Deliverable1

Montserrrat Martinez i Aleix Costa

21 de febrero de 2019

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')

Package loading and set Working directory

Carreguem els paquets necessaris i definim el nostre directori de treball

Loading data

Upload and select data

A partir del banc de dades proposat, hem de seleccionar una mostra de 5000 registres de manera aleatoria per poder començar a analitzar les nostres dades

```
#setwd("C:/Users/montserrat.martinez.santamaria/Documents/ADEI/bank-
additional/bank-additional")
#dirwd<-"C:/Users/montserrat.martinez.santamaria/Documents/ADEI/bank-
additional/bank-additional"

setwd("/Users/montsee/Desktop/ADEI/bank-additional/bank-additional")
dirwd<-"/Users/montsee/Desktop/ADEI/bank-additional/bank-additional"

# Data file already

df<-read.table(paste0(dirwd,"/bank-additional-
full.csv"),header=TRUE,sep=";",na.strings = "999")

# Select your 5000 register sample (random sample)

#nrow(df)
#ncol(df)
#dim(df)

set.seed(25071997)
mostra<-as.vector(sort(sample(1:nrow(df),5000)))
df<-df[mostra,]

#Verificacio i guardat de la mostra
```

```r
dim(df) #Mostra la dimensi? de la mostra
```

```
## [1] 5000   21
```

```r
names(df) #Mostra els noms de les variables de la mostra
```

```
##  [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"
```

```r
summary(df)
```

```
##       age                 job           marital
##  Min.   :17.00   admin.     :1315   divorced: 574
##  1st Qu.:32.00   blue-collar:1157   married :3029
##  Median :38.00   technician : 789   single  :1390
##  Mean   :40.16   services   : 477   unknown :   7
##  3rd Qu.:47.00   management : 348
##  Max.   :98.00   retired    : 212
##                  (Other)    : 702
##                education       default        housing          loan
##  university.degree  :1503   no     :3958   no     :2206   no     :4055
##  high.school        :1133   unknown:1042   unknown: 129   unknown: 129
##  basic.9y           : 765   yes    :   0   yes    :2665   yes    : 816
##  professional.course: 600
##  basic.4y           : 514
##  basic.6y           : 268
##  (Other)            : 217
##       contact          month      day_of_week    duration
##  cellular :3148   may    :1633   fri: 979    Min.   :   1.0
##  telephone:1852   jul    : 911   mon:1039    1st Qu.: 102.0
##                   aug    : 754   thu:1064    Median : 180.0
##                   jun    : 663   tue: 911    Mean   : 264.7
##                   nov    : 514   wed:1007    3rd Qu.: 329.0
##                   apr    : 282               Max.   :3253.0
##                   (Other): 243
##     campaign          pdays           previous            poutcome
##  Min.   : 1.000   Min.   : 0.000   Min.   :0.000   failure    : 502
##  1st Qu.: 1.000   1st Qu.: 3.000   1st Qu.:0.000   nonexistent:4330
##  Median : 2.000   Median : 5.000   Median :0.000   success    : 168
##  Mean   : 2.598   Mean   : 5.821   Mean   :0.169
##  3rd Qu.: 3.000   3rd Qu.: 6.000   3rd Qu.:0.000
##  Max.   :40.000   Max.   :20.000   Max.   :5.000
##                   NA's   :4816
##   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 :93.92   Median :-41.80   Median :4.857
##  Mean   : 0.1184   Mean   :93.59   Mean   :-40.45   Mean   :3.661
##  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 :4394
##  1st Qu.:5099   yes: 606
##  Median :5191
##  Mean   :5168
##  3rd Qu.:5228
##  Max.   :5228
##
```

```
save.image("DadesBank_5000.RData")
```

Inicialització dels vectors de missings, errors i outliers

Inicialitzarem tres vectors per poder tenir un recompte del total dels errors, missings i outliers:

```
num_total_missings<-rep(0,21)
num_total_errors<-rep(0,21)
num_total_outliers<-rep(0,21)
```

Inicialitzem les variables de contadors individuals per missings, errors i outliers:

```
df$missings_indiv <- 0
df$errors_indiv <- 0
df$outliers_indiv <- 0
```

Univariate Descriptive Analysis & Data Quality Report

Qualitative Variables (Factors) / Categorical

Hem de fer un analisi de totes les variables per poder identificar missings, errors i els outliers. Tamba tractarem de factoritzar cada variable per a que sigui mes facil entendre la mostra

2. Job

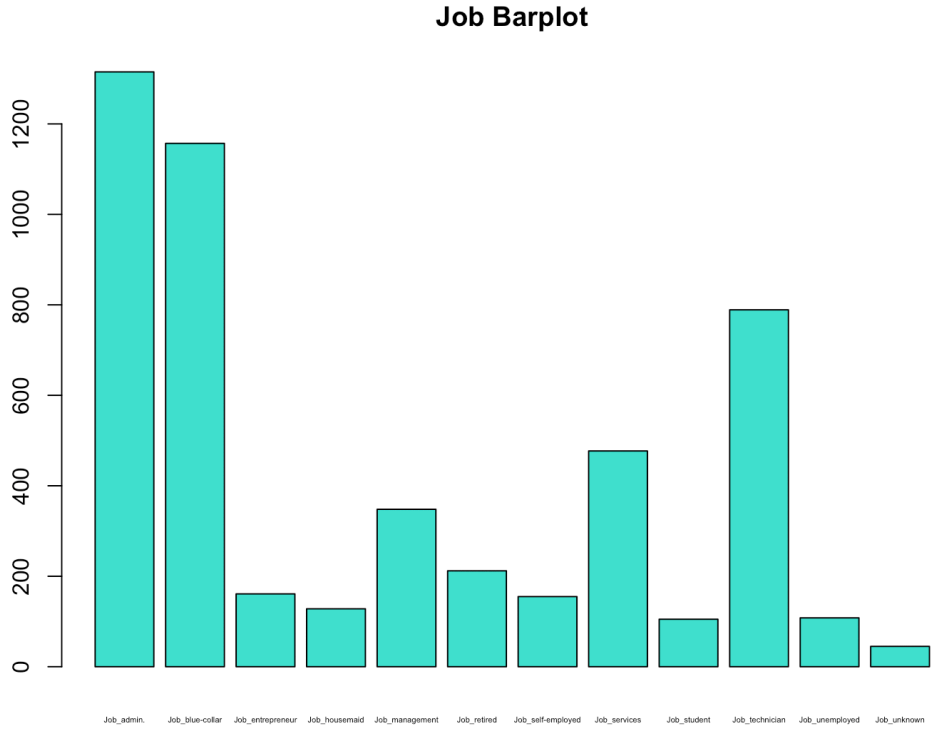
Type of job?

```
df$job<-factor(df$job)
levels(df$job)<-paste("Job_",sep="",levels(df$job))
summary(df$job)
```

```
##         Job_admin.    Job_blue-collar  Job_entrepreneur      Job_housemaid
##               1315               1157               161               128
##    Job_management       Job_retired Job_self-employed      Job_services
##                348               212               155               477
##       Job_student    Job_technician    Job_unemployed       Job_unknown
##                105               789               108                45
```

```
barplot(summary(df$job),main="Job Barplot",col = "turquoise",cex.names=0.35)
```

```
#Amb la comanda "factor" el que estem fent és factoritzar la variable que li passem i el valor que surt amb el "levels" és el numero total de les nostres 5000 observacions que tenen cada tipus de job i com podem veure tots els factors tenen valor i no tenim cap NA (data missing)
```



3. Marital

Marital status?

```
df$marital<-factor(df$marital)
levels(df$marital)<-paste("Marital_",sep="",levels(df$marital))
summary(df$marital)
```

```
## Marital_divorced  Marital_married   Marital_single  Marital_unknown
##              574             3029             1390                7
```

```
barplot(summary(df$marital),main="Marital Barplot",col = "turquoise")
```

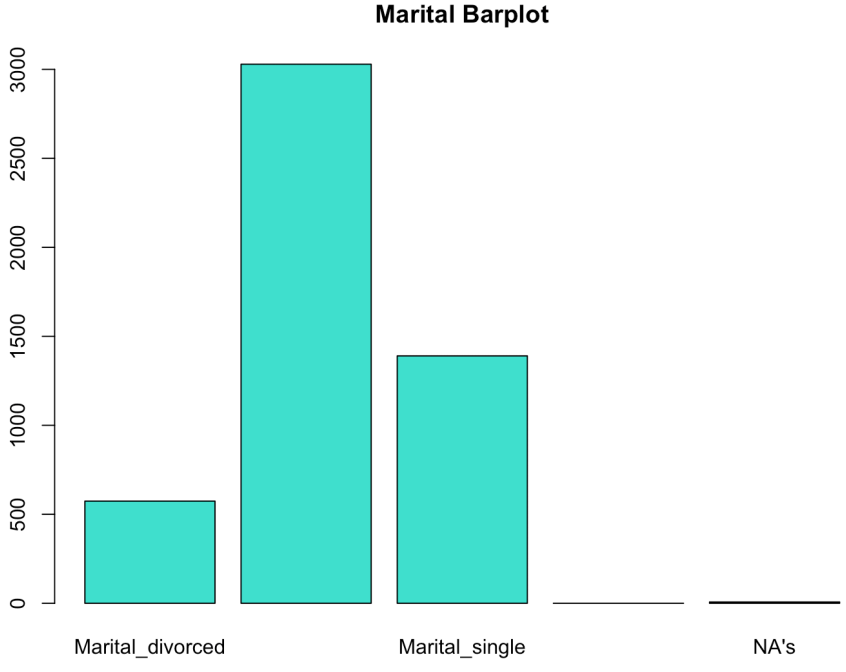
```
sel<-which(df$marital=="Marital_unknown");length(sel)
```

```
## [1] 7
```

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

```
## Marital_divorced  Marital_married   Marital_single  Marital_unknown
##              574             3029             1390                0
##             NA's
##                7
```

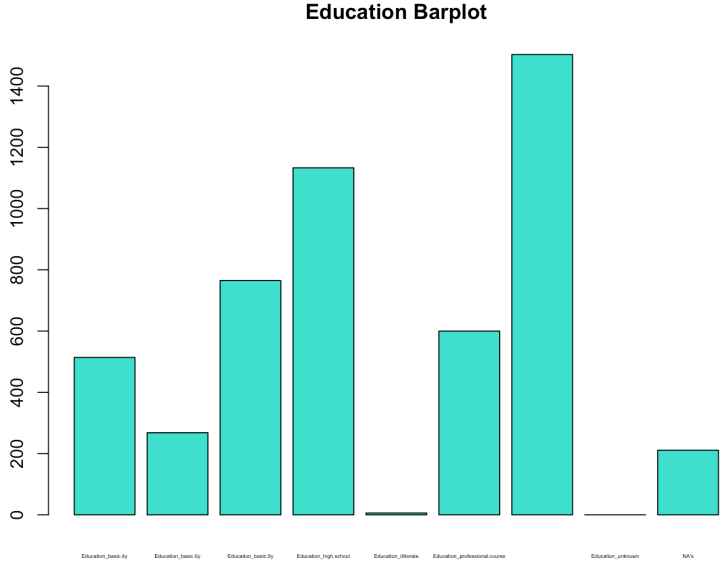
#Podem veure que de la nostra mostra no tenim cap factor incorrecte i com en la nostra mostra la variable "marital_unkown" és molt petita s'han de posar com a NA



4. Education

Type of education?

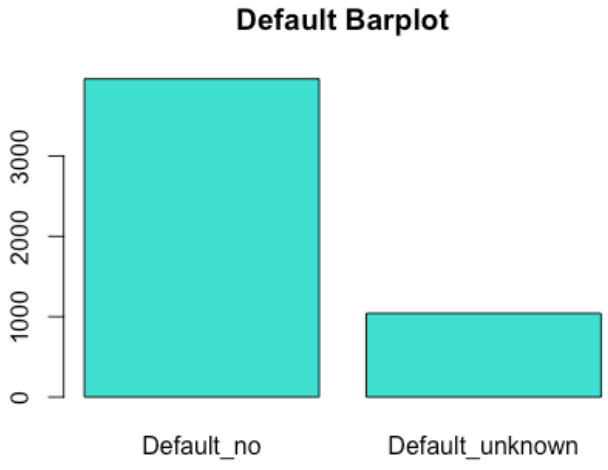
```
df$education<-factor(df$education)
levels(df$education)<-paste("Education_",sep="",levels(df$education))

barplot(summary(df$education),main="Education
Barplot",col="turquoise",cex.names = 0.3)

sel<-which(df$education=="Education_unknown");length(sel)

## [1] 211

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

##            Education_basic.4y            Education_basic.6y
##                           514                           268
##            Education_basic.9y         Education_high.school
##                           765                          1133
##          Education_illiterate Education_professional.course
##                             6                           600
##   Education_university.degree             Education_unknown
##                          1503                             0
##                          NA's
##                           211

#Quan observem tots els factors ens podem adonar que no hi ha cap NA (data
missing) ni cap factor no contemplat, llavors no tenim cap error, però els
unknown els posem com a NA's.
```



5. Default

Has credit in default?

```
df$default<-factor(df$default)
levels(df$default)<-paste("Default_",sep="",levels(df$default))
summary(df$default)

##      Default_no Default_unknown
##            3958            1042

barplot(summary(df$default),main="Default Barplot",col = "turquoise")
```

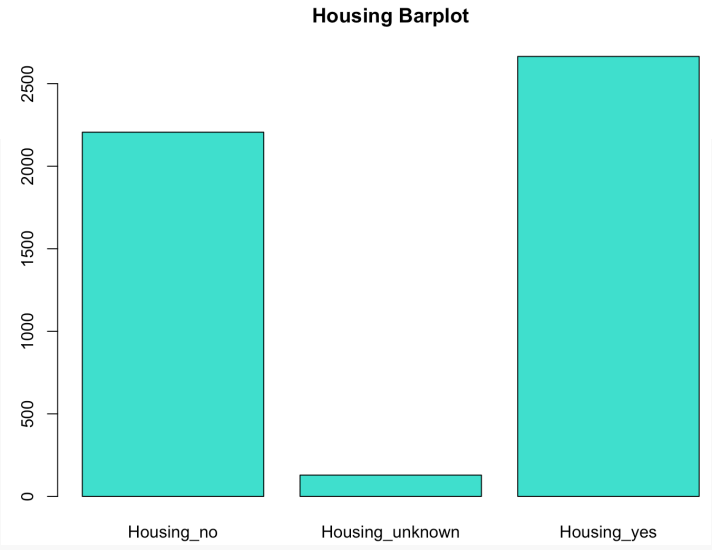



```
#Quan acabem d'analitzar la mostra veiem que com en els casos anteriors no tenim cap NA (data missing) ni cap factor incomplet, llavors la nostra mostra és correcta i com en els casos anteriors hem posat nom al nostre barplot per tenir una millor visualització
```

6. Housing

Has housing loan?

```
df$housing<-factor(df$housing)
levels(df$housing)<-paste("Housing_",sep="",levels(df$housing))
summary(df$housing)

##      Housing_no Housing_unknown     Housing_yes
##            2206             129            2665

barplot(summary(df$housing),main="Housing Barplot",col = "turquoise")
```




```
#Com podem veure anteriorment tampoc tenim cap data missing ni cap factor amb valors estranys, però podem veure que el factor "Housing_unknown" podria ser un possible outlier
```

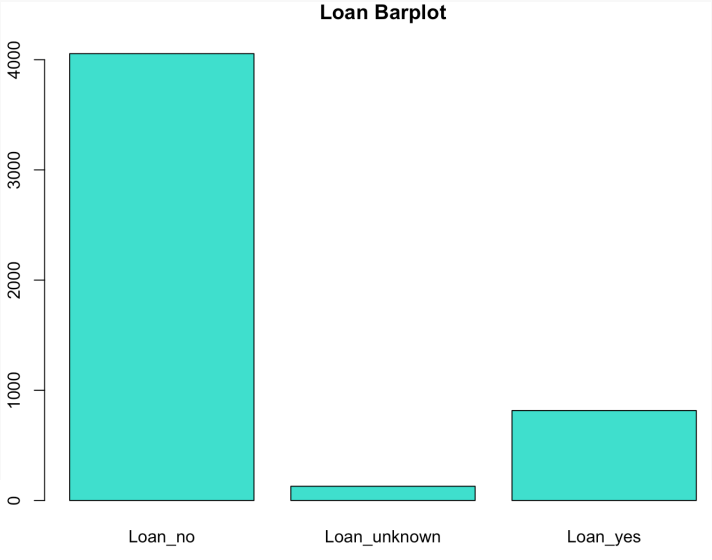
7. Loan

Has personal loan?

```
df$loan<-factor(df$loan)
levels(df$loan)<-paste("Loan_",sep="",levels(df$loan))
summary(df$loan)
```

```
##       Loan_no Loan_unknown      Loan_yes 
##          4055           129           816 
```

```
barplot(summary(df$loan),main="Loan Barplot",col = "turquoise")
```



```
#Quan acabem d'analitzar la mostra veiem que com en els casos anteriors no tenim cap NA (data missing) ni cap factor incomplet, llavors la nostra mostra és correcta i com en els casos anteriors hem posat nom al nostre barplot per tenir una millor visualització
```

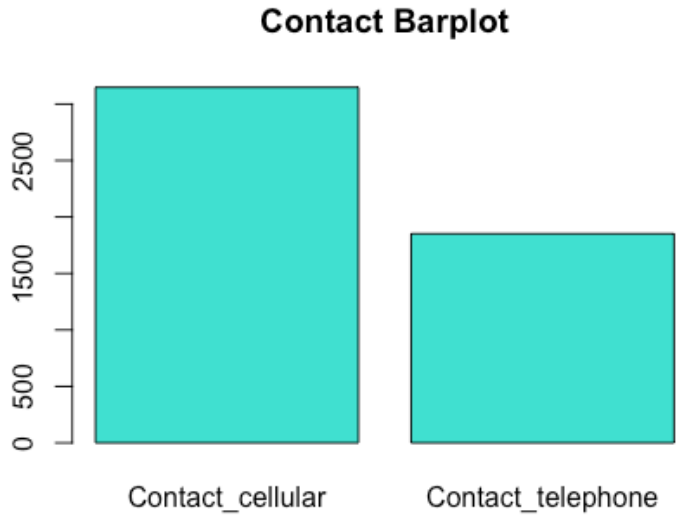
8. Contact

Contact communication type?

```
df$contact<-factor(df$contact)
levels(df$contact)<-paste("Contact_",sep="",levels(df$contact))
summary(df$contact)
```

```
##  Contact_cellular Contact_telephone
##              3148              1852
```

```
barplot(summary(df$contact),main="Contact Barplot",col = "turquoise")
```

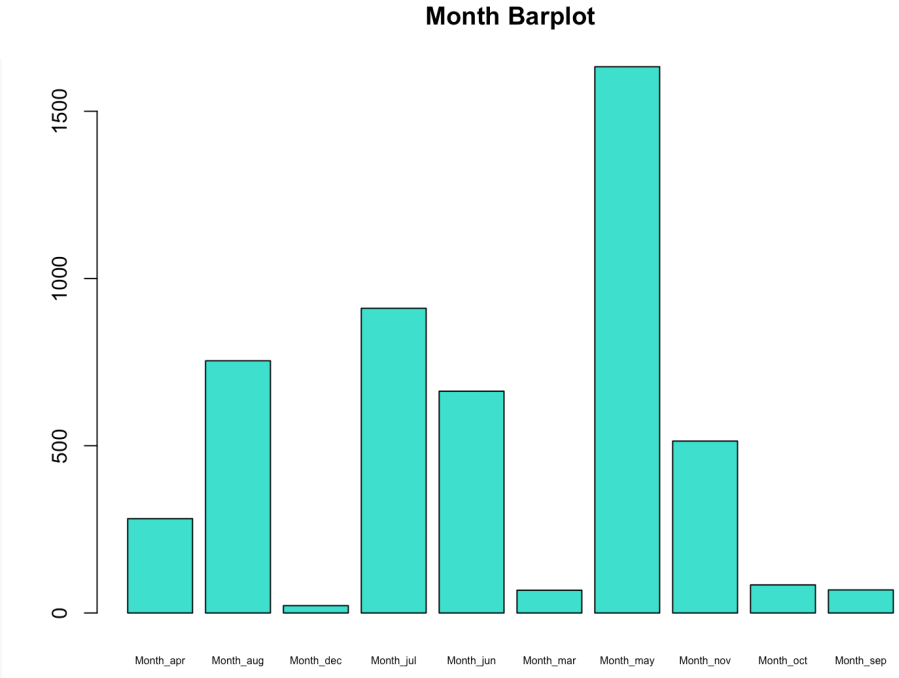



```
#Quan acabem d'analitzar la mostra veiem que com en els casos anteriors no tenim cap NA (data missing) ni cap factor incomplet, llavors la nostra mostra és correcta i com en els casos anteriors hem posat nom al nostre barplot per tenir una millor visualització
```

9. Month

Last contact month of the year?

```
df$month<-factor(df$month)
levels(df$month)<-paste("Month_",sep="",levels(df$month))
summary(df$month)

## Month_apr Month_aug Month_dec Month_jul Month_jun Month_mar Month_may
##       282       754        22       911       663        68      1633
## Month_nov Month_oct Month_sep
##       514        84        69

barplot(summary(df$month),main="Month Barplot",col = "turquoise",cex.names =
0.5)
```



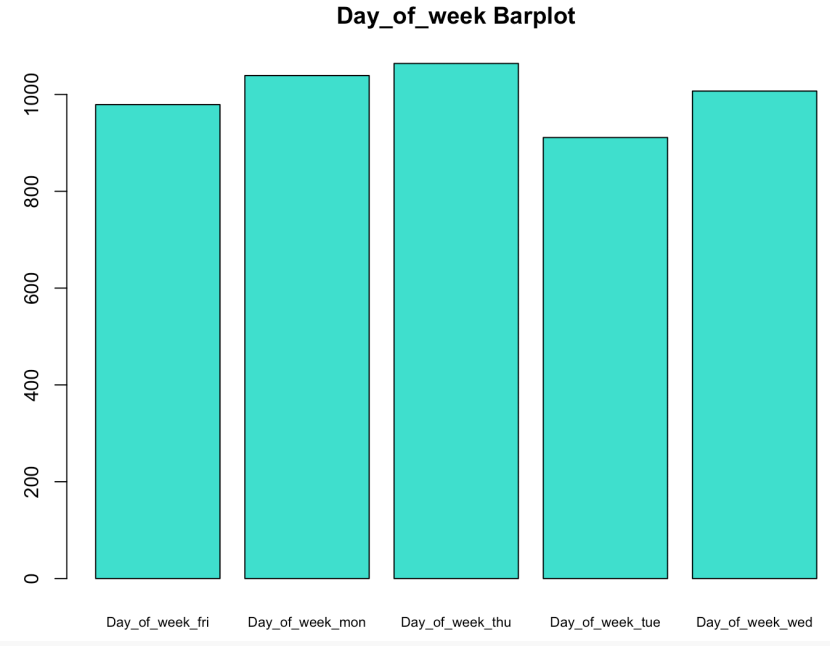
10. Day_of_week

Last contact day of the week?

```
df$day_of_week<-factor(df$day_of_week)
levels(df$day_of_week)<-paste("Day_of_week_",sep="",levels(df$day_of_week))
summary(df$day_of_week)
```

```
## Day_of_week_fri Day_of_week_mon Day_of_week_thu Day_of_week_tue
##             979             1039             1064              911
## Day_of_week_wed
##            1007
```

```
barplot(summary(df$day_of_week),main="Day_of_week Barplot",col =
"turquoise",cex.names=0.7)
```




```
#Com podem observar en els nostres factors no tenim cap valor missing i
segons les nostres observacions tampoc tenim cap outlier destacat
```

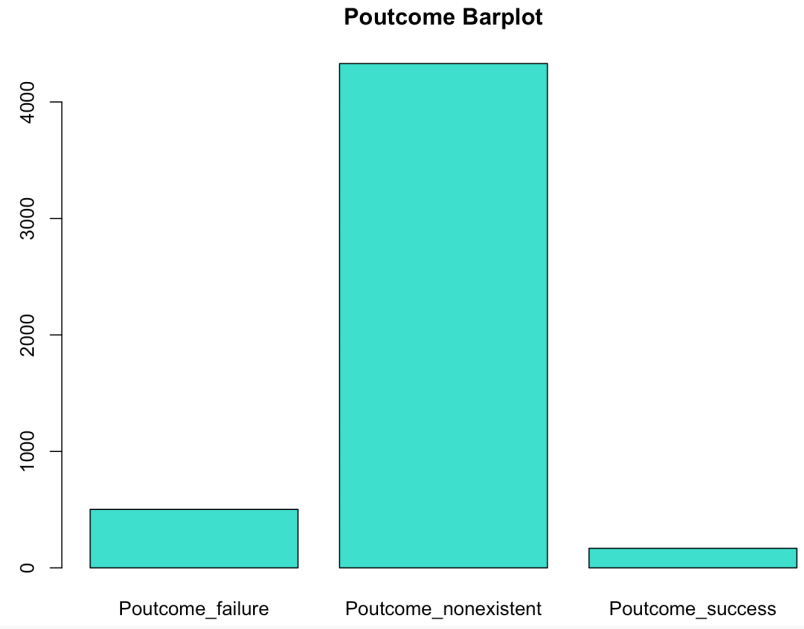
15. Poutcome

Outcome of the previous marketing campaign?

```
df$poutcome<-factor(df$poutcome)
levels(df$poutcome)<-paste("Poutcome_",sep="",levels(df$poutcome))
summary(df$poutcome)
```

```
##     Poutcome_failure Poutcome_nonexistent     Poutcome_success
##                  502                 4330                  168
```

```
barplot(summary(df$poutcome),main="Poutcome Barplot",col = "turquoise")
```




```
#Quan acabem d'analitzar la mostra veiem que com en els casos anteriors no tenim cap NA (data missing) ni cap factor incomplet, llavors la nostra mostra és correcta i com en els casos anteriors hem posat nom al nostre barplot per tenir una millor visualització
```

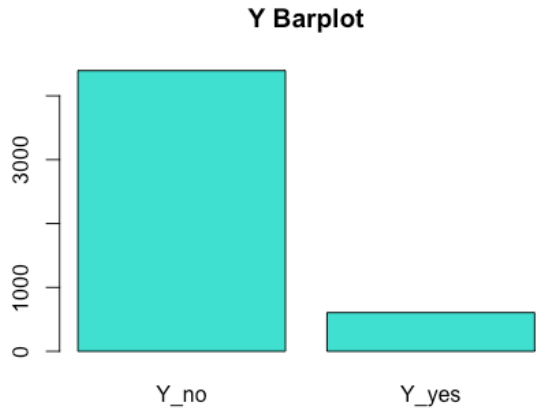
21. Y

Has the client subscribed a term deposit?

```
df$y<-factor(df$y)
levels(df$y)<-paste("Y_",sep="",levels(df$y))
summary(df$y)
```

```
##  Y_no Y_yes
##  4394   606
```

```
barplot(summary(df$y),main="Y Barplot",col = "turquoise")
```




```
#Quan acabem d'analitzar la mostra veiem que com en els casos anteriors no tenim cap NA (data missing) ni cap factor incomplet, llavors la nostra mostra és correcta i com en els casos anteriors hem posat nom al nostre barplot per tenir una millor visualització
```

Quantitative Variables (Numerical)

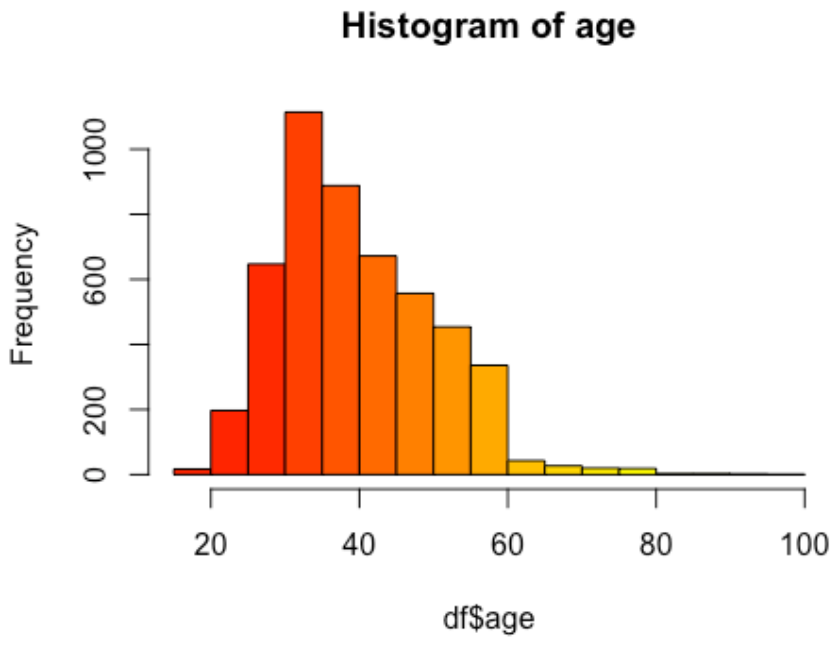
Hem de fer un analisi de totes les variables per poder identificar missings, errors i els outliers. Tambe farem una serie de boxplots i histogrames per analitzar i visualitzar millor les dades de la nostra mostra

1. Age

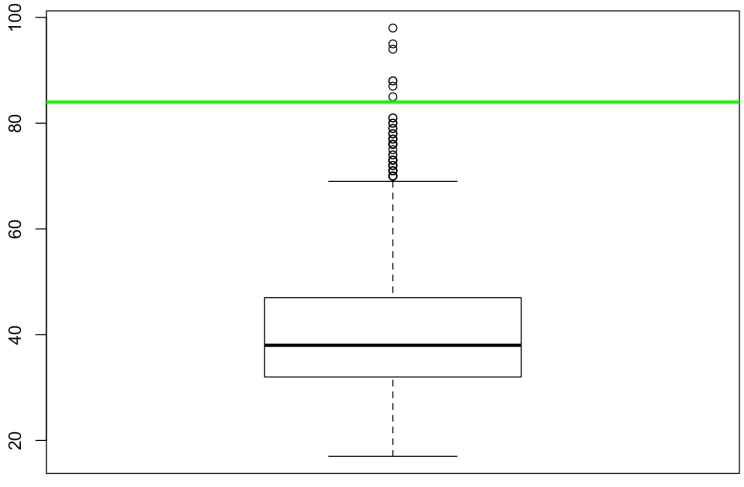
```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   17.00   32.00   38.00   40.16   47.00   98.00
```

```
hist(df$age,15,main="Histogram of age",col=heat.colors(17,alpha=1))
```




```
#A partir del summary veiem que no hi ha cap mostra que contingui un NA (missing data) ni tampoc cap possible error ja que l'edat mínima (17) i la màxima (98) son valors que s'adhereixen a la realitat.
boxplot(df$age)
abline(h=84,col="green",lwd=3)
```



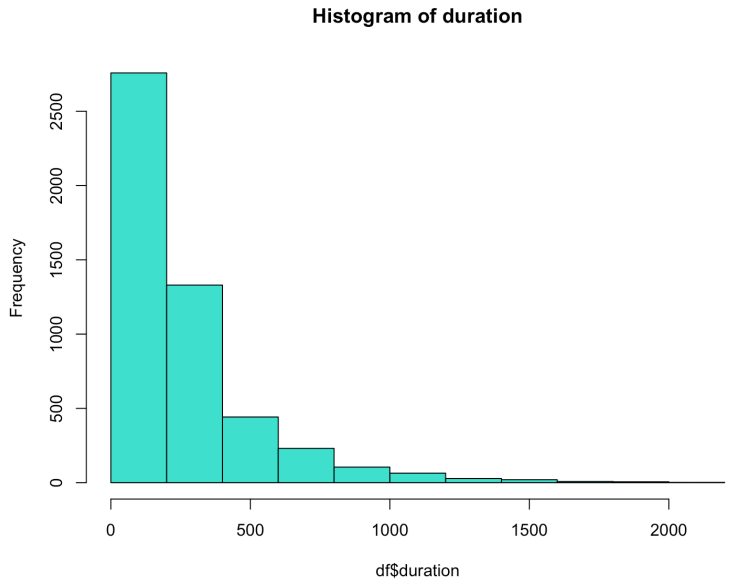
```
#Amb la comanda abline el que volem fer es poder identificar de una manera més fàcil els possibles outliers i poder tenir una millor visualització, per això marco a l'altura dels 84 anys la nostra mostra, ja que aquests valors sén els que s'allunyen una mica de la resta, llavors s'ahuran de fer una sèrie d'imputacions

sel <- which(df$age >= 84);length(sel);sel

## [1] 7

## [1] 3434 3436 3439 4564 4646 4714 4781

summary(df$age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   32.00   38.00   40.16   47.00   98.00

num_total_outliers[1] <- length(sel)
df[sel, "age"] <- NA
#Cuando eliminamos nuestros outliers lo que nos queda es que la edad máxima ahora es de 81 años y tenemos 7 NA's
df[sel, "outliers_indiv"] <- df[sel, "outliers_indiv"] + 1
summary(df$age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   17.00   32.00   38.00   40.09   47.00   81.00       7

#Un cop els hem identificat, actualitzem les variables de control per tal de portar un seguiment correcte de la mostra i eliminem els 7 outliers considerats.
```

11. Duration

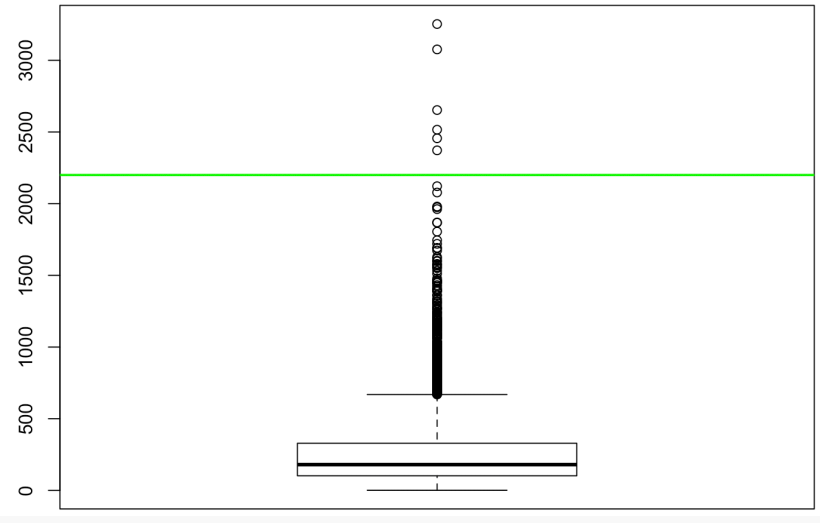
Last contact duration?

```
summary(df$duration)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   102.0   180.0   264.7   329.0  3253.0

hist(df$duration,15,main="Histogram of duration",col="turquoise")
```

```
#A partir del summary executat podem observar que el temps mínim de la durada de una trucada és d'1 segon, i ja ens podem adonar que aquest valor no té molt sentit a l'hora de tractar-se una trucada no? No dóna temps de que el client escolti i penji i la durada màxima és de 3253 segons que son aproximadament uns 54 minuts i pot ser un valor real
```




```r
boxplot(df$duration)
abline(h=2200,col="green",lwd=2)
```



#Per tal d'identificar possibles outliers utilitzem l'eina Boxplot, tenant en compte el significat de la variable marquem amb una línia vermella el valor 2200, a partir del qual definim els possibles outliers ja que considerem que les observacions que prenen un valor a partir de 2200 es desvien significativament de la resta

```r
sel <- which(df$duration >= 2200);length(sel);sel
```

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

```
## [1] 1013 1140 2197 2919 2969 3440
```

```r
num_total_outliers[11] <- length(sel)
df[sel, "outliers_indiv"] <- df[sel, "outliers_indiv"] + 1
df <- df[-sel,]
summary(df$duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   102.0   180.0   261.8   328.0  2122.0
```

```r
#Un cop els hem identificat, actualitzem les variables de control per tal de portar un seguiment
#correcte de la mostra i eliminem els 18 outliers del nostre traget numèric.
```

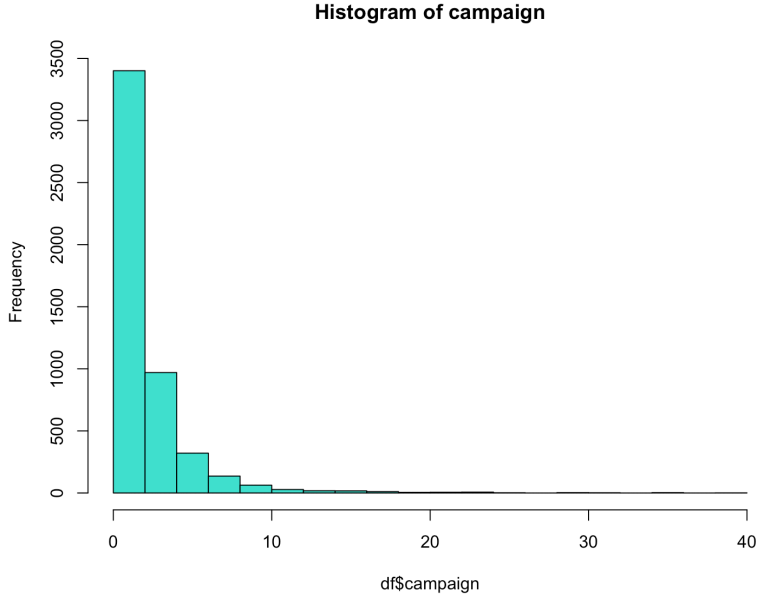
12. Campaign

Number of contacts performed during this campaign?

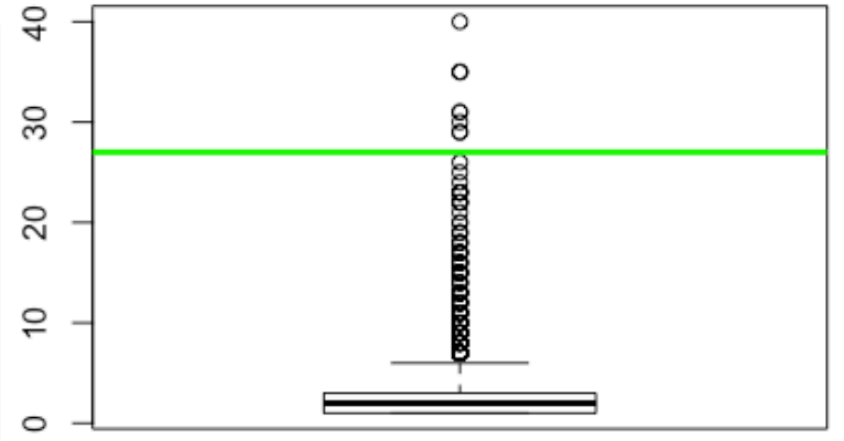
```r
summary(df$campaign)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.599   3.000  40.000
```

```r
hist(df$campaign,15,main="Histogram of campaign",col="turquoise")
```



```
#Quan fem el summary i el boxplot veiem que no hi ha cap mostra que contingui un NA (missing data) però amb el boxplot si que veiem que hi han alguns valors que poden no ser molt realistes, ja que és una mica sopitos que una campanya es contacti unes 40 vegades amb una mateixa persona, comptant que la mitjana són dues vegades, llavors eliminarem a partir d'unes 27 vegades/ persona que és el que te mes sentit comu i és on veiem que disten de la resta
#Aquestes dades de la mostra les considerem errors i les eliminarem de la mostra

boxplot(df$campaign)
abline(h=27,col="green",lwd=3)
```




```
sel <- which(df$campaign > 27)
length(sel);sel
```

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

```
## [1]  509 1116 1216 1278 1279 2311 2312 2318 2325
```

```
num_total_errors[12] <- length(sel)
df[sel, "campaign"] <- NA
df[sel, "errors_indiv"] <- df[sel, "errors_indiv"] + 1

#Després de fer l'anàlisi de la mostra podem arribar a la conclusió que no és molt normal rebre contacte de la mateixa campanya més de 15 cops, llavors haurem d'eliminar els possibles outliers de la mostra per tenir correcte el nostre traget numèric i veiem que eliminem 57 observacions
sel <- which(df$campaign >= 15)
length(sel);sel
```

```
## [1] 48

##  [1]  326  418  452  467  484  665  710  778  874  875  908  922  979 1005
## [15] 1039 1181 1219 1241 1276 1283 1284 1353 1401 1433 1458 1565 1651 1787
## [29] 2049 2095 2128 2155 2179 2182 2214 2242 2246 2270 2276 2279 2314 2321
## [43] 2795 2886 2908 2917 3685 4183
```

```
num_total_outliers[12] <- length(sel)
df[sel, "campaign"] <- NA
df[sel, "outliers_indiv"] <- df[sel, "outliers_indiv"] + 1
df<-df[-sel,]
summary(df$campaign)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   2.000   2.388   3.000  14.000       9
```

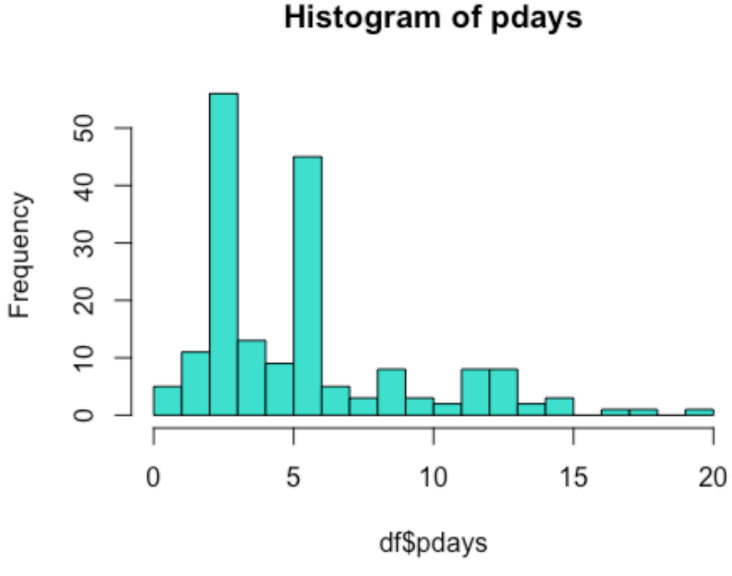
13. Pdays

Number of days that passed by after the client was last contacted from a previous campaign?

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   3.000   5.000   5.821   6.000  20.000    4762
```

```
hist(df$pdays,15,main="Histogram of pdays",col="turquoise")
```



```r
#Si analitzem aquesta variable veiem que tenir valor 0 significa que no ha passat cap dia des de que s'ha finalitzat la campanya anterior i s'ha contactat amb l'individu per aquesta campanya la qual cosa considerem que es tracta de un error per aixo procedim a identificar i comptabilitzar aquest error a continuació.

sel <- which(df$pdays == 0)
length(sel);sel
```

```
## [1] 2
```

```
## [1] 4844 4847
```

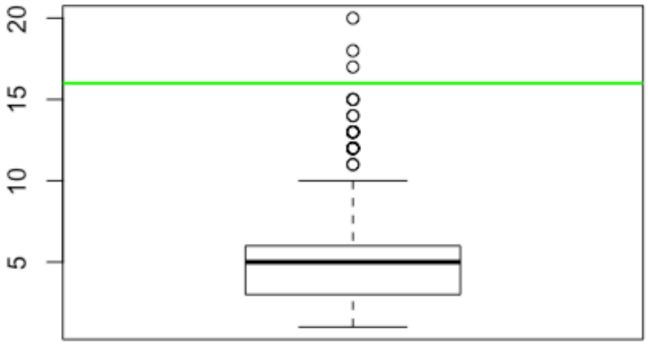
```r
#A partir del summary veiem que hi han 2 observacions que tenen valor 0.
num_total_errors[13] = length(sel)
df[sel, "pdays"] <- NA
df[sel, "errors_indiv"] <- df[sel, "errors_indiv"] + 1

#També podem observem que aquesta variable té un nombre molt elevat de NA's(missing data) aquestes situacions signifiquen que no s'ha contactat amb l'individu prèviament en cap altre campanya per això no pot existir cap valor amb els dies des de la última vegada que es va contactar.

sel <- which(is.na(df$pdays))
length(sel);#sel
```

```
## [1] 4764
```

```r
num_total_missings[13] = length(sel)
df[sel, "missings_indiv"] <- df[sel, "missings_indiv"] + 1
summary(df$pdays)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   3.000   5.000   5.885   6.000  20.000    4764
```

```r
boxplot(df$pdays)
abline(h=16,col="green",lwd=2)
```



```r
sel <- which(df$pdays >= 16)
length(sel);sel
```

```
## [1] 3
```

```
## [1] 4846 4870 4912
```

```r
num_total_outliers[13] = length(sel)
df[sel, "pdays"] <- NA
df[sel, "outliers_indiv"] <- df[sel, "outliers_indiv"] + 1
summary(df$pdays)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   3.000   5.000   5.676   6.000  15.000    4767
```

```r
#Un cop els hem identificat, actualitzem les variables de control per tal de portar un seguiment
#correcte de la mostra i eliminem els outliers del nostre target numèric.
```

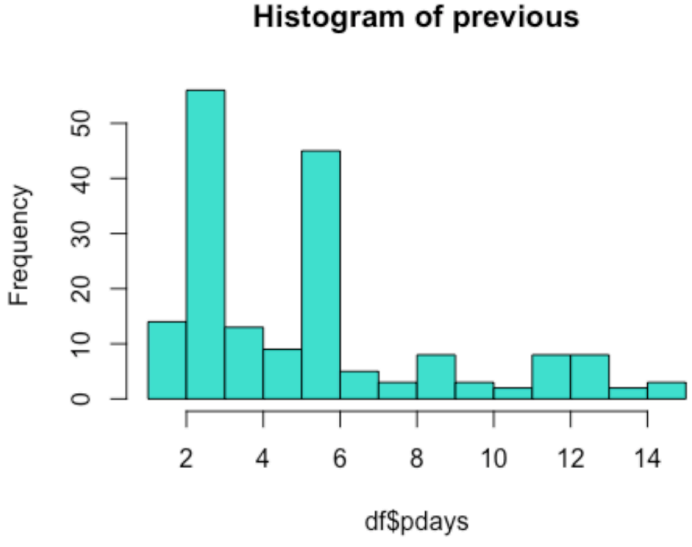
14. Previous

Number of contacts performed before this campaign and for this client?

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1708  0.0000  5.0000
```

```r
hist(df$pdays,15,main="Histogram of previous",col="turquoise")
```



```
#A partir del summary efectuat sobre la variable "Previous" podem veure que no tenim cap NA i podriem considerar que tampoc error perquè ja que el nombre mínim de contactes previs a la campanya actual amb l'individu és 0 i el màxim trobat és 5, que poden ser valors reals

#Quan observem el boxplot i el summary veiem que la majoria de les nostres observacions son 0 i llavors no podem tenir o identificar rapidament els possibles outliers
```

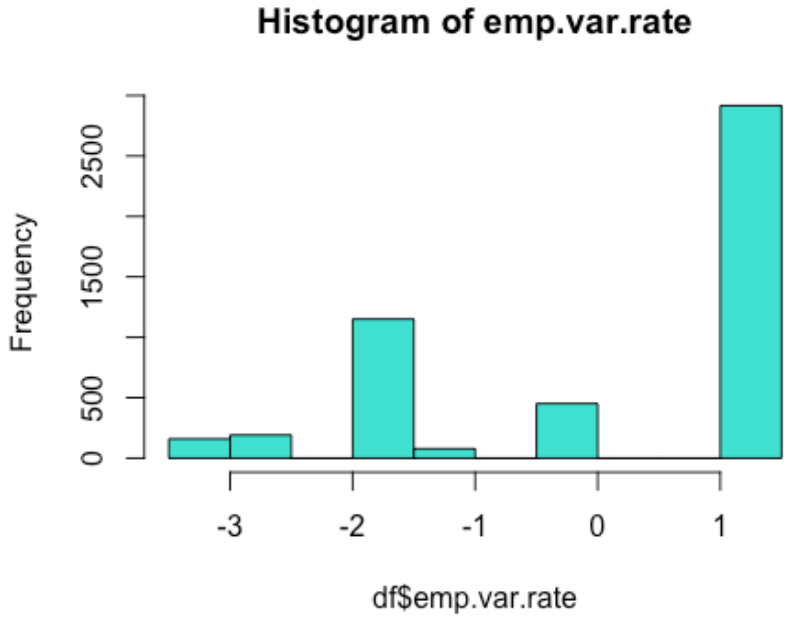
16. Emp.var.rate

Employment variation rate?

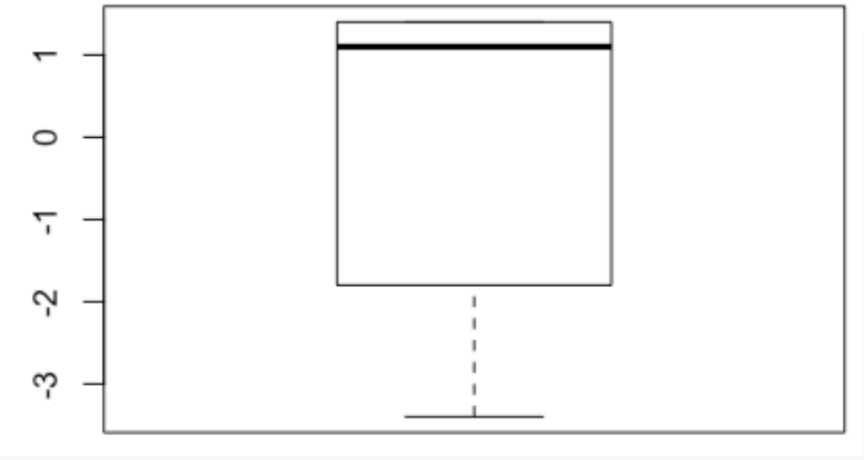
```
summary(df$emp.var.rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -3.4000 -1.8000  1.1000  0.1074  1.4000  1.4000
```

```
hist(df$emp.var.rate,15,main="Histogram of emp.var.rate",col="turquoise")
```



```r
boxplot(df$emp.var.rate)
```




```r
#A partir del summary, l'histograma i el boxplot podem afirmar que no tenim cap missing ni error ni outlier, perquè tots els valors agafats són realistes
```

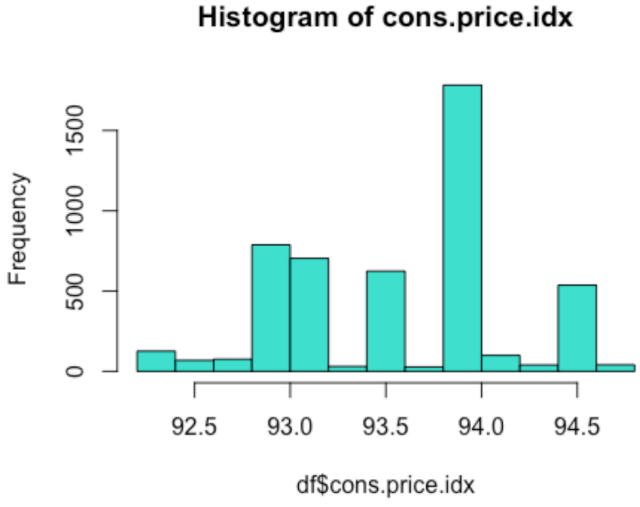
17. Cons.price.idx

Consumer price index - monthly indicator?

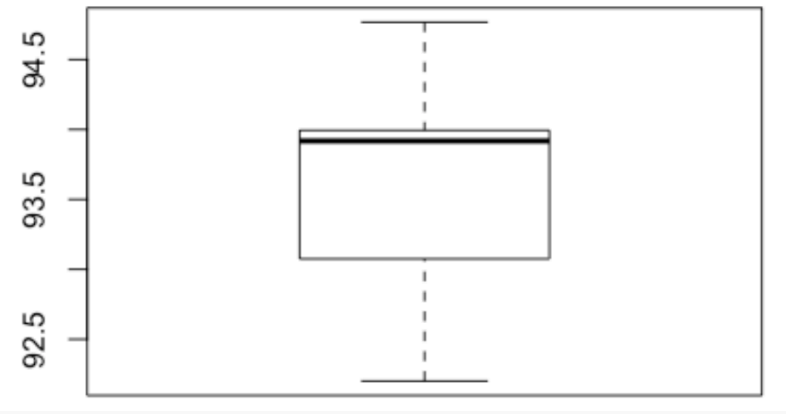
```r
summary(df$cons.price.idx)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   92.20   93.08   93.92   93.59   93.99   94.77
```

```r
hist(df$cons.price.idx,15,main="Histogram of cons.price.idx",col="turquoise")
```



```r
boxplot(df$cons.price.idx)
```




```r
#A partir del summary, l'histograma i el boxplot podem afirmar que no tenim cap missing ni error ni outlier, perquè tots els valors agafats són realistes
```

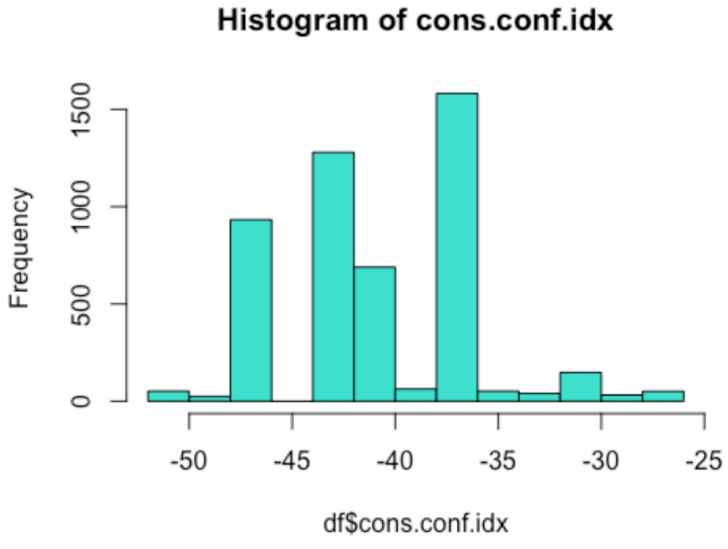
18. Cons.conf.idx

Consumer confidence index - monthly indicator?

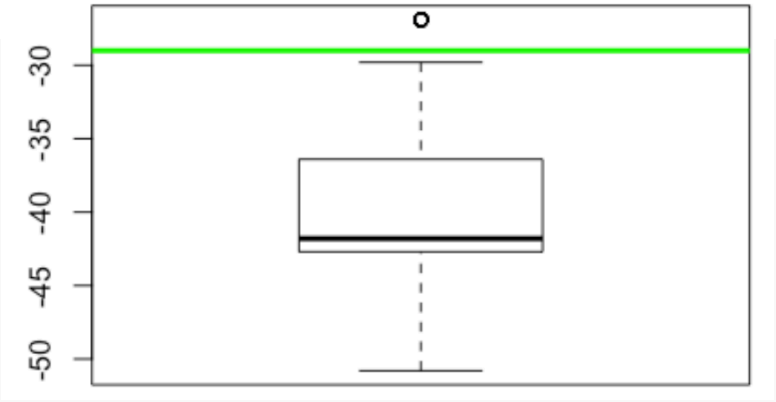
```r
summary(df$cons.conf.idx)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -50.80  -42.70  -41.80  -40.44  -36.40  -26.90
```

```r
hist(df$cons.conf.idx,15,main="Histogram of cons.conf.idx",col="turquoise")
```



```r
boxplot(df$cons.conf.idx)
#Com podem veure després del boxplot hi han algunes observacions que podrien considerarse possibles outliers, llavors marquem -29 amb el abline

abline(h=-29,col="green",lwd=3)
```




```r
sel <- which(df$cons.conf.idx >= -29)
length(sel);
```

```
## [1] 51
```

```r
num_total_outliers[18] = length(sel)
df[sel, "cons.conf.idx"] <- NA
df[sel, "outliers_indiv"] <- df[sel, "outliers_indiv"] + 1
summary(df$cons.conf.idx)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -50.80  -42.70  -41.80  -40.58  -36.40  -29.80      51
```

```r
#Ara el que hem fet és veure que hi han uns 51 possibles outliers, llavors el que hem de fer és imputar-los i posar-los com a NA (missing values) i llavors els posem en el vector creat per tenir tots els outliers a ma i després incrementem el contador d'outliers
```

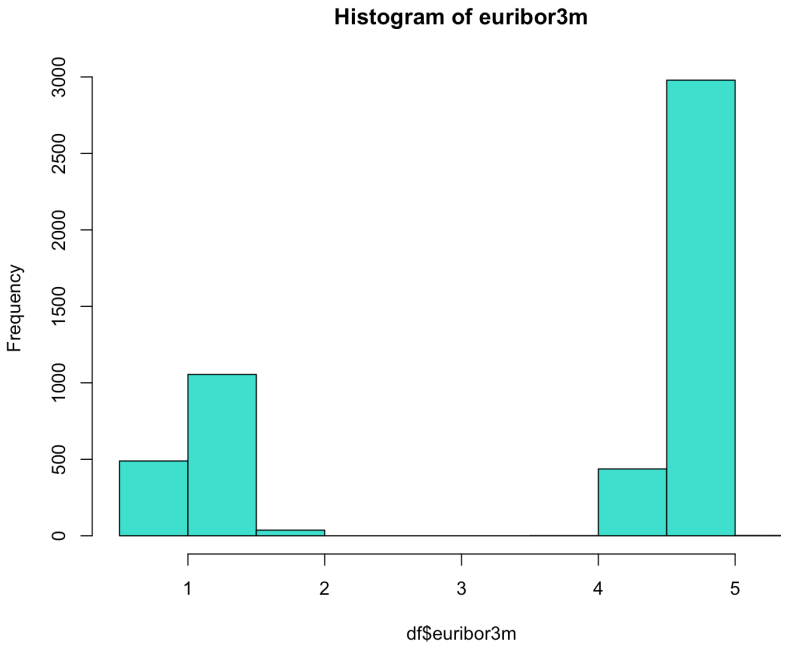
19. Euribor3m

Euribor 3 month rate - daily indicator?

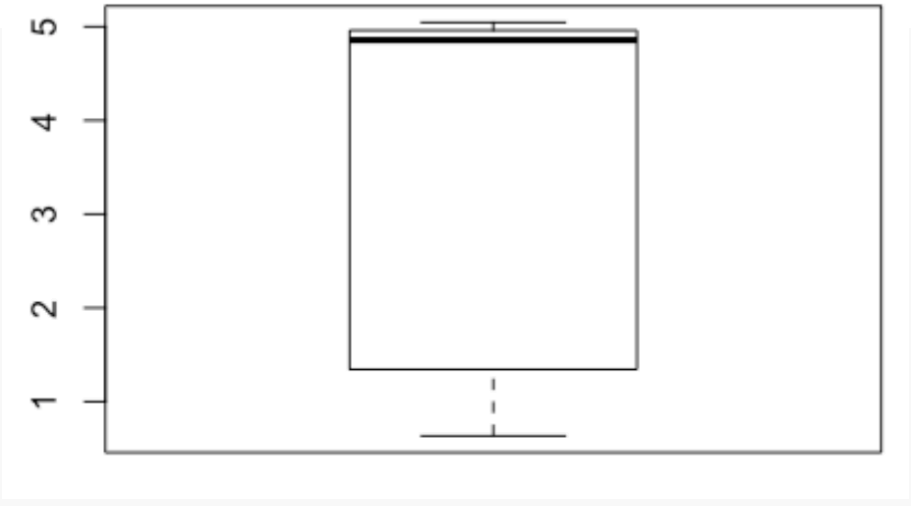
```r
summary(df$euribor3m)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.634   1.344   4.857   3.649   4.961   5.045
```

```r
hist(df$euribor3m,15,main="Histogram of euribor3m",col="turquoise")
```




```r
boxplot(df$euribor3m)
```




```r
#A partir del boxplot efectuat podem veure que els valors obtinguts són majoritàriament menors que 5 i com s'observa la mitjana es troba molt a prop del màxim obtingut
```

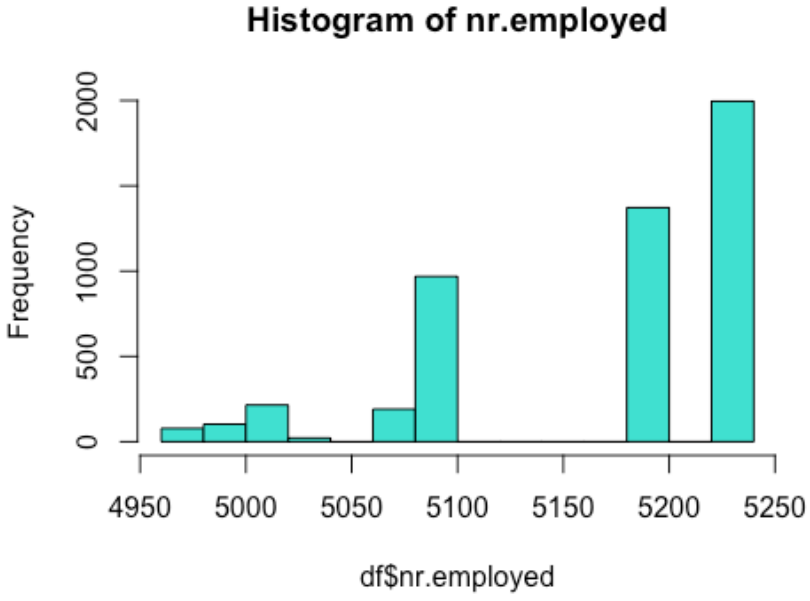
20. Nr.employed

Number of employees - quarterly indicator?

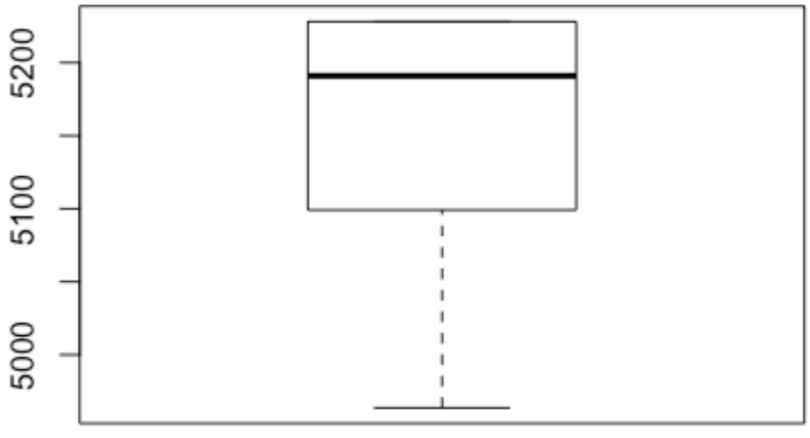
```
summary(df$nr.employed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4964    5099    5191    5168    5228    5228
```

```
hist(df$nr.employed,15,main="Histogram of nr.employed",col="turquoise")
```




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



```
#A partir del summary, l'histograma i el boxplot podem afirmar que no tenim cap missing ni error ni outlier.
```

CONTAR NA's

```
#Hem de contar el numero de NA's despres d'analitzar les dades i marcta els outliers, missings i errors
miss_row <- rowSums(is.na(df))
miss_col <- colSums(is.na(df))
miss_col
```

```
##            age            job        marital      education        default 
##              7              0              7            210              0 
##        housing           loan        contact          month    day_of_week 
##              0              0              0              0              0 
##       duration       campaign          pdays       previous       poutcome 
##              0              9           4767              0              0 
##   emp.var.rate cons.price.idx  cons.conf.idx      euribor3m    nr.employed 
##              0              0             51              0              0 
##              y missings_indiv   errors_indiv outliers_indiv 
##              0              0              0              0
```

```
#Podem veure el numero de NA que tenim per cada variable
summary(miss_row)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   1.000   1.000   1.021   1.000   3.000
```

Rank of variables

Com hem fet abans ja tenim creades les variables on tenim emmagatzemats els errors, missing values i els outliers i ara el que farem es un ranking amb aquestes variables

Per individuals:

```
#errors (la majoria de registres no tenen errors i els que tenen errors com a màxim només en tenen 1)
summary(df$errors_indiv)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## 0.000000 0.000000 0.000000 0.002224 0.000000 1.000000
```

```
#outliers (el registres amb outliers com a màxim tenen 2 variables amb outlier)
summary(df$outliers_indiv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.01233 0.00000 2.00000
```

```
#missings abans d'introduir manualment NA's per cada registre, només la variable pdays tenia missings des d'un principi
summary(df$missings_indiv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  1.0000  1.0000  0.9632  1.0000  1.0000
```

```
#Després de depurar les dades i introduïr els NA's
#miss_col<-colSums(is.na(df))
NAs_indiv <- rowSums(is.na(df))
summary(df$NAs_indiv)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

Per variable:

```
#Després de calcular tots el missings, outliers i errors fem el resum d'ells


#num total missings
data <- t(c(num_total_missings[13]))
data
```

```
##      [,1]
## [1,] 4764
```

```
barplot(data, main="Total missings", col=("turquoise"))
```

```r
#num total errors
data <- t(c(num_total_errors[12:13]))
data
```

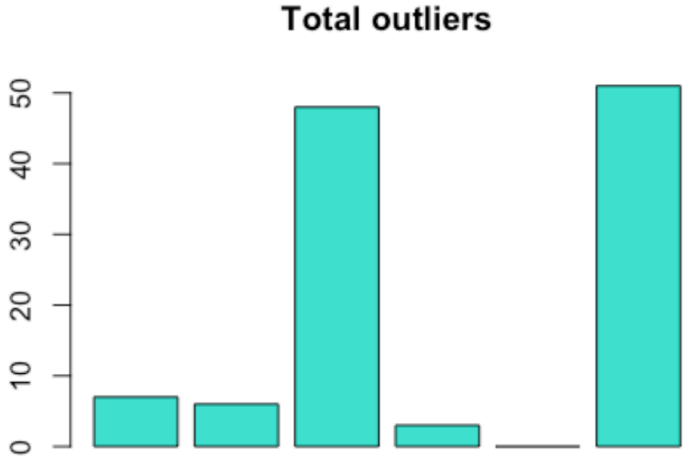
```
##      [,1] [,2]
## [1,]    9    2
```

```r
barplot(data, main="Total errors", col=("turquoise"))
```


```r
#num total outliers
data <-
t(c(num_total_outliers[1],num_total_outliers[11:14],num_total_outliers[18]))
data
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    7    6   48    3    0   51
```

```r
barplot(data, main="Total outliers", col=("turquoise"))
```



Imputation

Ara farem l'estudi per variables i tractarem d'imoutar les observacions que siguin necesasaries

```
library(missMDA)
# Numeric imputation
vars_con<-names(df)[c(1,11:14,16:20)]
vars_dis<-names(df)[c(2:10,15,21)] #solo 21
summary(df[,vars_con])
```

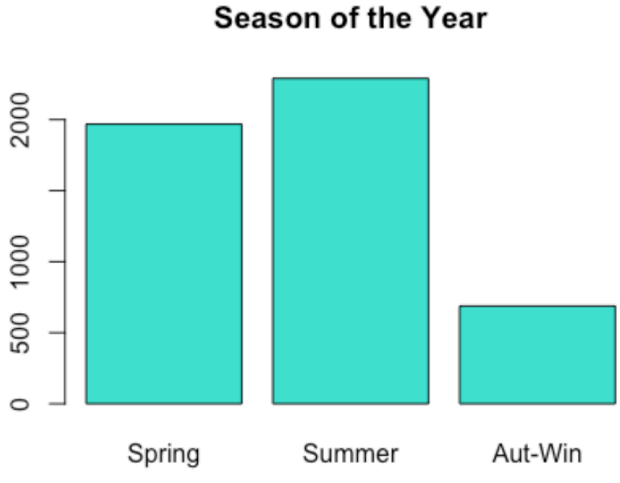
```
##       age            duration         campaign          pdays
##  Min.   :17.00   Min.   :   1.0   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:32.00   1st Qu.: 104.0   1st Qu.: 1.000   1st Qu.: 3.000
##  Median :38.00   Median : 182.0   Median : 2.000   Median : 5.000
##  Mean   :40.05   Mean   : 262.8   Mean   : 2.388   Mean   : 5.676
##  3rd Qu.:47.00   3rd Qu.: 329.0   3rd Qu.: 3.000   3rd Qu.: 6.000
##  Max.   :81.00   Max.   :2122.0   Max.   :14.000   Max.   :15.000
##  NA's   :7                        NA's   :9        NA's   :4767
##     previous       emp.var.rate      cons.price.idx  cons.conf.idx
##  Min.   :0.0000   Min.   :-3.4000   Min.   :92.20   Min.   :-50.80
##  1st Qu.:0.0000   1st Qu.:-1.8000   1st Qu.:93.08   1st Qu.:-42.70
##  Median :0.0000   Median : 1.1000   Median :93.92   Median :-41.80
##  Mean   :0.1708   Mean   : 0.1074   Mean   :93.59   Mean   :-40.58
##  3rd Qu.:0.0000   3rd Qu.: 1.4000   3rd Qu.:93.99   3rd Qu.:-36.40
##  Max.   :5.0000   Max.   : 1.4000   Max.   :94.77   Max.   :-29.80
##                                                     NA's   :51
##    euribor3m      nr.employed
##  Min.   :0.634   Min.   :4964
##  1st Qu.:1.344   1st Qu.:5099
##  Median :4.857   Median :5191
##  Mean   :3.649   Mean   :5168
##  3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :5.045   Max.   :5228
```

```
res.impn<-imputePCA(df[,vars_con],ncp=5) #vars_con=numericas
#res.impn<-imputePCA(df[,vars_dis],ncp=5)
attributes(res.impn)
```

```
## $names
## [1] "completeObs" "fittedX"
```

```
#data.frame with all NA imputed: res.impn$completeObs
#summary(res.impn$completeObs)


df[,"age"] <- res.impn$completeObs[,"age"]
df[,"campaign"] <- res.impn$completeObs[,"campaign"]
df[,"pdays"] <- res.impn$completeObs[,"pdays"]
df[,"cons.conf.idx"] <- res.impn$completeObs[,"cons.conf.idx"]
```

```r
df[,"euribor3m"] <- res.impn$completeObs[,"euribor3m"]
miss_row <- rowSums(is.na(df))
miss_col <- colSums(is.na(df))

table (df$month)
## Month_apr Month_aug Month_dec Month_jul Month_jun Month_mar Month_may
##       281       749        22       893       648        67      1620
## Month_nov Month_oct Month_sep
##       514        83        69

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

# 1 level - spring
sel<-which(df$month %in% c("Month_mar","Month_apr","Month_may"))
df$season[sel] <-1

# 2 level - Summer
sel<-which(df$month %in% c("Month_jun","Month_jul","Month_aug"))
df$season[sel] <-2

table(df$season)
##    1    2    3
## 1968 2290  688

summary(df$season)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.741   2.000   3.000

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

barplot(summary(df$season), main="Season of the Year", col=("turquoise"))
```



Discretitzation

Ara el que farrea será la discretització de les variables numeriques i això ho farem convertint en factors els diferents rangs que tenim de les observacions corresponents a una variable numèrica per tenir una visualització més clara

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

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

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

```
##       age           duration         campaign          pdays
##  Min.   :17.00   Min.   :   1.0   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:32.00   1st Qu.: 104.0   1st Qu.: 1.000   1st Qu.: 5.348
##  Median :38.00   Median : 182.0   Median : 2.000   Median : 5.657
##  Mean   :40.05   Mean   : 262.8   Mean   : 2.389   Mean   : 5.706
##  3rd Qu.:47.00   3rd Qu.: 329.0   3rd Qu.: 3.000   3rd Qu.: 5.990
##  Max.   :81.00   Max.   :2122.0   Max.   :14.000   Max.   :15.000
##     previous       emp.var.rate     cons.price.idx  cons.conf.idx
##  Min.   :0.0000   Min.   :-3.4000   Min.   :92.20   Min.   :-50.80
##  1st Qu.:0.0000   1st Qu.:-1.8000   1st Qu.:93.08   1st Qu.:-42.70
##  Median :0.0000   Median : 1.1000   Median :93.92   Median :-41.80
##  Mean   :0.1708   Mean   : 0.1074   Mean   :93.59   Mean   :-40.62
##  3rd Qu.:0.0000   3rd Qu.: 1.4000   3rd Qu.:93.99   3rd Qu.:-36.40
##  Max.   :5.0000   Max.   : 1.4000   Max.   :94.77   Max.   :-29.80
##    euribor3m      nr.employed
##  Min.   :0.634   Min.   :4964
##  1st Qu.:1.344   1st Qu.:5099
##  Median :4.857   Median :5191
##  Mean   :3.649   Mean   :5168
##  3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :5.045   Max.   :5228
```

Factor Age

```
# Trend and dispersion statistics
quantile(df$age,na.rm=TRUE)
```

```
quantile(df$age,seq(0,1,0.2),na.rm=TRUE)
```

```
##   0%  20%  40%  60%  80% 100%
##   17   31   36   41   49   81
```

```
#Es crea una variable auxiliar per tenir els diferents rangs d'edat i fem els intervals per a que sigui més sencilla i fàcil la visualització de les diferents mostres
```

```r
df$varauxiliar<-factor(cut(df$age,include.lowest=T,breaks=c(17,31,36,41,49,81)))
summary(df$varauxiliar)
```

```
## [17,31] (31,36] (36,41] (41,49] (49,81]
##    1113    1062     830     953     988
```

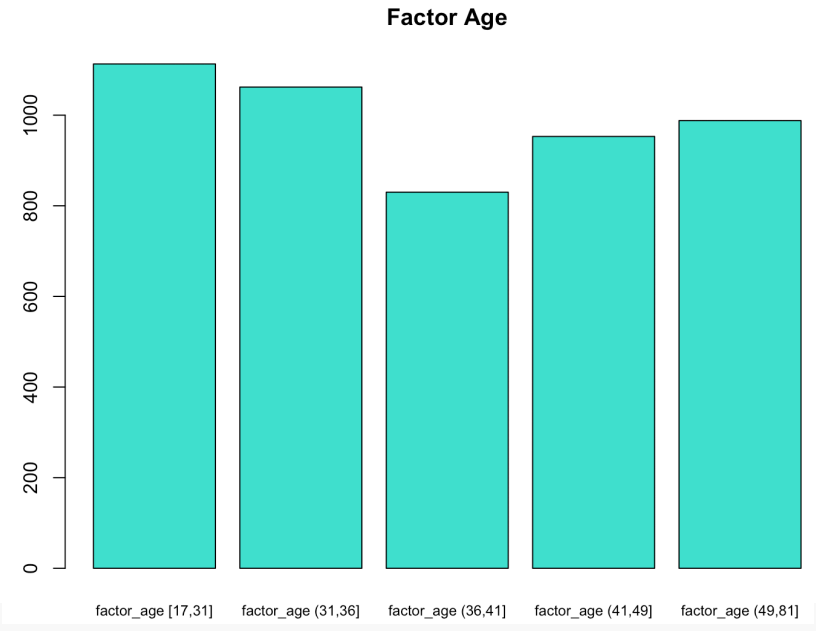
```r
#Fem la mitjana amb els valors de les edats i els nostres intervals
tapply(df$age,df$varauxiliar,median)
```

```
## [17,31] (31,36] (36,41] (41,49] (49,81]
##      29      34      39      45      55
```

```r
#Ara li posem el nom de "factor_age" a la nostra variable per poder tenir una
millor interpretación i tornem a fer el mateix procés
df$factor_age<-factor(cut(df$age,include.lowest=T,breaks=c(17,31,36,41,49,81)))
levels(df$factor_age)<-paste("factor_age ",levels(df$factor_age),sep="")
table(df$factor_age)
```

```
## factor_age [17,31] factor_age (31,36] factor_age (36,41]
##               1113               1062                830
## factor_age (41,49] factor_age (49,81]
##                953                988
```

```r
barplot(summary(df$factor_age), main="Factor Age",col=("turquoise"),cex.names=0.75)
```



Factor Duration

```
# Trend and dispersion statistics
quantile(df$duration,seq(0,1,0.125),na.rm=TRUE)
```

```
##    0% 12.5%   25% 37.5%   50% 62.5%   75% 87.5%  100%
##     1    68   104   139   182   236   329   504  2122
```

```
df$factor_duration<-
factor(cut(df$duration,include.lowest=T,breaks=c(1,68,104,139,182,236,329,504,2122)))
summary(df$factor_duration)
```

```
##         [1,68]       (68,104]      (104,139]      (139,182]      (182,236]
##            629            623            612            620            608
##      (236,329]      (329,504] (504,2.12e+03]
##            619            618            617
```

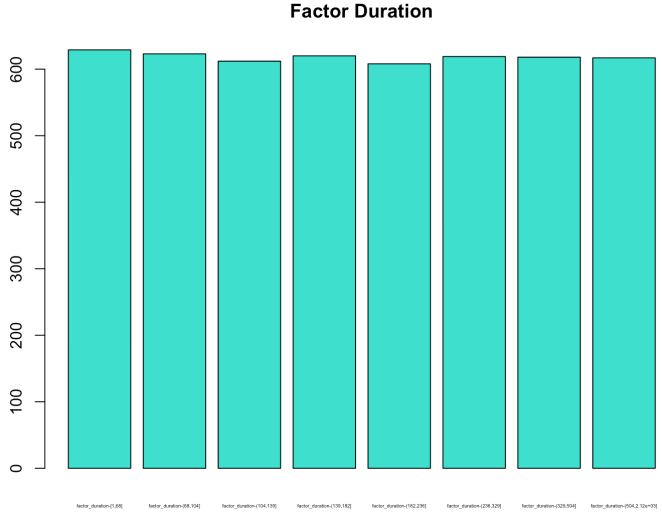
```
tapply(df$duration,df$factor_duration,median)
```

```
##         [1,68]       (68,104]      (104,139]      (139,182]      (182,236]
##             44             86            122            160            206
##      (236,329]      (329,504] (504,2.12e+03]
##            277            396            716
```

```
levels(df$factor_duration)<-paste("factor_duration-",levels(df$factor_duration),sep="")
table(df$factor_duration)
```

```
##         factor_duration-[1,68]       factor_duration-(68,104]
##                            629                            623
##      factor_duration-(104,139]      factor_duration-(139,182]
##                            612                            620
##      factor_duration-(182,236]      factor_duration-(236,329]
##                            608                            619
##      factor_duration-(329,504] factor_duration-(504,2.12e+03]
##                            618                            617
```

```
barplot(summary(df$factor_duration), main="Factor Duration",col=("turquoise"),cex.names=0.3)
```



Factor Campaign

```
# Trend and dispersion statistics
quantile(df$campaign,seq(0,1,0.2),na.rm=TRUE)
```

```
##   0%  20%  40%  60%  80% 100%
##    1    1    1    2    3   14
```

```
df$factor_campaign<-factor(cut(df$campaign,include.lowest=T,breaks=c(1,2,3,14)))
summary(df$factor_campaign)
```

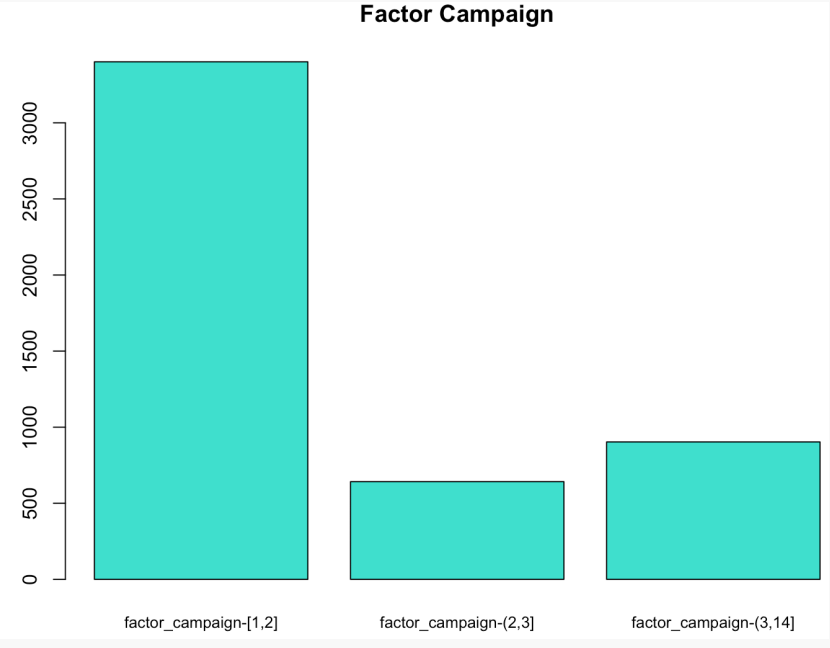
```
##  [1,2]  (2,3] (3,14]
##   3401    642    903
```

```
tapply(df$campaign,df$factor_campaign,median)
```

```
##  [1,2]  (2,3] (3,14]
##      1      3      5
```

```
levels(df$factor_campaign)<-paste("factor_campaign-",levels(df$factor_campaign),sep="")
table(df$factor_campaign)
##  factor_campaign-[1,2]  factor_campaign-(2,3] factor_campaign-(3,14]
##                   3401                    642                    903
```

```
barplot(summary(df$factor_campaign), main="Factor
Campaign",col=("turquoise"),cex.names=0.8)
```



Factor PDays

```
quantile(df$pdays,seq(0,1,0.25),na.rm=TRUE)
```

```
##        0%       25%       50%       75%      100%
##  1.000000  5.347945  5.657500  5.990268 15.000000
```

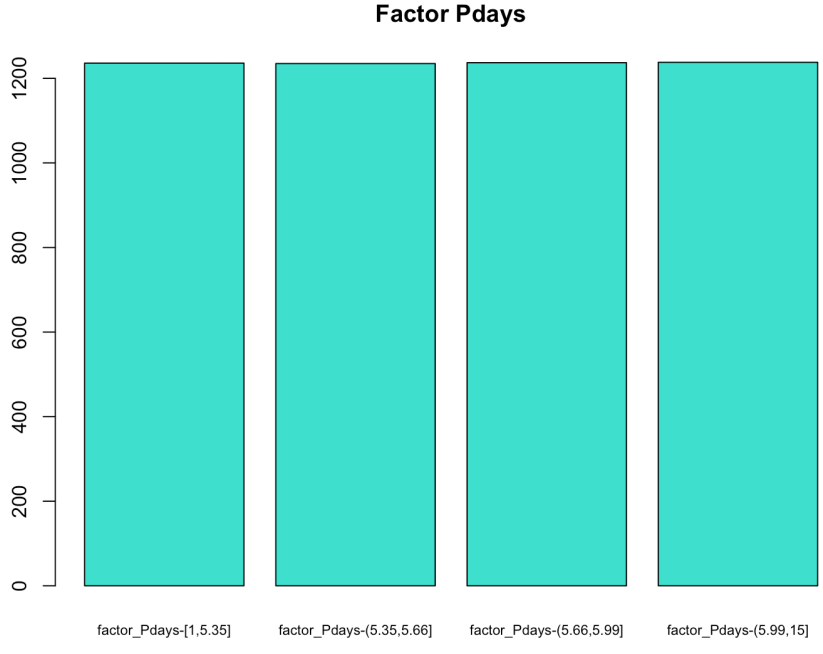
```
df$factor_Pdays<-
factor(cut(df$pdays,include.lowest=T,breaks=c(1,5.347,5.657,5.99,15)))

summary(df$factor_Pdays)
```

```
##    [1,5.35] (5.35,5.66] (5.66,5.99]   (5.99,15]
##        1236        1235        1237        1238
```

```
tapply(df$pdays,df$factor_Pdays,median)
```

```
##    [1,5.35] (5.35,5.66] (5.66,5.99]   (5.99,15]
##    5.093576    5.513674    5.805277    6.297859
```

```
levels(df$factor_Pdays)<-paste("factor_Pdays-",levels(df$factor_Pdays),sep="")
table(df$factor_Pdays)
##    factor_Pdays-[1,5.35] factor_Pdays-(5.35,5.66] factor_Pdays-(5.66,5.99]
##                    1236                    1235                    1237
##   factor_Pdays-(5.99,15]
##                    1238

barplot(summary(df$factor_Pdays), main="Factor
Pdays",col=("turquoise"),cex.names=0.7)
```



Factor Previous

```
quantile(df$previous,seq(0,1,0.1),na.rm=TRUE)
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    0    0    0    0    0    0    0    0    0    1    5
```

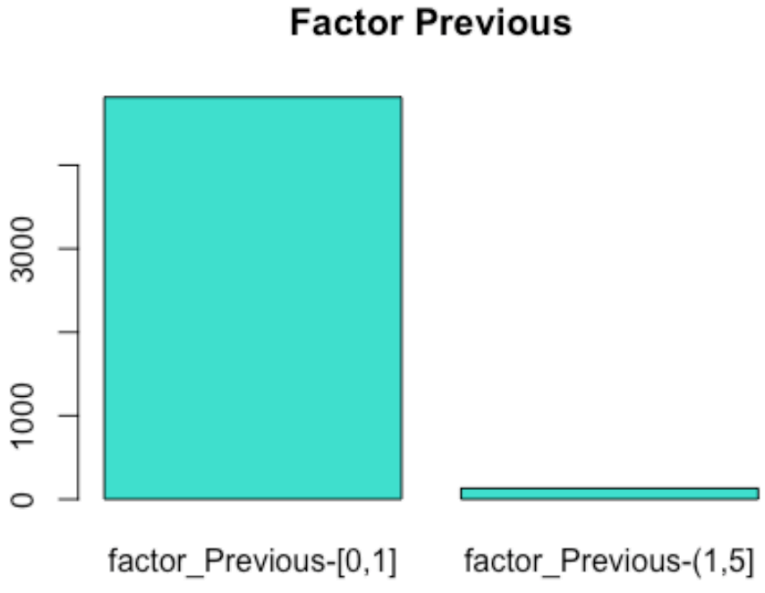
```
df$factor_Previous<-factor(cut(df$previous,include.lowest=T,breaks=c(0,1,5)))

summary(df$factor_Previous)
```

```
## [0,1] (1,5]
##  4815   131
```

```
tapply(df$previous,df$factor_Previous,median)
```

```
## [0,1] (1,5]
##     0     2
```

```
levels(df$factor_Previous)<-
paste("factor_Previous-",levels(df$factor_Previous),sep="")
table(df$factor_Previous)
## factor_Previous-[0,1] factor_Previous-(1,5]
##                  4815                   131

barplot(summary(df$factor_Previous), main="Factor
Previous",col=("turquoise"),cex.names=1.0)
```



```
#Amb aquesta discretització podem comprobar que el nombre de cops que s'ha contactat prèviament amb l'individu és majoritariament 0 o 1 i com a màxim una mitja de 5 cops.
```

Factor emp.var.rate

```
quantile(df$emp.var.rate,seq(0,1,0.2),na.rm=TRUE)
```

```
##   0%  20%  40%  60%  80% 100%
## -3.4 -1.8 -0.1  1.4  1.4  1.4
```

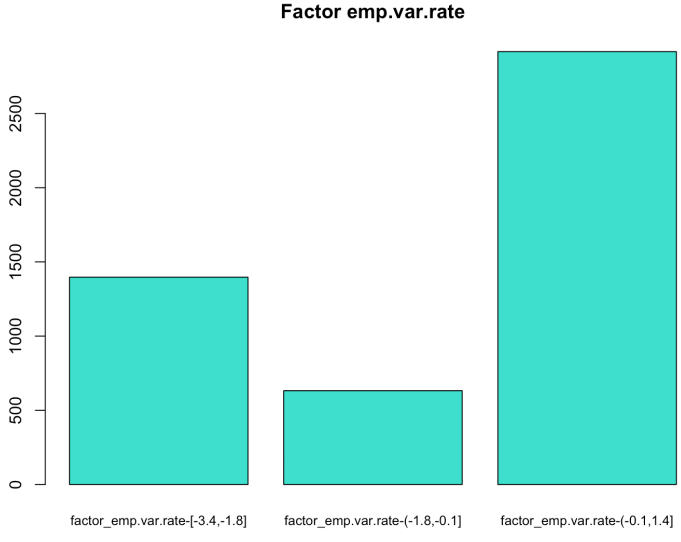
```
df$factor_emp.var.rate<-
factor(cut(df$emp.var.rate,include.lowest=T,breaks=c(-3.4,-1.8,-0.1,1.4)))

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

```
## [-3.4,-1.8] (-1.8,-0.1]  (-0.1,1.4]
##        1397         632        2917
```

```
tapply(df$emp.var.rate,df$factor_emp.var.rate,median)
```

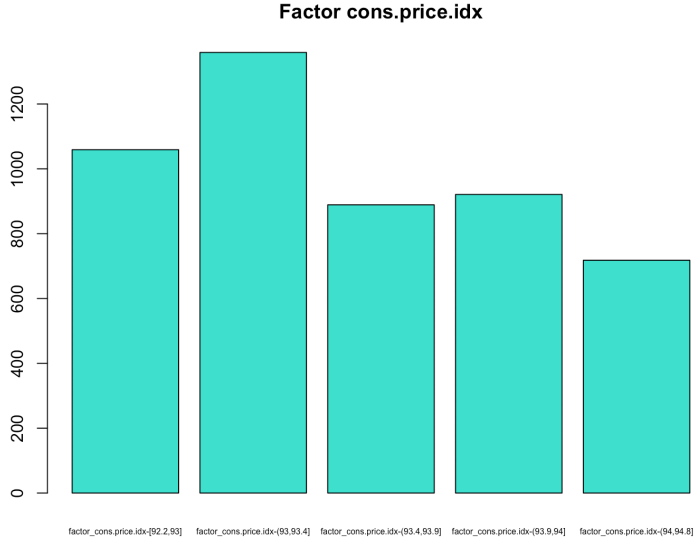
```
## [-3.4,-1.8] (-1.8,-0.1]  (-0.1,1.4]
##        -1.8        -0.1         1.4
```

```
levels(df$factor_emp.var.rate)<-
paste("factor_emp.var.rate-",levels(df$factor_emp.var.rate),sep="")
table(df$factor_emp.var.rate)
## factor_emp.var.rate-[-3.4,-1.8] factor_emp.var.rate-(-1.8,-0.1]
##                            1397                             632
##  factor_emp.var.rate-(-0.1,1.4]
##                            2917

barplot(summary(df$factor_emp.var.rate), main="Factor
emp.var.rate",col=("turquoise"),cex.names=0.8)
```



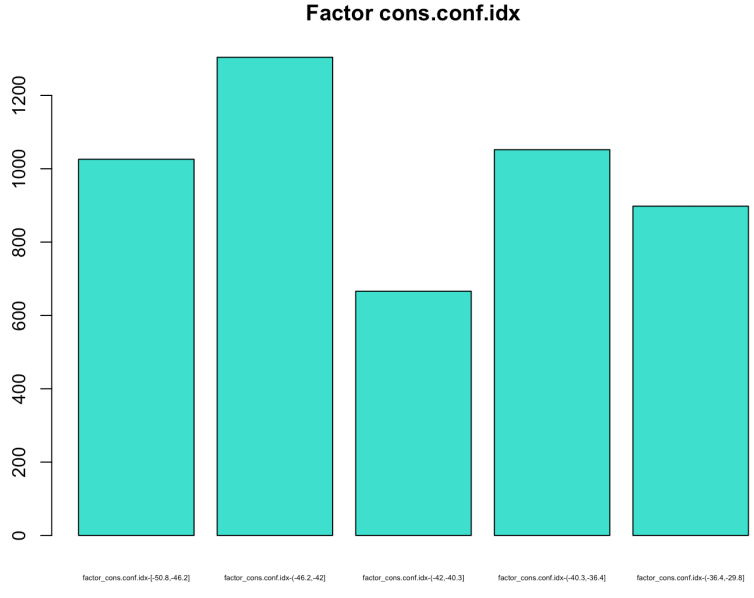
Factor cons.price.idx

```
quantile(df$cons.price.idx,seq(0,1,0.2),na.rm=TRUE)

##     0%    20%    40%    60%    80%   100%
## 92.201 92.963 93.444 93.918 93.994 94.767

df$factor_cons.price.idx<-
factor(cut(df$cons.price.idx,include.lowest=T,breaks=c(92.201,92.963,93.444,9
3.918,93.994,94.767)))

summary(df$factor_cons.price.idx)

##   [92.2,93]   (93,93.4] (93.4,93.9]   (93.9,94]   (94,94.8]
##        1059        1359         889         921         718

tapply(df$cons.price.idx,df$factor_cons.price.idx,median)

##   [92.2,93]   (93,93.4] (93.4,93.9]   (93.9,94]   (94,94.8]
##      92.893      93.200      93.918      93.994      94.465

levels(df$factor_cons.price.idx)<-
paste("factor_cons.price.idx-",levels(df$factor_cons.price.idx),sep="")
table(df$factor_cons.price.idx)
##   factor_cons.price.idx-[92.2,93]   factor_cons.price.idx-(93,93.4]
##                              1059                              1359
## factor_cons.price.idx-(93.4,93.9]   factor_cons.price.idx-(93.9,94]
##                               889                               921
##   factor_cons.price.idx-(94,94.8]
##                               718

barplot(summary(df$factor_cons.price.idx), main="Factor
cons.price.idx",col=("turquoise"),cex.names=0.5)
```



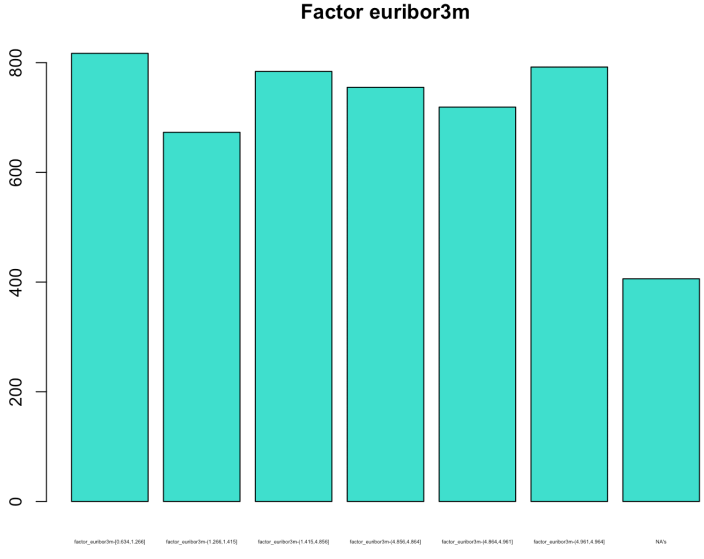
Factor cons.conf.idx

```
quantile(df$cons.conf.idx,seq(0,1,0.2),na.rm=TRUE)

##    0%   20%   40%   60%   80%  100%
## -50.8 -46.2 -42.0 -40.3 -36.4 -29.8

df$factor_cons.conf.idx<-
factor(cut(df$cons.conf.idx,include.lowest=T,breaks=c(-50.8,-46.2,-42,-40.3,-
36.4,-29.8)))

summary(df$factor_cons.conf.idx)

## [-50.8,-46.2]   (-46.2,-42]   (-42,-40.3] (-40.3,-36.4] (-36.4,-29.8]
##          1026          1304           666          1052           898

tapply(df$cons.conf.idx,df$factor_cons.conf.idx,median)

## [-50.8,-46.2]   (-46.2,-42]   (-42,-40.3] (-40.3,-36.4] (-36.4,-29.8]
##         -46.2         -42.7         -41.8         -36.4         -36.1

levels(df$factor_cons.conf.idx)<-
paste("factor_cons.conf.idx-",levels(df$factor_cons.conf.idx),sep="")
table(df$factor_cons.conf.idx)
## factor_cons.conf.idx-[-50.8,-46.2]   factor_cons.conf.idx-(-46.2,-42]
##                               1026                               1304
##   factor_cons.conf.idx-(-42,-40.3] factor_cons.conf.idx-(-40.3,-36.4]
##                                666                               1052
## factor_cons.conf.idx-(-36.4,-29.8]
##                                898

barplot(summary(df$factor_cons.conf.idx), main="Factor
cons.conf.idx",col=("turquoise"),cex.names=0.4)
```



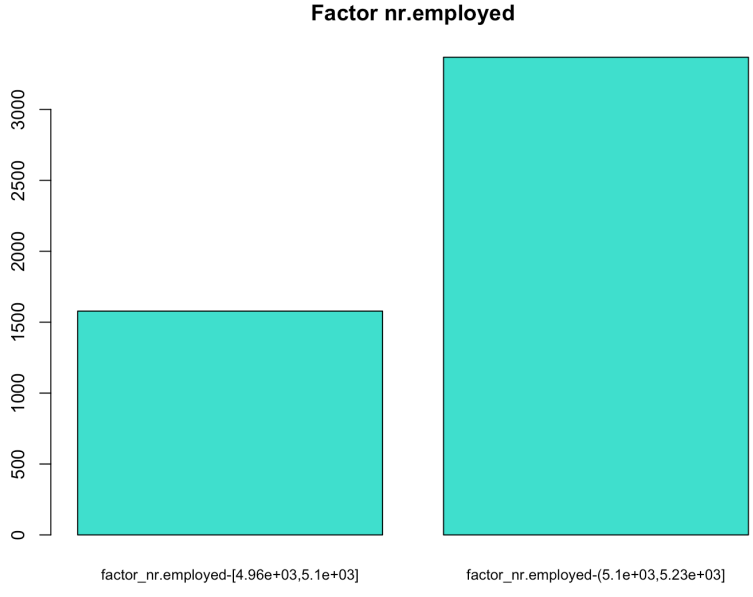
Factor euribor3m

```
quantile(df$euribor3m,seq(0,1,0.15),na.rm=TRUE)

##    0%   15%   30%   45%   60%   75%   90%
## 0.634 1.266 1.415 4.856 4.864 4.961 4.964

df$factor_euribor3m<-
factor(cut(df$euribor3m,include.lowest=T,breaks=c(0.634,1.266,1.415,4.856,4.8
64,4.961,4.964)))

summary(df$factor_euribor3m)

## [0.634,1.266] (1.266,1.415] (1.415,4.856] (4.856,4.864] (4.864,4.961]
##           817           673           784           755           719
## (4.961,4.964]          NA's
##           792           406

tapply(df$euribor3m,df$factor_euribor3m,median)

## [0.634,1.266] (1.266,1.415] (1.415,4.856] (4.856,4.864] (4.864,4.961]
##         0.884         1.334         4.153         4.858         4.960
## (4.961,4.964]
##         4.963

levels(df$factor_euribor3m)<-paste("factor_euribor3m-",levels(df$factor_euribor3m),sep="")
table(df$factor_euribor3m)
## factor_euribor3m-[0.634,1.266] factor_euribor3m-(1.266,1.415]
##                            817                            673
## factor_euribor3m-(1.415,4.856] factor_euribor3m-(4.856,4.864]
##                            784                            755
## factor_euribor3m-(4.864,4.961] factor_euribor3m-(4.961,4.964]
##                            719                            792

barplot(summary(df$factor_euribor3m), main="Factor euribor3m",col=("turquoise"),cex.names=0.3)
```



Factor nr.employed

```
quantile(df$nr.employed,seq(0,1,0.3),na.rm=TRUE)

##     0%    30%    60%    90%
## 4963.6 5099.1 5228.1 5228.1

df$factor_nr.employed<-
factor(cut(df$nr.employed,include.lowest=T,breaks=c(4963.6,5099.1,5228.1)))

summary(df$factor_nr.employed)

## [4.96e+03,5.1e+03] (5.1e+03,5.23e+03]
##               1578                3368

tapply(df$nr.employed,df$factor_nr.employed,median)

## [4.96e+03,5.1e+03] (5.1e+03,5.23e+03]
##             5099.1             5228.1

levels(df$factor_nr.employed)<-
paste("factor_nr.employed-",levels(df$factor_nr.employed),sep="")
table(df$factor_nr.employed)
## factor_nr.employed-[4.96e+03,5.1e+03]
##                                  1578
## factor_nr.employed-(5.1e+03,5.23e+03]
##                                  3368

barplot(summary(df$factor_nr.employed), main="Factor
nr.employed",col=("turquoise"),cex.names=0.8)
```



PROFILING

Numeric target (Duration)

El profiling s'utilitza per acabar de perfilar la nostra mostra

Ara procedirem a fer el profiling que ens demana del nostre target numeric (duration) i llavors hem d'utilitzar les variables originals i els factors menys el factor_duration, ja que es una variable que prove de la variable original i no volem aquesta informació

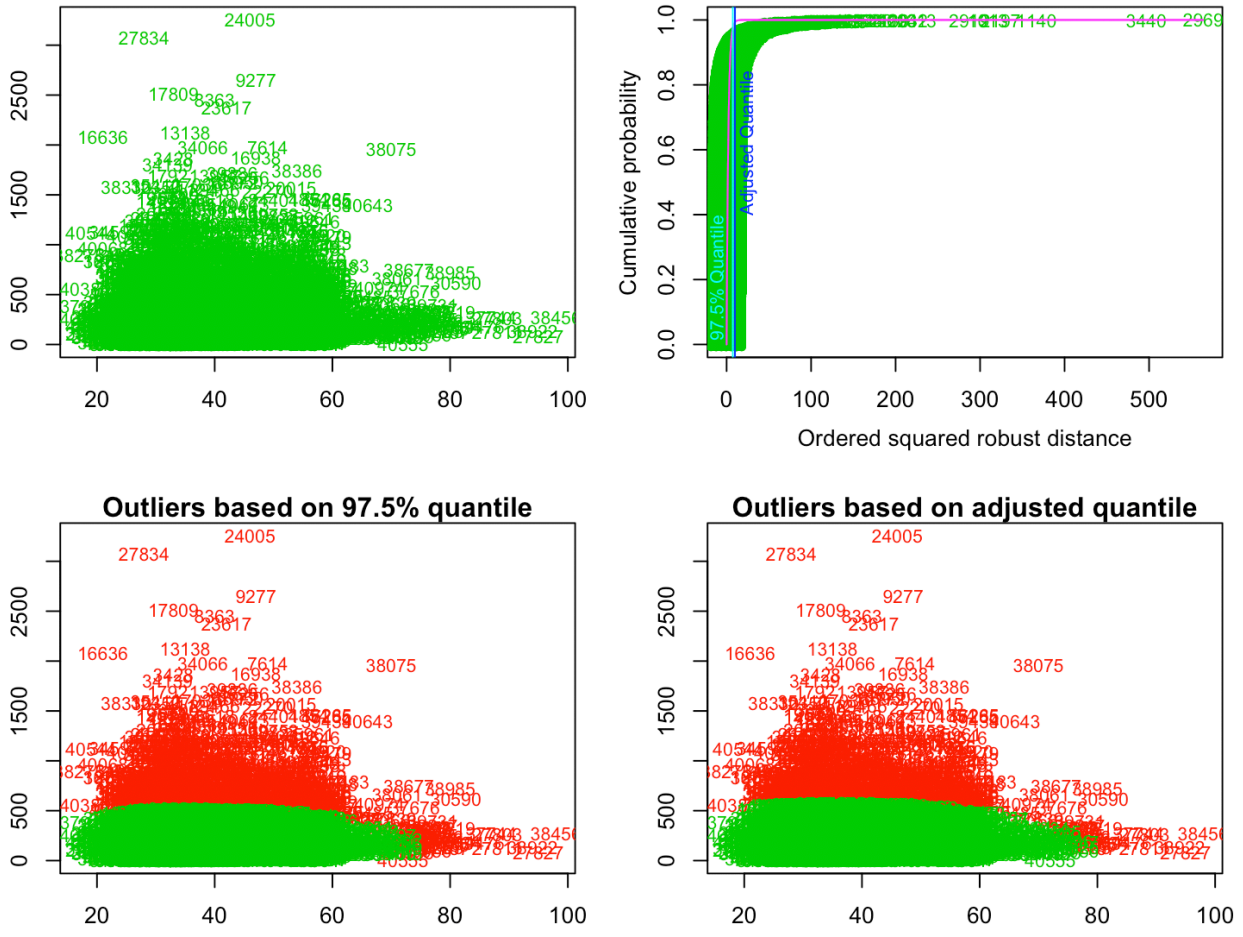
Per tal de observar la relacio del nostre target numeric amb les altres variables utilitzem la eina condes que ens proporciona informacio de les relacions entre les variables indicades i el target.

```
df$varauxiliar <- NULL #borrem la variable auxiliar creada
df$aux <- NULL
#Despres de discretitzar les nostres variables tenim un total de 35 variables
#names(df)

#Description continuous by quantitative variables and/or by categorical
variables
library(FactoMineR)
summary(df[,vars_con])
```

```
##       age           duration         campaign          pdays
##  Min.   :17.00   Min.   :   1.0   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:32.00   1st Qu.: 104.0   1st Qu.: 1.000   1st Qu.: 5.348
##  Median :38.00   Median : 182.0   Median : 2.000   Median : 5.657
##  Mean   :40.05   Mean   : 262.8   Mean   : 2.389   Mean   : 5.706
##  3rd Qu.:47.00   3rd Qu.: 329.0   3rd Qu.: 3.000   3rd Qu.: 5.990
##  Max.   :81.00   Max.   :2122.0   Max.   :14.000   Max.   :15.000
##     previous       emp.var.rate      cons.price.idx  cons.conf.idx
##  Min.   :0.0000   Min.   :-3.4000   Min.   :92.20   Min.   :-50.80
##  1st Qu.:0.0000   1st Qu.:-1.8000   1st Qu.:93.08   1st Qu.:-42.70
##  Median :0.0000   Median : 1.1000   Median :93.92   Median :-41.80
##  Mean   :0.1708   Mean   : 0.1074   Mean   :93.59   Mean   :-40.62
##  3rd Qu.:0.0000   3rd Qu.: 1.4000   3rd Qu.:93.99   3rd Qu.:-36.40
##  Max.   :5.0000   Max.   : 1.4000   Max.   :94.77   Max.   :-29.80
##    euribor3m      nr.employed
##  Min.   :0.634   Min.   :4964
##  1st Qu.:1.344   1st Qu.:5099
##  Median :4.857   Median :5191
##  Mean   :3.649   Mean   :5168
##  3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :5.045   Max.   :5228
```

```
library(mvoutlier)

vars_resu <-names(df)[c(1,11)]
```

```
vars_resu

aq.plot(df[,vars_resu])
```




```
#vars_res<-names(df)[c(11,21)]
vars<-unique(c(vars_con,vars_dis))
#vars
condes(df, which(names(df) == "duration"))

## $quanti
##                  correlation      p.value
## pdays             0.52693895 0.000000e+00
## previous          0.02859224 4.435374e-02
## errors_indiv     -0.03476735 1.447588e-02
## nr.employed      -0.03619203 1.091224e-02
## campaign         -0.04179341 3.284450e-03
## missings_indiv -0.07328498 2.474678e-07
```

```
##
## $quali
##                                R2       p.value
## factor_duration       0.8271873066  0.000000e+00
## factor_Pdays          0.4046346310  0.000000e+00
## y                     0.1863696068 9.891372e-224
## poutcome              0.0041874670  3.132625e-05
## month                 0.0073478185  3.327154e-05
## factor_cons.price.idx 0.0039803615  5.696640e-04
## factor_Previous       0.0019228074  2.038492e-03
## day_of_week           0.0029955473  5.075577e-03
## factor_cons.conf.idx  0.0026002247  1.194404e-02
## contact               0.0011105265  1.909343e-02
## default               0.0009897216  2.693284e-02
## factor_campaign       0.0013152237  3.866909e-02
##
## $category
##                                       Estimate       p.value
## factor_Pdays-(5.99,15]              277.390363  0.000000e+00
## factor_duration-(504,2.12e+03]      547.162252  0.000000e+00
## Y_yes                               169.675531 9.891372e-224
## factor_duration-(329,504]           138.462468  3.985182e-48
## Poutcome_success                     62.641078  7.933875e-06
## factor_cons.price.idx-(93.4,93.9]    27.117765  2.010384e-04
## Month_jul                            12.946601  2.986551e-04
## factor_Previous-(1,5]                34.966136  2.038492e-03
## Contact_cellular                      8.850090  1.909343e-02
## Default_no                            9.913335  2.693284e-02
## Month_dec                           104.090396  2.868142e-02
## Day_of_week_tue                      14.917687  4.872420e-02
## Education_illiterate                178.585152  4.932974e-02
## Education_university.degree         -38.308971  3.857651e-02
## factor_cons.conf.idx-(-36.4,-29.8]  -13.574401  3.768483e-02
## factor_cons.conf.idx-(-42,-40.3]    -17.926886  2.695593e-02
## Default_unknown                      -9.913335  2.693284e-02
## Contact_telephone                    -8.850090  1.909343e-02
## Month_jun                           -37.404273  1.736971e-02
## factor_campaign-(3,14]              -16.741883  1.148865e-02
## Job_technician                      -25.341033  1.106827e-02
## Day_of_week_mon                     -19.239047  7.577039e-03
## Month_aug                           -39.248662  5.073298e-03
## factor_cons.price.idx-(93,93.4]     -19.809889  2.312144e-03
## factor_Previous-[0,1]               -34.966136  2.038492e-03
## factor_duration-(182,236]           -56.414720  8.764699e-09
## factor_Pdays-(5.66,5.99]            -45.516643  3.987470e-13
## factor_duration-(139,182]          -103.067426  8.297196e-27
## factor_duration-(104,139]          -141.910732  3.245807e-49
## factor_Pdays-(5.35,5.66]           -106.671095  4.639847e-66
## factor_duration-(68,104]           -177.221056  2.195363e-78
```

```
## factor_Pdays-[1,5.35]              -125.202625  2.136961e-91
## factor_duration-[1,68]             -222.636796 8.250905e-127
## Y_no                               -169.675531 9.891372e-224
```

```
#S'utilitza per fer totes les combinacions possibles de variables numèriques i factorials
#Tindrem les variables que tenen un pvalor a partir d'un llindar del pvalor acceptat. No ens surten totes les variables estudiades, només les que tenen una mena de relació
#Con el p valor muy bajo entonces rechazamos la hipòtesi nula

#$quanti: Com podem observar la variable pdays es la que te mes relacio amb la nostra variable target (duration), es a dir, quant mes gran sigui la duracio de la trucada tenim una correlacio mes gran amb aquesta i veiem que com a relació inversament proporcional tenim campaign
#$quali: La variable qualitativa que té més realció amb el nostre target es el seu mateix factor (factor_duration) com és obvi, pero seguidament tenim el factor_Pdays i la nostra variable y
#$category: Podem observar que tenim una relacio dependent molt forta dels mesos i últims contactes, podem veure que ha tingut exit i majoritariament la y és yes
```

Y (target qual)

Per analitzar les relacions de la nostre variable qualitativa utilitzem l'eina catdes que de la mateixa manera que el condes ens mostrar? les seves relacions.

```
df_catdes<-df[c(1:21)]
catdes(df_catdes,21)
```

```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## =================================================================================
##                     p.value df
## poutcome    2.884978e-155  2
## month        2.020968e-82  9
## contact      8.049707e-27  1
## job          5.149262e-24 11
## default      7.888260e-14  1
## education    1.246599e-05  7
## marital      4.868728e-03  3
## day_of_week  3.137547e-02  4
##
## Description of each cluster by the categories
## =============================================
## $Y_no
##                                          Cla/Mod    Mod/Cla     Global
## poutcome=Poutcome_nonexistent           91.01964 89.4918372 86.4537000
## contact=Contact_telephone               94.44444 39.4803403 36.7569753
```

```
## default=Default_unknown                  94.67054 22.4649345 20.8653457
## month=Month_may                          92.83951 34.5826627 32.7537404
## job=Job_blue-collar                      92.74476 24.3964130 23.1298019
## education=Education_basic.9y             92.09486 16.0726604 15.3457339
## month=Month_jul                          90.92945 18.6709588 18.0549939
## education=Education_basic.6y             93.28358  5.7484479  5.4185200
## marital=Marital_married                  88.96667 61.3704300 60.6550748
## job=Job_services                         91.54334  9.9563118  9.5632835
## job=Job_technician                       90.17857 16.2566107 15.8511929
## day_of_week=Day_of_week_mon              89.79592 21.2462635 20.8046907
## education=NA                             83.33333  4.0239135  4.2458552
## education=Education_professional.course 85.21008 11.6578524 12.0299232
## day_of_week=Day_of_week_tue              85.16058 17.6822258 18.2571775
## education=Education_university.degree    85.93540 29.3630720 30.0444804
## marital=Marital_single                   85.47567 27.0636928 27.8406793
## poutcome=Poutcome_failure                83.26693  9.6114049 10.1496159
## job=Job_admin.                           85.16526 25.4771212 26.3040841
## month=Month_apr                          78.29181  5.0586342  5.6813587
## month=Month_dec                          45.45455  0.2299379  0.4448039
## job=Job_student                          65.71429  1.5865716  2.1229276
## job=Job_retired                          72.81553  3.4490688  4.1649818
## month=Month_mar                          50.74627  0.7817889  1.3546300
## month=Month_sep                          50.72464  0.8047827  1.3950667
## default=Default_no                       86.15227 77.5350655 79.1346543
## month=Month_oct                          48.19277  0.9197517  1.6781237
## contact=Contact_cellular                 84.14322 60.5196597 63.2430247
## poutcome=Poutcome_success                23.21429  0.8967579  3.3966842
##                                               p.value     v.test
## poutcome=Poutcome_nonexistent           3.543373e-50  14.895160
## contact=Contact_telephone               1.650430e-29  11.279842
## default=Default_unknown                 6.847442e-16   8.073209
## month=Month_may                         1.529311e-14   7.685055
## job=Job_blue-collar                     2.309977e-09   5.974358
## education=Education_basic.9y            6.478104e-05   3.994682
## month=Month_jul                         1.804548e-03   3.120646
## education=Education_basic.6y            3.345680e-03   2.934052
## marital=Marital_married                 5.727878e-03   2.762966
## job=Job_services                        8.657080e-03   2.625307
## job=Job_technician                      3.216891e-02   2.142305
## day_of_week=Day_of_week_mon             3.661258e-02   2.090058
## education=NA                            4.459048e-02  -2.008497
## education=Education_professional.course 3.369438e-02  -2.123710
## day_of_week=Day_of_week_tue             5.704442e-03  -2.764304
## education=Education_university.degree   5.300406e-03  -2.788186
## marital=Marital_single                  1.198449e-03  -3.239249
## poutcome=Poutcome_failure               1.167715e-03  -3.246651
## job=Job_admin.                          4.654028e-04  -3.499917
## month=Month_apr                         2.649823e-06  -4.696249
## month=Month_dec                         1.944834e-06  -4.759074
```

```
## job=Job_student                            2.045387e-09  -5.994161
## job=Job_retired                            1.710143e-09  -6.023188
## month=Month_mar                            6.474585e-14  -7.498107
## month=Month_sep                            2.609525e-14  -7.616349
## default=Default_no                         6.847442e-16  -8.073209
## month=Month_oct                            6.812368e-19  -8.877918
## contact=Contact_cellular                   1.650430e-29 -11.279842
## poutcome=Poutcome_success                  2.944669e-88 -19.916208
##
## $Y_yes
##                                              Cla/Mod   Mod/Cla     Global
## poutcome=Poutcome_success                  76.785714 21.608040  3.3966842
## contact=Contact_cellular                   15.856777 83.082077 63.2430247
## month=Month_oct                            51.807229  7.202680  1.6781237
## default=Default_no                         13.847726 90.787270 79.1346543
## month=Month_sep                            49.275362  5.695142  1.3950667
## month=Month_mar                            49.253731  5.527638  1.3546300
## job=Job_retired                            27.184466  9.380235  4.1649818
## job=Job_student                            34.285714  6.030151  2.1229276
## month=Month_dec                            54.545455  2.010050  0.4448039
## month=Month_apr                            21.708185 10.217755  5.6813587
## job=Job_admin.                             14.834743 32.328308 26.3040841
## poutcome=Poutcome_failure                  16.733068 14.070352 10.1496159
## marital=Marital_single                     14.524328 33.500838 27.8406793
## education=Education_university.degree      14.064603 35.008375 30.0444804
## day_of_week=Day_of_week_tue                14.839424 22.445561 18.2571775
## education=Education_professional.course    14.789916 14.740369 12.0299232
## education=NA                               16.666667  5.862647  4.2458552
## day_of_week=Day_of_week_mon                10.204082 17.587940 20.8046907
## job=Job_technician                          9.821429 12.897822 15.8511929
## job=Job_services                            8.456660  6.700168  9.5632835
## marital=Marital_married                    11.033333 55.443886 60.6550748
## education=Education_basic.6y                6.716418  3.015075  5.4185200
## month=Month_jul                             9.070549 13.567839 18.0549939
## education=Education_basic.9y                7.905138 10.050251 15.3457339
## job=Job_blue-collar                         7.255245 13.902848 23.1298019
## month=Month_may                             7.160494 19.430486 32.7537404
## default=Default_unknown                     5.329457  9.212730 20.8653457
## contact=Contact_telephone                   5.555556 16.917923 36.7569753
## poutcome=Poutcome_nonexistent               8.980355 64.321608 86.4537000
##                                                p.value     v.test
## poutcome=Poutcome_success                  2.944669e-88  19.916208
## contact=Contact_cellular                   1.650430e-29  11.279842
## month=Month_oct                            6.812368e-19   8.877918
## default=Default_no                         6.847442e-16   8.073209
## month=Month_sep                            2.609525e-14   7.616349
## month=Month_mar                            6.474585e-14   7.498107
## job=Job_retired                            1.710143e-09   6.023188
## job=Job_student                            2.045387e-09   5.994161
```

```
## month=Month_dec                            1.944834e-06   4.759074
## month=Month_apr                            2.649823e-06   4.696249
## job=Job_admin.                             4.654028e-04   3.499917
## poutcome=Poutcome_failure                  1.167715e-03   3.246651
## marital=Marital_single                     1.198449e-03   3.239249
## education=Education_university.degree      5.300406e-03   2.788186
## day_of_week=Day_of_week_tue                5.704442e-03   2.764304
## education=Education_professional.course    3.369438e-02   2.123710
## education=NA                               4.459048e-02   2.008497
## day_of_week=Day_of_week_mon                3.661258e-02  -2.090058
## job=Job_technician                         3.216891e-02  -2.142305
## job=Job_services                           8.657080e-03  -2.625307
## marital=Marital_married                    5.727878e-03  -2.762966
## education=Education_basic.6y               3.345680e-03  -2.934052
## month=Month_jul                            1.804548e-03  -3.120646
## education=Education_basic.9y               6.478104e-05  -3.994682
## job=Job_blue-collar                        2.309977e-09  -5.974358
## month=Month_may                            1.529311e-14  -7.685055
## default=Default_unknown                    6.847442e-16  -8.073209
## contact=Contact_telephone                  1.650430e-29 -11.279842
## poutcome=Poutcome_nonexistent              3.543373e-50 -14.895160
##
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                        Eta2       P-value
## duration       0.186369607 9.891372e-224
## nr.employed    0.139052649 5.557605e-163
## euribor3m      0.104758799 5.493737e-121
## emp.var.rate   0.099078243 3.487741e-114
## previous       0.070648755  9.329422e-81
## pdays          0.032371630  2.943423e-37
## cons.price.idx 0.019937283  1.907193e-23
## campaign       0.005057924  5.536389e-07
##
## Description of each cluster by quantitative variables
## =====================================================
## $Y_no
##                    v.test Mean in category Overall mean sd in category
## nr.employed     26.222421     5177.8744999 5167.8073595     64.2441089
## euribor3m       22.760322        3.8560536    3.6487535      1.6188731
## emp.var.rate    22.134632        0.2901587    0.1073999      1.4661991
## cons.price.idx   9.929243       93.6160205   93.5857345      0.5562445
## campaign         5.001143        2.4413845    2.3891187      2.0381577
## pdays          -12.652182        5.6490528    5.7062970      0.6031064
## previous       -18.691123        0.1230168    0.1708451      0.3957657
## duration       -30.357828      221.8063923  262.7672867    200.3541053
##                 Overall sd       p.value
## nr.employed     72.8658491 1.475237e-151
```

```
## euribor3m        1.7286683 1.134100e-114
## emp.var.rate     1.5670994 1.467071e-108
## cons.price.idx   0.5789159  3.106051e-23
## campaign         1.9835304  5.699132e-07
## pdays            0.8587295  1.088103e-36
## previous         0.4856692  5.846876e-78
## duration       256.0881160 1.980616e-202
##
## $Y_yes
##                   v.test Mean in category Overall mean sd in category
## duration        30.357828       561.157454  262.7672867    386.8354045
## previous        18.691123         0.519263    0.1708451      0.8216383
## pdays           12.652182         6.123307    5.7062970      1.8060480
## campaign        -5.001143         2.008375    2.3891187      1.4727896
## cons.price.idx  -9.929243        93.365109   93.5857345      0.6835676
## emp.var.rate   -22.134632        -1.223953    0.1073999      1.6338789
## euribor3m      -22.760322         2.138623    3.6487535      1.7527742
## nr.employed    -26.222421      5094.470687 5167.8073595     88.3423897
##                  Overall sd       p.value
## duration       256.0881160 1.980616e-202
## previous         0.4856692  5.846876e-78
## pdays            0.8587295  1.088103e-36
## campaign         1.9835304  5.699132e-07
## cons.price.idx   0.5789159  3.106051e-23
## emp.var.rate     1.5670994 1.467071e-108
## euribor3m        1.7286683 1.134100e-114
## nr.employed     72.8658491 1.475237e-151
```

#Podem veure que els factors que afecten més a l'hora de que el individu contracti el producte promocionat (var Y = yes) son el èxit o no de les anteriors campanyes, el nombre de contactes, la duració i altres factors relacionats amb les èpoques/mesos de l'any i l'status de l'individu.