lbls <- paste(lbls,"%",sep="") # add % to labels  
pie(aux,labels = lbls, main="Marital Situation")
```

Marital Situation



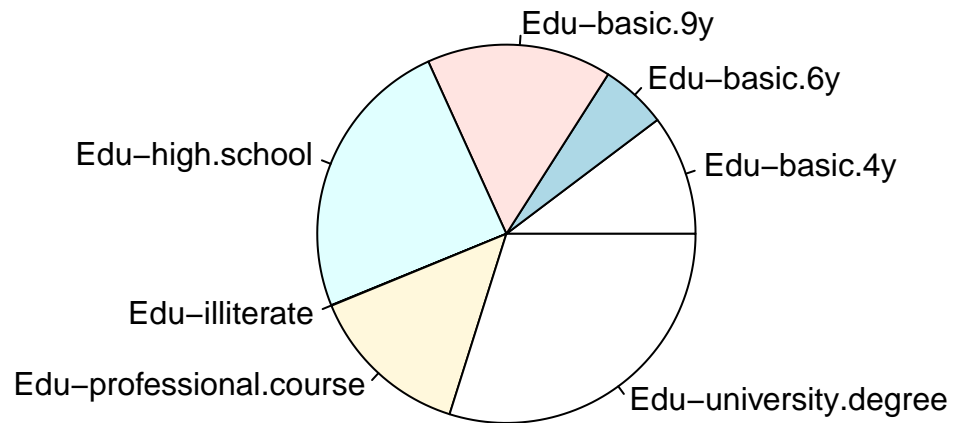
2.4.4 Education

```
table(df$education)
```

```
##  
##      Edu-basic.4y      Edu-basic.6y      Edu-basic.9y  
##           493           272           758  
##      Edu-high.school      Edu-illiterate      Edu-professional.course  
##           1169              2           668  
##      Edu-university.degree  
##           1431
```

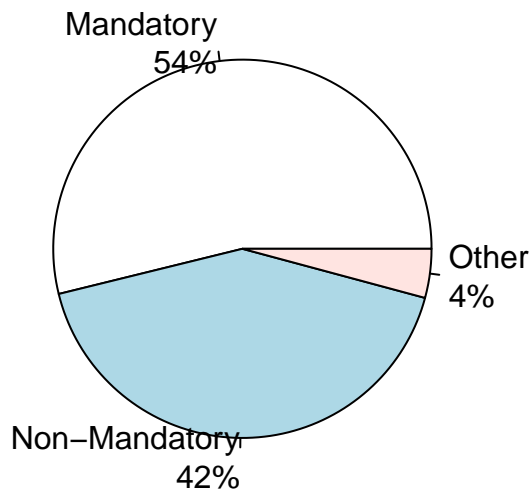
```
pie(table(df$education), main= "Education")
```

Education



```
aux <- table(df$f.education)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # add % to labels
pie(aux,labels = lbls, main="Education Level")
```


Education Level



2.4.5 Default-Housing-Loan

```
table(df$default)
```

```
##  
##   no  
## 3939
```

```
table(df$housing)
```

```
##  
##   no  yes  
## 2226 2662
```

```
table(df$loan)
```

```
##  
##   no  yes  
## 4138  750
```

```
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
```

```
##
```

```
##   mpg
```

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

```
aux <- table(df$loan)
```

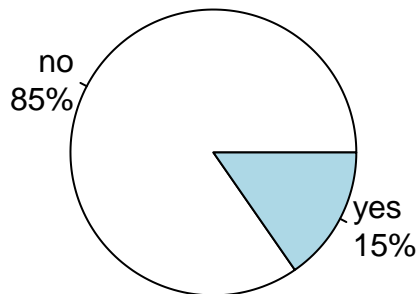
```
pct <- round(aux/sum(aux)*100)
```

```

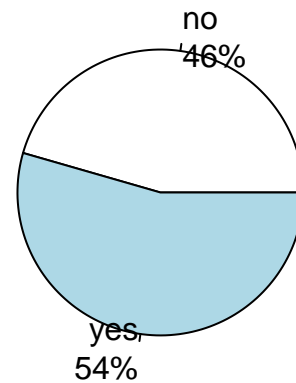
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls, "%", sep="") # ad % to labels
pie(aux, labels = lbls, main="Personal Loan")
aux <- table(df$housing)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls, "%", sep="") # ad % to labels
pie(aux, labels = lbls, main="Housing Loan")

```

Personal Loan



Housing Loan



```

# Retornar l'attach a l'estat predeterminat
attach(mtcars)

```

```

## The following objects are masked from mtcars (pos = 3):
##
##      am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
##
## The following object is masked from package:ggplot2:
##
##      mpg

```

```

par(mfrow=c(1,1))

```

Com es pot veure no hem el gràfic de deutes, ja que el 100% d'individus que han contestat a l'enquesta no en tenien.

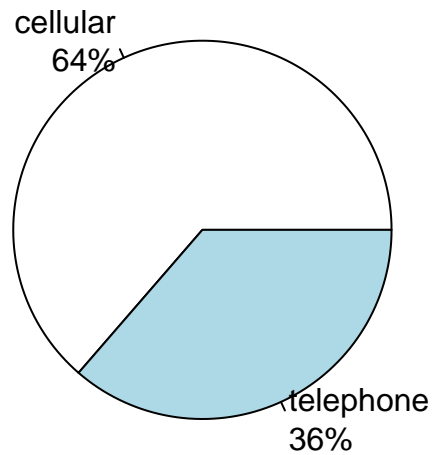
2.4.6 Contact Device

```
table(df$contact)

##
##  cellular telephone
##      3182      1818

aux <- table(df$contact)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # add % to labels
pie(aux,labels = lbls, main="Contact Device")
```

Contact Device



2.4.7 Date - Month and season

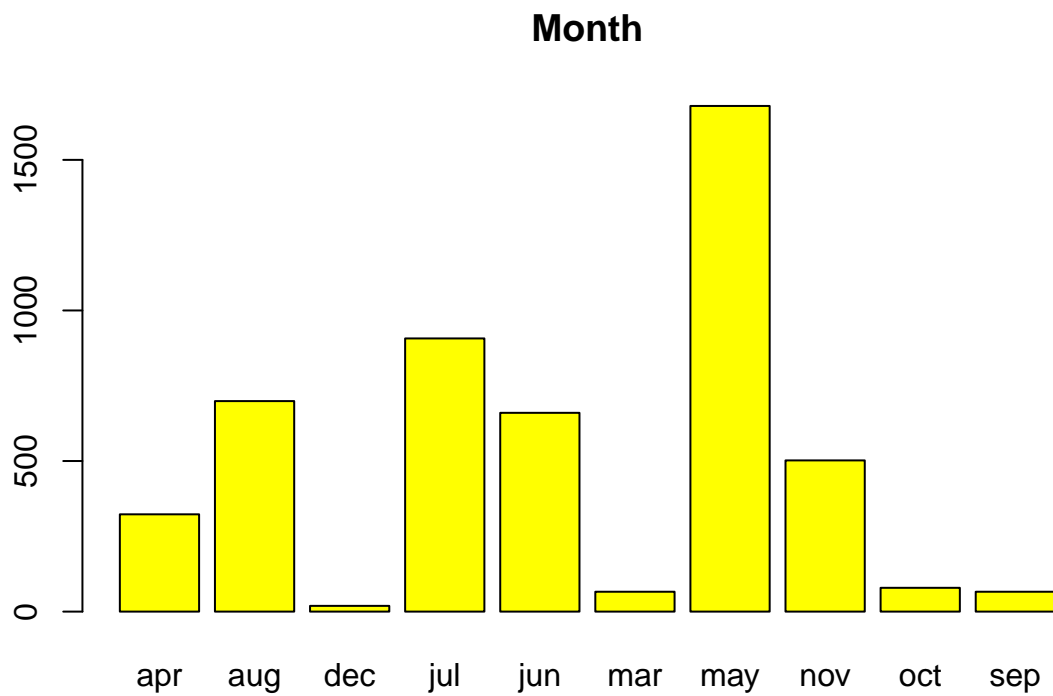
```
table(df$month)

##
##  apr  aug  dec  jul  jun  mar  may  nov  oct  sep
##  323  699   19  907  660   66 1679  502   79   66

table(df$f.season)

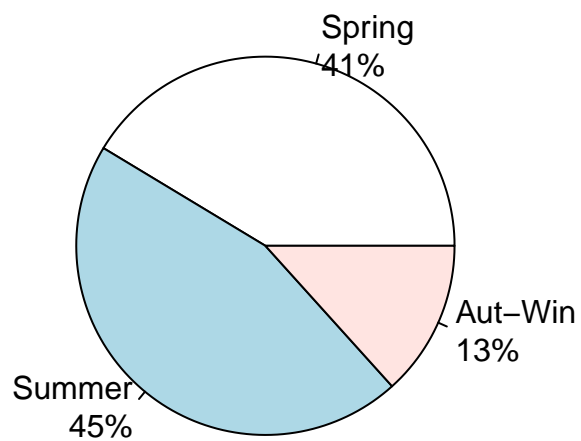
##
##  Spring  Summer Aut-Win
##    2068    2266     666
```

```
barplot(table(df$month), main= "Month", col="yellow")
```



```
aux <- table(df$f.season)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(aux,labels = lbls,
    main="Season")
```

Season

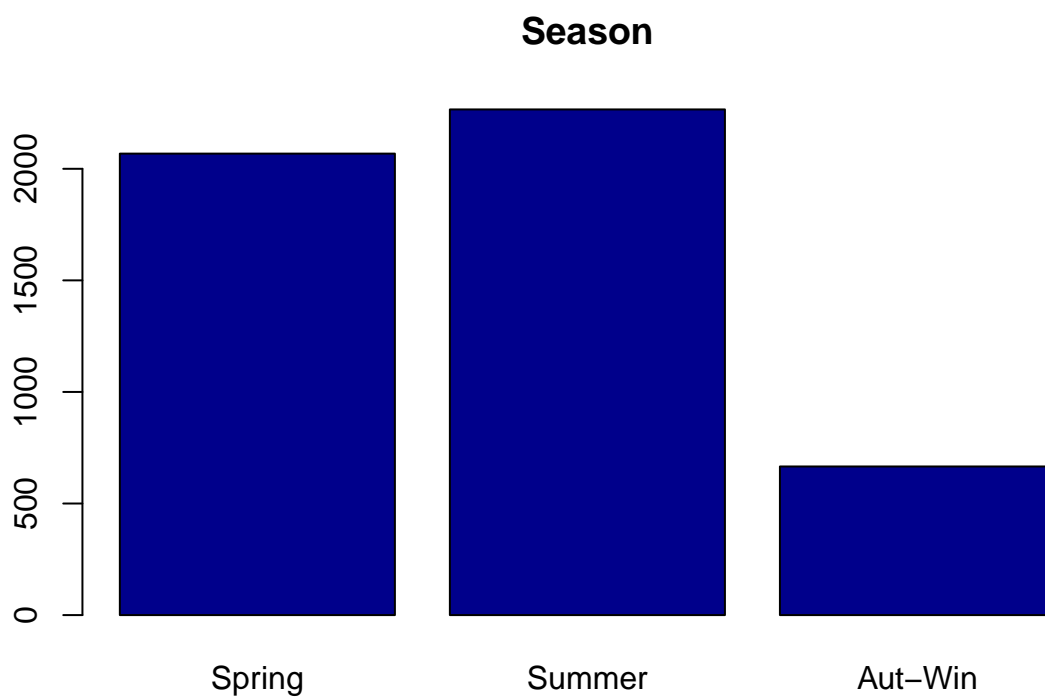


2.4.8 Date - Day of the week

```
table(df$day_of_week)
```

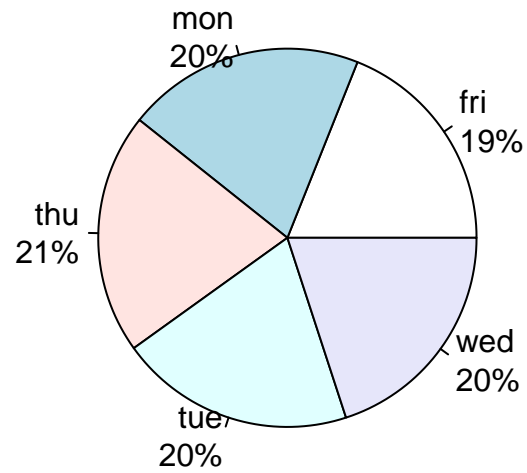
```
##  
##  fri  mon  thu  tue  wed  
##  948 1017 1031 1005  999
```

```
barplot(table(df$f.season), main= "Season", col="darkblue")
```



```
aux <- table(df$day_of_week)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # add % to labels
pie(aux,labels = lbls,
     main="Day of the week")
```

Day of the week



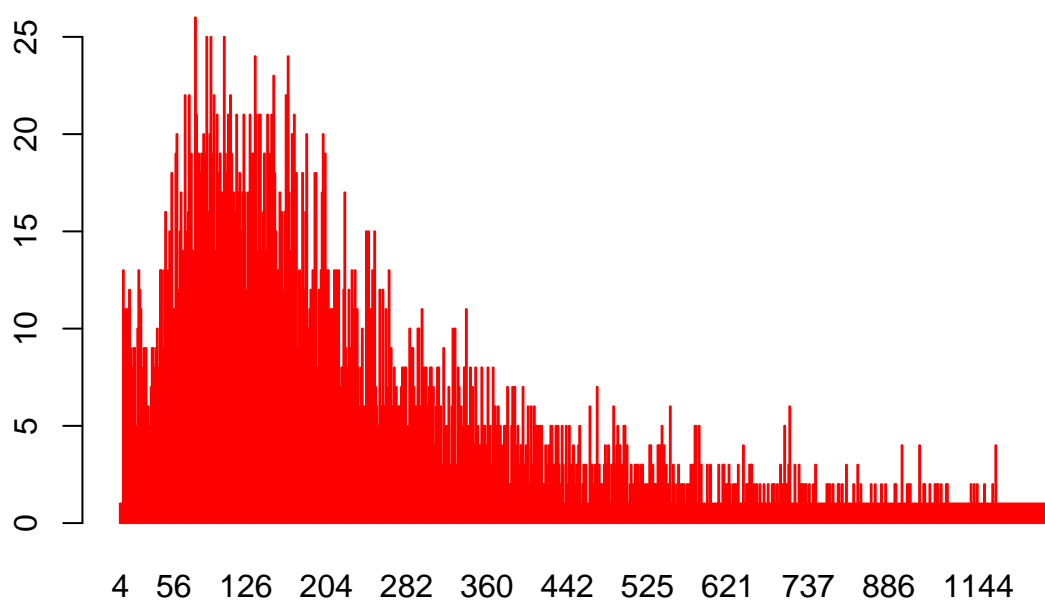
2.4.9 Duration

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       4.0   104.0   181.0   263.7   328.0   3078.0
```

```
barplot(table(df$duration),col="yellow",border="red", main="Call duration")
```

Call duration



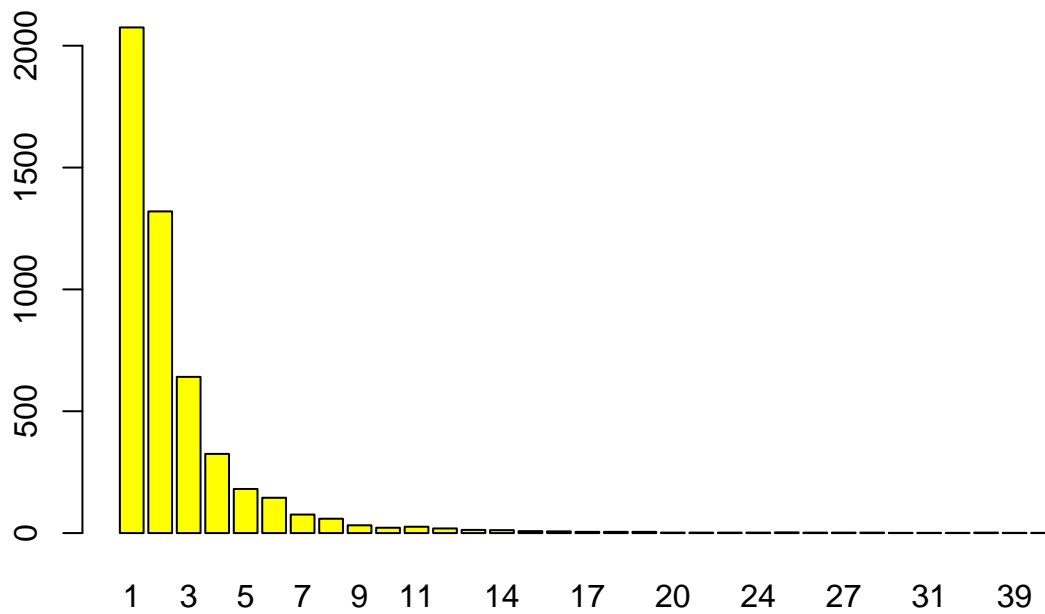
2.4.10 Campaign

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000   1.000   2.000   2.647   3.000  42.000
```

```
barplot(table(df$campaign),col="yellow", main="Number of campaigns previously contacted")
```


Number of campaigns previously contacted



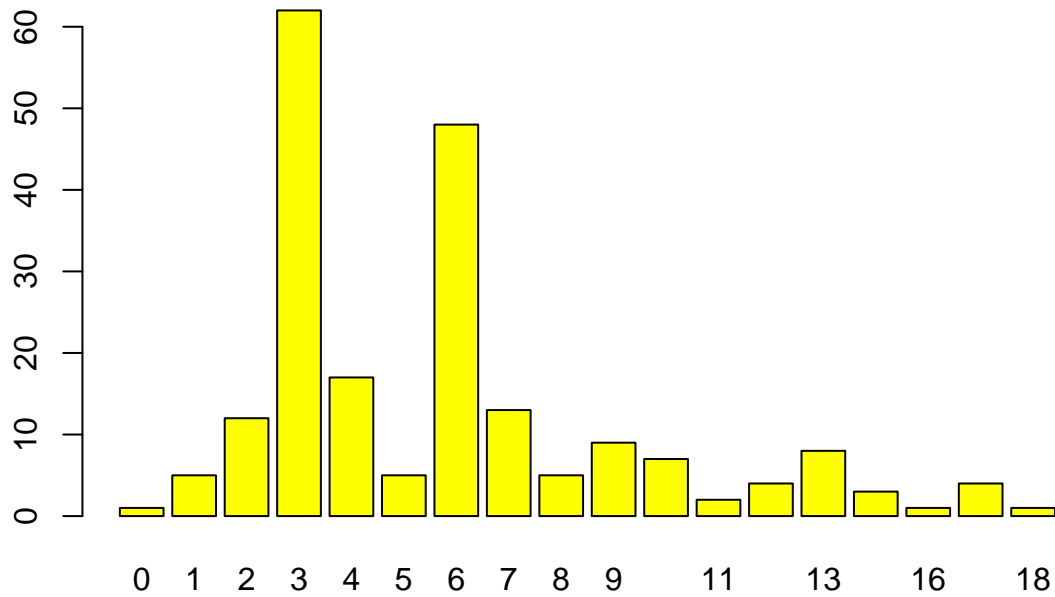
2.4.11 PDays

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
##    0.000   3.000   6.000   5.792   7.000  18.000  4793
```

```
barplot(table(df$pdays),col="yellow", main="Number of days between the last contact")
```

Number of days between the last contact

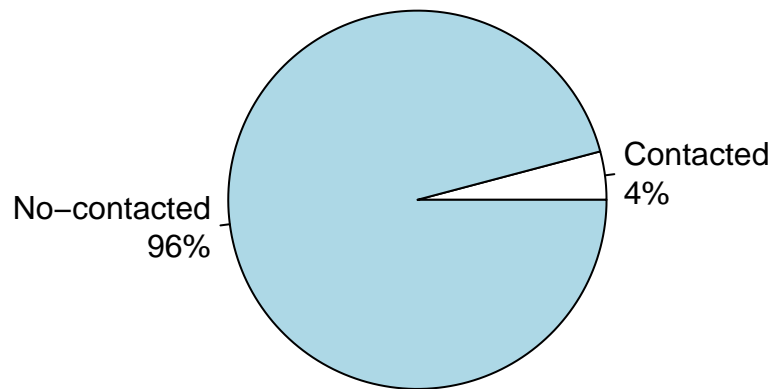


```
table(df$f.prev_contacted)
```

```
##  
##      Contacted No-contacted  
##           206           4794
```

```
aux <- table(df$f.prev_contacted)  
pct <- round(aux/sum(aux)*100)  
lbls <- paste(names(aux), "\n", pct, sep="")  
lbls <- paste(lbls,"%",sep="") # ad % to labels  
pie(aux,labels = lbls,  
     main="Was previously contacted?")
```

Was previously contacted?



2.4.12 Previously

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

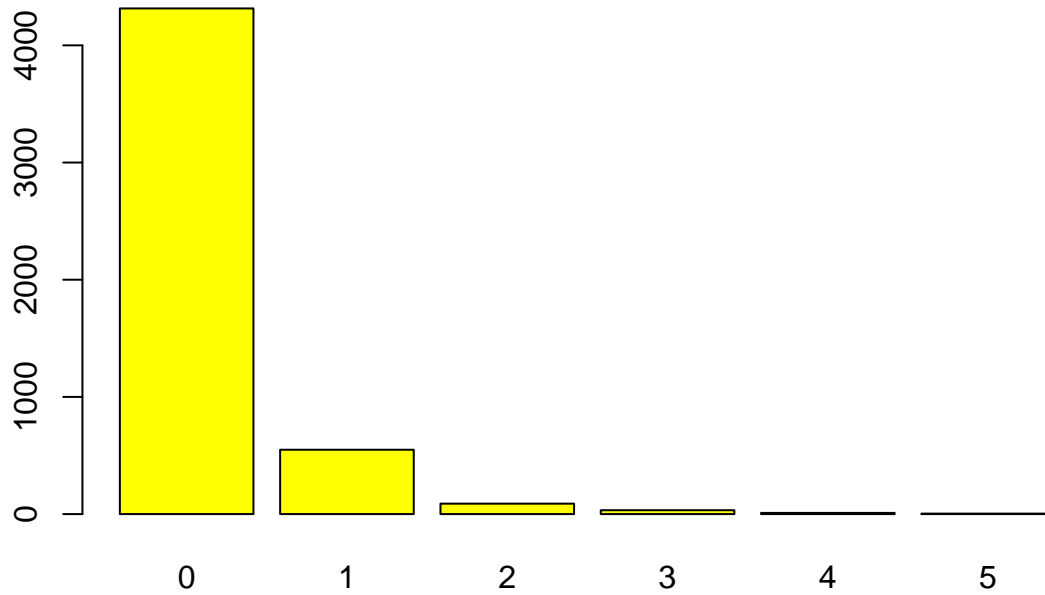
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000  0.0000  0.0000  0.1772  0.0000  5.0000
```

```
table(df$previous)
```

```
##
##      0      1      2      3      4      5
## 4315  549   89   33   10    4
```

```
barplot(table(df$previous),col="yellow", main="Number of contacts before this campaign")
```

Number of contacts before this campaign

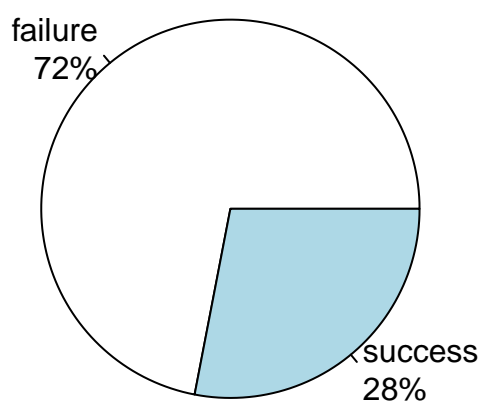


2.4.13 POutcome

```
table(df$poutcome)
```

```
##  
## failure success  
##      493      192  
  
aux <- table(df$poutcome)  
pct <- round(aux/sum(aux)*100)  
lbls <- paste(names(aux), "\n", pct, sep="")  
lbls <- paste(lbls,"%",sep="") # ad % to labels  
pie(aux,labels = lbls,  
     main="Outcome of the previous marketing campaign")
```

Outcome of the previous marketing campaign



2.4.14 Y

```
table(df$y)
```

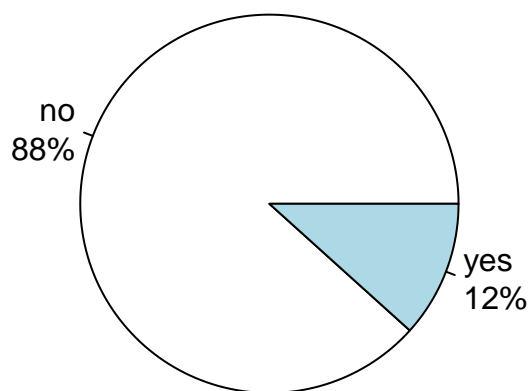
```
##
```

```
## no yes
```

```
## 4416 584
```

```
aux <- table(df$y)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(aux,labels = lbls,
     main="Binary target")
```

Binary target



3 Data Quality Report

3.1 Variables

3.1.1 Missing Values

```
vmiss<-rep(0,nrow(df))

nInitialVariables<- 21
nmiss<-rep(0,nInitialVariables)

initialVariables <- 0:21
names(nmiss) <- names(df[initialVariables])
names(df[initialVariables])

## [1] "age"          "job"          "marital"      "education"
## [5] "default"      "housing"      "loan"         "contact"
## [9] "month"        "day_of_week"  "duration"     "campaign"
## [13] "pdays"       "previous"     "poutcome"     "emp.var.rate"
## [17] "cons.price.idx" "cons.conf.idx" "euribor3m"    "nr.employed"
## [21] "y"

for(j in 1:21) {
  nmiss[j]<-nmiss[j]+sum(is.na(df[,j]))
}
```

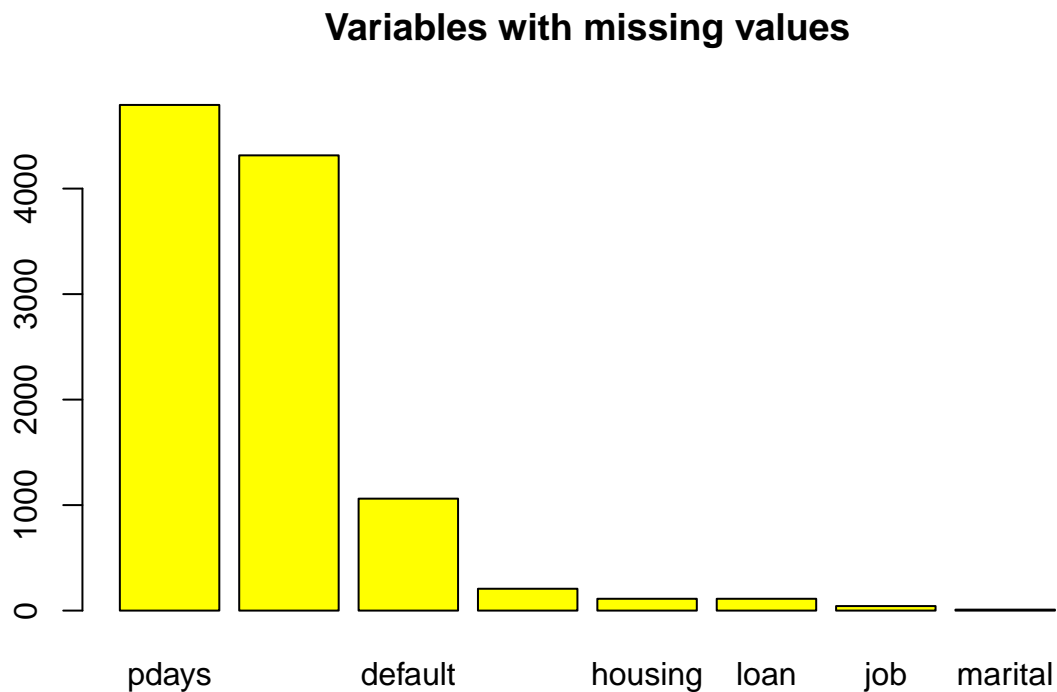
```

nmiss_aux <- nmiss[ nmiss != 0 ]
nmiss_aux <- sort(nmiss_aux, decreasing = TRUE)

table(nmiss_aux)

## nmiss_aux
##      9  43 112 207 1061 4315 4793
##      1   1   2   1   1   1   1
barplot(nmiss_aux, col="yellow", main="Variables with missing values");

```



Al barplot sols apareixen les variables amb dades mancants.

3.1.2 Errors

```

verrs<-rep(0, nrow(df))

nInitialVariables<- 21
nerrs<-rep(0, nInitialVariables)

initialVariables <- 0:21
names(nerrs) <- names(df[initialVariables])
names(df[initialVariables])

## [1] "age"          "job"          "marital"      "education"
## [5] "default"      "housing"      "loan"         "contact"
## [9] "month"       "day_of_week" "duration"     "campaign"

```

```
## [13] "pdays"          "previous"        "poutcome"        "emp.var.rate"
## [17] "cons.price.idx" "cons.conf.idx"  "euribor3m"       "nr.employed"
## [21] "y"
```

3.1.2.1 Job

```
v<-c("admin.", "blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "se.")
llista<-which(!is.element(df[2], v));
verrs[llista]<-verrs[llista]+1
nerrs[2]<-nerrs[2]+sum(!is.element(df[,2], v))
```

3.1.2.2 Marital

```
v<-c("divorced", "married", "single", NA)
llista<-which(!is.element(df[3], v));
verrs[llista]<-verrs[llista]+1
nerrs[3]<-nerrs[3]+sum(!is.element(df[,3], v))
```

3.1.2.3 Education

```
v<-c("Edu-basic.4y", "Edu-basic.6y", "Edu-basic.9y", "Edu-high.school", "Edu-illiterate", "Edu-professional")
llista<-which(!is.element(df[4], v));
verrs[llista]<-verrs[llista]+1
nerrs[4]<-nerrs[4]+sum(!is.element(df[,4], v))
```

3.1.2.4 Default

```
v<-c("no", "yes", NA)
llista<-which(!is.element(df[5], v));
verrs[llista]<-verrs[llista]+1
nerrs[5]<-nerrs[5]+sum(!is.element(df[,5], v))
```

3.1.2.5 Housing

```
v<-c("no", "yes", NA)
llista<-which(!is.element(df[6], v));
verrs[llista]<-verrs[llista]+1
nerrs[6]<-nerrs[6]+sum(!is.element(df[,6], v))
```

3.1.2.6 Loan

```
v<-c("no", "yes", NA)
llista<-which(!is.element(df[7], v));
verrs[llista]<-verrs[llista]+1
nerrs[7]<-nerrs[7]+sum(!is.element(df[,7], v))
```

3.1.2.7 Contact

```
v<-c("cellular", "telephone", NA)
llista<-which(!is.element(df[8], v));
verrs[llista]<-verrs[llista]+1
nerrs[8]<-nerrs[8]+sum(!is.element(df[,8], v))
```


3.1.2.8 Month

```
v<-c("apr", "aug", "dec", "jul", "jun", "mar", "may", "nov", "oct", "sep", "jan", "feb", NA)
llista<-which(!is.element(df[9], v));
verrs[llista]<-verrs[llista]+1
nerrs[9]<-nerrs[9]+sum(!is.element(df[,9], v))
```

3.1.2.9 Day of week

```
v<-c("mon", "tue", "wed", "thu", "fri", NA)
llista<-which(!is.element(df[10], v));
verrs[llista]<-verrs[llista]+1
nerrs[10]<-nerrs[10]+sum(!is.element(df[,10], v))
```

3.1.2.10 Poutcome

```
v<-c("failure", "success", NA)
llista<-which(!is.element(df[,15], v));
verrs[llista]<-verrs[llista]+1
nerrs[15]<-nerrs[15]+sum(!is.element(df[,15], v))
```

3.1.2.11 Y

```
v<-c("yes", "no", NA)
llista<-which(!is.element(df[21], v));
verrs[llista]<-verrs[llista]+1
nerrs[21]<-nerrs[21]+sum(!is.element(df[,21], v))
```

Així els errors queden:

```
nerrs
```

```
##          age          job          marital          education          default
##           0           0           0           0           0
##      housing          loan          contact          month      day_of_week
##           0           0           0           0           0
##      duration          campaign          pdays          previous          poutcome
##           0           0           0           0           0
## emp.var.rate cons.price.idx cons.conf.idx      euribor3m      nr.employed
##           0           0           0           0           0
##           y
##           0
```

3.1.2.12 Outliers

```
vout<-rep(0,nrow(df))

nInitialVariables<- 21
nout<-rep(0,nInitialVariables)

initialVariables <- 0:21
names(nout) <- names(df[initialVariables])
names(df[initialVariables])
```

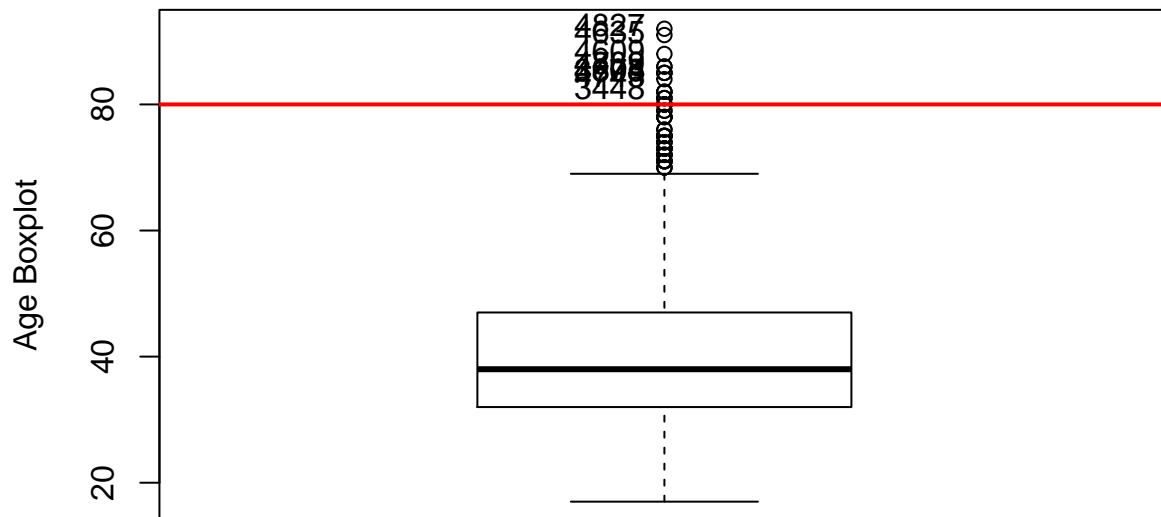
```
## [1] "age"          "job"          "marital"      "education"
## [5] "default"      "housing"      "loan"         "contact"
## [9] "month"       "day_of_week" "duration"     "campaign"
```

```
## [13] "pdays"          "previous"        "poutcome"        "emp.var.rate"
## [17] "cons.price.idx" "cons.conf.idx"   "euribor3m"        "nr.employed"
## [21] "y"
```

3.1.2.12.1 Age

```
Boxplot(df$age, ylab = "Age Boxplot")
```

```
## [1] 4827 4635 4609 4732 4869 3675 4803 4804 4743 3448
sout <- 80
abline(h=sout,col="red",lwd=2)
```



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

```
## [1] 15
```

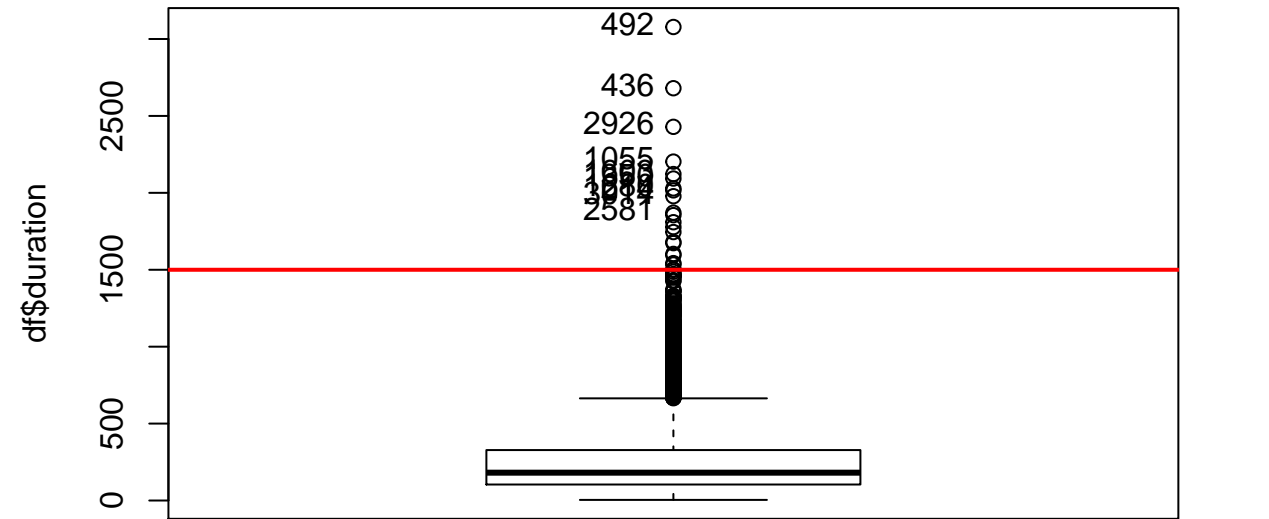
```
df$age[outliers] <- NA;
if(length(outliers)>0){
vout[outliers]<-vout[outliers]+1
nout["age"]<-length(outliers)}
```

3.1.2.12.2 Duration

```
Boxplot(df$duration)
```

```
## [1] 492 436 2926 1055 1603 1350 1680 214 3014 2581
```

```
sout <- 1500
abline(h=sout,col="red",lwd=2)
```



```
outliers<-which(df$duration>sout);length(outliers);
```

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

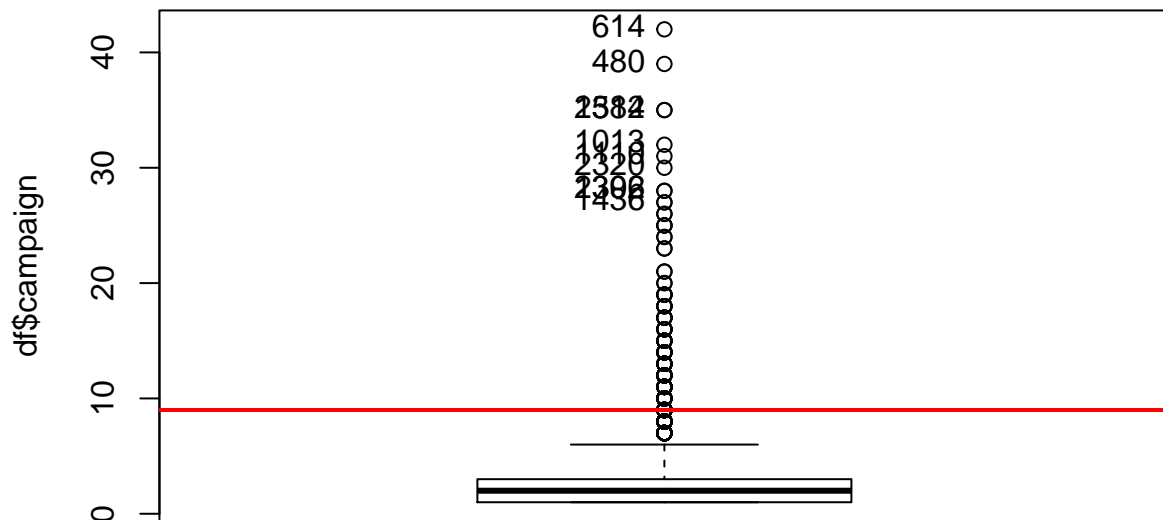
```
if(length(outliers)>0){
  vout[outliers]<-vout[outliers]+1
  nout["duration"]<-length(outliers)}
```

3.1.2.12.3 Campaign

```
Boxplot(df$campaign)
```

```
## [1] 614 480 1584 2312 1013 1110 2320 1392 2306 1436
```

```
sout <- 9
abline(h=sout,col="red",lwd=2)
```



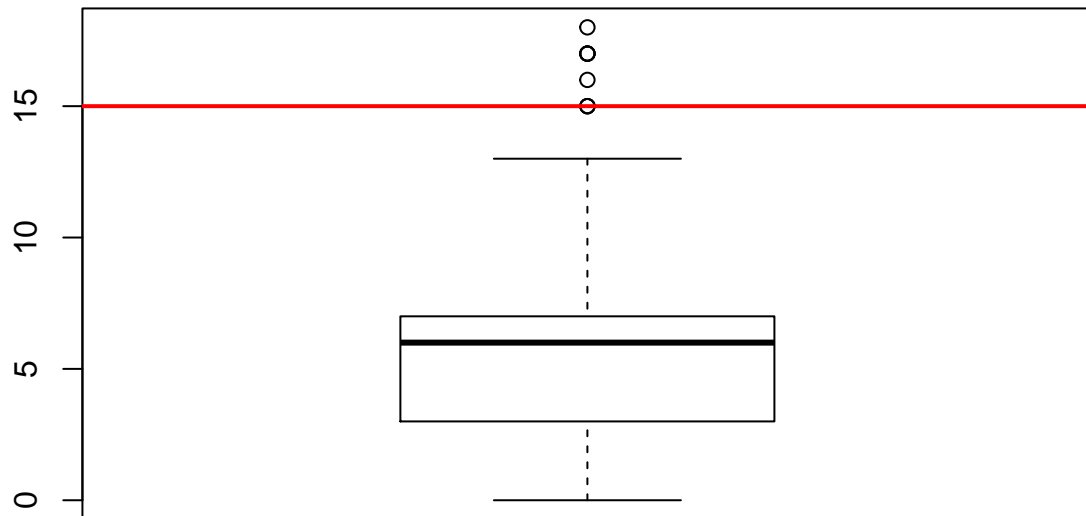
```
outliers<-which(df$campaign>sout);length(outliers);
```

```
## [1] 146
```

```
df$campaign[outliers] <- NA;
if(length(outliers)>0){
  vout[outliers]<-vout[outliers]+1
  nout["campaign"]<-length(outliers)}
```

3.1.2.12.4 Pdays

```
boxplot(df$pdays);
sout <- 15;
abline(h=sout,col="red",lwd=2);
```



```
outliers<-which(df$pdays> sout); length(outliers);
```

```
## [1] 6
```

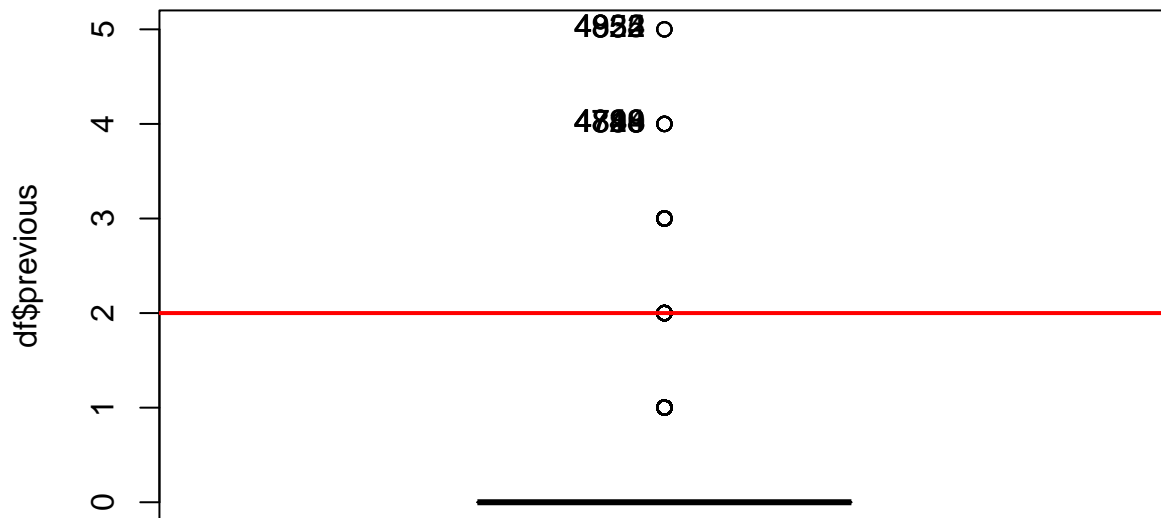
```
df$pdays[outliers] <- NA;
if(length(outliers)>0){
vout[outliers]<-vout[outliers]+1
nout["pdays"]<-length(outliers)}
```

3.1.2.12.5 Previous

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

```
## [1] 4822 4835 4952 4954 4719 4783 4790 4828 4844 4848
```

```
sout <- 2
abline(h=sout,col="red",lwd=2)
```



```
outliers<-which(df$previous> sout);
df$previous[outliers] <- NA;
length(outliers);
```

```
## [1] 47
```

```
if(length(outliers)>0){
vout[outliers]<-vout[outliers]+1
nout["previous"]<-length(outliers)}
```

Així els outliers queden:

```
nout
```

```
##      age      job      marital      education      default
##      15        0          0          0          0
##      housing    loan      contact      month      day_of_week
##      0          0          0          0          0
##      duration    campaign      pdays      previous      poutcome
##      21         146          6          47          0
##      emp.var.rate cons.price.idx cons.conf.idx      euribor3m      nr.employed
##      0            0            0          0          0
##      y
##      0
```

3.1.2.13 Rank Variables

```
miss <- sort(nmiss, decreasing = TRUE)
miss
```

```
##      pdays      poutcome      default      education      housing
##      4793        4315        1061        207        112
##      loan        job        marital      age        contact
##      112         43         9         0         0
##      month      day_of_week      duration      campaign      previous
##      0          0          0          0          0
##      emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed
##      0          0          0          0          0
##      y
##      0
```

```
err <- sort(nerrs, decreasing = TRUE)
err
```

```
##      age      job      marital      education      default
##      0        0        0        0        0
##      housing      loan      contact      month      day_of_week
##      0          0        0        0        0
##      duration      campaign      pdays      previous      poutcome
##      0          0        0        0        0
##      emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed
##      0          0        0        0        0
##      y
##      0
```

```
miss <- sort(nmiss, decreasing = TRUE)
miss
```

```
##      pdays      poutcome      default      education      housing
##      4793        4315        1061        207        112
##      loan        job        marital      age        contact
##      112         43         9         0         0
##      month      day_of_week      duration      campaign      previous
##      0          0          0          0          0
##      emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed
##      0          0          0          0          0
##      y
##      0
```

```
out <- sort(nout, decreasing = TRUE)
out
```

```
##      campaign      previous      duration      age      pdays
##      146         47        21        15        6
##      job        marital      education      default      housing
##      0          0        0        0        0
##      loan        contact      month      day_of_week      poutcome
##      0          0        0        0        0
##      emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed
##      0          0        0        0        0
##      y
##      0
```

```
ranking <- nmiss + nerrs + nout;
ranking <- sort(ranking, decreasing = TRUE);
ranking
```

```
##      pdays      poutcome      default      education      campaign
##      4799        4315        1061        207        146
##      housing      loan      previous      job      duration
##      112         112         47        43        21
##      age      marital      contact      month      day_of_week
##      15          9          0          0          0
##      emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed
##      0            0            0            0            0
##      y
##      0
```

3.2 Individual

```
vmis<-rep(0, nrow(df))
nmis<-rep(0, ncol(df))
for(i in 1:nrow(df)) {
  vmis[i]<-vmis[i]+sum(is.na(df[i,]))
}
### Create variable adding the total number missing values, outliers and errors
df$outliers<-vout
df$errors<-verrs
df$missings<-vmis
```

4 Correlation

```
##Outliers
condes(df, num.var = 35)
```

```
## $quanti
##      correlation      p.value
## cons.price.idx  0.09277707 4.935483e-11
## duration        0.09217400 6.578690e-11
## missings         0.08750383 5.725813e-10
## cons.conf.idx    0.04206987 2.926427e-03
## campaign         -0.03485868 1.515141e-02
## nr.employed      -0.06723109 1.953740e-06
##
## $quali
##      R2      p.value
## f.prev_contacted 0.0227976441 6.937572e-27
## month            0.0177019926 2.370453e-15
## f.month          0.0177019926 2.370453e-15
## poutcome         0.0126995605 1.354103e-14
## y                0.0058927611 5.488243e-08
## job              0.0078449930 4.672357e-05
## f.job            0.0078449930 4.672357e-05
## f.jobsituation   0.0036459260 1.088007e-04
## f.season         0.0026809265 1.221971e-03
```


## education	0.0035745603	1.249920e-02
## contact	0.0007892906	4.698187e-02
## f.contact	0.0007892906	4.698187e-02
## loan	0.0012042854	4.925618e-02
## f.loan	0.0012042854	4.925618e-02
##		
## \$category		
##	Estimate	p.value
## Contacted	0.081812419	6.937572e-27
## success	0.072116782	5.134633e-14
## Month-sep	0.132112071	3.375808e-10
## sep	0.132112071	3.375808e-10
## yes	0.025739217	5.488243e-08
## Other	0.014226777	4.381288e-05
## Job-retired	0.049221530	1.038034e-04
## retired	0.049221530	1.038034e-04
## Month-oct	0.059231366	1.236697e-04
## oct	0.059231366	1.236697e-04
## Job-student	0.062269658	6.151821e-04
## student	0.062269658	6.151821e-04
## Summer	0.009445982	1.934820e-03
## Edu-basic.4y	0.027994349	1.707508e-02
## f.no	0.016447061	1.881531e-02
## no	0.016447061	1.881531e-02
## f.single	0.020057898	3.525089e-02
## single	0.020057898	3.525089e-02
## f.married	0.004782537	3.900623e-02
## married	0.004782537	3.900623e-02
## f.age-(50,95]	0.011668659	4.206059e-02
## f.telephone	0.006289677	4.698187e-02
## telephone	0.006289677	4.698187e-02
## f.cellular	-0.006289677	4.698187e-02
## cellular	-0.006289677	4.698187e-02
## failure	-0.022404431	3.040848e-02
## Edu-basic.6y	-0.022588815	2.423028e-02
## Job-services	-0.032252635	2.269021e-02
## services	-0.032252635	2.269021e-02
## Month-jul	-0.017164598	1.433353e-02
## jul	-0.017164598	1.433353e-02
## Month-jun	-0.013342474	1.180245e-02
## jun	-0.013342474	1.180245e-02
## Job-blue-collar	-0.024858369	1.096798e-02
## blue-collar	-0.024858369	1.096798e-02
## Edu-basic.9y	-0.013266682	7.404904e-03
## Month-nov	-0.064072886	6.542548e-04
## nov	-0.064072886	6.542548e-04
## Month-may	-0.048442732	3.132088e-04
## may	-0.048442732	3.132088e-04
## Spring	-0.014074703	2.863111e-04
## Worker	-0.013828160	4.717187e-05
## no	-0.025739217	5.488243e-08
## NA	-0.049712351	1.179041e-09
## No-contacted	-0.081812419	6.937572e-27

```
##Errors
##condes(df, num.var = 36)
##Missings
condes(df, num.var = 37)
```

```
## $quanti
##          correlation      p.value
## emp.var.rate    0.32059379 6.915468e-120
## euribor3m       0.31925495 7.539951e-119
## nr.employed     0.31676931 6.159986e-117
## cons.price.idx  0.25210903 2.394819e-73
## age             0.13225565 6.805599e-21
## outliers        0.08750383 5.725813e-10
## cons.conf.idx   0.04099752 3.738104e-03
## campaign        0.03483158 1.523054e-02
## previous        -0.42315303 2.030358e-214
##
## $quali
##          R2      p.value
## default    0.535780204 0.000000e+00
## f.default  0.535780204 0.000000e+00
## loan       0.241643238 7.313059e-301
## f.loan     0.241643238 7.313059e-301
## housing    0.241517653 1.106051e-300
## f.housing  0.241517653 1.106051e-300
## poutcome   0.204236107 1.286462e-248
## f.prev_contacted 0.113007932 2.380717e-132
## education  0.080884953 6.578439e-87
## job        0.072449068 9.771474e-74
## f.job      0.072449068 9.771474e-74
## f.education 0.061370385 1.891864e-69
## contact    0.046652810 7.324880e-54
## f.contact  0.046652810 7.324880e-54
## month      0.049126070 4.716899e-49
## f.month    0.049126070 4.716899e-49
## y          0.029025116 7.096717e-34
## f.season   0.028568002 3.547259e-32
## f.age      0.022838134 7.424390e-25
## marital    0.022022645 5.860022e-24
## f.marital  0.022022645 5.860022e-24
## f.jobsituation 0.007767691 3.455343e-09
##
## $category
##          Estimate      p.value
## NA          1.087223059 0.000000e+00
## NA          1.087223059 0.000000e+00
## NA          2.698104360 2.243802e-302
## NA          2.688966220 2.243802e-302
## NA          2.698104360 2.243802e-302
## NA          2.688966220 2.243802e-302
## NA          1.158397313 1.944798e-227
## No-contacted 1.027183048 2.380717e-132
## NA          1.151292413 5.804433e-57
## Other        0.903536808 4.449317e-56
```

## f.telephone	0.272689164	7.324880e-54
## telephone	0.272689164	7.324880e-54
## NA	2.521893692	1.678143e-49
## NA	2.521893692	1.678143e-49
## no	0.322138053	7.096717e-34
## Job-blue-collar	0.098621784	1.020782e-18
## blue-collar	0.098621784	1.020782e-18
## Summer	0.274890250	5.082947e-18
## Edu-basic.4y	0.249890173	1.760364e-14
## NA	2.130389105	5.531075e-12
## NA	2.130389105	5.531075e-12
## f.age-(40,50]	0.184697514	9.683462e-11
## Month-jun	0.626592566	1.109488e-10
## jun	0.626592566	1.109488e-10
## f.age-(50,95]	0.197797300	5.304722e-08
## Worker	0.084409151	8.550677e-08
## Edu-basic.6y	0.209956096	5.955419e-07
## Month-jul	0.515360065	2.256065e-06
## jul	0.515360065	2.256065e-06
## Job-housemaid	0.177813354	6.666548e-04
## housemaid	0.177813354	6.666548e-04
## Month-may	0.413964198	3.280969e-03
## may	0.413964198	3.280969e-03
## Job-management	-0.307496264	3.868777e-02
## management	-0.307496264	3.868777e-02
## Job-technician	-0.266257317	2.039955e-02
## technician	-0.266257317	2.039955e-02
## Month-dec	-0.308176174	1.922055e-02
## dec	-0.308176174	1.922055e-02
## f.yes	-1.368490878	2.567887e-03
## yes	-1.368490878	2.567887e-03
## f.age-(30,40]	-0.087930286	1.231375e-04
## Edu-high.school	-0.270472752	7.333131e-05
## Mandatory	-0.310407397	1.913837e-05
## f.no	-1.341862621	4.952234e-06
## no	-1.341862621	4.952234e-06
## Job-student	-0.747641192	2.037418e-06
## student	-0.747641192	2.037418e-06
## Month-apr	-0.014058527	4.532814e-08
## apr	-0.014058527	4.532814e-08
## f.yes	-1.347103599	7.963467e-09
## yes	-1.347103599	7.963467e-09
## Month-oct	-0.439422010	7.506767e-09
## oct	-0.439422010	7.506767e-09
## Month-mar	-0.521892282	5.489624e-09
## mar	-0.521892282	5.489624e-09
## Other	-0.143651719	4.699821e-10
## Month-sep	-0.643104404	2.884680e-11
## sep	-0.643104404	2.884680e-11
## Month-nov	-0.008008424	7.921110e-12
## nov	-0.008008424	7.921110e-12
## f.married	-0.550730444	3.578250e-13
## married	-0.550730444	3.578250e-13
## f.single	-0.855957411	5.880094e-14

```
## single -0.855957411 5.880094e-14
## f.age-[17,30] -0.294564528 6.798179e-15
## Job-admin. -0.404908272 4.641734e-15
## admin. -0.404908272 4.641734e-15
## Edu-university.degree -0.426412445 4.461532e-25
## Non-Mandatory -0.593129411 1.085083e-26
## f.no -1.329613482 2.633646e-27
## no -1.329613482 2.633646e-27
## Aut-Win -0.361408450 1.598668e-27
## yes -0.322138053 7.096717e-34
## f.cellular -0.272689164 7.324880e-54
## cellular -0.272689164 7.324880e-54
## failure -0.104236689 2.835968e-95
## success -1.054160624 5.664133e-127
## Contacted -1.027183048 2.380717e-132
## f.no -1.087223059 0.000000e+00
## no -1.087223059 0.000000e+00
```

```
aggregate(df$missings, by=list(df$f.age), FUN=mean)
```

```
##          Group.1          x
## 1 f.age-[17,30] 2.153326
## 2 f.age-(30,40] 2.359960
## 3 f.age-(40,50] 2.632588
## 4 f.age-(50,95] 2.645688
```

```
aggregate(df$outliers, by=list(df$f.age), FUN=mean)
```

```
##          Group.1          x
## 1 f.age-[17,30] 0.04847802
## 2 f.age-(30,40] 0.04193709
## 3 f.age-(40,50] 0.04472843
## 4 f.age-(50,95] 0.06060606
```

```
aggregate(df$missings, by=list(df$f.jobsituation), FUN=mean)
```

```
##          Group.1          x
## 1 Self-employed 2.496774
## 2          Worker 2.521941
## 3          Other 2.293880
```

```
aggregate(df$outliers, by=list(df$f.jobsituation), FUN=mean)
```

```
##          Group.1          x
## 1 Self-employed 0.04946237
## 2          Worker 0.03603282
## 3          Other 0.06408776
```

```
aggregate(df$missings, by=list(df$f.education), FUN=mean)
```

```
##          Group.1          x
## 1      Mandatory 2.508544
## 2 Non-Mandatory 2.225822
## 3          Other 3.722488
```

```
aggregate(df$outliers, by=list(df$f.education), FUN=mean)
```

```
##          Group.1          x
```

```
## 1      Mandatory 0.04309064
## 2 Non-Mandatory 0.05050024
## 3           Other 0.06220096
```

```
aggregate(df$missings, by=list(df$f.marital), FUN=mean)
```

```
##      Group.1      x
## 1 f.divorced 2.368132
## 2 f.married  2.541103
## 3 f.single  2.235876
```

```
aggregate(df$outliers, by=list(df$f.marital), FUN=mean)
```

```
##      Group.1      x
## 1 f.divorced 0.04945055
## 2 f.married  0.04192803
## 3 f.single  0.05720339
```

5 Imputation

5.1 Numeric Variables

```
#Outliers -> missings
```

```
#Delete duration outliers
```

```
outliers<-which(df$duration>1500);length(outliers);
```

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

```
df <- df[-outliers, ]
```

```
var_num <-names(df)[c(1, 12:14)] ## age,campaign,pdays,previous
length(var_num)
```

```
## [1] 4
```

```
summary(df[,var_num])
```

```
##      age      campaign      pdays      previous
##  Min.   :17.00   Min.   :1.000   Min.   : 0.000   Min.   :0.000
## 1st Qu.:32.00   1st Qu.:1.000   1st Qu.: 3.000   1st Qu.:0.000
##  Median :38.00   Median :2.000   Median : 5.000   Median :0.000
##  Mean   :39.83   Mean   :2.269   Mean   : 5.458   Mean   :0.147
## 3rd Qu.:47.00   3rd Qu.:3.000   3rd Qu.: 7.000   3rd Qu.:0.000
##  Max.   :80.00   Max.   :9.000   Max.   :15.000   Max.   :2.000
## NA's   :15     NA's   :145   NA's   :4778   NA's   :47
```

```
res <- imputePCA(df[,var_num],ncp=2)
summary(res$completeObs)
```

```
##      age      campaign      pdays      previous
##  Min.   :17.00   Min.   :1.000   Min.   : 0.000   Min.   :-0.007828
## 1st Qu.:32.00   1st Qu.:1.000   1st Qu.: 5.338   1st Qu.: 0.000000
##  Median :38.00   Median :2.000   Median : 5.394   Median : 0.000000
##  Mean   :39.83   Mean   :2.271   Mean   : 5.389   Mean   : 0.146942
## 3rd Qu.:47.00   3rd Qu.:3.000   3rd Qu.: 5.433   3rd Qu.: 0.000000
##  Max.   :80.00   Max.   :9.000   Max.   :15.000   Max.   : 2.000000
```

```
#S'han imputat valors negatius a previous, els posem a 0
```

```
sel <- which(res$completeObs[, "previous"] < 0)
res$completeObs[sel, "previous"] <- 0
```

```
df$age <- res$completeObs[, "age"]
df$campaign <- res$completeObs[, "campaign"]
df$pdays <- res$completeObs[, "pdays"]
df$previous <- res$completeObs[, "previous"]
```

5.2 Factors

```
factors <- names(df)[c(24, 28, 29, 31)]; # f.job, f.housing, f.marital, f.loan
summary(df[, factors])
```

```
##           f.job      f.housing      f.marital      f.loan
## Job-admin.      :1283   f.no :2220   f.divorced: 545   f.no :4120
## Job-blue-collar:1154   f.yes:2647   f.married  :3018   f.yes: 747
## Job-technician : 829   NA's : 112   f.single  :1407   NA's : 112
## Job-services   : 469                      NA's      : 9
## Job-management : 343
## (Other)        : 860
## NA's           : 41
```

```
resfact <- imputeMCA(df[, factors], ncp=3);
summary(resfact$completeObs)
```

```
##           f.job      f.housing      f.marital      f.loan
## Job-admin.      :1303   f.no :2223   f.divorced: 545   f.no :4232
## Job-blue-collar:1175   f.yes:2756   f.married  :3027   f.yes: 747
## Job-technician : 829                      f.single  :1407
## Job-services   : 469
## Job-management : 343
## Job-retired    : 186
## (Other)        : 674
```

```
df$f.housing <- resfact$completeObs[, "f.housing"]
df$f.marital <- resfact$completeObs[, "f.marital"]
df$f.loan <- resfact$completeObs[, "f.loan"]
df$f.job <- resfact$completeObs[, "f.job"]
```

Imputem manualment poutcome ja que pensem que els que no han respós a la pregunta molt probablement tampoc havien sigut contactats amb anterioritat:

```
sel <- which(is.na(df$poutcome))
```

```
df$poutcome <- factor(df$poutcome, labels=paste("Pout", sep="-", levels(df$poutcome)))
```

```
table(df$poutcome)
```

```
##
## Pout-failure Pout-success
##           491           192
```

```
df$f.poutcome<-2
```

```

# 1 level - failure
sel<-which(df$poutcome %in% c("Pout-failure"))
df$f.poutcome[sel] <- 1
sel<- which(is.na(df$poutcome))
df$f.poutcome[sel] <- 1
table(df$f.poutcome)

##
##      1      2
## 4787  192

# 2 level - success
sel<-which(df$poutcome %in% c("Pout-success"))
df$f.poutcome[sel] <- 2
summary(df$f.education)

##      Mandatory Non-Mandatory      Other
##      2685      2086      208

df$f.poutcome<-factor(df$f.poutcome,levels=1:2,labels=c("f.Pout-failure","f.Pout-success"))
summary(df$f.poutcome)

## f.Pout-failure f.Pout-success
##      4787      192

Imputem manualment default ja que pensem que els que no han respós a la pregunta no poden ser imputats
a la repsosta afirmativa, tot i ser concients que seria totalment probable:

sel <- which(is.na(df$default))
df$f.default[sel] <- "f.no"
table(df$f.default)

##
## f.no f.si
## 4979    0

```

6 Profiling

```

condes(df[c(1:29, 31:34,38)],11)

## $quanti
##      correlation      p.value
## pdays      0.02993732 0.0346537787
## nr.employed -0.03189122 0.0244288564
## campaign    -0.04723473 0.0008560601
##
## $quali
##      R2      p.value
## y      0.160738690 1.169792e-191
## month    0.006371350 2.170869e-04
## f.month    0.006371350 2.170869e-04
## day_of_week 0.002933693 5.566255e-03
## f.day      0.002933693 5.566255e-03
## f.prev_contacted 0.001373939 8.903355e-03
## f.poutcome    0.001278653 1.162430e-02
## contact      0.001010183 2.491598e-02

```

```
## f.contact      0.001010183  2.491598e-02
## f.housing      0.001007355  2.511947e-02
## poutcome      0.001405185  3.024097e-02
##
## $category
##              Estimate      p.value
## yes          148.5619189 1.169792e-191
## f.day.wed     23.8425938  3.704865e-04
## wed          23.8425938  3.704865e-04
## Contacted     21.9755934  8.903355e-03
## Month-dec     122.9070700  9.727753e-03
## dec          122.9070700  9.727753e-03
## Job-self-employed 45.0705179 1.128480e-02
## self-employed 42.4765578 1.128480e-02
## f.Pout-success 21.9270496 1.162430e-02
## Pout-success  31.6018377 1.162430e-02
## Month-jul      0.4019466 1.503004e-02
## jul           0.4019466 1.503004e-02
## no           10.6116297 2.191781e-02
## f.cellular     7.8044172 2.491598e-02
## cellular      7.8044172 2.491598e-02
## f.no           7.5382090 2.511947e-02
## Mandatory     11.4647925 2.667720e-02
## f.day.mon     -13.0962186 4.765549e-02
## mon          -13.0962186 4.765549e-02
## yes           -4.7751417 2.952173e-02
## Summer        -9.4533836 2.921958e-02
## f.yes         -7.5382090 2.511947e-02
## f.telephone   -7.8044172 2.491598e-02
## telephone    -7.8044172 2.491598e-02
## Job-housemaid -48.9550116 2.360381e-02
## housemaid    -51.5489717 2.360381e-02
## f.Pout-failure -21.9270496 1.162430e-02
## No-contacted -21.9755934 8.903355e-03
## Month-aug     -40.1682151 4.976285e-03
## aug          -40.1682151 4.976285e-03
## Month-jun     -43.5063607 1.930997e-03
## jun          -43.5063607 1.930997e-03
## no          -148.5619189 1.169792e-191
```

```
catdes(df, num.var = 21)
```

```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## =====
##              p.value df
## f.default      0.000000e+00 1
## f.prev_contacted 1.746438e-113 1
## poutcome       1.256455e-110 2
## f.poutcome     6.599570e-109 1
## month          2.092803e-78 9
## f.month        2.092803e-78 9
## f.job          2.831986e-27 10
## job           6.520196e-27 11
## contact       7.944988e-25 1
```



```

## f.contact          7.944988e-25  1
## default            1.313876e-11  1
## f.jobssituation    3.313476e-08  2
## f.age              4.789647e-08  3
## f.season           5.088671e-08  2
## f.marital          1.549949e-05  2
## marital            3.916274e-05  3
## education          8.492460e-05  7
## f.education        7.801545e-03  2
##
## Description of each cluster by the categories
## =====
## $no
##
## Cla/Mod    Mod/Cla    Global
## f.prev_contacted=No-contacted  90.65577  98.1624319  95.8626230
## f.poutcome=f.Pout-failure      90.53687  98.3212341  96.1438040
## poutcome=NA                    91.13128  88.8157895  86.2823860
## f.contact=f.telephone          94.68733  38.8157895  36.2924282
## contact=telephone              94.68733  38.8157895  36.2924282
## f.month=Month-may              93.12612  35.3448276  33.6011247
## month=may                      93.12612  35.3448276  33.6011247
## default=NA                     94.41816  22.6406534  21.2291625
## job=blue-collar                94.02080  24.6143376  23.1773448
## f.job=Job-blue-collar          93.95745  25.0453721  23.5991163
## f.jobssituation=Worker         90.30411  57.2595281  56.1357702
## f.marital=f.married           89.85795  61.7059891  60.7953404
## marital=married               89.82770  61.5018149  60.6145812
## f.age=f.age-(30,40]           90.27569  40.8575318  40.0682868
## education=Edu-basic.9y        91.80978  15.7667877  15.2038562
## f.age=f.age-(40,50]           90.93098  25.7032668  25.0251054
## f.education=Mandatory          89.68343  54.6279492  53.9264913
## f.month=Month-nov             91.58317  10.3675136  10.0220928
## month=nov                     91.58317  10.3675136  10.0220928
## f.season=Summer               89.63226  45.8938294  45.3303876
## f.month=Month-jun             90.85366  13.5208711  13.1753364
## month=jun                     90.85366  13.5208711  13.1753364
## f.jobssituation=Self-employed  91.32321  9.5508167   9.2588873
## f.job=Job-management          85.13120  6.6243194   6.8889335
## job=management                85.13120  6.6243194   6.8889335
## poutcome=Pout-failure         85.33605  9.5054446   9.8614180
## f.month=Month-dec             63.15789  0.2722323   0.3816027
## month=dec                     63.15789  0.2722323   0.3816027
## f.age=f.age-(50,95]           85.12881  16.4927405  17.1520386
## education=Edu-university.degree 86.07595  27.7676951  28.5599518
## f.age=f.age-[17,30]           84.50226  16.9464610  17.7545692
## f.job=Job-retired              76.88172  3.2441016   3.7356899
## job=retired                   76.88172  3.2441016   3.7356899
## f.marital=f.single            85.14570  27.1778584  28.2586865
## marital=single                85.14570  27.1778584  28.2586865
## f.month=Month-apr             78.57143  5.7395644   6.4671621
## month=apr                     78.57143  5.7395644   6.4671621
## f.season=Aut-Win              81.84569  12.2731397  13.2757582
## f.jobssituation=Other         84.91004  33.1896552  34.6053424
## default=no                    86.94544  77.3593466  78.7708375

```

## f.job=Job-student	59.59596	1.3384755	1.9883511
## job=student	59.59596	1.3384755	1.9883511
## f.month=Month-sep	50.00000	0.7486388	1.3255674
## f.month=Month-mar	50.00000	0.7486388	1.3255674
## month=sep	50.00000	0.7486388	1.3255674
## month=mar	50.00000	0.7486388	1.3255674
## f.month=Month-oct	50.64935	0.8847550	1.5464953
## month=oct	50.64935	0.8847550	1.5464953
## f.contact=f.cellular	85.02522	61.1842105	63.7075718
## contact=cellular	85.02522	61.1842105	63.7075718
## f.poutcome=f.Pout-success	38.54167	1.6787659	3.8561960
## poutcome=Pout-success	38.54167	1.6787659	3.8561960
## f.prev_contacted=Contacted	39.32039	1.8375681	4.1373770
##		p.value	v.test
## f.prev_contacted=No-contacted	1.227915e-68		17.508783
## f.poutcome=f.Pout-failure	1.666964e-65		17.093224
## poutcome=NA	5.763783e-38		12.880929
## f.contact=f.telephone	2.440539e-27		10.831526
## contact=telephone	2.440539e-27		10.831526
## f.month=Month-may	6.034473e-14		7.507332
## month=may	6.034473e-14		7.507332
## default=NA	4.251760e-13		7.247295
## job=blue-collar	1.334762e-12		7.090658
## f.job=Job-blue-collar	1.422485e-12		7.081844
## f.jobsituation=Worker	9.925677e-06		4.418786
## f.marital=f.married	2.858318e-04		3.627813
## marital=married	4.111959e-04		3.532792
## f.age=f.age-(30,40]	1.478138e-03		3.178942
## education=Edu-basic.9y	1.493774e-03		3.175890
## f.age=f.age-(40,50]	1.761035e-03		3.127827
## f.education=Mandatory	5.915694e-03		2.752418
## f.month=Month-nov	2.035659e-02		2.319710
## month=nov	2.035659e-02		2.319710
## f.season=Summer	2.628277e-02		2.222008
## f.month=Month-jun	4.135002e-02		2.040003
## month=jun	4.135002e-02		2.040003
## f.jobsituation=Self-employed	4.326162e-02		2.021175
## f.job=Job-management	4.686185e-02		-1.987547
## job=management	4.686185e-02		-1.987547
## poutcome=Pout-failure	2.293362e-02		-2.274539
## f.month=Month-dec	4.353220e-03		-2.851363
## month=dec	4.353220e-03		-2.851363
## f.age=f.age-(50,95]	8.677261e-04		-3.330235
## education=Edu-university.degree	7.167173e-04		-3.383103
## f.age=f.age-[17,30]	6.154673e-05		-4.006801
## f.job=Job-retired	4.756560e-06		-4.575248
## job=retired	4.756560e-06		-4.575248
## f.marital=f.single	4.173805e-06		-4.602534
## marital=single	4.173805e-06		-4.602534
## f.month=Month-apr	1.145794e-07		-5.301939
## month=apr	1.145794e-07		-5.301939
## f.season=Aut-Win	4.630214e-08		-5.464956
## f.jobsituation=Other	9.853123e-09		-5.733238
## default=no	4.251760e-13		-7.247295

## f.job=Job-student	9.483114e-14	-7.447909	
## job=student	9.483114e-14	-7.447909	
## f.month=Month-sep	8.048922e-15	-7.766807	
## f.month=Month-mar	8.048922e-15	-7.766807	
## month=sep	8.048922e-15	-7.766807	
## month=mar	8.048922e-15	-7.766807	
## f.month=Month-oct	1.184930e-16	-8.284614	
## month=oct	1.184930e-16	-8.284614	
## f.contact=f.cellular	2.440539e-27	-10.831526	
## contact=cellular	2.440539e-27	-10.831526	
## f.poutcome=f.Pout-success	1.666964e-65	-17.093224	
## poutcome=Pout-success	1.666964e-65	-17.093224	
## f.prev_contacted=Contacted	1.227915e-68	-17.508783	
##			
## \$yes			
##	Cla/Mod	Mod/Cla	Global
## f.prev_contacted=Contacted	60.679612	21.891419	4.1373770
## f.poutcome=f.Pout-success	61.458333	20.665499	3.8561960
## poutcome=Pout-success	61.458333	20.665499	3.8561960
## f.contact=f.cellular	14.974779	83.187391	63.7075718
## contact=cellular	14.974779	83.187391	63.7075718
## f.month=Month-oct	49.350649	6.654991	1.5464953
## month=oct	49.350649	6.654991	1.5464953
## f.month=Month-sep	50.000000	5.779335	1.3255674
## f.month=Month-mar	50.000000	5.779335	1.3255674
## month=sep	50.000000	5.779335	1.3255674
## month=mar	50.000000	5.779335	1.3255674
## f.job=Job-student	40.404040	7.005254	1.9883511
## job=student	40.404040	7.005254	1.9883511
## default=no	13.054564	89.667250	78.7708375
## f.jobsituation=Other	15.089959	45.534151	34.6053424
## f.season=Aut-Win	18.154312	21.015762	13.2757582
## f.month=Month-apr	21.428571	12.084063	6.4671621
## month=apr	21.428571	12.084063	6.4671621
## f.marital=f.single	14.854300	36.602452	28.2586865
## marital=single	14.854300	36.602452	28.2586865
## f.job=Job-retired	23.118280	7.530648	3.7356899
## job=retired	23.118280	7.530648	3.7356899
## f.age=f.age-[17,30]	15.497738	23.992995	17.7545692
## education=Edu-university.degree	13.924051	34.676007	28.5599518
## f.age=f.age-(50,95]	14.871194	22.241681	17.1520386
## f.month=Month-dec	36.842105	1.225919	0.3816027
## month=dec	36.842105	1.225919	0.3816027
## poutcome=Pout-failure	14.663951	12.609457	9.8614180
## f.job=Job-management	14.868805	8.931699	6.8889335
## job=management	14.868805	8.931699	6.8889335
## f.jobsituation=Self-employed	8.676790	7.005254	9.2588873
## f.month=Month-jun	9.146341	10.507881	13.1753364
## month=jun	9.146341	10.507881	13.1753364
## f.season=Summer	10.367745	40.980736	45.3303876
## f.month=Month-nov	8.416834	7.355517	10.0220928
## month=nov	8.416834	7.355517	10.0220928
## f.education=Mandatory	10.316574	48.511384	53.9264913
## f.age=f.age-(40,50]	9.069021	19.789842	25.0251054

## education=Edu-basic.9y	8.190225	10.858144	15.2038562
## f.age=f.age-(30,40]	9.724311	33.975482	40.0682868
## marital=married	10.172300	53.765324	60.6145812
## f.marital=f.married	10.142055	53.765324	60.7953404
## f.job=situation=Worker	9.695886	47.460595	56.1357702
## f.job=Job-blue-collar	6.042553	12.434326	23.5991163
## job=blue-collar	5.979203	12.084063	23.1773448
## default=NA	5.581835	10.332750	21.2291625
## f.month=Month-may	6.873879	20.140105	33.6011247
## month=may	6.873879	20.140105	33.6011247
## f.contact=f.telephone	5.312673	16.812609	36.2924282
## contact=telephone	5.312673	16.812609	36.2924282
## poutcome=NA	8.868715	66.725044	86.2823860
## f.poutcome=f.Pout-failure	9.463129	79.334501	96.1438040
## f.prev_contacted=No-contacted	9.344228	78.108581	95.8626230
##	p.value	v.test	
## f.prev_contacted=Contacted	1.227915e-68	17.508783	
## f.poutcome=f.Pout-success	1.666964e-65	17.093224	
## poutcome=Pout-success	1.666964e-65	17.093224	
## f.contact=f.cellular	2.440539e-27	10.831526	
## contact=cellular	2.440539e-27	10.831526	
## f.month=Month-oct	1.184930e-16	8.284614	
## month=oct	1.184930e-16	8.284614	
## f.month=Month-sep	8.048922e-15	7.766807	
## f.month=Month-mar	8.048922e-15	7.766807	
## month=sep	8.048922e-15	7.766807	
## month=mar	8.048922e-15	7.766807	
## f.job=Job-student	9.483114e-14	7.447909	
## job=student	9.483114e-14	7.447909	
## default=no	4.251760e-13	7.247295	
## f.job=situation=Other	9.853123e-09	5.733238	
## f.season=Aut-Win	4.630214e-08	5.464956	
## f.month=Month-apr	1.145794e-07	5.301939	
## month=apr	1.145794e-07	5.301939	
## f.marital=f.single	4.173805e-06	4.602534	
## marital=single	4.173805e-06	4.602534	
## f.job=Job-retired	4.756560e-06	4.575248	
## job=retired	4.756560e-06	4.575248	
## f.age=f.age-[17,30]	6.154673e-05	4.006801	
## education=Edu-university.degree	7.167173e-04	3.383103	
## f.age=f.age-(50,95]	8.677261e-04	3.330235	
## f.month=Month-dec	4.353220e-03	2.851363	
## month=dec	4.353220e-03	2.851363	
## poutcome=Pout-failure	2.293362e-02	2.274539	
## f.job=Job-management	4.686185e-02	1.987547	
## job=management	4.686185e-02	1.987547	
## f.job=situation=Self-employed	4.326162e-02	-2.021175	
## f.month=Month-jun	4.135002e-02	-2.040003	
## month=jun	4.135002e-02	-2.040003	
## f.season=Summer	2.628277e-02	-2.222008	
## f.month=Month-nov	2.035659e-02	-2.319710	
## month=nov	2.035659e-02	-2.319710	
## f.education=Mandatory	5.915694e-03	-2.752418	
## f.age=f.age-(40,50]	1.761035e-03	-3.127827	

```

## education=Edu-basic.9y      1.493774e-03 -3.175890
## f.age=f.age-(30,40]        1.478138e-03 -3.178942
## marital=married            4.111959e-04 -3.532792
## f.marital=f.married        2.858318e-04 -3.627813
## f.job=situation=Worker      9.925677e-06 -4.418786
## f.job=Job-blue-collar      1.422485e-12 -7.081844
## job=blue-collar            1.334762e-12 -7.090658
## default=NA                 4.251760e-13 -7.247295
## f.month=Month-may          6.034473e-14 -7.507332
## month=may                   6.034473e-14 -7.507332
## f.contact=f.telephone      2.440539e-27 -10.831526
## contact=telephone          2.440539e-27 -10.831526
## poutcome=NA                5.763783e-38 -12.880929
## f.poutcome=f.Pout-failure  1.666964e-65 -17.093224
## f.prev_contacted=No-contacted 1.227915e-68 -17.508783
##
##
## Link between the cluster variable and the quantitative variables
## =====
##
##              Eta2      P-value
## duration      0.160738690 1.169792e-191
## nr.employed   0.120745600 2.760101e-141
## euribor3m     0.087576045 3.394474e-101
## emp.var.rate  0.081696867 3.070308e-94
## previous      0.042285141 1.108497e-48
## missings      0.030232449 4.294241e-35
## cons.price.idx 0.017977169 2.070129e-21
## cons.conf.idx  0.008055104 2.236211e-10
## campaign      0.005633367 1.143924e-07
## outliers      0.002285655 7.393671e-04
##
## Description of each cluster by quantitative variables
## =====
## $no
##
##      v.test Mean in category Overall mean sd in category
## nr.employed    24.516762    5177.09015426 5168.16794537    63.4164811
## euribor3m      20.879501      3.82261162    3.63896766    1.6252539
## emp.var.rate    20.166482      0.26197822    0.10236995    1.4608943
## missings        12.267727      2.51610708    2.44004820    1.1967189
## cons.price.idx   9.459934     93.61071461    93.58315164    0.5520673
## campaign         5.295555      2.31459362    2.27063047    1.6629733
## outliers        -3.373128      0.03924682    0.04277967    0.1976554
## cons.conf.idx   -6.332323     -40.74344374   -40.59754971    4.2611089
## previous       -14.508461      0.11734053    0.14694311    0.3499570
## duration       -28.287050     222.56442831   256.63908415   194.8113004
##
##      Overall sd      p.value
## nr.employed    71.3410455 9.788565e-133
## euribor3m       1.7241963 8.224659e-97
## emp.var.rate     1.5515130 1.928802e-90
## missings         1.2153909 1.349895e-34
## cons.price.idx   0.5711736 3.081386e-21
## campaign         1.6274493 1.186556e-07
## outliers         0.2053159 7.431944e-04
## cons.conf.idx    4.5165276 2.414979e-10

```

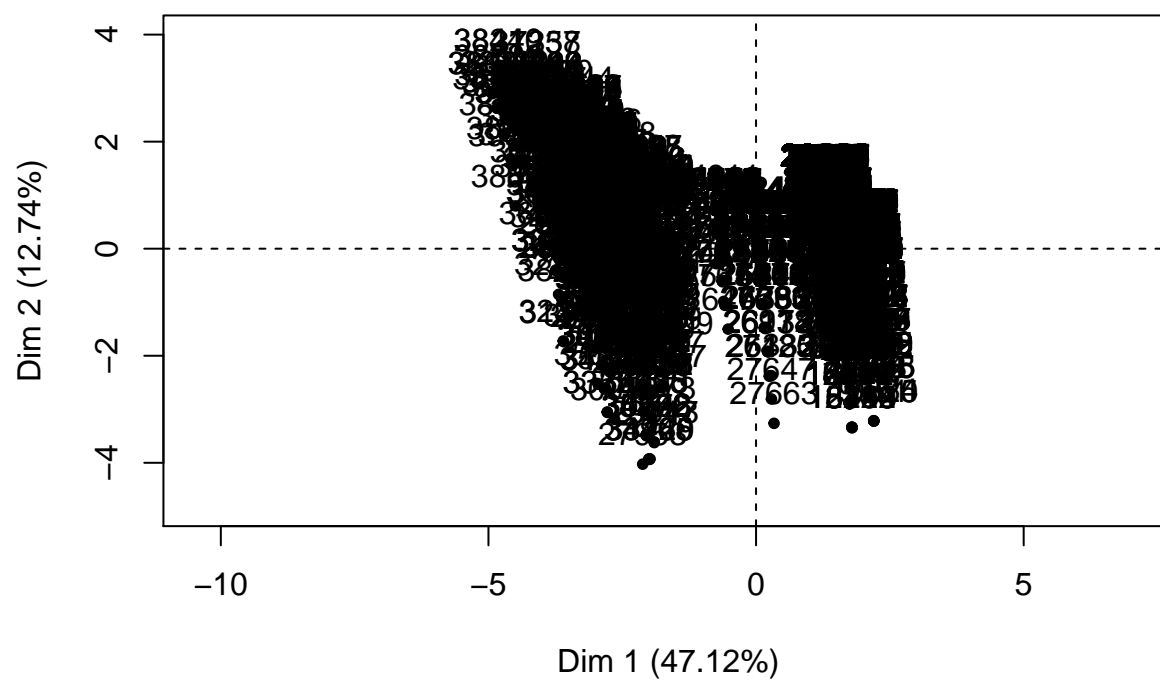
```
## previous      0.3999801 1.070970e-47
## duration      236.1424174 4.987506e-176
##
## $yes
##              v.test Mean in category Overall mean sd in category
## duration      28.287050      519.68826620 256.63908415 339.2762889
## previous      14.508461       0.37546879   0.14694311   0.6249775
## cons.conf.idx   6.332323     -39.47127846 -40.59754971   6.0227719
## outliers        3.373128      0.07005254   0.04277967   0.2552355
## campaign       -5.295555      1.93124423   2.27063047   1.2712707
## cons.price.idx -9.459934      93.37037128  93.58315164   0.6639079
## missings       -12.267727      1.85288967   2.44004820   1.1981032
## emp.var.rate   -20.166482     -1.12977233   0.10236995   1.6732331
## euribor3m      -20.879501      2.22127496   3.63896766   1.8058259
## nr.employed    -24.516762    5099.29036778 5168.16794537 89.3017729
##              Overall sd      p.value
## duration      236.1424174 4.987506e-176
## previous      0.3999801 1.070970e-47
## cons.conf.idx  4.5165276 2.414979e-10
## outliers       0.2053159 7.431944e-04
## campaign       1.6274493 1.186556e-07
## cons.price.idx 0.5711736 3.081386e-21
## missings       1.2153909 1.349895e-34
## emp.var.rate   1.5515130 1.928802e-90
## euribor3m      1.7241963 8.224659e-97
## nr.employed    71.3410455 9.788565e-133
```

7 PCA analysis

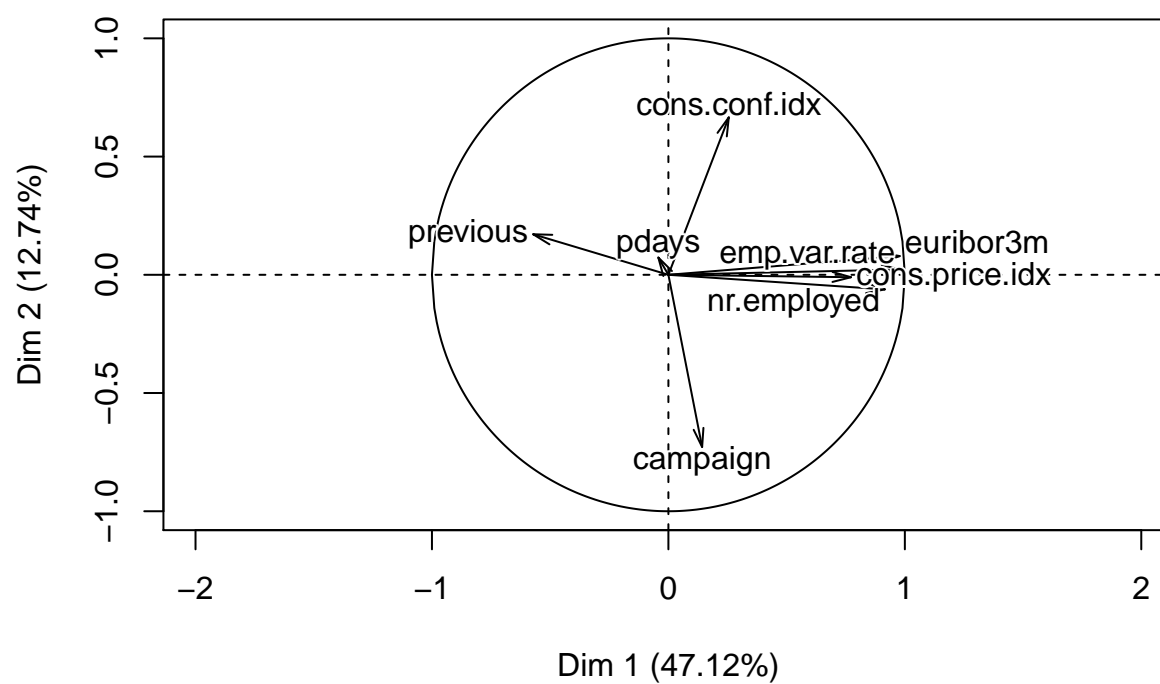
7.1 Eigenvalues and dominant axes analysis

```
#PCA Y analysis (11)
vfact <- names(df[c(23,25:29,31:34)])
vnum <- names(df[c(12:14,16:20)])
res.pca <- PCA(df[,vnum])
```

Individuals factor map (PCA)

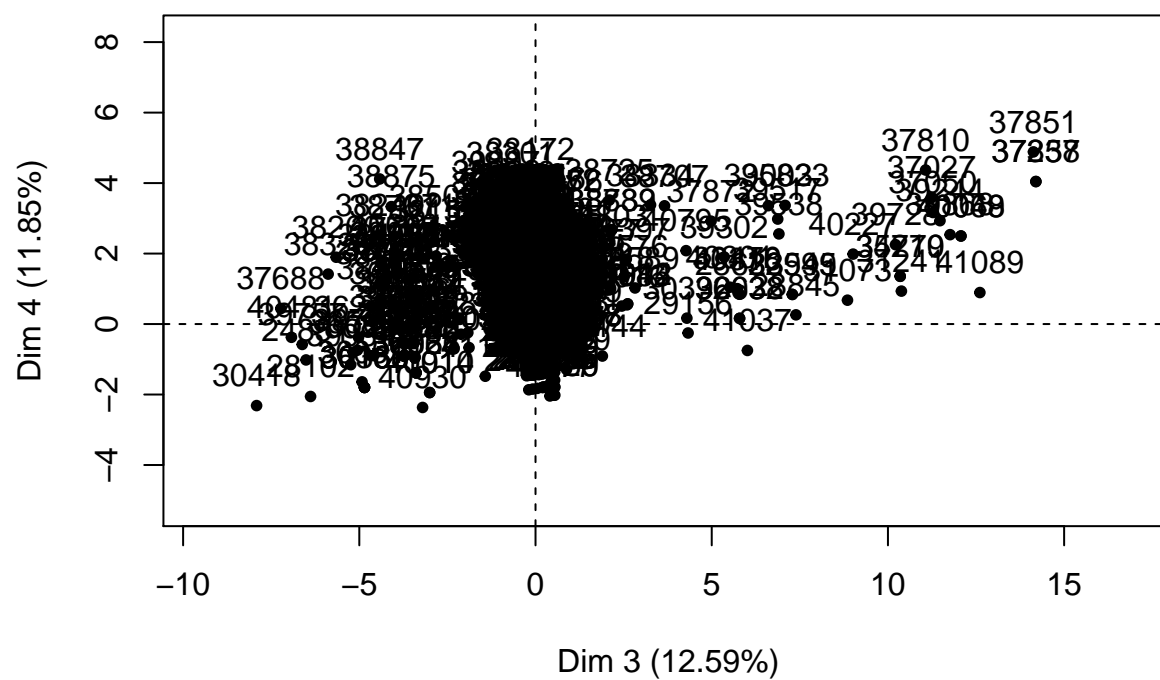


Variables factor map (PCA)

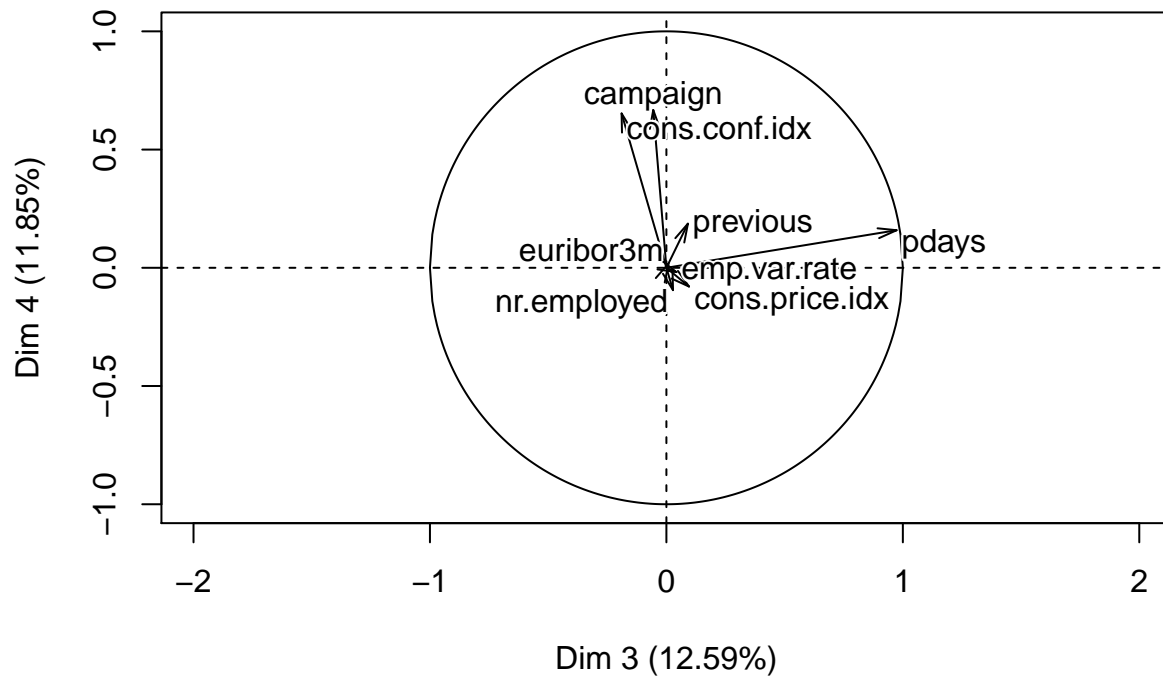


```
res.pca <-PCA(df[,vnum], axes=c(3,4))
```


Individuals factor map (PCA)



Variables factor map (PCA)



```
summary(res.pca, nb.dec = 2, nbelements = 10)
```

```
##
## Call:
## PCA(X = df[, vnum], axes = c(3, 4))
##
##
## Eigenvalues
##          Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7
## Variance      3.77   1.02   1.01   0.95   0.74   0.48   0.03
## % of var.     47.12  12.74  12.59  11.85   9.25   6.01   0.32
## Cumulative % of var. 47.12  59.86  72.45  84.30  93.55  99.55  99.87
##          Dim.8
## Variance      0.01
## % of var.     0.13
## Cumulative % of var. 100.00
##
## Individuals (the 10 first)
##          Dist  Dim.1  ctr  cos2  Dim.2  ctr  cos2  Dim.3
## 4          | 1.77 | 1.29 0.01 0.53 | 1.16 0.03 0.43 | -0.06
## 9          | 1.78 | 1.29 0.01 0.52 | 1.17 0.03 0.43 | 0.09
## 22         | 1.78 | 1.29 0.01 0.53 | 1.15 0.03 0.41 | -0.20
## 47         | 1.78 | 1.29 0.01 0.53 | 1.14 0.03 0.41 | -0.22
## 55         | 1.61 | 1.34 0.01 0.69 | 0.70 0.01 0.19 | -0.24
## 56         | 1.78 | 1.29 0.01 0.53 | 1.15 0.03 0.41 | -0.20
## 62         | 1.78 | 1.29 0.01 0.53 | 1.15 0.03 0.42 | -0.16
```

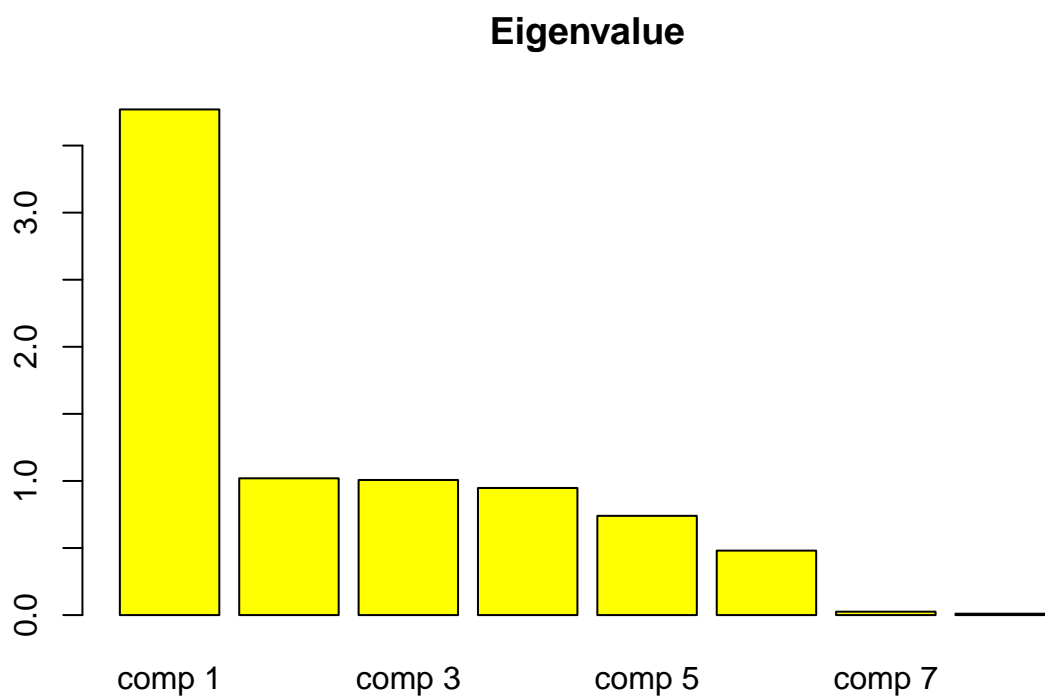
```

## 71          |  1.77 |  1.29  0.01  0.53 |  1.16  0.03  0.43 | -0.07
## 77          |  1.78 |  1.29  0.01  0.53 |  1.15  0.03  0.41 | -0.20
## 79          |  1.78 |  1.29  0.01  0.53 |  1.15  0.03  0.41 | -0.20
##            ctr  cos2
## 4            0.00  0.00 |
## 9            0.00  0.00 |
## 22           0.00  0.01 |
## 47           0.00  0.02 |
## 55           0.00  0.02 |
## 56           0.00  0.01 |
## 62           0.00  0.01 |
## 71           0.00  0.00 |
## 77           0.00  0.01 |
## 79           0.00  0.01 |
##
## Variables
##            Dim.1  ctr  cos2  Dim.2  ctr  cos2  Dim.3  ctr  cos2
## campaign        |  0.14  0.54  0.02 | -0.73 52.04  0.53 | -0.06  0.31  0.00
## pdays           | -0.04  0.05  0.00 |  0.07  0.52  0.01 |  0.97 94.13  0.95
## previous        | -0.57  8.68  0.33 |  0.17  2.88  0.03 |  0.09  0.82  0.01
## emp.var.rate    |  0.98 25.52  0.96 |  0.02  0.05  0.00 |  0.04  0.18  0.00
## cons.price.idx  |  0.77 15.83  0.60 | -0.01  0.01  0.00 |  0.10  0.91  0.01
## cons.conf.idx   |  0.26  1.73  0.07 |  0.67 43.50  0.44 | -0.19  3.57  0.04
## euribor3m       |  0.98 25.40  0.96 |  0.08  0.62  0.01 |  0.01  0.01  0.00
## nr.employed     |  0.92 22.25  0.84 | -0.06  0.37  0.00 |  0.03  0.08  0.00
##
## campaign        |
## pdays           |
## previous        |
## emp.var.rate    |
## cons.price.idx  |
## cons.conf.idx   |
## euribor3m       |
## nr.employed     |

```

Segons criteri de Khaiser realitzarem la interpretació de les 3 primeres dimensions, ja que la quarta ja te un egigenvalue menor a 1.

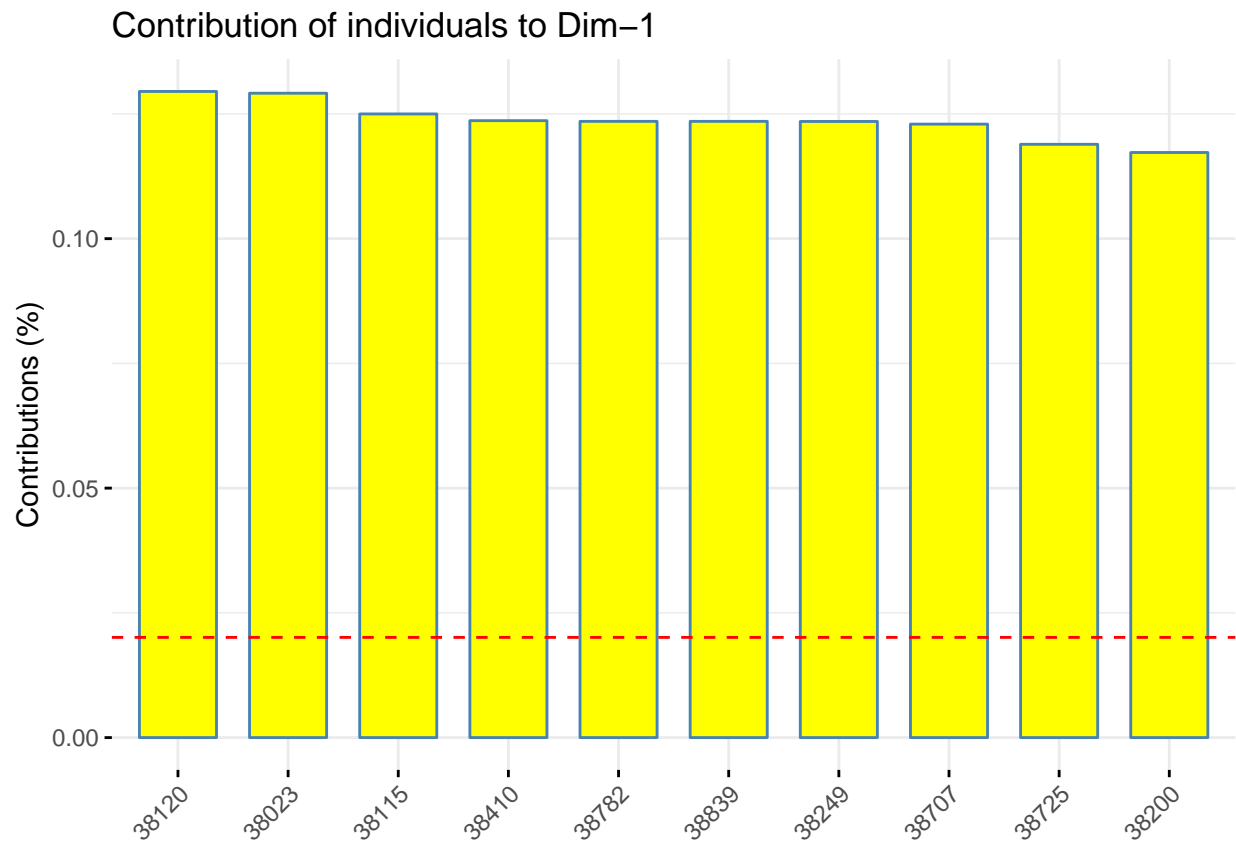
```
barplot(res.pca$eig[,1], col = "yellow", main= "Eigenvalue")
```



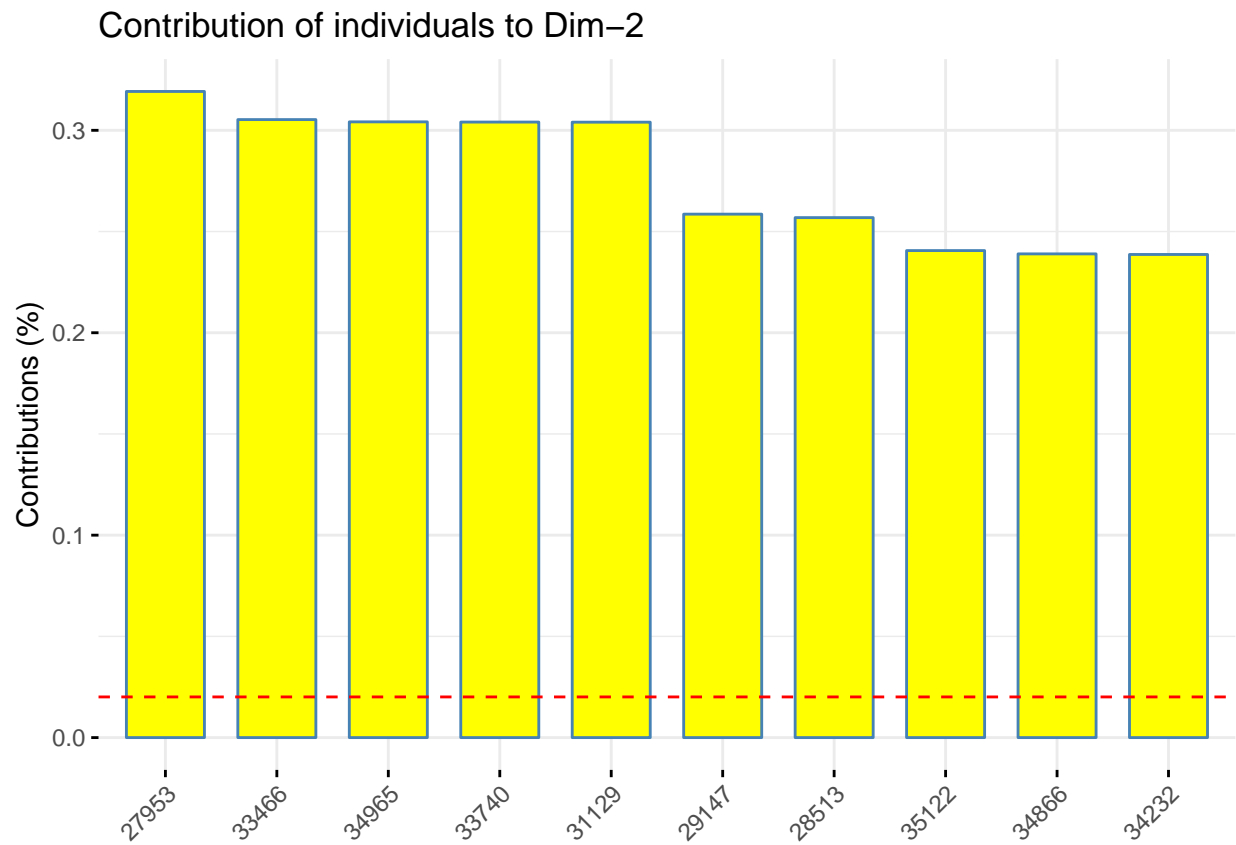
En canvi, interpretarem 6 dimensions per Elbow's rule ja que notem una baixada considerable en a partir de la component numero 7.

7.2 Individuals point of view

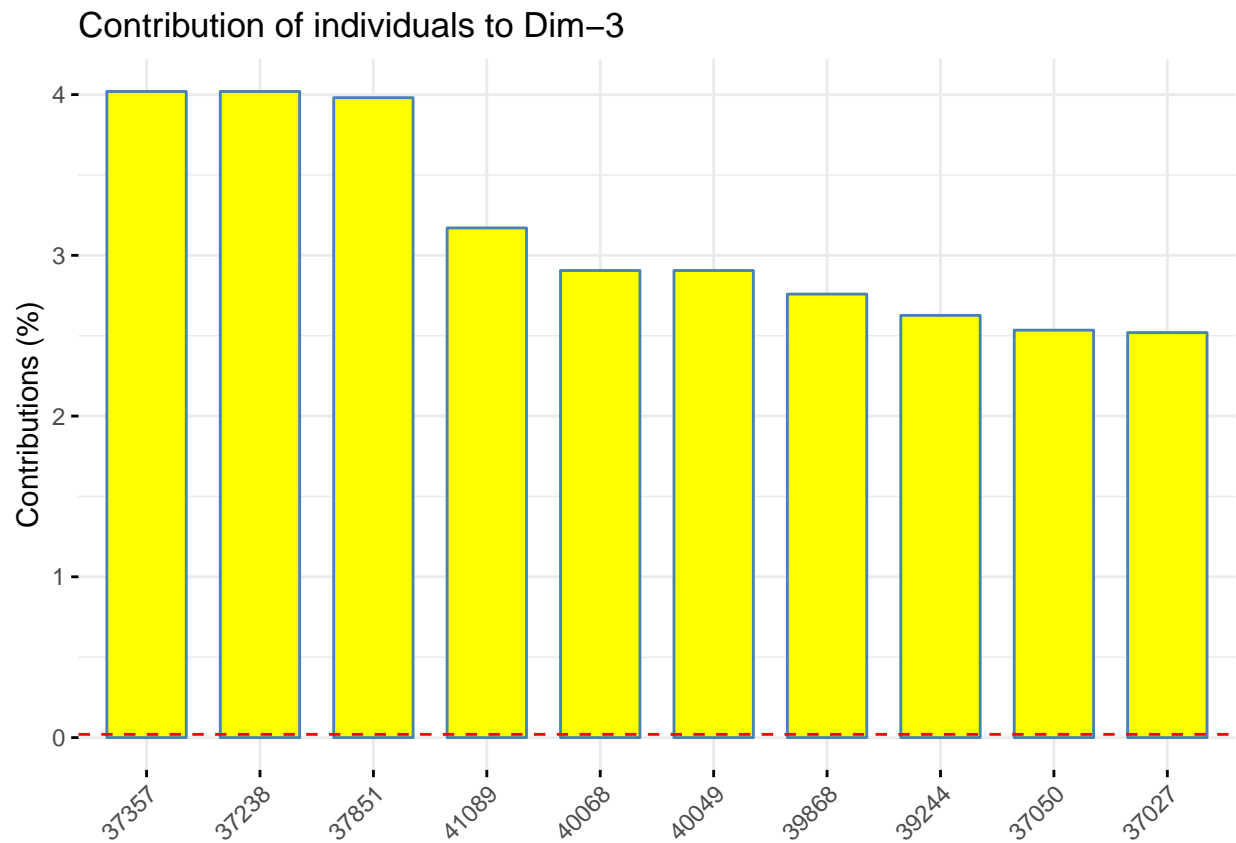
```
#Individus que contribueixen més a la dimensió 1
fviz_contrib(res.pca, choice = "ind", top = 10, fill = "yellow", axes = 1); # Dimensió 1
```



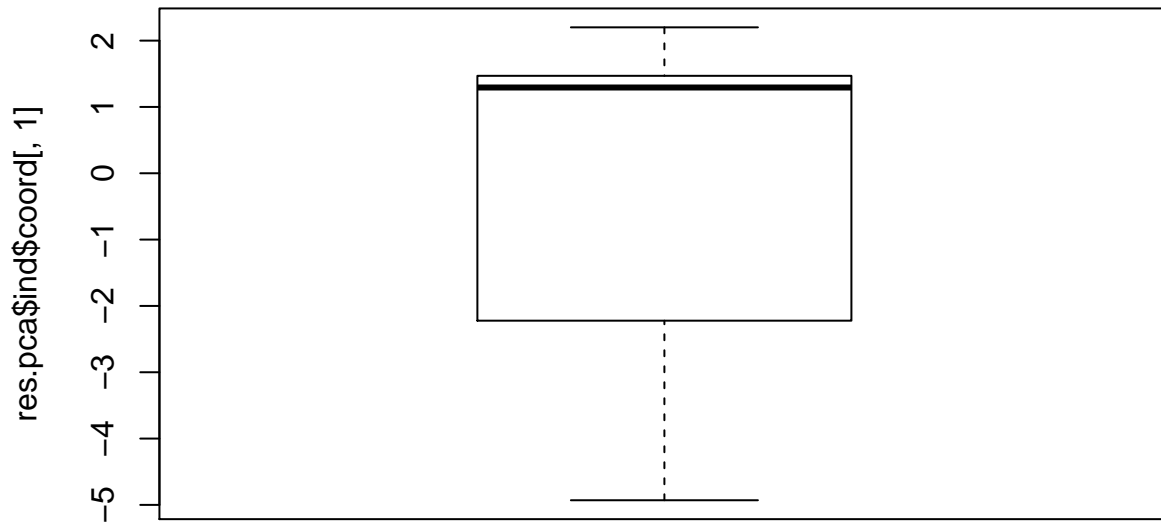
```
#Individus que contribueixen més a la dimensió 2  
fviz_contrib(res.pca, choice = "ind", top = 10, fill = "yellow", axes = 2); # Dimensió 2
```



```
#Individus que contribueixen més a la dimensió 3
fviz_contrib(res.pca, choice = "ind", top = 10, fill = "yellow", axes = 3); # Dimensió 3
```



```
#Ara observem els individus més extrems del nostre data frame.  
indiv_out.d1<-Boxplot(res.pca$ind$coord[,1]); indiv_out.d1; # Dimensió 1
```



```
## NULL
```

```
#En la dimensió 1 no trobem cap extrem
```

```
# Dimensió 2
```

```
indiv_out.d2<-Boxplot(res.pca$ind$coord[,2]); indiv_out.d2;
```

```
## [1] 3394 4068 4246 4099 3786 3535 3464 4264 4231 4152 4662 4641 4547 4535
```

```
## [15] 4699 4703 4634 4615 4650 4680
```

```
q1 = quantile(res.pca$ind$coord[,1])[2];q1;
```

```
## 25%
```

```
## -2.222655
```

```
q3 = quantile(res.pca$ind$coord[,1])[4];q3;
```

```
## 75%
```

```
## 1.469477
```

```
mild.threshold.upper = (q3-q1) * 1.5 + q3;mild.threshold.upper;
```

```
## 75%
```

```
## 7.007677
```

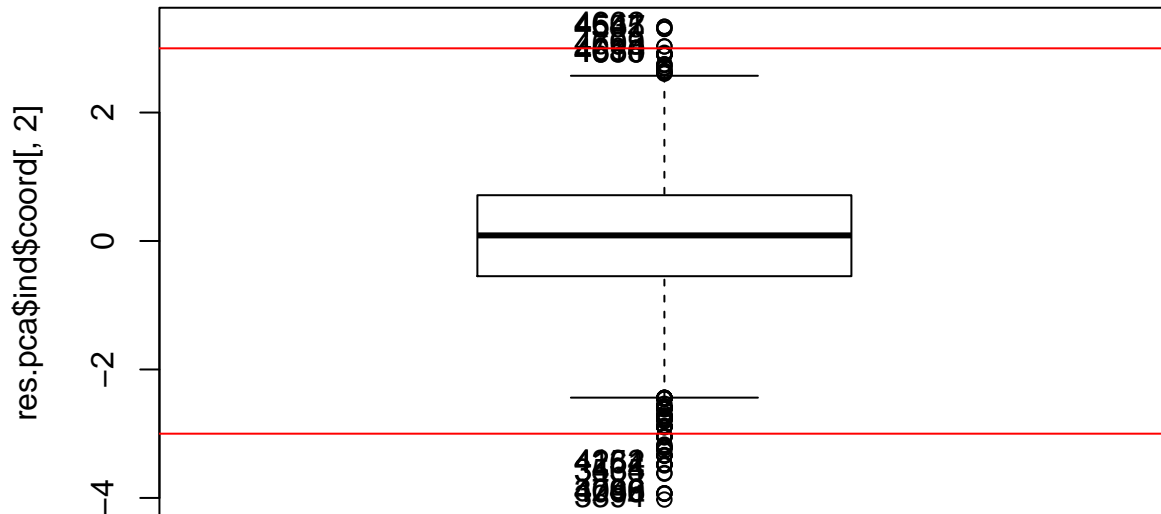
```
mild.threshold.lower = q1 -(q3-q1) * 1.5;mild.threshold.lower;
```

```
## 25%
```

```
## -7.760854
```



```
abline(h=c(3, -3), col = "red")
```



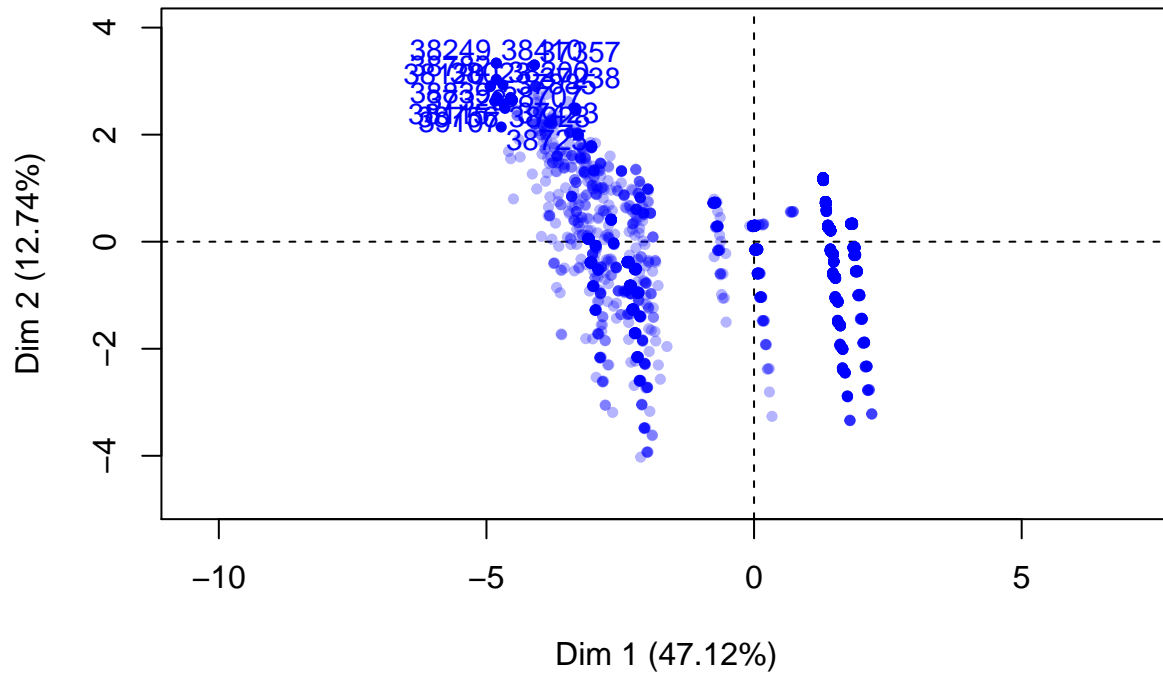
```
indiv_sup.d2 <- which(res.pca$ind$coord[,2] >= 3 | res.pca$ind$coord[,2] <= -3);
aux <- sort(indiv_sup.d2, decreasing= TRUE)
df[aux[1:5], vfact]
```

```
##      f.season f.jobssituation f.prev_contacted  f.education f.housing
## 38839 Aut-Win      Worker      Contacted Non-Mandatory    f.no
## 38782 Aut-Win      Other      Contacted Non-Mandatory    f.no
## 38410 Aut-Win      Other    No-contacted Non-Mandatory    f.no
## 38249 Aut-Win      Worker    No-contacted Non-Mandatory    f.no
## 37357 Summer      Other      Contacted Non-Mandatory    f.no
##      f.marital f.loan  f.contact  f.day      f.age
## 38839 f.single  f.no f.cellular f.day.mon f.age-(30,40]
## 38782 f.single  f.no f.cellular f.day.thu f.age-(30,40]
## 38410 f.single  f.yes f.cellular f.day.wed f.age-(30,40]
## 38249 f.divorced f.yes f.cellular f.day.tue f.age-(30,40]
## 37357 f.single  f.no f.cellular f.day.tue f.age-[17,30]
```

En la dimensió 2 podem veure una petita mostra que les coordenades més extremes ens apareixen en individus amb un nivell educatiu basic, sense parella i en una edat compresa entre 30 i 40 anys.

```
plot.PCA(res.pca,choix=c("ind"),cex=0.95, col.ind="blue",select = "contrib 18")
```

Individuals factor map (PCA)



#Dimensió 3

```
indiv_out.d3<-Boxplot(res.pca$ind$coord[,3]); indiv_out.d3;
```

```
## [1] 3699 4578 4901 4818 3013 3415 4651 4634 4844 4451 4547 4535 4597 4969
```

```
## [15] 4856 4853 4828 4745 4512 4510
```

```
q1 = quantile(res.pca$ind$coord[,1])[2];q1;
```

```
## 25%
```

```
## -2.222655
```

```
q3 = quantile(res.pca$ind$coord[,1])[4];q3;
```

```
## 75%
```

```
## 1.469477
```

```
mild.threshold.upper = (q3-q1) * 1.5 + q3;mild.threshold.upper;
```

```
## 75%
```

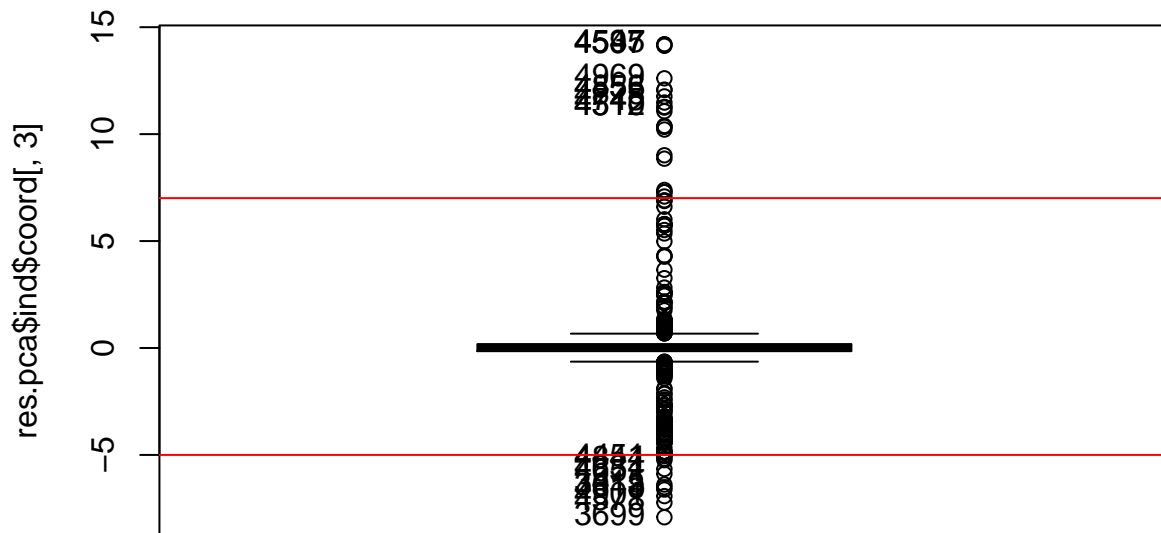
```
## 7.007677
```

```
mild.threshold.lower = q1 -(q3-q1) * 1.5;mild.threshold.lower;
```

```
## 25%
```

```
## -7.760854
```

```
abline(h=c(mild.threshold.upper, -5), col = "red")
```



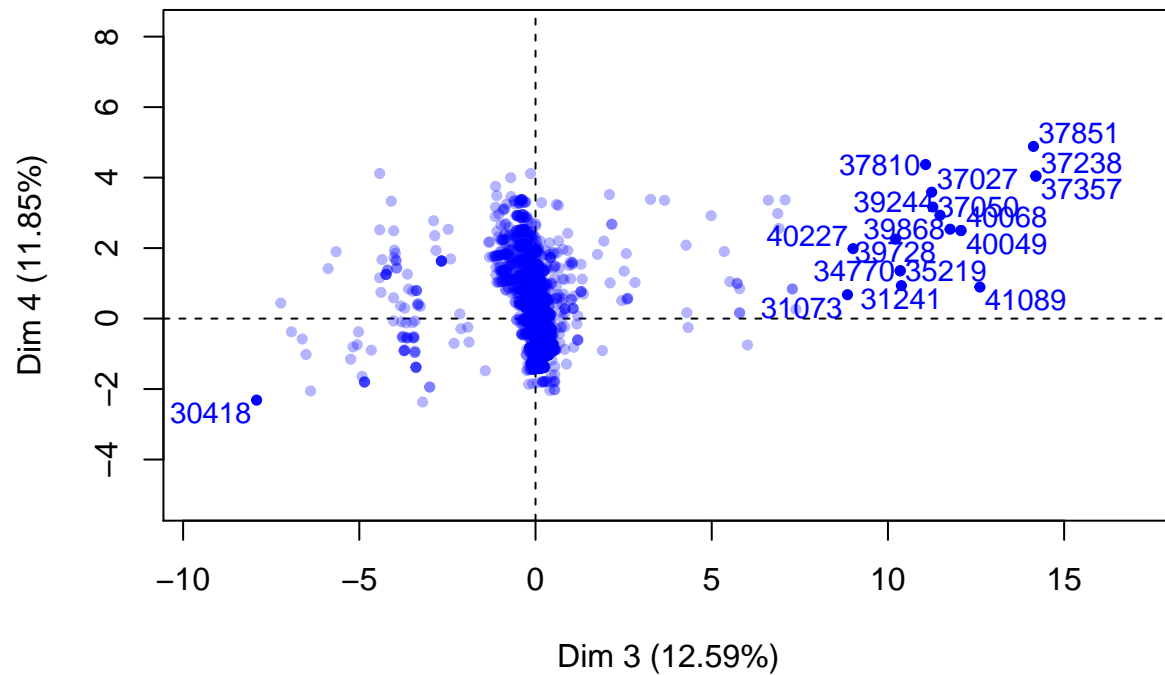
```
indiv_sup.d3 <- which(res.pca$ind$coord[,3] >= mild.threshold.upper | res.pca$ind$coord[,3] <= -5);
aux <- sort(indiv_sup.d3, decreasing= TRUE)
df[aux[1:7], vfact]
```

```
##      f.season f.jobssituation f.prev_contacted  f.education f.housing
## 41089 Aut-Win          Other      Contacted      Mandatory    f.yes
## 40481 Summer          Worker      Contacted Non-Mandatory    f.yes
## 40227 Summer          Worker      Contacted      Mandatory    f.yes
## 40068 Summer          Other      Contacted          Other    f.yes
## 40049 Summer          Other      Contacted      Mandatory    f.yes
## 39984 Summer Self-employed      Contacted      Mandatory    f.no
## 39868 Summer          Other      Contacted      Mandatory    f.yes
##      f.marital f.loan  f.contact    f.day      f.age
## 41089 f.single  f.no  f.telephone f.day.tue f.age-[17,30]
## 40481 f.married f.no  f.cellular  f.day.thu f.age-(50,95]
## 40227 f.single  f.no  f.cellular  f.day.thu f.age-[17,30]
## 40068 f.single  f.no  f.cellular  f.day.thu f.age-[17,30]
## 40049 f.married f.yes f.cellular  f.day.tue f.age-(40,50]
## 39984 f.married f.no  f.cellular  f.day.tue f.age-(30,40]
## 39868 f.married f.no  f.cellular  f.day.tue f.age-[17,30]
```

En la dimensió 3 en canvi podem veure que les coordenades més extremes ens apareixen en individus amb un nivell educatiu superior, que han estat previamment contactats, amb una hipoteca i que han estat contactats durant la temporada d'estiu.

```
plot.PCA(res.pca,choix=c("ind"),cex=0.95, col.ind="blue",select = "contrib 18", axes = 3:4)
```

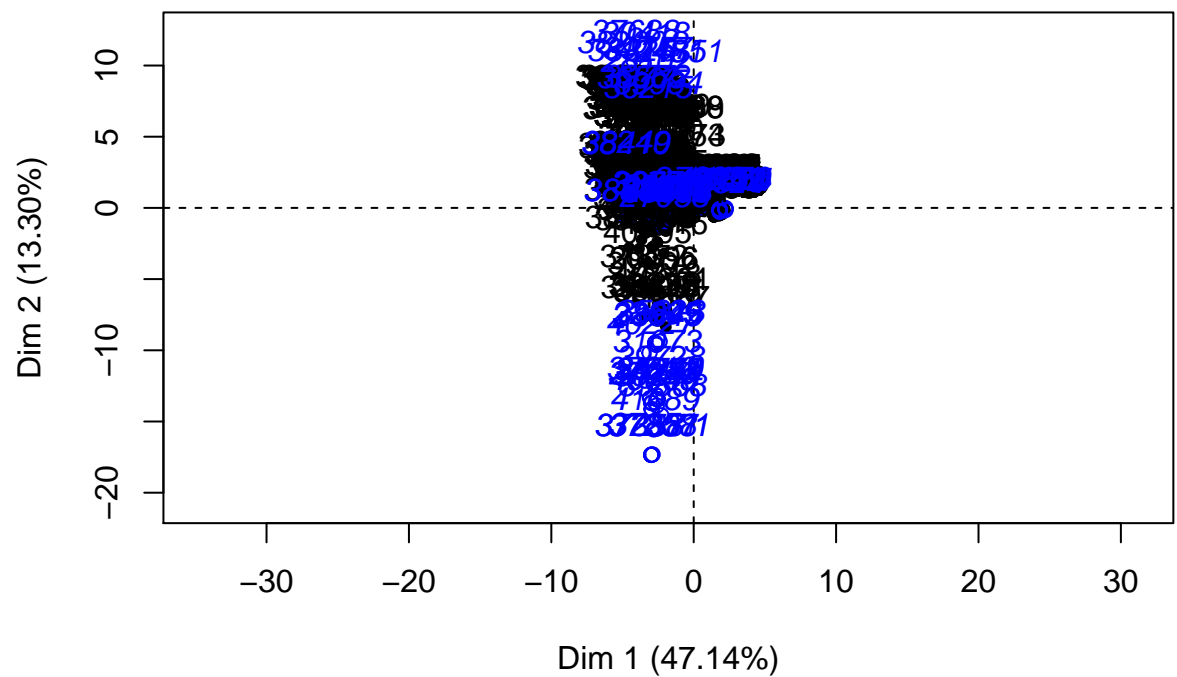
Individuals factor map (PCA)



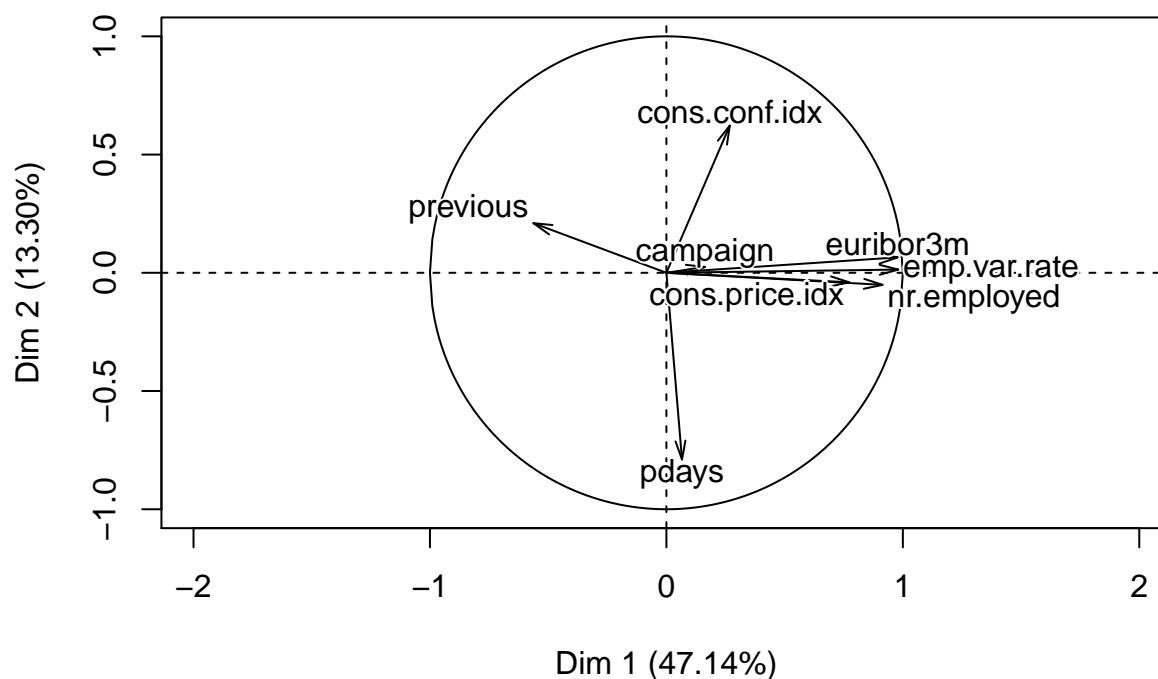
Tornem a realitzar el calcul dels PCA ara tenint en compte que els nostres individus considerats outliers de la mostra, realitzaran la funció d'individus suplementaris:

```
newres.pca <- PCA(df[,vnum], ind.sup = c(indiv_sup.d2, indiv_sup.d3))
```

Individuals factor map (PCA)



Variables factor map (PCA)



Podem veure que en utilitzar els outliers individus com a individus suplementaris els eigenvalues canvien significativament modificant inclús el nombre de dimensions necessaries per a la correcta interpretació de les dades.

```
summary(newres.pca, nb.dec = 2, nbelements = 10)
```

```
##
## Call:
## PCA(X = df[, vnum], ind.sup = c(indiv_sup.d2, indiv_sup.d3))
##
##
## Eigenvalues
##
##          Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7
## Variance      3.77   1.06   0.98   0.91   0.75   0.48   0.03
## % of var.     47.14  13.30  12.29  11.37   9.40   6.04   0.32
## Cumulative % of var. 47.14  60.44  72.73  84.10  93.50  99.55  99.87
##
##          Dim.8
## Variance      0.01
## % of var.      0.13
## Cumulative % of var. 100.00
##
## Individuals (the 10 first)
##
##          Dist  Dim.1  ctr  cos2  Dim.2  ctr  cos2  Dim.3
## 4          | 1.76 | 1.27 0.01 0.51 | 0.45 0.00 0.06 | -1.02
## 9          | 1.79 | 1.27 0.01 0.51 | 0.26 0.00 0.02 | -1.04
## 22         | 1.78 | 1.26 0.01 0.50 | 0.63 0.01 0.13 | -0.99
## 47         | 1.78 | 1.26 0.01 0.50 | 0.65 0.01 0.13 | -0.99
```

```

## 55      | 1.60 | 1.31 0.01 0.67 | 0.65 0.01 0.17 | -0.36
## 56      | 1.78 | 1.26 0.01 0.50 | 0.63 0.01 0.13 | -0.99
## 62      | 1.77 | 1.26 0.01 0.51 | 0.58 0.01 0.11 | -1.00
## 71      | 1.76 | 1.26 0.01 0.51 | 0.46 0.00 0.07 | -1.02
## 77      | 1.78 | 1.26 0.01 0.50 | 0.63 0.01 0.13 | -0.99
## 79      | 1.78 | 1.26 0.01 0.50 | 0.63 0.01 0.13 | -0.99
##          ctr  cos2
## 4          0.02 0.33 |
## 9          0.02 0.34 |
## 22         0.02 0.31 |
## 47         0.02 0.31 |
## 55         0.00 0.05 |
## 56         0.02 0.31 |
## 62         0.02 0.32 |
## 71         0.02 0.33 |
## 77         0.02 0.31 |
## 79         0.02 0.31 |
##
## Supplementary individuals (the 10 first)
##          Dist  Dim.1  cos2  Dim.2  cos2  Dim.3  cos2
## 9951      | 4.84 | 2.23 0.21 | -0.02 0.00 | 4.20 0.76 |
## 10574     | 4.83 | 2.23 0.21 | -0.15 0.00 | 4.18 0.75 |
## 10825     | 4.83 | 2.23 0.21 | -0.07 0.00 | 4.20 0.75 |
## 11050     | 4.84 | 2.23 0.21 | -0.02 0.00 | 4.20 0.76 |
## 12452     | 4.64 | 1.82 0.15 | -0.13 0.00 | 4.26 0.84 |
## 15324     | 4.63 | 1.82 0.15 | -0.27 0.00 | 4.24 0.84 |
## 16243     | 4.64 | 1.82 0.15 | -0.10 0.00 | 4.26 0.84 |
## 18119     | 4.64 | 1.82 0.15 | -0.09 0.00 | 4.26 0.84 |
## 18738     | 4.63 | 1.82 0.15 | -0.22 0.00 | 4.24 0.84 |
## 27663     | 4.45 | 0.35 0.01 | 0.00 0.00 | 4.31 0.94 |
##
## Variables
##          Dim.1  ctr  cos2  Dim.2  ctr  cos2  Dim.3  ctr  cos2
## campaign      | 0.16 0.70 0.03 | 0.02 0.06 0.00 | 0.97 95.39 0.94
## pdays         | 0.07 0.12 0.00 | -0.79 58.58 0.62 | -0.11 1.13 0.01
## previous      | -0.56 8.39 0.32 | 0.21 4.13 0.04 | 0.02 0.03 0.00
## emp.var.rate  | 0.98 25.49 0.96 | 0.01 0.02 0.00 | -0.02 0.06 0.00
## cons.price.idx | 0.77 15.87 0.60 | -0.04 0.15 0.00 | -0.02 0.05 0.00
## cons.conf.idx | 0.27 1.90 0.07 | 0.62 36.41 0.39 | -0.17 3.08 0.03
## euribor3m     | 0.98 25.37 0.96 | 0.07 0.40 0.00 | -0.05 0.26 0.00
## nr.employed   | 0.91 22.16 0.84 | -0.05 0.24 0.00 | 0.00 0.00 0.00
##
## campaign      |
## pdays         |
## previous      |
## emp.var.rate  |
## cons.price.idx |
## cons.conf.idx |
## euribor3m     |
## nr.employed   |

```

7.3 Interpreting the axes

Comprovem de manera més exhaustiva quines variables afecten més als diferents eixos.

```
dimdesc(newres.pca, axes = 1:3)
```

```
## $Dim.1
## $Dim.1$quanti
##          correlation      p.value
## emp.var.rate    0.98049415 0.000000e+00
## euribor3m       0.97805412 0.000000e+00
## nr.employed     0.91405151 0.000000e+00
## cons.price.idx  0.77352829 0.000000e+00
## cons.conf.idx   0.26799918 1.557909e-81
## campaign        0.16290029 1.506009e-30
## pdays           0.06603261 3.641267e-06
## previous        -0.56248739 0.000000e+00
##
##
## $Dim.2
## $Dim.2$quanti
##          correlation      p.value
## cons.conf.idx    0.62247943 0.000000e+00
## previous         0.20972163 6.242514e-50
## euribor3m        0.06518106 4.853424e-06
## cons.price.idx   -0.04039870 4.636959e-03
## nr.employed      -0.05103148 3.472055e-04
## pdays            -0.78955218 0.000000e+00
##
##
## $Dim.3
## $Dim.3$quanti
##          correlation      p.value
## campaign         0.96854829 0.000000e+00
## euribor3m        -0.05087184 3.623859e-04
## pdays            -0.10532598 1.371656e-13
## cons.conf.idx    -0.17391113 1.207461e-34
```

Pel que fa a la primera dimensió, les variables socioeconòmiques són les que ens mostren una major correlació. Per altra banda veiem que el nombre de contactes efectuats abans de l'últim contacte té una relació negativa amb el primer eix.

Pel que fa a la segona dimensió, el més destacable és la relació inversament proporcional que el segon eix de coordenades té amb el nombre de dies que han passat des de l'últim contacte.

En canvi el tercer eix de dimensions està altament relacionat amb el nombre de vegades que un client ha estat contactat durant aquesta campanya.

8 K-Means Classification

```
set.seed(676489)
```

```
summary(res.pca)
```

```
##
## Call:
## PCA(X = df[, vnum], axes = c(3, 4))
##
```



```

##
## Eigenvalues
##          Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance      3.769   1.020   1.007   0.948   0.740   0.481
## % of var.     47.117  12.744  12.591  11.845   9.248   6.008
## Cumulative % of var. 47.117 59.861 72.452 84.297 93.545 99.553
##          Dim.7   Dim.8
## Variance      0.025   0.010
## % of var.     0.317   0.129
## Cumulative % of var. 99.871 100.000
##
## Individuals (the 10 first)
##          Dist   Dim.1   ctr   cos2   Dim.2   ctr   cos2
## 4          | 1.772 | 1.291 0.009 0.531 | 1.156 0.026 0.426 |
## 9          | 1.778 | 1.288 0.009 0.525 | 1.167 0.027 0.431 |
## 22         | 1.779 | 1.295 0.009 0.530 | 1.146 0.026 0.415 |
## 47         | 1.780 | 1.295 0.009 0.529 | 1.144 0.026 0.413 |
## 55         | 1.607 | 1.340 0.010 0.695 | 0.702 0.010 0.191 |
## 56         | 1.779 | 1.295 0.009 0.530 | 1.146 0.026 0.415 |
## 62         | 1.776 | 1.294 0.009 0.531 | 1.148 0.026 0.418 |
## 71         | 1.772 | 1.292 0.009 0.531 | 1.155 0.026 0.425 |
## 77         | 1.779 | 1.295 0.009 0.530 | 1.146 0.026 0.415 |
## 79         | 1.779 | 1.295 0.009 0.530 | 1.146 0.026 0.415 |
##          Dim.3   ctr   cos2
## 4          -0.062 0.000 0.001 |
## 9           0.088 0.000 0.002 |
## 22         -0.201 0.001 0.013 |
## 47         -0.220 0.001 0.015 |
## 55         -0.243 0.001 0.023 |
## 56         -0.201 0.001 0.013 |
## 62         -0.164 0.001 0.009 |
## 71         -0.071 0.000 0.002 |
## 77         -0.201 0.001 0.013 |
## 79         -0.201 0.001 0.013 |
##
## Variables
##          Dim.1   ctr   cos2   Dim.2   ctr   cos2   Dim.3
## campaign    | 0.143 0.540 0.020 | -0.728 52.045 0.531 | -0.056
## pdays      | -0.043 0.048 0.002 | 0.073 0.524 0.005 | 0.974
## previous    | -0.572 8.683 0.327 | 0.171 2.882 0.029 | 0.091
## emp.var.rate | 0.981 25.520 0.962 | 0.022 0.049 0.000 | 0.043
## cons.price.idx | 0.773 15.833 0.597 | -0.010 0.010 0.000 | 0.096
## cons.conf.idx | 0.255 1.732 0.065 | 0.666 43.501 0.443 | -0.189
## euribor3m    | 0.978 25.398 0.957 | 0.079 0.615 0.006 | 0.009
## nr.employed  | 0.916 22.246 0.839 | -0.062 0.373 0.004 | 0.028
##          ctr   cos2
## campaign    0.306 0.003 |
## pdays      94.130 0.948 |
## previous    0.823 0.008 |
## emp.var.rate 0.180 0.002 |
## cons.price.idx 0.910 0.009 |
## cons.conf.idx 3.565 0.036 |
## euribor3m    0.008 0.000 |
## nr.employed 0.078 0.001 |

```

```

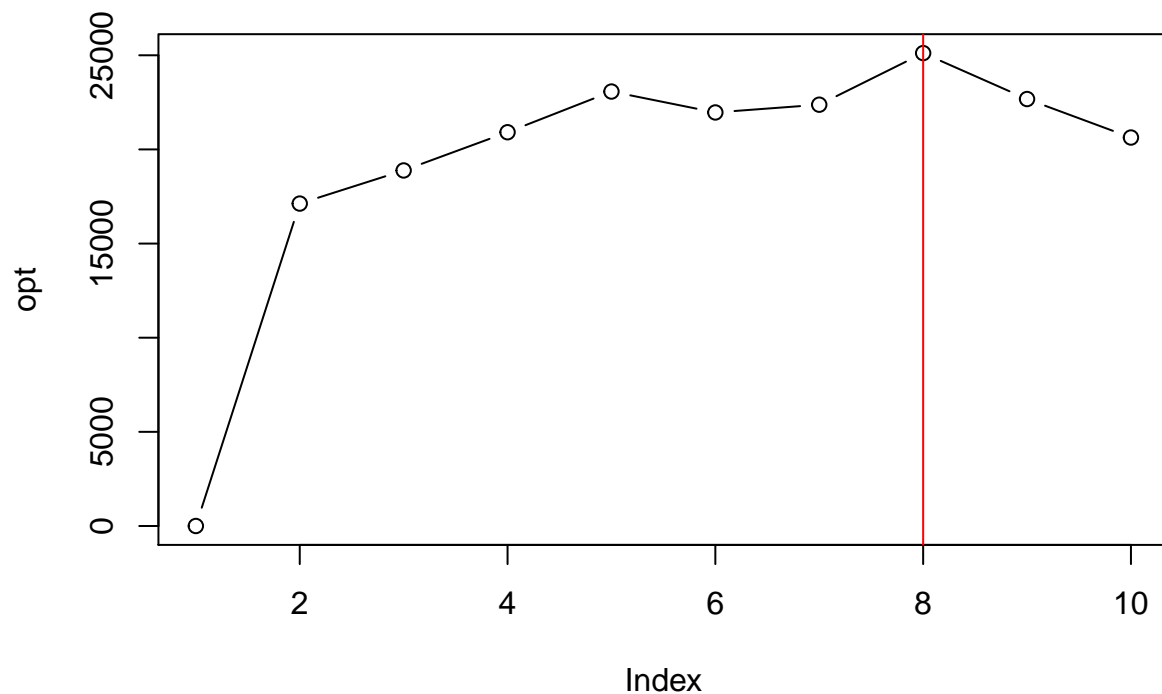
opt <- kmeans(res.pca$ind$coord[,1:3], centers = 1)$betweenss
opt[2] <- kmeans(res.pca$ind$coord[,1:3], centers = 2)$betweenss
opt[3] <- kmeans(res.pca$ind$coord[,1:3], centers = 3)$betweenss
opt[4] <- kmeans(res.pca$ind$coord[,1:3], centers = 4)$betweenss
opt[5] <- kmeans(res.pca$ind$coord[,1:3], centers = 5)$betweenss
opt[6] <- kmeans(res.pca$ind$coord[,1:3], centers = 6)$betweenss
opt[7] <- kmeans(res.pca$ind$coord[,1:3], centers = 7)$betweenss
opt[8] <- kmeans(res.pca$ind$coord[,1:3], centers = 8)$betweenss
opt[9] <- kmeans(res.pca$ind$coord[,1:3], centers = 9)$betweenss
opt[10] <- kmeans(res.pca$ind$coord[,1:3], centers = 10)$betweenss

```

```

plot(opt, type = "b")
abline(v = 8, col = "red")

```



Com podem comprovar en el gràfic de forma visual, trobem un pic en el moment en que el nombre de clusters és optim. En aquest cas n'utilitzarem 7.

8.1 Description of clusters

```

set.seed(676489)
def <- kmeans(res.pca$ind$coord[,1:3], centers = 8)
df$kmeansclu<-factor(def$cluster)

vclust <- c(vfact, vnum, "kmeansclu")
targ <- which(vclust == "kmeansclu")

```

```
catdes(df[,vclust], targ)
```

```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## =====
##                p.value df
## f.season      0.000000e+00 14
## f.prev_contacted 0.000000e+00 7
## f.contact      9.529352e-253 7
## f.age          6.450985e-39 21
## f.jobssituation 2.112793e-23 14
## f.marital      8.985336e-14 14
## f.education    1.878068e-11 14
## f.day          2.731759e-07 28
## f.housing      5.320342e-05 7
##
## Description of each cluster by the categories
## =====
## $`1`
##                Cla/Mod  Mod/Cla  Global  p.value
## f.prev_contacted=Contacted 16.5048544 100.00000 4.137377 5.822212e-49
## f.jobssituation=Other      1.2768427  64.70588 34.605342 3.998384e-04
## f.contact=f.cellular      0.9457755  88.23529 63.707572 1.599915e-03
## f.marital=f.single        1.1371713  47.05882 28.258686 2.068110e-02
## f.age=f.age-[17,30]       1.2443439  32.35294 17.754569 3.924935e-02
## f.marital=f.married       0.4625041  41.17647 60.795340 2.213685e-02
## f.jobssituation=Worker    0.3935599  32.35294 56.135770 5.746337e-03
## f.contact=f.telephone     0.2213614  11.76471 36.292428 1.599915e-03
## f.prev_contacted=No-contacted 0.0000000  0.00000 95.862623 5.822212e-49
##                v.test
## f.prev_contacted=Contacted 14.706890
## f.jobssituation=Other      3.540190
## f.contact=f.cellular      3.155922
## f.marital=f.single        2.313756
## f.age=f.age-[17,30]       2.061563
## f.marital=f.married      -2.288012
## f.jobssituation=Worker    -2.761915
## f.contact=f.telephone    -3.155922
## f.prev_contacted=No-contacted -14.706890
##
## $`2`
##                Cla/Mod  Mod/Cla  Global  p.value
## f.prev_contacted=Contacted 58.252427 35.190616 4.137377 2.099940e-92
## f.season=Aut-Win          20.574887 39.882698 13.275758 1.027456e-37
## f.contact=f.cellular      9.709962 90.322581 63.707572 4.143143e-31
## f.jobssituation=Other     11.839814 59.824047 34.605342 7.969224e-23
## f.age=f.age-(50,95]      14.168618 35.483871 17.152039 2.320965e-17
## f.education=Non-Mandatory  8.916587 54.545455 41.895963 1.166636e-06
## f.education=Other        11.538462  7.038123  4.177546 1.130290e-02
## f.marital=f.single       8.102345 33.431085 28.258686 3.021896e-02
## f.marital=f.married      6.243806 55.425220 60.795340 3.655493e-02
## f.day=f.day.fri          5.196182 14.369501 18.939546 2.240708e-02
## f.season=Summer          5.494019 36.363636 45.330388 5.288761e-04
## f.age=f.age-(30,40]      5.062657 29.618768 40.068287 3.457015e-05
```

```

## f.age=f.age-(40,50]          3.611557 13.196481 25.025105 3.022398e-08
## f.education=Mandatory        4.878957 38.416422 53.926491 2.781501e-09
## f.season=Spring              3.930131 23.753666 41.393854 1.691600e-12
## f.jobssituation=Worker       4.042934 33.137830 56.135770 9.075160e-19
## f.contact=f.telephone        1.826231  9.677419 36.292428 4.143143e-31
## f.prev_contacted=No-contacted 4.630212 64.809384 95.862623 2.099940e-92
##                               v.test
## f.prev_contacted=Contacted   20.388857
## f.season=Aut-Win             12.836239
## f.contact=f.cellular         11.599537
## f.jobssituation=Other        9.834842
## f.age=f.age-(50,95]         8.476487
## f.education=Non-Mandatory    4.861221
## f.education=Other            2.533192
## f.marital=f.single           2.167208
## f.marital=f.married          -2.090700
## f.day=f.day.fri              -2.283396
## f.season=Summer              -3.465688
## f.age=f.age-(30,40]         -4.141062
## f.age=f.age-(40,50]         -5.540134
## f.education=Mandatory        -5.944000
## f.season=Spring              -7.057799
## f.jobssituation=Worker       -8.845952
## f.contact=f.telephone        -11.599537
## f.prev_contacted=No-contacted -20.388857
##
## $`3`
##                               Cla/Mod   Mod/Cla   Global   p.value
## f.season=Spring              50.072780 83.766234 41.39385 7.180797e-275
## f.contact=f.cellular         35.403531 91.152597 63.70757 1.041593e-137
## f.age=f.age-[17,30]         34.049774 24.431818 17.75457 6.099324e-12
## f.marital=f.single           31.485430 35.957792 28.25869 9.741118e-12
## f.housing=f.yes              26.560232 59.415584 55.35248 9.249461e-04
## f.education=Mandatory        26.443203 57.629870 53.92649 2.621338e-03
## f.day=f.day.fri              27.995758 21.428571 18.93955 1.091708e-02
## f.marital=f.divorced         20.000000  8.847403 10.94597 5.691355e-03
## f.education=Non-Mandatory    22.674976 38.392857 41.89596 3.993617e-03
## f.day=f.day.tue              21.257485 17.288961 20.12452 3.835554e-03
## f.housing=f.no               22.492128 40.584416 44.64752 9.249461e-04
## f.age=f.age-(40,50]         20.786517 21.022727 25.02511 1.557078e-04
## f.age=f.age-(50,95]         19.203747 13.311688 17.15204 2.599294e-05
## f.marital=f.married          22.464486 55.194805 60.79534 3.872533e-06
## f.season=Aut-Win             8.925870  4.788961 13.27576 2.195961e-28
## f.contact=f.telephone        6.032097  8.847403 36.29243 1.041593e-137
## f.season=Summer              6.247231 11.444805 45.33039 8.570615e-187
##                               v.test
## f.season=Spring              35.424441
## f.contact=f.cellular         24.978711
## f.age=f.age-[17,30]         6.877310
## f.marital=f.single           6.810277
## f.housing=f.yes              3.312414
## f.education=Mandatory        3.008971
## f.day=f.day.fri              2.545342
## f.marital=f.divorced         -2.765053

```

```
## f.education=Non-Mandatory -2.878665
## f.day=f.day.tue -2.891379
## f.housing=f.no -3.312414
## f.age=f.age-(40,50] -3.781785
## f.age=f.age-(50,95] -4.206001
## f.marital=f.married -4.618109
## f.season=Aut-Win -11.049844
## f.contact=f.telephone -24.978711
## f.season=Summer -29.149123
```

```
##
```

```
## $`4`
```

	Cla/Mod	Mod/Cla	Global	p.value
## f.season=Aut-Win	27.8366112	100.000000	13.275758	7.469427e-173
## f.contact=f.cellular	5.2963430	91.304348	63.707572	2.118037e-18
## f.education=Non-Mandatory	5.0814957	57.608696	41.895963	1.334185e-05
## f.day=f.day.fri	5.4082715	27.717391	18.939546	3.111384e-03
## f.jobsituation=Self-employed	6.2906725	15.760870	9.258887	4.131217e-03
## f.day=f.day.thu	4.8780488	27.173913	20.586463	2.889605e-02
## f.education=Other	0.9615385	1.086957	4.177546	1.776257e-02
## f.age=f.age-[17,30]	2.2624434	10.869565	17.754569	9.262198e-03
## f.day=f.day.tue	2.2954092	12.500000	20.124523	6.118809e-03
## f.day=f.day.mon	2.2682446	12.500000	20.365535	4.814364e-03
## f.education=Mandatory	2.8305400	41.304348	53.926491	4.895913e-04
## f.contact=f.telephone	0.8854455	8.695652	36.292428	2.118037e-18
## f.season=Spring	0.0000000	0.000000	41.393854	1.692208e-44
## f.season=Summer	0.0000000	0.000000	45.330388	3.044426e-50

```
##
```

	v.test
## f.season=Aut-Win	28.027706
## f.contact=f.cellular	8.750822
## f.education=Non-Mandatory	4.354422
## f.day=f.day.fri	2.956514
## f.jobsituation=Self-employed	2.867965
## f.day=f.day.thu	2.184902
## f.education=Other	-2.370531
## f.age=f.age-[17,30]	-2.602222
## f.day=f.day.tue	-2.741346
## f.day=f.day.mon	-2.819199
## f.education=Mandatory	-3.486387
## f.contact=f.telephone	-8.750822
## f.season=Spring	-13.994159
## f.season=Summer	-14.905300

```
##
```

```
## $`5`
```

	Cla/Mod	Mod/Cla	Global
## f.season=Aut-Win	39.6369138	100.000000	13.275758
## f.contact=f.cellular	7.4716267	90.4580153	63.707572
## f.prev_contacted=No-contacted	5.4473078	99.2366412	95.862623
## f.housing=f.yes	6.1683599	64.8854962	55.352480
## f.age=f.age-(30,40]	6.5162907	49.6183206	40.068287
## f.marital=f.divorced	7.1559633	14.8854962	10.945973
## f.education=Other	2.4038462	1.9083969	4.177546
## f.housing=f.no	4.1385515	35.1145038	44.647520
## f.prev_contacted=Contacted	0.9708738	0.7633588	4.137377
## f.age=f.age-[17,30]	2.9411765	9.9236641	17.754569

```

## f.contact=f.telephone      1.3835086   9.5419847 36.292428
## f.season=Spring            0.0000000   0.0000000 41.393854
## f.season=Summer            0.0000000   0.0000000 45.330388
##                               p.value      v.test
## f.season=Aut-Win           1.186073e-253  34.018519
## f.contact=f.cellular       3.372737e-24   10.148275
## f.prev_contacted=No-contacted 1.108220e-03   3.261506
## f.housing=f.yes            1.326847e-03   3.210109
## f.age=f.age-(30,40]        1.339483e-03   3.207383
## f.marital=f.divorced       4.327817e-02   2.021015
## f.education=Other          4.489318e-02  -2.005654
## f.housing=f.no             1.326847e-03  -3.210109
## f.prev_contacted=Contacted 1.108220e-03  -3.261506
## f.age=f.age-[17,30]        3.059024e-04  -3.610250
## f.contact=f.telephone      3.372737e-24 -10.148275
## f.season=Spring            9.752018e-64 -16.854322
## f.season=Summer            4.859935e-72 -17.949300
##
## $`6`
##                               Cla/Mod      Mod/Cla      Global      p.value
## f.season=Summer            15.152858   85.929648  45.330388  7.576666e-69
## f.prev_contacted=No-contacted 8.338571 100.000000 95.862623  2.408690e-08
## f.contact=f.telephone      10.680686   48.492462 36.292428  2.141525e-07
## f.day=f.day.tue            10.179641   25.628141 20.124523  5.414381e-03
## f.housing=f.no             8.951867   50.000000 44.647520  2.564174e-02
## f.housing=f.yes            7.220610   50.000000 55.352480  2.564174e-02
## f.day=f.day.fri            5.938494   14.070352 18.939546  7.954203e-03
## f.contact=f.cellular       6.462799   51.507538 63.707572  2.141525e-07
## f.prev_contacted=Contacted 0.000000   0.000000  4.137377  2.408690e-08
## f.season=Aut-Win           1.361573   2.261307 13.275758  1.558582e-15
## f.season=Spring            2.280446   11.809045 41.393854  3.687233e-41
##                               v.test
## f.season=Summer            17.536249
## f.prev_contacted=No-contacted 5.579747
## f.contact=f.telephone      5.186613
## f.day=f.day.tue            2.781287
## f.housing=f.no             2.231595
## f.housing=f.yes            -2.231595
## f.day=f.day.fri            -2.654008
## f.contact=f.cellular       -5.186613
## f.prev_contacted=Contacted -5.579747
## f.season=Aut-Win           -7.972213
## f.season=Spring            -13.436657
##
## $`7`
##                               Cla/Mod      Mod/Cla      Global
## f.contact=f.telephone      38.074156   65.5238095 36.292428
## f.season=Spring            32.508491   63.8095238 41.393854
## f.prev_contacted=No-contacted 21.998743 100.0000000 95.862623
## f.marital=f.married        23.851999   68.7619048 60.795340
## f.age=f.age-(40,50]        25.441413   30.1904762 25.025105
## f.day=f.day.tue            25.149701   24.0000000 20.124523
## f.jobsituation=Worker       22.826476   60.7619048 56.135770
## f.day=f.day.mon            18.639053   18.0000000 20.365535

```

```

## f.jobssituation=Other      18.746373  30.7619048 34.605342
## f.day=f.day.thu           17.560976  17.1428571 20.586463
## f.marital=f.single        15.565032  20.8571429 28.258686
## f.season=Summer           16.526362  35.5238095 45.330388
## f.age=f.age-[17,30]      11.199095   9.4285714 17.754569
## f.prev_contacted=Contacted 0.000000  0.0000000  4.137377
## f.season=Aut-Win          1.059002   0.6666667 13.275758
## f.contact=f.cellular      11.412358  34.4761905 63.707572
##                               p.value      v.test
## f.contact=f.telephone     2.097715e-105 21.804613
## f.season=Spring           3.474101e-61 16.503295
## f.prev_contacted=No-contacted 2.007732e-22  9.741398
## f.marital=f.married       1.825860e-09  6.012587
## f.age=f.age-(40,50]       1.824763e-05  4.285320
## f.day=f.day.tue           5.175943e-04  3.471482
## f.jobssituation=Worker     6.513133e-04  3.409294
## f.day=f.day.mon           3.080611e-02 -2.159568
## f.jobssituation=Other      3.042168e-03 -2.963445
## f.day=f.day.thu           1.631759e-03 -3.150169
## f.marital=f.single        8.962670e-10 -6.126866
## f.season=Summer           4.723635e-13 -7.233022
## f.age=f.age-[17,30]      4.141770e-17 -8.408814
## f.prev_contacted=Contacted 2.007732e-22 -9.741398
## f.season=Aut-Win          8.694864e-62 -16.586714
## f.contact=f.cellular      2.097715e-105 -21.804613
##
## $`8`
##                               Cla/Mod    Mod/Cla    Global      p.value
## f.season=Summer           56.00354   85.52097 45.330388 2.869286e-319
## f.contact=f.telephone     40.89651   50.00000 36.292428 2.765253e-38
## f.prev_contacted=No-contacted 30.96585 100.00000 95.862623 4.873451e-33
## f.education=Mandatory     31.54562   57.30717 53.926491 1.860217e-03
## f.housing=f.no            31.44399   47.29364 44.647520 1.479675e-02
## f.jobssituation=Worker     30.87657   58.38972 56.135770 3.721353e-02
## f.jobssituation=Self-employed 33.83948 10.55480  9.258887 4.238536e-02
## f.housing=f.yes           28.26560   52.70636 55.352480 1.479675e-02
## f.day=f.day.tue           26.44711   17.92963 20.124523 1.154144e-02
## f.education=Non-Mandatory 27.27709   38.49797 41.895963 1.561059e-03
## f.jobssituation=Other      26.63958   31.05548 34.605342 5.914279e-04
## f.prev_contacted=Contacted 0.000000  0.000000  4.137377 4.873451e-33
## f.contact=f.cellular      23.29760   50.00000 63.707572 2.765253e-38
## f.season=Aut-Win          0.000000  0.000000 13.275758 7.705202e-111
## f.season=Spring           10.38331   14.47903 41.393854 1.235639e-151
##                               v.test
## f.season=Summer           38.199510
## f.contact=f.telephone     12.937488
## f.prev_contacted=No-contacted 11.973816
## f.education=Mandatory     3.111687
## f.housing=f.no            2.437316
## f.jobssituation=Worker     2.083414
## f.jobssituation=Self-employed 2.029717
## f.housing=f.yes           -2.437316
## f.day=f.day.tue           -2.525863
## f.education=Non-Mandatory -3.163087

```

```

## f.jobssituation=Other          -3.435515
## f.prev_contacted=Contacted    -11.973816
## f.contact=f.cellular          -12.937488
## f.season=Aut-Win              -22.370049
## f.season=Spring               -26.229169
##
##
## Link between the cluster variable and the quantitative variables
## =====
##                               Eta2 P-value
## campaign      0.4922609      0
## pdays        0.6167808      0
## previous      0.3030681      0
## emp.var.rate  0.9672291      0
## cons.price.idx 0.6084482      0
## cons.conf.idx  0.6521304      0
## euribor3m     0.9945648      0
## nr.employed   0.8854108      0
##
## Description of each cluster by quantitative variables
## =====
## $`1`
##               v.test Mean in category Overall mean sd in category
## pdays        52.866518      11.1176471    5.3888741    1.9965368
## previous      13.984999      1.1030733    0.1469431    0.4115099
## cons.price.idx -3.398825     93.2513235   93.5831516    0.7591678
## emp.var.rate   -7.860980     -1.9823529   0.1023700    0.5506523
## euribor3m     -8.954788      0.9998529    3.6389677    0.2525389
## nr.employed   -9.859286     5047.9411765 5168.1679454   50.1994910
##               Overall sd      p.value
## pdays        0.6339641 0.000000e+00
## previous      0.3999801 1.924809e-44
## cons.price.idx 0.5711736 6.767606e-04
## emp.var.rate   1.5515130 3.811407e-15
## euribor3m     1.7241963 3.403881e-19
## nr.employed   71.3410455 6.249222e-23
##
## $`2`
##               v.test Mean in category Overall mean sd in category
## cons.conf.idx  30.151936     -33.4791789 -40.5975497    5.2193589
## previous      27.786925      0.7278942    0.1469431    0.7377902
## campaign      -8.145421      1.5777126    2.2706305    0.9304094
## pdays       -17.571354      4.8065957    5.3888741    1.2570502
## cons.price.idx -25.142023     92.8325161   93.5831516    0.6401125
## euribor3m    -31.067188      0.8390235    3.6389677    0.2573116
## emp.var.rate  -34.052249     -2.6592375    0.1023700    0.7328957
## nr.employed   -35.463784     5035.9211144 5168.1679454   36.8844199
##               Overall sd      p.value
## cons.conf.idx  4.5165276 1.011749e-199
## previous      0.3999801 6.241502e-170
## campaign      1.6274493 3.779638e-16
## pdays        0.6339641 4.083064e-69
## cons.price.idx 0.5711736 1.727655e-139
## euribor3m     1.7241963 6.685843e-212

```



```

## emp.var.rate      1.5515130 3.759369e-254
## nr.employed       71.3410455 1.778609e-275
##
## $`3`
##               v.test Mean in category Overall mean sd in category
## previous      14.19336      0.2872673    0.1469431    0.4909763
## pdays          2.40260      5.4265232    5.3888741    0.2314812
## cons.price.idx -35.49852     93.0819789   93.5831516    0.4123410
## cons.conf.idx  -43.29709    -45.4311688  -40.5975497    2.9919241
## nr.employed    -47.60642    5084.2192370 5168.1679454   37.2331422
## emp.var.rate   -50.59823     -1.8380682    0.1023700    0.3208205
## euribor3m      -55.54677      1.2716607    3.6389677    0.1791309
##               Overall sd      p.value
## previous        0.3999801   1.007190e-45
## pdays            0.6339641   1.627900e-02
## cons.price.idx   0.5711736   5.180812e-276
## cons.conf.idx    4.5165276   0.000000e+00
## nr.employed      71.3410455   0.000000e+00
## emp.var.rate     1.5515130   0.000000e+00
## euribor3m        1.7241963   0.000000e+00
##
## $`4`
##               v.test Mean in category Overall mean sd in category
## nr.employed      5.332697    5195.69402174 5168.1679454    1.43364734
## euribor3m        3.624386      4.09111413    3.6389677    0.06933538
## campaign         2.917167      2.61413043    2.2706305    0.94869264
## previous        -2.823985      0.06521739    0.1469431    0.24690906
## cons.conf.idx    -4.356517    -42.02119565  -40.5975497    0.28672947
## cons.price.idx   -9.329774     93.19758696   93.5831516    0.03264305
##               Overall sd      p.value
## nr.employed      71.3410455   9.676480e-08
## euribor3m        1.7241963   2.896483e-04
## campaign         1.6274493   3.532267e-03
## previous         0.3999801   4.743056e-03
## cons.conf.idx    4.5165276   1.321487e-05
## cons.price.idx   0.5711736   1.060965e-20
##
## $`5`
##               v.test Mean in category Overall mean sd in category
## previous        6.585110      0.3053435    0.1469431    0.46055277
## nr.employed      6.440489    5195.8000000 5168.1679454    0.00000000
## euribor3m        4.708953      4.1272443    3.6389677    0.09619296
## emp.var.rate     -2.168881     -0.1000000    0.1023700    0.00000000
## cons.conf.idx    -5.163307    -42.0000000  -40.5975497    0.00000000
## cons.price.idx  -11.154438     93.2000000   93.5831516    0.00000000
## campaign        -11.500555      1.1450382    2.2706305    0.43898108
##               Overall sd      p.value
## previous        0.3999801   4.545480e-11
## nr.employed      71.3410455   1.190890e-10
## euribor3m        1.7241963   2.489922e-06
## emp.var.rate     1.5515130   3.009170e-02
## cons.conf.idx    4.5165276   2.426246e-07
## cons.price.idx   0.5711736   6.812203e-29
## campaign         1.6274493   1.310700e-30

```

```

##
## $`6`
##          v.test Mean in category Overall mean sd in category
## campaign      46.564277      5.914573    2.2706305    1.5491086
## emp.var.rate   16.463871      1.330653    0.1023700    0.2381146
## nr.employed    15.980573    5222.988442  5168.1679454   12.6493020
## euribor3m      15.464656      4.921116    3.6389677    0.1350120
## cons.price.idx 14.693416     93.986706   93.5831516    0.3529350
## previous      -7.640111      0.000000    0.1469431    0.0000000
##          Overall sd      p.value
## campaign      1.6274493 0.000000e+00
## emp.var.rate   1.5515130 6.669491e-61
## nr.employed    71.3410455 1.745356e-57
## euribor3m      1.7241963 6.009340e-54
## cons.price.idx 0.5711736 7.104069e-49
## previous       0.3999801 2.170339e-14
##
## $`7`
##          v.test Mean in category Overall mean sd in category
## cons.conf.idx   34.54314     -36.320095  -40.5975497    0.36366650
## euribor3m       26.59982      4.896399    3.6389677    0.05189715
## emp.var.rate    25.77010      1.198571    0.1023700    0.17847998
## nr.employed     18.42753    5204.211333  5168.1679454   17.73605193
## cons.price.idx   13.67580     93.797312   93.5831516    0.26273487
## previous       -13.39960      0.000000    0.1469431    0.00000000
## campaign       -18.01188      1.466946    2.2706305    0.51895636
##          Overall sd      p.value
## cons.conf.idx   4.5165276 1.806939e-261
## euribor3m       1.7241963 6.820926e-156
## emp.var.rate    1.5515130 1.919347e-146
## nr.employed     71.3410455 7.900327e-76
## cons.price.idx   0.5711736 1.416375e-42
## previous        0.3999801 6.078735e-41
## campaign        1.6274493 1.571975e-72
##
## $`8`
##          v.test Mean in category Overall mean sd in category
## emp.var.rate    37.057622      1.356563    0.1023700    0.10556677
## cons.price.idx   36.412511     94.036832   93.5831516    0.31318902
## nr.employed     35.059537    5222.728281  5168.1679454   13.05509087
## euribor3m       34.526536      4.937555    3.6389677    0.04338448
## cons.conf.idx    -2.171606    -40.811502  -40.5975497    2.67049033
## previous       -16.841449      0.000000    0.1469431    0.00000000
##          Overall sd      p.value
## emp.var.rate    1.5515130 1.353721e-300
## cons.price.idx   0.5711736 2.698575e-290
## nr.employed     71.3410455 2.790375e-269
## euribor3m       1.7241963 3.207593e-261
## cons.conf.idx    4.5165276 2.988543e-02
## previous        0.3999801 1.212306e-63

```

Cluster 1: Es troba la gent jove treballadora que han sigut prèviament contactats

Cluster 2: Indivídues que han estat prèviament contactats i d'edat avançada.

Cluster 3: En aquest cas és difícil veure un patró que tots els individus segueixin.

Cluster 4: En aquest cluster es troba la gent jove que treballa com a autònom.

Cluster 5: Hi trobam individus que no són d'edat avançada, contactats durant la tardor-hivern i que estan divorciats.

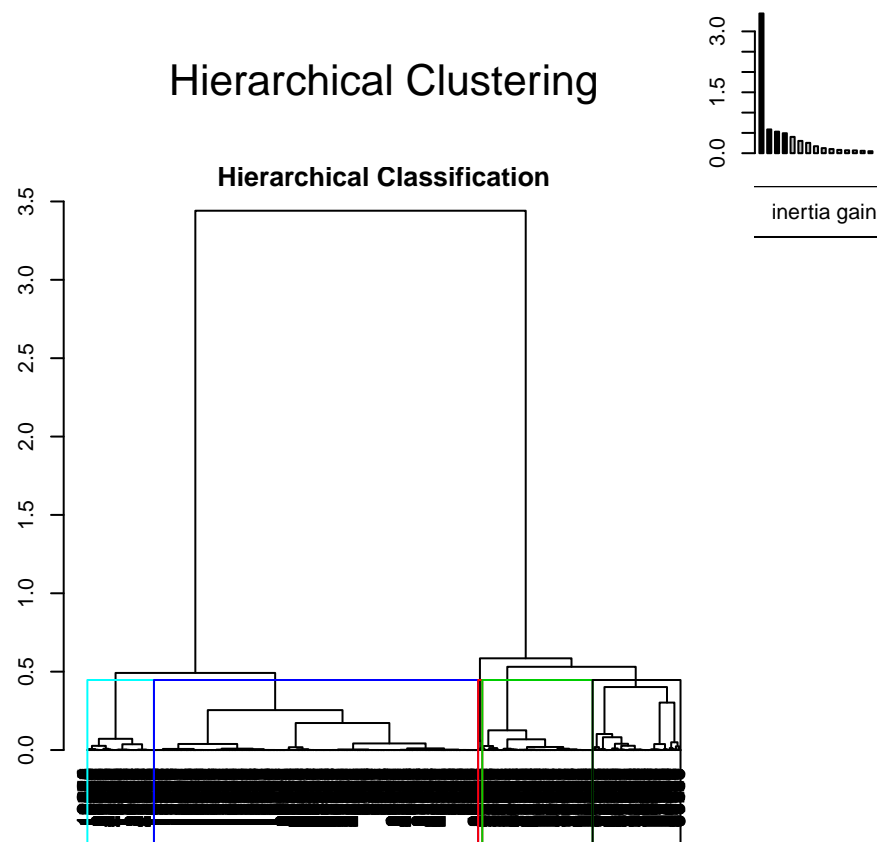
Cluster 6: Aquest cluster segueix la línia del tercer, ja que no es pot veure un patró clar a simple vista.

Cluster 7: Hi podem trobar individus que no són ancians (fins els 50 anys) que no han estat previament contactats.

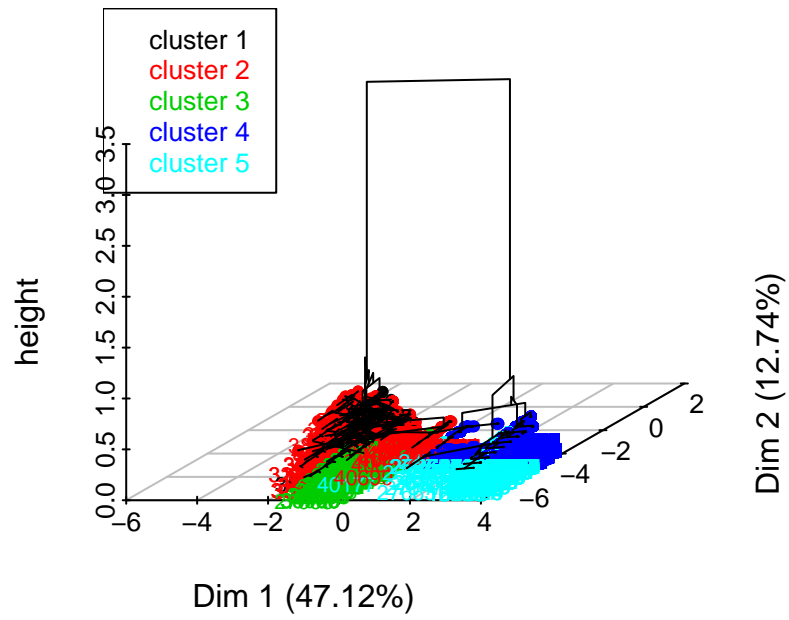
9 Hierarchical Clustering

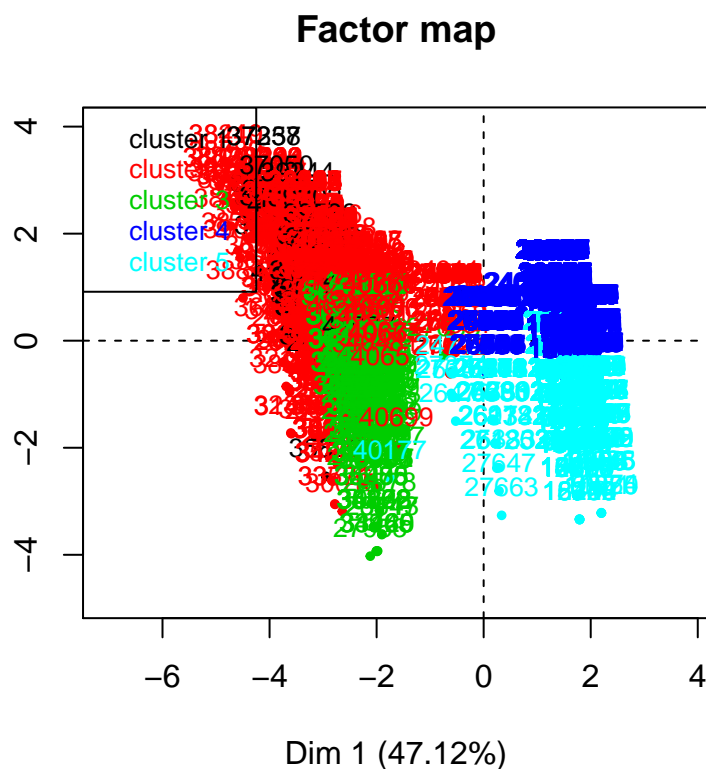
Després de comprovar el nombre optim de cluster a definir interaccionant amb el plot, hem vist que és 5.

```
clust <- HCPC(res.pca, nb.clust = 5, order = TRUE)
```



Hierarchical clustering on the factor map





9.1 Description of clusters

```
clust$desc.var
```

```
##
## Link between the cluster variable and the quantitative variables
## =====
##               Eta2 P-value
## campaign      0.5135931      0
## pdays         0.5736642      0
## previous       0.6463952      0
## emp.var.rate   0.8735959      0
## cons.price.idx 0.4666171      0
## cons.conf.idx  0.2876758      0
## euribor3m      0.9277479      0
## nr.employed    0.8034334      0
##
## Description of each cluster by quantitative variables
## =====
## $`1`
##               v.test Mean in category Overall mean sd in category
## pdays         52.936829      11.212121      5.3888741      1.9502537
## previous        13.384792       1.075894      0.1469431      0.3864648
## cons.price.idx -3.637888     93.222606     93.5831516      0.7521708
## emp.var.rate   -7.843050     -2.009091      0.1023700      0.5367487
## euribor3m      -8.808883       1.003545      3.6389677      0.2554308
```

```

## nr.employed      -9.505769      5050.496970 5168.1679454      48.7263943
## Overall sd      p.value
## pdays          0.6339641 0.000000e+00
## previous        0.3999801 7.420521e-41
## cons.price.idx  0.5711736 2.748824e-04
## emp.var.rate    1.5515130 4.397334e-15
## euribor3m       1.7241963 1.263986e-18
## nr.employed     71.3410455 1.985758e-21
##
## $`2`
## v.test Mean in category Overall mean sd in category
## previous      54.601436      0.8460976      0.1469431      0.5770173
## cons.conf.idx  7.400120     -39.5275735     -40.5975497      6.7866064
## campaign      -6.558706      1.9289216      2.2706305      1.3448693
## pdays        -9.172576      5.2027139      5.3888741      0.9283452
## cons.price.idx -29.371876     93.0460821     93.5831516      0.5792142
## euribor3m     -41.268815      1.3610429      3.6389677      0.9937786
## emp.var.rate   -41.621890     -1.9649510      0.1023700      0.9301312
## nr.employed   -43.247602     5069.3963235   5168.1679454     62.5784272
## Overall sd      p.value
## previous        0.3999801 0.000000e+00
## cons.conf.idx    4.5165276 1.360616e-13
## campaign         1.6274493 5.427680e-11
## pdays           0.6339641 4.618470e-20
## cons.price.idx   0.5711736 1.256157e-189
## euribor3m        1.7241963 0.000000e+00
## emp.var.rate      1.5515130 0.000000e+00
## nr.employed       71.3410455 0.000000e+00
##
## $`3`
## v.test Mean in category Overall mean sd in category
## previous      -11.65178      8.398215e-04      0.1469431      0.01060273
## cons.price.idx -31.01940      9.302772e+01     93.5831516      0.33423662
## nr.employed    -34.67972      5.090607e+03   5168.1679454     27.85533532
## cons.conf.idx  -37.82280     -4.595290e+01     -40.5975497      2.13113449
## emp.var.rate   -40.45508     -1.865325e+00      0.1023700      0.31941892
## euribor3m     -43.15513      1.306323e+00      3.6389677      0.16092179
## Overall sd      p.value
## previous        0.3999801 2.247220e-31
## cons.price.idx   0.5711736 2.951813e-211
## nr.employed       71.3410455 1.592855e-263
## cons.conf.idx    4.5165276 0.000000e+00
## emp.var.rate      1.5515130 0.000000e+00
## euribor3m        1.7241963 0.000000e+00
##
## $`4`
## v.test Mean in category Overall mean sd in category
## euribor3m      52.64690      4.823374      3.6389677      0.2696129
## emp.var.rate    50.51197      1.124935      0.1023700      0.4802150
## nr.employed     48.29887     5213.127050   5168.1679454     17.6645147
## cons.price.idx  37.02986      93.859121     93.5831516      0.3835478
## cons.conf.idx   18.99887     -39.477922     -40.5975497      2.9964714
## campaign       -26.18570      1.714582      2.2706305      0.7436116
## previous       -28.15591      0.000000      0.1469431      0.0000000

```

```

##          Overall sd      p.value
## euribor3m      1.7241963 0.000000e+00
## emp.var.rate    1.5515130 0.000000e+00
## nr.employed     71.3410455 0.000000e+00
## cons.price.idx  0.5711736 3.789174e-300
## cons.conf.idx   4.5165276 1.742638e-80
## campaign        1.6274493 3.867302e-151
## previous        0.3999801 2.028395e-174
##
## $`5`
##          v.test Mean in category Overall mean sd in category
## campaign      49.984100      5.415254e+00      2.2706305      1.49533892
## emp.var.rate   18.750173      1.226949e+00      0.1023700      0.39454223
## euribor3m     18.432707      4.867553e+00      3.6389677      0.27199300
## nr.employed    17.733867      5.217075e+03 5168.1679454     18.89857603
## cons.price.idx 13.490940      9.388103e+01     93.5831516      0.37532374
## cons.conf.idx   6.104696     -3.953169e+01    -40.5975497      3.03354056
## previous       -8.955367      8.474576e-03      0.1469431      0.09166656
##          Overall sd      p.value
## campaign        1.6274493 0.000000e+00
## emp.var.rate     1.5515130 1.929612e-78
## euribor3m        1.7241963 7.179853e-76
## nr.employed      71.3410455 2.296785e-70
## cons.price.idx   0.5711736 1.768316e-41
## cons.conf.idx    4.5165276 1.029966e-09
## previous         0.3999801 3.386078e-19

```

```

clust$desc.axes

```

```

##
## Link between the cluster variable and the quantitative variables
## =====
##          Eta2 P-value
## Dim.1 0.9191789      0
## Dim.2 0.4450225      0
## Dim.3 0.5497222      0
## Dim.4 0.4098434      0
## Dim.5 0.2984519      0
##
## Description of each cluster by quantitative variables
## =====
## $`1`
##          v.test Mean in category Overall mean sd in category Overall sd
## Dim.3 51.398588      8.951013 -7.344825e-13      2.8976185      1.0036410
## Dim.4 11.930668      2.015231  9.709110e-13      1.3810674      0.9734587
## Dim.2  6.944085      1.216613  4.468775e-13      1.2383845      1.0097057
## Dim.1 -10.049127     -3.385357 -3.915432e-12      0.4548905      1.9414825
##          p.value
## Dim.3 0.000000e+00
## Dim.4 8.191371e-33
## Dim.2 3.809211e-12
## Dim.1 9.268528e-24
##
## $`2`
##          v.test Mean in category Overall mean sd in category Overall sd

```

```
## Dim.5 30.34007      0.8354515 -2.563362e-12      1.5992376 0.8601496
## Dim.2 16.83824      0.5442796 4.468775e-13      1.1477283 1.0097057
## Dim.4 16.50144      0.5142450 9.709110e-13      1.1257317 0.9734587
## Dim.3 -11.07431     -0.3558163 -7.344825e-13      1.4938443 1.0036410
## Dim.1 -45.99890     -2.8589830 -3.915432e-12      0.9059308 1.9414825
##
##          p.value
## Dim.5 3.397393e-202
## Dim.2 1.279947e-63
## Dim.4 3.582524e-61
## Dim.3 1.671594e-28
## Dim.1 0.000000e+00
##
## $`3`
##          v.test Mean in category Overall mean sd in category Overall sd
## Dim.4 -23.38948      -0.7137853 9.709110e-13      0.7313931 0.9734587
## Dim.2 -27.41583      -0.8678120 4.468775e-13      0.7683096 1.0097057
## Dim.5 -29.19190      -0.7871652 -2.563362e-12      0.2547806 0.8601496
## Dim.1 -37.36288      -2.2740669 -3.915432e-12      0.1550095 1.9414825
##
##          p.value
## Dim.4 5.467981e-121
## Dim.2 1.776165e-165
## Dim.5 2.456920e-187
## Dim.1 1.566731e-305
##
## $`4`
##          v.test Mean in category Overall mean sd in category Overall sd
## Dim.1 50.714279      1.2847096 -3.915432e-12      0.5067072 1.9414825
## Dim.2 28.208382      0.3716330 4.468775e-13      0.5530912 1.0097057
## Dim.5 -2.183965      -0.0245110 -2.563362e-12      0.3374144 0.8601496
## Dim.4 -18.983718      -0.2411238 9.709110e-13      0.5634599 0.9734587
##
##          p.value
## Dim.1 0.000000e+00
## Dim.2 4.615007e-175
## Dim.5 2.896480e-02
## Dim.4 2.325448e-80
##
## $`5`
##          v.test Mean in category Overall mean sd in category Overall sd
## Dim.4 34.539170      1.29974688 9.709110e-13      0.8023248 0.9734587
## Dim.1 20.413868      1.53210358 -3.915432e-12      0.4658507 1.9414825
## Dim.5 2.633085      0.08755244 -2.563362e-12      0.3695755 0.8601496
## Dim.3 -3.061863      -0.11879371 -7.344825e-13      0.2086902 1.0036410
## Dim.2 -32.677417      -1.27547477 4.468775e-13      0.7901854 1.0097057
##
##          p.value
## Dim.4 2.072702e-261
## Dim.1 1.259146e-92
## Dim.5 8.461313e-03
## Dim.3 2.199642e-03
## Dim.2 3.269755e-234
```

```
#Individus que són més propers i més lluny al centre dels diferents clusters.
clust$desc.ind
```

```
## $para
## Cluster: 1
```



```

##      39728      31073      34770      35219      31241
## 1.305872 1.448076 1.798985 1.799196 1.870738
## -----
## Cluster: 2
##      36450      36741      36890      36686      36836
## 0.6644885 0.7354002 0.7747697 0.8770443 0.8895529
## -----
## Cluster: 3
##      33221      35442      30983      31837      31859
## 0.1280541 0.1383877 0.1450067 0.1464073 0.1464073
## -----
## Cluster: 4
##      1562      1496      1529      1754      1970
## 0.6944152 0.6944408 0.6944408 0.6944408 0.6944408
## -----
## Cluster: 5
##      3890      5049      7050      2762      3306
## 0.7413756 0.7413900 0.7413907 0.7414199 0.7416763
##
## $dist
## Cluster: 1
##      37851      37238      37357      41089      40049
## 15.61562 15.60960 15.60957 13.18075 12.80459
## -----
## Cluster: 2
##      30418      38200      37688      39757      38875
## 8.869425 8.589632 8.126727 7.608229 7.495552
## -----
## Cluster: 3
##      27953      34965      33740      33466      31129
## 4.631202 4.577766 4.569714 4.568274 4.556035
## -----
## Cluster: 4
##      10886      9586      11274      11299      9144
## 2.884114 2.883392 2.882757 2.882757 2.882483
## -----
## Cluster: 5
##      27663      11050      9951      10825      10574
## 4.716676 4.577871 4.577766 4.576886 4.575903

```

En la divisió de clusters per les seves variables podem trobar dos primers clusters altament relacionats amb la variable previous, persones que havien estat contactades anteriorment. En el cas del primer, també es important veure la relació que el cluster té amb els dies que han passat des de l'últim contacte. Del segon cluster també obtenim relacions socioeconòmiques negatives i amb una important presència.

Per altra banda el cluster numero 3 unicament obté relacions en el sentit negatiu i principalment de variables socioeconòmiques. Totalment el contrari que el cluster numero 4.

Acabem amb el cluster 5. En aquest cas els individus majoritariament havien estat contactats més d'una vegada durant la campanya actual, aquests també gaudeixen d'una relació positiva amb les variables socioeconòmiques.

Pel que fa als eixos resumidament podem dir que el cluster numero 1 està ben representat en la dimensió 3 i 4. El cluster 2 està ben representat en la dimensió 2,4 i 5 però molt malt representat en la dimensió 1. Seguidament el cluster 3 veiem que estpa mal representat en totes les dimensions existents. El 4 gaudeix de bona representació en les dues primeres dimencions. I per acabar el cluster numero 5 gaudeix de bona

representació en la dimensió 4 i 1 per contra no li beneficia la dimensió 2.

```
summary(df$duration)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      4.0   103.0   181.0   256.6   326.0  1499.0

qulist<-quantile(df$duration,seq(0,1,0.125),na.rm=TRUE)
df$f.duration<-factor(cut(df$duration,breaks=c(4,64,103,140,181,235,326,496,1499),include.lowest=T))

summary(df$f.duration)

##      [4,64]      (64,103]      (103,140]      (140,181]      (181,235]
##      576      672      627      629      609
##      (235,326]      (326,496]      (496,1.5e+03]
##      622      621      623

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

10 CA analysis

Per a realitzar l'anàlisi de correspondència simples utilitzarem dos factors sorgits de la neteja de les dades. En aquest cas farem l'estudi amb el factor f.jobituation i el factor f.age.

10.1 f.duration - f.job

```
t <- table(df$f.job, df$f.duration); t;

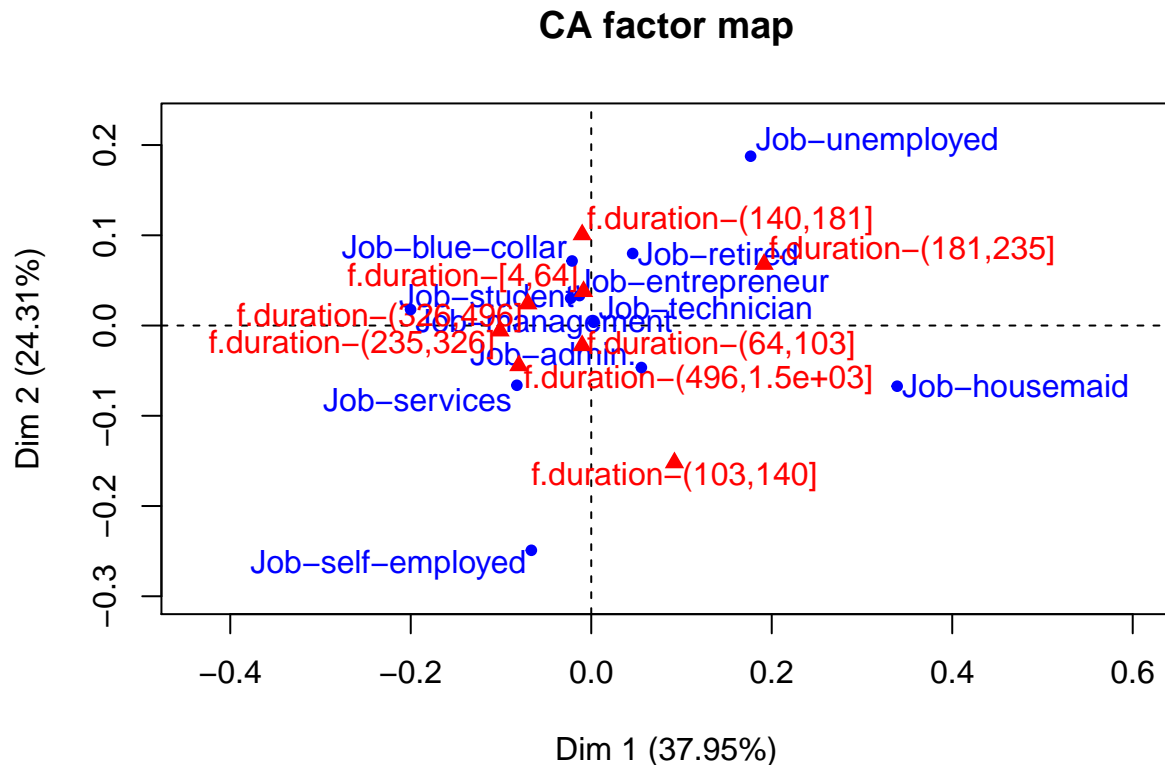
##
##      f.duration-[4,64] f.duration-(64,103]
##  Job-admin.           153           170
##  Job-blue-collar       138           149
##  Job-entrepreneur       21           26
##  Job-housemaid         13           22
##  Job-management        43           53
##  Job-retired           16           25
##  Job-self-employed     13           20
##  Job-services          51           68
##  Job-student           8           11
##  Job-technician       104          114
##  Job-unemployed        16           14
##
##      f.duration-(103,140] f.duration-(140,181]
##  Job-admin.           195           154
##  Job-blue-collar       128           174
##  Job-entrepreneur       18           23
##  Job-housemaid         24           14
##  Job-management        34           46
##  Job-retired           21           32
##  Job-self-employed     28           18
##  Job-services          58           47
##  Job-student          10            8
##  Job-technician       101           96
##  Job-unemployed        10           17
##
```

```
##          f.duration-(181,235] f.duration-(235,326]
## Job-admin.                168                163
## Job-blue-collar           144                155
## Job-entrepreneur           23                 23
## Job-housemaid              26                 11
## Job-management             22                 53
## Job-retired                26                 17
## Job-self-employed          11                 17
## Job-services               48                 63
## Job-student               16                 16
## Job-technician            103                 96
## Job-unemployed             22                  8
##
##          f.duration-(326,496] f.duration-(496,1.5e+03]
## Job-admin.                154                146
## Job-blue-collar           146                141
## Job-entrepreneur           17                 28
## Job-housemaid              12                  9
## Job-management             50                 42
## Job-retired                22                 27
## Job-self-employed          14                 30
## Job-services               65                 69
## Job-student               18                 12
## Job-technician            111                104
## Job-unemployed             12                 15
```

```
chisq.test(t)
```

```
##
## Pearson's Chi-squared test
##
## data:  t
## X-squared = 108.72, df = 70, p-value = 0.002084
```

```
t <- CA(t)
```



Com que el p-value és més petit que 0,05 rebutgem la hipòtesis de independència entre els dos factors. Trobem una certa relació entre les trucades amb més durada i els treballadors en el sector dels serveis. Per altre banda, retirats i estudiants son els que més relació tenen amb les trucades curtes.

10.1.1 Eigenvalues and dominant axes analysis

```
t(t$eig)
```

```
##                dim 1        dim 2        dim 3
## eigenvalue      0.008287759  0.005308914  0.003718622
## percentage of variance 37.954044478 24.312333499 17.029540434
## cumulative percentage of variance 37.954044478 62.266377977 79.295918411
##                dim 4        dim 5        dim 6
## eigenvalue      0.002121133  0.001192038  0.000707237
## percentage of variance 9.713794962  5.458972020  3.238813160
## cumulative percentage of variance 89.009713373 94.468685393 97.707498554
##                dim 7
## eigenvalue      5.005975e-04
## percentage of variance 2.292501e+00
## cumulative percentage of variance 1.000000e+02
```

En aquest punt escollim utilitzar les 4 primeres dimensions, aquestes superen en escriure el límit del 80% de representació:

```
t$col$cos2[,1:4]
```

```
##                Dim 1        Dim 2        Dim 3        Dim 4
## f.duration-[4,64] 0.007255284 0.143539053 8.759853e-02 0.10367470
```

```
## f.duration-(64,103]      0.017662737 0.084975935 8.333259e-03 0.01217276
## f.duration-(103,140]    0.251000251 0.681835067 3.673305e-05 0.04971561
## f.duration-(140,181]    0.004478179 0.451220658 2.990549e-01 0.19867418
## f.duration-(181,235]    0.822078155 0.103433663 5.339270e-03 0.06100023
## f.duration-(235,326]    0.567974086 0.001926722 2.330852e-01 0.01501167
## f.duration-(326,496]    0.318967399 0.037656024 3.891893e-01 0.11573080
## f.duration-(496,1.5e+03] 0.254086183 0.077750091 4.603492e-01 0.19768454
```

Destaquem que la dimensió 3 i 4 realitzen una representació molt lleu de les nostres dades. En canvi la dimensió 1 té en compte les trucades de durada mitjana i en canvi en la tercera dimensió i juguen un paper bastant important les trucades de major durada de la mostra.

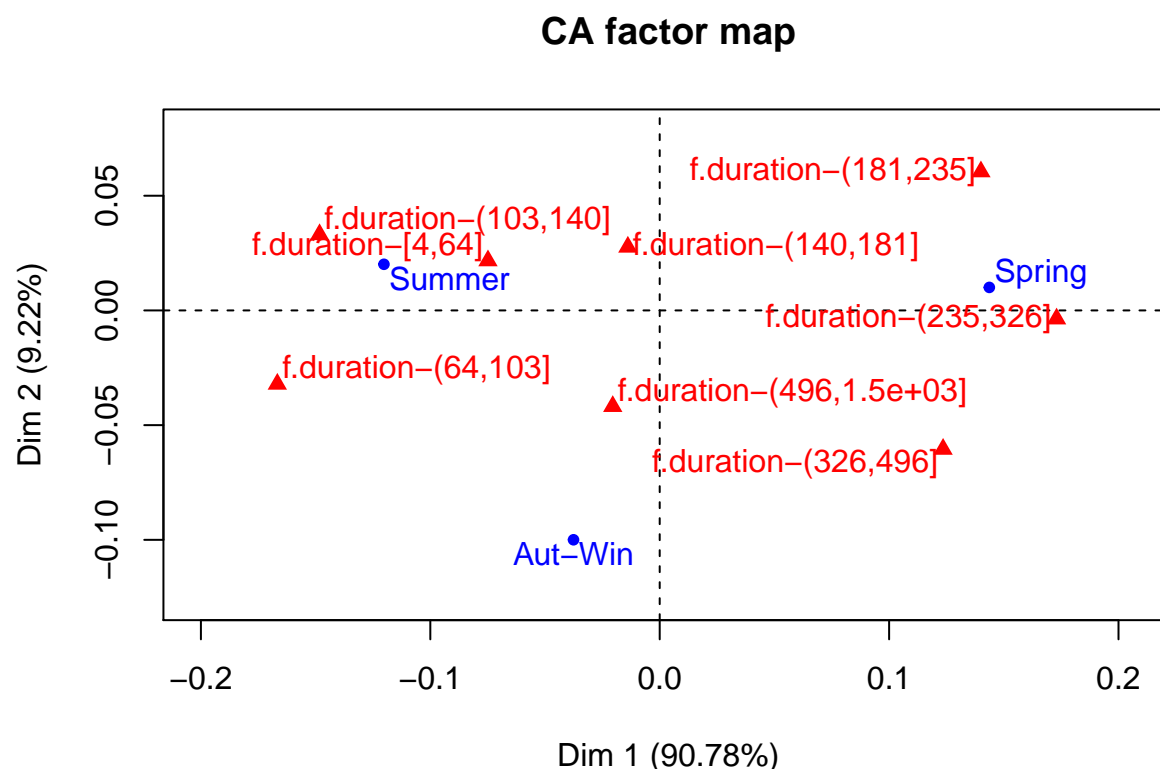
10.2 f.duration - f.age

```
t <- table(df$f.season, df$f.duration); t;
```

```
##
##          f.duration-[4,64] f.duration-(64,103] f.duration-(103,140]
## Spring                219                222                217
## Summer                283                349                330
## Aut-Win                74                101                80
##
##          f.duration-(140,181] f.duration-(181,235] f.duration-(235,326]
## Spring                258                297                309
## Summer                293                247                234
## Aut-Win                78                65                79
##
##          f.duration-(326,496] f.duration-(496,1.5e+03]
## Spring                290                249
## Summer                239                282
## Aut-Win                92                92
```

```
chisq.test(t)
```

```
##
## Pearson's Chi-squared test
##
## data:  t
## X-squared = 83.826, df = 14, p-value = 5.487e-12
t <- CA(t)
```



Com que el p-value és més petit que 0,05 rebutgem la hipòtesis de independència entre els dos factors. En conjunt podem veure que al voltant de la primavera es situen les trucades amb una durada superior. En canvi, a l'estiu passa totalment el contrari.

10.2.1 Eigenvalues and dominant axes analysis

```
t(t$eig)
```

```
##                dim 1      dim 2
## eigenvalue      0.01528409 1.551901e-03
## percentage of variance 90.78224362 9.217756e+00
## cumulative percentage of variance 90.78224362 1.000000e+02
```

En aquest punt escollim utilitzar les dues úniques dimensions per a poder fer-ne una posterior valoració.

```
t$col$cos2[,1:2]
```

```
##                Dim 1      Dim 2
## f.duration-[4,64]  0.9225859 0.0774141059
## f.duration-(64,103) 0.9642447 0.0357552505
## f.duration-(103,140) 0.9525546 0.0474454013
## f.duration-(140,181) 0.2014826 0.7985173967
## f.duration-(181,235) 0.8427807 0.1572192538
## f.duration-(235,326) 0.9995529 0.0004471099
## f.duration-(326,496) 0.8066524 0.1933475905
## f.duration-(496,1.5e+03) 0.1932805 0.8067194776
```

Tot i que la primera dimensió representa en gran escala la major part de les dades, podem comentar que en la dimensió número dos es veuen molt millor representades les trucades amb una major durada i també les de

durades intermitges.