Deliverable Final

Guillem Valls, Sergio Mazzariol

Table of Contents

[Preparación de la muestra 3](#_Toc11151369)

[Inicializamos datos y funciones 3](#_Toc11151370)

[Análisis y exploración de datos 4](#_Toc11151371)

[Tratamiento de las Variables target 6](#_Toc11151372)

[Y 6](#_Toc11151373)

[Duration 6](#_Toc11151374)

[Tratamiento de variables no-Target Categóricas 8](#_Toc11151375)

[Análisis de errores y missings 8](#_Toc11151376)

[Creación de nuevos niveles de los factores 12](#_Toc11151377)

[Tratamiento de variables no-Target Numéricas 14](#_Toc11151378)

[Age 14](#_Toc11151379)

[Campaing 14](#_Toc11151380)

[Verificación de inconsistencias en pdays/previous/poutcome 16](#_Toc11151381)

[Pdays 17](#_Toc11151382)

[Previous 17](#_Toc11151383)

[Comprobación de inconsistencias en los índices trimestrales/mensuales 18](#_Toc11151384)

[Emp.var.rate,cons.price.idx, cons.conf.idx, euribor3m, nr.employed 19](#_Toc11151385)

[Resumen del Data Quality Report y Ranking 21](#_Toc11151386)

[Creación de factores adicionales para cada variable cuantitativa 22](#_Toc11151387)

[Age 22](#_Toc11151388)

[Duration 23](#_Toc11151389)

[Campaign 24](#_Toc11151390)

[Pdays 25](#_Toc11151391)

[Previous 25](#_Toc11151392)

[Profiling 26](#_Toc11151393)

[Nombres de niveles más informativos 26](#_Toc11151394)

[Resultado del CONDES 26](#_Toc11151395)

[Resultado del CATDES 27](#_Toc11151396)

[Perfil de persona más propensa a que acepte el producto: 35](#_Toc11151397)

[Perfil de llamada más propensa a que se acepte el producto: 35](#_Toc11151398)

[Valores propios y ejes dominantes 35](#_Toc11151399)

[Eigenvalues and dominant axes analysis. How many axes we have to interpret according to Kaiser and Elbow’s rule? 35](#_Toc11151400)

[Individuals point of view: Are they any individuals “too contributive”? To better understand the axes meaning use the extreme individuals. Detection of multivariant outliers and influent data. 38](#_Toc11151401)

[Interpreting the axes: Variables point of view coordinates, quality of representation, contribution of the variables 41](#_Toc11151402)

[Perform a PCA taking into account also supplementary variables the supplementary variables can be quantitative and/or categorical 46](#_Toc11151403)

[K-Means Classification 53](#_Toc11151404)

[Description of clusters 55](#_Toc11151405)

[Hierarchical Clustering 66](#_Toc11151406)

[CA analysis for your data should contain your factor version of the numeric target (duration) in K= 7 (maximum 10) levels and 2 factors: 77](#_Toc11151407)

[Eigenvalues and dominant axes analysis. How many axes we have to consider are there any row categories that can be combined/avoided to explain Duration target. 77](#_Toc11151408)

[CA - duration vs f.age 77](#_Toc11151409)

[CA - Education vs f.duration 81](#_Toc11151410)

[Modelización con target numérico 87](#_Toc11151411)

[Modelización con variables explicativas numéricas 87](#_Toc11151412)

[Modelo simple 87](#_Toc11151413)

[Modelo con transformaciones 92](#_Toc11151414)

[Modelo de regresión polinómica 95](#_Toc11151415)

[Modelización con variables explicativas numéricas y categóricas 97](#_Toc11151416)

[Interacciones 99](#_Toc11151417)

[Validación 100](#_Toc11151418)

[Modelización con target binario 107](#_Toc11151419)

[Modelización con variables explicativas numéricas 108](#_Toc11151420)

[Modelo simple 108](#_Toc11151421)

[Modelo de regresión polinómica 111](#_Toc11151422)

[Modelización con variables explicativas numéricas y categóricas 114](#_Toc11151423)

[Interacciones 124](#_Toc11151424)

[Validación 127](#_Toc11151425)

# 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=integer(), 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], 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 ) }  
  
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"

# 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

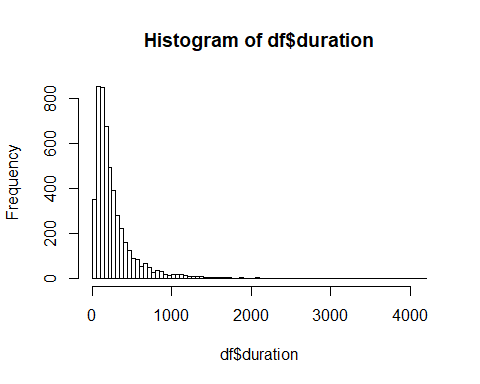
### Duration

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

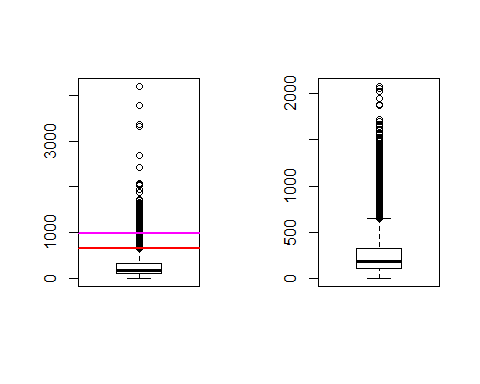
hist(df$duration,100)



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

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 individuos que se consideraban como outliers o errores en las variables target.  
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

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

# 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.

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,no\_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 las 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 - 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  
  
df$f.education<-factor(df$f.education,levels=1:3,labels=c("Basic","High School","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.

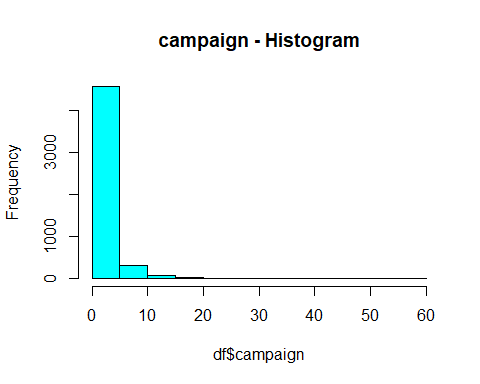
### Campaing

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

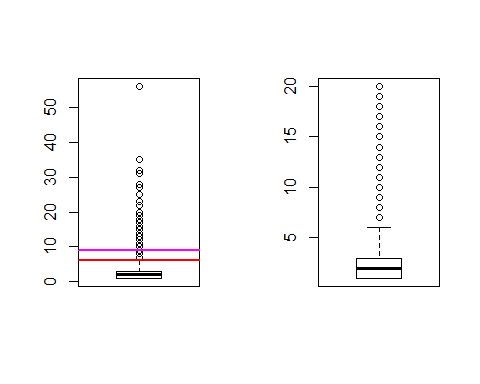
## integer(0)

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

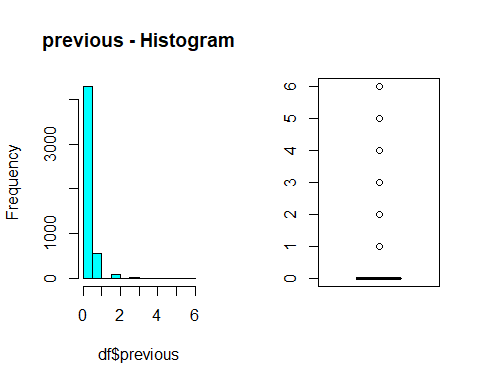
### Previous

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))#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

### 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 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)  
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 235.97615   
## 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

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 0.2184369   
## 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-square 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 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

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:

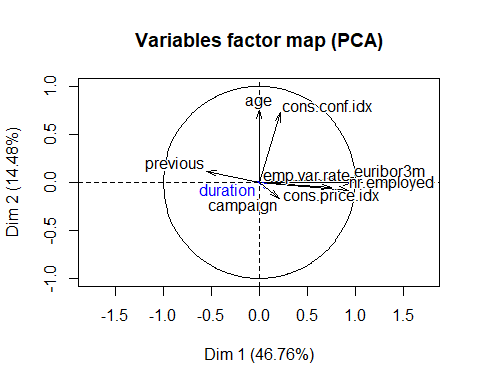
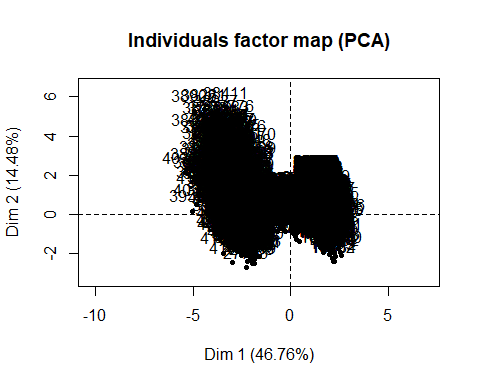
1. Abril, duración larga (más de 300 segundos) y hechas a un móvil.

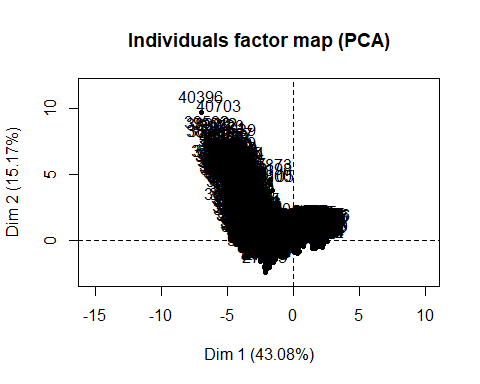
## Valores propios y ejes dominantes

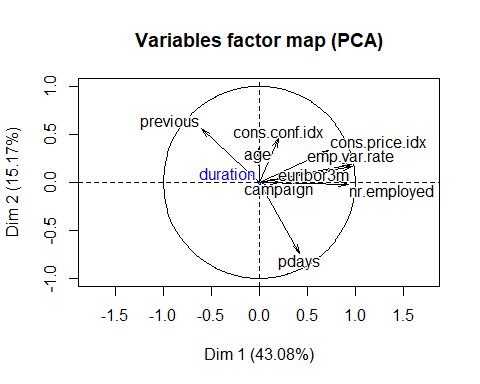
### Eigenvalues and dominant axes analysis. How many axes we have to interpret according to Kaiser and Elbow’s rule?

Hemos decidido probar como se ve el PCA sin y con la variable pdays, ya que consideramos que es una variable con bastantes missings, aún así aporta información por lo tanto la vamos a considerar.

vars\_num\_sin\_pday = vars\_num[-3];  
res2.pca<-PCA(df[,c('duration',vars\_num\_sin\_pday)],quanti.sup=1)



res.pca<-PCA(df[,c('duration',vars\_num)],quanti.sup=1)



Vemos que con pdays existe una relación inversa con previous, respecto a los dos ejes factoriales, sin pdays se puede ver que la contribución de la variable age con el segundo eje factorial es mayor, ya que gráficamente tiene mayor magnitud además que las variables socio económicas, se ven mejor representadas en el primer eje factorial.

Por la ley de Kaiser, deberíamos utilizar los 3 primeros ejes factoriales, los cuales son mayores a 1. Por la ley de ElBow, al realizar el gráfico podemos ver que la gráfica empieza a ser plana a partir de la 2da dimensión, es decir que se cogen las 2 primeras dimensiones.

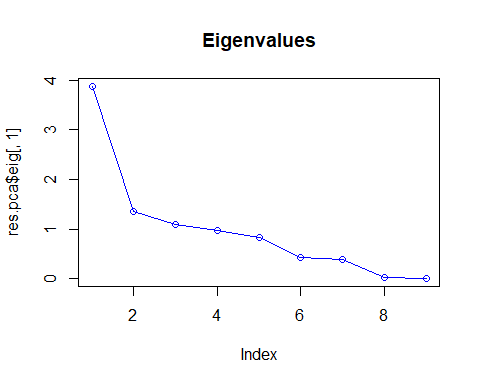
Si tomamos en cuenta el criterio del 80% se deberían coger las 4 primeras dimensiones.

Para realizar el futuro análisis, conviene utilizar dimensiones pares, por lo que decidimos solo usar 2.

summary(res.pca,ncp=4,nb.dec=2)

##   
## Call:  
## PCA(X = df[, c("duration", vars\_num)], quanti.sup = 1)   
##   
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7  
## Variance 3.88 1.36 1.10 0.97 0.83 0.43 0.39  
## % of var. 43.08 15.17 12.28 10.74 9.25 4.82 4.29  
## Cumulative % of var. 43.08 58.24 70.52 81.26 90.50 95.33 99.62  
## Dim.8 Dim.9  
## Variance 0.02 0.01  
## % of var. 0.27 0.12  
## Cumulative % of var. 99.88 100.00  
##   
## Individuals (the 10 first)  
## Dist Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3  
## 20 | 1.74 | 1.29 0.01 0.55 | 0.44 0.00 0.06 | 0.41  
## 21 | 1.99 | 1.30 0.01 0.42 | 0.18 0.00 0.01 | -0.13  
## 30 | 2.24 | 1.28 0.01 0.33 | 0.90 0.01 0.16 | 1.38  
## 33 | 1.93 | 1.29 0.01 0.44 | 0.73 0.01 0.14 | 1.02  
## 48 | 1.74 | 1.29 0.01 0.55 | 0.47 0.00 0.07 | 0.47  
## 56 | 2.24 | 1.28 0.01 0.33 | 0.90 0.01 0.16 | 1.38  
## 61 | 1.85 | 1.29 0.01 0.48 | 0.67 0.01 0.13 | 0.90  
## 62 | 2.02 | 1.29 0.01 0.41 | 0.78 0.01 0.15 | 1.14  
## 65 | 1.89 | 1.29 0.01 0.46 | 0.70 0.01 0.14 | 0.96  
## 84 | 1.93 | 1.29 0.01 0.44 | 0.73 0.01 0.14 | 1.02  
## ctr cos2 Dim.4 ctr cos2   
## 20 0.00 0.06 | -0.81 0.01 0.22 |  
## 21 0.00 0.00 | -1.06 0.02 0.28 |  
## 30 0.03 0.38 | -0.38 0.00 0.03 |  
## 33 0.02 0.28 | -0.54 0.01 0.08 |  
## 48 0.00 0.07 | -0.79 0.01 0.21 |  
## 56 0.03 0.38 | -0.38 0.00 0.03 |  
## 61 0.01 0.23 | -0.60 0.01 0.10 |  
## 62 0.02 0.32 | -0.49 0.00 0.06 |  
## 65 0.02 0.26 | -0.57 0.01 0.09 |  
## 84 0.02 0.28 | -0.54 0.01 0.08 |  
##   
## Variables  
## Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2  
## age | -0.01 0.00 0.00 | 0.35 8.93 0.12 | 0.67 40.10 0.44  
## campaign | 0.21 1.13 0.04 | 0.00 0.00 0.00 | -0.23 4.74 0.05  
## pdays | 0.42 4.63 0.18 | -0.74 40.21 0.55 | 0.28 7.10 0.08  
## previous | -0.60 9.37 0.36 | 0.56 22.84 0.31 | -0.30 8.38 0.09  
## emp.var.rate | 0.96 23.97 0.93 | 0.19 2.54 0.03 | -0.11 1.00 0.01  
## cons.price.idx | 0.72 13.43 0.52 | 0.33 8.11 0.11 | -0.30 8.08 0.09  
## cons.conf.idx | 0.20 1.06 0.04 | 0.46 15.36 0.21 | 0.58 30.44 0.34  
## euribor3m | 0.97 24.20 0.94 | 0.16 1.98 0.03 | -0.01 0.01 0.00  
## nr.employed | 0.93 22.20 0.86 | -0.02 0.03 0.00 | -0.04 0.16 0.00  
## Dim.4 ctr cos2   
## age | 0.28 8.23 0.08 |  
## campaign | 0.93 89.89 0.87 |  
## pdays | 0.04 0.17 0.00 |  
## previous | -0.02 0.05 0.00 |  
## emp.var.rate | -0.06 0.39 0.00 |  
## cons.price.idx | -0.06 0.41 0.00 |  
## cons.conf.idx | -0.04 0.13 0.00 |  
## euribor3m | -0.07 0.51 0.00 |  
## nr.employed | -0.05 0.24 0.00 |  
##   
## Supplementary continuous variable  
## Dim.1 cos2 Dim.2 cos2 Dim.3 cos2 Dim.4 cos2   
## duration | -0.02 0.00 | 0.02 0.00 | 0.00 0.00 | -0.05 0.00 |

plot(res.pca$eig[,1],main="Eigenvalues",type="o", col="blue")

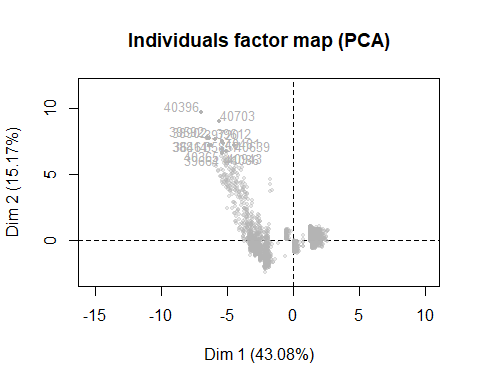


## Individuals point of view: Are they any individuals “too contributive”? To better understand the axes meaning use the extreme individuals. Detection of multivariant outliers and influent data.

Primero graficamos en rp los 15 individuos más contributivos en ambos ejes, luego analizamos los 5 individuos más contributivos en la dimensión 1 y 2. Al ver si estos tienen alguna relación significativa, podemos decir que para los 5 individuos de la dimension 1, vemos que principalmente son gente mayor de 45años, todos han comprado el producto, han sido contactados mediante el móvil, han sido contactados previamente, comprado un producto en una campaña anterior y la duración de la llamada ha sido mayor a los 300s.

Para la dimensión 2 podemos ver prácticamente las mismas características menos la duración que ha sido menor. Cabe destacar sin embargo que hay dos individuos que son muy contributivos en ambos ejes, eso hace pensar que pueden ser posibles outliers pero de igual forma los dejamos en los datos.

plot.PCA(res.pca, choix=c("ind"),cex=0.8,col.ind="grey70",select="contrib15",axes=c(1,2))



mas\_ctr\_dim1 <- sort(res.pca$ind$contrib[,1], decreasing = TRUE)[1:5]  
mas\_ctr\_dim2 <- sort(res.pca$ind$contrib[,2], decreasing = TRUE)[1:5]  
df[names(mas\_ctr\_dim1),]

## age job marital education  
## 40396 48 job.admin. marital.divorced education.university.degree  
## 39592 24 job.student marital.single education.high.school  
## 38902 83 job.retired marital.divorced education.basic.4y  
## 38814 65 job.retired marital.married education.university.degree  
## 36461 55 job.retired marital.married education.basic.4y  
## default housing loan contact month  
## 40396 default.no housing.no loan.no contact.cellular month.aug  
## 39592 default.no housing.yes loan.no contact.cellular month.may  
## 38902 default.no housing.no loan.no contact.cellular month.nov  
## 38814 default.no housing.no loan.no contact.cellular month.nov  
## 36461 default.no housing.no loan.no contact.cellular month.jun  
## day\_of\_week duration campaign pdays previous poutcome  
## 40396 day\_of\_week.thu 172 3 3 6 poutcome.success  
## 39592 day\_of\_week.wed 258 1 3 5 poutcome.success  
## 38902 day\_of\_week.tue 242 1 3 3 poutcome.success  
## 38814 day\_of\_week.fri 226 1 3 3 poutcome.success  
## 36461 day\_of\_week.tue 553 2 3 4 poutcome.failure  
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed  
## 40396 -1.7 94.027 -38.3 0.904 4991.6  
## 39592 -1.8 93.876 -40.0 0.672 5008.7  
## 38902 -3.4 92.649 -30.1 0.716 5017.5  
## 38814 -3.4 92.649 -30.1 0.714 5017.5  
## 36461 -2.9 92.963 -40.8 1.262 5076.2  
## y f.job f.season  
## 40396 y.yes f.job.Admin-Managment f.season.Jun-Aug  
## 39592 y.yes f.job.Not-working f.season.Mar-May  
## 38902 y.yes f.job.Entrep-Retired-selfEmpl f.season.Sep-Dec  
## 38814 y.yes f.job.Entrep-Retired-selfEmpl f.season.Sep-Dec  
## 36461 y.no f.job.Entrep-Retired-selfEmpl f.season.Jun-Aug  
## f.education f.age f.duration  
## 40396 f.education.Professional f.age-(40,50] f.duration-(120,180]  
## 39592 f.education.High School f.age-[18,30] f.duration-(180,300]  
## 38902 f.education.Basic f.age-(50,92] f.duration-(180,300]  
## 38814 f.education.Professional f.age-(50,92] f.duration-(180,300]  
## 36461 f.education.Basic f.age-(50,92] f.duration-(300,2.1e+03]  
## f.campaign f.pdays f.previous  
## 40396 f.campaign-(2,20] f.pdays-[0,22] f.previous-(1,6]  
## 39592 f.campaign-[0,1] f.pdays-[0,22] f.previous-(1,6]  
## 38902 f.campaign-[0,1] f.pdays-[0,22] f.previous-(1,6]  
## 38814 f.campaign-[0,1] f.pdays-[0,22] f.previous-(1,6]  
## 36461 f.campaign-(1,2] f.pdays-[0,22] f.previous-(1,6]

df[names(mas\_ctr\_dim2),]

## age job marital education  
## 40396 48 job.admin. marital.divorced education.university.degree  
## 40703 82 job.retired marital.married education.university.degree  
## 39592 24 job.student marital.single education.high.school  
## 38902 83 job.retired marital.divorced education.basic.4y  
## 39612 52 job.technician marital.married education.university.degree  
## default housing loan contact month  
## 40396 default.no housing.no loan.no contact.cellular month.aug  
## 40703 default.unknown housing.no loan.no contact.cellular month.sep  
## 39592 default.no housing.yes loan.no contact.cellular month.may  
## 38902 default.no housing.no loan.no contact.cellular month.nov  
## 39612 default.no housing.no loan.no contact.cellular month.may  
## day\_of\_week duration campaign pdays previous poutcome  
## 40396 day\_of\_week.thu 172 3 3 6 poutcome.success  
## 40703 day\_of\_week.mon 81 3 3 4 poutcome.success  
## 39592 day\_of\_week.wed 258 1 3 5 poutcome.success  
## 38902 day\_of\_week.tue 242 1 3 3 poutcome.success  
## 39612 day\_of\_week.thu 211 1 3 4 poutcome.success  
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed  
## 40396 -1.7 94.027 -38.3 0.904 4991.6  
## 40703 -1.1 94.199 -37.5 0.879 4963.6  
## 39592 -1.8 93.876 -40.0 0.672 5008.7  
## 38902 -3.4 92.649 -30.1 0.716 5017.5  
## 39612 -1.8 93.876 -40.0 0.677 5008.7  
## y f.job f.season  
## 40396 y.yes f.job.Admin-Managment f.season.Jun-Aug  
## 40703 y.no f.job.Entrep-Retired-selfEmpl f.season.Sep-Dec  
## 39592 y.yes f.job.Not-working f.season.Mar-May  
## 38902 y.yes f.job.Entrep-Retired-selfEmpl f.season.Sep-Dec  
## 39612 y.yes f.job.Serv-Tech-BlueC f.season.Mar-May  
## f.education f.age f.duration  
## 40396 f.education.Professional f.age-(40,50] f.duration-(120,180]  
## 40703 f.education.Professional f.age-(50,92] f.duration-[5,120]  
## 39592 f.education.High School f.age-[18,30] f.duration-(180,300]  
## 38902 f.education.Basic f.age-(50,92] f.duration-(180,300]  
## 39612 f.education.Professional f.age-(50,92] f.duration-(180,300]  
## f.campaign f.pdays f.previous  
## 40396 f.campaign-(2,20] f.pdays-[0,22] f.previous-(1,6]  
## 40703 f.campaign-(2,20] f.pdays-[0,22] f.previous-(1,6]  
## 39592 f.campaign-[0,1] f.pdays-[0,22] f.previous-(1,6]  
## 38902 f.campaign-[0,1] f.pdays-[0,22] f.previous-(1,6]  
## 39612 f.campaign-[0,1] f.pdays-[0,22] f.previous-(1,6]

## Interpreting the axes: Variables point of view coordinates, quality of representation, contribution of the variables

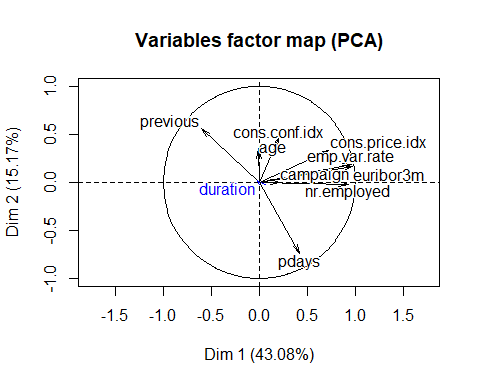
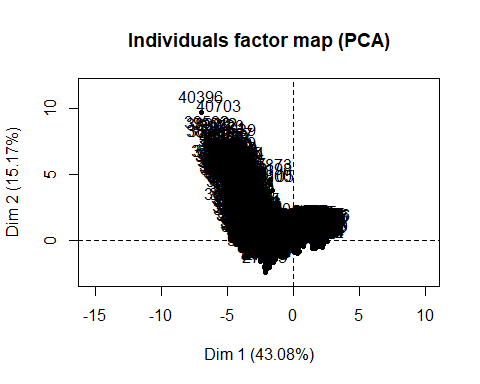
Al hacer el PCA con la variable target duration como suplementaria, podemos ver que su módulo es prácticamente nulo, esto quiere decir que la variable no se ve representada en niguno de los ejes factoriales.

El eje horizontal está muy relacionado con las variables socio económicas, mirando el cos2 del summary podemos ver que las variables que están mejor representadas con la dimensión 1 son: euribor3m, emp.var.rate, nr.emplyed Para el eje vertical: pdays y previous

Para el eje vertical, podemos decir que está relacionado con las campañas previas.

Al hacer el PCA con la variable target Y como suplementaria, podemos ver que en el gráfico de rp, el factor NO, esta muy cerca del centro, por lo que no se ve representada en estos ejes factoriales. En cambio el factor SI, está a una distancia mayor del centro, aunque poco significativa.

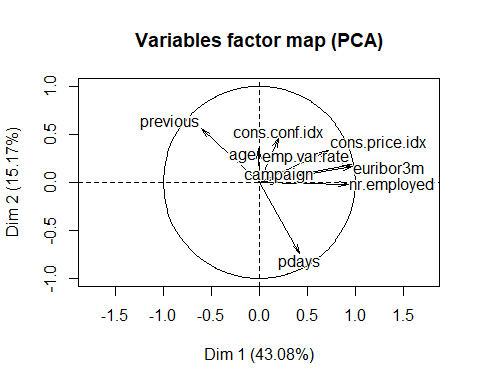
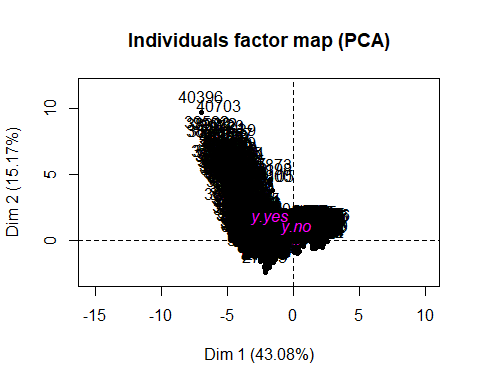
res.pca<-PCA(df[,c('duration',vars\_num)],quanti.sup=1)



summary(res.pca,ncp=4,nb.dec=2)

##   
## Call:  
## PCA(X = df[, c("duration", vars\_num)], quanti.sup = 1)   
##   
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7  
## Variance 3.88 1.36 1.10 0.97 0.83 0.43 0.39  
## % of var. 43.08 15.17 12.28 10.74 9.25 4.82 4.29  
## Cumulative % of var. 43.08 58.24 70.52 81.26 90.50 95.33 99.62  
## Dim.8 Dim.9  
## Variance 0.02 0.01  
## % of var. 0.27 0.12  
## Cumulative % of var. 99.88 100.00  
##   
## Individuals (the 10 first)  
## Dist Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3  
## 20 | 1.74 | 1.29 0.01 0.55 | 0.44 0.00 0.06 | 0.41  
## 21 | 1.99 | 1.30 0.01 0.42 | 0.18 0.00 0.01 | -0.13  
## 30 | 2.24 | 1.28 0.01 0.33 | 0.90 0.01 0.16 | 1.38  
## 33 | 1.93 | 1.29 0.01 0.44 | 0.73 0.01 0.14 | 1.02  
## 48 | 1.74 | 1.29 0.01 0.55 | 0.47 0.00 0.07 | 0.47  
## 56 | 2.24 | 1.28 0.01 0.33 | 0.90 0.01 0.16 | 1.38  
## 61 | 1.85 | 1.29 0.01 0.48 | 0.67 0.01 0.13 | 0.90  
## 62 | 2.02 | 1.29 0.01 0.41 | 0.78 0.01 0.15 | 1.14  
## 65 | 1.89 | 1.29 0.01 0.46 | 0.70 0.01 0.14 | 0.96  
## 84 | 1.93 | 1.29 0.01 0.44 | 0.73 0.01 0.14 | 1.02  
## ctr cos2 Dim.4 ctr cos2   
## 20 0.00 0.06 | -0.81 0.01 0.22 |  
## 21 0.00 0.00 | -1.06 0.02 0.28 |  
## 30 0.03 0.38 | -0.38 0.00 0.03 |  
## 33 0.02 0.28 | -0.54 0.01 0.08 |  
## 48 0.00 0.07 | -0.79 0.01 0.21 |  
## 56 0.03 0.38 | -0.38 0.00 0.03 |  
## 61 0.01 0.23 | -0.60 0.01 0.10 |  
## 62 0.02 0.32 | -0.49 0.00 0.06 |  
## 65 0.02 0.26 | -0.57 0.01 0.09 |  
## 84 0.02 0.28 | -0.54 0.01 0.08 |  
##   
## Variables  
## Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2  
## age | -0.01 0.00 0.00 | 0.35 8.93 0.12 | 0.67 40.10 0.44  
## campaign | 0.21 1.13 0.04 | 0.00 0.00 0.00 | -0.23 4.74 0.05  
## pdays | 0.42 4.63 0.18 | -0.74 40.21 0.55 | 0.28 7.10 0.08  
## previous | -0.60 9.37 0.36 | 0.56 22.84 0.31 | -0.30 8.38 0.09  
## emp.var.rate | 0.96 23.97 0.93 | 0.19 2.54 0.03 | -0.11 1.00 0.01  
## cons.price.idx | 0.72 13.43 0.52 | 0.33 8.11 0.11 | -0.30 8.08 0.09  
## cons.conf.idx | 0.20 1.06 0.04 | 0.46 15.36 0.21 | 0.58 30.44 0.34  
## euribor3m | 0.97 24.20 0.94 | 0.16 1.98 0.03 | -0.01 0.01 0.00  
## nr.employed | 0.93 22.20 0.86 | -0.02 0.03 0.00 | -0.04 0.16 0.00  
## Dim.4 ctr cos2   
## age | 0.28 8.23 0.08 |  
## campaign | 0.93 89.89 0.87 |  
## pdays | 0.04 0.17 0.00 |  
## previous | -0.02 0.05 0.00 |  
## emp.var.rate | -0.06 0.39 0.00 |  
## cons.price.idx | -0.06 0.41 0.00 |  
## cons.conf.idx | -0.04 0.13 0.00 |  
## euribor3m | -0.07 0.51 0.00 |  
## nr.employed | -0.05 0.24 0.00 |  
##   
## Supplementary continuous variable  
## Dim.1 cos2 Dim.2 cos2 Dim.3 cos2 Dim.4 cos2   
## duration | -0.02 0.00 | 0.02 0.00 | 0.00 0.00 | -0.05 0.00 |

res.pca<-PCA(df[,c('y',vars\_num)],quali.sup=1)



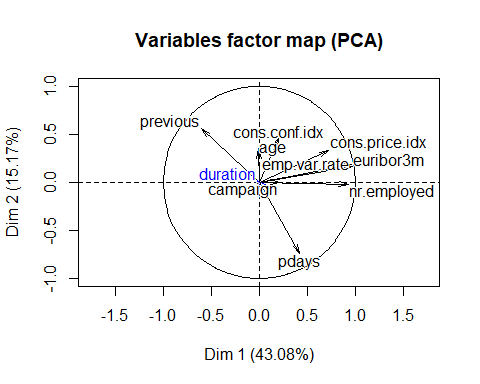
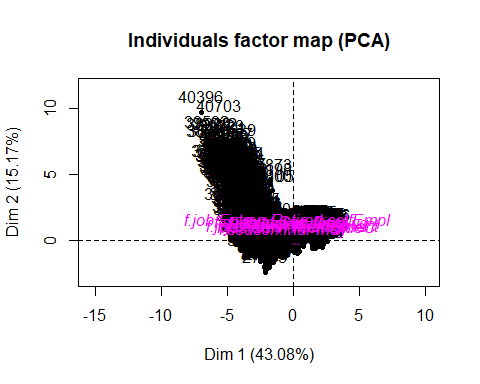
summary(res.pca,ncp=4,nb.dec=2)

##   
## Call:  
## PCA(X = df[, c("y", vars\_num)], quali.sup = 1)   
##   
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7  
## Variance 3.88 1.36 1.10 0.97 0.83 0.43 0.39  
## % of var. 43.08 15.17 12.28 10.74 9.25 4.82 4.29  
## Cumulative % of var. 43.08 58.24 70.52 81.26 90.50 95.33 99.62  
## Dim.8 Dim.9  
## Variance 0.02 0.01  
## % of var. 0.27 0.12  
## Cumulative % of var. 99.88 100.00  
##   
## Individuals (the 10 first)  
## Dist Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3  
## 20 | 1.74 | 1.29 0.01 0.55 | 0.44 0.00 0.06 | 0.41  
## 21 | 1.99 | 1.30 0.01 0.42 | 0.18 0.00 0.01 | -0.13  
## 30 | 2.24 | 1.28 0.01 0.33 | 0.90 0.01 0.16 | 1.38  
## 33 | 1.93 | 1.29 0.01 0.44 | 0.73 0.01 0.14 | 1.02  
## 48 | 1.74 | 1.29 0.01 0.55 | 0.47 0.00 0.07 | 0.47  
## 56 | 2.24 | 1.28 0.01 0.33 | 0.90 0.01 0.16 | 1.38  
## 61 | 1.85 | 1.29 0.01 0.48 | 0.67 0.01 0.13 | 0.90  
## 62 | 2.02 | 1.29 0.01 0.41 | 0.78 0.01 0.15 | 1.14  
## 65 | 1.89 | 1.29 0.01 0.46 | 0.70 0.01 0.14 | 0.96  
## 84 | 1.93 | 1.29 0.01 0.44 | 0.73 0.01 0.14 | 1.02  
## ctr cos2 Dim.4 ctr cos2   
## 20 0.00 0.06 | -0.81 0.01 0.22 |  
## 21 0.00 0.00 | -1.06 0.02 0.28 |  
## 30 0.03 0.38 | -0.38 0.00 0.03 |  
## 33 0.02 0.28 | -0.54 0.01 0.08 |  
## 48 0.00 0.07 | -0.79 0.01 0.21 |  
## 56 0.03 0.38 | -0.38 0.00 0.03 |  
## 61 0.01 0.23 | -0.60 0.01 0.10 |  
## 62 0.02 0.32 | -0.49 0.00 0.06 |  
## 65 0.02 0.26 | -0.57 0.01 0.09 |  
## 84 0.02 0.28 | -0.54 0.01 0.08 |  
##   
## Variables  
## Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2  
## age | -0.01 0.00 0.00 | 0.35 8.93 0.12 | 0.67 40.10 0.44  
## campaign | 0.21 1.13 0.04 | 0.00 0.00 0.00 | -0.23 4.74 0.05  
## pdays | 0.42 4.63 0.18 | -0.74 40.21 0.55 | 0.28 7.10 0.08  
## previous | -0.60 9.37 0.36 | 0.56 22.84 0.31 | -0.30 8.38 0.09  
## emp.var.rate | 0.96 23.97 0.93 | 0.19 2.54 0.03 | -0.11 1.00 0.01  
## cons.price.idx | 0.72 13.43 0.52 | 0.33 8.11 0.11 | -0.30 8.08 0.09  
## cons.conf.idx | 0.20 1.06 0.04 | 0.46 15.36 0.21 | 0.58 30.44 0.34  
## euribor3m | 0.97 24.20 0.94 | 0.16 1.98 0.03 | -0.01 0.01 0.00  
## nr.employed | 0.93 22.20 0.86 | -0.02 0.03 0.00 | -0.04 0.16 0.00  
## Dim.4 ctr cos2   
## age | 0.28 8.23 0.08 |  
## campaign | 0.93 89.89 0.87 |  
## pdays | 0.04 0.17 0.00 |  
## previous | -0.02 0.05 0.00 |  
## emp.var.rate | -0.06 0.39 0.00 |  
## cons.price.idx | -0.06 0.41 0.00 |  
## cons.conf.idx | -0.04 0.13 0.00 |  
## euribor3m | -0.07 0.51 0.00 |  
## nr.employed | -0.05 0.24 0.00 |  
##   
## Supplementary categories  
## Dist Dim.1 cos2 v.test Dim.2 cos2 v.test   
## y.no | 0.23 | 0.21 0.84 21.77 | -0.08 0.13 -14.15 |  
## y.yes | 1.89 | -1.74 0.84 -21.77 | 0.67 0.13 14.15 |  
## Dim.3 cos2 v.test Dim.4 cos2 v.test   
## y.no -0.01 0.00 -1.74 | 0.01 0.00 1.09 |  
## y.yes 0.07 0.00 1.74 | -0.04 0.00 -1.09 |

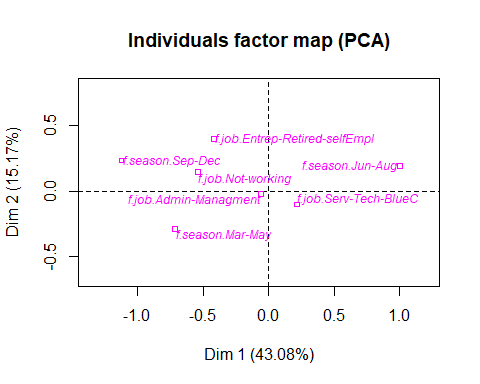
## Perform a PCA taking into account also supplementary variables the supplementary variables can be quantitative and/or categorical

Hemos dividido el plot en diferentes partes, para así poder entender y ver mejor el resultado. Para la primera dimensión podemos ver que para los niveles mejor representados son: f.season.jun-aug, f.previous-(0.9,1], y.no, y.yes.

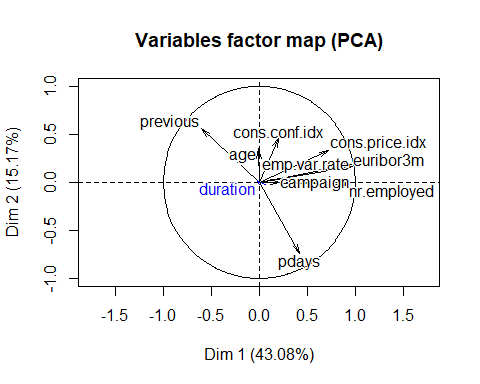
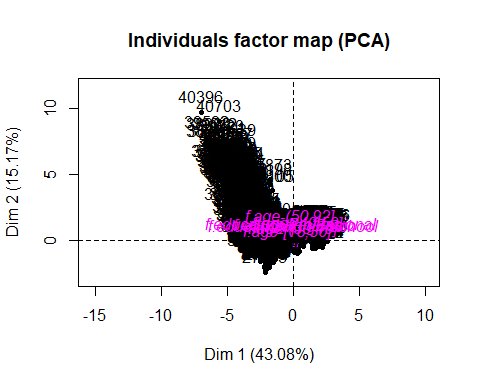
Para la segunda dimsensión, las que se ven mejor representadas son: f.pdays-[0,22],f.pdays-(22,23], f.previous-(1,6]

vars\_factorizadas<- c("f.job","f.season","f.education","f.age","f.duration","f.campaign","f.pdays","f.previous","y");  
  
res.pca<-PCA(df[,c('duration',vars\_num, "f.job","f.season")],quanti.sup=1, quali.sup = c(11:12))

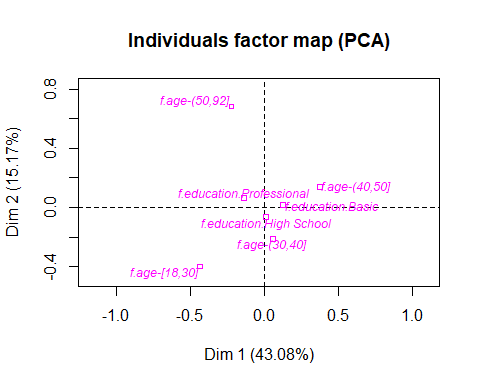
plot.PCA(res.pca,choix="ind",invisible="ind",cex=0.75)

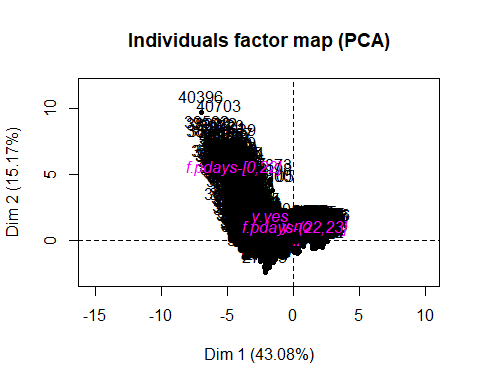
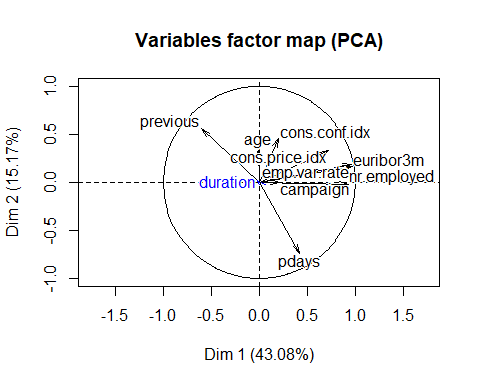


res.pca<-PCA(df[,c('duration',vars\_num,"f.education","f.age")],quanti.sup=1, quali.sup = c(11:12))

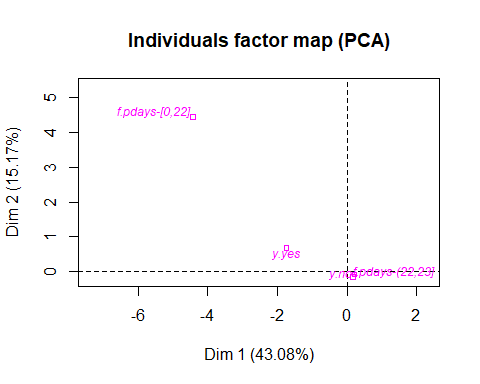


plot.PCA(res.pca,choix="ind",invisible="ind",cex=0.75)

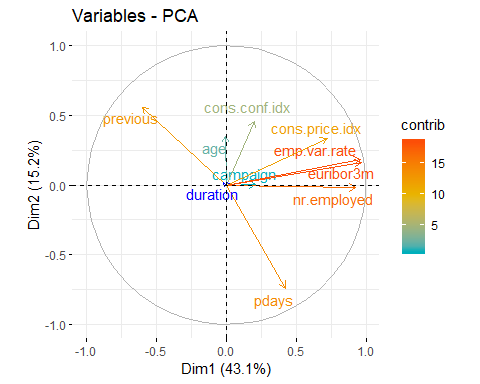


res.pca<-PCA(df[,c('duration',vars\_num,"f.pdays","y")],quanti.sup=1, quali.sup = c(11:12))

plot.PCA(res.pca,choix="ind",invisible="ind",cex=0.75)



res.pca<-PCA(df[,c('duration',vars\_num,vars\_factorizadas)],quanti.sup=1, quali.sup = c(11:19),graph=FALSE)  
  
fviz\_pca\_var(res.pca, col.var = "contrib",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)



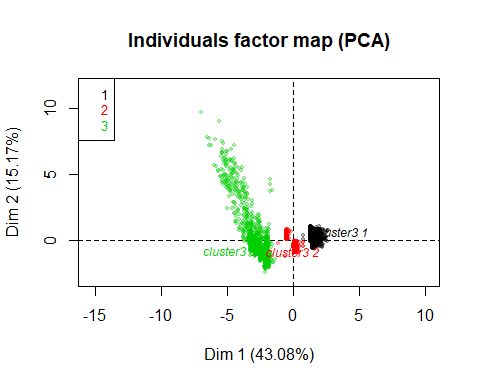
summary(res.pca,dig = 2, nbelements= 30, nbind=3, ncp=2)

##   
## Call:  
## PCA(X = df[, c("duration", vars\_num, vars\_factorizadas)], quanti.sup = 1,   
## quali.sup = c(11:19), graph = FALSE)   
##   
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6  
## Variance 3.877 1.365 1.105 0.966 0.832 0.434  
## % of var. 43.078 15.166 12.276 10.736 9.246 4.823  
## Cumulative % of var. 43.078 58.244 70.521 81.257 90.503 95.326  
## Dim.7 Dim.8 Dim.9  
## Variance 0.386 0.024 0.010  
## % of var. 4.291 0.266 0.116  
## Cumulative % of var. 99.618 99.884 100.000  
##   
## Individuals (the 3 first)  
## Dist Dim.1 ctr cos2 Dim.2 ctr  
## 20 | 1.739 | 1.291 0.009 0.551 | 0.441 0.003  
## 21 | 1.990 | 1.295 0.009 0.424 | 0.184 0.000  
## 30 | 2.240 | 1.284 0.009 0.328 | 0.898 0.012  
## cos2   
## 20 0.064 |  
## 21 0.009 |  
## 30 0.161 |  
##   
## Variables  
## Dim.1 ctr cos2 Dim.2 ctr  
## age | -0.009 0.002 0.000 | 0.349 8.933  
## campaign | 0.210 1.132 0.044 | 0.003 0.001  
## pdays | 0.424 4.633 0.180 | -0.741 40.212  
## previous | -0.603 9.366 0.363 | 0.558 22.836  
## emp.var.rate | 0.964 23.967 0.929 | 0.186 2.544  
## cons.price.idx | 0.722 13.431 0.521 | 0.333 8.109  
## cons.conf.idx | 0.203 1.063 0.041 | 0.458 15.356  
## euribor3m | 0.969 24.202 0.938 | 0.164 1.978  
## nr.employed | 0.928 22.205 0.861 | -0.020 0.030  
## cos2   
## age 0.122 |  
## campaign 0.000 |  
## pdays 0.549 |  
## previous 0.312 |  
## emp.var.rate 0.035 |  
## cons.price.idx 0.111 |  
## cons.conf.idx 0.210 |  
## euribor3m 0.027 |  
## nr.employed 0.000 |  
##   
## Supplementary continuous variable  
## Dim.1 cos2 Dim.2 cos2   
## duration | -0.022 0.000 | 0.017 0.000 |  
##   
## Supplementary categories  
## Dist Dim.1 cos2 v.test   
## f.job.Admin-Managment | 0.125 | -0.061 0.237 -1.532 |  
## f.job.Entrep-Retired-selfEmpl | 1.020 | -0.418 0.168 -5.418 |  
## f.job.Not-working | 0.690 | -0.541 0.615 -5.266 |  
## f.job.Serv-Tech-BlueC | 0.272 | 0.217 0.637 7.570 |  
## f.season.Mar-May | 0.793 | -0.719 0.821 -21.657 |  
## f.season.Jun-Aug | 1.026 | 1.002 0.954 32.508 |  
## f.season.Sep-Dec | 1.257 | -1.125 0.801 -16.040 |  
## f.education.Basic | 0.294 | 0.125 0.181 3.071 |  
## f.education.High School | 0.155 | 0.010 0.005 0.285 |  
## f.education.Professional | 0.255 | -0.140 0.300 -3.384 |  
## f.age-[18,30] | 1.312 | -0.436 0.111 -7.188 |  
## f.age-(30,40] | 0.485 | 0.055 0.013 1.600 |  
## f.age-(40,50] | 0.606 | 0.373 0.380 7.755 |  
## f.age-(50,92] | 1.667 | -0.225 0.018 -3.730 |  
## f.duration-[5,120] | 0.274 | 0.190 0.483 4.596 |  
## f.duration-(120,180] | 0.138 | 0.076 0.299 1.332 |  
## f.duration-(180,300] | 0.222 | -0.180 0.658 -3.418 |  
## f.duration-(300,2.1e+03] | 0.166 | -0.126 0.571 -2.781 |  
## f.campaign-[0,1] | 0.707 | -0.356 0.254 -10.989 |  
## f.campaign-(1,2] | 0.221 | -0.006 0.001 -0.127 |  
## f.campaign-(2,20] | 1.083 | 0.474 0.192 11.738 |  
## f.pdays-[0,22] | 6.580 | -4.417 0.451 -30.206 |  
## f.pdays-(22,23] | 0.239 | 0.161 0.451 30.206 |  
## f.previous-[0,0.9] | 0.543 | 0.487 0.804 43.179 |  
## f.previous-(0.9,1] | 2.787 | -2.633 0.893 -33.713 |  
## f.previous-(1,6] | 6.104 | -4.396 0.519 -26.498 |  
## y.no | 0.231 | 0.212 0.844 21.769 |  
## y.yes | 1.892 | -1.738 0.844 -21.769 |  
## Dim.2 cos2 v.test   
## f.job.Admin-Managment -0.020 0.027 -0.869 |  
## f.job.Entrep-Retired-selfEmpl 0.399 0.153 8.730 |  
## f.job.Not-working 0.143 0.043 2.353 |  
## f.job.Serv-Tech-BlueC -0.101 0.139 -5.958 |  
## f.season.Mar-May -0.288 0.132 -14.633 |  
## f.season.Jun-Aug 0.194 0.036 10.604 |  
## f.season.Sep-Dec 0.234 0.035 5.626 |  
## f.education.Basic 0.015 0.003 0.613 |  
## f.education.High School -0.066 0.180 -3.025 |  
## f.education.Professional 0.062 0.059 2.527 |  
## f.age-[18,30] -0.400 0.093 -11.108 |  
## f.age-(30,40] -0.213 0.193 -10.495 |  
## f.age-(40,50] 0.139 0.053 4.874 |  
## f.age-(50,92] 0.682 0.168 19.031 |  
## f.duration-[5,120] -0.094 0.118 -3.822 |  
## f.duration-(120,180] 0.014 0.010 0.414 |  
## f.duration-(180,300] 0.085 0.147 2.721 |  
## f.duration-(300,2.1e+03] 0.029 0.030 1.081 |  
## f.campaign-[0,1] 0.003 0.000 0.167 |  
## f.campaign-(1,2] 0.014 0.004 0.486 |  
## f.campaign-(2,20] -0.015 0.000 -0.627 |  
## f.pdays-[0,22] 4.435 0.454 51.122 |  
## f.pdays-(22,23] -0.161 0.454 -51.122 |  
## f.previous-[0,0.9] -0.181 0.111 -27.027 |  
## f.previous-(0.9,1] 0.442 0.025 9.534 |  
## f.previous-(1,6] 3.839 0.396 38.998 |  
## y.no -0.082 0.126 -14.153 |  
## y.yes 0.671 0.126 14.153 |

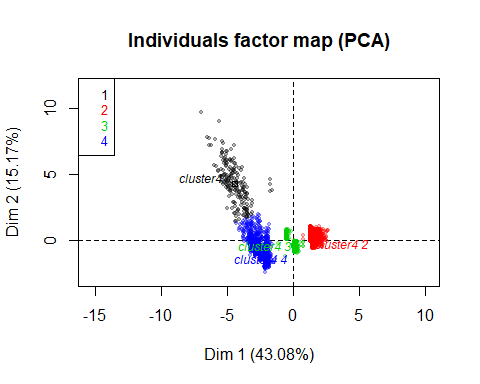
## K-Means Classification

Hemos graficado los grupos separados en 3, 4, 5 y 6 clusters, para los cuales nos parece que gráficamente con 4 clusters los grupos están bien definidos, por lo que decidimos usar 4 clusters los cuales analizaremos seguidamente.

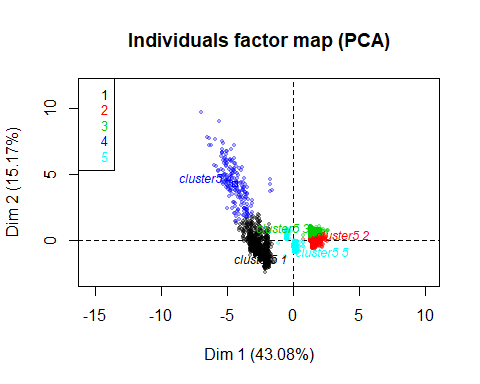
dclu<- res.pca$ind$coord[,1:2]; # los dos ejes  
  
kcla<- kmeans(dclu,4);  
  
df$cluster3 = factor(kmeans(dclu,3)$cluster);  
df$cluster4 = factor(kmeans(dclu,4)$cluster);  
df$cluster5 = factor(kmeans(dclu,5)$cluster);  
df$cluster6 = factor(kmeans(dclu,6)$cluster);  
  
res.pca<-PCA(df[,c('duration',vars\_num, "cluster3")],quanti.sup=1, quali.sup = 11, graph=FALSE)  
plot.PCA(res.pca,choix="ind",habillage=11,select=0 ,cex=0.75)



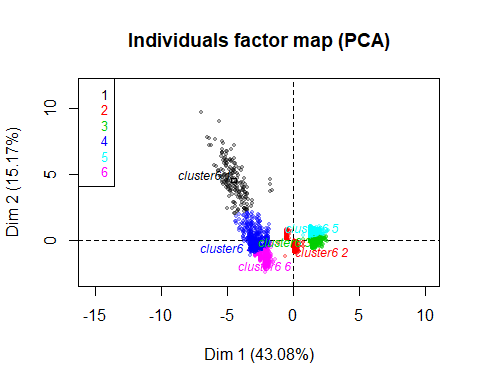
res.pca<-PCA(df[,c('duration',vars\_num, "cluster4")],quanti.sup=1, quali.sup = 11, graph=FALSE)  
plot.PCA(res.pca,choix="ind",habillage=11,select=0 ,cex=0.75)



res.pca<-PCA(df[,c('duration',vars\_num, "cluster5")],quanti.sup=1, quali.sup = 11, graph=FALSE)  
plot.PCA(res.pca,choix="ind",habillage=11,select=0 ,cex=0.75)



res.pca<-PCA(df[,c('duration',vars\_num, "cluster6")],quanti.sup=1, quali.sup = 11, graph=FALSE)  
plot.PCA(res.pca,choix="ind",habillage=11,select=0 ,cex=0.75)



df <- df[,c(1:29, 31)] # guardamos la clasificación en 4 clusters

## Description of clusters

Viendo el chi-square test, podemos saber qué variables se utilizarán para caracterizar nuestros 4 clusters. Viendo las categorías donde el P-value es casi 0 podemos ver que las categorías que cumplen estas caracteristicas son: month (y por extensión también f.season), poutcome, f.pdays, f.previous, contact, y, job(f.job), default, f.age, f.campaign, marital. Ahora veremos que categorías de estas variables son las que caracterizan estos clusters.

Para el catdes del cluster 1, hemos podido ver las categorías que mejor lo definen, la temporada de verano es la que mejor lo caracteriza, f.season.Jun-Aug, tenemos también un poutcome que nos indica que ninguno de los individuos ha sido contactado previamente, podemos ver que y.no representa el 95% de este cluster, han sido contactados en su mayoría por teléfono fijo, también podemos ver que este cluster está ligeramente relacionado con la categoría f.job.Serv-Tech-BlueC. En conclusión podemos decir que este cluster está caracterizado por: - Meses de jun-Ago - No han sido contactados previamente - Han sido contactados más de una vez en la campaña actual - Contactados por teléfono fijo - Trabajo normalmente es, servicio, técnicos o blue collar. - No compraron el producto en su mayoría.

Para el cluster 2, tenemos: - La temporada de mar-may están sobrerepresentadas en este cluster - Han sido contacados en su mayoría por teléfono móvil - No han comprado el producto en campañas anteriores - Han sido contacados en camapañas previas - La categoría student está sobrerepresentada - La aceptación del producto está sobrerepresentada también

Para el cluster 3, tenemos: - Temporada de Sep-Dec - Contactados por móvil en su mayoría - La categoría de job.management está sobrerepresentada - No han adquirido el producto - Una gran cantidad de individuos rechazó el producto (y.no)

Para el cluster 4, podemos ver que aglutina individuos muy bien caracterizados por las siguientes variables: - Han sido contactados previamente f.pdays[0,22] - Han comprado el producto en una campaña previa - Han comprado el producto y.yes - Temporada de Sep-Dec - Han sido contactados por móvil - Una parte importante son job.retired - Una edad de f.age-(50,92]

catdes(df, 30, proba = 0.001)

##   
## Link between the cluster variable and the categorical variables (chi-square test)  
## =================================================================================  
## p.value df  
## month 0.000000e+00 27  
## poutcome 0.000000e+00 6  
## f.season 0.000000e+00 6  
## f.pdays 0.000000e+00 3  
## f.previous 0.000000e+00 6  
## contact 4.750967e-213 3  
## y 3.629487e-144 3  
## job 4.108692e-67 30  
## default 6.986972e-43 3  
## f.age 1.030850e-26 9  
## f.campaign 6.092760e-24 6  
## f.job 1.806095e-19 9  
## marital 1.705029e-16 6  
## f.duration 1.919242e-10 9  
## education 5.661192e-09 18  
## f.education 8.669827e-09 6  
## housing 1.403613e-07 3  
## day\_of\_week 4.039589e-04 12  
##   
## Description of each cluster by the categories  
## =============================================  
## $`1`  
## Cla/Mod Mod/Cla Global  
## f.pdays=f.pdays-[0,22] 96.5714286 88.947368 3.507014  
## poutcome=poutcome.success 96.1538462 78.947368 3.126253  
## f.previous=f.previous-(1,6] 75.1824818 54.210526 2.745491  
## y=y.yes 21.0332103 60.000000 10.861723  
## f.previous=f.previous-(0.9,1] 15.4255319 45.789474 11.302605  
## f.season=f.season.Sep-Dec 11.1600587 40.000000 13.647295  
## month=month.oct 32.5000000 13.684211 1.603206  
## contact=contact.cellular 5.4011864 91.052632 64.188377  
## default=default.no 4.6954315 97.368421 78.957916  
## job=job.retired 15.7657658 18.421053 4.448898  
## month=month.sep 25.7142857 9.473684 1.402806  
## f.age=f.age-(50,92] 7.4200913 34.210526 17.555110  
## month=month.mar 20.6349206 6.842105 1.262525  
## job=job.student 15.5963303 8.947368 2.184369  
## f.job=f.job.Entrep-Retired-selfEmpl 7.6256499 23.157895 11.563126  
## f.duration=f.duration-(180,300] 6.0550459 34.736842 21.843687  
## poutcome=poutcome.failure 7.3394495 21.052632 10.921844  
## f.education=f.education.Professional 5.4347826 44.736842 31.342685  
## education=education.university.degree 5.4347826 44.736842 31.342685  
## f.job=f.job.Not-working 8.1871345 14.736842 6.853707  
## f.campaign=f.campaign-[0,1] 4.9504950 55.263158 42.505010  
## month=month.jun 1.6233766 5.263158 12.344689  
## f.education=f.education.Basic 2.5046963 21.052632 32.004008  
## f.season=f.season.Mar-May 2.7131783 29.473684 41.362725  
## marital=marital.married 3.0040053 47.368421 60.040080  
## f.age=f.age-(40,50] 1.9952115 13.157895 25.110220  
## f.season=f.season.Jun-Aug 2.5835189 30.526316 44.989980  
## education=education.basic.9y 1.3580247 5.789474 16.232465  
## f.campaign=f.campaign-(2,20] 2.0496894 17.368421 32.264529  
## month=month.jul 1.1560694 5.263158 17.334669  
## month=month.may 1.8452381 16.315789 33.667335  
## f.duration=f.duration-[5,120] 1.7341040 14.210526 31.202405  
## f.job=f.job.Serv-Tech-BlueC 1.9801980 25.263158 48.577154  
## job=job.blue-collar 0.8650519 5.263158 23.166333  
## default=default.unknown 0.4761905 2.631579 21.042084  
## contact=contact.telephone 0.9513151 8.947368 35.811623  
## y=y.no 1.7086331 40.000000 89.138277  
## f.previous=f.previous-[0,0.9] 0.0000000 0.000000 85.951904  
## poutcome=poutcome.nonexistent 0.0000000 0.000000 85.951904  
## f.pdays=f.pdays-(22,23] 0.4361371 11.052632 96.492986  
## p.value v.test  
## f.pdays=f.pdays-[0,22] 1.259643e-281 35.860546  
## poutcome=poutcome.success 3.796871e-240 33.092604  
## f.previous=f.previous-(1,6] 1.553305e-129 24.214815  
## y=y.yes 2.494385e-64 16.934745  
## f.previous=f.previous-(0.9,1] 1.156226e-34 12.280262  
## f.season=f.season.Sep-Dec 3.391033e-20 9.205810  
## month=month.oct 3.163903e-18 8.705424  
## contact=contact.cellular 4.127348e-18 8.675225  
## default=default.no 4.787962e-14 7.537574  
## job=job.retired 2.103831e-13 7.342027  
## month=month.sep 5.701121e-11 6.551371  
## f.age=f.age-(50,92] 1.767166e-08 5.633375  
## month=month.mar 5.214842e-07 5.018235  
## job=job.student 6.469157e-07 4.976658  
## f.job=f.job.Entrep-Retired-selfEmpl 4.357820e-06 4.593544  
## f.duration=f.duration-(180,300] 3.266772e-05 4.154025  
## poutcome=poutcome.failure 3.513520e-05 4.137342  
## f.education=f.education.Professional 8.193961e-05 3.938655  
## education=education.university.degree 8.193961e-05 3.938655  
## f.job=f.job.Not-working 1.066886e-04 3.874854  
## f.campaign=f.campaign-[0,1] 3.261253e-04 3.593612  
## month=month.jun 9.887768e-04 -3.293701  
## f.education=f.education.Basic 6.774908e-04 -3.398530  
## f.season=f.season.Mar-May 5.767248e-04 -3.442331  
## marital=marital.married 3.331185e-04 -3.588083  
## f.age=f.age-(40,50] 4.090062e-05 -4.102332  
## f.season=f.season.Jun-Aug 3.480205e-05 -4.139528  
## education=education.basic.9y 1.044307e-05 -4.407790  
## f.campaign=f.campaign-(2,20] 2.688717e-06 -4.693270  
## month=month.jul 4.872344e-07 -5.031273  
## month=month.may 5.199683e-08 -5.444343  
## f.duration=f.duration-[5,120] 4.101428e-08 -5.486428  
## f.job=f.job.Serv-Tech-BlueC 2.251008e-11 -6.688740  
## job=job.blue-collar 1.081400e-11 -6.795230  
## default=default.unknown 4.787962e-14 -7.537574  
## contact=contact.telephone 4.127348e-18 -8.675225  
## y=y.no 2.494385e-64 -16.934745  
## f.previous=f.previous-[0,0.9] 2.196177e-173 -28.071291  
## poutcome=poutcome.nonexistent 2.196177e-173 -28.071291  
## f.pdays=f.pdays-(22,23] 1.259643e-281 -35.860546  
##   
## $`2`  
## Cla/Mod Mod/Cla Global  
## f.season=f.season.Jun-Aug 87.0824053 68.47635727 44.9899800  
## f.previous=f.previous-[0,0.9] 66.5656330 100.00000000 85.9519038  
## poutcome=poutcome.nonexistent 66.5656330 100.00000000 85.9519038  
## contact=contact.telephone 86.6256295 54.22066550 35.8116232  
## month=month.jul 93.5260116 28.33625219 17.3346693  
## f.pdays=f.pdays-(22,23] 59.2938733 100.00000000 96.4929860  
## month=month.aug 82.3298429 22.03152364 15.3106212  
## y=y.no 61.0161871 95.06129597 89.1382766  
## month=month.jun 83.9285714 18.10858144 12.3446894  
## default=default.unknown 75.8095238 27.88091068 21.0420842  
## f.campaign=f.campaign-(2,20] 67.0807453 37.82837128 32.2645291  
## f.age=f.age-(40,50] 65.5227454 28.75656743 25.1102204  
## f.job=f.job.Serv-Tech-BlueC 62.1699670 52.78458844 48.5771543  
## marital=marital.married 60.8144192 63.81786340 60.0400802  
## housing=housing.no 61.6498464 49.21190893 45.6713427  
## job=job.technician 63.7698898 18.24868651 16.3727455  
## job=job.blue-collar 62.2837370 25.21891419 23.1663327  
## f.education=f.education.Basic 61.1145899 34.18563923 32.0040080  
## job=job.housemaid 72.2689076 3.01225919 2.3847695  
## f.education=f.education.Professional 53.7084399 29.42206655 31.3426854  
## education=education.university.degree 53.7084399 29.42206655 31.3426854  
## f.job=f.job.Not-working 47.6608187 5.70928196 6.8537074  
## month=month.may 53.4523810 31.45359019 33.6673347  
## f.job=f.job.Entrep-Retired-selfEmpl 48.5268631 9.80735552 11.5631263  
## job=job.retired 41.8918919 3.25744308 4.4488978  
## month=month.dec 0.0000000 0.00000000 0.3607214  
## f.age=f.age-[18,30] 48.6206897 14.81611208 17.4348697  
## housing=housing.yes 53.4857986 50.78809107 54.3286573  
## marital=marital.single 49.8936924 24.65849387 28.2765531  
## job=job.student 14.6788991 0.56042032 2.1843687  
## f.campaign=f.campaign-[0,1] 49.6463932 36.88266200 42.5050100  
## month=month.mar 0.0000000 0.00000000 1.2625251  
## month=month.oct 2.5000000 0.07005254 1.6032064  
## month=month.sep 0.0000000 0.00000000 1.4028056  
## default=default.no 52.2588832 72.11908932 78.9579158  
## f.previous=f.previous-(1,6] 0.0000000 0.00000000 2.7454910  
## y=y.yes 26.0147601 4.93870403 10.8617234  
## poutcome=poutcome.success 0.0000000 0.00000000 3.1262525  
## f.season=f.season.Mar-May 43.5077519 31.45359019 41.3627255  
## f.pdays=f.pdays-[0,22] 0.0000000 0.00000000 3.5070140  
## month=month.apr 0.0000000 0.00000000 6.4328657  
## month=month.nov 0.0000000 0.00000000 10.2805611  
## poutcome=poutcome.failure 0.0000000 0.00000000 10.9218437  
## f.previous=f.previous-(0.9,1] 0.0000000 0.00000000 11.3026052  
## contact=contact.cellular 40.8054948 45.77933450 64.1883768  
## f.season=f.season.Sep-Dec 0.2936858 0.07005254 13.6472946  
## p.value v.test  
## f.season=f.season.Jun-Aug 0.000000e+00 Inf  
## f.previous=f.previous-[0,0.9] 3.019215e-293 36.598522  
## poutcome=poutcome.nonexistent 3.019215e-293 36.598522  
## contact=contact.telephone 1.051844e-235 32.782323  
## month=month.jul 1.306801e-149 26.051114  
## f.pdays=f.pdays-(22,23] 4.293494e-67 17.305259  
## month=month.aug 3.823967e-57 15.931608  
## y=y.no 8.988442e-55 15.586531  
## month=month.jun 2.951209e-51 15.060372  
## default=default.unknown 8.402225e-45 14.043849  
## f.campaign=f.campaign-(2,20] 1.196425e-22 9.793857  
## f.age=f.age-(40,50] 4.665985e-12 6.915384  
## f.job=f.job.Serv-Tech-BlueC 5.923144e-12 6.881486  
## marital=marital.married 3.139148e-10 6.291746  
## housing=housing.no 6.199688e-09 5.811280  
## job=job.technician 3.130573e-05 4.163754  
## job=job.blue-collar 6.671792e-05 3.987697  
## f.education=f.education.Basic 1.296174e-04 3.827184  
## job=job.housemaid 6.364492e-04 3.415586  
## f.education=f.education.Professional 7.361711e-04 -3.375741  
## education=education.university.degree 7.361711e-04 -3.375741  
## f.job=f.job.Not-working 2.382353e-04 -3.674586  
## month=month.may 1.335878e-04 -3.819748  
## f.job=f.job.Entrep-Retired-selfEmpl 8.231372e-06 -4.459076  
## job=job.retired 2.872930e-06 -4.679701  
## month=month.dec 2.215519e-07 -5.180281  
## f.age=f.age-[18,30] 2.026128e-08 -5.609755  
## housing=housing.yes 6.199688e-09 -5.811280  
## marital=marital.single 6.116596e-11 -6.540860  
## job=job.student 2.122476e-20 -9.255993  
## f.campaign=f.campaign-[0,1] 1.628849e-20 -9.284227  
## month=month.mar 3.477079e-24 -10.145300  
## month=month.oct 8.502567e-27 -10.716652  
## month=month.sep 8.041008e-27 -10.721815  
## default=default.no 8.402225e-45 -14.043849  
## f.previous=f.previous-(1,6] 2.340302e-52 -15.227016  
## y=y.yes 8.988442e-55 -15.586531  
## poutcome=poutcome.success 1.057860e-59 -16.295757  
## f.season=f.season.Mar-May 8.078887e-61 -16.452266  
## f.pdays=f.pdays-[0,22] 4.293494e-67 -17.305259  
## month=month.apr 1.591937e-125 -23.831058  
## month=month.nov 3.512409e-207 -30.715702  
## poutcome=poutcome.failure 2.239752e-221 -31.760980  
## f.previous=f.previous-(0.9,1] 6.871049e-230 -32.371692  
## contact=contact.cellular 1.051844e-235 -32.782323  
## f.season=f.season.Sep-Dec 8.033924e-278 -35.615604  
##   
## $`3`  
## Cla/Mod Mod/Cla Global  
## f.season=f.season.Sep-Dec 66.079295 100.000000 13.647295  
## month=month.nov 86.549708 98.666667 10.280561  
## contact=contact.cellular 12.425851 88.444444 64.188377  
## job=job.management 17.948718 14.000000 7.034068  
## f.pdays=f.pdays-(22,23] 9.345794 100.000000 96.492986  
## default=default.no 10.076142 88.222222 78.957916  
## y=y.no 9.622302 95.111111 89.138277  
## poutcome=poutcome.failure 14.311927 17.333333 10.921844  
## f.campaign=f.campaign-[0,1] 11.032532 52.000000 42.505010  
## f.duration=f.duration-[5,120] 11.560694 40.000000 31.202405  
## f.previous=f.previous-(0.9,1] 13.829787 17.333333 11.302605  
## f.education=f.education.Professional 11.125320 38.666667 31.342685  
## education=education.university.degree 11.125320 38.666667 31.342685  
## job=job.entrepreneur 16.402116 6.888889 3.787575  
## education=education.basic.4y 5.048544 5.777778 10.320641  
## job=job.retired 2.702703 1.333333 4.448898  
## f.age=f.age-[18,30] 5.747126 11.111111 17.434870  
## job=job.student 0.000000 0.000000 2.184369  
## job=job.blue-collar 5.795848 14.888889 23.166333  
## y=y.yes 4.059041 4.888889 10.861723  
## f.previous=f.previous-(1,6] 0.000000 0.000000 2.745491  
## f.education=f.education.Basic 6.199123 22.000000 32.004008  
## f.campaign=f.campaign-(2,20] 6.211180 22.222222 32.264529  
## poutcome=poutcome.success 0.000000 0.000000 3.126253  
## default=default.unknown 5.047619 11.777778 21.042084  
## f.pdays=f.pdays-[0,22] 0.000000 0.000000 3.507014  
## month=month.apr 0.000000 0.000000 6.432866  
## month=month.jun 0.000000 0.000000 12.344689  
## contact=contact.telephone 2.909905 11.555556 35.811623  
## month=month.aug 0.000000 0.000000 15.310621  
## month=month.jul 0.000000 0.000000 17.334669  
## month=month.may 0.000000 0.000000 33.667335  
## f.season=f.season.Mar-May 0.000000 0.000000 41.362725  
## f.season=f.season.Jun-Aug 0.000000 0.000000 44.989980  
## p.value v.test  
## f.season=f.season.Sep-Dec 0.000000e+00 Inf  
## month=month.nov 0.000000e+00 Inf  
## contact=contact.cellular 4.457861e-34 12.170601  
## job=job.management 4.606444e-08 5.465869  
## f.pdays=f.pdays-(22,23] 4.814305e-08 5.458036  
## default=default.no 8.895002e-08 5.347962  
## y=y.no 3.321039e-06 4.649895  
## poutcome=poutcome.failure 1.761402e-05 4.293170  
## f.campaign=f.campaign-[0,1] 2.207930e-05 4.242754  
## f.duration=f.duration-[5,120] 3.465489e-05 4.140500  
## f.previous=f.previous-(0.9,1] 6.332305e-05 4.000073  
## f.education=f.education.Professional 5.526671e-04 3.453839  
## education=education.university.degree 5.526671e-04 3.453839  
## job=job.entrepreneur 9.628257e-04 3.301170  
## education=education.basic.4y 4.140424e-04 -3.530968  
## job=job.retired 1.504256e-04 -3.790366  
## f.age=f.age-[18,30] 1.055513e-04 -3.877463  
## job=job.student 2.982402e-05 -4.174806  
## job=job.blue-collar 5.480188e-06 -4.545512  
## y=y.yes 3.321039e-06 -4.649895  
## f.previous=f.previous-(1,6] 1.972148e-06 -4.756258  
## f.education=f.education.Basic 9.408844e-07 -4.903616  
## f.campaign=f.campaign-(2,20] 9.107926e-07 -4.909994  
## poutcome=poutcome.success 3.093148e-07 -5.117681  
## default=default.unknown 8.895002e-08 -5.347962  
## f.pdays=f.pdays-[0,22] 4.814305e-08 -5.458036  
## month=month.apr 2.294108e-14 -7.632967  
## month=month.jun 8.435071e-28 -10.928369  
## contact=contact.telephone 4.457861e-34 -12.170601  
## month=month.aug 6.622682e-35 -12.325263  
## month=month.jul 6.599006e-40 -13.221438  
## month=month.may 8.848336e-86 -19.628385  
## f.season=f.season.Mar-May 8.311269e-112 -22.469179  
## f.season=f.season.Jun-Aug 2.151192e-125 -23.818443  
##   
## $`4`  
## Cla/Mod Mod/Cla Global  
## f.season=f.season.Mar-May 53.779070 74.2474916 41.3627255  
## month=month.apr 96.261682 20.6688963 6.4328657  
## contact=contact.cellular 41.367468 88.6287625 64.1883768  
## poutcome=poutcome.failure 78.348624 28.5618729 10.9218437  
## f.previous=f.previous-(0.9,1] 70.744681 26.6889632 11.3026052  
## month=month.may 44.702381 50.2341137 33.6673347  
## y=y.yes 48.892989 17.7257525 10.8617234  
## default=default.no 32.969543 86.8896321 78.9579158  
## f.pdays=f.pdays-(22,23] 30.924195 99.5986622 96.4929860  
## job=job.student 69.724771 5.0836120 2.1843687  
## month=month.mar 79.365079 3.3444816 1.2625251  
## month=month.sep 74.285714 3.4782609 1.4028056  
## marital=marital.single 37.774628 35.6521739 28.2765531  
## f.age=f.age-[18,30] 40.574713 23.6120401 17.4348697  
## f.campaign=f.campaign-[0,1] 34.370580 48.7625418 42.5050100  
## month=month.oct 60.000000 3.2107023 1.6032064  
## housing=housing.yes 32.312800 58.5953177 54.3286573  
## month=month.dec 72.222222 0.8695652 0.3607214  
## housing=housing.no 27.161036 41.4046823 45.6713427  
## f.season=f.season.Sep-Dec 22.466960 10.2341137 13.6472946  
## job=job.technician 22.888617 12.5083612 16.3727455  
## marital=marital.married 27.202937 54.5150502 60.0400802  
## f.campaign=f.campaign-(2,20] 24.658385 26.5551839 32.2645291  
## f.age=f.age-(40,50] 22.984836 19.2642140 25.1102204  
## poutcome=poutcome.success 3.846154 0.4013378 3.1262525  
## f.pdays=f.pdays-[0,22] 3.428571 0.4013378 3.5070140  
## default=default.unknown 18.666667 13.1103679 21.0420842  
## month=month.jun 14.448052 5.9531773 12.3446894  
## y=y.no 27.652878 82.2742475 89.1382766  
## month=month.aug 12.696335 6.4882943 15.3106212  
## month=month.nov 7.797271 2.6755853 10.2805611  
## f.previous=f.previous-[0,0.9] 24.761017 71.0367893 85.9519038  
## poutcome=poutcome.nonexistent 24.761017 71.0367893 85.9519038  
## month=month.jul 5.317919 3.0769231 17.3346693  
## contact=contact.telephone 9.513151 11.3712375 35.8116232  
## f.season=f.season.Jun-Aug 10.334076 15.5183946 44.9899800  
## p.value v.test  
## f.season=f.season.Mar-May 1.214684e-211 31.047983  
## month=month.apr 2.443356e-153 26.378112  
## contact=contact.cellular 7.335999e-138 24.992718  
## poutcome=poutcome.failure 5.641457e-137 24.911094  
## f.previous=f.previous-(0.9,1] 3.761358e-101 21.351684  
## month=month.may 2.068828e-57 15.969971  
## y=y.yes 9.791179e-23 9.814096  
## default=default.no 1.504249e-20 9.292699  
## f.pdays=f.pdays-(22,23] 7.889588e-20 9.114690  
## job=job.student 6.601535e-18 8.621617  
## month=month.mar 5.354652e-16 8.103167  
## month=month.sep 1.931681e-14 7.655096  
## marital=marital.single 7.928406e-14 7.471503  
## f.age=f.age-[18,30] 1.851132e-13 7.359130  
## f.campaign=f.campaign-[0,1] 5.524740e-09 5.830543  
## month=month.oct 2.441289e-08 5.577408  
## housing=housing.yes 7.419015e-05 3.962432  
## month=month.dec 2.959629e-04 3.618808  
## housing=housing.no 7.419015e-05 -3.962432  
## f.season=f.season.Sep-Dec 2.698987e-06 -4.692491  
## job=job.technician 8.712805e-07 -4.918684  
## marital=marital.married 2.090160e-07 -5.191135  
## f.campaign=f.campaign-(2,20] 1.226294e-08 -5.696030  
## f.age=f.age-(40,50] 2.434489e-10 -6.331081  
## poutcome=poutcome.success 4.435265e-17 -8.400779  
## f.pdays=f.pdays-[0,22] 7.889588e-20 -9.114690  
## default=default.unknown 1.504249e-20 -9.292699  
## month=month.jun 1.863347e-21 -9.512388  
## y=y.no 9.791179e-23 -9.814096  
## month=month.aug 1.221397e-33 -12.088058  
## month=month.nov 8.917776e-38 -12.847202  
## f.previous=f.previous-[0,0.9] 4.885281e-80 -18.944683  
## poutcome=poutcome.nonexistent 4.885281e-80 -18.944683  
## month=month.jul 1.393898e-85 -19.605278  
## contact=contact.telephone 7.335999e-138 -24.992718  
## f.season=f.season.Jun-Aug 1.401958e-178 -28.493704  
##   
##   
## Link between the cluster variable and the quantitative variables  
## ================================================================  
## Eta2 P-value  
## pdays 0.833431759 0.000000e+00  
## previous 0.501141537 0.000000e+00  
## emp.var.rate 0.949447313 0.000000e+00  
## cons.price.idx 0.574427755 0.000000e+00  
## euribor3m 0.989874994 0.000000e+00  
## nr.employed 0.866883329 0.000000e+00  
## cons.conf.idx 0.167721928 3.916150e-198  
## campaign 0.027687626 3.757869e-30  
## age 0.008632187 2.197904e-09  
##   
## Description of each cluster by quantitative variables  
## =====================================================  
## $`1`  
## v.test Mean in category Overall mean sd in category  
## previous 45.971997 1.8210526 0.17855711 0.9567373  
## cons.conf.idx 6.937479 -38.2494737 -40.54192385 7.1162952  
## age 5.645279 44.3789474 40.17755511 16.9858634  
## campaign -4.322398 1.7894737 2.51503006 1.2598641  
## cons.price.idx -4.817260 93.3651368 93.56373427 0.8341906  
## emp.var.rate -19.115707 -2.0936842 0.05212425 0.8565304  
## euribor3m -21.054798 0.9672263 3.58457355 0.5121394  
## nr.employed -27.122883 5025.3821053 5165.87569138 51.3340730  
## pdays -64.481704 7.8473684 22.41362725 6.4451425  
## Overall sd p.value  
## previous 0.5020810 0.000000e+00  
## cons.conf.idx 4.6436681 3.991593e-12  
## age 10.4585324 1.649135e-08  
## campaign 2.3588988 1.543422e-05  
## cons.price.idx 0.5793439 1.455430e-06  
## emp.var.rate 1.5774788 1.868722e-81  
## euribor3m 1.7469207 2.066633e-98  
## nr.employed 72.7919889 5.290316e-162  
## pdays 3.1744936 0.000000e+00  
##   
## $`2`  
## v.test Mean in category Overall mean sd in category  
## emp.var.rate 64.85040 1.304588 0.05212425 0.14415633  
## euribor3m 62.64096 4.924314 3.58457355 0.05025276  
## nr.employed 56.70185 5216.408091 5165.87569138 17.23254124  
## cons.price.idx 52.54683 93.936445 93.56373427 0.32429720  
## cons.conf.idx 25.36681 -39.099755 -40.54192385 3.04458233  
## pdays 15.08725 23.000000 22.41362725 0.00000000  
## campaign 11.60168 2.850088 2.51503006 2.73344964  
## previous -29.04787 0.000000 0.17855711 0.00000000  
## Overall sd p.value  
## emp.var.rate 1.5774788 0.000000e+00  
## euribor3m 1.7469207 0.000000e+00  
## nr.employed 72.7919889 0.000000e+00  
## cons.price.idx 0.5793439 0.000000e+00  
## cons.conf.idx 4.6436681 5.863605e-142  
## pdays 3.1744936 1.964506e-51  
## campaign 2.3588988 4.040627e-31  
## previous 0.5020810 1.637199e-185  
##   
## $`3`  
## v.test Mean in category Overall mean sd in category  
## nr.employed 9.115196 5195.713333 5165.875691 1.29710789  
## euribor3m 6.818350 4.120207 3.584574 0.11552217  
## pdays 4.107565 23.000000 22.413627 0.00000000  
## campaign -4.708571 2.015556 2.515030 1.58527328  
## cons.conf.idx -6.997289 -42.003111 -40.541924 0.30057591  
## cons.price.idx -13.833237 93.203342 93.563734 0.06359003  
## Overall sd p.value  
## nr.employed 72.7919889 7.852866e-20  
## euribor3m 1.7469207 9.209193e-12  
## pdays 3.1744936 3.998529e-05  
## campaign 2.3588988 2.494597e-06  
## cons.conf.idx 4.6436681 2.609633e-12  
## cons.price.idx 0.5793439 1.606263e-43  
##   
## $`4`  
## v.test Mean in category Overall mean sd in category  
## previous 12.312403 0.3123746 0.17855711 0.5101781  
## pdays 8.075532 22.9685619 22.41362725 0.5687059  
## age -4.052007 39.2602007 40.17755511 12.0436159  
## campaign -7.780683 2.1177258 2.51503006 1.6827435  
## cons.conf.idx -25.921260 -43.1475585 -40.54192385 5.9781301  
## cons.price.idx -46.092491 92.9856876 93.56373427 0.4315899  
## nr.employed -55.611005 5078.2481605 5165.87569138 38.1020523  
## emp.var.rate -60.712801 -2.0210702 0.05212425 0.5356747  
## euribor3m -63.124684 1.1974843 3.58457355 0.2469009  
## Overall sd p.value  
## previous 0.5020810 7.767580e-35  
## pdays 3.1744936 6.718328e-16  
## age 10.4585324 5.078016e-05  
## campaign 2.3588988 7.213395e-15  
## cons.conf.idx 4.6436681 3.835830e-148  
## cons.price.idx 0.5793439 0.000000e+00  
## nr.employed 72.7919889 0.000000e+00  
## emp.var.rate 1.5774788 0.000000e+00  
## euribor3m 1.7469207 0.000000e+00

## Hierarchical Clustering

Al hacer el HCPC podemos ver que el gráfico de ganancia de inercia nos da la mayoría en dos variables, y luego dos picos más pequeños.

Ahora vemos el clustering no supervisado. Vamos a clasificar estos clusters.

Para el cluster 1:  
- Está caracterizado por personas que han sido contactados previamente  
- Han aceptado el producto  
- Se han contactado en f.season.Sep-Dec  
- Tienen una sobrerepresentación de f.job.Entrep-Retired-selfEmpl  
- Llamadas de duración mayor a 3min

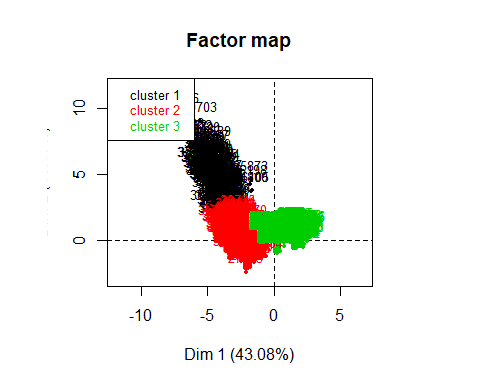
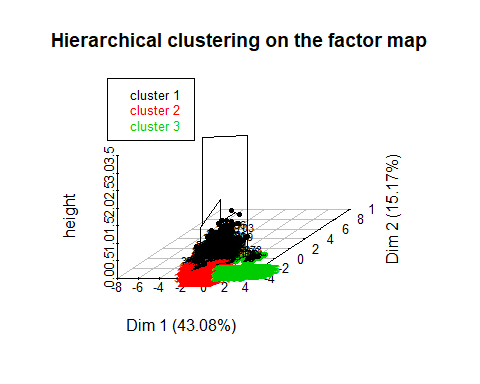
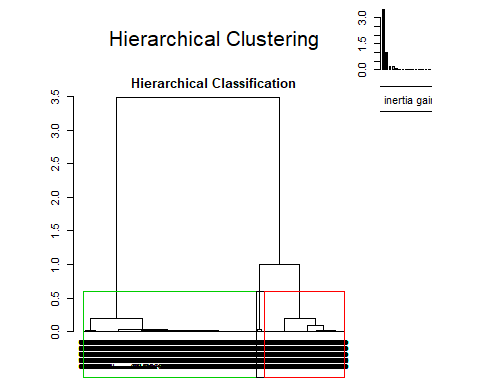
Para el cluster 2:  
- Han sido contactados f.season.Mar-May  
- Han sido contactados en campañas previas  
- Han aceptado el producto (y.yes)

Para el cluster 3:  
- Han sido contactados previamente  
- No han sido contactados en campañas previas  
- Han sido contactados en la temporada de f.season.Jun-Aug  
- Han rechazado el producto  
- Tienen una leve representación de f.job.Serv-Tech-BlueC

Además vemos que los parengones tienen mucho parecido a lo descrito para cada cluster.

Utilizamos el atributo nb.clust=3, después de haber visto que era el mejor corte.

res.pca<-PCA(df[,c('duration',vars\_num,vars\_factorizadas)],quanti.sup=1, quali.sup = c(11:19), ncp=2, graph=FALSE)  
res.hcpc<-HCPC(res.pca,order=TRUE, nb.clust = 3)



attributes(res.hcpc)

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

summary(res.hcpc$data.clust)

## duration age campaign pdays   
## Min. : 5.0 Min. :18.00 Min. : 1.000 Min. : 0.00   
## 1st Qu.: 103.0 1st Qu.:32.00 1st Qu.: 1.000 1st Qu.:23.00   
## Median : 178.5 Median :38.00 Median : 2.000 Median :23.00   
## Mean : 259.9 Mean :40.18 Mean : 2.515 Mean :22.41   
## 3rd Qu.: 321.0 3rd Qu.:47.00 3rd Qu.: 3.000 3rd Qu.:23.00   
## Max. :2078.0 Max. :92.00 Max. :20.000 Max. :23.00   
## previous emp.var.rate cons.price.idx cons.conf.idx   
## Min. :0.0000 Min. :-3.40000 Min. :92.20 Min. :-50.80   
## 1st Qu.:0.0000 1st Qu.:-1.80000 1st Qu.:93.08 1st Qu.:-42.70   
## Median :0.0000 Median : 1.10000 Median :93.44 Median :-41.80   
## Mean :0.1786 Mean : 0.05212 Mean :93.56 Mean :-40.54   
## 3rd Qu.:0.0000 3rd Qu.: 1.40000 3rd Qu.:93.99 3rd Qu.:-36.40   
## Max. :6.0000 Max. : 1.40000 Max. :94.77 Max. :-26.90   
## euribor3m nr.employed f.job   
## Min. :0.634 Min. :4964 f.job.Admin-Managment :1647   
## 1st Qu.:1.334 1st Qu.:5099 f.job.Entrep-Retired-selfEmpl: 577   
## Median :4.857 Median :5191 f.job.Not-working : 342   
## Mean :3.585 Mean :5166 f.job.Serv-Tech-BlueC :2424   
## 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :5.045 Max. :5228   
## f.season f.education   
## f.season.Mar-May:2064 f.education.Basic :1597   
## f.season.Jun-Aug:2245 f.education.High School :1829   
## f.season.Sep-Dec: 681 f.education.Professional:1564   
##   
##   
##   
## f.age f.duration   
## f.age-[18,30]: 870 f.duration-[5,120] :1557   
## f.age-(30,40]:1991 f.duration-(120,180] : 966   
## f.age-(40,50]:1253 f.duration-(180,300] :1090   
## f.age-(50,92]: 876 f.duration-(300,2.1e+03]:1377   
##   
##   
## f.campaign f.pdays f.previous   
## f.campaign-[0,1] :2121 f.pdays-[0,22] : 175 f.previous-[0,0.9]:4289   
## f.campaign-(1,2] :1259 f.pdays-(22,23]:4815 f.previous-(0.9,1]: 564   
## f.campaign-(2,20]:1610 f.previous-(1,6] : 137   
##   
##   
##   
## y clust   
## y.no :4448 1: 190   
## y.yes: 542 2:1497   
## 3:3303   
##   
##   
##

attributes(res.hcpc$desc.var)

## $names  
## [1] "test.chi2" "category" "quanti.var" "quanti" "call"   
##   
## $class  
## [1] "catdes" "list "

# Factors globally related to clustering partition  
res.hcpc$desc.var$test.chi2

## p.value df  
## f.pdays 0.000000e+00 2  
## f.previous 0.000000e+00 4  
## f.season 3.418306e-239 4  
## y 2.373371e-145 2  
## f.age 4.054913e-27 6  
## f.job 5.294443e-17 6  
## f.campaign 6.236440e-15 4  
## f.duration 2.126018e-09 6  
## f.education 7.523481e-04 4

# Categories over/under represented in each cluster  
res.hcpc$desc.var$category

## $`1`  
## Cla/Mod Mod/Cla Global  
## f.pdays=f.pdays-[0,22] 96.5714286 88.94737 3.507014  
## f.previous=f.previous-(1,6] 75.1824818 54.21053 2.745491  
## y=y.yes 21.0332103 60.00000 10.861723  
## f.previous=f.previous-(0.9,1] 15.4255319 45.78947 11.302605  
## f.season=f.season.Sep-Dec 11.1600587 40.00000 13.647295  
## f.age=f.age-(50,92] 7.4200913 34.21053 17.555110  
## f.job=f.job.Entrep-Retired-selfEmpl 7.6256499 23.15789 11.563126  
## f.duration=f.duration-(180,300] 6.0550459 34.73684 21.843687  
## f.education=f.education.Professional 5.4347826 44.73684 31.342685  
## f.job=f.job.Not-working 8.1871345 14.73684 6.853707  
## f.campaign=f.campaign-[0,1] 4.9504950 55.26316 42.505010  
## f.duration=f.duration-(300,2.1e+03] 4.9382716 35.78947 27.595190  
## f.age=f.age-[18,30] 5.0574713 23.15789 17.434870  
## f.age=f.age-(30,40] 2.8126570 29.47368 39.899800  
## f.education=f.education.Basic 2.5046963 21.05263 32.004008  
## f.season=f.season.Mar-May 2.7131783 29.47368 41.362725  
## f.age=f.age-(40,50] 1.9952115 13.15789 25.110220  
## f.season=f.season.Jun-Aug 2.5835189 30.52632 44.989980  
## f.campaign=f.campaign-(2,20] 2.0496894 17.36842 32.264529  
## f.duration=f.duration-[5,120] 1.7341040 14.21053 31.202405  
## f.job=f.job.Serv-Tech-BlueC 1.9801980 25.26316 48.577154  
## y=y.no 1.7086331 40.00000 89.138277  
## f.previous=f.previous-[0,0.9] 0.0000000 0.00000 85.951904  
## f.pdays=f.pdays-(22,23] 0.4361371 11.05263 96.492986  
## p.value v.test  
## f.pdays=f.pdays-[0,22] 1.259643e-281 35.860546  
## f.previous=f.previous-(1,6] 1.553305e-129 24.214815  
## y=y.yes 2.494385e-64 16.934745  
## f.previous=f.previous-(0.9,1] 1.156226e-34 12.280262  
## f.season=f.season.Sep-Dec 3.391033e-20 9.205810  
## f.age=f.age-(50,92] 1.767166e-08 5.633375  
## f.job=f.job.Entrep-Retired-selfEmpl 4.357820e-06 4.593544  
## f.duration=f.duration-(180,300] 3.266772e-05 4.154025  
## f.education=f.education.Professional 8.193961e-05 3.938655  
## f.job=f.job.Not-working 1.066886e-04 3.874854  
## f.campaign=f.campaign-[0,1] 3.261253e-04 3.593612  
## f.duration=f.duration-(300,2.1e+03] 1.193267e-02 2.514129  
## f.age=f.age-[18,30] 4.007463e-02 2.052979  
## f.age=f.age-(30,40] 2.422784e-03 -3.032822  
## f.education=f.education.Basic 6.774908e-04 -3.398530  
## f.season=f.season.Mar-May 5.767248e-04 -3.442331  
## f.age=f.age-(40,50] 4.090062e-05 -4.102332  
## f.season=f.season.Jun-Aug 3.480205e-05 -4.139528  
## f.campaign=f.campaign-(2,20] 2.688717e-06 -4.693270  
## f.duration=f.duration-[5,120] 4.101428e-08 -5.486428  
## f.job=f.job.Serv-Tech-BlueC 2.251008e-11 -6.688740  
## y=y.no 2.494385e-64 -16.934745  
## f.previous=f.previous-[0,0.9] 2.196177e-173 -28.071291  
## f.pdays=f.pdays-(22,23] 1.259643e-281 -35.860546  
##   
## $`2`  
## Cla/Mod Mod/Cla Global p.value  
## f.season=f.season.Mar-May 53.779070 74.1482966 41.362725 9.639926e-211  
## f.previous=f.previous-(0.9,1] 70.921986 26.7201069 11.302605 8.620072e-102  
## y=y.yes 48.892989 17.7020708 10.861723 1.228275e-22  
## f.pdays=f.pdays-(22,23] 30.965732 99.5991984 96.492986 7.200421e-20  
## f.age=f.age-[18,30] 40.574713 23.5804943 17.434870 2.310583e-13  
## f.campaign=f.campaign-[0,1] 34.464875 48.8309953 42.505010 3.637684e-09  
## f.job=f.job.Not-working 37.134503 8.4836339 6.853707 3.388459e-03  
## f.duration=f.duration-[5,120] 28.066795 29.1917168 31.202405 4.431130e-02  
## f.job=f.job.Serv-Tech-BlueC 28.135314 45.5577822 48.577154 5.209983e-03  
## f.season=f.season.Sep-Dec 22.760646 10.3540414 13.647295 5.972406e-06  
## f.campaign=f.campaign-(2,20] 24.658385 26.5197061 32.264529 9.621448e-09  
## f.age=f.age-(40,50] 22.984836 19.2384770 25.110220 1.946669e-10  
## f.pdays=f.pdays-[0,22] 3.428571 0.4008016 3.507014 7.200421e-20  
## y=y.no 27.697842 82.2979292 89.138277 1.228275e-22  
## f.previous=f.previous-[0,0.9] 24.784332 71.0086840 85.951904 1.676131e-80  
## f.season=f.season.Jun-Aug 10.334076 15.4976620 44.989980 3.526427e-179  
## v.test  
## f.season=f.season.Mar-May 30.981263  
## f.previous=f.previous-(0.9,1] 21.420424  
## y=y.yes 9.791202  
## f.pdays=f.pdays-(22,23] 9.124596  
## f.age=f.age-[18,30] 7.329474  
## f.campaign=f.campaign-[0,1] 5.899879  
## f.job=f.job.Not-working 2.930106  
## f.duration=f.duration-[5,120] -2.011134  
## f.job=f.job.Serv-Tech-BlueC -2.793756  
## f.season=f.season.Sep-Dec -4.527364  
## f.campaign=f.campaign-(2,20] -5.737271  
## f.age=f.age-(40,50] -6.365490  
## f.pdays=f.pdays-[0,22] -9.124596  
## y=y.no -9.791202  
## f.previous=f.previous-[0,0.9] -19.000910  
## f.season=f.season.Jun-Aug -28.542041  
##   
## $`3`  
## Cla/Mod Mod/Cla Global  
## f.previous=f.previous-[0,0.9] 75.21567 97.668786 85.951904  
## f.season=f.season.Jun-Aug 87.08241 59.188616 44.989980  
## f.pdays=f.pdays-(22,23] 68.59813 100.000000 96.492986  
## y=y.no 70.59353 95.065092 89.138277  
## f.age=f.age-(40,50] 75.01995 28.458977 25.110220  
## f.campaign=f.campaign-(2,20] 73.29193 35.725098 32.264529  
## f.job=f.job.Serv-Tech-BlueC 69.88449 51.286709 48.577154  
## f.duration=f.duration-[5,120] 70.19910 33.091129 31.202405  
## f.duration=f.duration-(180,300] 62.66055 20.678171 21.843687  
## f.duration=f.duration-(300,2.1e+03] 63.03558 26.279140 27.595190  
## f.job=f.job.Entrep-Retired-selfEmpl 58.92548 10.293672 11.563126  
## f.job=f.job.Not-working 54.67836 5.661520 6.853707  
## f.campaign=f.campaign-[0,1] 60.58463 38.904027 42.505010  
## f.age=f.age-[18,30] 54.36782 14.320315 17.434870  
## f.previous=f.previous-(1,6] 0.00000 0.000000 2.745491  
## y=y.yes 30.07380 4.934908 10.861723  
## f.pdays=f.pdays-[0,22] 0.00000 0.000000 3.507014  
## f.previous=f.previous-(0.9,1] 13.65248 2.331214 11.302605  
## f.season=f.season.Mar-May 43.50775 27.187405 41.362725  
## p.value v.test  
## f.previous=f.previous-[0,0.9] 6.011990e-240 33.078726  
## f.season=f.season.Jun-Aug 3.447463e-187 29.180312  
## f.pdays=f.pdays-(22,23] 7.182640e-86 19.638980  
## y=y.no 2.742424e-74 18.234602  
## f.age=f.age-(40,50] 8.378911e-15 7.761714  
## f.campaign=f.campaign-(2,20] 1.432305e-13 7.393298  
## f.job=f.job.Serv-Tech-BlueC 8.211297e-08 5.362423  
## f.duration=f.duration-[5,120] 5.127384e-05 4.049743  
## f.duration=f.duration-(180,300] 5.558862e-03 -2.772728  
## f.duration=f.duration-(300,2.1e+03] 3.761527e-03 -2.897497  
## f.job=f.job.Entrep-Retired-selfEmpl 1.078588e-04 -3.872197  
## f.job=f.job.Not-working 4.940469e-06 -4.567300  
## f.campaign=f.campaign-[0,1] 6.903561e-13 -7.181332  
## f.age=f.age-[18,30] 1.568440e-15 -7.971434  
## f.previous=f.previous-(1,6] 6.723232e-67 -17.279410  
## y=y.yes 2.742424e-74 -18.234602  
## f.pdays=f.pdays-[0,22] 7.182640e-86 -19.638980  
## f.previous=f.previous-(0.9,1] 1.915585e-167 -27.580337  
## f.season=f.season.Mar-May 1.553439e-179 -28.570715

# Numeric variables globally related to clustering partition  
res.hcpc$desc.var$quanti.var

## Eta2 P-value  
## pdays 0.833431720 0.000000e+00  
## previous 0.492162105 0.000000e+00  
## emp.var.rate 0.887391268 0.000000e+00  
## cons.price.idx 0.451020756 0.000000e+00  
## euribor3m 0.972756054 0.000000e+00  
## nr.employed 0.859965803 0.000000e+00  
## cons.conf.idx 0.137891619 2.106281e-161  
## campaign 0.018064870 1.813245e-20  
## age 0.008646255 3.945874e-10

res.hcpc$desc.var$quanti

## $`1`  
## v.test Mean in category Overall mean sd in category  
## previous 45.971997 1.8210526 0.17855711 0.9567373  
## cons.conf.idx 6.937479 -38.2494737 -40.54192385 7.1162952  
## age 5.645279 44.3789474 40.17755511 16.9858634  
## duration 2.110371 297.6368421 259.85110220 204.0561577  
## campaign -4.322398 1.7894737 2.51503006 1.2598641  
## cons.price.idx -4.817260 93.3651368 93.56373427 0.8341906  
## emp.var.rate -19.115707 -2.0936842 0.05212425 0.8565304  
## euribor3m -21.054798 0.9672263 3.58457355 0.5121394  
## nr.employed -27.122883 5025.3821053 5165.87569138 51.3340730  
## pdays -64.481704 7.8473684 22.41362725 6.4451425  
## Overall sd p.value  
## previous 0.5020810 0.000000e+00  
## cons.conf.idx 4.6436681 3.991593e-12  
## age 10.4585324 1.649135e-08  
## duration 251.6124483 3.482644e-02  
## campaign 2.3588988 1.543422e-05  
## cons.price.idx 0.5793439 1.455430e-06  
## emp.var.rate 1.5774788 1.868722e-81  
## euribor3m 1.7469207 2.066633e-98  
## nr.employed 72.7919889 5.290316e-162  
## pdays 3.1744936 0.000000e+00  
##   
## $`2`  
## v.test Mean in category Overall mean sd in category  
## previous 12.347248 0.3126253 0.17855711 0.5102106  
## pdays 8.083856 22.9686039 22.41362725 0.5683270  
## age -4.063321 39.2585170 40.17755511 12.0357116  
## campaign -7.817387 2.1162325 2.51503006 1.6821146  
## cons.conf.idx -25.982635 -43.1512358 -40.54192385 5.9749813  
## cons.price.idx -46.161007 92.9853808 93.56373427 0.4313831  
## nr.employed -55.580906 5078.3791583 5165.87569138 38.2446628  
## emp.var.rate -60.699472 -2.0186373 0.05212425 0.5394336  
## euribor3m -63.098140 1.2007649 3.58457355 0.2625532  
## Overall sd p.value  
## previous 0.5020810 5.040624e-35  
## pdays 3.1744936 6.275018e-16  
## age 10.4585324 4.837945e-05  
## campaign 2.3588988 5.393129e-15  
## cons.conf.idx 4.6436681 7.782048e-149  
## cons.price.idx 0.5793439 0.000000e+00  
## nr.employed 72.7919889 0.000000e+00  
## emp.var.rate 1.5774788 0.000000e+00  
## euribor3m 1.7469207 0.000000e+00  
##   
## $`3`  
## v.test Mean in category Overall mean sd in category  
## euribor3m 69.642425 4.81553255 3.58457355 0.2815584  
## emp.var.rate 66.534303 1.11407811 0.05212425 0.4992563  
## nr.employed 64.815247 5213.61292764 5165.87569138 17.5063561  
## cons.price.idx 46.665983 93.83728217 93.56373427 0.3924268  
## cons.conf.idx 22.363260 -39.49118983 -40.54192385 2.9986228  
## pdays 18.255882 23.00000000 22.41362725 0.0000000  
## campaign 9.321541 2.73751135 2.51503006 2.6231360  
## previous -30.559574 0.02331214 0.17855711 0.1508930  
## Overall sd p.value  
## euribor3m 1.7469207 0.000000e+00  
## emp.var.rate 1.5774788 0.000000e+00  
## nr.employed 72.7919889 0.000000e+00  
## cons.price.idx 0.5793439 0.000000e+00  
## cons.conf.idx 4.6436681 8.971345e-111  
## pdays 3.1744936 1.857861e-74  
## campaign 2.3588988 1.146631e-20  
## previous 0.5020810 4.219155e-205

### desc.ind ###  
### C. The description of the clusters by the individuals ###  
names(res.hcpc$desc.ind)

## [1] "para" "dist"

res.hcpc$desc.ind$para # Close to center of gravity

## Cluster: 1  
## 36296 36721 40892 41007 36907   
## 0.03946875 0.05735028 0.13237750 0.21737224 0.22738201   
## --------------------------------------------------------   
## Cluster: 2  
## 30951 36346 36347 36427 36864   
## 0.06134805 0.09336824 0.09336824 0.09552183 0.09624810   
## --------------------------------------------------------   
## Cluster: 3  
## 1467 1752 331 6185 18926   
## 0.01424926 0.01424926 0.01478019 0.01478019 0.01491243

res.hcpc$desc.ind$dist

## Cluster: 1  
## 40396 40703 39592 38902 39612   
## 11.543929 10.404680 9.652239 9.431592 9.270355   
## --------------------------------------------------------   
## Cluster: 2  
## 37956 38026 38148 37904 38051   
## 4.742178 4.688806 4.646346 4.645320 4.644823   
## --------------------------------------------------------   
## Cluster: 3  
## 11696 8484 11485 11056 10761   
## 5.265453 5.246245 5.197173 5.138562 5.009706

## CA analysis for your data should contain your factor version of the numeric target (duration) in K= 7 (maximum 10) levels and 2 factors:

### Eigenvalues and dominant axes analysis. How many axes we have to consider are there any row categories that can be combined/avoided to explain Duration target.

Para experimentar y para que tenga más sentido el análisis de correspondencias, refactorizaremos a 8 niveles la variable duration.

Ahora hacemos el análisis de correspondencias entre nuestra nueva duration factorizada y f.age. Para saber cuantas dimensiones debemos considerar, obtenemos la media de los eigenvalues. Vemos que solamente tiene sentido considerar el primer eje, ya que este es el único valor mayor a la media (kaiser).

Al graficar el CA, podemos ver que los 2 niveles con menores edades, son los que menos representados en ese eje. Para duration, los niveles mejor representados en el eje son los de mayor y menor duración.

Al ejecutar la función del chisq.test podemos ver que el pvalue es muy grande, lo que nos puede decir que la probabilidad de que no tengan relación es muy grande.

### CA - duration vs f.age

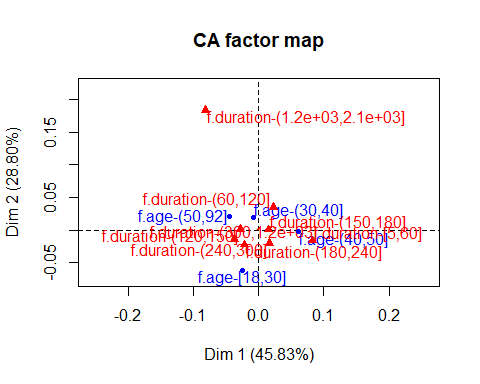
# Para duration  
aux2<-c(5,60,120,150,180,240,300,1200,2100) # Niveles "naturales"  
duration\_k8<-factor(cut(df$duration,breaks=aux2,include.lowest=T))  
table(duration\_k8)

## duration\_k8  
## [5,60] (60,120] (120,150] (150,180]   
## 490 1067 496 470   
## (180,240] (240,300] (300,1.2e+03] (1.2e+03,2.1e+03]   
## 606 484 1311 66

levels(duration\_k8)<-paste0("f.duration-",levels(duration\_k8)) # Hacemos las etiquetas más informativas  
summary(duration\_k8)

## f.duration-[5,60] f.duration-(60,120]   
## 490 1067   
## f.duration-(120,150] f.duration-(150,180]   
## 496 470   
## f.duration-(180,240] f.duration-(240,300]   
## 606 484   
## f.duration-(300,1.2e+03] f.duration-(1.2e+03,2.1e+03]   
## 1311 66

res.ca<-CA(table(df$f.age,duration\_k8))



attributes(res.ca)

## $names  
## [1] "eig" "call" "row" "col" "svd"   
##   
## $class  
## [1] "CA" "list"

res.ca$eig

## eigenvalue percentage of variance  
## dim 1 0.0014130984 45.82949  
## dim 2 0.0008880983 28.80273  
## dim 3 0.0007821856 25.36778  
## cumulative percentage of variance  
## dim 1 45.82949  
## dim 2 74.63222  
## dim 3 100.00000

mean(res.ca$eig[,1]) # Mean of eigenvalues

## [1] 0.001027794

sum(res.ca$eig[,1]) # Total inertia

## [1] 0.003083382

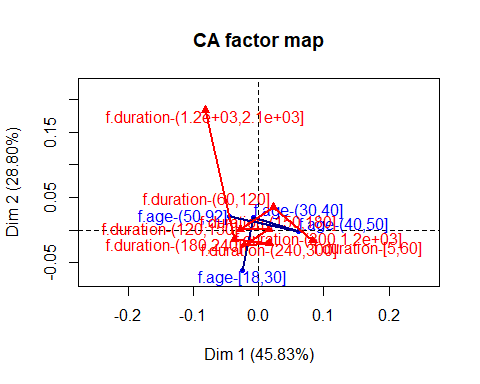
# Rows  
res.ca$row

## $coord  
## Dim 1 Dim 2 Dim 3  
## f.age-[18,30] -0.024672376 -0.061587071 0.00515479  
## f.age-(30,40] -0.007883325 0.019128850 0.02866207  
## f.age-(40,50] 0.060905585 -0.002487498 -0.01655528  
## f.age-(50,92] -0.044696383 0.021246627 -0.04658344  
##   
## $contrib  
## Dim 1 Dim 2 Dim 3  
## f.age-[18,30] 7.510490 74.4623554 0.5922851  
## f.age-(30,40] 1.754758 16.4394554 41.9059874  
## f.age-(40,50] 65.916230 0.1749504 8.7985988  
## f.age-(50,92] 24.818522 8.9232388 48.7031287  
##   
## $cos2  
## Dim 1 Dim 2 Dim 3  
## f.age-[18,30] 0.13746379 0.856535696 0.006000512  
## f.age-(30,40] 0.04973439 0.292830101 0.657435510  
## f.age-(40,50] 0.92975389 0.001550884 0.068695230  
## f.age-(50,92] 0.43249165 0.097726640 0.469781706  
##   
## $inertia  
## [1] 0.0007720623 0.0004985776 0.0010018363 0.0008109061

# Columns: the same  
res.ca$col

## $coord  
## Dim 1 Dim 2 Dim 3  
## f.duration-[5,60] 0.08319396 -0.0161245945 2.458294e-02  
## f.duration-(60,120] 0.02249525 0.0348536919 -2.577791e-02  
## f.duration-(120,150] -0.02735408 0.0013898644 1.373241e-02  
## f.duration-(150,180] 0.01565127 0.0006091358 6.835979e-05  
## f.duration-(180,240] -0.02116464 -0.0227191287 4.705510e-02  
## f.duration-(240,300] 0.01669583 -0.0202658085 3.700409e-04  
## f.duration-(300,1.2e+03] -0.03694459 -0.0143453208 -2.112106e-02  
## f.duration-(1.2e+03,2.1e+03] -0.08146358 0.1836311251 1.153212e-01  
##   
## $contrib  
## Dim 1 Dim 2 Dim 3  
## f.duration-[5,60] 48.095757 2.87482957 7.586705e+00  
## f.duration-(60,120] 7.657254 29.24828562 1.816559e+01  
## f.duration-(120,150] 5.263239 0.02162046 2.396431e+00  
## f.duration-(150,180] 1.632767 0.00393518 5.627155e-05  
## f.duration-(180,240] 3.849653 7.05820696 3.437761e+01  
## f.duration-(240,300] 1.913324 4.48550815 1.697988e-03  
## f.duration-(300,1.2e+03] 25.376484 6.08781771 1.498386e+01  
## f.duration-(1.2e+03,2.1e+03] 6.211521 50.21979635 2.248806e+01  
##   
## $cos2  
## Dim 1 Dim 2 Dim 3  
## f.duration-[5,60] 0.8889838 0.033395490 7.762072e-02  
## f.duration-(60,120] 0.2121463 0.509274053 2.785796e-01  
## f.duration-(120,150] 0.7970605 0.002057746 2.008818e-01  
## f.duration-(150,180] 0.9984686 0.001512388 1.904743e-05  
## f.duration-(180,240] 0.1409384 0.162401758 6.966599e-01  
## f.duration-(240,300] 0.4042264 0.595574996 1.985675e-04  
## f.duration-(300,1.2e+03] 0.6767699 0.102037507 2.211926e-01  
## f.duration-(1.2e+03,2.1e+03] 0.1236833 0.628458833 2.478578e-01  
##   
## $inertia  
## [1] 7.645138e-04 5.100466e-04 9.331130e-05 2.310800e-05 3.859799e-04  
## [6] 6.688615e-05 5.298621e-04 7.096744e-04

# Link levels in rows  
plot.CA(res.ca)  
lines(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="darkblue")  
lines(res.ca$col$coord[,1],res.ca$col$coord[,2],lwd=2,col="red")



# Phi2 = Intensity of the association Chisq/nbobservations  
sum(res.ca$eig[,1]) # Total Inertia = Phi2

## [1] 0.003083382

# H0: f.duration - f.age independency  
chisq.test(table(df$f.age,duration\_k8))

##   
## Pearson's Chi-squared test  
##   
## data: table(df$f.age, duration\_k8)  
## X-squared = 15.386, df = 21, p-value = 0.8031

### CA - Education vs f.duration

Para la segunda prueba decidimos utilizar duration junto con education. Para education usaremos la variable original con todos sus niveles menos el nivel illiterate el cual nos puede causar inconvenientes.

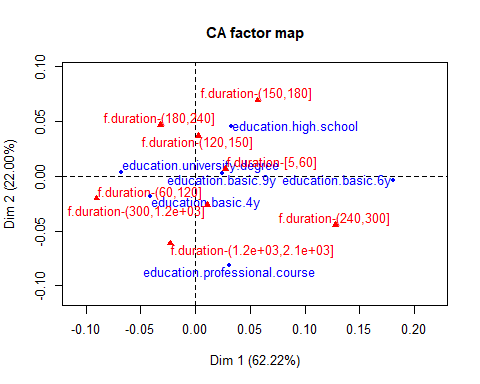
Por kaiser vemos que las primeras dos dimensiones están por encima de la media, por lo que son las que cogemos.

Podemos ver que para la primera dimensión los valores más lejanos del centro son los niveles de education.basic\_6y, education.university.degree, para la primera dimensión. Para la segunda tenemos, f.duration(150,180], education.professioal.course, esto nos puede decir qué niveles se ven mejor representados en las dimensiones.

#Education  
table(df$education)

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

education\_k6<-df$education  
education\_k6[which(education\_k6=="education.illiterate")]<-"education.basic.4y"  
education\_k6=factor(education\_k6)  
  
par(cex=0.8)  
res.ca<-CA(table(education\_k6,duration\_k8))



res.ca$eig

## eigenvalue percentage of variance  
## dim 1 3.865234e-03 62.2179391  
## dim 2 1.366656e-03 21.9987982  
## dim 3 7.963697e-04 12.8190136  
## dim 4 1.318112e-04 2.1217398  
## dim 5 5.234014e-05 0.8425093  
## cumulative percentage of variance  
## dim 1 62.21794  
## dim 2 84.21674  
## dim 3 97.03575  
## dim 4 99.15749  
## dim 5 100.00000

mean(res.ca$eig[,1])

## [1] 0.001242482

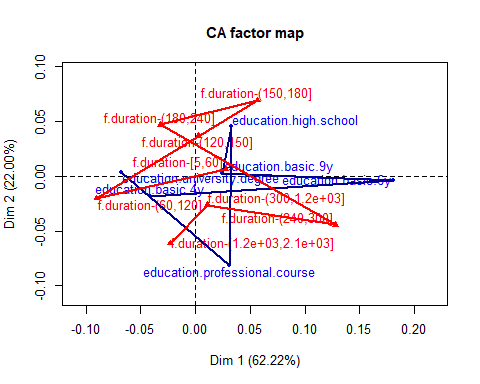
# Rows  
res.ca$row

## $coord  
## Dim 1 Dim 2 Dim 3  
## education.basic.4y -0.04133205 -0.018121606 -0.058412332  
## education.basic.6y 0.18102401 -0.004110569 0.042729415  
## education.basic.9y 0.02483545 0.002685184 -0.027046691  
## education.high.school 0.03234069 0.045050136 -0.004194745  
## education.professional.course 0.03069724 -0.081359937 0.007155999  
## education.university.degree -0.06774789 0.003779150 0.026186751  
## Dim 4 Dim 5  
## education.basic.4y 0.010678244 -0.012189412  
## education.basic.6y -0.003590873 -0.018490687  
## education.basic.9y -0.022417259 0.003721054  
## education.high.school 0.010664921 0.005274952  
## education.professional.course 0.008039829 0.008059669  
## education.university.degree -0.002700335 -0.001997413  
##   
## $contrib  
## Dim 1 Dim 2 Dim 3 Dim 4  
## education.basic.4y 4.570327 2.48475408 44.3040567 8.9453332  
## education.basic.6y 46.043183 0.06714497 12.4511048 0.5312717  
## education.basic.9y 2.590316 0.08563935 14.9106984 61.8866872  
## education.high.school 6.485648 35.59293113 0.5295743 20.6820512  
## education.professional.course 3.092616 61.44198967 0.8156972 6.2207761  
## education.university.degree 37.217909 0.32754080 26.9888687 1.7338807  
## Dim 5  
## education.basic.4y 29.354847  
## education.basic.6y 35.476433  
## education.basic.9y 4.294194  
## education.high.school 12.741859  
## education.professional.course 15.743551  
## education.university.degree 2.389116  
##   
## $cos2  
## Dim 1 Dim 2 Dim 3  
## education.basic.4y 0.2991135 0.057498362 0.597408243  
## education.basic.6y 0.9371554 0.000483218 0.052214742  
## education.basic.9y 0.3295023 0.003851787 0.390789233  
## education.high.school 0.3233541 0.627439873 0.005439900  
## education.professional.course 0.1217065 0.854941382 0.006613871  
## education.university.degree 0.8658182 0.002694161 0.129359520  
## Dim 4 Dim 5  
## education.basic.4y 0.0199646527 0.0260152290  
## education.basic.6y 0.0003687563 0.0097778938  
## education.basic.9y 0.2684598606 0.0073968422  
## education.high.school 0.0351637465 0.0086023464  
## education.professional.course 0.0083485069 0.0083897596  
## education.university.degree 0.0013755312 0.0007526117  
##   
## $inertia  
## [1] 0.0005905913 0.0018990198 0.0003038576 0.0007752660 0.0009821731  
## [6] 0.0016615026

# Columns: the same  
res.ca$col

## $coord  
## Dim 1 Dim 2 Dim 3  
## f.duration-[5,60] 0.027029115 0.006192711 0.035619391  
## f.duration-(60,120] -0.090112862 -0.020538411 0.017086167  
## f.duration-(120,150] 0.002502632 0.036564696 -0.049483431  
## f.duration-(150,180] 0.056880902 0.069080817 0.004342747  
## f.duration-(180,240] -0.032420109 0.046836380 0.019636428  
## f.duration-(240,300] 0.127883956 -0.044401648 0.030160547  
## f.duration-(300,1.2e+03] 0.010833855 -0.026353481 -0.028797208  
## f.duration-(1.2e+03,2.1e+03] -0.023054621 -0.061622025 -0.029181692  
## Dim 4 Dim 5  
## f.duration-[5,60] 0.0049402640 0.001195881  
## f.duration-(60,120] -0.0054537007 0.003195489  
## f.duration-(120,150] -0.0047413221 0.015186085  
## f.duration-(150,180] -0.0004840916 -0.004166702  
## f.duration-(180,240] 0.0067808780 -0.005449376  
## f.duration-(240,300] -0.0044409552 0.006544906  
## f.duration-(300,1.2e+03] -0.0015812648 -0.007781878  
## f.duration-(1.2e+03,2.1e+03] 0.0922854018 0.011622914  
##   
## $contrib  
## Dim 1 Dim 2 Dim 3 Dim 4  
## f.duration-[5,60] 1.85602333 0.2755485 15.6442142 1.81820805  
## f.duration-(60,120] 44.92226773 6.5999164 7.8385985 4.82496535  
## f.duration-(120,150] 0.01610644 9.7240230 30.5622963 1.69522671  
## f.duration-(150,180] 7.88414319 32.8891869 0.2230549 0.01674558  
## f.duration-(180,240] 3.30236637 19.4930424 5.8800674 4.23634841  
## f.duration-(240,300] 41.03941655 13.9921319 11.0792023 1.45126036  
## f.duration-(300,1.2e+03] 0.79779697 13.3511603 27.3582403 0.49837825  
## f.duration-(1.2e+03,2.1e+03] 0.18187944 3.6749905 1.4143262 85.45886729  
## Dim 5  
## f.duration-[5,60] 0.2683099  
## f.duration-(60,120] 4.1716088  
## f.duration-(120,150] 43.7963518  
## f.duration-(150,180] 3.1242607  
## f.duration-(180,240] 6.8901823  
## f.duration-(240,300] 7.9381048  
## f.duration-(300,1.2e+03] 30.3973777  
## f.duration-(1.2e+03,2.1e+03] 3.4138039  
##   
## $cos2  
## Dim 1 Dim 2 Dim 3  
## f.duration-[5,60] 0.354045576 0.01858477 0.614849014  
## f.duration-(60,120] 0.915064957 0.04753484 0.032897861  
## f.duration-(120,150] 0.001548393 0.33053014 0.605350244  
## f.duration-(150,180] 0.402214863 0.59325319 0.002344522  
## f.duration-(180,240] 0.283613187 0.59192152 0.104045298  
## f.duration-(240,300] 0.847459925 0.10216102 0.047137383  
## f.duration-(300,1.2e+03] 0.068871807 0.40752236 0.486604635  
## f.duration-(1.2e+03,2.1e+03] 0.038426381 0.27452723 0.061565082  
## Dim 4 Dim 5  
## f.duration-[5,60] 1.182758e-02 0.0006930612  
## f.duration-(60,120] 3.351668e-03 0.0011506756  
## f.duration-(120,150] 5.557584e-03 0.0570136422  
## f.duration-(150,180] 2.913267e-05 0.0021582909  
## f.duration-(180,240] 1.240707e-02 0.0080129231  
## f.duration-(240,300] 1.021974e-03 0.0022196981  
## f.duration-(300,1.2e+03] 1.467184e-03 0.0355340175  
## f.duration-(1.2e+03,2.1e+03] 6.157147e-01 0.0097666035  
##   
## $inertia  
## [1] 0.0002026283 0.0018975163 0.0004020629 0.0007576562 0.0004500643  
## [6] 0.0018717928 0.0004477408 0.0001829489

# Link levels in rows  
plot.CA(res.ca)  
lines(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="darkblue")  
lines(res.ca$col$coord[,1],res.ca$col$coord[,2],lwd=2,col="red")



# Phi2 = Intensity of the association Chisq/nbobservations  
sum(res.ca$eig[,1]) # Total Inertia = Phi2

## [1] 0.00621241

chisq.test(table(education\_k6,duration\_k8))

## Warning in chisq.test(table(education\_k6, duration\_k8)): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(education\_k6, duration\_k8)  
## X-squared = 31, df = 35, p-value = 0.6617

# Traditional analysis  
table(df$y,duration\_k8)

## duration\_k8  
## f.duration-[5,60] f.duration-(60,120] f.duration-(120,150]  
## y.no 490 1047 481  
## y.yes 0 20 15  
## duration\_k8  
## f.duration-(150,180] f.duration-(180,240] f.duration-(240,300]  
## y.no 441 550 431  
## y.yes 29 56 53  
## duration\_k8  
## f.duration-(300,1.2e+03] f.duration-(1.2e+03,2.1e+03]  
## y.no 983 25  
## y.yes 328 41

chisq.test(table(df$y,duration\_k8))

##   
## Pearson's Chi-squared test  
##   
## data: table(df$y, duration\_k8)  
## X-squared = 643.03, df = 7, p-value < 2.2e-16

# Modelización con target numérico

## Modelización con variables explicativas numéricas

### Modelo simple

El primer paso es decidir con cuantas variables contamos para el modelo. Si tuviéramos muchas variables explicativas podríamos utilizar el resultado del condes para saber cuáles de ellas utilizar, aunque también sería posible seleccionarlas a partir del análisis de componentes principales. Dado que tenemos poca cantidad de variables usamos todas.

Empezamos utilizando **lm** para crear un modelo inicial del cual podemos ir descartando aquellas variables explicativas que nos parecen irrelevantes. Después contrastaremos nuestra selección usando el método Akaike o BIC, que en una sucesión de pasos va descartando variables.

m1<-lm(duration~.,data=df[,c("duration",vars\_num)])  
summary(m1)

##   
## Call:  
## lm(formula = duration ~ ., data = df[, c("duration", vars\_num)])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -328.23 -154.46 -82.08 61.30 1842.65   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 777.28481 2613.90120 0.297 0.7662   
## age 0.03205 0.34459 0.093 0.9259   
## campaign -6.21960 1.53172 -4.061 4.97e-05 \*\*\*  
## pdays -2.37020 1.40614 -1.686 0.0919 .   
## previous -17.62769 9.52959 -1.850 0.0644 .   
## emp.var.rate 3.48261 13.07499 0.266 0.7900   
## cons.price.idx 11.61303 15.53269 0.748 0.4547   
## cons.conf.idx -0.51158 1.24917 -0.410 0.6822   
## euribor3m 3.62210 16.39663 0.221 0.8252   
## nr.employed -0.30339 0.28145 -1.078 0.2811   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 251.1 on 4980 degrees of freedom  
## Multiple R-squared: 0.006393, Adjusted R-squared: 0.004597   
## F-statistic: 3.56 on 9 and 4980 DF, p-value: 0.0002021

Anova(m1)

## Anova Table (Type II tests)  
##   
## Response: duration  
## Sum Sq Df F value Pr(>F)   
## age 545 1 0.0087 0.92589   
## campaign 1039241 1 16.4879 4.971e-05 \*\*\*  
## pdays 179087 1 2.8413 0.09193 .   
## previous 215671 1 3.4217 0.06440 .   
## emp.var.rate 4472 1 0.0709 0.78998   
## cons.price.idx 35233 1 0.5590 0.45471   
## cons.conf.idx 10571 1 0.1677 0.68216   
## euribor3m 3076 1 0.0488 0.82518   
## nr.employed 73240 1 1.1620 0.28111   
## Residuals 313891375 4980   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Viendo este volcado, vemos que todas las variables menos, campaign tienen un p-value superior al 0.05, sin embargo, pdays y previous están por debajo de 0.1 lo que podríamos llegar a incorporarlas al modelo. El r-square es de 0.006393 lo que nos dice que nuestro modelo no se ajusta bien.

Al ver el resultado de Anova, podemos ver resultados muy parecidos.

Ahora probaremos seleccionando las variables a partir de la criba anterior:

m2<-lm(duration~campaign+pdays+previous,data=df)  
summary(m2)

##   
## Call:  
## lm(formula = duration ~ campaign + pdays + previous, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -264.50 -156.27 -82.24 61.80 1840.02   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 344.591 32.204 10.700 < 2e-16 \*\*\*  
## campaign -6.304 1.513 -4.167 3.14e-05 \*\*\*  
## pdays -2.991 1.377 -2.172 0.0299 \*   
## previous -10.391 8.726 -1.191 0.2337   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 251.1 on 4986 degrees of freedom  
## Multiple R-squared: 0.004472, Adjusted R-squared: 0.003873   
## F-statistic: 7.465 on 3 and 4986 DF, p-value: 5.52e-05

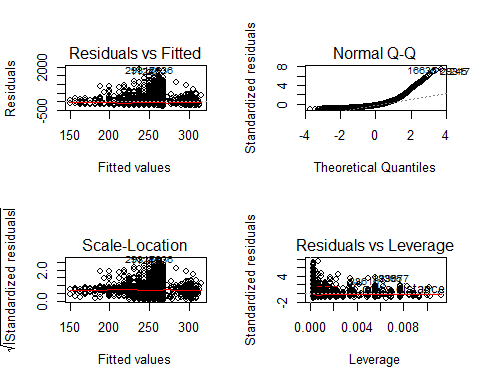
m3<-lm(duration~campaign+pdays,data=df)  
summary(m3)

##   
## Call:  
## lm(formula = duration ~ campaign + pdays, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -263.02 -156.25 -82.58 60.87 1840.89   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 321.114 25.467 12.609 < 2e-16 \*\*\*  
## campaign -6.183 1.510 -4.095 4.28e-05 \*\*\*  
## pdays -2.040 1.122 -1.818 0.0691 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 251.2 on 4987 degrees of freedom  
## Multiple R-squared: 0.004189, Adjusted R-squared: 0.003789   
## F-statistic: 10.49 on 2 and 4987 DF, p-value: 2.848e-05

vif(m3)

## campaign pdays   
## 1.003138 1.003138

par(mfrow=c(2,2))  
plot(m3)



par(mfrow=c(1,1))  
m=m3;

Viendo el resultado del lm con estas variables, podemos ver que previous da por encima de 0.2, por lo que también descartamos esta variable. También podemos ver que el r-square sigue siendo muy bajo.

Al realizar nuevamente el lm con estas dos variables restantes, vemos que su p-value es inferior al 0.1, por lo que daríamos por concluida la criba.

Finalmente hacemos el análisis de residuos con vif, el cual nos dice si existen problemas de colinealidad es decir si existen variables que pueden explicar a otras. Si nos da valores por debajo de 3 son buenos y por encima de 5 que las variables elegidas tienen redundancia y que inflará las varianzas. En nuestro caso, el resultado de las dos variables es inferior a 3.

Viendo el plot de la normal Q-Q, vemos que los valores distan mucho de la recta de referencia, con que podemos decir que su distribución no es para nada normal.

Para quitar las variables redundantes probamos con la versión bayesiana del step (del BIC):

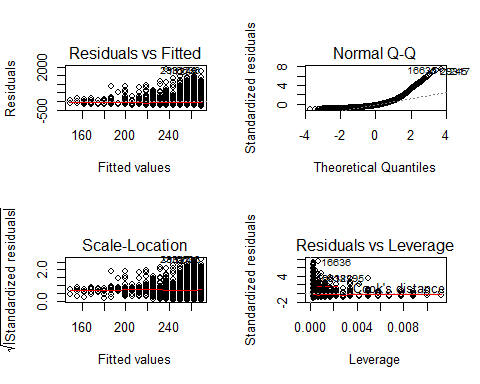
m5<-step(m,k=log(nrow(df)))

## Start: AIC=55172.94  
## duration ~ campaign + pdays  
##   
## Df Sum of Sq RSS AIC  
## - pdays 1 208524 314796334 55168  
## <none> 314587810 55173  
## - campaign 1 1058016 315645826 55181  
##   
## Step: AIC=55167.73  
## duration ~ campaign  
##   
## Df Sum of Sq RSS AIC  
## <none> 314796334 55168  
## - campaign 1 1114698 315911032 55177

summary(m5)

##   
## Call:  
## lm(formula = duration ~ campaign, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -264.45 -156.69 -82.45 61.14 1840.23   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 275.786 5.199 53.051 < 2e-16 \*\*\*  
## campaign -6.336 1.508 -4.203 2.68e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 251.2 on 4988 degrees of freedom  
## Multiple R-squared: 0.003529, Adjusted R-squared: 0.003329   
## F-statistic: 17.66 on 1 and 4988 DF, p-value: 2.684e-05

par(mfrow=c(2,2))  
plot(m5)



par(mfrow=c(1,1))

La versión bayesiana es conveniente usarla en casos de muestras grandes. En este caso vemos que se queda con una sola variable (campaign), ya que en el primer step del volcado vemos que sin la variable p-days el valor AIC, en este caso BIC, es menor.

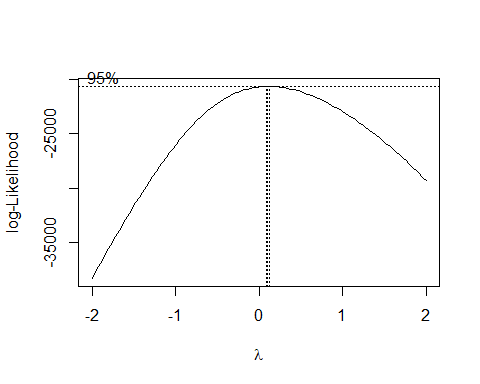
En este caso no podemos hacer el análisis de residuos con vif porque solo tenemos 1 variable.

Al igual que en nuestro caso nos da una plot Q-Q totalmente desviada de las dist normal.

### Modelo con transformaciones

Mediante la función boxcox descartamos la posibilidad de elevar el target al cuadrado, pero sí contemplamos aplicarle el logaritmo, pues el pico de la curva está entre 0 y 1, bastante cerca del 0.

boxcox(m,data=df)



Ahora procedemos a la transformación polinómica.

Como solo tenemos una variable explicativa podemos empezar desde cero, pero si tuviéramos ya un modelo no volveríamos a empezar.

m6<-lm(log(duration)~.,data=df[,c("duration",vars\_num)])   
Anova(m6)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## age 0.1 1 0.1176 0.73162   
## campaign 97.4 1 120.3195 < 2e-16 \*\*\*  
## pdays 4.0 1 4.9361 0.02635 \*   
## previous 0.2 1 0.1873 0.66523   
## emp.var.rate 0.2 1 0.1976 0.65665   
## cons.price.idx 0.4 1 0.4944 0.48201   
## cons.conf.idx 0.1 1 0.1082 0.74227   
## euribor3m 1.6 1 1.9413 0.16359   
## nr.employed 2.7 1 3.3650 0.06666 .   
## Residuals 4030.4 4980   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Viendo el resultado del Anova, procedemos a descartar las variables cuyo valor de Pr es mayor a 0.1

m7<-lm(log(duration)~campaign+pdays+nr.employed,data=df)  
summary(m7)

##   
## Call:  
## lm(formula = log(duration) ~ campaign + pdays + nr.employed,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6815 -0.5509 -0.0106 0.5858 2.6860   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.3887200 0.9445802 6.764 1.5e-11 \*\*\*  
## campaign -0.0598301 0.0054664 -10.945 < 2e-16 \*\*\*  
## pdays -0.0135538 0.0042873 -3.161 0.00158 \*\*   
## nr.employed -0.0001463 0.0001888 -0.775 0.43843   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9007 on 4986 degrees of freedom  
## Multiple R-squared: 0.02798, Adjusted R-squared: 0.0274   
## F-statistic: 47.84 on 3 and 4986 DF, p-value: < 2.2e-16

Anova(m7)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## campaign 97.2 1 119.7932 < 2e-16 \*\*\*  
## pdays 8.1 1 9.9945 0.00158 \*\*   
## nr.employed 0.5 1 0.6005 0.43843   
## Residuals 4044.5 4986   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Viendo los p-values, nos encontramos que la variable nr.employed es mayor a 0.1, por lo que procedemos a eliminarla de nuestro modelo.

Relativo al gráfico, podemos ver como la Normal Q-Q ha mejorado bastante acercándose a la recta ideal.

Ahora procedemos a quitar nr.employed.

m9<-lm(log(duration)~campaign+pdays,data=df)  
summary(m9)

##   
## Call:  
## lm(formula = log(duration) ~ campaign + pdays, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6522 -0.5521 -0.0090 0.5858 2.6797   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.660184 0.091319 61.982 < 2e-16 \*\*\*  
## campaign -0.060418 0.005413 -11.161 < 2e-16 \*\*\*  
## pdays -0.014703 0.004023 -3.655 0.00026 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9006 on 4987 degrees of freedom  
## Multiple R-squared: 0.02786, Adjusted R-squared: 0.02747   
## F-statistic: 71.47 on 2 and 4987 DF, p-value: < 2.2e-16

Anova(m9)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## campaign 101.0 1 124.57 < 2.2e-16 \*\*\*  
## pdays 10.8 1 13.36 0.0002597 \*\*\*  
## Residuals 4045.0 4987   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

vif(m9)

## campaign pdays   
## 1.003138 1.003138

Viendo el valor final del r-square, podemos ver que este no es un buen modelo. También los que no puede decir es que las variables no representan a nuestro target, esto ya lo pudimos ver en el deliverable2.

El resultado del vif nos da valores aceptables, diciendo que no hay colinealidad entre variables.

### Modelo de regresión polinómica

Ahora podemos probar con las versiones cuadráticas de las variables explicativas, partiendo de nuestro mejor modelo:

m20<-lm(log(duration)~poly(campaign,2)+poly(pdays,2),data=df)  
summary(m20)

##   
## Call:  
## lm(formula = log(duration) ~ poly(campaign, 2) + poly(pdays,   
## 2), data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6353 -0.5534 -0.0100 0.5842 2.6431   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.17868 0.01274 406.451 < 2e-16 \*\*\*  
## poly(campaign, 2)1 -10.03807 0.90154 -11.134 < 2e-16 \*\*\*  
## poly(campaign, 2)2 -1.79572 0.90036 -1.994 0.046158 \*   
## poly(pdays, 2)1 -3.34605 0.90176 -3.711 0.000209 \*\*\*  
## poly(pdays, 2)2 -1.90923 0.90014 -2.121 0.033968 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9 on 4985 degrees of freedom  
## Multiple R-squared: 0.02951, Adjusted R-squared: 0.02873   
## F-statistic: 37.89 on 4 and 4985 DF, p-value: < 2.2e-16

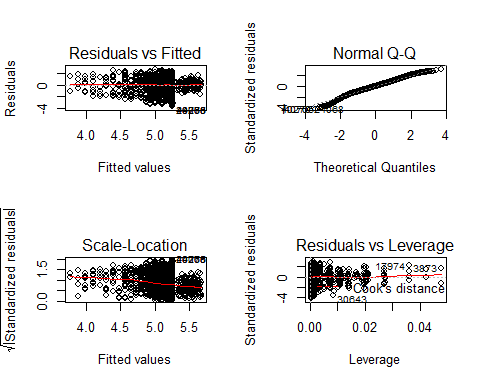
Anova(m20)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## poly(campaign, 2) 103.7 2 64.0104 < 2.2e-16 \*\*\*  
## poly(pdays, 2) 14.8 2 9.1263 0.0001106 \*\*\*  
## Residuals 4038.2 4985   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

vif(m20)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(campaign, 2) 1.004062 2 1.001014  
## poly(pdays, 2) 1.004062 2 1.001014

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



par(mfrow=c(1,1))

## Modelización con variables explicativas numéricas y categóricas

Creamos una variable que contiene las variables categóricas y categóricas factorizadas además de las categóricas.

vars\_cat\_total = c(vars\_cat, names(df[,22:29]))  
condes(df[,c("duration",vars\_cat\_total)],1, proba= 0.05)

## $quali  
## R2 p.value  
## f.duration 0.621168787 0.000000e+00  
## f.campaign 0.003783221 7.858324e-05  
## month 0.004450289 8.185248e-03  
## poutcome 0.001675246 1.528736e-02  
## f.pdays 0.001120416 1.805086e-02  
## f.season 0.001469523 2.555430e-02  
##   
## $category  
## Estimate p.value  
## f.duration-(300,2.1e+03] 310.351061 0.000000e+00  
## f.campaign-(1,2] 23.010415 3.895001e-05  
## month.apr 35.257830 4.865526e-03  
## f.season.Mar-May 13.191703 6.782891e-03  
## poutcome.success 38.426383 1.231875e-02  
## f.pdays-[0,22] 22.891544 1.805086e-02  
## day\_of\_week.wed 14.619928 3.788283e-02  
## job.retired 34.239467 3.904250e-02  
## marital.divorced -15.444147 4.653367e-02  
## f.season.Jun-Aug -6.204726 4.445499e-02  
## f.pdays-(22,23] -22.891544 1.805086e-02  
## month.aug -25.222251 7.943838e-03  
## f.campaign-(2,20] -17.351641 3.316706e-03  
## f.duration-(180,300] -20.507215 3.927333e-04  
## f.duration-(120,180] -106.753548 5.404997e-53  
## f.duration-[5,120] -183.090298 1.278559e-312

Al hacer condes, con todas las variables categóricas, contemplamos el uso de f.campaign y month para nuestro modelo, ya que la probabilidad de que no tengan relación con el target está por debajo del 0.01. Como nos sale la versión categórica de campaign que también nos sale en el modelo numérico, debemos elegir entre una u otra, pero nunca las dos a la vez.

En vista de que la variable numérica pdays aporta una información errante ya que aquellos que no fueron contactados tienen asignados un valor que no les corresponde, decidimos utilizar f.pdays porque contiene una información más rigurosa, ya que se clasifican entre contactados y no contactados.

Debido a que la variable month es una variable con muchos niveles y eso no es bueno para la modelización, decidimos reagruparla.

#chunk 115  
# Months to groups  
df$f.influentMonth <- 3  
# 1 level - mar-may   
aux<-which(df$month %in% c("month.apr","month.jun","month.aug"))  
df$f.influentMonth[aux] <-1  
  
# 2 level - jun-ago  
aux<-which(df$month %in% c("month.sep","month.may","month.jul"))  
df$f.influentMonth[aux] <-2  
  
# 3 level - aug-feb  
aux<-which(df$month %in% c("month.mar","month.dec","month.oct","month.nov"))  
df$f.influentMonth[aux] <-3  
  
df$f.influentMonth<-factor(df$f.influentMonth,levels=1:3,labels=c("apr-jun-aug","sep-may-jul","mar-dec-oct-nov"))  
levels(df$f.influentMonth)<-paste0("f.influentMonth.",levels(df$f.influentMonth)) # Hacemos las etiquetas m?s informativas  
summary(df$f.influentMonth)

## f.influentMonth.apr-jun-aug f.influentMonth.sep-may-jul   
## 1701 2615   
## f.influentMonth.mar-dec-oct-nov   
## 674

Contrastamos un modelo con campaign o con f.campaign para ver cuál es mejor.

m22<-lm(log(duration)~campaign+f.pdays+f.influentMonth,data=df)  
m23<-lm(log(duration)~f.pdays+f.campaign+f.influentMonth,data=df)  
BIC(m23,m22)

## df BIC  
## m23 7 13214.68  
## m22 6 13150.71

# Ya que nos quedamos con el modelo m22  
Anova(m22)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## campaign 102.8 1 126.9775 < 2.2e-16 \*\*\*  
## f.pdays 15.2 1 18.7951 1.484e-05 \*\*\*  
## f.influentMonth 8.4 2 5.1938 0.005581 \*\*   
## Residuals 4033.9 4985   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Haciendo BIC para comparar modelos, podemos ver que el que da un menor BIC es m22, por lo que decidimos quedarnos con este modelo. Viendo el resultado del Anova, podemos ver que los p-values son inferiores a 0.1

## Interacciones

m30<-lm(log(duration)~(campaign+f.pdays+f.influentMonth)^2,data=df)  
Anova(m30)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## campaign 103.1 1 127.5736 < 2.2e-16 \*\*\*  
## f.pdays 15.0 1 18.5584 1.68e-05 \*\*\*  
## f.influentMonth 8.5 2 5.2517 0.005268 \*\*   
## campaign:f.pdays 2.2 1 2.7306 0.098506 .   
## campaign:f.influentMonth 5.2 2 3.1929 0.041136 \*   
## f.pdays:f.influentMonth 1.2 2 0.7427 0.475884   
## Residuals 4025.4 4980   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Vemos que la interacción entre campaign y nuestra nueva variable factor month es significativa, por lo tanto, creamos un nuevo modelo m31 con esa interacción. Por otro lado, aunque f.pdays con f.influentMoth tiene un p-value muy alto de 0.4, realizamos la interacción porque lo pide el enunciado.

#chunk 140  
m31<-lm(log(duration)~(f.influentMonth\*campaign+f.pdays),data=df)  
Anova(m31)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## f.influentMonth 8.4 2 5.1981 0.005557 \*\*   
## campaign 102.8 1 127.0831 < 2.2e-16 \*\*\*  
## f.pdays 15.0 1 18.5535 1.684e-05 \*\*\*  
## f.influentMonth:campaign 5.0 2 3.0728 0.046377 \*   
## Residuals 4028.9 4983   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

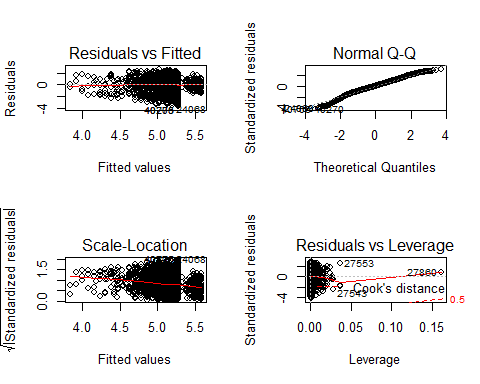
m32<-lm(log(duration)~(f.influentMonth\*f.pdays+campaign),data=df)  
Anova(m32)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## f.influentMonth 8.4 2 5.1932 0.005584 \*\*   
## f.pdays 15.2 1 18.7930 1.486e-05 \*\*\*  
## campaign 103.1 1 127.4200 < 2.2e-16 \*\*\*  
## f.influentMonth:f.pdays 1.2 2 0.7228 0.485455   
## Residuals 4032.7 4983   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Vemos que el modelo 31 es aceptable, sus p-values son aceptables, mientras como ya era previsible el modelo m32 lo descartamos.

## Validación

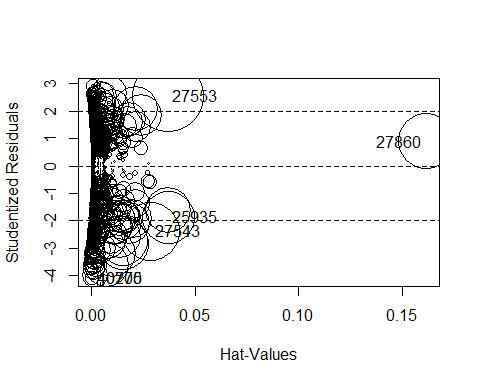
par(mfrow=c(2,2))  
plot(m31)



par(mfrow=c(1,1))

Analizando los gráficos:  
- Residual VS Fitted. En este gráfico muestra los residuos de los valores predichos. Lo deseable es que los puntos estén uniformemente dispersos, para poderlo contrastar el gráfico está provisto de una recta smoother que conviene que sea horizontal, y uniforme. A pesar de que podemos ver un patrón en el gráfico, podemos decir que el resultado no es aceptable. - Normal Q-Q. Este plot nos muestra la tendencia a una distribución normal de los residuos, esta provista de una recta diagonal de referencia en la que se espera que los residuos se ajusten lo máximo posible. En nuestro caso, apreciamos ciertas desviaciones en los extremos de la recta, aunque si lo comparamos con plots anteriores, se acerca más a la normal, pero sigue siendo poco aceptable.  
- Scale-Location. Este plot hace referencia a la varianza de los valores de la predicción, si se mantiene constante implica homocedasticidad, de lo contrario heterocedasticidad que se vería reflejada en una nube de puntos en forma de cono. Para nuestro caso, podemos ver que el gráfico tiene una tendencia a cono que además se evidencia con la desviación de la smoother line. Pero es una heterocedasticidad que es imposible de corregir de manera fácil, es una réplica del primer plot. - Residuals Vs Leverage. Vemos que hay un individuo con mucho leverage, el 27860. Utilizaremos el influencePlot para poder ver con más detalles los individuos influyentes.

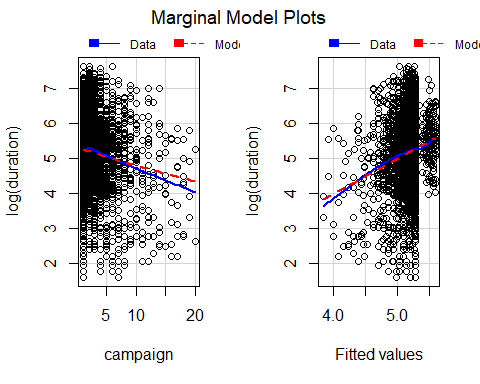
#chunk 150  
influencePlot(m31)



## StudRes Hat CookD  
## 25935 -1.866796 0.0369674028 0.019101054  
## 27543 -2.380400 0.0286710547 0.023871117  
## 27553 2.548824 0.0369674028 0.035586127  
## 27860 0.882873 0.1613511871 0.021424449  
## 40270 -4.092113 0.0005533216 0.001320216  
## 40705 -4.092113 0.0005533216 0.001320216

marginalModelPlots(m31)

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



which(row.names(df)==27860)

## [1] 3329

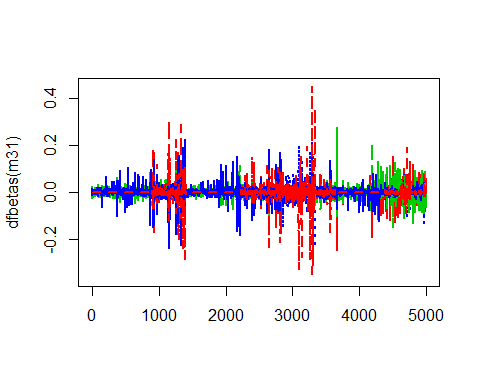
which(row.names(df)==27553)

## [1] 3293

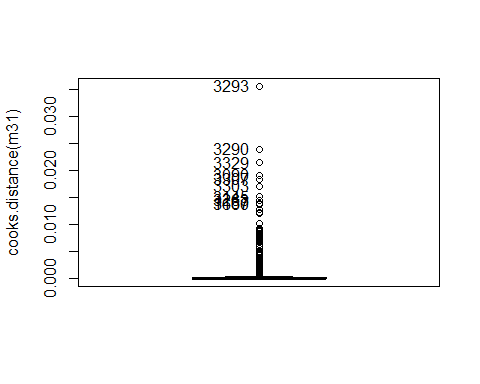
* InfluenPLot. Nos muestra los individuos más influentes, esto se puede ver gráficamente a través del radio de las circunferencias. En nuestro caso, viendo el gráfico podemos ver que hay individuos bastante influyentes, el 3329 y 3293 que para nuestra muestra serían los individuos.
* MarginalModelPlot. Nos muestra las discrepancias entre las predicciones de nuestro modelo y los resultados reales de nuestras observaciones desglosado por variables, utiliza dos líneas de soporte, una roja para la tendencia del modelo y otra azul referente a cada variable. Podemos ver que, para nuestro modelo, las líneas tienen un poco de desviación entre ellas, pero nada muy relevante.

Trabajamos con el mejor modelo obtenido, y vemos que individuos influyen más en nuestros datos para saber si están afectando nuestro resultado.

matplot(dfbetas(m31), type="l", col=2:4,lwd=2)



Boxplot(cooks.distance(m31))

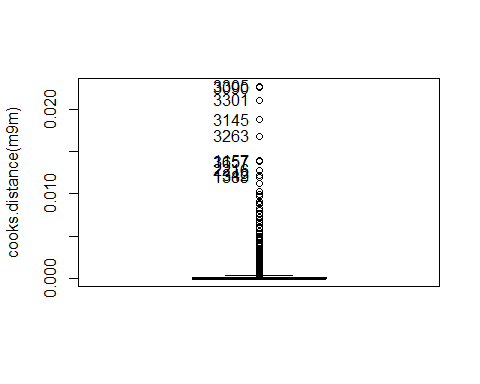


## [1] 3293 3290 3329 3090 3307 3303 3145 3263 1157 3660

Consideramos que hay un individuo que repercute demasiado en los datos (3293), aun así, no lo eliminaremos.

m9m<-lm(log(duration)~(f.influentMonth\*campaign+f.pdays),data=df[c(-3293,-3290,-3329),])

Boxplot(cooks.distance(m9m))



## [1] 3305 3090 3301 3145 3263 1157 3657 2216 1342 1389

summary(m9m)

##   
## Call:  
## lm(formula = log(duration) ~ (f.influentMonth \* campaign + f.pdays),   
## data = df[c(-3293, -3290, -3329), ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6728 -0.5449 -0.0105 0.5877 2.6092   
##   
## Coefficients:  
## Estimate  
## (Intercept) 5.631270  
## f.influentMonthf.influentMonth.sep-may-jul 0.001571  
## f.influentMonthf.influentMonth.mar-dec-oct-nov -0.057482  
## campaign -0.078799  
## f.pdaysf.pdays-(22,23] -0.300195  
## f.influentMonthf.influentMonth.sep-may-jul:campaign 0.028354  
## f.influentMonthf.influentMonth.mar-dec-oct-nov:campaign 0.011485  
## Std. Error t value  
## (Intercept) 0.073735 76.372  
## f.influentMonthf.influentMonth.sep-may-jul 0.040830 0.038  
## f.influentMonthf.influentMonth.mar-dec-oct-nov 0.066125 -0.869  
## campaign 0.008971 -8.784  
## f.pdaysf.pdays-(22,23] 0.069722 -4.306  
## f.influentMonthf.influentMonth.sep-may-jul:campaign 0.011481 2.470  
## f.influentMonthf.influentMonth.mar-dec-oct-nov:campaign 0.025573 0.449  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## f.influentMonthf.influentMonth.sep-may-jul 0.9693   
## f.influentMonthf.influentMonth.mar-dec-oct-nov 0.3847   
## campaign < 2e-16 \*\*\*  
## f.pdaysf.pdays-(22,23] 1.7e-05 \*\*\*  
## f.influentMonthf.influentMonth.sep-may-jul:campaign 0.0136 \*   
## f.influentMonthf.influentMonth.mar-dec-oct-nov:campaign 0.6534   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8983 on 4980 degrees of freedom  
## Multiple R-squared: 0.03195, Adjusted R-squared: 0.03078   
## F-statistic: 27.39 on 6 and 4980 DF, p-value: < 2.2e-16

summary(m31)

##   
## Call:  
## lm(formula = log(duration) ~ (f.influentMonth \* campaign + f.pdays),   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6728 -0.5452 -0.0098 0.5884 2.6092   
##   
## Coefficients:  
## Estimate  
## (Intercept) 5.631659  
## f.influentMonthf.influentMonth.sep-may-jul 0.001578  
## f.influentMonthf.influentMonth.mar-dec-oct-nov -0.077716  
## campaign -0.078797  
## f.pdaysf.pdays-(22,23] -0.300603  
## f.influentMonthf.influentMonth.sep-may-jul:campaign 0.028353  
## f.influentMonthf.influentMonth.mar-dec-oct-nov:campaign 0.022602  
## Std. Error t value  
## (Intercept) 0.073805 76.304  
## f.influentMonthf.influentMonth.sep-may-jul 0.040870 0.039  
## f.influentMonthf.influentMonth.mar-dec-oct-nov 0.063074 -1.232  
## campaign 0.008980 -8.775  
## f.pdaysf.pdays-(22,23] 0.069788 -4.307  
## f.influentMonthf.influentMonth.sep-may-jul:campaign 0.011493 2.467  
## f.influentMonthf.influentMonth.mar-dec-oct-nov:campaign 0.022954 0.985  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## f.influentMonthf.influentMonth.sep-may-jul 0.9692   
## f.influentMonthf.influentMonth.mar-dec-oct-nov 0.2180   
## campaign < 2e-16 \*\*\*  
## f.pdaysf.pdays-(22,23] 1.68e-05 \*\*\*  
## f.influentMonthf.influentMonth.sep-may-jul:campaign 0.0137 \*   
## f.influentMonthf.influentMonth.mar-dec-oct-nov:campaign 0.3248   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8992 on 4983 degrees of freedom  
## Multiple R-squared: 0.03173, Adjusted R-squared: 0.03057   
## F-statistic: 27.22 on 6 and 4983 DF, p-value: < 2.2e-16

Podemos ver que el nuevo modelo sin los individuos influyentes tiene una mejora en el r-square, aunque este sigue siendo muy bajo.

# Modelización con target binario

Empezamos dividiendo nuestra muestra en una muestra de trabajo y una muestra de testeo, para ello seleccionaremos aleatoriamente el 25% de la muestra para crear la muestra de testeo.

set.seed(19101990)  
sam <-sample(1:nrow(df),0.75\*nrow(df))   
  
dfw<-df[sam,]  
dft<-df[-sam,]

## Modelización con variables explicativas numéricas

### Modelo simple

Para empezar, hacemos un catdes con todas las variables numéricas para ver cuáles son las que están más relacionadas con nuestro target. Las utilizamos para hacer un modelo lineal general con variables explicativas numéricas. Este modelo es de la familia binomial ya que nuestro target es binario.

catdes(dfw[,c("y",vars\_num,"duration")],1)

##   
## Link between the cluster variable and the quantitative variables  
## ================================================================  
## Eta2 P-value  
## duration 0.176628371 4.526217e-160  
## nr.employed 0.107146172 3.620826e-94  
## pdays 0.098200453 4.721537e-86  
## euribor3m 0.077763912 8.384377e-68  
## emp.var.rate 0.072681695 2.518903e-63  
## previous 0.043535295 4.410367e-38  
## cons.price.idx 0.012743864 4.345713e-12  
## campaign 0.006955344 3.241195e-07  
## age 0.004418712 4.712507e-05  
## cons.conf.idx 0.003847937 1.464648e-04  
##   
## Description of each cluster by quantitative variables  
## =====================================================  
## $y.no  
## v.test Mean in category Overall mean sd in category  
## nr.employed 20.020835 5174.1172673 5165.6679316 66.5132140  
## pdays 19.166844 22.7627628 22.4142170 2.0015446  
## euribor3m 17.056225 3.7481483 3.5760190 1.6726437  
## emp.var.rate 16.489458 0.1990691 0.0485302 1.5112677  
## cons.price.idx 6.904694 93.5892261 93.5660259 0.5683953  
## campaign 5.100975 2.5990991 2.5299305 2.4284226  
## cons.conf.idx -3.794092 -40.6458859 -40.5445216 4.4151245  
## age -4.065760 39.8219219 40.0652058 9.8012900  
## previous -12.761878 0.1405405 0.1774452 0.4123568  
## duration -25.705383 219.4867868 255.9438803 196.7693288  
## Overall sd p.value  
## nr.employed 73.3850752 3.625944e-89  
## pdays 3.1621071 7.003107e-82  
## euribor3m 1.7548478 3.142184e-65  
## emp.var.rate 1.5874850 4.368514e-61  
## cons.price.idx 0.5842716 5.031186e-12  
## campaign 2.3578875 3.379091e-07  
## cons.conf.idx 4.6456264 1.481848e-04  
## age 10.4049243 4.787623e-05  
## previous 0.5028450 2.676679e-37  
## duration 246.6183107 1.017624e-145  
##   
## $y.yes  
## v.test Mean in category Overall mean sd in category  
## duration 25.705383 550.6092233 255.9438803 376.687736  
## previous 12.761878 0.4757282 0.1774452 0.906767  
## age 4.065760 42.0315534 40.0652058 14.230300  
## cons.conf.idx 3.794092 -39.7252427 -40.5445216 6.140692  
## campaign -5.100975 1.9708738 2.5299305 1.574717  
## cons.price.idx -6.904694 93.3785097 93.5660259 0.670650  
## emp.var.rate -16.489458 -1.1682039 0.0485302 1.662956  
## euribor3m -17.056225 2.1847791 3.5760190 1.783746  
## pdays -19.166844 19.5970874 22.4142170 7.036851  
## nr.employed -20.020835 5097.3759709 5165.6679316 88.965074  
## Overall sd p.value  
## duration 246.6183107 1.017624e-145  
## previous 0.5028450 2.676679e-37  
## age 10.4049243 4.787623e-05  
## cons.conf.idx 4.6456264 1.481848e-04  
## campaign 2.3578875 3.379091e-07  
## cons.price.idx 0.5842716 5.031186e-12  
## emp.var.rate 1.5874850 4.368514e-61  
## euribor3m 1.7548478 3.142184e-65  
## pdays 3.1621071 7.003107e-82  
## nr.employed 73.3850752 3.625944e-89

gm1<-glm( y ~   
 duration +  
 nr.employed +  
 pdays +  
 euribor3m +  
 emp.var.rate +  
 previous +  
 cons.price.idx +  
 campaign +  
 age +  
 cons.conf.idx, family = binomial, data = dfw)  
summary(gm1)

##   
## Call:  
## glm(formula = y ~ duration + nr.employed + pdays + euribor3m +   
## emp.var.rate + previous + cons.price.idx + campaign + age +   
## cons.conf.idx, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.6937 -0.3319 -0.1897 -0.1238 2.9794   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.530e+01 5.242e+01 -0.673 0.50070   
## duration 5.008e-03 2.518e-04 19.887 < 2e-16 \*\*\*  
## nr.employed -5.899e-03 4.944e-03 -1.193 0.23281   
## pdays -1.205e-01 1.702e-02 -7.079 1.45e-12 \*\*\*  
## euribor3m 3.016e-02 2.839e-01 0.106 0.91542   
## emp.var.rate -6.405e-01 2.289e-01 -2.797 0.00515 \*\*   
## previous -3.306e-01 1.316e-01 -2.512 0.01201 \*   
## cons.price.idx 6.955e-01 3.308e-01 2.102 0.03553 \*   
## campaign -1.280e-01 4.381e-02 -2.922 0.00348 \*\*   
## age 1.356e-02 5.361e-03 2.530 0.01141 \*   
## cons.conf.idx 3.157e-02 1.987e-02 1.589 0.11208   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2594.9 on 3741 degrees of freedom  
## Residual deviance: 1584.3 on 3731 degrees of freedom  
## AIC: 1606.3  
##   
## Number of Fisher Scoring iterations: 6

Anova(gm1)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## duration 563.62 1 < 2.2e-16 \*\*\*  
## nr.employed 1.44 1 0.230622   
## pdays 55.07 1 1.161e-13 \*\*\*  
## euribor3m 0.01 1 0.915436   
## emp.var.rate 7.77 1 0.005314 \*\*   
## previous 6.63 1 0.010024 \*   
## cons.price.idx 4.29 1 0.038251 \*   
## campaign 9.81 1 0.001740 \*\*   
## age 6.38 1 0.011547 \*   
## cons.conf.idx 2.52 1 0.112437   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Viendo el resultado de summary, podemos ver variables que tienen el p-value mayor a 0.1 (cons.cinf.idx, euribor3m), por lo que procedemos a quitarlas de nuestro modelo. Podemos ver que el deviance es inferior al null deviance.

gm2<-glm( y ~   
 duration +  
 nr.employed +  
 pdays +  
 emp.var.rate +  
 previous +  
 campaign +  
 age +  
 cons.conf.idx, family = binomial, data = dfw)  
vif(gm2)

## duration nr.employed pdays emp.var.rate previous   
## 1.283771 3.979725 1.829567 3.518890 2.048827   
## campaign age cons.conf.idx   
## 1.029761 1.037190 1.057214

Haciendo vif podemos ver que emp.var.rate tiene un valor mayor a 3, por lo que decidimos sacarla de nuestro modelo.

gm3<-glm( y ~   
 duration +  
 nr.employed +  
 pdays +  
 previous +  
 campaign +  
 age +  
 cons.conf.idx, family = binomial, data = dfw)  
vif(gm3)

## duration nr.employed pdays previous campaign   
## 1.241533 1.496925 1.820608 2.031438 1.021478   
## age cons.conf.idx   
## 1.034254 1.056345

### Modelo de regresión polinómica

Hacemos un tanteo aplicando una transformación polinómica de segundo grado a cada una de las variables.

gm4<-glm(y~  
 poly(duration,2) +  
 poly(nr.employed,2) +  
 poly(pdays,2) +  
 poly(previous,2) +  
 poly(campaign,2) +  
 poly(age,2) +  
 poly(cons.conf.idx,2), family = binomial, data = dfw  
 )  
summary(gm4)

##   
## Call:  
## glm(formula = y ~ poly(duration, 2) + poly(nr.employed, 2) +   
## poly(pdays, 2) + poly(previous, 2) + poly(campaign, 2) +   
## poly(age, 2) + poly(cons.conf.idx, 2), family = binomial,   
## data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8033 -0.3187 -0.1672 -0.1044 2.9358   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.3334 0.1132 -29.449 < 2e-16 \*\*\*  
## poly(duration, 2)1 82.4000 3.9367 20.931 < 2e-16 \*\*\*  
## poly(duration, 2)2 -24.3172 3.0845 -7.884 3.18e-15 \*\*\*  
## poly(nr.employed, 2)1 -61.5257 5.4420 -11.306 < 2e-16 \*\*\*  
## poly(nr.employed, 2)2 4.9535 3.9118 1.266 0.205411   
## poly(pdays, 2)1 -26.0379 3.5679 -7.298 2.92e-13 \*\*\*  
## poly(pdays, 2)2 -0.9880 2.5784 -0.383 0.701581   
## poly(previous, 2)1 -16.9490 4.7079 -3.600 0.000318 \*\*\*  
## poly(previous, 2)2 8.2807 3.2481 2.549 0.010790 \*   
## poly(campaign, 2)1 -12.3096 5.9479 -2.070 0.038493 \*   
## poly(campaign, 2)2 10.7221 5.9999 1.787 0.073929 .   
## poly(age, 2)1 5.7959 3.6339 1.595 0.110721   
## poly(age, 2)2 6.4103 3.3367 1.921 0.054714 .   
## poly(cons.conf.idx, 2)1 4.8461 3.7665 1.287 0.198225   
## poly(cons.conf.idx, 2)2 5.7061 3.7847 1.508 0.131638   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2594.9 on 3741 degrees of freedom  
## Residual deviance: 1523.1 on 3727 degrees of freedom  
## AIC: 1553.1  
##   
## Number of Fisher Scoring iterations: 7

En vista del summary, podemos omitir el termino cuadrático de las variables nr.employed, pdays, age, con.conf.idx.

gm5<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
summary(gm5)

##   
## Call:  
## glm(formula = y ~ poly(duration, 2) + nr.employed + pdays + poly(previous,   
## 2) + poly(campaign, 2) + age + cons.conf.idx, family = binomial,   
## data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8321 -0.3202 -0.1672 -0.1022 2.9323   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 79.166461 5.222892 15.158 < 2e-16 \*\*\*  
## poly(duration, 2)1 82.458219 3.923723 21.015 < 2e-16 \*\*\*  
## poly(duration, 2)2 -23.887690 3.100519 -7.704 1.31e-14 \*\*\*  
## nr.employed -0.015261 0.001019 -14.975 < 2e-16 \*\*\*  
## pdays -0.133120 0.017975 -7.406 1.30e-13 \*\*\*  
## poly(previous, 2)1 -15.732038 4.526575 -3.475 0.00051 \*\*\*  
## poly(previous, 2)2 8.049464 3.229608 2.492 0.01269 \*   
## poly(campaign, 2)1 -12.349248 5.936024 -2.080 0.03749 \*   
## poly(campaign, 2)2 10.937572 6.001919 1.822 0.06840 .   
## age 0.011624 0.005552 2.094 0.03629 \*   
## cons.conf.idx 0.028578 0.011999 2.382 0.01724 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2594.9 on 3741 degrees of freedom  
## Residual deviance: 1530.6 on 3731 degrees of freedom  
## AIC: 1552.6  
##   
## Number of Fisher Scoring iterations: 7

Anova(gm5)

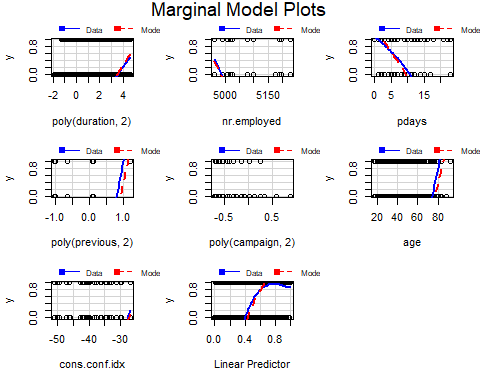
## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## poly(duration, 2) 612.00 2 < 2.2e-16 \*\*\*  
## nr.employed 255.09 1 < 2.2e-16 \*\*\*  
## pdays 59.71 1 1.102e-14 \*\*\*  
## poly(previous, 2) 16.12 2 0.0003158 \*\*\*  
## poly(campaign, 2) 12.48 2 0.0019487 \*\*   
## age 4.38 1 0.0364587 \*   
## cons.conf.idx 5.67 1 0.0172915 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

vif(gm5)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(duration, 2) 1.377914 2 1.083442  
## nr.employed 1.641804 1 1.281329  
## pdays 1.808186 1 1.344688  
## poly(previous, 2) 2.079898 2 1.200910  
## poly(campaign, 2) 1.044713 2 1.010996  
## age 1.037165 1 1.018413  
## cons.conf.idx 1.060481 1 1.029796

marginalModelPlots(gm5)

## Warning in mmps(...): Splines and/or polynomials replaced by a fitted  
## linear combination

 Podemos ver que los p-values son inferiores a 0.1 para todas las variables, también vemos que el resultado del vif no presenta colinealidad.

Generalmente podemos ver que el modelo no se acerca tanto a los valores reales.

## Modelización con variables explicativas numéricas y categóricas

# duration y f.duration  
gm6<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
gm7<-glm(y~  
 f.duration +  
 nr.employed +  
 pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
BIC(gm7,gm6)

## df BIC  
## gm7 12 1847.037  
## gm6 11 1621.054

# pdays y f.pdays  
gm8<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
gm9<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 f.pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
BIC(gm9,gm8)

## df BIC  
## gm9 11 1620.968  
## gm8 11 1621.054

# previous y f.previous  
gm10<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
gm11<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 pdays +  
 f.previous +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
BIC(gm11,gm10)

## df BIC  
## gm11 11 1621.220  
## gm10 11 1621.054

# campaign vs f.campaign  
gm12<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
gm13<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 pdays +  
 poly(previous,2) +  
 f.campaign +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
BIC(gm13,gm12)

## df BIC  
## gm13 11 1624.202  
## gm12 11 1621.054

# age vs f.age  
gm14<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
gm15<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 f.age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
BIC(gm15,gm14)

## df BIC  
## gm15 13 1630.939  
## gm14 11 1621.054

A partir de los resultados de los BICs, nos quedamos con las versiones de las variables numéricas o de factores cuyo valor de BIC es menor.

Con el resultado obtenido anteriormente, creamos un nuevo modelo.

gm16<-glm(y~  
 poly(duration,2) +  
 nr.employed +  
 f.pdays +  
 poly(previous,2) +  
 poly(campaign,2) +  
 age +  
 cons.conf.idx, family = binomial, data = dfw  
 )  
summary(gm16)

##   
## Call:  
## glm(formula = y ~ poly(duration, 2) + nr.employed + f.pdays +   
## poly(previous, 2) + poly(campaign, 2) + age + cons.conf.idx,   
## family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8290 -0.3203 -0.1678 -0.1034 2.9250   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 77.867008 5.208078 14.951 < 2e-16 \*\*\*  
## poly(duration, 2)1 82.135385 3.916635 20.971 < 2e-16 \*\*\*  
## poly(duration, 2)2 -23.637806 3.097309 -7.632 2.32e-14 \*\*\*  
## nr.employed -0.015156 0.001018 -14.892 < 2e-16 \*\*\*  
## f.pdaysf.pdays-(22,23] -2.326819 0.314062 -7.409 1.27e-13 \*\*\*  
## poly(previous, 2)1 -16.405010 4.606619 -3.561 0.000369 \*\*\*  
## poly(previous, 2)2 9.381292 3.275937 2.864 0.004187 \*\*   
## poly(campaign, 2)1 -12.215658 5.905733 -2.068 0.038599 \*   
## poly(campaign, 2)2 10.952908 5.964041 1.836 0.066285 .   
## age 0.012244 0.005557 2.203 0.027560 \*   
## cons.conf.idx 0.028649 0.011975 2.392 0.016738 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2594.9 on 3741 degrees of freedom  
## Residual deviance: 1530.5 on 3731 degrees of freedom  
## AIC: 1552.5  
##   
## Number of Fisher Scoring iterations: 7

Anova(gm16)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## poly(duration, 2) 608.02 2 < 2.2e-16 \*\*\*  
## nr.employed 251.42 1 < 2.2e-16 \*\*\*  
## f.pdays 59.79 1 1.055e-14 \*\*\*  
## poly(previous, 2) 17.95 2 0.0001263 \*\*\*  
## poly(campaign, 2) 12.34 2 0.0020925 \*\*   
## age 4.85 1 0.0276839 \*   
## cons.conf.idx 5.72 1 0.0167883 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

vif(gm16)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(duration, 2) 1.374363 2 1.082743  
## nr.employed 1.632340 1 1.277631  
## f.pdays 1.911480 1 1.382563  
## poly(previous, 2) 2.167088 2 1.213303  
## poly(campaign, 2) 1.044674 2 1.010986  
## age 1.037314 1 1.018486  
## cons.conf.idx 1.057644 1 1.028418

Comprobamos el resultado y son correctos.

Ahora añadimos el resto de factores, utilizamos un catdes para ver cuales están más relacionadas con nuestro target.

catdes(dfw[,c("y",vars\_cat)],1)

##   
## Link between the cluster variable and the categorical variables (chi-square test)  
## =================================================================================  
## p.value df  
## poutcome 8.905412e-84 2  
## month 9.664852e-42 9  
## job 5.444613e-17 10  
## contact 2.379692e-14 1  
## default 1.290603e-09 1  
## marital 8.221254e-04 2  
## housing 4.148800e-03 1  
## education 8.424425e-03 6  
##   
## Description of each cluster by the categories  
## =============================================  
## $y.no  
## Cla/Mod Mod/Cla Global  
## poutcome=poutcome.nonexistent 90.93168 87.9279279 86.0502405  
## contact=contact.telephone 94.20074 38.0480480 35.9433458  
## default=default.unknown 94.96855 22.6726727 21.2453234  
## month=month.may 92.54210 34.6546547 33.3244254  
## job=job.blue-collar 93.39513 24.2042042 23.0625334  
## job=job.services 94.67456 9.6096096 9.0326029  
## marital=marital.married 90.30817 60.7207207 59.8343132  
## education=education.basic.9y 92.32026 16.9669670 16.3548904  
## housing=housing.no 90.57259 47.0270270 46.2052378  
## month=month.dec 66.66667 0.3003003 0.4008552  
## housing=housing.yes 87.63040 52.9729730 53.7947622  
## education=education.university.degree 86.43042 30.0300300 30.9192945  
## marital=marital.single 85.95506 27.5675676 28.5408872  
## month=month.apr 80.08850 5.4354354 6.0395510  
## job=job.student 70.65217 1.9519520 2.4585783  
## month=month.sep 62.74510 0.9609610 1.3629075  
## job=job.retired 74.56647 3.8738739 4.6231962  
## month=month.mar 54.16667 0.7807808 1.2827365  
## default=default.no 87.37699 77.3273273 78.7546766  
## month=month.oct 54.83871 1.0210210 1.6568680  
## contact=contact.cellular 86.06592 61.9519520 64.0566542  
## poutcome=poutcome.success 33.33333 1.1711712 3.1266702  
## p.value v.test  
## poutcome=poutcome.nonexistent 1.033863e-17 8.570110  
## contact=contact.telephone 1.723437e-15 7.959781  
## default=default.unknown 6.533397e-11 6.530995  
## month=month.may 4.627585e-07 5.041143  
## job=job.blue-collar 7.948561e-07 4.936626  
## job=job.services 1.615234e-04 3.772649  
## marital=marital.married 1.803002e-03 3.120898  
## education=education.basic.9y 2.923747e-03 2.975643  
## housing=housing.no 4.058668e-03 2.873566  
## month=month.dec 2.219854e-02 -2.286953  
## housing=housing.yes 4.058668e-03 -2.873566  
## education=education.university.degree 9.941775e-04 -3.292169  
## marital=marital.single 2.417762e-04 -3.670817  
## month=month.apr 5.137216e-05 -4.049295  
## job=job.student 1.134551e-06 -4.866738  
## month=month.sep 8.387879e-07 -4.926119  
## job=job.retired 4.689017e-08 -5.462717  
## month=month.mar 9.910614e-10 -6.110843  
## default=default.no 6.533397e-11 -6.530995  
## month=month.oct 7.066133e-12 -6.856311  
## contact=contact.cellular 1.723437e-15 -7.959781  
## poutcome=poutcome.success 7.526687e-49 -14.689500  
##   
## $y.yes  
## Cla/Mod Mod/Cla Global  
## poutcome=poutcome.success 66.666667 18.932039 3.1266702  
## contact=contact.cellular 13.934084 81.067961 64.0566542  
## month=month.oct 45.161290 6.796117 1.6568680  
## default=default.no 12.623006 90.291262 78.7546766  
## month=month.mar 45.833333 5.339806 1.2827365  
## job=job.retired 25.433526 10.679612 4.6231962  
## month=month.sep 37.254902 4.611650 1.3629075  
## job=job.student 29.347826 6.553398 2.4585783  
## month=month.apr 19.911504 10.922330 6.0395510  
## marital=marital.single 14.044944 36.407767 28.5408872  
## education=education.university.degree 13.569576 38.106796 30.9192945  
## housing=housing.yes 12.369598 60.436893 53.7947622  
## month=month.dec 33.333333 1.213592 0.4008552  
## housing=housing.no 9.427415 39.563107 46.2052378  
## education=education.basic.9y 7.679739 11.407767 16.3548904  
## marital=marital.married 9.691827 52.669903 59.8343132  
## job=job.services 5.325444 4.368932 9.0326029  
## job=job.blue-collar 6.604867 13.834951 23.0625334  
## month=month.may 7.457899 22.572816 33.3244254  
## default=default.unknown 5.031447 9.708738 21.2453234  
## contact=contact.telephone 5.799257 18.932039 35.9433458  
## poutcome=poutcome.nonexistent 9.068323 70.873786 86.0502405  
## p.value v.test  
## poutcome=poutcome.success 7.526687e-49 14.689500  
## contact=contact.cellular 1.723437e-15 7.959781  
## month=month.oct 7.066133e-12 6.856311  
## default=default.no 6.533397e-11 6.530995  
## month=month.mar 9.910614e-10 6.110843  
## job=job.retired 4.689017e-08 5.462717  
## month=month.sep 8.387879e-07 4.926119  
## job=job.student 1.134551e-06 4.866738  
## month=month.apr 5.137216e-05 4.049295  
## marital=marital.single 2.417762e-04 3.670817  
## education=education.university.degree 9.941775e-04 3.292169  
## housing=housing.yes 4.058668e-03 2.873566  
## month=month.dec 2.219854e-02 2.286953  
## housing=housing.no 4.058668e-03 -2.873566  
## education=education.basic.9y 2.923747e-03 -2.975643  
## marital=marital.married 1.803002e-03 -3.120898  
## job=job.services 1.615234e-04 -3.772649  
## job=job.blue-collar 7.948561e-07 -4.936626  
## month=month.may 4.627585e-07 -5.041143  
## default=default.unknown 6.533397e-11 -6.530995  
## contact=contact.telephone 1.723437e-15 -7.959781  
## poutcome=poutcome.nonexistent 1.033863e-17 -8.570110

Viendo el resultado del catdes, obtenemos que las variables que están más relacionadas son outcome, month, job, contact, default, marital, housing y education. Como month tiene muchos niveles decidimos usar el month factorizado.

gm17<-glm(y~poly(duration,2) +nr.employed +f.pdays +poly(previous,2) +poly(campaign,2) +age +cons.conf.idx+poutcome+ f.influentMonth + job+ contact+ default+ marital+ housing+ education, family = binomial, data = dfw)  
Anova(gm17)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## poly(duration, 2) 625.07 2 < 2.2e-16 \*\*\*  
## nr.employed 177.78 1 < 2.2e-16 \*\*\*  
## f.pdays 0.02 1 0.895857   
## poly(previous, 2) 1.21 2 0.546061   
## poly(campaign, 2) 9.00 2 0.011102 \*   
## age 6.51 1 0.010722 \*   
## cons.conf.idx 1.54 1 0.214577   
## poutcome 9.22 2 0.009954 \*\*   
## f.influentMonth 12.92 2 0.001564 \*\*   
## job 12.51 10 0.252449   
## contact 3.60 1 0.057727 .   
## default 5.75 1 0.016536 \*   
## marital 4.60 2 0.100507   
## housing 3.43 1 0.063919 .   
## education 5.17 6 0.522155   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Seguimos cribando dado el resultado del Anova

gm18<-glm(y~poly(duration,2) +nr.employed +poly(campaign,2) +age +poutcome+ f.influentMonth + contact+ default+ housing, family = binomial, data = dfw)  
Anova(gm18)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## poly(duration, 2) 623.25 2 < 2.2e-16 \*\*\*  
## nr.employed 213.89 1 < 2.2e-16 \*\*\*  
## poly(campaign, 2) 9.65 2 0.008020 \*\*   
## age 8.49 1 0.003566 \*\*   
## poutcome 70.85 2 4.116e-16 \*\*\*  
## f.influentMonth 20.46 2 3.614e-05 \*\*\*  
## contact 3.80 1 0.051222 .   
## default 8.33 1 0.003909 \*\*   
## housing 3.21 1 0.073099 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

vif(gm18)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(duration, 2) 1.447940 2 1.096952  
## nr.employed 1.721167 1 1.311932  
## poly(campaign, 2) 1.056114 2 1.013743  
## age 1.033641 1 1.016681  
## poutcome 1.296327 2 1.067035  
## f.influentMonth 1.106236 2 1.025562  
## contact 1.123197 1 1.059810  
## default 1.075544 1 1.037084  
## housing 1.010610 1 1.005291

Ahora las variables nos dan aceptables, con p-values menores a 0.1 y sin colinealidad.

Ahora hacemos un step con el criterio bayesiano, para validar el modelo

gm19<-step(gm18,k=log(nrow(dfw)))

## Start: AIC=1601.84  
## y ~ poly(duration, 2) + nr.employed + poly(campaign, 2) + age +   
## poutcome + f.influentMonth + contact + default + housing  
##   
## Df Deviance AIC  
## - poly(campaign, 2) 2 1496.3 1595.0  
## - housing 1 1489.9 1596.8  
## - contact 1 1490.5 1597.4  
## <none> 1486.7 1601.8  
## - default 1 1495.0 1601.9  
## - age 1 1495.1 1602.1  
## - f.influentMonth 2 1507.1 1605.8  
## - poutcome 2 1557.5 1656.2  
## - nr.employed 1 1700.5 1807.5  
## - poly(duration, 2) 2 2109.9 2208.6  
##   
## Step: AIC=1595.03  
## y ~ poly(duration, 2) + nr.employed + age + poutcome + f.influentMonth +   
## contact + default + housing  
##   
## Df Deviance AIC  
## - housing 1 1499.3 1589.8  
## - contact 1 1500.8 1591.3  
## <none> 1496.3 1595.0  
## - age 1 1505.1 1595.6  
## - default 1 1505.2 1595.7  
## - f.influentMonth 2 1518.2 1600.4  
## - poutcome 2 1567.5 1649.7  
## - nr.employed 1 1717.9 1808.4  
## - poly(duration, 2) 2 2117.1 2199.4  
##   
## Step: AIC=1589.77  
## y ~ poly(duration, 2) + nr.employed + age + poutcome + f.influentMonth +   
## contact + default  
##   
## Df Deviance AIC  
## - contact 1 1504.1 1586.4  
## <none> 1499.3 1589.8  
## - age 1 1507.6 1589.8  
## - default 1 1508.4 1590.7  
## - f.influentMonth 2 1522.0 1596.1  
## - poutcome 2 1570.1 1644.2  
## - nr.employed 1 1721.4 1803.7  
## - poly(duration, 2) 2 2120.4 2194.5  
##   
## Step: AIC=1586.4  
## y ~ poly(duration, 2) + nr.employed + age + poutcome + f.influentMonth +   
## default  
##   
## Df Deviance AIC  
## - age 1 1512.3 1586.3  
## <none> 1504.1 1586.4  
## - default 1 1514.4 1588.5  
## - f.influentMonth 2 1529.8 1595.6  
## - poutcome 2 1573.6 1639.5  
## - nr.employed 1 1747.3 1821.3  
## - poly(duration, 2) 2 2128.2 2194.0  
##   
## Step: AIC=1586.33  
## y ~ poly(duration, 2) + nr.employed + poutcome + f.influentMonth +   
## default  
##   
## Df Deviance AIC  
## <none> 1512.3 1586.3  
## - default 1 1520.5 1586.3  
## - f.influentMonth 2 1539.5 1597.1  
## - poutcome 2 1582.0 1639.6  
## - nr.employed 1 1767.7 1833.5  
## - poly(duration, 2) 2 2136.3 2193.9

summary(gm19)

##   
## Call:  
## glm(formula = y ~ poly(duration, 2) + nr.employed + poutcome +   
## f.influentMonth + default, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6811 -0.3085 -0.1597 -0.1003 3.0735   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 75.15851 5.20488  
## poly(duration, 2)1 83.77127 3.98039  
## poly(duration, 2)2 -24.63315 3.12149  
## nr.employed -0.01528 0.00103  
## poutcomepoutcome.nonexistent 0.89824 0.22100  
## poutcomepoutcome.success 2.40968 0.30331  
## f.influentMonthf.influentMonth.sep-may-jul -0.68808 0.15291  
## f.influentMonthf.influentMonth.mar-dec-oct-nov 0.06799 0.19438  
## defaultdefault.unknown -0.57789 0.20843  
## z value Pr(>|z|)   
## (Intercept) 14.440 < 2e-16 \*\*\*  
## poly(duration, 2)1 21.046 < 2e-16 \*\*\*  
## poly(duration, 2)2 -7.891 2.99e-15 \*\*\*  
## nr.employed -14.831 < 2e-16 \*\*\*  
## poutcomepoutcome.nonexistent 4.064 4.81e-05 \*\*\*  
## poutcomepoutcome.success 7.945 1.95e-15 \*\*\*  
## f.influentMonthf.influentMonth.sep-may-jul -4.500 6.80e-06 \*\*\*  
## f.influentMonthf.influentMonth.mar-dec-oct-nov 0.350 0.72651   
## defaultdefault.unknown -2.773 0.00556 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2594.9 on 3741 degrees of freedom  
## Residual deviance: 1512.3 on 3733 degrees of freedom  
## AIC: 1530.3  
##   
## Number of Fisher Scoring iterations: 7

Hay que ver que todos los coeficientes sean calculables y que no tengamos ningún NA en el summary, en nuestro caso no tenemos ninguno.

## Interacciones

Primero probamos con todas las interacciones posibles de orden 2 para hacernos una idea de las interacciones que podemos usar de muestra.

gm20<-glm(y~ (poly(duration,2) +nr.employed +poly(campaign,2) +age +poutcome+ f.influentMonth + contact+ default+ housing)^2, family = binomial, data = dfw)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Anova(gm20)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## poly(duration, 2) 606.44 2 < 2.2e-16 \*\*\*  
## nr.employed 219.61 1 < 2.2e-16 \*\*\*  
## poly(campaign, 2) 11.93 2 0.0025650 \*\*   
## age 7.41 1 0.0064886 \*\*   
## poutcome 63.00 2 2.083e-14 \*\*\*  
## f.influentMonth 16.74 2 0.0002317 \*\*\*  
## contact 6.26 1 0.0123735 \*   
## default 8.37 1 0.0038152 \*\*   
## housing 4.99 1 0.0255125 \*   
## poly(duration, 2):nr.employed 12.60 2 0.0018372 \*\*   
## poly(duration, 2):poly(campaign, 2) 2.17 4 0.7042833   
## poly(duration, 2):age 0.20 2 0.9040617   
## poly(duration, 2):poutcome 3.63 4 0.4578341   
## poly(duration, 2):f.influentMonth 16.14 4 0.0028348 \*\*   
## poly(duration, 2):contact 0.70 2 0.7038259   
## poly(duration, 2):default 1.43 2 0.4886817   
## poly(duration, 2):housing 6.20 2 0.0450133 \*   
## nr.employed:poly(campaign, 2) 2.43 2 0.2968249   
## nr.employed:age 0.00 1 0.9476946   
## nr.employed:poutcome 5.67 2 0.0588327 .   
## nr.employed:f.influentMonth 0.84 2 0.6558761   
## nr.employed:contact 0.01 1 0.9381053   
## nr.employed:default 0.48 1 0.4868526   
## nr.employed:housing 0.02 1 0.8895965   
## poly(campaign, 2):age 8.90 2 0.0116869 \*   
## poly(campaign, 2):poutcome 13.38 4 0.0095456 \*\*   
## poly(campaign, 2):f.influentMonth 18.34 4 0.0010595 \*\*   
## poly(campaign, 2):contact 1.19 2 0.5509442   
## poly(campaign, 2):default 0.06 2 0.9705383   
## poly(campaign, 2):housing 3.48 2 0.1755176   
## age:poutcome 1.69 2 0.4304841   
## age:f.influentMonth 0.46 2 0.7937762   
## age:contact 0.15 1 0.6986240   
## age:default 0.83 1 0.3622380   
## age:housing 3.63 1 0.0567390 .   
## poutcome:f.influentMonth 7.12 4 0.1297656   
## poutcome:contact 3.67 2 0.1594200   
## poutcome:default 3.36 2 0.1866776   
## poutcome:housing 1.76 2 0.4153831   
## f.influentMonth:contact 9.26 2 0.0097760 \*\*   
## f.influentMonth:default 4.77 2 0.0920652 .   
## f.influentMonth:housing 2.12 2 0.3465305   
## contact:default 0.64 1 0.4233159   
## contact:housing 0.30 1 0.5831223   
## default:housing 3.33 1 0.0678341 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Elegiremos una interacción factor por factor y factor por numérica de entre todas las interacciones, cuyos p-values son los más convincentes. Comprobamos la bondad de la interacción factor

gm21<-glm(y~ poly(duration,2) +nr.employed +poly(campaign,2) +age +poutcome+ f.influentMonth\*contact+ default+ housing, family = binomial, data = dfw)  
  
gm22<-glm(y~ poly(duration,2) +nr.employed +poly(campaign,2)\*f.influentMonth +age +poutcome+ contact+ default+ housing, family = binomial, data = dfw)  
  
BIC(gm21,gm20)

## df BIC  
## gm21 16 1613.523  
## gm20 88 2052.909

BIC(gm22,gm20)

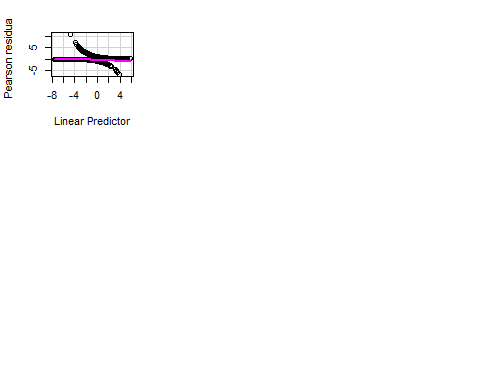
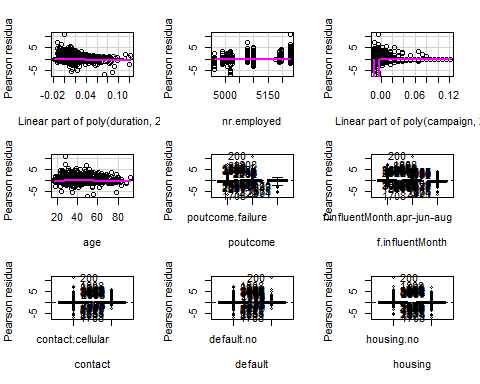
## df BIC  
## gm22 18 1618.316  
## gm20 88 2052.909

Vemos que los dos modelos con interacciones dan mejor que nuestro modelo, entre estos vemos que el de menor BIC es el de la interacción de f.influentMonth\*contact.

## Validación

Para la validación analizamos los gráficos

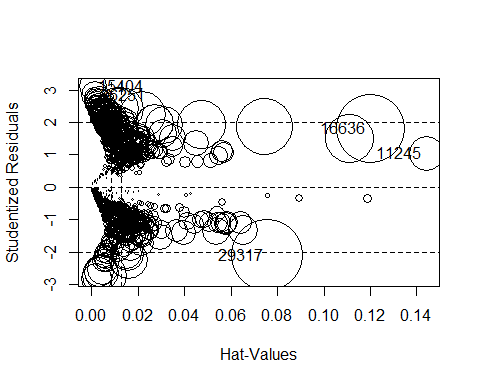
residualPlots(gm21)



## Test stat Pr(>|Test stat|)   
## poly(duration, 2)   
## nr.employed 0.8698 0.35102   
## poly(campaign, 2)   
## age 3.4946 0.06157 .  
## poutcome   
## f.influentMonth   
## contact   
## default   
## housing   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Con el residualPlots, podemos ver que tenemos una observación en común que es muy influente, como ahora vamos a hacer el influencePlot, podremos determinar si efectivamente esta observación es demasiado influyente.

influencePlot(gm21)



## StudRes Hat CookD  
## 15404 3.116369 0.0006512188 0.004991063  
## 36251 2.835739 0.0015232026 0.005018080  
## 11245 1.043963 0.1440726950 0.007802044  
## 29317 -2.103560 0.0758185513 0.033909413  
## 16636 1.817313 0.1200968043 0.030955448

which(row.names(df)==11245)

## [1] 1317

which(row.names(df)==16636)

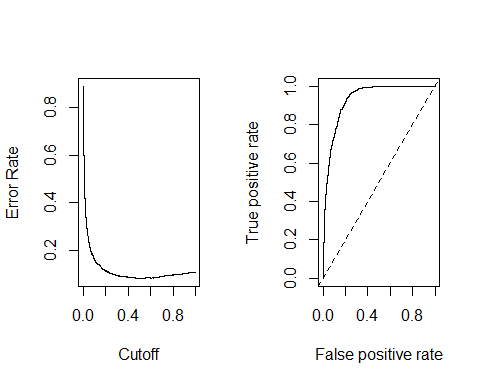
## [1] 1940

which(row.names(df)==29317)

## [1] 3498

Viendo el resultado del influentPlot, no vemos al individuo 200, que nos sale en la gráfica de residuos, lo que nos puede decir que no influye demasiado en nuestro modelo.

dataroc<-prediction(predict(gm21, type="response"),dfw$y)  
par(mfrow=c(1,2))  
plot(performance(dataroc,"err"))  
plot(performance(dataroc,"tpr","fpr"))  
abline(0,1,lty=2)



Estamos cogiendo las betas de este modelo y aplicándolos a las variables explicativas del dft, para así obtener las predicciones según nuestro modelo. Montamos una tabla con las predicciones y los datos reales a modo de matriz de confusión, del cual su diagonal nos indica la cantidad de aciertos.

p<-factor(ifelse(predict(gm21, dft, type = "response") < 0.4, 0, 1 ))  
tabConfusion<-table(p, dft[,"y"])

Para calcular la capacidad predictiva del modelo, bastará con sumar la diagonal de la matriz de confusión y dividirla entre el número de observaciones.

capacidadPredictiva <- (tabConfusion[1,1] + tabConfusion[2,2])/nrow(dft)

Tenemos un 91,42% de aciertos con nuestro modelo.

Nos damos cuenta que por los datos que tenemos no es posible que tengamos una capacidad predictiva tan grande, por lo que decidimos comparar con el modelo null.

gmnull<-glm(y~ 1, family = binomial, data = dfw)  
pnull<-factor(ifelse(predict(gmnull, dft, type = "response") < 0.4, 0, 1 ))  
tabConfusionNull<-table(pnull, dft[,"y"])  
capacidadPredictivaNull <- (tabConfusionNull[1,1] + 0)/nrow(dft)

Con el modelo Null tenemos un 89,58% de aciertos, ahora viendo la diferencia entre nuestro modelo y el null tenemos que

MejoraModelo <- capacidadPredictiva - capacidadPredictivaNull  
MejoraModelo\*100

## [1] 1.842949

Tenemos que nuestro modelo es 1.84% mejor que el modelo más básico. El hecho de que la capacidad predictiva sea tan alta en ambos casos, es debido a que la gran mayoría de las observaciones tienen como valor de respuesta “no”, esto hace que cualquier modelo por tonto que sea tenga una buena capacidad predictiva.