deliverable-1

Guillem Valls, Sergio Mazzariol

Table of Contents

# Loading packages

# Loading data

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,]  
save.image("set-datos.RData")

# Loading data

Cargamos los datos guardados en el fichero set-Datos.RData. En este fichero hay la muestra de 5000 individuos que se usará para el análisis. Podemos comprobar que la dimensión de los datos es correcto, 5000 filas y 21 columnas.

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

# Inicializamos datos y funciones

#rm(list=ls())#Limpiamos el workspace  
#load("data-INI.RData", envir = parent.frame(), verbose = FALSE)#cargamos los datos  
  
#Comprobar el tamaño de la muestra  
dim(df)

## [1] 5000 21

#Creamos dataframe para almacenar missings, errors, outliers  
dqr <- data.frame(variable=character(), missings=integer(), errors=integer(), outliers=integer())  
dqr[length(names(df)),]<-NA  
dqr$variable <-names(df)  
  
#También para individuales  
dqri <- data.frame(missings=integer(), errors=integer(), outliers=integer())  
  
calcQ <- function(x) {# Declaramos una función para discriminar outliers leves y severos  
 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], q2=s.x[3],   
 q3=s.x[5], max=s.x[6], mouts=s.x[5]+1.5\*iqr, souts=s.x[5]+3\*iqr ) }  
  
#Separamos variables target/no-target-numéricas/no-target-categóricas (observamos un individuo cualquiera)  
  
df[1,]

## age job marital education default housing loan contact month  
## 20 39 management single basic.9y unknown no no telephone may  
## day\_of\_week duration campaign pdays previous poutcome emp.var.rate  
## 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"

#Guardamos datos parciales en este punto  
#save.image("data-step-1.RData")

# Step 1: Exploratory Data Analysis

#rm(list=ls())#Limpiamos el workspace  
#load("data-step-1.RData", envir = parent.frame(), verbose = FALSE)#cargamos los datos  
  
# Graphics  
names(df)

## [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"

# Overview of data frame   
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   
##

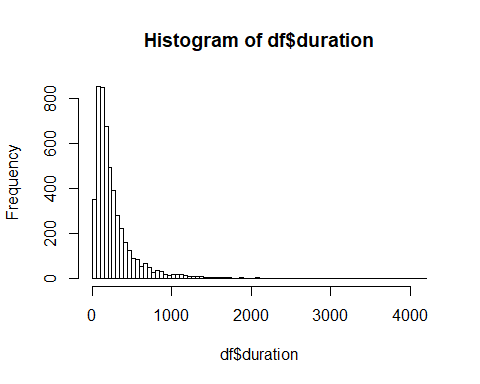
#Analizamos variables según si son cat/num o target/no-target (obtenemos 4 clasificaciones)  
  
#target+categorica (miramos qué niveles tiene)  
  
summary(df$y) #solo hay yes/no, por tanto no hay que hacer nada

## no yes   
## 4455 545

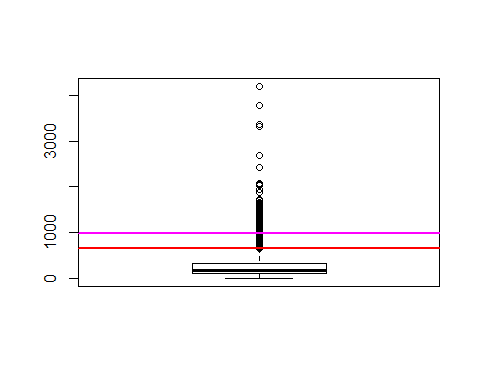
# Para el data analisis guardamos los missings  
dqr[dqr$variable=='y','missings']<-0  
# Para el data analisis guardamos los outliers  
dqr[dqr$variable=='y','errors']<-0  
# Para el data analisis guardamos los errores  
dqr[dqr$variable=='y','outliers']<-0  
  
  
#target+numerica (miramos qué valores toma)  
  
# Para el data analisis guardamos los missings (no hay)  
dqr[dqr$variable=="duration","missings"]<-sum(is.na(df[,"duration"]))  
  
summary(df$duration)#Vemos que hay valores muy pequeños, incluso 0, tambien de muy grandes.

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

hist(df$duration,100) #Miramos distribución



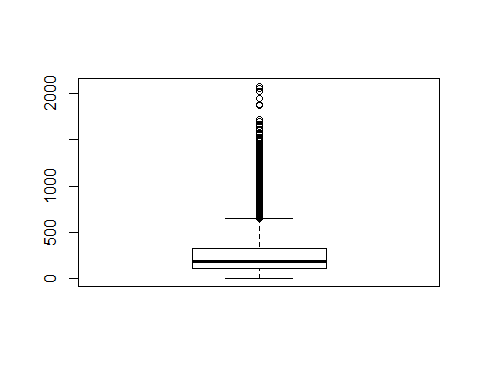
boxplot(df$duration) #Hacemos boxplot para ver outliers. Solo se contemplan outliers superiores.  
aux<-calcQ(df$duration)  
abline(h=aux[8],col="red",lwd=2) #Fijamos límite soft  
abline(h=aux[9],col="magenta",lwd=2) #Fijamos límite extreme



aux<-order(df$duration,decreasing=TRUE)[1:10];df[aux,'duration'] #Echamos un ojo a los 10 valores más extremos

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

df<-df[-aux[1:6],]#decidimos eliminar los 6 elementos más extremos   
boxplot(df$duration) #Hacemos boxplot de nuevo



# Para el data analisis guardamos los outliers  
dqr[dqr$variable=="duration","outliers"]<-6  
  
aux<-which(df$duration<5);length(aux);df[aux,'duration'] # Ahora miramos los que son inferiores a 5s (no puede considerarse llamada. Se considerarán errores)

## [1] 4

## [1] 0 4 0 1

df<-df[-aux,] # decidimos eliminarlos todos  
  
# Para el data analisis guardamos los errores  
dqr[dqr$variable=="duration","errors"]<-length(aux)  
dqri[nrow(df),]<-0  
dqri[,]<-0  
  
  
  
#no-target+categorica  
  
for(i in vars\_cat){  
 cat("################ ",i," ##################\n")  
 print(summary(df[,i]))  
} #Vemos un resumen de los niveles de cada variable categorica

## ################ 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

for(i in vars\_cat){ # Vamos a poner NA's en los unknows  
 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

# Para el data analisis guardamos los missings de las variables categoricas  
for(i in vars\_cat){   
 dqr[dqr$variable==i,"missings"]<-sum(is.na(df[,i]))  
 dqr[dqr$variable==i,"outliers"]<-0  
 dqr[dqr$variable==i,"errors"]<-0  
}  
  
aux2<-imputeMCA(df[,vars\_cat],ncp=10)  
  
for(i in vars\_cat){ # Contrastamos las imputaciones antes de modificar los datos  
 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,no\_imputadas)] # Aceptamos las imputaciones realizadas, solo guardamos las variables que tenian NA's  
  
# Guardamos los datos en este punto  
#save.image("data-step-2.RData")

# Step 2: REFACTORIZACIÓN

#rm(list=ls())#Limpiamos el workspace  
#load("data-step-2.RData", envir = parent.frame(), verbose = FALSE)  
  
# Refactorización (Agrupar subcategorias en menos categorias)  
  
for(i in vars\_cat){  
 cat("################ ",i," ##################\n")  
 print(summary(df[,i]))  
}

## ################ 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   
## ################ marital ##################  
## divorced married single   
## 583 2996 1411   
## ################ education ##################  
## 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   
## ################ housing ##################  
## no yes   
## 2279 2711   
## ################ loan ##################  
## no yes   
## 4245 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

# preguntar como criterio para las agrupaciones categoricas y en numericas  
  
#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

# admin. blue-collar entrepreneur housemaid management retired self-employed   
# 1296 1156 189 119 351 222 166   
# services student technician unemployed   
# 451 109 817 114   
# Define new factor categories: 1-Admin 2-Bussines 3-Not working 4-Serv-Tech 5-Blue-Collar  
  
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  
  
  
summary(df$f.job)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 1.00 3.00 2.71 4.00 4.00

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)) # Hacemos las etiquetas más informativas  
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

# 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

# apr aug dec jul jun mar may nov oct sep   
# 321 764 18 865 616 63 1680 513 80 70   
  
# Define new factor categories: 1- Spring 2-Summer 3-Resta  
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-jul   
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 las etiquetas más informativas  
summary(df$f.season)

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

#Education  
  
# education a Season  
  
table(df$education)

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

# basic.4y basic.6y basic.9y high.school illiterate professional.course university.degree   
# 515 271 810 1196 1 633 1564  
  
# Define new factor categories: 1-Basic 2-High School 3-Professional  
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 - Higb 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  
  
table(df$f.education);summary(df$f.education)

##   
## 1 2 3   
## 1597 1196 2197

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 1.00 2.00 2.12 3.00 3.00

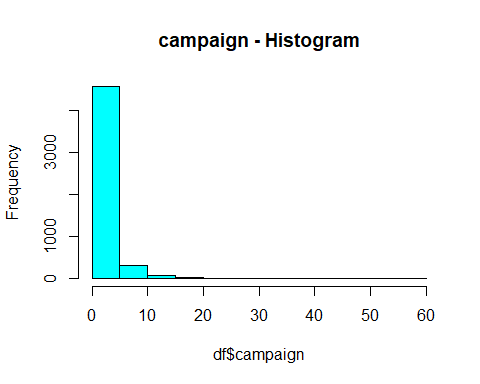
df$f.education<-factor(df$f.education,levels=1:3,labels=c("Basic","High School","Professional"))  
  
  
# Guardamos los datos en este punto  
#save.image("data-step-3.RData")

# Step 3:

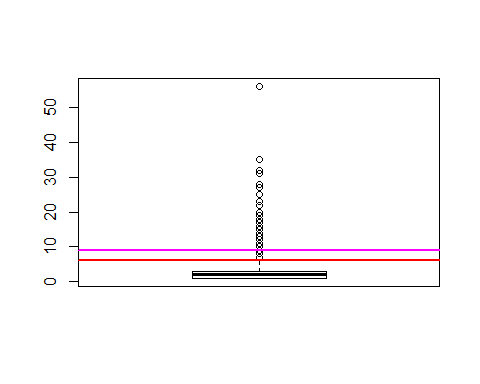
#rm(list=ls())#Limpiamos el workspace  
#load("data-step-3.RData", envir = parent.frame(), verbose = FALSE)  
  
#no-target+numerica: Revisión de cada variable con summary y boxplots  
  
 # age - Consideramos que no presenta ningún outlier # preguntar si esta bien  
   
 # Para el data analisis guardamos los missings  
 dqr[dqr$variable=='age','missings']<-0  
 # Para el data analisis guardamos los outliers  
 dqr[dqr$variable=='age','errors']<-0  
 # Para el data analisis guardamos los errores  
 dqr[dqr$variable=='age','outliers']<-0  
   
  
  
 # campaign - Consideramos que en los 10 meses que dura la campaña, contactar a un cliente como maximo cada 15 días puede ser posible  
 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")



boxplot(df$campaign, labels=row.names(df))  
  
 aux<-calcQ(df$campaign);  
 abline(h=aux[8],col="red",lwd=2) #Fijamos límite soft  
 abline(h=aux[9],col="magenta",lwd=2) #Fijamos límite extreme



aux<-which(df$campaign<=0);aux #Si se incluye el último contacto, este valor no puede ser 0

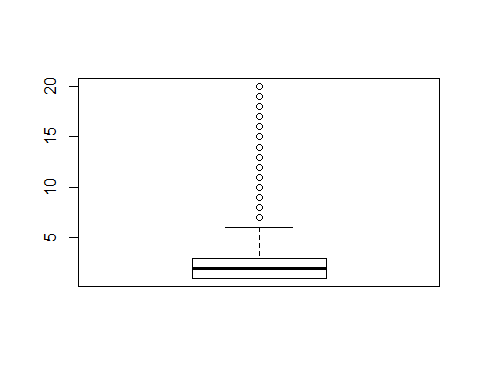
## integer(0)

aux<-which(df$campaign>20);length(aux);df[aux,'campaign'] #Hechamos un ojo a los outliers mayores de 20

## [1] 11

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

df[aux,"campaign"]<-NA # los ponemos como NA's  
 boxplot(df$campaign) #Hacemos boxplot de nuevo



# Para el data analisis guardamos los missings  
 dqr[dqr$variable=='campaing','missings']<-sum(is.na(df[,"campaign"]))  
 # Para el data analisis guardamos los outliers  
 dqr[dqr$variable=='campaing','errors']<-0   
 # Para el data analisis guardamos los errores  
 dqr[dqr$variable=='campaing','outliers']<-0  
 # Para los individuales  
 dqri[aux,'outliers']<-dqri[aux,'outliers']+1  
  
# Verificamos si hay algun error o outlier en pdays   
  
#Para pdays/previous/outcome creemos que estan relacionadas, por lo que podemos detectar errores si encontramos algún valor nonexistent ó 0 y que no sea 0 ó nonexistent respectivamente y además pdays=999  
  
 rel\_pdays<-which(df$pdays==999)  
 rel\_previous<-which(df$previous==0)  
 rel\_poutcome<-which(df$poutcome=='nonexistent')  
 length(setdiff(rel\_poutcome, rel\_previous)) # Cuantos en rel\_poutcome no están en rel\_previous y el resultado es 0

## [1] 0

length(setdiff(rel\_previous, rel\_poutcome)) # Ahora en el modo inverso vemos que el resultado es 0

## [1] 0

length(setdiff(rel\_previous, rel\_pdays)) # Los individuos que no han sido contactados en la campañas previa, tienen el pdays a 999

## [1] 0

length(setdiff(rel\_pdays, rel\_previous)) # Los individuos que nunca han sido contactados

## [1] 526

# preguntar porque la ultima diferencia da 527 - se toman como si no han participado nunnca en otra campaña por ejemplo ponerles 995 para identificar el error  
   
   
 #Vemos que hay pdays=999 que tienen previous>0 y poutcome!='nonexistent'  
 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

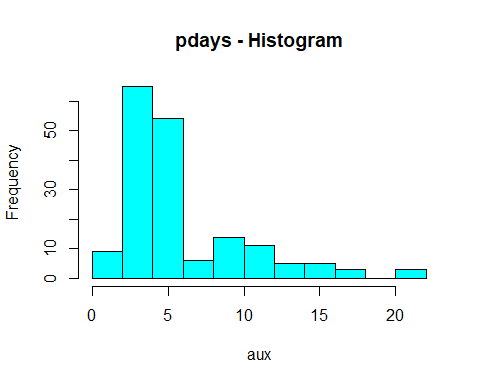
#Les asignamos NA como missing values, pues parece que hayan tenido campaña previa  
 #df[setdiff(aux,aux2),'pdays']<-NA  
 #guardamos los missingsindividuales  
 #dqri[setdiff(aux,aux2),'missings']<-dqri[setdiff(aux,aux2),'missings']+1  
  
 # pdays - preguntar creemos que no hay outliers, solo se han contactado a 175 personas en la campaña previa  
 summary(df$pdays)

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

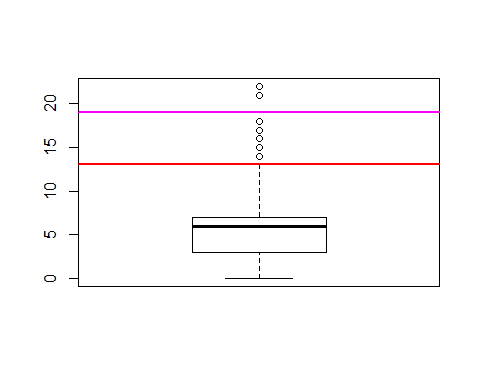
aux2<-which(df$pdays<999);summary(df$pdays[aux2])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 3.00 6.00 6.28 7.00 22.00

aux<-df$pdays[aux2]  
 hist(aux,col="cyan",main="pdays - Histogram")



boxplot(aux, labels=row.names(df))  
  
 aux2<-calcQ(aux);  
 abline(h=aux2[8],col="red",lwd=2) #Fijamos límite soft  
 abline(h=aux2[9],col="magenta",lwd=2) #Fijamos límite extreme



aux<-which( (df$pdays<999 & df$pdays>20) | (df$pdays<0));length(aux);df[aux,'pdays'] #Hechamos un ojo a los valores mayores a 20

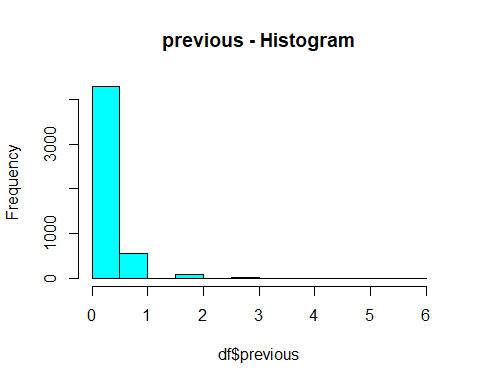
## [1] 3

## [1] 21 22 21

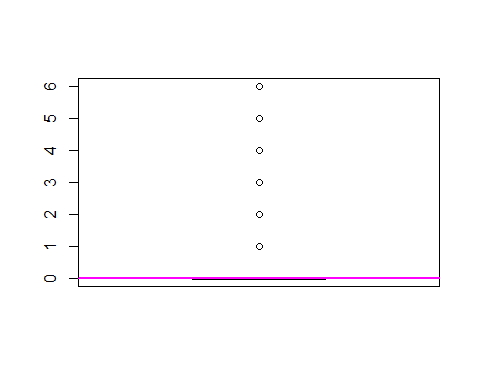
# Para el data analisis guardamos los missings  
 dqr[dqr$variable=='pdays','missings']<-sum(is.na(df[,"pdays"]))  
 # Para el data analisis guardamos los outliers  
 dqr[dqr$variable=='pdays','errors']<-0 # no se ve ningun numero negativo o algo raro  
 # Para el data analisis guardamos los errores  
 dqr[dqr$variable=='pdays','outliers']<-0  
  
  
  
 # previous - Vemos que de los outliers del boxplot   
   
 # Para el data analisis guardamos los missings  
 aux<-which(is.na(df$previous))  
 dqr[dqr$variable=='previous','missings']<-length(aux)  
 dqri[aux,'missings']<-dqri[aux,'missings']+1  
   
 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

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



boxplot(df$previous, labels=row.names(df))  
  
 aux<-calcQ(df$previous)  
 abline(h=aux[8],col="red",lwd=2) #Fijamos límite soft  
 abline(h=aux[9],col="magenta",lwd=2) #Fijamos límite extreme



# consideramos que para esta variable no hay outliers, ya que por los valores se ve que pueden haber sido contactado hasta en 6 capañas previas, lo que tiene sentido.  
   
 # Para el data analisis guardamos los outliers  
 dqr[dqr$variable=='previous','errors']<-0 # no se ve ningun numero negativo o algo raro  
 # Para el data analisis guardamos los errores  
 dqr[dqr$variable=='previous','outliers']<-0  
   
  
 #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, si no es que son errores  
 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")  
 }  
 }#Aparecen muchas discordancias #preguntar si es necesario cosiderar errores para estas variables pues o bien son índices locales que varían según la zona o bien son errores en introducir el mes o el valor -> No complicarse con estas variables solo anotarlo como observación

## ################ 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

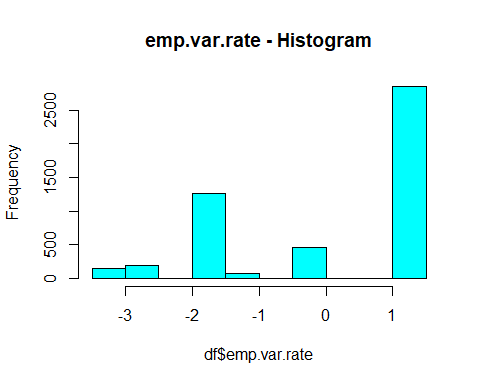
# explicar que es una incoherencia   
  
 # emp.var.rate  
 length(levels(factor(df$emp.var.rate)))#Comprobamos si solamente hay 10 valores distintos

## [1] 10

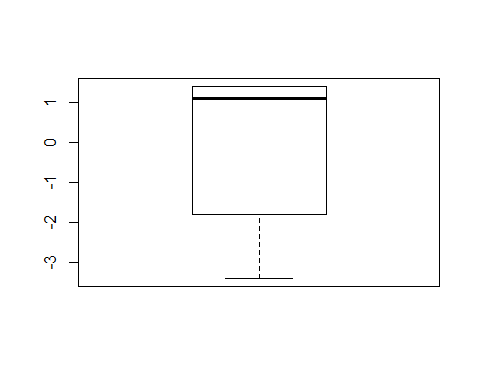
summary(df$emp.var.rate)

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

hist(df$emp.var.rate,col="cyan",main="emp.var.rate - Histogram")



boxplot(df$emp.var.rate, labels=row.names(df))  
  
 aux<-calcQ(df$emp.var.rate)  
 abline(h=aux[8],col="red",lwd=2) #Fijamos límite soft  
 abline(h=aux[9],col="magenta",lwd=2) #Fijamos límite extreme



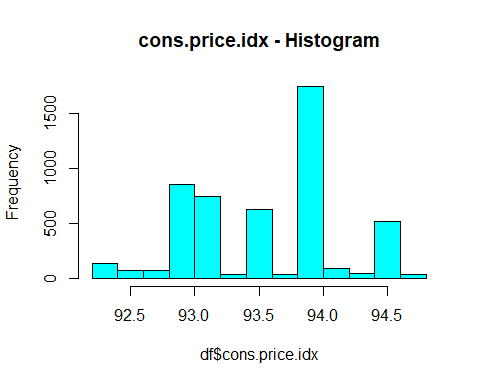
# Para el data analisis guardamos los missings  
 dqr[dqr$variable=='emp.var.rate','missings']<-sum(is.na(df[,"emp.var.rate"]))  
 # Para el data analisis guardamos los outliers  
 dqr[dqr$variable=='emp.var.rate','errors']<-0 # no se ve ningun numero negativo o algo raro  
 # Para el data analisis guardamos los errores  
 dqr[dqr$variable=='emp.var.rate','outliers']<-0  
  
   
 # cons.price.idx  
 length(levels(factor(df$cons.price.idx)))

## [1] 26

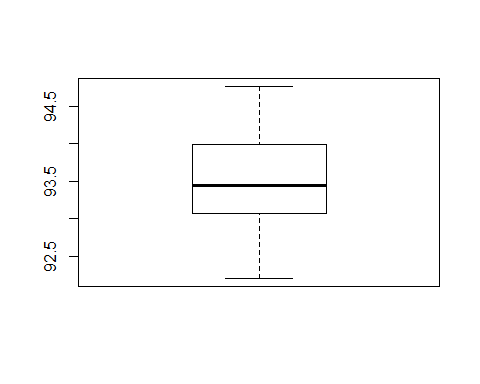
summary(df$cons.price.idx)

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

hist(df$cons.price.idx,col="cyan",main="cons.price.idx - Histogram")



boxplot(df$cons.price.idx, labels=row.names(df))  
  
 aux<-calcQ(df$cons.price.idx)  
 abline(h=aux[8],col="red",lwd=2) #Fijamos límite soft  
 abline(h=aux[9],col="magenta",lwd=2) #Fijamos límite extreme



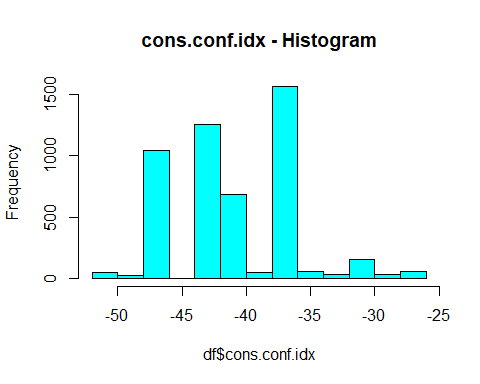
# Para el data analisis guardamos los missings  
 dqr[dqr$variable=='cons.price.idx','missings']<-sum(is.na(df[,"cons.price.idx"]))  
 # Para el data analisis guardamos los outliers  
 dqr[dqr$variable=='cons.price.idx','errors']<-0 # no se ve ningun numero negativo o algo raro  
 # Para el data analisis guardamos los errores  
 dqr[dqr$variable=='cons.price.idx','outliers']<-0  
  
  
 # cons.conf.idx  
 levels(factor(df$cons.conf.idx))

## [1] "-50.8" "-50" "-49.5" "-47.1" "-46.2" "-45.9" "-42.7" "-42"   
## [9] "-41.8" "-40.8" "-40.4" "-40.3" "-40" "-39.8" "-38.3" "-37.5"  
## [17] "-36.4" "-36.1" "-34.8" "-34.6" "-33.6" "-33" "-31.4" "-30.1"  
## [25] "-29.8" "-26.9"

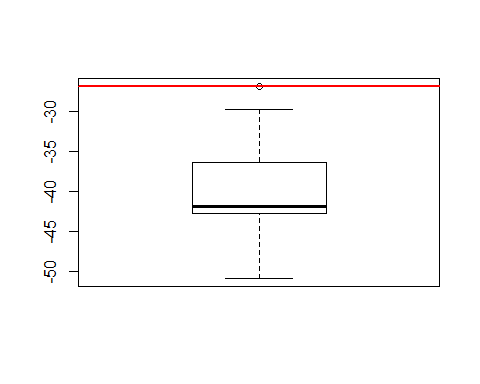
summary(df$cons.conf.idx)

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

hist(df$cons.conf.idx,col="cyan",main="cons.conf.idx - Histogram")



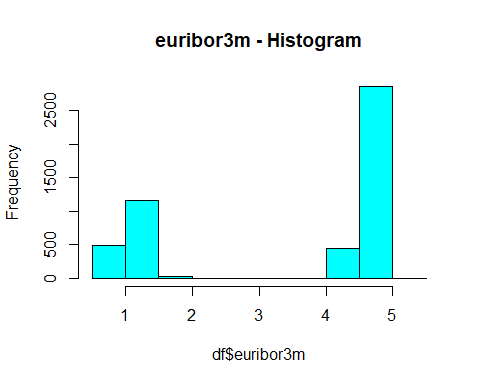
boxplot(df$cons.conf.idx, labels=row.names(df))  
  
 aux<-calcQ(df$cons.conf.idx)  
 abline(h=aux[8],col="red",lwd=2) #Fijamos límite soft  
 abline(h=aux[9],col="magenta",lwd=2) #Fijamos límite extreme



# Para el data analisis guardamos los missings  
 dqr[dqr$variable=='cons.conf.idx','missings']<-sum(is.na(df[,"cons.conf.idx"]))  
 # Para el data analisis guardamos los outliers  
 dqr[dqr$variable=='cons.conf.idx','errors']<-0 # no se ve ningun numero negativo o algo raro  
 # Para el data analisis guardamos los errores  
 dqr[dqr$variable=='cons.conf.idx','outliers']<-0  
  
  
 # euribor3m  
 summary(df$euribor3m)

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

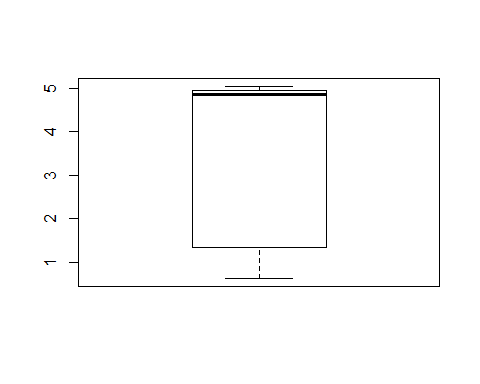
hist(df$euribor3m,col="cyan",main="euribor3m - Histogram")



boxplot(df$euribor3m, labels=row.names(df))  
  
 aux<-calcQ(df$euribor3m);aux

## $souti  
## 1st Qu.   
## -9.547   
##   
## $mouti  
## 1st Qu.   
## -4.1065   
##   
## $min  
## Min.   
## 0.634   
##   
## $q1  
## 1st Qu.   
## 1.334   
##   
## $q2  
## Median   
## 4.857   
##   
## $q3  
## 3rd Qu.   
## 4.961   
##   
## $max  
## Max.   
## 5.045   
##   
## $mouts  
## 3rd Qu.   
## 10.4015   
##   
## $souts  
## 3rd Qu.   
## 15.842

abline(h=aux[8],col="red",lwd=2) #Fijamos límite soft  
 abline(h=aux[9],col="magenta",lwd=2) #Fijamos límite extreme



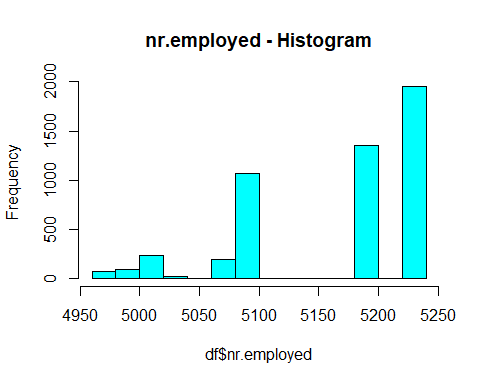
# Para el data analisis guardamos los missings  
 dqr[dqr$variable=='euribor3m','missings']<-sum(is.na(df[,"euribor3m"]))  
 # Para el data analisis guardamos los outliers  
 dqr[dqr$variable=='euribor3m','errors']<-0 # no se ve ningun numero negativo o algo raro  
 # Para el data analisis guardamos los errores  
 dqr[dqr$variable=='euribor3m','outliers']<-0  
  
 # nr.employed  
 levels(factor(df$nr.employed))

## [1] "4963.6" "4991.6" "5008.7" "5017.5" "5023.5" "5076.2" "5099.1"  
## [8] "5176.3" "5191" "5195.8" "5228.1"

summary(df$nr.employed)

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

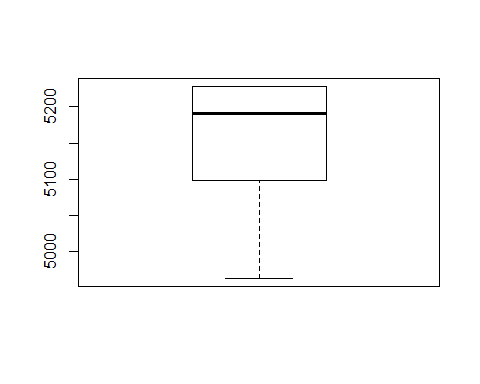
hist(df$nr.employed,col="cyan",main="nr.employed - Histogram")



boxplot(df$nr.employed, labels=row.names(df))  
  
 aux<-calcQ(df$nr.employed);aux

## $souti  
## 1st Qu.   
## 4712.1   
##   
## $mouti  
## 1st Qu.   
## 4905.6   
##   
## $min  
## Min.   
## 4963.6   
##   
## $q1  
## 1st Qu.   
## 5099.1   
##   
## $q2  
## Median   
## 5191   
##   
## $q3  
## 3rd Qu.   
## 5228.1   
##   
## $max  
## Max.   
## 5228.1   
##   
## $mouts  
## 3rd Qu.   
## 5421.6   
##   
## $souts  
## 3rd Qu.   
## 5615.1

abline(h=aux[8],col="red",lwd=2) #Fijamos límite soft  
 abline(h=aux[9],col="magenta",lwd=2) #Fijamos límite extreme



# Para el data analisis guardamos los missings  
 dqr[dqr$variable=='nr.employed','missings']<-sum(is.na(df[,"nr.employed"]))  
 # Para el data analisis guardamos los outliers  
 dqr[dqr$variable=='nr.employed','errors'] <-0 # no se ve ningun numero negativo o algo raro  
 # Para el data analisis guardamos los errores  
 dqr[dqr$variable=='nr.employed','outliers']<-0  
   
 #Imputacion de variables numericas  
 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   
##

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)#Observamos que da valores razonados, solamente que debemos redondearlos en ambos casos ya que se trata de "número de contactos"

## 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]  
 unique(df[,'previous'])

## [1] 0 1 2 3 4 5 6

unique(df[,'campaign'])

## [1] 1.000000 2.000000 3.000000 4.000000 5.000000 6.000000 8.000000  
## [8] 9.000000 13.000000 7.000000 10.000000 2.813103 2.813113 16.000000  
## [15] 15.000000 18.000000 20.000000 2.812980 12.000000 14.000000 11.000000  
## [22] 2.812942 17.000000 19.000000 2.929980 2.930073 2.849238 2.849686  
## [29] 2.849693 2.849669 2.849727

aux<-c('previous','campaign')  
 df[,aux]<-round(df[,aux])  
 unique(df[,'previous'])

## [1] 0 1 2 3 4 5 6

unique(df[,'campaign'])

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

# Guardamos los datos en este punto  
# save.image("data-step-4.RData")

# Step 4: Crear factores adicionales para cada variable cuantitativa

#rm(list=ls())#Limpiamos el workspace  
#load("data-step-4.RData", envir = parent.frame(), verbose = FALSE)  
  
# Para Age  
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))#Nos quedamos con los niveles naturales  
levels(df$f.age)<-paste0("f.age-",levels(df$f.age))#Hacemos las etiquetas más informativas  
summary(df$f.age)

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

# 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))#Hacemos 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

# Para campaign  
#aux<-quantile(df$campaign,seq(0,1,0.25),na.rm=TRUE)#Niveles por quartiles  
aux<-levels(factor(df$campaign))#En este caso no funcionan los niveles por quartiles. Miramos todos los niveles  
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 23   
## (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 10   
## (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 19   
## (19,20]   
## 20

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

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

#tapply(df$campaign,aux,median)#no tiene sentido ver la mediana en valores tan pequeños  
df$f.campaign<-factor(cut(df$campaign,breaks=aux2,include.lowest=T))#Nos quedamos con los niveles naturales  
levels(df$f.campaign)<-paste0("f.campaign-",levels(df$f.campaign))#Hacemos las etiquetas más informativas  
summary(df$f.campaign)

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

# Para pdays - Para pdays no tiene sentido ver outliers - revisar solo dos niveles  
aux2<-c(0,998,999) # Nos quedamos solo con dos niveles ya que no tiene sentido hacer más  
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$f.pdays<-pdays\_cutted # Nos quedamos con los niveles naturales  
levels(df$f.pdays)<-paste0("f.pdays-",levels(df$f.pdays))#Hacemos las etiquetas más informativas  
summary(df$f.pdays)

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

# Para previous - Vemos que solo hay 4 niveles por lo que los pasamos directamente a factores  
table(df$previous)

##   
## 0 1 2 3 4 5 6   
## 4289 564 98 30 5 3 1

df$f.previous<-factor(df$previous)  
summary(df$f.previous)

## 0 1 2 3 4 5 6   
## 4289 564 98 30 5 3 1

# Guardamos los datos en este punto  
# save.image("data-step-5.RData")

# Step 5: Profiling

#rm(list=ls())#Limpiamos el workspace  
#load("data-step-5.RData", envir = parent.frame(), verbose = FALSE)  
  
# Añadimos nombres a los subniveles de las factores  
vars\_cat\_con\_y<-c(vars\_cat,"y")  
for (i in vars\_cat\_con\_y){  
 levels(df[,i])<-paste0(i,".",levels(df[,i]))  
}  
  
  
sink("out.txt")  
cat("=================================================================================  
 CONDES  
=================================================================================  
")

## =================================================================================  
## CONDES  
## =================================================================================

# El 11 es la posición de la variable duration  
condes(df,11,proba=0.01) # Ya que con el valor de 0.001 solo tenemos relación con campaing, decidimos probar con un valor mayor

## $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 235.97615   
## 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

cat("=================================================================================  
 CATDES  
=================================================================================  
")

## =================================================================================  
## CATDES  
## =================================================================================

catdes(df,21,proba=0.001)

##   
## Link between the cluster variable and the categorical variables (chi-square test)  
## =================================================================================  
## p.value df  
## f.duration 1.038223e-118 3  
## poutcome 5.738265e-111 2  
## f.pdays 9.773367e-110 1  
## f.previous 3.390142e-53 6  
## month 7.431682e-53 9  
## 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  
##   
## Description of each cluster by the categories  
## =============================================  
## $y.no  
## Cla/Mod Mod/Cla Global  
## f.duration=f.duration-[5,120] 98.71548 34.5548561 31.2024048  
## f.pdays=f.pdays-(998,999] 91.00727 98.5161871 96.4929860  
## f.previous=0 91.11681 87.8597122 85.9519038  
## poutcome=poutcome.nonexistent 91.11681 87.8597122 85.9519038  
## contact=contact.telephone 94.34807 37.9046763 35.8116232  
## f.duration=f.duration-(120,180] 95.44513 20.7284173 19.3587174  
## default=default.unknown 95.04762 22.4370504 21.0420842  
## f.job=f.job.Serv-Tech-BlueC 92.07921 50.1798561 48.5771543  
## job=job.blue-collar 93.85813 24.3929856 23.1663327  
## month=month.may 92.55952 34.9595324 33.6673347  
## f.campaign=f.campaign-(2,20] 92.17391 33.3633094 32.2645291  
## f.age=f.age-(40,50] 92.41820 26.0341727 25.1102204  
## education=education.basic.9y 92.71605 16.8839928 16.2324649  
## f.education=Basic 91.42142 32.8237410 32.0040080  
## job=job.services 93.56984 9.4874101 9.0380762  
## marital=marital.single 86.60524 27.4730216 28.2765531  
## education=education.university.degree 86.70077 30.4856115 31.3426854  
## f.education=Professional 87.07328 43.0080935 44.0280561  
## f.campaign=f.campaign-[0,1] 86.89298 41.4343525 42.5050100  
## month=month.apr 80.37383 5.8003597 6.4328657  
## f.previous=1 82.80142 10.4991007 11.3026052  
## f.job=f.job.Entrep-Retired-selfEmpl 82.84229 10.7464029 11.5631263  
## job=job.student 72.47706 1.7760791 2.1843687  
## f.season=f.season.Sep-Dec 82.81938 12.6798561 13.6472946  
## month=month.sep 62.85714 0.9892086 1.4028056  
## f.age=f.age-(50,92] 83.21918 16.3893885 17.5551102  
## f.previous=2 63.26531 1.3938849 1.9639279  
## month=month.mar 55.55556 0.7868705 1.2625251  
## month=month.oct 57.50000 1.0341727 1.6032064  
## job=job.retired 71.62162 3.5746403 4.4488978  
## default=default.no 87.56345 77.5629496 78.9579158  
## f.previous=3 26.66667 0.1798561 0.6012024  
## contact=contact.cellular 86.23166 62.0953237 64.1883768  
## f.pdays=f.pdays-[0,998] 37.71429 1.4838129 3.5070140  
## poutcome=poutcome.success 33.97436 1.1915468 3.1262525  
## f.duration=f.duration-(300,2.1e+03] 73.20261 22.6618705 27.5951904  
## 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=0 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  
## 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=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  
## f.previous=2 9.665060e-12 -6.811404  
## 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  
## f.previous=3 1.097534e-15 -8.015429  
## contact=contact.cellular 2.596533e-20 -9.234435  
## 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 27.5951904  
## poutcome=poutcome.success 66.025641 19.003690 3.1262525  
## f.pdays=f.pdays-[0,998] 62.285714 20.110701 3.5070140  
## contact=contact.cellular 13.768342 81.365314 64.1883768  
## f.previous=3 73.333333 4.059041 0.6012024  
## default=default.no 12.436548 90.405904 78.9579158  
## job=job.retired 28.378378 11.623616 4.4488978  
## month=month.oct 42.500000 6.273063 1.6032064  
## month=month.mar 44.444444 5.166052 1.2625251  
## f.previous=2 36.734694 6.642066 1.9639279  
## f.age=f.age-(50,92] 16.780822 27.121771 17.5551102  
## month=month.sep 37.142857 4.797048 1.4028056  
## f.season=f.season.Sep-Dec 17.180617 21.586716 13.6472946  
## job=job.student 27.522936 5.535055 2.1843687  
## f.job=f.job.Entrep-Retired-selfEmpl 17.157712 18.265683 11.5631263  
## f.previous=1 17.198582 17.896679 11.3026052  
## month=month.apr 19.626168 11.623616 6.4328657  
## f.campaign=f.campaign-[0,1] 13.107025 51.291513 42.5050100  
## f.education=Professional 12.926718 52.398524 44.0280561  
## education=education.university.degree 13.299233 38.376384 31.3426854  
## marital=marital.single 13.394755 34.870849 28.2765531  
## job=job.services 6.430155 5.350554 9.0380762  
## f.education=Basic 8.578585 25.276753 32.0040080  
## education=education.basic.9y 7.283951 10.885609 16.2324649  
## f.age=f.age-(40,50] 7.581804 17.527675 25.1102204  
## f.campaign=f.campaign-(2,20] 7.826087 23.247232 32.2645291  
## month=month.may 7.440476 23.062731 33.6673347  
## job=job.blue-collar 6.141869 13.099631 23.1663327  
## f.job=f.job.Serv-Tech-BlueC 7.920792 35.424354 48.5771543  
## default=default.unknown 4.952381 9.594096 21.0420842  
## f.duration=f.duration-(120,180] 4.554865 8.118081 19.3587174  
## contact=contact.telephone 5.651931 18.634686 35.8116232  
## f.previous=0 8.883190 70.295203 85.9519038  
## poutcome=poutcome.nonexistent 8.883190 70.295203 85.9519038  
## f.pdays=f.pdays-(998,999] 8.992731 79.889299 96.4929860  
## f.duration=f.duration-[5,120] 1.284522 3.690037 31.2024048  
## 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  
## contact=contact.cellular 2.596533e-20 9.234435  
## f.previous=3 1.097534e-15 8.015429  
## 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.previous=2 9.665060e-12 6.811404  
## 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=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  
## 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=0 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 category  
## nr.employed 23.279681 5174.2504272 5165.87569138 65.9756532  
## pdays 22.264832 984.2796763 964.18517034 119.9458512  
## euribor3m 19.999540 3.7572383 3.58457355 1.6629196  
## emp.var.rate 19.282392 0.2024505 0.05212425 1.5004109  
## cons.price.idx 8.330259 93.5875852 93.56373427 0.5618398  
## campaign 5.633986 2.5807104 2.51503006 2.4292611  
## cons.conf.idx -4.309630 -40.6408273 -40.54192385 4.4122246  
## age -4.853184 39.9267086 40.17755511 9.8426481  
## previous -15.060507 0.1411871 0.17855711 0.4120855  
## duration -29.721802 222.8923112 259.85110220 201.9718952  
## 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 category  
## duration 29.721802 563.1586716 259.85110220 380.6385060  
## previous 15.060507 0.4852399 0.17855711 0.9064976  
## age 4.853184 42.2361624 40.17755511 14.3956848  
## cons.conf.idx 4.309630 -39.7302583 -40.54192385 6.1664495  
## campaign -5.633986 1.9760148 2.51503006 1.5727665  
## cons.price.idx -8.330259 93.3679982 93.56373427 0.6757139  
## emp.var.rate -19.282392 -1.1815498 0.05212425 1.6515762  
## euribor3m -19.999540 2.1675756 3.58457355 1.7747720  
## pdays -22.264832 799.2767528 964.18517034 398.0744161  
## nr.employed -23.279681 5097.1472325 5165.87569138 88.1024662  
## 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

sink()