Deliverable 1

Guillem Valls, Sergio Mazzariol

Table of Contents

Preparación de la muestra

Establecemos el directorio de trabajo, luego importamos todos los datos del archivo csv bank-additional-full y establecemos una semilla para obtener siempre la misma muestra "aleatoria". Obtenemos 5000 individuos que se usarán para el análisis a lo largo de toda la asignatura. Partimos siempre del mismo fichero, data-INI.RData, para asegurarnos que se usa siempre la misma muestra ya generada.

```
#setwd("C:/Users/Sergio/Dropbox/UPC/FIB/Analisis de datos y explotacion de la informacion (ADEI)/FIB-ADEI-Big-Data-Analysis")
setwd("C:/Users/usuario/Documents/ADEI/FIB-ADEI-Big-Data-Analysis")

# Data file alread
df<-read.table('bank-additional-full.csv',header=TRUE,sep=";")

# Select your 5000 register sample (random sample)
set.seed(19101990)
llista<-sample(size=5000,x=1:nrow(df),replace=FALSE)
llista<-sort(llista)

#llista
df<-df[llista,]
dim(df)

## [1] 5000   21

#save.image("set-datos.RData")
load("data-INI.RData")
```

Inicializamos datos y funciones

Creamos un dataframe que llamamos data quality report "dqr" para almacenar missings, errors, outliers. También creamos uno para los datos individuales "dqri". Inicializamos el "dqr" todo a 0, y el dqri lo inicializamos a 0 pero después de eliminar los individuos que nos dan outliers o errores en las variables target. Declaramos la función calcQ que nos permitirá discriminar los outliers leves y severos en los boxplots. Para poder tratar los datos con mayor facilidad separamos las variables en tres grupos, las variables target "duration, y", las variables categóricas "job", "marital", "education", "default", "housing", "loan", "contact", "month", "day_of_week", "poutcome" y las variables númericas "age", "campaign", "pdays", "previous", "emp.var.rate", "cons.price.idx", "cons.conf.idx", "euribor3m", "nr.employed"

```
dqr <- data.frame(variable=character(), missings=integer(), errors=intege
r(), outliers=integer())
dqr[length(names(df)),2:4]<-0
dqr$variable <-names(df)
dqr[,2:4]<-0

dqri <- data.frame(missings=integer(), errors=integer(), outliers=integer
())

calcQ <- function(x) {
  s.x <- summary(x)
  iqr<-s.x[5]-s.x[2]
  list(souti=s.x[2]-3*iqr, mouti=s.x[2]-1.5*iqr, min=s.x[1], q1=s.x[2], q
2=s.x[3],
       q3=s.x[5], max=s.x[6], mouts=s.x[5]+1.5*iqr, souts=s.x[5]+3*iqr )
}

df[1,]

##    age        job marital education default housing loan   contact mon
th
## 20  39 management  single  basic.9y unknown      no   no telephone   m
ay
##    day_of_week duration campaign pdays previous    poutcome emp.var.ra
te
## 20         mon      195        1   999        0 nonexistent          1
.1
##    cons.price.idx cons.conf.idx euribor3m nr.employed  y
## 20         93.994         -36.4     4.857        5191 no

vars_target<-c("duration","y");vars_target

## [1] "duration" "y"

vars_cat<-c("job", "marital", "education", "default", "housing", "loan",
"contact", "month", "day_of_week", "poutcome");vars_cat

##  [1] "job"         "marital"     "education"   "default"     "housing"
##  [6] "loan"        "contact"     "month"       "day_of_week" "poutcome
"

vars_num<-c("age", "campaign", "pdays", "previous", "emp.var.rate", "cons
.price.idx", "cons.conf.idx", "euribor3m", "nr.employed");vars_num

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

Análisis y exploración de datos

Empezamos con la exploración de datos, verificamos los nombres de las variables, también un summary para comprobar que los datos son correctos.

```
summary(df)
```

```
##       age                  job           marital
##  Min.   :18.00   admin.     :1285   divorced: 584
##  1st Qu.:32.00   blue-collar:1130   married :2995
##  Median :38.00   technician : 816   single  :1413
##  Mean   :40.18   services   : 451   unknown :   8
##  3rd Qu.:47.00   management : 352
##  Max.   :92.00   retired    : 223
##                  (Other)    : 743
##                education       default        housing          loan
##  university.degree  :1469   no     :3949   no     :2244   no     :4141
##  high.school        :1142   unknown:1051   unknown: 113   unknown: 113
##  basic.9y           : 756   yes    :   0   yes    :2643   yes    : 746
##  professional.course: 610
##  basic.4y           : 510
##  basic.6y           : 271
##  (Other)            : 242
##       contact          month      day_of_week    duration
##  cellular :3207   may    :1682   fri: 960    Min.   :   0.0
##  telephone:1793   jul    : 866   mon:1058    1st Qu.: 103.0
##                   aug    : 767   thu:1008    Median : 179.0
##                   jun    : 617   tue: 954    Mean   : 263.3
##                   nov    : 514   wed:1020    3rd Qu.: 322.0
##                   apr    : 322               Max.   :4199.0
##                   (Other): 232
##     campaign          pdays          previous             poutcome
##  Min.   : 1.000   Min.   :  0.0   Min.   :0.0000   failure    : 546
##  1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000   nonexistent:4298
##  Median : 2.000   Median :999.0   Median :0.0000   success    : 156
##  Mean   : 2.579   Mean   :964.3   Mean   :0.1784
##  3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :56.000   Max.   :999.0   Max.   :6.0000
##
##   emp.var.rate      cons.price.idx  cons.conf.idx      euribor3m
##  Min.   :-3.40000   Min.   :92.20   Min.   :-50.80   Min.   :0.634
##  1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.334
##  Median : 1.10000   Median :93.44   Median :-41.80   Median :4.857
##  Mean   : 0.05264   Mean   :93.56   Mean   :-40.54   Mean   :3.585
##  3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961
##  Max.   : 1.40000   Max.   :94.77   Max.   :-26.90   Max.   :5.045
##
##   nr.employed     y
##  Min.   :4964   no :4455
##  1st Qu.:5099   yes: 545
```

```
##  Median :5191
##  Mean   :5166
##  3rd Qu.:5228
##  Max.   :5228
##
```

Tratamiento de las Variables target

En primer lugar trataremos las variables target porque de estas se pueden desprender errores y outliers que implicarán eliminación de individuos ya que estos errores no pueden imputarse, sería falsificación de la variable target. Tenemos dos variables targets, una categórica y otra numérica, empezamos con la categórica.

Y

Hacemos un summary de la variable y podemos ver que los únicos valores que toma es yes o no, de los cuales podemos decir que no hay errores, outliers o missings.

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

```
##   no  yes
## 4455  545
```

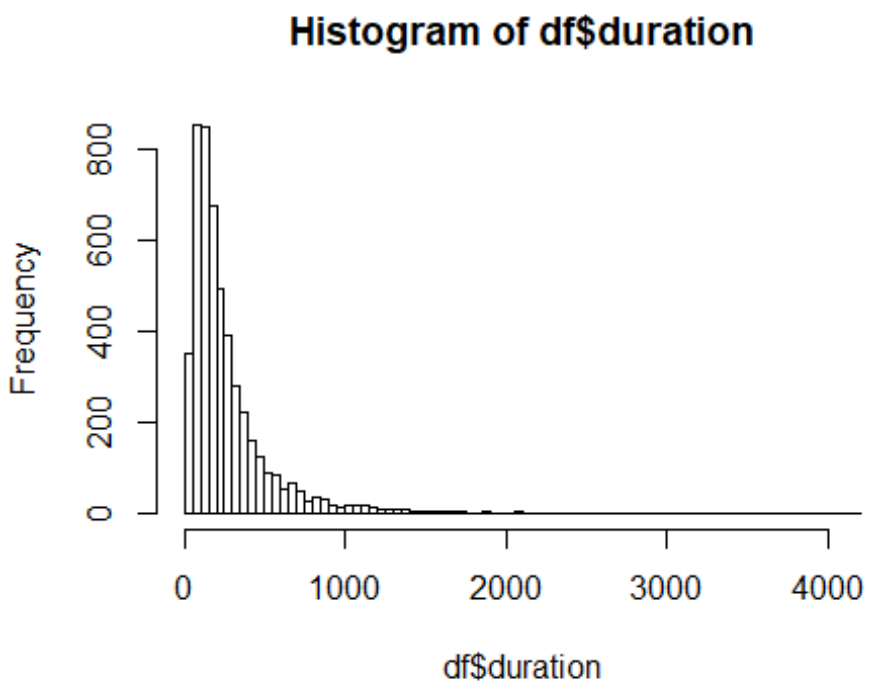
Duration

Vemos que hay valores muy pequeños, incluso 0, también valores muy grandes. Miramos distribución en el histograma. De él se desprende que las llamadas son mayormente de aproximadamente 250 minutos, como ya anticipaba el summary. Hacemos boxplot para ver outliers y solo se contemplan outliers superiores con la función calcQ que fija límite soft y extremo. Detectar outliers aplicando el linde proporcionado por calcQ echaría a perder la muestra, así que mejor se revisan los 10 valores más extremos y vemos que los últimos 6 abarcan un intervalo superior de duración al resto de la muestra, es decir, 4994 individuos están en el intervalo [0,2078] mientras que estos 6 abarcan un intervalo más extenso, [2079,4199]. Hacemos boxplot nuevamente para ver el resultado el cual almacenamos en nuestro data frame. Luego procedemos a revisar los errores, los cuales consideramos que pueden ser llamadas con una duración inferior a 5 segundos. Tanto errores como outliers son eliminados de la muestra.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   103.0   179.0   263.3   322.0  4199.0
```

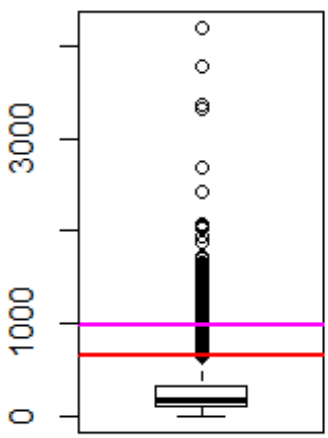
```
hist(df$duration,100)
```

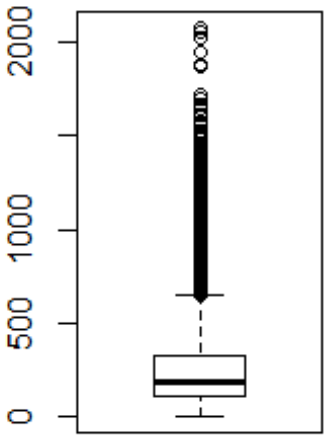



```r
par(mfrow=c(1,2))
boxplot(df$duration)
aux<-calcQ(df$duration)
abline(h=aux[8],col="red",lwd=2)
abline(h=aux[9],col="magenta",lwd=2)
aux<-order(df$duration,decreasing=TRUE)[1:10];df[aux,'duration']
```

```
##  [1] 4199 3785 3366 3322 2692 2420 2078 2053 2028 1946
```

```r
df<-df[-aux[1:6],]
boxplot(df$duration)
```






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

aux<-which(df$duration<5);length(aux);df[aux,'duration']

## [1] 4

## [1] 0 4 0 1

df<-df[-aux,]

dqr[dqr$variable=="duration","outliers"]<-6
dqr[dqr$variable=="duration","errors"]<-length(aux)

# Inicializamos el dqri ya que en este punto hemos eliminado todos los in
dividuos que se consideraban como outliers o errores en las variables tar
get.
dqri[nrow(df),]<-0
dqri[,]<-0
```

Tratamiento de variables no-Target Categóricas

Análisis de errores y missings

Primero realizamos un summary de todas las variables categóricas, para analizar sus valores. En este análisis podemos ver que la variable default tiene una cantidad alta de valores unknows, por lo que nos da indicios de que esta variable no nos será útil. Vemos que todas los factores con niveles unknow, menos "default" se pueden

considerar como missings, por lo que procedemos a pasar estos valores a NA´s, para esto utilizamos un bucle for. Para evitar realizar el cambio de variables que tengan una cantidad de uknowns mayor a 300, ya que en estos casos debe permanecer como un nivel más, como lo es en el caso de la variable “default”.

```
for(i in vars_cat){
  cat("################ ",i," ##################\n")
  print(summary(df[,i]))
}
```

```
## ################  job  ##################
##        admin.   blue-collar  entrepreneur     housemaid    management
##          1281          1128           189           119           351
##       retired self-employed      services       student    technician
##           222           166           451           109           815
##    unemployed       unknown
##           114            45
## ################  marital  ##################
## divorced  married   single  unknown
##      583     2988     1411        8
## ################  education  ##################
##            basic.4y            basic.6y            basic.9y
##                 508                 271                 756
##         high.school          illiterate professional.course
##                1138                   1                 609
##   university.degree             unknown
##                1466                 241
## ################  default  ##################
##      no unknown     yes
##    3940    1050       0
## ################  housing  ##################
##      no unknown     yes
##    2239     113    2638
## ################  loan  ##################
##      no unknown     yes
##    4132     113     745
## ################  contact  ##################
##  cellular telephone
##      3203      1787
## ################  month  ##################
##  apr  aug  dec  jul  jun  mar  may  nov  oct  sep
##  321  764   18  865  616   63 1680  513   80   70
## ################  day_of_week  ##################
##  fri  mon  thu  tue  wed
##  957 1058 1005  952 1018
## ################  poutcome  ##################
##     failure nonexistent     success
##         545        4289         156
```

```r
for(i in vars_cat){
  aux<-which(df[,i]=="unknown")
  if(length(aux)>0 && length(aux)<300){ # Solo si como máximo la variable tiene 300 unknowns (Para filtrar a default)
    cat(i, " -- ", length(aux), "\n")
    df[aux,i]<-NA
    dqri[aux,"missings"]<-dqri[aux,"missings"]+1
    df[,i]<-factor(df[,i])
  }
}
```

```
## job  --  45
## marital  --  8
## education  --  241
## housing  --  113
## loan  --  113
```

```r
# Para el data análisis guardamos los missings de las variables categóricas
for(i in vars_cat){
  dqr[dqr$variable==i,"missings"]<-sum(is.na(df[,i]))
}
```

Ahora realizamos la imputación de las variables categóricas. Contrastamos los summmaries originales e imputados, para comprobar que la imputación se hizo correctamente. Vemos que todo ha sido correcto y aceptamos estos datos, por lo que procedemos a almacenarlo en nuestro data frame que, por seguridad, solo sobrescribimos aquellas variables que han sido modificadas.

```r
aux2<-imputeMCA(df[,vars_cat],ncp=10)

for(i in vars_cat){
  cat("################ ",i," ##################\n")
  print(summary(df[,i]))
  print("--- --- --- ---")
  print(summary(aux2$completeObs[,i]))
}
```

```
## ################  job  ##################
##        admin.   blue-collar  entrepreneur     housemaid    management
##          1281          1128           189           119           351
##       retired self-employed      services       student    technician
##           222           166           451           109           815
##    unemployed          NA's
##           114            45
## [1] "--- --- --- ---"
##        admin.   blue-collar  entrepreneur     housemaid    management
##          1296          1156           189           119           351
##       retired self-employed      services       student    technician
##           222           166           451           109           817
##    unemployed
```

```
##          114
## ################  marital  ##################
## divorced  married   single     NA's
##      583     2988     1411        8
## [1] "--- --- --- ---"
## divorced  married   single
##      583     2996     1411
## ################  education  ##################
##            basic.4y            basic.6y            basic.9y
##                 508                 271                 756
##         high.school          illiterate professional.course
##                1138                   1                 609
##   university.degree                NA's
##                1466                 241
## [1] "--- --- --- ---"
##            basic.4y            basic.6y            basic.9y
##                 515                 271                 810
##         high.school          illiterate professional.course
##                1196                   1                 633
##   university.degree
##                1564
## ################  default  ##################
##      no unknown     yes
##    3940    1050       0
## [1] "--- --- --- ---"
##      no unknown
##    3940    1050
## ################  housing  ##################
##   no  yes NA's
## 2239 2638  113
## [1] "--- --- --- ---"
##   no  yes
## 2279 2711
## ################  loan  ##################
##   no  yes NA's
## 4132  745  113
## [1] "--- --- --- ---"
##   no  yes
## 4245  745
## ################  contact  ##################
##  cellular telephone
##      3203      1787
## [1] "--- --- --- ---"
##  cellular telephone
##      3203      1787
## ################  month  ##################
##  apr  aug  dec  jul  jun  mar  may  nov  oct  sep
##  321  764   18  865  616   63 1680  513   80   70
## [1] "--- --- --- ---"
##  apr  aug  dec  jul  jun  mar  may  nov  oct  sep
```

```
##  321  764   18  865  616   63 1680  513   80   70
## ################  day_of_week  ##################
##  fri  mon  thu  tue  wed
##  957 1058 1005  952 1018
## [1] "--- --- --- ---"
##  fri  mon  thu  tue  wed
##  957 1058 1005  952 1018
## ################  poutcome  ##################
##     failure nonexistent     success
##         545        4289         156
## [1] "--- --- --- ---"
##     failure nonexistent     success
##         545        4289         156

no_imputadas<-c("poutcome","day_of_week","month","contact","default")
df[,setdiff(vars_cat,no_imputadas)]<-aux2$completeObs[,setdiff(vars_cat,n
o_imputadas)]
```

Creación de nuevos niveles de los factores

Agrupamos subcategorías en menos categorías. El resumen anterior de las variables categóricas nos sirve como referencia para ver como reagruparlas. En jobs realizamos la agrupación en función del posible ingreso monetario. Finalmente vemos la reagrupación final la cual no ha quedado uniformemente distribuida, sin embargo los grupos tienen una relación más significativa.

```
# Job

table(df$job)

##
##        admin.   blue-collar  entrepreneur     housemaid    management
##          1296          1156           189           119           351
##       retired self-employed      services       student    technician
##           222           166           451           109           817
##    unemployed
##           114

df$f.job <- 4
# 1 level - Admin-Managment
aux<-which(df$job %in% c("admin.", "management"))
df$f.job[aux] <-1

# 2 level - Entrep-Retired-selfEmpl
aux<-which(df$job %in% c("entrepreneur", "retired", "self-employed"))
df$f.job[aux] <-2

# 3 level - Not working
aux<-which(df$job %in% c("housemaid","unemployed","student"))
df$f.job[aux] <-3
```

```
# 4 level - Serv-Tech-BlueC
aux<-which(df$job %in% c("services","technician","blue-collar"))
df$f.job[aux] <-4

df$f.job<-factor(df$f.job,levels=1:4,labels=c("Admin-Managment", "Entrep-
Retired-selfEmpl", "Not-working", "Serv-Tech-BlueC"))
levels(df$f.job)<-paste0("f.job.",levels(df$f.job))
summary(df$f.job)

##         f.job.Admin-Managment f.job.Entrep-Retired-selfEmpl
##                          1647                           577
##             f.job.Not-working         f.job.Serv-Tech-BlueC
##                           342                          2424
```

En months realizamos la agrupación en función de las temporadas aunque no tan estrictamente.

```
# Months to groups
table(df$month)

##
##  apr  aug  dec  jul  jun  mar  may  nov  oct  sep
##  321  764   18  865  616   63 1680  513   80   70

df$f.season <- 3
# 1 level - mar-may
aux<-which(df$month %in% c("mar","apr","may"))
df$f.season[aux] <-1

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

# 3 level - aug-feb
aux<-which(df$month %in% c("dec","sep","oct","nov"))
df$f.season[aux] <-3

summary(df$f.season)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.723   2.000   3.000

df$f.season<-factor(df$f.season,levels=1:3,labels=c("Mar-May","Jun-Aug","
Sep-Dec"))
levels(df$f.season)<-paste0("f.season.",levels(df$f.season)) # Hacemos la
s etiquetas más informativas
summary(df$f.season)

## f.season.Mar-May f.season.Jun-Aug f.season.Sep-Dec
##             2064             2245              681
```

En Education realizamos la agrupación en función del nivel de estudios de cada individuo. Hemos puesto la categoría illiterate dentro de la que consideramos que el nivel de estudios es inferior. Al realizar la agrupación los niveles quedaron relativamente bien equilibrados.

```
#Education
table(df$education)

##
##            basic.4y            basic.6y            basic.9y
##                 515                 271                 810
##         high.school          illiterate professional.course
##                1196                   1                 633
##   university.degree
##                1564

df$f.education <- 3
# 1 level - Basic
aux<-which(df$education %in% c("illiterate","basic.4y","basic.6y","basic.
9y"))
df$f.education[aux] <-1

# 2 level - High School
aux<-which(df$education %in% c("high.school"))
df$f.education[aux] <-2

# 3 level - Professional
aux<-which(df$education %in% c("professional.course","university.degree")
)
df$f.education[aux] <-3

df$f.education<-factor(df$f.education,levels=1:3,labels=c("Basic","High S
chool","Professional"))
table(df$f.education);

##
##        Basic  High School Professional
##         1597         1196         2197
```

Tratamiento de variables no-Target Numéricas

Age

Consideramos que no presenta ningún outlier, ya que las edades comprendidas entre 18 y 92 años, son considerados normal.

Campaing

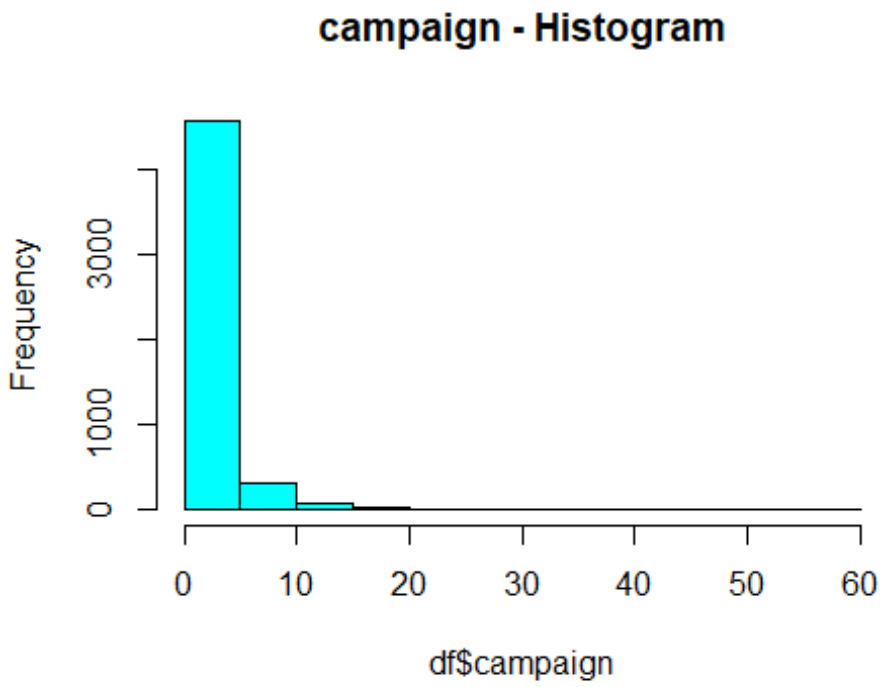
Para sopesar los outliers consideramos que en los 10 meses que dura la campaña, un máximo de 20 contactos es aceptable puesto que eso implica una media de un

contacto cada 15 días. Como errores se han buscado aquellos valores menores a 1 ya que se incluye la presente campaña. No se han detectado errores.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   1.000   2.000   2.575   3.000  56.000
```

```
  hist(df$campaign,col="cyan",main="campaign - Histogram")
```




```
  par(mfrow=c(1,2))
  boxplot(df$campaign, labels=row.names(df))
  aux<-calcQ(df$campaign);
  abline(h=aux[8],col="red",lwd=2)
  abline(h=aux[9],col="magenta",lwd=2)
  aux<-which(df$campaign<1);aux  # Si se incluye el último contacto, este valor no puede ser 0
```

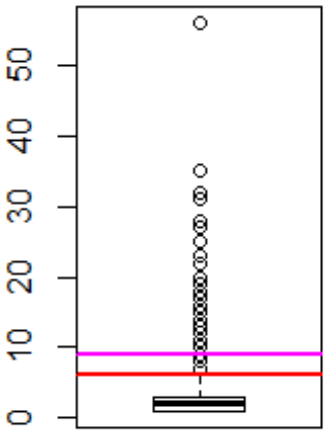
```
## integer(0)
```

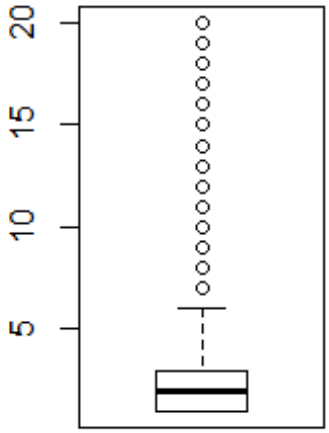
```
  aux<-which(df$campaign>20);length(aux);df[aux,'campaign']
```

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

```
##  [1] 23 25 56 32 35 31 28 27 22 28 25
```

```
  df[aux,"campaign"]<-NA
  boxplot(df$campaign)
```






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

  # Para el data analisis guardamos los missings
  dqr[dqr$variable=='campaing','missings']<-sum(is.na(df[,"campaign"]))
  # Para los individuales
  dqri[aux,'missings']<-dqri[aux,'missings']+1
```

Verificación de inconsistencias en pdays/previous/poutcome

Para pdays/previous/poutcome debería existir la relación directa entre previous=0, outcome=nonexistent y pdays=999 por lo que podemos detectar errores. Al ver el resultado podemos decir que hay inconsistencias entre el pdays y previous, ya que todos los que son pdays = 999, deberían ser previous = nonexistent, lo que en este caso nos dan 526 individuos que no cumplen esta condición. Como suponen más de un 10% de la muestra y nuestro trabajo no es exhaustivo vamos a ignorarlo.

```
  rel_pdays<-which(df$pdays==999)
  rel_previous<-which(df$previous==0)
  rel_poutcome<-which(df$poutcome=='nonexistent')
  length(setdiff(rel_poutcome, rel_previous))
```

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

```
  length(setdiff(rel_previous, rel_poutcome))
```

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

```
  length(setdiff(rel_previous, rel_pdays))
```

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

```
  length(setdiff(rel_pdays,    rel_previous))
```

```
## [1] 526
```

```
  summary(df[setdiff(rel_pdays,rel_previous),c('previous','poutcome')]) #
Miramos el perfil de esos individuos
```

```
##     previous            poutcome
##  Min.   :1.000   failure    :526
##  1st Qu.:1.000   nonexistent:  0
##  Median :1.000   success    :  0
##  Mean   :1.118
##  3rd Qu.:1.000
##  Max.   :5.000
```

Pdays

Con el summary podemos ver que no tenemos outliers ni errores, tampoco missings.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   999.0   999.0   964.2   999.0   999.0
```

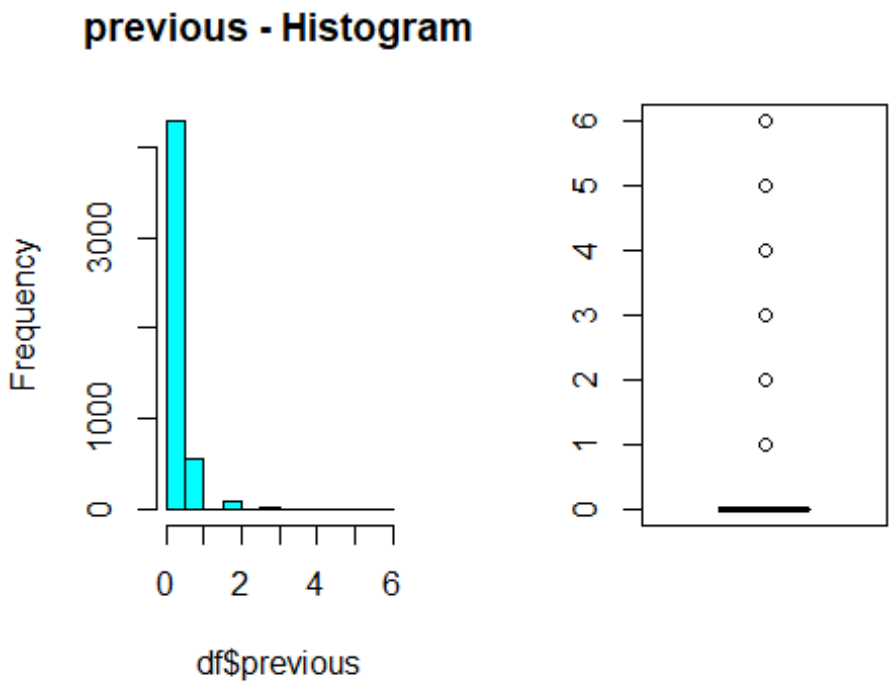
Previous

Consideramos que para esta variable no hay outliers, ya que por los valores se ve que pueden haber sido contactado hasta en 6 campañas previas, lo que tiene sentido.

```
  summary(df$previous)#Vemos que gran parte de los valores es 0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1786  0.0000  6.0000
```

```
  par(mfrow=c(1,2))
  hist(df$previous,col="cyan",main="previous - Histogram")
  boxplot(df$previous, labels=row.names(df))
```




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

Comprobación de inconsistencias en los índices trimestrales/mensuales

Para los índices trimestrales/mensuales (emp.var.rate/nr.employed/cons.prize.idx/ cons.conf.idx) cabe esperar que tengan los mismos valores para cada mes, de lo contrario pueden considerarse errores. Aparecen muchas discordancias, ya que para cada individuo y para un mismo mes el valor debería ser el mismo y en este caso no lo son. Nuestro trabajo no es exhaustivo, así que vamos a ignorar esta inconsistencia. A continuación se muestra, para cada variable y para cada mes el número de niveles, que en el caso ideal debería haber un solo nivel.

```
  aux<-c('emp.var.rate','nr.employed','cons.price.idx','cons.conf.idx')
  for(i in aux){
    cat("################ ",i," ##################\n")
    for(j in levels(df$month)){
      #cat("-- ",j,"--\n")
      aux2<-unique(df[which(df$month==j),i])
      cat(j,": ",aux2,"\n")
    }
  }
```

```
## ################  emp.var.rate  ##################
## apr :  -1.8
## aug :  1.4 -2.9 -1.7
## dec :  -0.2 -3
## jul :  1.4 -2.9 -1.7
```

```
## jun :  1.4 -2.9 -1.7
## mar :  -1.8
## may :  1.1 -1.8
## nov :  -0.1 -3.4 -1.1
## oct :  -0.1 -3.4 -1.1
## sep :  -3.4 -1.1
## ################  nr.employed  ##################
## apr :  5099.1 5008.7
## aug :  5228.1 5076.2 4991.6
## dec :  5176.3 5023.5
## jul :  5228.1 5076.2 4991.6
## jun :  5228.1 5076.2 4991.6
## mar :  5099.1 5008.7
## may :  5191 5099.1 5008.7
## nov :  5195.8 5017.5 4963.6
## oct :  5195.8 5017.5 4963.6
## sep :  5017.5 4963.6
## ################  cons.price.idx  ##################
## apr :  93.075 93.749
## aug :  93.444 92.201 94.027
## dec :  92.756 92.713
## jul :  93.918 92.469 94.215
## jun :  94.465 92.963 94.055
## mar :  92.843 93.369
## may :  93.994 92.893 93.876
## nov :  93.2 92.649 94.767
## oct :  93.798 92.431 94.601
## sep :  92.379 94.199
## ################  cons.conf.idx  ##################
## apr :  -47.1 -34.6
## aug :  -36.1 -31.4 -38.3
## dec :  -45.9 -33
## jul :  -42.7 -33.6 -40.3
## jun :  -41.8 -40.8 -39.8
## mar :  -50 -34.8
## may :  -36.4 -46.2 -40
## nov :  -42 -30.1 -50.8
## oct :  -40.4 -26.9 -49.5
## sep :  -29.8 -37.5
```

Emp.var.rate,cons.price.idx, cons.conf.idx, euribor3m, nr.employed

Necesitamos saber cómo se han obtenido estos datos para poder validarlos, como no tenemos esa información solo podemos comprobar los missings values. En este caso al hacer summary de cada variable, podemos ver que no existen missings.

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

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.40000 -1.80000  1.10000  0.05212  1.40000  1.40000
```

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   92.20   93.08   93.44   93.56   93.99   94.77

  summary(df$cons.conf.idx)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -50.80  -42.70  -41.80  -40.54  -36.40  -26.90

  summary(df$euribor3m)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.634   1.334   4.857   3.585   4.961   5.045

  summary(df$nr.employed)

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

Realizamos la imputación de las variables numéricas y comparamos los datos imputados con los originales. Observamos que da valores razonados, solamente que debemos redondearlos en ambos casos ya que se trata de "número de contactos" de las variables 'previous' y 'campaign'. Igual que en el caso anterior, solo se sobrescriben las variables imputadas en nuestro df.

```
  vars_num_imp<-imputePCA(df[,vars_num],npc=5)

  summary(df[,vars_num])

##       age           campaign          pdays          previous
##  Min.   :18.00   Min.   : 1.000   Min.   :  0.0   Min.   :0.0000
##  1st Qu.:32.00   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000
##  Median :38.00   Median : 2.000   Median :999.0   Median :0.0000
##  Mean   :40.18   Mean   : 2.514   Mean   :964.2   Mean   :0.1786
##  3rd Qu.:47.00   3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :92.00   Max.   :20.000   Max.   :999.0   Max.   :6.0000
##                  NA's   :11
##   emp.var.rate      cons.price.idx  cons.conf.idx      euribor3m
##  Min.   :-3.40000   Min.   :92.20   Min.   :-50.80   Min.   :0.634
##  1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.334
##  Median : 1.10000   Median :93.44   Median :-41.80   Median :4.857
##  Mean   : 0.05212   Mean   :93.56   Mean   :-40.54   Mean   :3.585
##  3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961
##  Max.   : 1.40000   Max.   :94.77   Max.   :-26.90   Max.   :5.045
##
##   nr.employed
##  Min.   :4964
##  1st Qu.:5099
##  Median :5191
##  Mean   :5166
##  3rd Qu.:5228
```

```
##  Max.   :5228
##
```

```
  summary(vars_num_imp$completeObs)
```

```
##       age           campaign          pdays          previous
##  Min.   :18.00   Min.   : 1.000   Min.   :  0.0   Min.   :0.0000
##  1st Qu.:32.00   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000
##  Median :38.00   Median : 2.000   Median :999.0   Median :0.0000
##  Mean   :40.18   Mean   : 2.515   Mean   :964.2   Mean   :0.1786
##  3rd Qu.:47.00   3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :92.00   Max.   :20.000   Max.   :999.0   Max.   :6.0000
##   emp.var.rate      cons.price.idx  cons.conf.idx      euribor3m
##  Min.   :-3.40000   Min.   :92.20   Min.   :-50.80   Min.   :0.634
##  1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.334
##  Median : 1.10000   Median :93.44   Median :-41.80   Median :4.857
##  Mean   : 0.05212   Mean   :93.56   Mean   :-40.54   Mean   :3.585
##  3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961
##  Max.   : 1.40000   Max.   :94.77   Max.   :-26.90   Max.   :5.045
##   nr.employed
##  Min.   :4964
##  1st Qu.:5099
##  Median :5191
##  Mean   :5166
##  3rd Qu.:5228
##  Max.   :5228
```

```
  df[,vars_num]<-vars_num_imp$completeObs[,vars_num]
  aux<-c('previous','campaign')
  df[,aux]<-round(df[,aux])
```

Resumen del Data Quality Report y Ranking

A continuación se muestra el ranking de missings, errors y outliers para cada variable que tiene por lo menos algún missing, error o outlier. Vemos que el valor más destacable, los missings de education, no alcanza el 5% de la muestra.

```
aux<-which(dqr$missings>0 | dqr$errors>0 | dqr$outliers>0)
dqr_subset<-dqr[aux,]
dqr_subset[order(-dqr_subset$missings),]
```

```
##     variable missings errors outliers
## 4  education      241      0        0
## 6    housing      113      0        0
## 7       loan      113      0        0
## 2        job       45      0        0
## 3    marital        8      0        0
## 11  duration        0      4        6
```

```
dqr[dqr$variable=="education",'missings']/nrow(df)
```

```
## [1] 0.04829659
```

Para el data quality report de individuales cabe destacar que se han ignorado errores y ouliers de la variable target duration, pues estos individuos se han eliminado resultando una muestra de 4990. Dicho esto, y viendo los resultados anteriores, bastará con supervisar los missings individuales. El summary revela poca incidencia con un escaso 0.1 missings de media, pero sí vemos que hay individuos con hasta 3 missings. Con prop.table se observa un 5% de la muestra con 1 missing, un 2,5% con dos y un 0,24% con tres. Lo consideramos valores razonables.

```
summary(dqri$missings)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1064  0.0000  3.0000

prop.table(table(dqri$missings))

##
##          0          1          2          3
## 0.92244489 0.05110220 0.02404810 0.00240481
```

Creación de factores adicionales para cada variable cuantitativa

Age

Primero miramos cuan distribuidos quedan aplicando unos cortes según los cuartiles, como estos no difieren demasiado con los niveles naturales (20 añeros, 30 añeros...) preferimos quedarnos con los niveles naturales.

```
aux<-quantile(df$age,seq(0,1,0.25),na.rm=TRUE) # Niveles por quartiles
aux<-factor(cut(df$age,breaks=aux,include.lowest=T))
table(aux)

## aux
## [18,32] (32,38] (38,47] (47,92]
##    1328    1188    1287    1187

tapply(df$age,aux,median)

## [18,32] (32,38] (38,47] (47,92]
##      30      35      43      54

aux2<-c(18,30,40,50,92) # Niveles "naturales"
aux<-factor(cut(df$age,breaks=aux2,include.lowest=T))
table(aux)

## aux
## [18,30] (30,40] (40,50] (50,92]
##     870    1991    1253     876

tapply(df$age,aux,median)
```

```
## [18,30] (30,40] (40,50] (50,92]
##      28      35      45      56

df$f.age<-factor(cut(df$age,breaks=aux2,include.lowest=T))
levels(df$f.age)<-paste0("f.age-",levels(df$f.age))
summary(df$f.age)

## f.age-[18,30] f.age-(30,40] f.age-(40,50] f.age-(50,92]
##           870          1991          1253           876
```

Duration

Hemos buscado una distribución más o menos equilibrada y hemos conseguido separarlo en niveles de 2min, 3min, 5min, y el resto.

```
# Para duration
aux<-quantile(df$duration,seq(0,1,0.25),na.rm=TRUE)#Niveles por quartiles
aux<-factor(cut(df$duration,breaks=aux,include.lowest=T))
table(aux)

## aux
##        [5,103]      (103,178]      (178,321] (321,2.08e+03]
##           1255           1240           1249           1246

tapply(df$duration,aux,median)

##        [5,103]      (103,178]      (178,321] (321,2.08e+03]
##             68            140            240            488

aux2<-c(5,120,180,300,2100)#Niveles "naturales"
aux<-factor(cut(df$duration,breaks=aux2,include.lowest=T))
table(aux)

## aux
##       [5,120]     (120,180]     (180,300] (300,2.1e+03]
##          1557           966          1090          1377

tapply(df$duration,aux,median)

##       [5,120]     (120,180]     (180,300] (300,2.1e+03]
##            77           150           235           461

df$f.duration<-factor(cut(df$duration,breaks=aux2,include.lowest=T))#Nos
quedamos con los niveles naturales
levels(df$f.duration)<-paste0("f.duration-",levels(df$f.duration))#Hacemo
s las etiquetas más informativas
summary(df$f.duration)

##       f.duration-[5,120]     f.duration-(120,180]     f.duration-(180,
300]
##                     1557                      966
1090
```

```
## f.duration-(300,2.1e+03]
##                     1377
```

Campaign

Como para esta variable la mayoría de los valores están entre 0 y 1, no se puede hacer la separación por cuartiles. Hemos realizado una factorización manual viendo la cantidad de valores en cada nivel.

```
aux<-levels(factor(df$campaign))
aux<-factor(cut(df$campaign,breaks=aux,include.lowest=T))
table(aux)

## aux
##   [1,2]   (2,3]   (3,4]   (4,5]   (5,6]   (6,7]   (7,8]   (8,9]  (9,10
]
##    3380     676     334     190     117      86      60      31      2
3
## (10,11] (11,12] (12,13] (13,14] (14,15] (15,16] (16,17] (17,18] (18,19
]
##      17      21       9      11       8       6       8       7
4
## (19,20]
##       2

tapply(df$campaign,aux,median)

##   [1,2]   (2,3]   (3,4]   (4,5]   (5,6]   (6,7]   (7,8]   (8,9]  (9,10
]
##       1       3       4       5       6       7       8       9      1
0
## (10,11] (11,12] (12,13] (13,14] (14,15] (15,16] (16,17] (17,18] (18,19
]
##      11      12      13      14      15      16      17      18      1
9
## (19,20]
##      20

aux2<-c(0,1,2,20)
aux<-factor(cut(df$campaign,breaks=aux2,include.lowest=T))
table(aux)

## aux
##  [0,1]  (1,2] (2,20]
##   2121   1259   1610

df$f.campaign<-factor(cut(df$campaign,breaks=aux2,include.lowest=T))
levels(df$f.campaign)<-paste0("f.campaign-",levels(df$f.campaign))
summary(df$f.campaign)

##  f.campaign-[0,1]  f.campaign-(1,2] f.campaign-(2,20]
##              2121              1259              1610
```

Pdays

Como en pdays hay 4815 valores de 999 que significa que no se han contactado en campañas previas, esto sería un 96% de los individuos por lo que decidimos realizar la agrupación en solo dos niveles, contactados y no-contactados.

```
aux2<-c(0,998,999)
pdays_cutted<-factor(cut(df$pdays,breaks=aux2,include.lowest=T))
table(pdays_cutted)

## pdays_cutted
##   [0,998] (998,999]
##       175      4815

tapply(df$pdays,pdays_cutted,median)

##   [0,998] (998,999]
##         6       999

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

##   f.pdays-[0,998] f.pdays-(998,999]
##               175              4815
```

Previous

Vemos que esta variable solo tiene 6 niveles por lo decidimos pasarlos a los tres niveles más relevantes, sin que sea binaria. Ya que pensamos que el grupo de individuos con un solo contacto en una campaña previa podría ser significativo con respecto a la variable target Y.

```
aux2<-c(0,0.9,1,6)
previous_cutted<-factor(cut(df$previous,breaks=aux2,include.lowest=T))
table(previous_cutted)

## previous_cutted
## [0,0.9] (0.9,1]   (1,6]
##    4289     564     137

tapply(df$previous,previous_cutted,median)

## [0,0.9] (0.9,1]   (1,6]
##       0       1       2

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

## f.previous-[0,0.9] f.previous-(0.9,1]   f.previous-(1,6]
##               4289                564                137
```

Profiling

Nombres de niveles más informativos

Para poder hacer profiling, necesitamos darle nombres a los subniveles de los factores, para esto hacemos un bucle que recorre cada variable categórica y le añade el nombre de la variable más un "." y el nombre del nivel. Luego procedemos a ejecutar la función condes con la variable target duration, la cual se encuentra en la posición 11 de nuestro data frame. Usamos una probabilidad de 0.01 que consideramos puede mostrarnos el resultado que queremos. Para la función catdes usamos la variable "Y" la cual se encuentra en la posición 21 de nuestro data frame.

```
vars_cat_con_y<-c(vars_cat,"y")
for (i in vars_cat_con_y){
  levels(df[,i])<-paste0(i,".",levels(df[,i]))
}
```

Resultado del CONDES

```
condes(df,11,proba=0.01)
```

```
## $quanti
##           correlation      p.value
## campaign -0.05940135 2.683764e-05
##
## $quali
##                       R2       p.value
## f.duration 0.621168787  0.000000e+00
## y          0.177066645 2.228224e-213
## f.campaign 0.003783221  7.858324e-05
## month      0.004450289  8.185248e-03
##
## $category
##                               Estimate       p.value
## f.duration-(300,2.1e+03]  310.35106  0.000000e+00
## y.yes                     170.13318 2.228224e-213
## f.campaign-(1,2]           23.01041  3.895001e-05
## month.apr                  35.25783  4.865526e-03
## f.season.Mar-May           13.19170  6.782891e-03
## month.aug                 -25.22225  7.943838e-03
## f.campaign-(2,20]         -17.35164  3.316706e-03
## f.duration-(180,300]      -20.50721  3.927333e-04
## f.duration-(120,180]     -106.75355  5.404997e-53
## y.no                     -170.13318 2.228224e-213
## f.duration-[5,120]       -183.09030 1.278559e-312
```

```
tapply(df$duration,df$f.dur,mean)
```

```
##       f.duration-[5,120]     f.duration-(120,180]     f.duration-(180,
300]
##                 73.39306                149.72981                235.9
```

```
7615
## f.duration-(300,2.1e+03]
##                 566.83442

summary(df$duration)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     5.0   103.0   178.5   259.9   321.0  2078.0

tapply(df$duration,df$y,mean)

##     y.no    y.yes
## 222.8923 563.1587
```

En el resultado de la correlación cuantitativa, podemos ver que la única variable posiblemente relacionada es campaing. Campaing aun estando inversamente relacionada con duration, su correlación es muy pequeña pues no llega al 6%. Traducido al lenguaje natural podemos decir: "hay indicios de que cuantas más campañas ha participado el individuo más corta será la duración de la llamada". Además el pvalor nos indica que la probabilidad de que la correlación sea cero, es muy baja, tanto es así que nos da cierta confianza de que la correlación indicada es la real.

Para las variables cualitativas, podemos ver que los factores de duration están muy relacionados lo cual tiene total sentido ya que se está comparando con ella misma. Para la variable Y, podemos ver que hay una relación con duration aunque 0.177 comparado con 1 es aparentemente poco, en este tipo de estudios es una relación relevante que cabe destacar. Además el pvalor es casi nulo, que nos da mucha confianza sobre este indicador. Para f.campaing y month, presentan ciertos indicios de relación pero con pvalores bastante ajustados.

Mirando el análisis por categorías que nos muestra condes, vemos en primer lugar que el f.duration con intervalo entre (300,2.1e+03] tiene una media estimada de 310 segundos sobre la media global lo cual no deja de ser una obviedad. Sin embargo, si nos fijamos en el y.yes podemos ver que los individuos están 170 segundos por encima de la media global, cosa que viene apoyada por la confianza de un pvalor casi nulo. Con esto podemos decir que los individuos propensos a comprar el producto, resulta que duran más tiempo al teléfono. Sin más información sobre el proceso de contacto en las campañas, nos hace pensar que puede ser por el hecho de que al comprar el producto, estos individuos deben permanecer más tiempo para poder dar todos sus datos.

Si comparamos los meses de abril y agosto podemos ver que en abril, los individuos duran un poco más de tiempo al teléfono respecto a la media, y esto, asumiendo lo anteriormente dicho, puede que sea un mes más propenso a la venta del producto. En cambio en el mes de agosto estos duraron menos tiempo al teléfono, podemos intuir que puede ser debido a las vacaciones.

Resultado del CATDES

```
prop.table(table(df$y)) # y
```

```
##
##      y.no     y.yes
## 0.8913828 0.1086172
```

```
prop.table(table(df$f.duration)) # f.duration
```

```
##
##       f.duration-[5,120]     f.duration-(120,180]     f.duration-(180,
300]
##                0.3120240                0.1935872                0.218
4369
## f.duration-(300,2.1e+03]
##                0.2759519
```

```
prop.table(table(df$f.duration,df$y),1)
```

```
##
##                                    y.no      y.yes
##   f.duration-[5,120]       0.98715478 0.01284522
##   f.duration-(120,180]     0.95445135 0.04554865
##   f.duration-(180,300]     0.90000000 0.10000000
##   f.duration-(300,2.1e+03] 0.73202614 0.26797386
```

```
prop.table(table(df$f.duration,df$y),2)
```

```
##
##                                    y.no      y.yes
##   f.duration-[5,120]       0.34554856 0.03690037
##   f.duration-(120,180]     0.20728417 0.08118081
##   f.duration-(180,300]     0.22054856 0.20110701
##   f.duration-(300,2.1e+03] 0.22661871 0.68081181
```

```
catdes(df,21,proba=0.01)
```

```
##
## Link between the cluster variable and the categorical variables (chi-s
quare test)
## ================================================================
===========
##                    p.value df
## f.duration  1.038223e-118  3
## poutcome    5.738265e-111  2
## f.pdays     9.773367e-110  1
## month        7.431682e-53  9
## f.previous   4.536325e-49  2
## job          1.524734e-25 10
## contact      1.007104e-18  1
## f.age        1.378066e-12  3
## default      4.342743e-12  1
## f.job        8.884797e-12  3
## f.season     4.127488e-08  2
## f.campaign   1.868723e-06  2
```

```
## f.education  7.638741e-05  2
## education    5.054754e-04  6
## marital      1.381426e-03  2
##
## Description of each cluster by the categories
## =============================================
## $y.no
##                                          Cla/Mod    Mod/Cla    Global
## f.duration=f.duration-[5,120]         98.71548 34.5548561 31.202405
## f.pdays=f.pdays-(998,999]             91.00727 98.5161871 96.492986
## f.previous=f.previous-[0,0.9]         91.11681 87.8597122 85.951904
## poutcome=poutcome.nonexistent         91.11681 87.8597122 85.951904
## contact=contact.telephone             94.34807 37.9046763 35.811623
## f.duration=f.duration-(120,180]       95.44513 20.7284173 19.358717
## default=default.unknown               95.04762 22.4370504 21.042084
## f.job=f.job.Serv-Tech-BlueC           92.07921 50.1798561 48.577154
## job=job.blue-collar                   93.85813 24.3929856 23.166333
## month=month.may                       92.55952 34.9595324 33.667335
## f.campaign=f.campaign-(2,20]          92.17391 33.3633094 32.264529
## f.age=f.age-(40,50]                   92.41820 26.0341727 25.110220
## education=education.basic.9y          92.71605 16.8839928 16.232465
## f.education=Basic                     91.42142 32.8237410 32.004008
## job=job.services                      93.56984  9.4874101  9.038076
## f.age=f.age-(30,40]                   90.85886 40.6699640 39.899800
## f.season=f.season.Jun-Aug             90.69042 45.7733813 44.989980
## marital=marital.married               90.22029 60.7688849 60.040080
## f.age=f.age-[18,30]                   86.43678 16.9064748 17.434870
## marital=marital.single                86.60524 27.4730216 28.276553
## education=education.university.degree 86.70077 30.4856115 31.342685
## f.education=Professional              87.07328 43.0080935 44.028056
## f.campaign=f.campaign-[0,1]           86.89298 41.4343525 42.505010
## month=month.apr                       80.37383  5.8003597  6.432866
## f.previous=f.previous-(0.9,1]         82.80142 10.4991007 11.302605
## f.job=f.job.Entrep-Retired-selfEmpl   82.84229 10.7464029 11.563126
## job=job.student                       72.47706  1.7760791  2.184369
## f.season=f.season.Sep-Dec             82.81938 12.6798561 13.647295
## month=month.sep                       62.85714  0.9892086  1.402806
## f.age=f.age-(50,92]                   83.21918 16.3893885 17.555110
## month=month.mar                       55.55556  0.7868705  1.262525
## month=month.oct                       57.50000  1.0341727  1.603206
## job=job.retired                       71.62162  3.5746403  4.448898
## default=default.no                    87.56345 77.5629496 78.957916
## contact=contact.cellular              86.23166 62.0953237 64.188377
## f.previous=f.previous-(1,6]           53.28467  1.6411871  2.745491
## f.pdays=f.pdays-[0,998]               37.71429  1.4838129  3.507014
## poutcome=poutcome.success             33.97436  1.1915468  3.126253
## f.duration=f.duration-(300,2.1e+03]   73.20261 22.6618705 27.595190
##                                            p.value     v.test
## f.duration=f.duration-[5,120]         2.229911e-64  16.941339
## f.pdays=f.pdays-(998,999]             7.719766e-64  16.868133
```

```
## f.previous=f.previous-[0,0.9]          6.319513e-24  10.086802
## poutcome=poutcome.nonexistent          6.319513e-24  10.086802
## contact=contact.telephone              2.596533e-20   9.234435
## f.duration=f.duration-(120,180]        2.258505e-14   7.634983
## default=default.unknown                8.276549e-14   7.465847
## f.job=f.job.Serv-Tech-BlueC            6.563448e-11   6.530308
## job=job.blue-collar                    5.519353e-10   6.203578
## month=month.may                        1.337581e-08   5.681193
## f.campaign=f.campaign-(2,20]           1.119511e-06   4.869376
## f.age=f.age-(40,50]                    8.643837e-06   4.448584
## education=education.basic.9y           1.987893e-04   3.720550
## f.education=Basic                      3.022238e-04   3.613386
## job=job.services                       8.082497e-04   3.349954
## f.age=f.age-(30,40]                    1.346805e-03   3.205815
## f.season=f.season.Jun-Aug              1.383606e-03   3.198049
## marital=marital.married                2.776091e-03   2.991502
## f.age=f.age-[18,30]                    5.943641e-03  -2.750874
## marital=marital.single                 3.906851e-04  -3.546297
## education=education.university.degree 2.321260e-04  -3.681214
## f.education=Professional               3.492236e-05  -4.138737
## f.campaign=f.campaign-[0,1]            1.327080e-05  -4.355592
## month=month.apr                        1.663724e-06  -4.790493
## f.previous=f.previous-(0.9,1]          1.344587e-06  -4.833047
## f.job=f.job.Entrep-Retired-selfEmpl    1.134068e-06  -4.866823
## job=job.student                        1.088887e-06  -4.874854
## f.season=f.season.Sep-Dec              7.466144e-08  -5.379576
## month=month.sep                        6.365587e-09  -5.806859
## f.age=f.age-(50,92]                    3.842509e-09  -5.890834
## month=month.mar                        9.598941e-12  -6.812392
## month=month.oct                        2.600541e-13  -7.313613
## job=job.retired                        1.403012e-13  -7.396044
## default=default.no                     8.276549e-14  -7.465847
## contact=contact.cellular               2.596533e-20  -9.234435
## f.previous=f.previous-(1,6]            3.828645e-27 -10.790222
## f.pdays=f.pdays-[0,998]                7.719766e-64 -16.868133
## poutcome=poutcome.success              5.851631e-64 -16.884494
## f.duration=f.duration-(300,2.1e+03]    2.022672e-97 -20.946423
##
## $y.yes
##                                         Cla/Mod   Mod/Cla    Global
## f.duration=f.duration-(300,2.1e+03]   26.797386 68.081181 27.595190
## poutcome=poutcome.success             66.025641 19.003690  3.126253
## f.pdays=f.pdays-[0,998]               62.285714 20.110701  3.507014
## f.previous=f.previous-(1,6]           46.715328 11.808118  2.745491
## contact=contact.cellular              13.768342 81.365314 64.188377
## default=default.no                    12.436548 90.405904 78.957916
## job=job.retired                       28.378378 11.623616  4.448898
## month=month.oct                       42.500000  6.273063  1.603206
## month=month.mar                       44.444444  5.166052  1.262525
## f.age=f.age-(50,92]                   16.780822 27.121771 17.555110
```

```
## month=month.sep                            37.142857  4.797048  1.402806
## f.season=f.season.Sep-Dec                  17.180617 21.586716 13.647295
## job=job.student                            27.522936  5.535055  2.184369
## f.job=f.job.Entrep-Retired-selfEmpl        17.157712 18.265683 11.563126
## f.previous=f.previous-(0.9,1]              17.198582 17.896679 11.302605
## month=month.apr                            19.626168 11.623616  6.432866
## f.campaign=f.campaign-[0,1]                13.107025 51.291513 42.505010
## f.education=Professional                   12.926718 52.398524 44.028056
## education=education.university.degree      13.299233 38.376384 31.342685
## marital=marital.single                     13.394755 34.870849 28.276553
## f.age=f.age-[18,30]                        13.563218 21.771218 17.434870
## marital=marital.married                     9.779706 54.059041 60.040080
## f.season=f.season.Jun-Aug                   9.309577 38.560886 44.989980
## f.age=f.age-(30,40]                         9.141135 33.579336 39.899800
## job=job.services                            6.430155  5.350554  9.038076
## f.education=Basic                           8.578585 25.276753 32.004008
## education=education.basic.9y                7.283951 10.885609 16.232465
## f.age=f.age-(40,50]                         7.581804 17.527675 25.110220
## f.campaign=f.campaign-(2,20]                7.826087 23.247232 32.264529
## month=month.may                             7.440476 23.062731 33.667335
## job=job.blue-collar                         6.141869 13.099631 23.166333
## f.job=f.job.Serv-Tech-BlueC                 7.920792 35.424354 48.577154
## default=default.unknown                     4.952381  9.594096 21.042084
## f.duration=f.duration-(120,180]             4.554865  8.118081 19.358717
## contact=contact.telephone                   5.651931 18.634686 35.811623
## f.previous=f.previous-[0,0.9]               8.883190 70.295203 85.951904
## poutcome=poutcome.nonexistent               8.883190 70.295203 85.951904
## f.pdays=f.pdays-(998,999]                   8.992731 79.889299 96.492986
## f.duration=f.duration-[5,120]               1.284522  3.690037 31.202405
##                                                 p.value     v.test
## f.duration=f.duration-(300,2.1e+03]   2.022672e-97  20.946423
## poutcome=poutcome.success             5.851631e-64  16.884494
## f.pdays=f.pdays-[0,998]               7.719766e-64  16.868133
## f.previous=f.previous-(1,6]           3.828645e-27  10.790222
## contact=contact.cellular              2.596533e-20   9.234435
## default=default.no                    8.276549e-14   7.465847
## job=job.retired                       1.403012e-13   7.396044
## month=month.oct                       2.600541e-13   7.313613
## month=month.mar                       9.598941e-12   6.812392
## f.age=f.age-(50,92]                   3.842509e-09   5.890834
## month=month.sep                       6.365587e-09   5.806859
## f.season=f.season.Sep-Dec             7.466144e-08   5.379576
## job=job.student                       1.088887e-06   4.874854
## f.job=f.job.Entrep-Retired-selfEmpl   1.134068e-06   4.866823
## f.previous=f.previous-(0.9,1]         1.344587e-06   4.833047
## month=month.apr                       1.663724e-06   4.790493
## f.campaign=f.campaign-[0,1]           1.327080e-05   4.355592
## f.education=Professional              3.492236e-05   4.138737
## education=education.university.degree 2.321260e-04   3.681214
## marital=marital.single                3.906851e-04   3.546297
```

```
## f.age=f.age-[18,30]                    5.943641e-03   2.750874
## marital=marital.married                2.776091e-03  -2.991502
## f.season=f.season.Jun-Aug              1.383606e-03  -3.198049
## f.age=f.age-(30,40]                    1.346805e-03  -3.205815
## job=job.services                       8.082497e-04  -3.349954
## f.education=Basic                      3.022238e-04  -3.613386
## education=education.basic.9y           1.987893e-04  -3.720550
## f.age=f.age-(40,50]                    8.643837e-06  -4.448584
## f.campaign=f.campaign-(2,20]           1.119511e-06  -4.869376
## month=month.may                        1.337581e-08  -5.681193
## job=job.blue-collar                    5.519353e-10  -6.203578
## f.job=f.job.Serv-Tech-BlueC            6.563448e-11  -6.530308
## default=default.unknown                8.276549e-14  -7.465847
## f.duration=f.duration-(120,180]        2.258505e-14  -7.634983
## contact=contact.telephone              2.596533e-20  -9.234435
## f.previous=f.previous-[0,0.9]          6.319513e-24 -10.086802
## poutcome=poutcome.nonexistent          6.319513e-24 -10.086802
## f.pdays=f.pdays-(998,999]              7.719766e-64 -16.868133
## f.duration=f.duration-[5,120]          2.229911e-64 -16.941339
##
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                         Eta2       P-value
## duration       0.177066645 2.228224e-213
## nr.employed    0.108627691 9.588810e-127
## pdays          0.099363145 1.586887e-115
## euribor3m      0.080172702  1.211844e-92
## emp.var.rate   0.074526086  5.345604e-86
## previous       0.045463793  2.111426e-52
## cons.price.idx 0.013909243  6.368783e-17
## campaign       0.006362358  1.679586e-08
## age            0.004721065  1.184599e-06
## cons.conf.idx  0.003722772  1.610540e-05
##
## Description of each cluster by quantitative variables
## ====================================================
## $y.no
##                      v.test Mean in category  Overall mean sd in categor
y
## nr.employed     23.279681     5174.2504272 5165.87569138     65.975653
2
## pdays           22.264832      984.2796763  964.18517034    119.945851
2
## euribor3m       19.999540        3.7572383    3.58457355      1.662919
6
## emp.var.rate    19.282392        0.2024505    0.05212425      1.500410
9
## cons.price.idx   8.330259       93.5875852   93.56373427      0.561839
8
```

```
## campaign         5.633986        2.5807104    2.51503006      2.429261
1
## cons.conf.idx   -4.309630      -40.6408273  -40.54192385      4.412224
6
## age             -4.853184       39.9267086   40.17755511      9.842648
1
## previous       -15.060507        0.1411871    0.17855711      0.412085
5
## duration       -29.721802      222.8923112  259.85110220    201.971895
2
##                 Overall sd       p.value
## nr.employed     72.7919889 7.122275e-120
## pdays          182.6196113 8.102628e-110
## euribor3m        1.7469207  5.558249e-89
## emp.var.rate     1.5774788  7.550292e-83
## cons.price.idx   0.5793439  8.066566e-17
## campaign         2.3588988  1.760909e-08
## cons.conf.idx    4.6436681  1.635282e-05
## age             10.4585324  1.214948e-06
## previous         0.5020810  2.945210e-51
## duration       251.6124483 4.014694e-194
##
## $y.yes
##                    v.test Mean in category  Overall mean sd in categor
y
## duration         29.721802      563.1586716  259.85110220    380.638506
0
## previous         15.060507        0.4852399    0.17855711      0.906497
6
## age               4.853184       42.2361624   40.17755511     14.395684
8
## cons.conf.idx     4.309630      -39.7302583  -40.54192385      6.166449
5
## campaign         -5.633986        1.9760148    2.51503006      1.572766
5
## cons.price.idx   -8.330259       93.3679982   93.56373427      0.675713
9
## emp.var.rate    -19.282392       -1.1815498    0.05212425      1.651576
2
## euribor3m       -19.999540        2.1675756    3.58457355      1.774772
0
## pdays           -22.264832      799.2767528  964.18517034    398.074416
1
## nr.employed     -23.279681     5097.1472325 5165.87569138     88.102466
2
##                 Overall sd       p.value
## duration       251.6124483 4.014694e-194
## previous         0.5020810  2.945210e-51
## age             10.4585324  1.214948e-06
## cons.conf.idx    4.6436681  1.635282e-05
```

```
## campaign         2.3588988  1.760909e-08
## cons.price.idx   0.5793439  8.066566e-17
## emp.var.rate     1.5774788  7.550292e-83
## euribor3m        1.7469207  5.558249e-89
## pdays          182.6196113 8.102628e-110
## nr.employed     72.7919889 7.122275e-120
```

En la descripción por categorías catdes nos da la relación que tiene cada categoría con nuestro target yes o no, de los cuales nos vamos a focalizar en los que respondieron yes.

Aquí de nuevo se corrobora lo que ya nos anticipaba el condes, ya que la categoría que contiene la mayor duración de tiempo de las llamadas, es la que esta más relacionada con que el individuo compre el producto.

Esto lo interpretamos de la columna Mod/Cla en la cual aquellos que compraron el producto, un 68% eran de las llamadas más prolongadas, sin embargo, y esto viene reflejado en la columna Cla/Mod, no podemos decir que todos los que duran un tiempo prolongado en el telefono, vayan a comprar el producto, pues solo un 26% de estos aceptaron el producto, que no es poco.

De la categoría de poutcome, podemos ver que aquellos que aceptaron en una campaña previa el producto, aceptarán con una probabilidad de un 66% el producto de esta campaña. Esto apoya la tesis que pregona el marketing: “Si el individuo ya es cliente de la empresa esto le da confianza para comprar de nuevo”.

En la misma línea nos indica la categoría f.pdays[0,988] que a fin de cuentas tiene el mismo significado que el poutcome y que previous, salvo como hemos visto en el anterior análisis hay ciertos individuos de pdays que no son consistentes con el poutcome.

Otro valor que nos llama la atención es el que da la categoría job, en su nivel retired, podemos ver que un 28% aceptó el producto, lo que es un buen indicador de que este es un tipo de individuo de interés.

En los meses de marzo y octubre, vemos un incremento relevante en las ventas, aunque vemos que estos meses son una muestra poco representativa de nuestra muestra (esto lo podemos ver en la columna global, donde estos meses tienen un valor inferior al 1.7% del total de individuos) lo que nos puede decir que no son valores muy representativos. En cambio para el mes de abril podemos ver que es una muestra mayor, con un 6% con respecto a la muestra global, de este porcentaje casi un 20% aceptó el producto, lo cual nos puede indicar, que sea un mes más propenso a la aceptación del mismo.

Además parece ser que la franja de edad más propensa a la compra corresponde al intervalo de más larga edad que es de mayores de 50 años.

Después de analizar estos datos, podemos crear algunos perfiles que pueden ser propensos a aceptar futuros productos.

Perfil de persona más propensa a que acepte el producto:

1- Persona entre 50 y 92 años, que esté retirada, que haya sido contactada en una campaña previa. 2- Persona mayor de 40 años, profesional, soltero, que haya sido contactada en una campaña previa.

Perfil de llamada más propensa a que se acepte el producto:

1- Abril, duración larga (más de 300 segundos) y hechas a un móvil.