FINAL Deliverable

Montserrrat Martinez i Aleix Costa

11 de Juny de 2019

INDEX

[**Package loading and set Working directory**](#_heading=h.88mk7h7udhrx) **5**

[**Loading data**](#_heading=h.9cakb8ghxxfx) **5**

[**Upload and select data**](#_heading=h.e95lb88qinp1) **5**

[**Inicialització dels vectors de missings, errors i outliers**](#_heading=h.efnlewgc85ax) **7**

[**Univariate Descriptive Analysis & Data Quality Report**](#_heading=h.2n43e99i9sd5) **7**

[Qualitative Variables (Factors) / Categorical](#_heading=h.qsrfm8pa99ad) 7

[Quantitative Variables (Numerical)](#_heading=h.6l5v1k6tz20c) 16

[CONTAR NA’s](#_heading=h.qspk92j845af) 30

[Imputation](#_heading=h.atqmbrhnqoxo) 33

[Discretitzation](#_heading=h.q1wziycbg9xq) 37

[PROFILING](#_heading=h.2pz60krnacn4) 49

[Numeric target (Duration)](#_heading=h.j272vbl1a9sa) 49

[Y (target qual)](#_heading=h.chstwa733rtf) 52

[**Principal Component Analysis (PCA)**](#_heading=h.a1atvv7fln8q) **56**

[Data format & analysis](#_heading=h.55ihkmm5u3xr) 57

[Create PCA](#_heading=h.vdhre0gikrcb) 57

[Eigenvalues and dominant axes Analysis](#_heading=h.6tlrcdy8s1iq) 59

[Kaiser Rule](#_heading=h.fi8nnio30iys) 59

[Elbow Rule](#_heading=h.g2s6533uc94f) 60

[Individuals point of view](#_heading=h.f09k7q5ss1jl) 61

[Individuals contribution](#_heading=h.il3fapuytxf1) 61

[Individuals best representation](#_heading=h.pgvk8s63g5c4) 65

[Variables contribution](#_heading=h.npsqflv5fcp5) 67

[Interpreting the axes](#_heading=h.4lxm1tmerptu) 68

[K-Means Classification](#_heading=h.6gfetca9gatt) 71

[Descripició dels clusters](#_heading=h.ksgt18sjmnh2) 72

[Hierarchical Clustering](#_heading=h.o8d6dxynolus) 88

[Descripcio dels clusters](#_heading=h.cux5vj2qiw2r) 92

[Correspondence Analysis (CA)](#_heading=h.k4tzqj2v5ol) 96

[Factor\_age i Factor\_duration](#_heading=h.hfxr6d4s9ueq) 96

[Eigenvalues and dominant axes analysis](#_heading=h.5oiwx0mcah41) 100

[Job i Factor\_duration](#_heading=h.23fdiayhmuqj) 103

[Eigenvalues and dominant axes analysis](#_heading=h.iutst4vot94s) 110

[Model construction only with numeric explanatory variables](#_heading=h.rxvme55m1bov) 114

[Multivariant Data Analysis](#_heading=h.7xq4ys817p1p) 114

[Model Construction](#_heading=h.aci4zfwws5gr) 114

[Initial modelling](#_heading=h.ukbqgc1gbxnl) 114

[Transforming variables](#_heading=h.pj2klpntvfk6) 121

[Adding factors as explanatory variables](#_heading=h.8acakkrmqkto) 123

[**Interactions between numeric variables and factors**](#_heading=h.ecx4rvv1k20t) **135**

[Model Additiu](#_heading=h.8dpjni5dn4j1) 135

[Model Interaccions](#_heading=h.poigxgw0claf) 136

[Binary Regression](#_heading=h.1naohl5eootd) 138

[Explanatory numeric variables](#_heading=h.4imdlccasd3p) 138

[**Initial modelling**](#_heading=h.g5m76g3hqoru) **138**

[**Transforming variables**](#_heading=h.154e0ptfywsl) **143**

[**Adding Factors**](#_heading=h.a4htwx5l2jyo) **144**

[**Adding new factors**](#_heading=h.9haxw2dv2jap) **146**

[**Add to the best model: INTERACTIONS**](#_heading=h.ybfkcyq1n73t) **165**

[Model final](#_heading=h.8y8gx722cbn6) 169

[Interpretació del model final](#_heading=h.sabjj5rbu9q8) 171

[**Validació del model**](#_heading=h.1tadf89xhour) **171**

[Anàlisi dels residus](#_heading=h.95a5yi9q7bqo) 171

[**Predicció**](#_heading=h.jeadq3h4xbsa) **173**

[WORK](#_heading=h.nnganwltt8dm) 173

[TEST](#_heading=h.igvp1dcn0br5) 174

*----------- DELIVERABLE 1 ------------*

*Input variables:*

1. age (numeric)
2. job : type of job (categorical: ‘admin.’,‘blue-collar’,‘entrepreneur’,‘housemaid’,‘management’,‘retired’,‘self-employed’,‘services’,‘student’,‘technician’,‘unemployed’,‘unknown’)
3. marital : marital status (categorical: ‘divorced’,‘married’,‘single’,‘unknown’; note: ‘divorced’ means divorced or widowed)
4. education (categorical:‘basic.4y’,‘basic.6y’,‘basic.9y’,‘high.school’,‘illiterate’,‘professional.course’,‘university.degree’,‘unknown’)
5. default: has credit in default? (categorical: ‘no’,‘yes’,‘unknown’)
6. housing: has housing loan? (categorical: ‘no’,‘yes’,‘unknown’)
7. loan: has personal loan? (categorical: ‘no’,‘yes’,‘unknown’)# related with the last contact of the current campaign:
8. contact: contact communication type (categorical:‘cellular’,‘telephone’)
9. month: last contact month of year (categorical: ‘jan’, ‘feb’, ‘mar’,…, ‘nov’, ‘dec’)
10. day\_of\_week: last contact day of the week (categorical:‘mon’,‘tue’,‘wed’,‘thu’,‘fri’)
11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y=‘no’). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14. previous: number of contacts performed before this campaign and for this client (numeric)
15. poutcome: outcome of the previous marketing campaign (categorical: ‘failure’,‘nonexistent’,‘success’)# social and economic context attributes
16. emp.var.rate: employment variation rate - quarterly indicator (numeric)
17. cons.price.idx: consumer price index - monthly indicator (numeric)
18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19. euribor3m: euribor 3 month rate - daily indicator (numeric)
20. nr.employed: number of employees - quarterly indicator (numeric)
21. y - has the client subscribed a term deposit? (binary: ‘yes’,‘no’)

# Package loading and set Working directory

Carreguem els paquets necessaris i definim el nostre directori de treball

# Loading data

# Upload and select data

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

*#setwd("C:/Users/montserrat.martinez.santamaria/Documents/ADEI/bank-additional/bank-additional")*  
*#dirwd<-"C:/Users/montserrat.martinez.santamaria/Documents/ADEI/bank-additional/bank-additional"*  
  
**setwd**("/Users/montsee/Desktop/ADEI/bank-additional/bank-additional")  
dirwd<-"/Users/montsee/Desktop/ADEI/bank-additional/bank-additional"  
  
*# Data file already*  
  
df<-**read.table**(**paste0**(dirwd,"/bank-additional-full.csv"),header=TRUE,sep=";",na.strings = "999")  
  
*# Select your 5000 register sample (random sample)*  
  
*#nrow(df)*  
*#ncol(df)*  
*#dim(df)*   
  
**set.seed**(25071997)  
mostra<-**as.vector**(**sort**(**sample**(1**:nrow**(df),5000)))  
df<-df[mostra,]  
  
*#Verificacio i guardat de la mostra*  
  
**dim**(df) *#Mostra la dimensi? de la mostra*

## [1] 5000 21

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

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

**summary**(df)

## age job marital   
## Min. :17.00 admin. :1315 divorced: 574   
## 1st Qu.:32.00 blue-collar:1157 married :3029   
## Median :38.00 technician : 789 single :1390   
## Mean :40.16 services : 477 unknown : 7   
## 3rd Qu.:47.00 management : 348   
## Max. :98.00 retired : 212   
## (Other) : 702   
## education default housing loan   
## university.degree :1503 no :3958 no :2206 no :4055   
## high.school :1133 unknown:1042 unknown: 129 unknown: 129   
## basic.9y : 765 yes : 0 yes :2665 yes : 816   
## professional.course: 600   
## basic.4y : 514   
## basic.6y : 268   
## (Other) : 217   
## contact month day\_of\_week duration   
## cellular :3148 may :1633 fri: 979 Min. : 1.0   
## telephone:1852 jul : 911 mon:1039 1st Qu.: 102.0   
## aug : 754 thu:1064 Median : 180.0   
## jun : 663 tue: 911 Mean : 264.7   
## nov : 514 wed:1007 3rd Qu.: 329.0   
## apr : 282 Max. :3253.0   
## (Other): 243   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : 0.000 Min. :0.000 failure : 502   
## 1st Qu.: 1.000 1st Qu.: 3.000 1st Qu.:0.000 nonexistent:4330   
## Median : 2.000 Median : 5.000 Median :0.000 success : 168   
## Mean : 2.598 Mean : 5.821 Mean :0.169   
## 3rd Qu.: 3.000 3rd Qu.: 6.000 3rd Qu.:0.000   
## Max. :40.000 Max. :20.000 Max. :5.000   
## NA's :4816   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m   
## Min. :-3.4000 Min. :92.20 Min. :-50.80 Min. :0.634   
## 1st Qu.:-1.8000 1st Qu.:93.08 1st Qu.:-42.70 1st Qu.:1.344   
## Median : 1.1000 Median :93.92 Median :-41.80 Median :4.857   
## Mean : 0.1184 Mean :93.59 Mean :-40.45 Mean :3.661   
## 3rd Qu.: 1.4000 3rd Qu.:93.99 3rd Qu.:-36.40 3rd Qu.:4.961   
## Max. : 1.4000 Max. :94.77 Max. :-26.90 Max. :5.045   
##   
## nr.employed y   
## Min. :4964 no :4394   
## 1st Qu.:5099 yes: 606   
## Median :5191   
## Mean :5168   
## 3rd Qu.:5228   
## Max. :5228   
##

**save.image**("DadesBank\_5000.RData")

# Inicialització dels vectors de missings, errors i outliers

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

num\_total\_missings<-**rep**(0,21)  
num\_total\_errors<-**rep**(0,21)  
num\_total\_outliers<-**rep**(0,21)

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

df**$**missings\_indiv <- 0  
df**$**errors\_indiv <- 0  
df**$**outliers\_indiv <- 0

# Univariate Descriptive Analysis & Data Quality Report

## Qualitative Variables (Factors) / Categorical

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

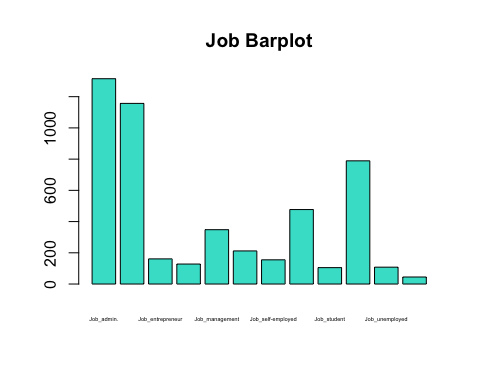
### 2. Job

### Type of job?

df**$**job<-**factor**(df**$**job)  
**levels**(df**$**job)<-**paste**("Job\_",sep="",**levels**(df**$**job))  
**summary**(df**$**job)

## Job\_admin. Job\_blue-collar Job\_entrepreneur Job\_housemaid   
## 1315 1157 161 128   
## Job\_management Job\_retired Job\_self-employed Job\_services   
## 348 212 155 477   
## Job\_student Job\_technician Job\_unemployed Job\_unknown   
## 105 789 108 45

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



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

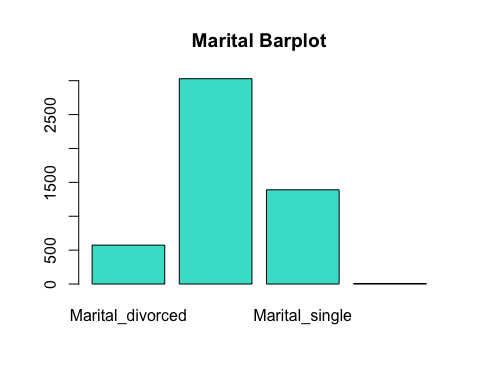
### 3. Marital

### Marital status?

df**$**marital<-**factor**(df**$**marital)  
**levels**(df**$**marital)<-**paste**("Marital\_",sep="",**levels**(df**$**marital))  
**summary**(df**$**marital)

## Marital\_divorced Marital\_married Marital\_single Marital\_unknown   
## 574 3029 1390 7

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



sel<-**which**(df**$**marital**==**"Marital\_unknown");**length**(sel)

## [1] 7

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

## Marital\_divorced Marital\_married Marital\_single Marital\_unknown   
## 574 3029 1390 0   
## NA's   
## 7

*#Podem veure que de la nostra mostra no tenim cap factor incorrecte i com en la nostra mostra la variable "marital\_unkown" es molt petita s'han de posar com a NA*

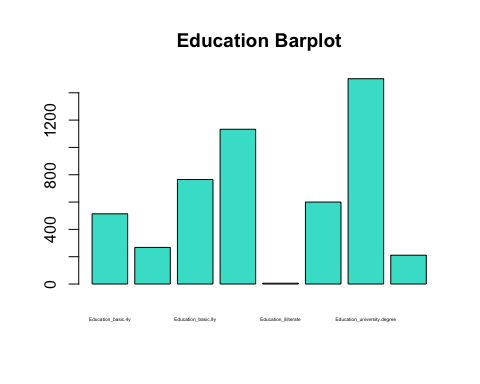
### 4. Education

### Type of education?

df**$**education<-**factor**(df**$**education)  
**levels**(df**$**education)<-**paste**("Education\_",sep="",**levels**(df**$**education))  
**summary**(df**$**education)

## Education\_basic.4y Education\_basic.6y   
## 514 268   
## Education\_basic.9y Education\_high.school   
## 765 1133   
## Education\_illiterate Education\_professional.course   
## 6 600   
## Education\_university.degree Education\_unknown   
## 1503 211

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



sel<-**which**(df**$**education**==**"Education\_unknown");**length**(sel)

## [1] 211

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

## Education\_basic.4y Education\_basic.6y   
## 514 268   
## Education\_basic.9y Education\_high.school   
## 765 1133   
## Education\_illiterate Education\_professional.course   
## 6 600   
## Education\_university.degree Education\_unknown   
## 1503 0   
## NA's   
## 211

*#Quan observem tots els factors ens podem adonar que no hi ha cap NA (data missing) ni cap factor no contemplat, llavors no tenim cap error*

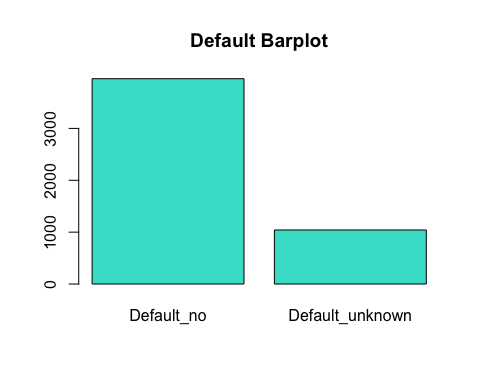
### 5. Default

### Has credit in default?

df**$**default<-**factor**(df**$**default)  
**levels**(df**$**default)<-**paste**("Default\_",sep="",**levels**(df**$**default))  
**summary**(df**$**default)

## Default\_no Default\_unknown   
## 3958 1042

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



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

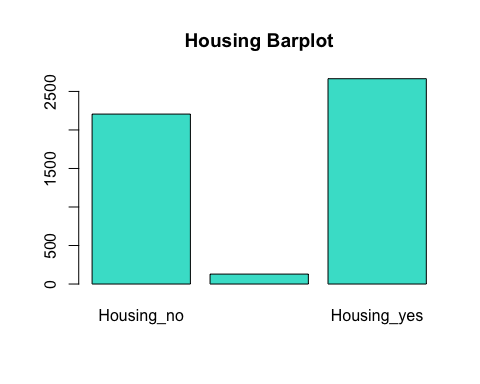
### 6. Housing

### Has housing loan?

df**$**housing<-**factor**(df**$**housing)  
**levels**(df**$**housing)<-**paste**("Housing\_",sep="",**levels**(df**$**housing))  
**summary**(df**$**housing)

## Housing\_no Housing\_unknown Housing\_yes   
## 2206 129 2665

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



*#Com podem veure anteriorment tampoc tenim cap data missing ni cap factor amb valors estranys, pero podem veure que el factor "Housing\_unknown" podria ser un possible outlier*

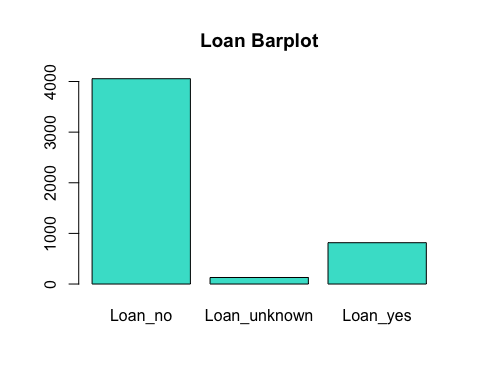
### 7. Loan

### Has personal loan?

df**$**loan<-**factor**(df**$**loan)  
**levels**(df**$**loan)<-**paste**("Loan\_",sep="",**levels**(df**$**loan))  
**summary**(df**$**loan)

## Loan\_no Loan\_unknown Loan\_yes   
## 4055 129 816

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



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

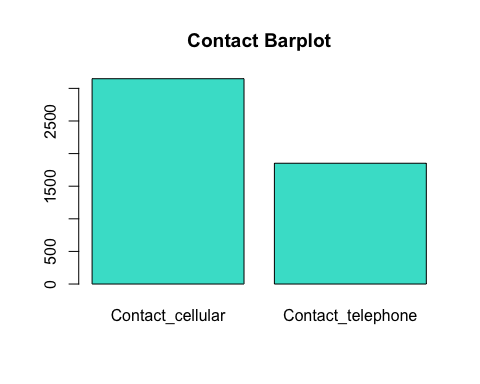
### 8. Contact

### Contact communication type?

df**$**contact<-**factor**(df**$**contact)  
**levels**(df**$**contact)<-**paste**("Contact\_",sep="",**levels**(df**$**contact))  
**summary**(df**$**contact)

## Contact\_cellular Contact\_telephone   
## 3148 1852

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



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

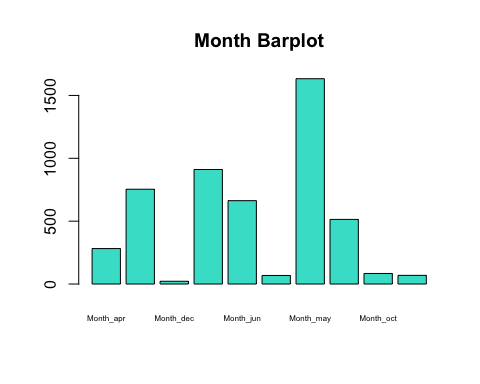
### 9. Month

### Last contact month of the year?

df**$**month<-**factor**(df**$**month)  
**levels**(df**$**month)<-**paste**("Month\_",sep="",**levels**(df**$**month))  
**summary**(df**$**month)

## Month\_apr Month\_aug Month\_dec Month\_jul Month\_jun Month\_mar Month\_may   
## 282 754 22 911 663 68 1633   
## Month\_nov Month\_oct Month\_sep   
## 514 84 69

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



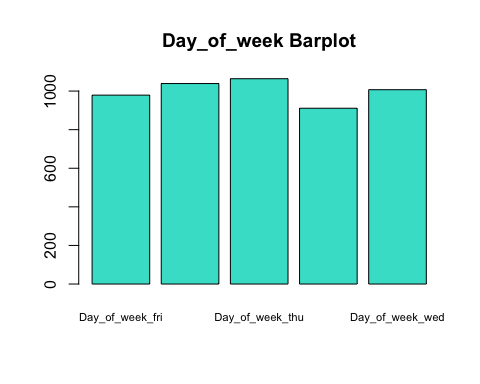
### 10. Day\_of\_week

### Last contact day of the week?

df**$**day\_of\_week<-**factor**(df**$**day\_of\_week)  
**levels**(df**$**day\_of\_week)<-**paste**("Day\_of\_week\_",sep="",**levels**(df**$**day\_of\_week))  
**summary**(df**$**day\_of\_week)

## Day\_of\_week\_fri Day\_of\_week\_mon Day\_of\_week\_thu Day\_of\_week\_tue   
## 979 1039 1064 911   
## Day\_of\_week\_wed   
## 1007

**barplot**(**summary**(df**$**day\_of\_week),main="Day\_of\_week Barplot",col = "turquoise",cex.names=0.7)



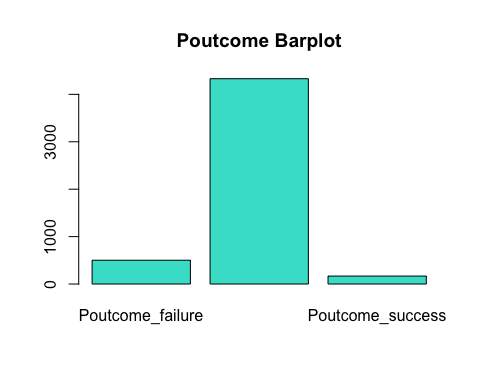
### 15. Poutcome

### Outcome of the previous marketing campaign?

df**$**poutcome<-**factor**(df**$**poutcome)  
**levels**(df**$**poutcome)<-**paste**("Poutcome\_",sep="",**levels**(df**$**poutcome))  
**summary**(df**$**poutcome)

## Poutcome\_failure Poutcome\_nonexistent Poutcome\_success   
## 502 4330 168

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



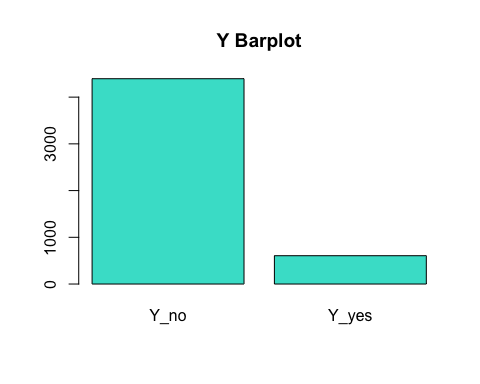
### 21. Y

### Has the client subscribed a term deposit?

df**$**y<-**factor**(df**$**y)  
**levels**(df**$**y)<-**paste**("Y\_",sep="",**levels**(df**$**y))  
**summary**(df**$**y)

## Y\_no Y\_yes   
## 4394 606

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



## Quantitative Variables (Numerical)

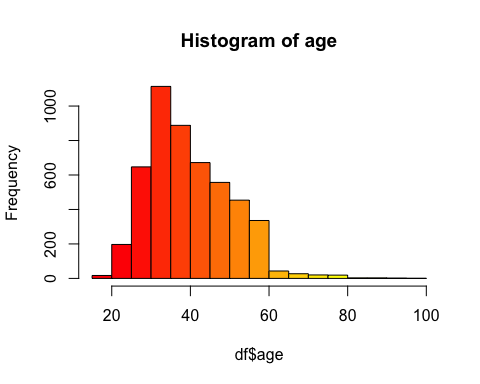
Hem de fer un analisi de totes les variables per poder identificar missings, errors i els outliers. Tambe farem una serie de boxplots i histogrames per analitzar i visualitzar millor les dades de la nostra mostra

### 1. Age

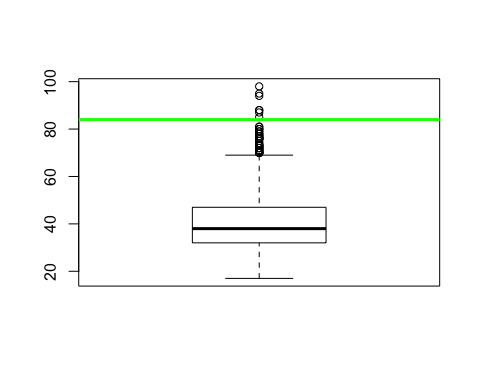
**summary**(df**$**age)

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

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



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



*#Amb la comanda abline el que volem fer es poder identificar de una manera mes facil els possibles outliers i poder tenir una millor visualitzacio, per aixo marco a l'altura dels 84 anys la nostra mostra, ja que aquests valors son els que s'allunyen una mica de la resta, llavors s'ahuran de fer una serie d'imputacions*  
  
sel <- **which**(df**$**age **>=** 84);**length**(sel);sel

## [1] 7

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

**summary**(df**$**age)

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

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

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

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

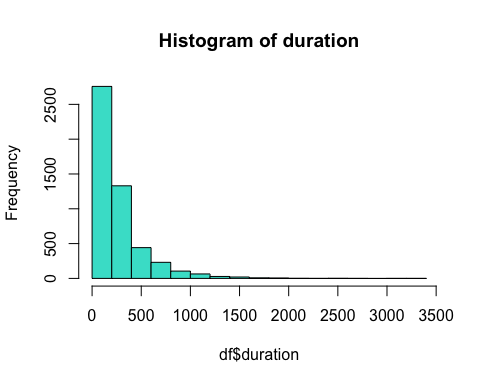
### 11. Duration

### Last contact duration?

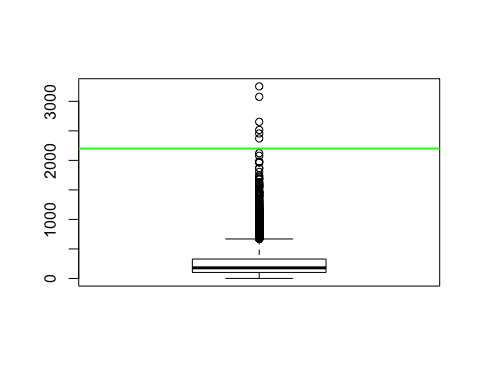
**summary**(df**$**duration)

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

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



*#A partir del summary executat podem observar que el temsp minim de la durada de una trucada es d'1 segon, i ja ens podem adonar que aquest valor no te molt sentit a l'hora de tractar-se una trucada no? No dona temps que el client escolti i penji i la durada maxima es de 3253 segons que son aproximadament uns 54 minuts i pot ser un valor real*  
  
**boxplot**(df**$**duration)  
**abline**(h=2200,col="green",lwd=2)



*#Per tal d'identificar possibles outliers utilitzem l'eina Boxplot, tinguent en compte el significat de la variable marquem amb una linia vermella el valor 2200, a partir del qual definim els possibles outliers ja que considerem que les observacions que prenen un valor a partir de 2200 es desvien significativament de la resta*  
  
sel <- **which**(df**$**duration **>=** 2200);**length**(sel);sel

## [1] 6

## [1] 1013 1140 2197 2919 2969 3440

num\_total\_outliers[11] <- **length**(sel)  
df[sel, "outliers\_indiv"] <- df[sel, "outliers\_indiv"] **+** 1  
df <- df[**-**sel,]   
**summary**(df**$**duration)

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

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

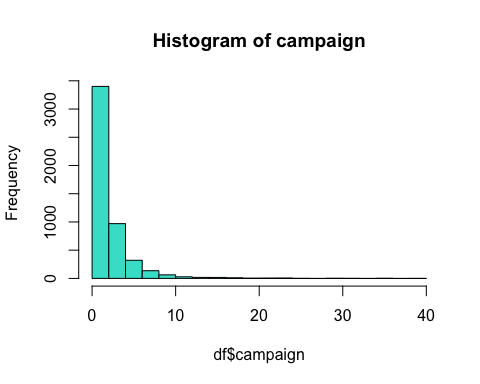
### 12. Campaign

### Number of contacts performed during this campaign?

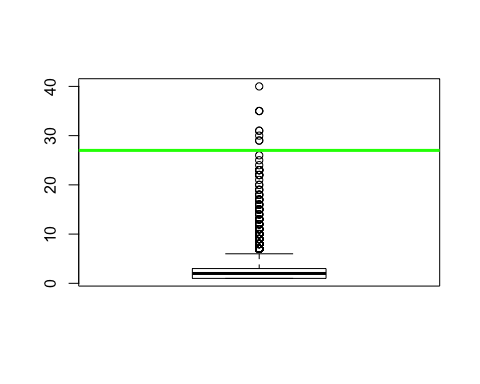
**summary**(df**$**campaign)

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

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



*#Quan fem el summary i el boxplot veiem que no hi ha cap mostra que contingui un NA (missing data) pero amb el boxplot si que veiem que hi han alguns valors que poden no ser molt realistes, ja que es una mica estrany que una campanya es contacti unes 40 vegades amb una mateixa persona, comptant que la mitjana són dues vegades, llavors eliminarem a partir d'unes 27 vegades/persona que es el que te mes sentit comu i es on veiem que disten de la resta*  
*#Aquestes dades de la mostra les considerem errors i les eliminarem de la mostra*  
  
**boxplot**(df**$**campaign)  
**abline**(h=27,col="green",lwd=3)

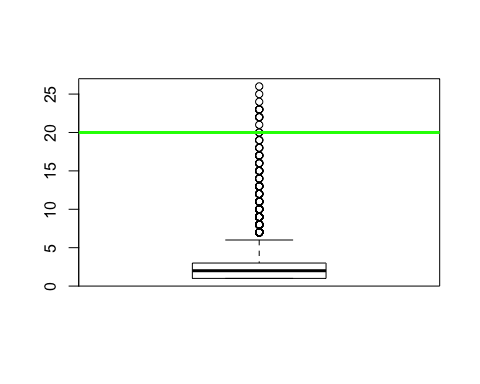


sel <- **which**(df**$**campaign **>** 27)  
**length**(sel);sel

## [1] 9

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

num\_total\_errors[12] <- **length**(sel)  
df[sel, "campaign"] <- NA  
df[sel, "errors\_indiv"] <- df[sel, "errors\_indiv"] **+** 1  
  
**boxplot**(df**$**campaign)  
**abline**(h=20,col="green",lwd=3)



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

## [1] 48

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

num\_total\_outliers[12] <- **length**(sel)  
df[sel, "campaign"] <- NA  
df[sel, "outliers\_indiv"] <- df[sel, "outliers\_indiv"] **+** 1  
df<-df[**-**sel,]  
**summary**(df**$**campaign)

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

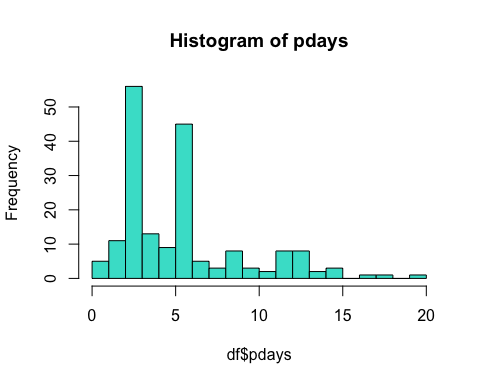
### 13. Pdays

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

**summary**(df**$**pdays)

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

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



*#Si analitzem aquesta variable veiem que tenir valor 0 significa que no ha passat cap dia des de que s'ha finalitzat la campanya anterior i s'ha contactat amb l'individu per aquesta campanya la qual cosa considerem que es tracta de un error per aixo procedim a identificar i comptabilitzar l'esmentat error a continuacio.*   
  
sel <- **which**(df**$**pdays **==** 0)  
**length**(sel);sel

## [1] 2

## [1] 4844 4847

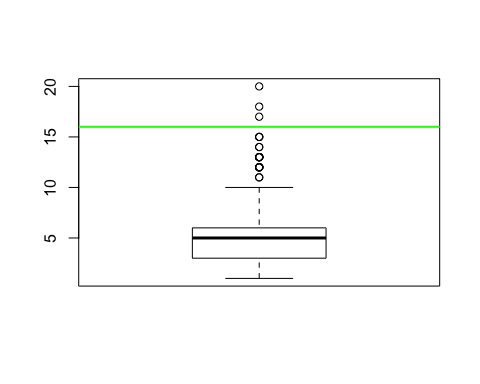
*#A partir del summary veiem que hi han 2 observacions que tenen valor 0.*  
num\_total\_errors[13] = **length**(sel)  
df[sel, "pdays"] <- NA  
df[sel, "errors\_indiv"] <- df[sel, "errors\_indiv"] **+** 1  
  
*#Tambe podem observem que aquesta variable te un nombre molt elevat de NA's(missing data) aquestes situacions signifiquen que no s'ha contactat amb l'individu previament en cap altre campanya per aixo no pot existir cap valor amb els dies des de la ultima vegada que es va contactar.*  
  
sel <- **which**(**is.na**(df**$**pdays))  
**length**(sel);*#sel*

## [1] 4764

num\_total\_missings[13] = **length**(sel)  
df[sel, "missings\_indiv"] <- df[sel, "missings\_indiv"] **+** 1  
**summary**(df**$**pdays)

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

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



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

## [1] 3

## [1] 4846 4870 4912

num\_total\_outliers[13] = **length**(sel)  
df[sel, "pdays"] <- NA  
df[sel, "outliers\_indiv"] <- df[sel, "outliers\_indiv"] **+** 1  
**summary**(df**$**pdays)

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

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

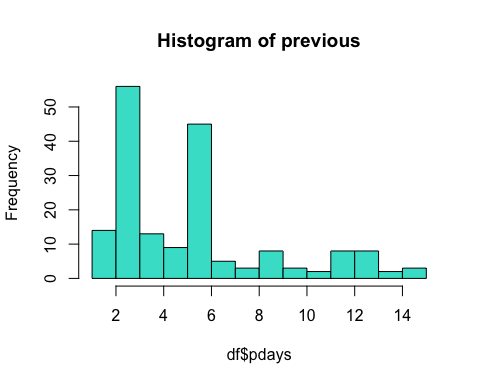
### 14. Previous

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

**summary**(df**$**previous)

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

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



*#A partir del summary efectuat sobre la variable "Previous" podem veure que no tenim cap NA i podriem considerar que tampoc error perque ja que el nombre minim de ocntactes previs a la campanya actual amb l'individu es 0 i el maxim trobat es 5, que poden ser valors reals*  
  
*#Quan observem el boxplot i el summary veiem que la majoria de les nostres observacions son 0 i llavors no podem tenir o identificar rapidament els possibles outliers*

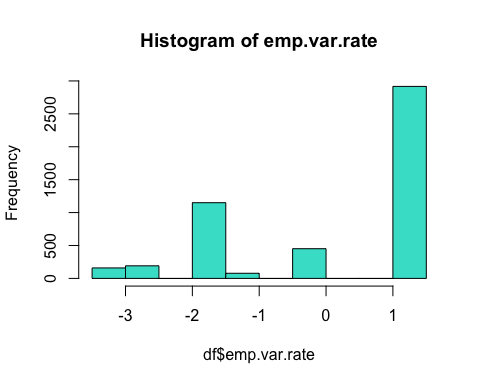
### 16. Emp.var.rate

### Employment variation rate?

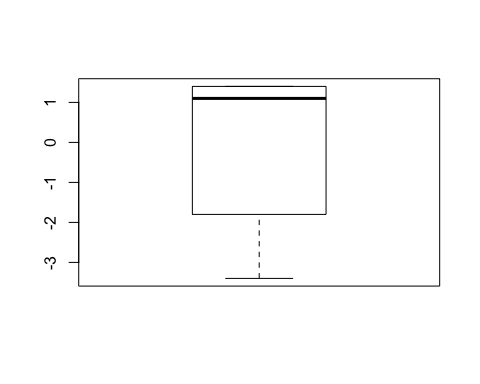
**summary**(df**$**emp.var.rate)

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

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



**boxplot**(df**$**emp.var.rate)



*#A partir del summary, l'histograma i el boxplot podem afirmar que no tenim cap missing ni error ni outlier, perque tots els valors agafats son realistes*

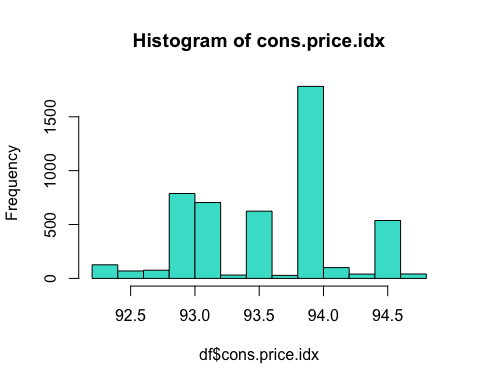
### 17. Cons.price.idx

### Consumer price index - monthly indicator?

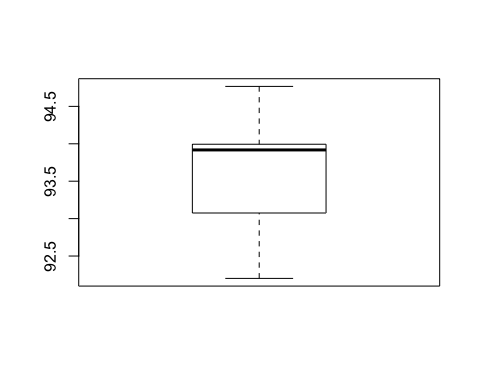
**summary**(df**$**cons.price.idx)

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

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



**boxplot**(df**$**cons.price.idx)



*#A partir del summary, l'histograma i el boxplot podem afirmar que no tenim cap missing ni error ni outlier, perque tots els valors agafats son realistes*

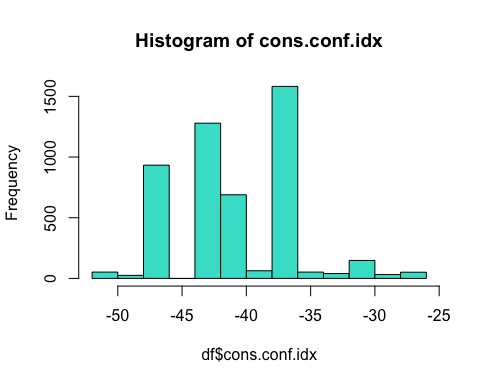
### 18. Cons.conf.idx

### Consumer confidence index - monthly indicator?

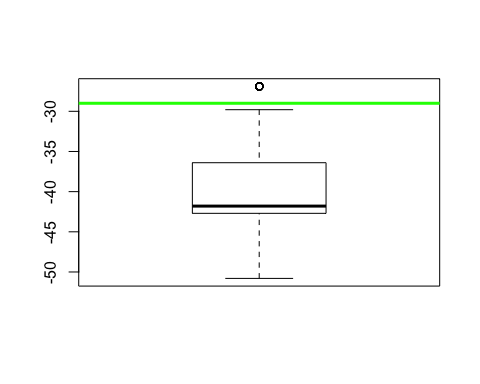
**summary**(df**$**cons.conf.idx)

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

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



**boxplot**(df**$**cons.conf.idx)  
*#Com podem veure després del boxplot hi han algunes observacions que podrien considerarse possibles outliers, llavors marquem -29 amb el abline*  
  
**abline**(h=**-**29,col="green",lwd=3)



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

## [1] 51

## [1] 4561 4562 4563 4564 4565 4566 4567 4568 4569 4570 4571 4572 4573 4574  
## [15] 4575 4576 4577 4578 4579 4580 4581 4582 4583 4584 4585 4586 4587 4588  
## [29] 4589 4590 4591 4592 4593 4594 4595 4596 4597 4598 4599 4600 4601 4602  
## [43] 4603 4604 4605 4606 4607 4608 4609 4610 4611

num\_total\_outliers[18] = **length**(sel)  
df[sel, "cons.conf.idx"] <- NA  
df[sel, "outliers\_indiv"] <- df[sel, "outliers\_indiv"] **+** 1  
**summary**(df**$**cons.conf.idx)

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

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

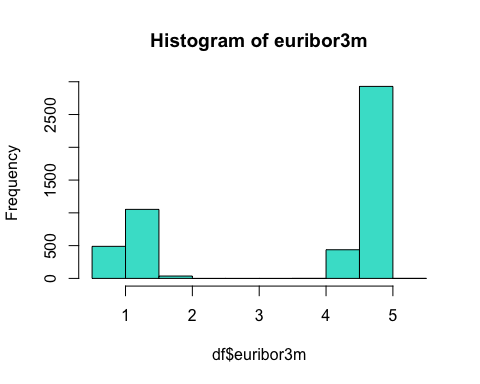
### 19. Euribor3m

### Euribor 3 month rate - daily indicator?

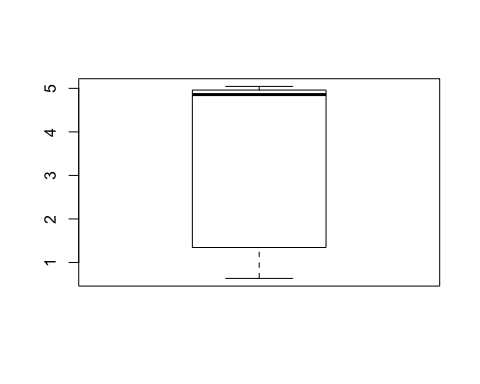
**summary**(df**$**euribor3m)

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

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



**boxplot**(df**$**euribor3m)



*#A partir del boxplot efectuat podem veure que els valors obtinguts son majoritariament menors que 5 i com s'observa la mitjana es troba molt a prop del maxim obtingut*

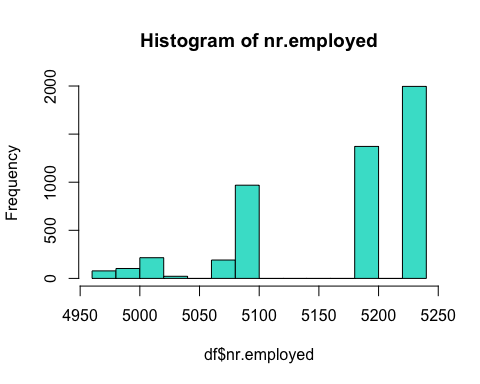
### 20. Nr.employed

### Number of employees - quarterly indicator?

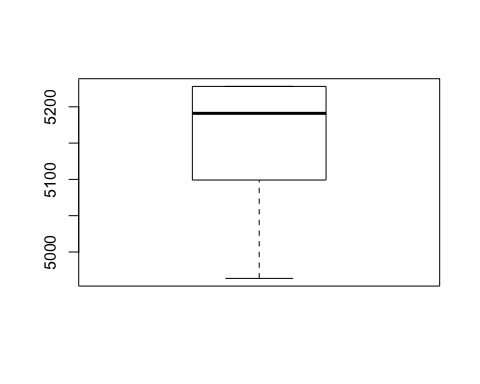
**summary**(df**$**nr.employed)

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

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



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



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

### CONTAR NA’s

*#Hem de contar el numero de NA's despres d'analitzar les dades i marcta els outliers, missings i errors*  
miss\_row <- **rowSums**(**is.na**(df))  
miss\_col <- **colSums**(**is.na**(df))  
miss\_col

## age job marital education default   
## 7 0 7 210 0   
## housing loan contact month day\_of\_week   
## 0 0 0 0 0   
## duration campaign pdays previous poutcome   
## 0 9 4767 0 0   
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed   
## 0 0 51 0 0   
## y missings\_indiv errors\_indiv outliers\_indiv   
## 0 0 0 0

*#Podem veure el numero de NA que tenim per cada variable*  
**summary**(miss\_row)

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

### Rank of variables

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

### Per individuals:

*#errors (la majoria de registres no tenen errors i els que tenen errors com a maxim nomes en tenen 1 )*  
**summary**(df**$**errors\_indiv)

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

*#outliers (el registres amb outliers com a maxim tenen 2 variables amb outlier)*  
**summary**(df**$**outliers\_indiv)

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

*#missings abans d'introduir manualment NA's per cada registre, nomes la variable pdays tenia missings des de un principi*  
**summary**(df**$**missings\_indiv)

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

*#despres de depurar les dades i introduir els NA`s*  
*#miss\_col<-colSums(is.na(df))*  
NAs\_indiv <- **rowSums**(**is.na**(df))  
**summary**(df**$**NAs\_indiv)

## Length Class Mode   
## 0 NULL NULL

### Per variable:

*#Després de calcular tots el missings, outliers i errors fem el resum d'ells*  
  
*#num total missings*  
data <- **t**(**c**(num\_total\_missings[13]))  
data

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

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



*#num total errors*  
data <- **t**(**c**(num\_total\_errors[12**:**13]))  
data

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

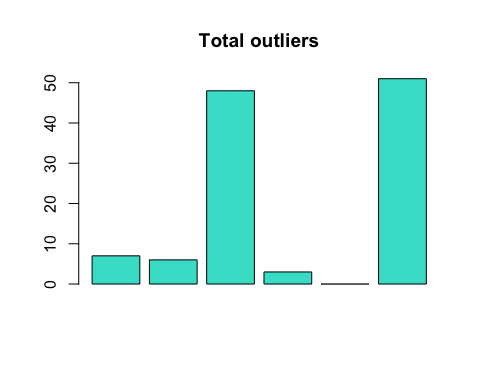
**barplot**(data, main="Total errors", col=("turquoise"))



*#num total outliers*  
data <- **t**(**c**(num\_total\_outliers[1],num\_total\_outliers[11**:**14],num\_total\_outliers[18]))  
data

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

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



### Imputation

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

**library**(missMDA)  
  
*# Numeric imputation*  
vars\_con<-**names**(df)[**c**(1,11**:**14,16**:**20)]  
vars\_dis<-**names**(df)[**c**(2**:**10,15,21)] *#solo 21*  
**summary**(df[,vars\_con])

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

**summary**(df[,vars\_dis])

## job marital   
## Job\_admin. :1301 Marital\_divorced: 562   
## Job\_blue-collar:1144 Marital\_married :3000   
## Job\_technician : 784 Marital\_single :1377   
## Job\_services : 473 Marital\_unknown : 0   
## Job\_management : 345 NA's : 7   
## Job\_retired : 206   
## (Other) : 693   
## education default   
## Education\_university.degree :1486 Default\_no :3914   
## Education\_high.school :1120 Default\_unknown:1032   
## Education\_basic.9y : 759   
## Education\_professional.course: 595   
## Education\_basic.4y : 502   
## (Other) : 274   
## NA's : 210   
## housing loan contact   
## Housing\_no :2179 Loan\_no :4020 Contact\_cellular :3128   
## Housing\_unknown: 126 Loan\_unknown: 126 Contact\_telephone:1818   
## Housing\_yes :2641 Loan\_yes : 800   
##   
##   
##   
##   
## month day\_of\_week poutcome   
## Month\_may:1620 Day\_of\_week\_fri: 967 Poutcome\_failure : 502   
## Month\_jul: 893 Day\_of\_week\_mon:1029 Poutcome\_nonexistent:4276   
## Month\_aug: 749 Day\_of\_week\_thu:1049 Poutcome\_success : 168   
## Month\_jun: 648 Day\_of\_week\_tue: 903   
## Month\_nov: 514 Day\_of\_week\_wed: 998   
## Month\_apr: 281   
## (Other) : 241   
## y   
## Y\_no :4349   
## Y\_yes: 597   
##   
##   
##   
##   
##

*#aq.plot(df[,vars\_con],delta=qchisq(0.995,df=ncol(x)))*  
  
res.impn<-**imputePCA**(df[,vars\_con],ncp=5) *#vars\_con=numericas*   
*#res.impn<-imputePCA(df[,vars\_dis],ncp=5)*  
**attributes**(res.impn)

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

*#data.frame with all NA imputed: res.impn$completeObs*  
*#summary(res.impn$completeObs)*  
  
df[,"age"] <- res.impn**$**completeObs[,"age"]  
df[,"campaign"] <- res.impn**$**completeObs[,"campaign"]  
*#df[,"pdays"] <- res.impn$completeObs[,"pdays"]*  
df[,"cons.conf.idx"] <- res.impn**$**completeObs[,"cons.conf.idx"]  
df[,"euribor3m"] <- res.impn**$**completeObs[,"euribor3m"]  
miss\_row <- **rowSums**(**is.na**(df))  
miss\_col <- **colSums**(**is.na**(df))

**summary**(df**$**month)

## Month\_apr Month\_aug Month\_dec Month\_jul Month\_jun Month\_mar Month\_may   
## 281 749 22 893 648 67 1620   
## Month\_nov Month\_oct Month\_sep   
## 514 83 69

**table** (df**$**month)

##   
## Month\_apr Month\_aug Month\_dec Month\_jul Month\_jun Month\_mar Month\_may   
## 281 749 22 893 648 67 1620   
## Month\_nov Month\_oct Month\_sep   
## 514 83 69

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

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3 3 3 3 3 3

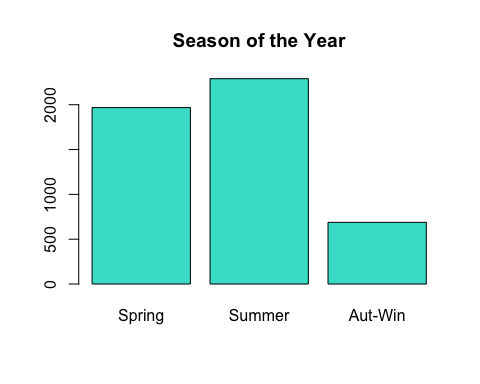
*# 1 level - spring*   
sel<-**which**(df**$**month **%in%** **c**("Month\_mar","Month\_apr","Month\_may"))  
df**$**season[sel] <-1  
  
*# 2 level - Summer*   
sel<-**which**(df**$**month **%in%** **c**("Month\_jun","Month\_jul","Month\_aug"))  
df**$**season[sel] <-2  
  
**table**(df**$**season)

##   
## 1 2 3   
## 1968 2290 688

**summary**(df**$**season)

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

df**$**season<-**factor**(df**$**season,levels=1**:**3,labels=**c**("Spring","Summer","Aut-Win"))  
  
**barplot**(**summary**(df**$**season), main="Season of the Year", col=("turquoise"))



*#IMPUTATION Pdays (Manual)*  
  
**table**(df**$**pdays)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15   
## 3 11 56 13 9 45 5 3 8 3 2 8 8 2 3

**summary**(df**$**pdays)

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

sel <- **which**(**is.na**(df**$**pdays))  
sel

**length**(sel)

## [1] 4767

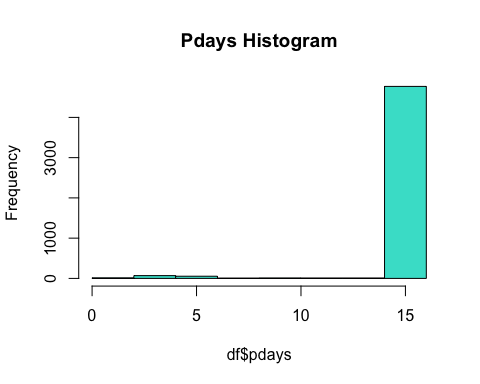
df[sel, "pdays"] <- 16  
**table**(df**$**pdays)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15   
## 3 11 56 13 9 45 5 3 8 3 2 8 8 2 3   
## 16   
## 4767

**summary**(df**$**pdays)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 16.00 16.00 15.63 16.00 16.00

**hist**(df**$**pdays, 10, main = "Pdays Histogram", col = "turquoise")



### Discretitzation

Ara el que farem será la discretització de les variables numeriques i aixo ho farem convertint en factors els diferents rangs que tenim de les observacions corresponents a una variable numerica per tenir una visualitzacio mes clara

vars\_con<-**names**(df)[**c**(1,11**:**14,16**:**20)];  
vars\_con

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

**summary**(df[,vars\_con])

## age duration campaign pdays   
## Min. :17.00 Min. : 1.0 Min. : 1.000 Min. : 1.00   
## 1st Qu.:32.00 1st Qu.: 104.0 1st Qu.: 1.000 1st Qu.:16.00   
## Median :38.00 Median : 182.0 Median : 2.000 Median :16.00   
## Mean :40.05 Mean : 262.8 Mean : 2.389 Mean :15.63   
## 3rd Qu.:47.00 3rd Qu.: 329.0 3rd Qu.: 3.000 3rd Qu.:16.00   
## Max. :81.00 Max. :2122.0 Max. :14.000 Max. :16.00   
## previous emp.var.rate cons.price.idx cons.conf.idx   
## Min. :0.0000 Min. :-3.4000 Min. :92.20 Min. :-50.80   
## 1st Qu.:0.0000 1st Qu.:-1.8000 1st Qu.:93.08 1st Qu.:-42.70   
## Median :0.0000 Median : 1.1000 Median :93.92 Median :-41.80   
## Mean :0.1708 Mean : 0.1074 Mean :93.59 Mean :-40.62   
## 3rd Qu.:0.0000 3rd Qu.: 1.4000 3rd Qu.:93.99 3rd Qu.:-36.40   
## Max. :5.0000 Max. : 1.4000 Max. :94.77 Max. :-29.80   
## euribor3m nr.employed   
## Min. :0.634 Min. :4964   
## 1st Qu.:1.344 1st Qu.:5099   
## Median :4.857 Median :5191   
## Mean :3.649 Mean :5168   
## 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :5.045 Max. :5228

### Factor Age

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

## 0% 25% 50% 75% 100%   
## 17 32 38 47 81

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

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

*#Es crea una variable auxiliar per tenir els diferents rangs d'edat i fem els intervals per a que sigui mes sencilla i facil la visualitzacio de les diferents mostres*  
df**$**varauxiliar<-**factor**(**cut**(df**$**age,include.lowest=T,breaks=**c**(17,31,36,41,49,81)))  
**summary**(df**$**varauxiliar)

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

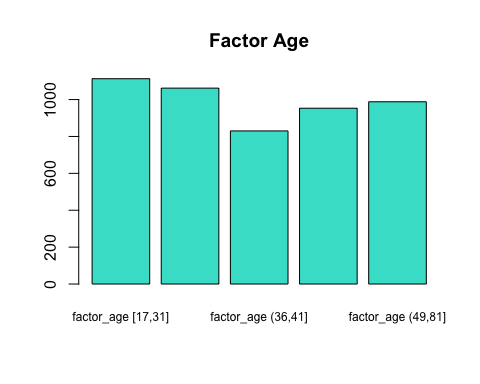
*#Fem la mitjana amb els valors de les edats i els nostres intervals*  
**tapply**(df**$**age,df**$**varauxiliar,median)

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

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

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

**barplot**(**summary**(df**$**factor\_age), main="Factor Age",col=("turquoise"),cex.names=0.75)



### Factor Duration

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

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

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

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

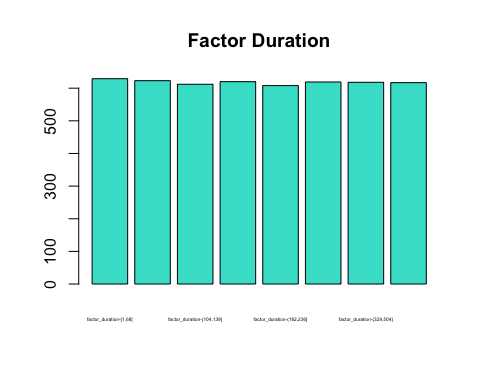
**tapply**(df**$**duration,df**$**factor\_duration,median)

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

**levels**(df**$**factor\_duration)<-**paste**("factor\_duration-",**levels**(df**$**factor\_duration),sep="")  
**table**(df**$**factor\_duration)

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

**barplot**(**summary**(df**$**factor\_duration), main="Factor Duration",col=("turquoise"),cex.names=0.3)



### Factor Campaign

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

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

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

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

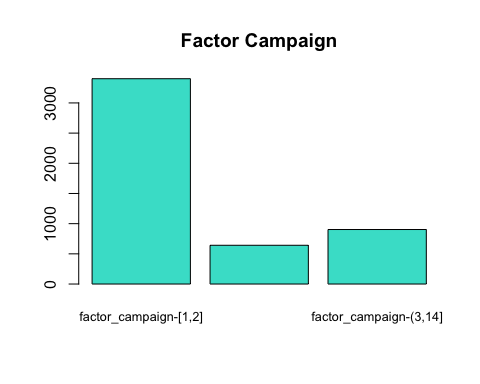
**tapply**(df**$**campaign,df**$**factor\_campaign,median)

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

**levels**(df**$**factor\_campaign)<-**paste**("factor\_campaign-",**levels**(df**$**factor\_campaign),sep="")  
**table**(df**$**factor\_campaign)

##   
## factor\_campaign-[1,2] factor\_campaign-(2,3] factor\_campaign-(3,14]   
## 3401 642 903

**barplot**(**summary**(df**$**factor\_campaign), main="Factor Campaign",col=("turquoise"),cex.names=0.8)



### Factor PDays

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

## 0% 25% 50% 75% 100%   
## 1 16 16 16 16

df**$**factor\_Pdays<-**factor**(**cut**(df**$**pdays,include.lowest=T,breaks=**c**(0,15,17)))  
  
**summary**(df**$**factor\_Pdays)

## [0,15] (15,17]   
## 179 4767

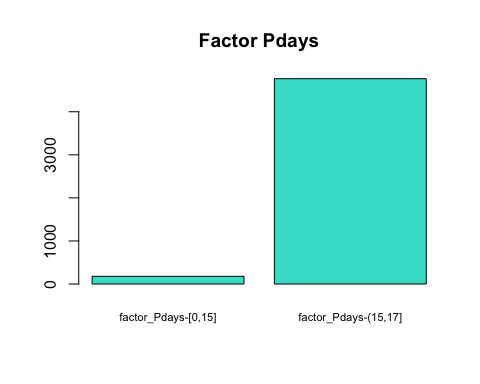
**tapply**(df**$**pdays,df**$**factor\_Pdays,median)

## [0,15] (15,17]   
## 5 16

**levels**(df**$**factor\_Pdays)<-**paste**("factor\_Pdays-",**levels**(df**$**factor\_Pdays),sep="")  
**table**(df**$**factor\_Pdays)

##   
## factor\_Pdays-[0,15] factor\_Pdays-(15,17]   
## 179 4767

**barplot**(**summary**(df**$**factor\_Pdays), main="Factor Pdays",col=("turquoise"),cex.names=0.7)



### Factor Previous

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

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

df**$**factor\_Previous<-**factor**(**cut**(df**$**previous,include.lowest=T,breaks=**c**(0,1,5)))  
  
**summary**(df**$**factor\_Previous)

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

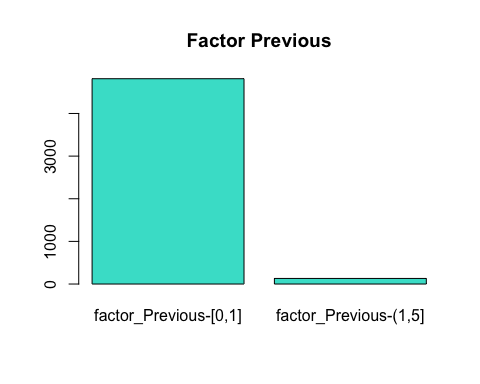
**tapply**(df**$**previous,df**$**factor\_Previous,median)

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

**levels**(df**$**factor\_Previous)<-**paste**("factor\_Previous-",**levels**(df**$**factor\_Previous),sep="")  
**table**(df**$**factor\_Previous)

##   
## factor\_Previous-[0,1] factor\_Previous-(1,5]   
## 4815 131

**barplot**(**summary**(df**$**factor\_Previous), main="Factor Previous",col=("turquoise"),cex.names=1.0)



*#Amb aquesta discretitzacio podem comprobar que el nombre de cops que s'ha contactat previament amb l'individu es majoritariament 0 o 1 i com a maxim una mitja de 5 cops.*

### Factor emp.var.rate

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

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

df**$**factor\_emp.var.rate<-**factor**(**cut**(df**$**emp.var.rate,include.lowest=T,breaks=**c**(**-**3.4,**-**1.8,**-**0.1,1.4)))  
  
**summary**(df**$**factor\_emp.var.rate)

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

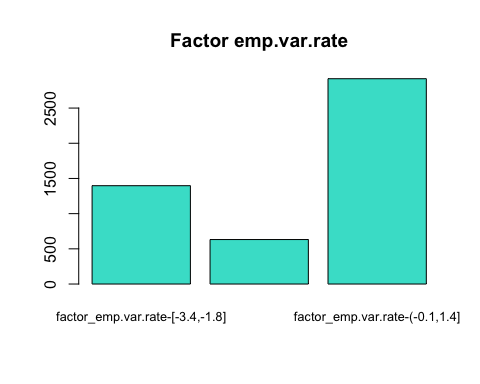
**tapply**(df**$**emp.var.rate,df**$**factor\_emp.var.rate,median)

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

**levels**(df**$**factor\_emp.var.rate)<-**paste**("factor\_emp.var.rate-",**levels**(df**$**factor\_emp.var.rate),sep="")  
**table**(df**$**factor\_emp.var.rate)

##   
## factor\_emp.var.rate-[-3.4,-1.8] factor\_emp.var.rate-(-1.8,-0.1]   
## 1397 632   
## factor\_emp.var.rate-(-0.1,1.4]   
## 2917

**barplot**(**summary**(df**$**factor\_emp.var.rate), main="Factor emp.var.rate",col=("turquoise"),cex.names=0.8)



### Factor cons.price.idx

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

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

df**$**factor\_cons.price.idx<-**factor**(**cut**(df**$**cons.price.idx,include.lowest=T,breaks=**c**(92.201,92.963,93.444,93.918,93.994,94.767)))  
  
**summary**(df**$**factor\_cons.price.idx)

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

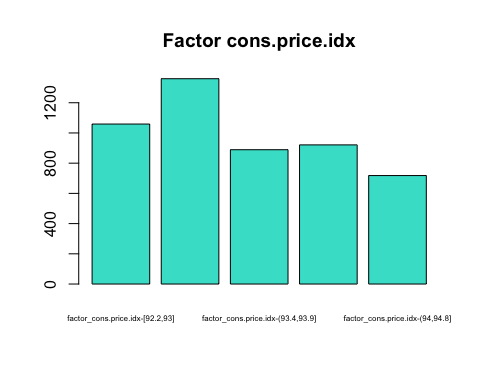
**tapply**(df**$**cons.price.idx,df**$**factor\_cons.price.idx,median)

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

**levels**(df**$**factor\_cons.price.idx)<-**paste**("factor\_cons.price.idx-",**levels**(df**$**factor\_cons.price.idx),sep="")  
**table**(df**$**factor\_cons.price.idx)

##   
## factor\_cons.price.idx-[92.2,93] factor\_cons.price.idx-(93,93.4]   
## 1059 1359   
## factor\_cons.price.idx-(93.4,93.9] factor\_cons.price.idx-(93.9,94]   
## 889 921   
## factor\_cons.price.idx-(94,94.8]   
## 718

**barplot**(**summary**(df**$**factor\_cons.price.idx), main="Factor cons.price.idx",col=("turquoise"),cex.names=0.5)



### Factor cons.conf.idx

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

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

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

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

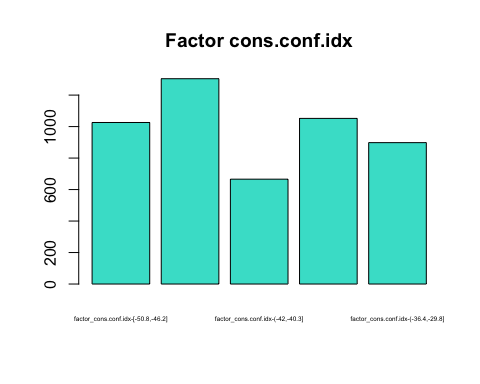
**tapply**(df**$**cons.conf.idx,df**$**factor\_cons.conf.idx,median)

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

**levels**(df**$**factor\_cons.conf.idx)<-**paste**("factor\_cons.conf.idx-",**levels**(df**$**factor\_cons.conf.idx),sep="")  
**table**(df**$**factor\_cons.conf.idx)

##   
## factor\_cons.conf.idx-[-50.8,-46.2] factor\_cons.conf.idx-(-46.2,-42]   
## 1026 1304   
## factor\_cons.conf.idx-(-42,-40.3] factor\_cons.conf.idx-(-40.3,-36.4]   
## 666 1052   
## factor\_cons.conf.idx-(-36.4,-29.8]   
## 898

**barplot**(**summary**(df**$**factor\_cons.conf.idx), main="Factor cons.conf.idx",col=("turquoise"),cex.names=0.4)



### Factor euribor3m

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

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

df**$**factor\_euribor3m<-**factor**(**cut**(df**$**euribor3m,include.lowest=T,breaks=**c**(0.634,1.266,1.415,4.856,4.864,4.961,4.964)))  
  
**summary**(df**$**factor\_euribor3m)

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

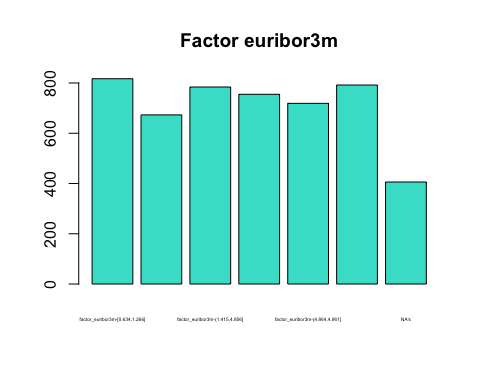
**tapply**(df**$**euribor3m,df**$**factor\_euribor3m,median)

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

**levels**(df**$**factor\_euribor3m)<-**paste**("factor\_euribor3m-",**levels**(df**$**factor\_euribor3m),sep="")  
**table**(df**$**factor\_euribor3m)

##   
## factor\_euribor3m-[0.634,1.266] factor\_euribor3m-(1.266,1.415]   
## 817 673   
## factor\_euribor3m-(1.415,4.856] factor\_euribor3m-(4.856,4.864]   
## 784 755   
## factor\_euribor3m-(4.864,4.961] factor\_euribor3m-(4.961,4.964]   
## 719 792

**barplot**(**summary**(df**$**factor\_euribor3m), main="Factor euribor3m",col=("turquoise"),cex.names=0.3)



### Factor nr.employed

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

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

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

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

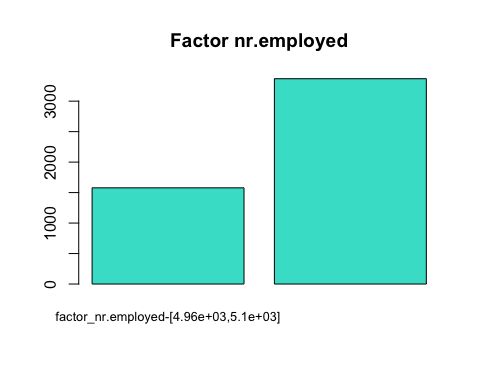
**tapply**(df**$**nr.employed,df**$**factor\_nr.employed,median)

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

**levels**(df**$**factor\_nr.employed)<-**paste**("factor\_nr.employed-",**levels**(df**$**factor\_nr.employed),sep="")  
**table**(df**$**factor\_nr.employed)

##   
## factor\_nr.employed-[4.96e+03,5.1e+03]   
## 1578   
## factor\_nr.employed-(5.1e+03,5.23e+03]   
## 3368

**barplot**(**summary**(df**$**factor\_nr.employed), main="Factor nr.employed",col=("turquoise"),cex.names=0.8)



### PROFILING

### Numeric target (Duration)

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

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

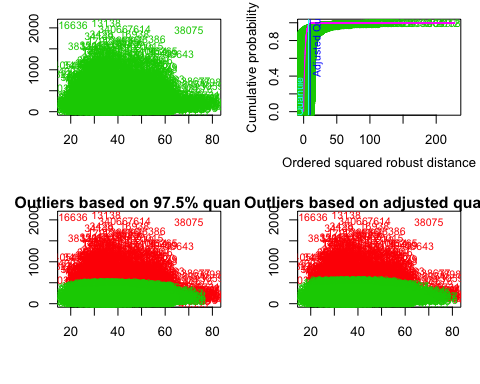
Per tal de observar la relacio del nostre target numeric amb les altres variables utilitzem la eina condes que ens proporciona informacio de les relacions entre les variables indicades i el target.

df**$**varauxiliar <- NULL *#borrem la variable auxiliar creada*  
df**$**aux <- NULL  
*#Despres de discretitzar les nostres variables tenim un total de 35 variables*  
*#names(df)*  
  
*#Description continuous by quantitative variables and/or by categorical variables*  
**library**(FactoMineR)  
  
**library**(mvoutlier)  
  
vars\_resu <-**names**(df)[**c**(1,11)]  
vars\_resu

## [1] "age" "duration"

**summary**(df[,vars\_con])

**aq.plot**(df[,vars\_resu])



## $outliers

*#vars\_res<-names(df)[c(11,21)]*  
vars<-**unique**(**c**(vars\_con,vars\_dis))  
*#vars*  
  
**condes**(df, **which**(**names**(df) **==** "duration"))

## $quanti  
## correlation p.value  
## previous 0.02859224 4.435374e-02  
## errors\_indiv -0.03476735 1.447588e-02  
## nr.employed -0.03619203 1.091224e-02  
## campaign -0.04179341 3.284450e-03  
## pdays -0.06147234 1.516945e-05  
## missings\_indiv -0.07328498 2.474678e-07  
##   
## $quali  
## R2 p.value  
## factor\_duration 0.8271873066 0.000000e+00  
## y 0.1863696068 9.891372e-224  
## factor\_Pdays 0.0051824450 4.017238e-07  
## poutcome 0.0041874670 3.132625e-05  
## month 0.0073478185 3.327154e-05  
## factor\_cons.price.idx 0.0039803615 5.696640e-04  
## factor\_Previous 0.0019228074 2.038492e-03  
## day\_of\_week 0.0029955473 5.075577e-03  
## factor\_cons.conf.idx 0.0026002247 1.194404e-02  
## contact 0.0011105265 1.909343e-02  
## default 0.0009897216 2.693284e-02  
## factor\_campaign 0.0013152237 3.866909e-02  
##   
## $category  
## Estimate p.value  
## factor\_duration-(504,2.12e+03] 547.162252 0.000000e+00  
## Y\_yes 169.675531 9.891372e-224  
## factor\_duration-(329,504] 138.462468 3.985182e-48  
## factor\_Pdays-[0,15] 49.355073 4.017238e-07  
## Poutcome\_success 62.641078 7.933875e-06  
## factor\_cons.price.idx-(93.4,93.9] 27.117765 2.010384e-04  
## Month\_jul 12.946601 2.986551e-04  
## factor\_Previous-(1,5] 34.966136 2.038492e-03  
## Contact\_cellular 8.850090 1.909343e-02  
## Default\_no 9.913335 2.693284e-02  
## Month\_dec 104.090396 2.868142e-02  
## Day\_of\_week\_tue 14.917687 4.872420e-02  
## Education\_illiterate 178.585152 4.932974e-02  
## Education\_university.degree -38.308971 3.857651e-02  
## factor\_cons.conf.idx-(-36.4,-29.8] -13.574401 3.768483e-02  
## factor\_cons.conf.idx-(-42,-40.3] -17.926886 2.695593e-02  
## Default\_unknown -9.913335 2.693284e-02  
## Contact\_telephone -8.850090 1.909343e-02  
## Month\_jun -37.404273 1.736971e-02  
## factor\_campaign-(3,14] -16.741883 1.148865e-02  
## Job\_technician -25.341033 1.106827e-02  
## Day\_of\_week\_mon -19.239047 7.577039e-03  
## Month\_aug -39.248662 5.073298e-03  
## factor\_cons.price.idx-(93,93.4] -19.809889 2.312144e-03  
## factor\_Previous-[0,1] -34.966136 2.038492e-03  
## factor\_Pdays-(15,17] -49.355073 4.017238e-07  
## factor\_duration-(182,236] -56.414720 8.764699e-09  
## factor\_duration-(139,182] -103.067426 8.297196e-27  
## factor\_duration-(104,139] -141.910732 3.245807e-49  
## factor\_duration-(68,104] -177.221056 2.195363e-78  
## factor\_duration-[1,68] -222.636796 8.250905e-127  
## Y\_no -169.675531 9.891372e-224

*#S'utilitza per fer totes les combinacions possibles de variables numeriques i factorials*  
*#Tindrem les variables que tenen un pvalor a partir d'un llindar del pvalor acceptat. No ens surten totes les variables estudiades, només les que tenen una mena de relació*  
*#Con el p valor muy bajo entonces rechazamos la hipotesi nula*  
  
*#$quanti: Com podem observar la variable pdays es la que te mes relacio amb la nostra variable target (duration), es a dir, quant mes gran sigui la duracio de la trucada tenim una correlacio mes gran amb aquesta i veiem que com a relacio inversament proporcional tenim campaign*  
*#$quali: La variable qualitativa que te mes realcio amb el nostre target es el seu mateix factor (factor\_duration) com es obvi, pero seguidament tenim el factor\_Pdays i la nostra variable y*  
*#$category: Podem observar que tenim una relacio dependent molt forta dels mesos i ultims contactes, podem veure que ha tingut exit i majoritariament la y es yes*

### Y (target qual)

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

df\_catdes<-df[**c**(1**:**21)]  
**catdes**(df\_catdes,21)

##   
## Link between the cluster variable and the categorical variables (chi-square test)  
## =================================================================================  
## p.value df  
## poutcome 2.884978e-155 2  
## month 2.020968e-82 9  
## contact 8.049707e-27 1  
## job 5.149262e-24 11  
## default 7.888260e-14 1  
## education 1.246599e-05 7  
## marital 4.868728e-03 3  
## day\_of\_week 3.137547e-02 4  
##   
## Description of each cluster by the categories  
## =============================================  
## $Y\_no  
## Cla/Mod Mod/Cla Global  
## poutcome=Poutcome\_nonexistent 91.01964 89.4918372 86.4537000  
## contact=Contact\_telephone 94.44444 39.4803403 36.7569753  
## default=Default\_unknown 94.67054 22.4649345 20.8653457  
## month=Month\_may 92.83951 34.5826627 32.7537404  
## job=Job\_blue-collar 92.74476 24.3964130 23.1298019  
## education=Education\_basic.9y 92.09486 16.0726604 15.3457339  
## month=Month\_jul 90.92945 18.6709588 18.0549939  
## education=Education\_basic.6y 93.28358 5.7484479 5.4185200  
## marital=Marital\_married 88.96667 61.3704300 60.6550748  
## job=Job\_services 91.54334 9.9563118 9.5632835  
## job=Job\_technician 90.17857 16.2566107 15.8511929  
## day\_of\_week=Day\_of\_week\_mon 89.79592 21.2462635 20.8046907  
## education=NA 83.33333 4.0239135 4.2458552  
## education=Education\_professional.course 85.21008 11.6578524 12.0299232  
## day\_of\_week=Day\_of\_week\_tue 85.16058 17.6822258 18.2571775  
## education=Education\_university.degree 85.93540 29.3630720 30.0444804  
## marital=Marital\_single 85.47567 27.0636928 27.8406793  
## poutcome=Poutcome\_failure 83.26693 9.6114049 10.1496159  
## job=Job\_admin. 85.16526 25.4771212 26.3040841  
## month=Month\_apr 78.29181 5.0586342 5.6813587  
## month=Month\_dec 45.45455 0.2299379 0.4448039  
## job=Job\_student 65.71429 1.5865716 2.1229276  
## job=Job\_retired 72.81553 3.4490688 4.1649818  
## month=Month\_mar 50.74627 0.7817889 1.3546300  
## month=Month\_sep 50.72464 0.8047827 1.3950667  
## default=Default\_no 86.15227 77.5350655 79.1346543  
## month=Month\_oct 48.19277 0.9197517 1.6781237  
## contact=Contact\_cellular 84.14322 60.5196597 63.2430247  
## poutcome=Poutcome\_success 23.21429 0.8967579 3.3966842  
## p.value v.test  
## poutcome=Poutcome\_nonexistent 3.543373e-50 14.895160  
## contact=Contact\_telephone 1.650430e-29 11.279842  
## default=Default\_unknown 6.847442e-16 8.073209  
## month=Month\_may 1.529311e-14 7.685055  
## job=Job\_blue-collar 2.309977e-09 5.974358  
## education=Education\_basic.9y 6.478104e-05 3.994682  
## month=Month\_jul 1.804548e-03 3.120646  
## education=Education\_basic.6y 3.345680e-03 2.934052  
## marital=Marital\_married 5.727878e-03 2.762966  
## job=Job\_services 8.657080e-03 2.625307  
## job=Job\_technician 3.216891e-02 2.142305  
## day\_of\_week=Day\_of\_week\_mon 3.661258e-02 2.090058  
## education=NA 4.459048e-02 -2.008497  
## education=Education\_professional.course 3.369438e-02 -2.123710  
## day\_of\_week=Day\_of\_week\_tue 5.704442e-03 -2.764304  
## education=Education\_university.degree 5.300406e-03 -2.788186  
## marital=Marital\_single 1.198449e-03 -3.239249  
## poutcome=Poutcome\_failure 1.167715e-03 -3.246651  
## job=Job\_admin. 4.654028e-04 -3.499917  
## month=Month\_apr 2.649823e-06 -4.696249  
## month=Month\_dec 1.944834e-06 -4.759074  
## job=Job\_student 2.045387e-09 -5.994161  
## job=Job\_retired 1.710143e-09 -6.023188  
## month=Month\_mar 6.474585e-14 -7.498107  
## month=Month\_sep 2.609525e-14 -7.616349  
## default=Default\_no 6.847442e-16 -8.073209  
## month=Month\_oct 6.812368e-19 -8.877918  
## contact=Contact\_cellular 1.650430e-29 -11.279842  
## poutcome=Poutcome\_success 2.944669e-88 -19.916208  
##   
## $Y\_yes  
## Cla/Mod Mod/Cla Global  
## poutcome=Poutcome\_success 76.785714 21.608040 3.3966842  
## contact=Contact\_cellular 15.856777 83.082077 63.2430247  
## month=Month\_oct 51.807229 7.202680 1.6781237  
## default=Default\_no 13.847726 90.787270 79.1346543  
## month=Month\_sep 49.275362 5.695142 1.3950667  
## month=Month\_mar 49.253731 5.527638 1.3546300  
## job=Job\_retired 27.184466 9.380235 4.1649818  
## job=Job\_student 34.285714 6.030151 2.1229276  
## month=Month\_dec 54.545455 2.010050 0.4448039  
## month=Month\_apr 21.708185 10.217755 5.6813587  
## job=Job\_admin. 14.834743 32.328308 26.3040841  
## poutcome=Poutcome\_failure 16.733068 14.070352 10.1496159  
## marital=Marital\_single 14.524328 33.500838 27.8406793  
## education=Education\_university.degree 14.064603 35.008375 30.0444804  
## day\_of\_week=Day\_of\_week\_tue 14.839424 22.445561 18.2571775  
## education=Education\_professional.course 14.789916 14.740369 12.0299232  
## education=NA 16.666667 5.862647 4.2458552  
## day\_of\_week=Day\_of\_week\_mon 10.204082 17.587940 20.8046907  
## job=Job\_technician 9.821429 12.897822 15.8511929  
## job=Job\_services 8.456660 6.700168 9.5632835  
## marital=Marital\_married 11.033333 55.443886 60.6550748  
## education=Education\_basic.6y 6.716418 3.015075 5.4185200  
## month=Month\_jul 9.070549 13.567839 18.0549939  
## education=Education\_basic.9y 7.905138 10.050251 15.3457339  
## job=Job\_blue-collar 7.255245 13.902848 23.1298019  
## month=Month\_may 7.160494 19.430486 32.7537404  
## default=Default\_unknown 5.329457 9.212730 20.8653457  
## contact=Contact\_telephone 5.555556 16.917923 36.7569753  
## poutcome=Poutcome\_nonexistent 8.980355 64.321608 86.4537000  
## p.value v.test  
## poutcome=Poutcome\_success 2.944669e-88 19.916208  
## contact=Contact\_cellular 1.650430e-29 11.279842  
## month=Month\_oct 6.812368e-19 8.877918  
## default=Default\_no 6.847442e-16 8.073209  
## month=Month\_sep 2.609525e-14 7.616349  
## month=Month\_mar 6.474585e-14 7.498107  
## job=Job\_retired 1.710143e-09 6.023188  
## job=Job\_student 2.045387e-09 5.994161  
## month=Month\_dec 1.944834e-06 4.759074  
## month=Month\_apr 2.649823e-06 4.696249  
## job=Job\_admin. 4.654028e-04 3.499917  
## poutcome=Poutcome\_failure 1.167715e-03 3.246651  
## marital=Marital\_single 1.198449e-03 3.239249  
## education=Education\_university.degree 5.300406e-03 2.788186  
## day\_of\_week=Day\_of\_week\_tue 5.704442e-03 2.764304  
## education=Education\_professional.course 3.369438e-02 2.123710  
## education=NA 4.459048e-02 2.008497  
## day\_of\_week=Day\_of\_week\_mon 3.661258e-02 -2.090058  
## job=Job\_technician 3.216891e-02 -2.142305  
## job=Job\_services 8.657080e-03 -2.625307  
## marital=Marital\_married 5.727878e-03 -2.762966  
## education=Education\_basic.6y 3.345680e-03 -2.934052  
## month=Month\_jul 1.804548e-03 -3.120646  
## education=Education\_basic.9y 6.478104e-05 -3.994682  
## job=Job\_blue-collar 2.309977e-09 -5.974358  
## month=Month\_may 1.529311e-14 -7.685055  
## default=Default\_unknown 6.847442e-16 -8.073209  
## contact=Contact\_telephone 1.650430e-29 -11.279842  
## poutcome=Poutcome\_nonexistent 3.543373e-50 -14.895160  
##   
##   
## Link between the cluster variable and the quantitative variables  
## ================================================================  
## Eta2 P-value  
## duration 0.186369607 9.891372e-224  
## nr.employed 0.139052649 5.557605e-163  
## pdays 0.124416618 7.349696e-145  
## euribor3m 0.104758799 5.493737e-121  
## emp.var.rate 0.099078243 3.487741e-114  
## previous 0.070648755 9.329422e-81  
## cons.price.idx 0.019937283 1.907193e-23  
## campaign 0.005057924 5.536389e-07  
##   
## Description of each cluster by quantitative variables  
## =====================================================  
## $Y\_no  
## v.test Mean in category Overall mean sd in category  
## nr.employed 26.222421 5177.8744999 5167.8073595 64.2441089  
## pdays 24.804035 15.8919292 15.6263647 1.1098761  
## euribor3m 22.760322 3.8560536 3.6487535 1.6188731  
## emp.var.rate 22.134632 0.2901587 0.1073999 1.4661991  
## cons.price.idx 9.929243 93.6160205 93.5857345 0.5562445  
## campaign 5.001143 2.4413845 2.3891187 2.0381577  
## previous -18.691123 0.1230168 0.1708451 0.3957657  
## duration -30.357828 221.8063923 262.7672867 200.3541053  
## Overall sd p.value  
## nr.employed 72.8658491 1.475237e-151  
## pdays 2.0320681 8.109757e-136  
## euribor3m 1.7286683 1.134100e-114  
## emp.var.rate 1.5670994 1.467071e-108  
## cons.price.idx 0.5789159 3.106051e-23  
## campaign 1.9835304 5.699132e-07  
## previous 0.4856692 5.846876e-78  
## duration 256.0881160 1.980616e-202  
##   
## $Y\_yes  
## v.test Mean in category Overall mean sd in category  
## duration 30.357828 561.157454 262.7672867 386.8354045  
## previous 18.691123 0.519263 0.1708451 0.8216383  
## campaign -5.001143 2.008375 2.3891187 1.4727896  
## cons.price.idx -9.929243 93.365109 93.5857345 0.6835676  
## emp.var.rate -22.134632 -1.223953 0.1073999 1.6338789  
## euribor3m -22.760322 2.138623 3.6487535 1.7527742  
## pdays -24.804035 13.691792 15.6263647 4.5804350  
## nr.employed -26.222421 5094.470687 5167.8073595 88.3423897  
## Overall sd p.value  
## duration 256.0881160 1.980616e-202  
## previous 0.4856692 5.846876e-78  
## campaign 1.9835304 5.699132e-07  
## cons.price.idx 0.5789159 3.106051e-23  
## emp.var.rate 1.5670994 1.467071e-108  
## euribor3m 1.7286683 1.134100e-114  
## pdays 2.0320681 8.109757e-136  
## nr.employed 72.8658491 1.475237e-151

**save.image**("DadesBank1\_5000.RData")

————– DELIVERABLE 2 —————-

# Principal Component Analysis (PCA)

L’analisi de components principals (a partir d’ara PCA) es una tecnica utilitzada per reduir la dimensionalitat d’un conjunt de dades per a poder-les representar graficament en grafics de dues o tres dimensions agrupant diverses variables de les dades en factors, o components, compostos per l’agrupacio de diverses variables.

Intuïtivament, la tècnica serveix per determinar el nombre de factors explicatius d’un conjunt de dades que determinen en major grau la variabilitat d’aquestes dades. Llavors podrem sintetitzar i visualitzar informacio util en un conjunt de dades que contindra observacions descrites per multiples variables quantitatives correlacionades.

Com hem pogut observar a la nostra mostra o conjunt de dades, tenim un elevat nombre de variables i aixo ens dificulta la visualitzacio de la informacio que volem tractar en un espai multi-dimensional.

Gracies al procediment explicat aconseguirem reduir la dimensionalitat de les nostes dades en un baix nombre de components que podrem visualitzar graficament amb la menor perdua de informacio i variança possible.

### Data format & analysis

Abans de res, prepararem les dades necessaries per realitzar l’analisi de components principals. Escollirem les variables actives que ens permetran realitzar el PCA i tambe seleccionarem un conjunt de variables suplementaries.

## Create PCA

Hem agrupat totes les variables numeriques, les quals utilitzarem com a variables actives menys el target numeric “duration” i com a variables suplementaries tenim “y”, “marital” y “job”, encara que havíem també seleccionat “education”, però la mostra no era del tot concluent.

**names**(df)

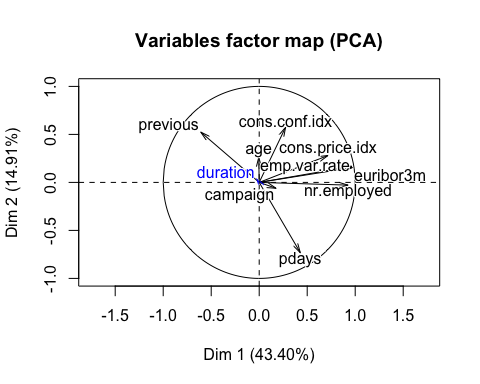
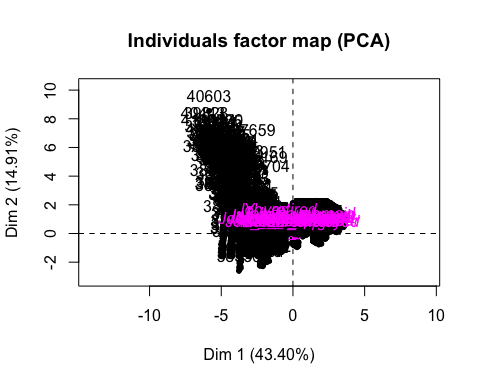
## [1] "age" "job"   
## [3] "marital" "education"   
## [5] "default" "housing"   
## [7] "loan" "contact"   
## [9] "month" "day\_of\_week"   
## [11] "duration" "campaign"   
## [13] "pdays" "previous"   
## [15] "poutcome" "emp.var.rate"   
## [17] "cons.price.idx" "cons.conf.idx"   
## [19] "euribor3m" "nr.employed"   
## [21] "y" "missings\_indiv"   
## [23] "errors\_indiv" "outliers\_indiv"   
## [25] "season" "factor\_age"   
## [27] "factor\_duration" "factor\_campaign"   
## [29] "factor\_Pdays" "factor\_Previous"   
## [31] "factor\_emp.var.rate" "factor\_cons.price.idx"  
## [33] "factor\_cons.conf.idx" "factor\_euribor3m"   
## [35] "factor\_nr.employed"

vars\_conaux <- **names**(df)[**c**(1,12**:**14,16**:**20)]  
vars\_conaux

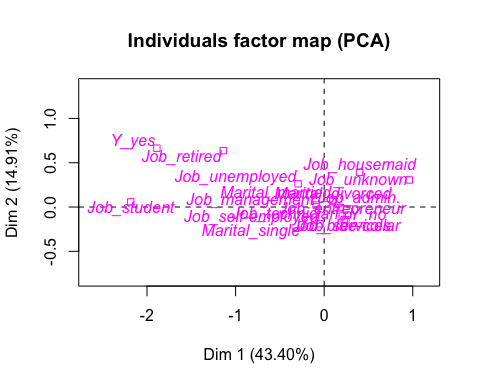
## [1] "age" "campaign" "pdays" "previous"   
## [5] "emp.var.rate" "cons.price.idx" "cons.conf.idx" "euribor3m"   
## [9] "nr.employed"

res.pca<-**PCA**(df[,**c**("duration","y","marital","job",vars\_conaux)],quanti.sup = 1,quali.sup = 2**:**4)

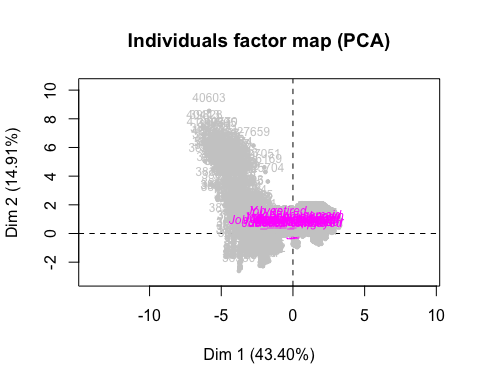
## Warning in PCA(df[, c("duration", "y", "marital", "job", vars\_conaux)], :  
## Missing values are imputed by the mean of the variable: you should use the  
## imputePCA function of the missMDA package



*#LES VARIABLES ACTIVES NO PODEN SER FACTORS!*  
  
**plot.PCA**(res.pca,choix = "ind", invisible = "ind")



**plot**(res.pca,choix="ind", cex=0.75, col.ind="grey80")



*#par(mfrow=c(1,2)) poner dos graficos juntos!*

La funcio PCA() ha realitzat el PCA del nostre conjunt de dades. Visualitzarem dos grafics, tenim el “Variables factor map” i el “Individuals factor map” que detallarem amb més profunditat posteriorment.

En el grafic “Variables factor map” podem observar que les variables “previous”" i “pdays” es troben totalment oposades i tambe veiem que el nostre target (variable quantitativa suplementaria) “duration” no te res a veure amb les variables numeriques ja que la fetxa es molt curta.

### Eigenvalues and dominant axes Analysis

En aquest apartat utilitzarem valors propis (Eigenvalues) per determinar quins components principals considerarem per el nostre analisi (denominat axes).

Concretament els valors propis mesuren la quantitat de variança proporcionada per cada component principal. A partir d’aquesta informacio i les regles de Kaiser i Elbow podrem determinar, com hem dit, els components a considerar i les dimensions necessaries a agafar.

## Kaiser Rule

res.pca**$**eig

## eigenvalue percentage of variance cumulative percentage of variance  
## comp 1 3.90643762 43.4048625 43.40486  
## comp 2 1.34224472 14.9138303 58.31869  
## comp 3 1.03534030 11.5037811 69.82247  
## comp 4 0.98070837 10.8967597 80.71923  
## comp 5 0.84014761 9.3349735 90.05421  
## comp 6 0.46176101 5.1306779 95.18488  
## comp 7 0.39576928 4.3974364 99.58232  
## comp 8 0.02438733 0.2709704 99.85329  
## comp 9 0.01320375 0.1467083 100.00000

Quan executem aquesta comanda podem visualitzar una taula on observem els valors propis (eigenvalues) de cada component principal.

La primera columna mostra el valor propi per cada component, la suma de tots els valors propis ens dona una variança de 9. En la segona columna podem observar la proporcio de variança de cada component i en la tercera el percentatge acomulat de variança obtingut a partir de la suma dels successius components.

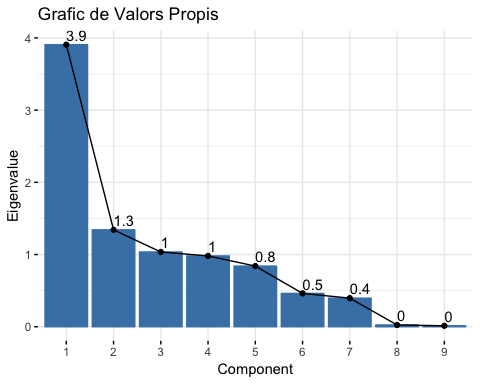
La regla de Kaiser diu que un valor propi (eigenvalue) amb valor superior a 1 indica que les components principals compten amb mes variança que una de les variables originals en dades estandaritzades.

Després de la execució, a partir de la taula de valors propis i seguint la regla de Kaiser hem decidit tenir en compte les 4 primeres components principals. Com podem veure el valor propi de la component numero 4 no supera el valor 1, però el seu valor es de 0.9807 que es molt proxim a 1, llavors tambe es podria considerar agafar-la. Amb el nostre percentatge de variança (69.822) podem dir que quasi tres quarts (75%) de les nostres dades queden representades amb aquestes 3 components principals i si agafessim les 4 components seria una mica mes de tres quarts de les nostres dades (80.719).

## Elbow Rule

Tambe tenim un altre metode d’interpretacio i validacio de les nostres components i aquest es el “Elbow Rule”, que utilitza un grafic dels valors propis ordenats de major a menor i determina el nombre de components principals a considerar fins al punt del grafic en el qual el valor propi es relativament petit.

**fviz\_eig**(res.pca, choice = "eigenvalue", addlabels = TRUE, main = "Grafic de Valors Propis", xlab = "Component", ylab = "Eigenvalue")



Com podem observar al grafic dels valors propis, segons la regla d’elbow hauriem de considerar les 7 primeres components principals. Tot i així, en el nostre cas, decidim considerar les 3 primeres components principals ja que ens proporcionen una variança totalment acceptable (80.71%) i en el cas d’utilitzar les 7 components obtindrem una dimensionalitat massa elevada, fet que no ens interessa molt.

### Individuals point of view

En aquest apartat estudiarem diferents aspectes del nostre conjunt de dades i de les nostres components principals a partir del individus de la nostra mostra.

## Individuals contribution

Ara el que farem es estudiar les possibles contribucions per part d’alguns individus.

*#Hacemos esto para poder ver los tres más contributivos al segundo eje de las 4 dimensiones que hemos cogido*  
  
**sort**(res.pca**$**ind**$**contrib[,1],decreasing = TRUE)[1**:**3]

## 40443 41004 38275   
## 0.2035832 0.2016805 0.1941485

*#Se ha de hacer con which*  
df["40443",]

## age job marital education default  
## 40443 26 Job\_admin. Marital\_single Education\_university.degree Default\_no  
## housing loan contact month day\_of\_week  
## 40443 Housing\_no Loan\_no Contact\_cellular Month\_aug Day\_of\_week\_mon  
## duration campaign pdays previous poutcome emp.var.rate  
## 40443 242 1 6 5 Poutcome\_success -1.7  
## cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 40443 94.027 -38.3 0.904 4991.6 Y\_yes  
## missings\_indiv errors\_indiv outliers\_indiv season factor\_age  
## 40443 0 0 0 Summer factor\_age [17,31]  
## factor\_duration factor\_campaign factor\_Pdays  
## 40443 factor\_duration-(236,329] factor\_campaign-[1,2] factor\_Pdays-[0,15]  
## factor\_Previous factor\_emp.var.rate  
## 40443 factor\_Previous-(1,5] factor\_emp.var.rate-(-1.8,-0.1]  
## factor\_cons.price.idx factor\_cons.conf.idx  
## 40443 factor\_cons.price.idx-(94,94.8] factor\_cons.conf.idx-(-40.3,-36.4]  
## factor\_euribor3m factor\_nr.employed  
## 40443 factor\_euribor3m-[0.634,1.266] factor\_nr.employed-[4.96e+03,5.1e+03]

**sort**(res.pca**$**ind**$**contrib[,2],decreasing = TRUE)[1**:**3]

## 40603 39828 40443   
## 1.1009452 0.8130194 0.8116665

df["40603",]

## age job marital education  
## 40603 59 Job\_services Marital\_married Education\_professional.course  
## default housing loan contact month  
## 40603 Default\_no Housing\_yes Loan\_no Contact\_cellular Month\_sep  
## day\_of\_week duration campaign pdays previous poutcome  
## 40603 Day\_of\_week\_fri 251 3 2 4 Poutcome\_success  
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 40603 -1.1 94.199 -37.5 0.883 4963.6 Y\_no  
## missings\_indiv errors\_indiv outliers\_indiv season  
## 40603 0 0 0 Aut-Win  
## factor\_age factor\_duration factor\_campaign  
## 40603 factor\_age (49,81] factor\_duration-(236,329] factor\_campaign-(2,3]  
## factor\_Pdays factor\_Previous  
## 40603 factor\_Pdays-[0,15] factor\_Previous-(1,5]  
## factor\_emp.var.rate factor\_cons.price.idx  
## 40603 factor\_emp.var.rate-(-1.8,-0.1] factor\_cons.price.idx-(94,94.8]  
## factor\_cons.conf.idx factor\_euribor3m  
## 40603 factor\_cons.conf.idx-(-40.3,-36.4] factor\_euribor3m-[0.634,1.266]  
## factor\_nr.employed  
## 40603 factor\_nr.employed-[4.96e+03,5.1e+03]

**sort**(res.pca**$**ind**$**contrib[,3],decreasing = TRUE)[1**:**3]

## 40930 41004 37819   
## 0.7201366 0.5128497 0.4860395

df["40930",]

## age job marital education default housing  
## 40930 20 Job\_student Marital\_single <NA> Default\_no Housing\_yes  
## loan contact month day\_of\_week duration  
## 40930 Loan\_yes Contact\_cellular Month\_oct Day\_of\_week\_tue 187  
## campaign pdays previous poutcome emp.var.rate cons.price.idx  
## 40930 1 3 4 Poutcome\_success -1.1 94.601  
## cons.conf.idx euribor3m nr.employed y missings\_indiv  
## 40930 -49.5 0.982 4963.6 Y\_yes 0  
## errors\_indiv outliers\_indiv season factor\_age  
## 40930 0 0 Aut-Win factor\_age [17,31]  
## factor\_duration factor\_campaign factor\_Pdays  
## 40930 factor\_duration-(182,236] factor\_campaign-[1,2] factor\_Pdays-[0,15]  
## factor\_Previous factor\_emp.var.rate  
## 40930 factor\_Previous-(1,5] factor\_emp.var.rate-(-1.8,-0.1]  
## factor\_cons.price.idx factor\_cons.conf.idx  
## 40930 factor\_cons.price.idx-(94,94.8] factor\_cons.conf.idx-[-50.8,-46.2]  
## factor\_euribor3m factor\_nr.employed  
## 40930 factor\_euribor3m-[0.634,1.266] factor\_nr.employed-[4.96e+03,5.1e+03]

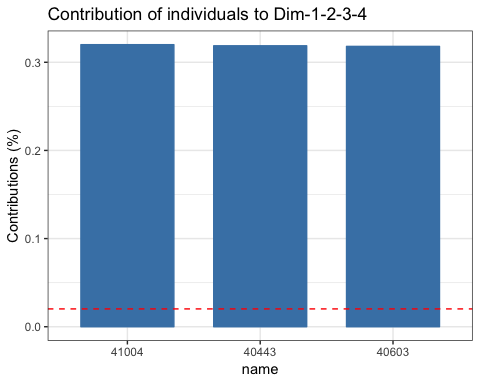
**sort**(res.pca**$**ind**$**contrib[,4],decreasing = TRUE)[1**:**3]

## 35442 33741 11630   
## 0.6914135 0.6822475 0.6640766

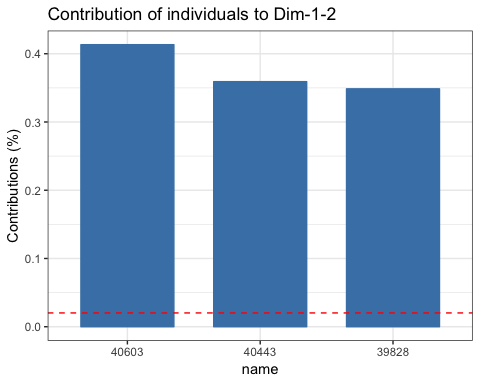
df["35442",]

## age job marital education default  
## 35442 36 Job\_admin. Marital\_married Education\_high.school Default\_unknown  
## housing loan contact month day\_of\_week  
## 35442 Housing\_no Loan\_no Contact\_cellular Month\_may Day\_of\_week\_mon  
## duration campaign pdays previous poutcome emp.var.rate  
## 35442 11 14 16 0 Poutcome\_nonexistent -1.8  
## cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 35442 92.893 -46.2 1.244 5099.1 Y\_no  
## missings\_indiv errors\_indiv outliers\_indiv season factor\_age  
## 35442 1 0 0 Spring factor\_age (31,36]  
## factor\_duration factor\_campaign factor\_Pdays  
## 35442 factor\_duration-[1,68] factor\_campaign-(3,14] factor\_Pdays-(15,17]  
## factor\_Previous factor\_emp.var.rate  
## 35442 factor\_Previous-[0,1] factor\_emp.var.rate-[-3.4,-1.8]  
## factor\_cons.price.idx factor\_cons.conf.idx  
## 35442 factor\_cons.price.idx-[92.2,93] factor\_cons.conf.idx-[-50.8,-46.2]  
## factor\_euribor3m factor\_nr.employed  
## 35442 factor\_euribor3m-[0.634,1.266] factor\_nr.employed-[4.96e+03,5.1e+03]

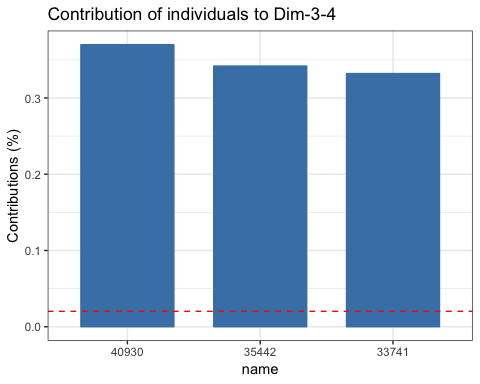
*#fviz\_pca\_var(res.pca)*  
**fviz\_contrib**(res.pca, choice = "ind", axes = 1**:**4, top = 3)**+theme\_bw**()



*#Aqui fem el mateix pero separant les dimensions per fer-ho en dos grafics diferents*  
**fviz\_contrib**(res.pca, choice = "ind", axes = 1**:**2, top = 3)**+theme\_bw**()



**fviz\_contrib**(res.pca, choice = "ind", axes = 3**:**4, top = 3)**+theme\_bw**()



A partir dels dos grafics anteriors veiem que per cada parell de dimensions hi ha individus determinats que tenen una contribucio elevada.

## Individuals best representation

Ara veurem els individuals que tenen una millor representació

*#Millor representats*  
  
**sort**(res.pca**$**ind**$**cos2[,1],decreasing = TRUE)[1**:**3]

## 38571 38490 38345   
## 0.8867685 0.8752577 0.8582645

df["38571",]

## age job marital education  
## 38571 34 Job\_technician Marital\_single Education\_university.degree  
## default housing loan contact month  
## 38571 Default\_no Housing\_no Loan\_no Contact\_cellular Month\_oct  
## day\_of\_week duration campaign pdays previous poutcome  
## 38571 Day\_of\_week\_thu 136 1 16 1 Poutcome\_failure  
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed  
## 38571 -3.4 92.431 -42.33883 0.722 5017.5  
## y missings\_indiv errors\_indiv outliers\_indiv season  
## 38571 Y\_yes 1 0 1 Aut-Win  
## factor\_age factor\_duration factor\_campaign  
## 38571 factor\_age (31,36] factor\_duration-(104,139] factor\_campaign-[1,2]  
## factor\_Pdays factor\_Previous  
## 38571 factor\_Pdays-(15,17] factor\_Previous-[0,1]  
## factor\_emp.var.rate factor\_cons.price.idx  
## 38571 factor\_emp.var.rate-[-3.4,-1.8] factor\_cons.price.idx-[92.2,93]  
## factor\_cons.conf.idx factor\_euribor3m  
## 38571 factor\_cons.conf.idx-(-46.2,-42] factor\_euribor3m-[0.634,1.266]  
## factor\_nr.employed  
## 38571 factor\_nr.employed-[4.96e+03,5.1e+03]

**sort**(res.pca**$**ind**$**cos2[,2],decreasing = TRUE)[1**:**3]

## 40603 39181 39505   
## 0.5929517 0.5861391 0.5856818

df["40603",]

## age job marital education  
## 40603 59 Job\_services Marital\_married Education\_professional.course  
## default housing loan contact month  
## 40603 Default\_no Housing\_yes Loan\_no Contact\_cellular Month\_sep  
## day\_of\_week duration campaign pdays previous poutcome  
## 40603 Day\_of\_week\_fri 251 3 2 4 Poutcome\_success  
## emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 40603 -1.1 94.199 -37.5 0.883 4963.6 Y\_no  
## missings\_indiv errors\_indiv outliers\_indiv season  
## 40603 0 0 0 Aut-Win  
## factor\_age factor\_duration factor\_campaign  
## 40603 factor\_age (49,81] factor\_duration-(236,329] factor\_campaign-(2,3]  
## factor\_Pdays factor\_Previous  
## 40603 factor\_Pdays-[0,15] factor\_Previous-(1,5]  
## factor\_emp.var.rate factor\_cons.price.idx  
## 40603 factor\_emp.var.rate-(-1.8,-0.1] factor\_cons.price.idx-(94,94.8]  
## factor\_cons.conf.idx factor\_euribor3m  
## 40603 factor\_cons.conf.idx-(-40.3,-36.4] factor\_euribor3m-[0.634,1.266]  
## factor\_nr.employed  
## 40603 factor\_nr.employed-[4.96e+03,5.1e+03]

**sort**(res.pca**$**ind**$**cos2[,3],decreasing = TRUE)[1**:**3]

## 37819 27018 26458   
## 0.7361513 0.6887437 0.6855514

df["37819",]

## age job marital education default  
## 37819 80 Job\_retired Marital\_married Education\_basic.4y Default\_no  
## housing loan contact month day\_of\_week  
## 37819 Housing\_yes Loan\_no Contact\_cellular Month\_aug Day\_of\_week\_wed  
## duration campaign pdays previous poutcome emp.var.rate  
## 37819 323 1 16 0 Poutcome\_nonexistent -2.9  
## cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 37819 92.201 -31.4 0.834 5076.2 Y\_yes  
## missings\_indiv errors\_indiv outliers\_indiv season factor\_age  
## 37819 1 0 0 Summer factor\_age (49,81]  
## factor\_duration factor\_campaign factor\_Pdays  
## 37819 factor\_duration-(236,329] factor\_campaign-[1,2] factor\_Pdays-(15,17]  
## factor\_Previous factor\_emp.var.rate  
## 37819 factor\_Previous-[0,1] factor\_emp.var.rate-[-3.4,-1.8]  
## factor\_cons.price.idx factor\_cons.conf.idx  
## 37819 factor\_cons.price.idx-[92.2,93] factor\_cons.conf.idx-(-36.4,-29.8]  
## factor\_euribor3m factor\_nr.employed  
## 37819 factor\_euribor3m-[0.634,1.266] factor\_nr.employed-[4.96e+03,5.1e+03]

**sort**(res.pca**$**ind**$**cos2[,4],decreasing = TRUE)[1**:**3]

## 26278 16663 12711   
## 0.8875421 0.8809677 0.8802130

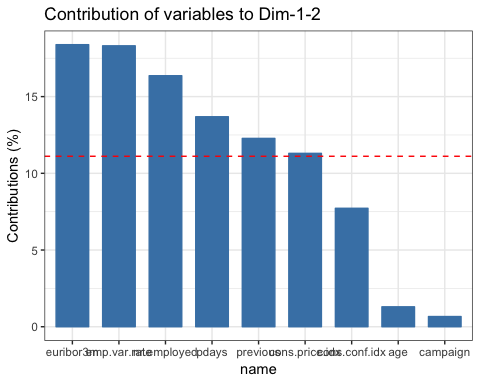
df["26278",]

## age job marital education  
## 26278 47 Job\_blue-collar Marital\_married Education\_basic.9y  
## default housing loan contact month  
## 26278 Default\_unknown Housing\_yes Loan\_no Contact\_telephone Month\_nov  
## day\_of\_week duration campaign pdays previous  
## 26278 Day\_of\_week\_thu 76 9 16 0  
## poutcome emp.var.rate cons.price.idx cons.conf.idx  
## 26278 Poutcome\_nonexistent -0.1 93.2 -42  
## euribor3m nr.employed y missings\_indiv errors\_indiv  
## 26278 4.076 5195.8 Y\_no 1 0  
## outliers\_indiv season factor\_age factor\_duration  
## 26278 0 Aut-Win factor\_age (41,49] factor\_duration-(68,104]  
## factor\_campaign factor\_Pdays factor\_Previous  
## 26278 factor\_campaign-(3,14] factor\_Pdays-(15,17] factor\_Previous-[0,1]  
## factor\_emp.var.rate factor\_cons.price.idx  
## 26278 factor\_emp.var.rate-(-1.8,-0.1] factor\_cons.price.idx-(93,93.4]  
## factor\_cons.conf.idx factor\_euribor3m  
## 26278 factor\_cons.conf.idx-(-46.2,-42] factor\_euribor3m-(1.415,4.856]  
## factor\_nr.employed  
## 26278 factor\_nr.employed-(5.1e+03,5.23e+03]

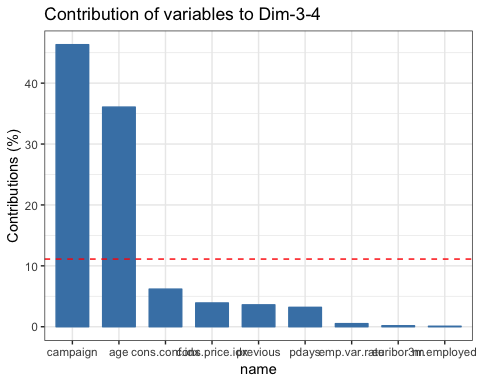
*# Quality of individuals*  
*# head(res.pca$ind$cos2)*

### Variables contribution

**fviz\_contrib**(res.pca, choice = "var", axes = 1**:**2)**+theme\_bw**()



**fviz\_contrib**(res.pca, choice = "var", axes = 3**:**4)**+theme\_bw**()



*#fviz\_contrib(res.pca, choice = "var", axes = 1:4)+theme\_bw()*

Com podem veure en els grafics que surten despres d’executar les comandes anteriors, podem veure que les variables que tenen mes contribucio o els individuals mes contributius son els corresponents a les variables “euribor3m”, “emp.var.rate”, i “nr.employed”, aixo pel que fa a la dim 1-2 i a la dim 3-4 tenim les variables “campaign” i “age” com les mes destacades.

### Interpreting the axes

*# summary(res.pca, nb.dec = 2,ncp = 4)*  
  
**dimdesc**(res.pca, axes = 1**:**4)

## $Dim.1  
## $Dim.1$quanti  
## correlation p.value  
## euribor3m 0.97012135 0.000000e+00  
## emp.var.rate 0.96596055 0.000000e+00  
## nr.employed 0.92622181 0.000000e+00  
## cons.price.idx 0.71732355 0.000000e+00  
## pdays 0.42778256 2.747395e-219  
## cons.conf.idx 0.27475758 2.220057e-86  
## campaign 0.17647126 6.925306e-36  
## duration -0.02789008 4.984006e-02  
## previous -0.60838071 0.000000e+00  
##   
## $Dim.1$quali  
## R2 p.value  
## y 0.125386566 4.727704e-146  
## job 0.050547845 1.431720e-48  
## marital 0.006555608 4.090296e-07  
##   
## $Dim.1$category  
## Estimate p.value  
## Y\_no 1.07413334 4.727704e-146  
## Job\_blue-collar 0.38172472 1.578463e-06  
## Marital\_married 0.31200050 1.111008e-05  
## Job\_unknown 1.09681411 1.361795e-03  
## Job\_technician 0.31629473 5.350108e-03  
## Job\_services 0.36150713 9.057754e-03  
## Job\_housemaid 0.53511934 2.165660e-02  
## Marital\_single -0.03928771 2.479330e-08  
## Job\_retired -1.00385338 2.180033e-17  
## Job\_student -2.04730655 1.106798e-30  
## Y\_yes -1.07413334 4.727704e-146  
##   
##   
## $Dim.2  
## $Dim.2$quanti  
## correlation p.value  
## cons.conf.idx 0.57422055 0.000000e+00  
## previous 0.52363024 0.000000e+00  
## cons.price.idx 0.28034870 5.339409e-90  
## age 0.26095722 8.309706e-78  
## emp.var.rate 0.16716817 2.513370e-32  
## euribor3m 0.15421659 1.052205e-27  
## duration 0.04037167 4.515730e-03  
## nr.employed -0.02841696 4.567319e-02  
## campaign -0.06123050 1.638747e-05  
## pdays -0.73167488 0.000000e+00  
##   
## $Dim.2$quali  
## R2 p.value  
## y 0.04515955 1.302046e-51  
## job 0.02376411 3.195565e-20  
## marital 0.01239773 2.598827e-13  
##   
## $Dim.2$category  
## Estimate p.value  
## Y\_yes 0.377861724 1.302046e-51  
## Job\_retired 0.538789877 6.759304e-16  
## Marital\_single 0.004356898 2.903848e-14  
## Marital\_married 0.287205205 6.730200e-10  
## Job\_housemaid 0.295587120 1.140921e-04  
## Job\_unemployed 0.164414804 1.813003e-02  
## Job\_self-employed -0.286908713 4.048030e-02  
## Job\_services -0.237667993 5.602952e-03  
## Job\_blue-collar -0.243007374 1.206646e-06  
## Y\_no -0.377861724 1.302046e-51  
##   
##   
## $Dim.3  
## $Dim.3$quanti  
## correlation p.value  
## age 0.82888610 0.000000e+00  
## cons.conf.idx 0.34042071 1.951278e-134  
## pdays 0.24888801 1.013084e-70  
## duration -0.03380074 1.744413e-02  
## emp.var.rate -0.09882178 3.278995e-12  
## campaign -0.11007792 8.282436e-15  
## previous -0.26339058 2.783976e-79  
## cons.price.idx -0.27971856 1.379085e-89  
##   
## $Dim.3$quali  
## R2 p.value  
## job 0.178569210 2.700620e-201  
## marital 0.108636234 7.136894e-123  
## y 0.002560617 3.706473e-04  
##   
## $Dim.3$category  
## Estimate p.value  
## Job\_retired 1.68149671 5.680668e-154  
## Marital\_married 0.20097034 4.022987e-60  
## Marital\_divorced 0.32318275 1.363916e-14  
## Job\_management 0.21386530 7.974613e-09  
## Job\_housemaid 0.26294711 7.749768e-05  
## Y\_no 0.07902348 3.706473e-04  
## Job\_unknown 0.32659224 6.931541e-03  
## Job\_technician -0.19259908 2.190366e-03  
## Job\_blue-collar -0.17391611 1.561623e-03  
## Y\_yes -0.07902348 3.706473e-04  
## Job\_admin. -0.19544362 1.450930e-05  
## Job\_services -0.28666694 1.024159e-05  
## Job\_student -1.31635556 2.904328e-36  
## Marital\_single -0.52342349 6.555178e-124  
##   
##   
## $Dim.4  
## $Dim.4$quanti  
## correlation p.value  
## campaign 0.96002062 0.000000e+00  
## age 0.20031553 6.085712e-46  
## previous 0.05584528 8.510511e-05  
## duration -0.03555363 1.239946e-02  
## nr.employed -0.03657601 1.009608e-02  
## pdays -0.04772882 7.858589e-04  
## euribor3m -0.05064684 3.662674e-04  
## cons.conf.idx -0.09302262 5.577676e-11  
##   
## $Dim.4$quali  
## R2 p.value  
## marital 0.006967409 1.511213e-07  
## job 0.008422687 1.773990e-05  
##   
## $Dim.4$category  
## Estimate p.value  
## Job\_retired 0.3107102 3.526031e-07  
## NA 0.5653440 4.474349e-02  
## Job\_student -0.3084071 3.956014e-03  
## Marital\_married -0.1436551 2.306144e-04  
## Marital\_single -0.3104477 3.333871e-08

Ara comentarem a partir de les comandes executades anteriorment quines variables son mes explicatives segons cada dimensio:

A la dimensio 1 les variables mes explicatives son les que mostren els diferents indicadors relacionats amb l’individu i l’estat de l’economia. Tambe podem veure que la variable previous (numero de cops que s’ha contactat amb el client anteriorment) es inversament proporcional.

A la dimensio 2 la variable mes clarament explicativa es “cons.conf.idx”, que es l’index de confiança del consumidor.

A la dimensio 3 veiem que “age”, “cons.conf.idx” i “pdays” tenen una alta contribucio, les dues variables relacionades amb la confiança i amb aspectes especifics d’aquest client abans de realitzar l’actual campanya.

Finalment a la dimensio 4, veiem que “campaign” i “age” son les variables mes explicatives.

### K-Means Classification

Ara farem un nou metode d’agrupament, que es el clustering i ens permetra buscar dins de les nostres observacions grups d’individus amb caracteristiques similars.

*# Fixed number of groups/clusters*  
  
dclu<-res.pca**$**ind**$**coord[,1**:**4] *# Significant axes*  
kcla <- **kmeans**(dclu,7) *# No less than 6 groups*  
  
*#names(kcla)*  
*#summary(kcla)*  
**table**(kcla**$**cluster)

##   
## 1 2 3 4 5 6 7   
## 312 1744 828 166 1245 376 275

kcla**$**totss *#inercia total*

## [1] 35931.36

kcla**$**betweenss *#inercia entre grups*

## [1] 28960.27

kcla**$**withinss *#inercia intra grups*

## [1] 809.6068 1638.6338 875.0929 811.3854 1164.1523 822.3318 849.8907

*#Set clusters m'expliquen una mica mes d'un 80% de l'informacio, es la qualitat de la representacio*  
info<-kcla**$**betweenss**/**kcla**$**totss  
info

## [1] 0.8059886

Sabem que no hi ha una manera del tot correcte per determinar el nombre de clusters, pero sabem que no hem d’agafar menys de 6, pero sabem que s’han d’agafar un minim per a que el nombre de clusters sigui mes optim i poder veure una bona representacio dels nostres clusters. Podem comprobar que amb set clusters tenim una mica més d’un 80% de qualitat en la representació de l’informació i aixo ho sabem amb la nostra nova variable creada “info”.

## Descripició dels clusters

nbcluster <- 7  
df**$**CLUSTER <- nbcluster  
df[**names**(kcla**$**cluster),"CLUSTER"]<-kcla**$**cluster  
  
df**$**f.CLUSTER <- **factor**(df**$**CLUSTER, labels = **c**("CLUSTER-1","CLUSTER-2","CLUSTER-3","CLUSTER-4","CLUSTER-5","CLUSTER-6","CLUSTER-7"))  
  
*#df$kcla<-factor(kcla$cluster)*  
*#names(df)*  
*#catdes(df,34,prob=0.005)*  
*#res.pca<-PCA(df[,c("duration","y","kcla",vars\_con)],quanti.sup=1,quali.sup=2:3,ncp=4)*  
*#plot.PCA(res.pca,choix="ind", habillage=3)*  
  
**names**(df)

## [1] "age" "job"   
## [3] "marital" "education"   
## [5] "default" "housing"   
## [7] "loan" "contact"   
## [9] "month" "day\_of\_week"   
## [11] "duration" "campaign"   
## [13] "pdays" "previous"   
## [15] "poutcome" "emp.var.rate"   
## [17] "cons.price.idx" "cons.conf.idx"   
## [19] "euribor3m" "nr.employed"   
## [21] "y" "missings\_indiv"   
## [23] "errors\_indiv" "outliers\_indiv"   
## [25] "season" "factor\_age"   
## [27] "factor\_duration" "factor\_campaign"   
## [29] "factor\_Pdays" "factor\_Previous"   
## [31] "factor\_emp.var.rate" "factor\_cons.price.idx"  
## [33] "factor\_cons.conf.idx" "factor\_euribor3m"   
## [35] "factor\_nr.employed" "CLUSTER"   
## [37] "f.CLUSTER"

sel <- **c**(1**:**21)  
  
vars\_km <- **names**(df[sel])   
  
vars <- **c**(vars\_km,"f.CLUSTER")  
targ <- **which**(vars **==** "f.CLUSTER")  
**catdes**(df[,vars],targ)

##   
## Link between the cluster variable and the categorical variables (chi-square test)  
## =================================================================================  
## p.value df  
## month 0.000000e+00 54  
## poutcome 0.000000e+00 12  
## y 7.309361e-189 6  
## job 2.405672e-165 66  
## contact 1.740516e-145 6  
## marital 4.107931e-75 18  
## default 1.164839e-52 6  
## education 1.497215e-20 42  
## day\_of\_week 3.433224e-05 24  
##   
## Description of each cluster by the categories  
## =============================================  
## $`CLUSTER-1`  
## Cla/Mod Mod/Cla Global  
## job=Job\_retired 44.660194 29.4871795 4.1649818  
## poutcome=Poutcome\_failure 21.115538 33.9743590 10.1496159  
## y=Y\_yes 17.420436 33.3333333 12.0703599  
## contact=Contact\_cellular 8.663683 86.8589744 63.2430247  
## month=Month\_sep 40.579710 8.9743590 1.3950667  
## default=Default\_no 7.332652 91.9871795 79.1346543  
## month=Month\_dec 50.000000 3.5256410 0.4448039  
## education=Education\_basic.4y 12.350598 19.8717949 10.1496159  
## month=Month\_oct 24.096386 6.4102564 1.6781237  
## month=Month\_mar 25.373134 5.4487179 1.3546300  
## marital=Marital\_married 7.566667 72.7564103 60.6550748  
## month=Month\_apr 12.455516 11.2179487 5.6813587  
## day\_of\_week=Day\_of\_week\_tue 9.191584 26.6025641 18.2571775  
## month=Month\_aug 9.212283 22.1153846 15.1435503  
## job=Job\_management 9.565217 10.5769231 6.9753336  
## marital=Marital\_divorced 8.718861 15.7051282 11.3627173  
## education=Education\_university.degree 7.402423 35.2564103 30.0444804  
## day\_of\_week=Day\_of\_week\_wed 4.809619 15.3846154 20.1779216  
## job=Job\_student 0.952381 0.3205128 2.1229276  
## day\_of\_week=Day\_of\_week\_thu 4.575786 15.3846154 21.2090578  
## education=Education\_high.school 4.553571 16.3461538 22.6445613  
## poutcome=Poutcome\_success 1.190476 0.6410256 3.3966842  
## job=Job\_technician 3.571429 8.9743590 15.8511929  
## education=Education\_basic.9y 3.425560 8.3333333 15.3457339  
## month=Month\_jun 2.932099 6.0897436 13.1014962  
## job=Job\_services 1.902748 2.8846154 9.5632835  
## month=Month\_may 3.641975 18.9102564 32.7537404  
## job=Job\_blue-collar 2.797203 10.2564103 23.1298019  
## default=Default\_unknown 2.422481 8.0128205 20.8653457  
## month=Month\_jul 1.791713 5.1282051 18.0549939  
## marital=Marital\_single 2.614379 11.5384615 27.8406793  
## contact=Contact\_telephone 2.255226 13.1410256 36.7569753  
## poutcome=Poutcome\_nonexistent 4.770814 65.3846154 86.4537000  
## y=Y\_no 4.782709 66.6666667 87.9296401  
## p.value v.test  
## job=Job\_retired 2.741695e-59 16.237431  
## poutcome=Poutcome\_failure 7.003146e-33 11.943706  
## y=Y\_yes 1.093042e-24 10.257677  
## contact=Contact\_cellular 8.513033e-22 9.593520  
## month=Month\_sep 1.333337e-16 8.270558  
## default=Default\_no 2.409876e-10 6.332649  
## month=Month\_dec 2.114315e-08 5.602377  
## education=Education\_basic.4y 1.043361e-07 5.319005  
## month=Month\_oct 1.381825e-07 5.267648  
## month=Month\_mar 5.658811e-07 5.002512  
## marital=Marital\_married 4.005160e-06 4.611114  
## month=Month\_apr 8.740549e-05 3.923131  
## day\_of\_week=Day\_of\_week\_tue 1.658807e-04 3.766005  
## month=Month\_aug 7.213479e-04 3.381334  
## job=Job\_management 1.486789e-02 2.435581  
## marital=Marital\_divorced 1.655982e-02 2.396338  
## education=Education\_university.degree 4.049047e-02 2.048710  
## day\_of\_week=Day\_of\_week\_wed 2.592471e-02 -2.227338  
## job=Job\_student 9.145980e-03 -2.606549  
## day\_of\_week=Day\_of\_week\_thu 7.514420e-03 -2.673143  
## education=Education\_high.school 4.743381e-03 -2.823963  
## poutcome=Poutcome\_success 1.357037e-03 -3.203637  
## job=Job\_technician 2.717358e-04 -3.640853  
## education=Education\_basic.9y 1.566510e-04 -3.780282  
## month=Month\_jun 3.957147e-05 -4.109968  
## job=Job\_services 2.723397e-06 -4.690649  
## month=Month\_may 2.011190e-08 -5.611036  
## job=Job\_blue-collar 1.784928e-09 -6.016260  
## default=Default\_unknown 2.409876e-10 -6.332649  
## month=Month\_jul 4.182234e-12 -6.930882  
## marital=Marital\_single 7.887592e-13 -7.163095  
## contact=Contact\_telephone 8.513033e-22 -9.593520  
## poutcome=Poutcome\_nonexistent 7.915205e-23 -9.835527  
## y=Y\_no 1.093042e-24 -10.257677  
##   
## $`CLUSTER-2`  
## Cla/Mod Mod/Cla Global  
## poutcome=Poutcome\_nonexistent 39.663237 97.2477064 86.4537000  
## month=Month\_jul 58.230683 29.8165138 18.0549939  
## contact=Contact\_telephone 46.149615 48.1077982 36.7569753  
## marital=Marital\_single 46.550472 36.7545872 27.8406793  
## y=Y\_no 37.686825 93.9793578 87.9296401  
## month=Month\_jun 53.240741 19.7821101 13.1014962  
## month=Month\_nov 47.859922 14.1055046 10.3922362  
## job=Job\_services 43.974630 11.9266055 9.5632835  
## education=Education\_high.school 39.821429 25.5733945 22.6445613  
## day\_of\_week=Day\_of\_week\_wed 39.979960 22.8784404 20.1779216  
## job=Job\_technician 40.051020 18.0045872 15.8511929  
## education=Education\_basic.9y 39.789196 17.3165138 15.3457339  
## day\_of\_week=Day\_of\_week\_thu 38.036225 22.8784404 21.2090578  
## marital=NA 0.000000 0.0000000 0.1415285  
## month=Month\_aug 32.042724 13.7614679 15.1435503  
## day\_of\_week=Day\_of\_week\_mon 32.458698 19.1513761 20.8046907  
## job=Job\_management 27.246377 5.3899083 6.9753336  
## marital=Marital\_divorced 28.291815 9.1169725 11.3627173  
## month=Month\_dec 0.000000 0.0000000 0.4448039  
## job=Job\_student 13.333333 0.8027523 2.1229276  
## education=Education\_basic.4y 22.709163 6.5366972 10.1496159  
## month=Month\_oct 4.819277 0.2293578 1.6781237  
## marital=Marital\_married 31.466667 54.1284404 60.6550748  
## month=Month\_mar 0.000000 0.0000000 1.3546300  
## month=Month\_sep 0.000000 0.0000000 1.3950667  
## y=Y\_yes 17.587940 6.0206422 12.0703599  
## month=Month\_may 24.012346 22.3050459 32.7537404  
## poutcome=Poutcome\_success 0.000000 0.0000000 3.3966842  
## contact=Contact\_cellular 28.932225 51.8922018 63.2430247  
## job=Job\_retired 1.456311 0.1720183 4.1649818  
## poutcome=Poutcome\_failure 9.561753 2.7522936 10.1496159  
## month=Month\_apr 0.000000 0.0000000 5.6813587  
## p.value v.test  
## poutcome=Poutcome\_nonexistent 9.983079e-74 18.163820  
## month=Month\_jul 2.245179e-54 15.527928  
## contact=Contact\_telephone 6.123716e-34 12.144659  
## marital=Marital\_single 1.820325e-24 10.208297  
## y=Y\_no 5.068766e-24 10.108435  
## month=Month\_jun 9.092724e-24 10.051014  
## month=Month\_nov 5.691155e-10 6.198755  
## job=Job\_services 3.921572e-05 4.112052  
## education=Education\_high.school 3.071926e-04 3.609158  
## day\_of\_week=Day\_of\_week\_wed 5.232636e-04 3.468556  
## job=Job\_technician 2.386167e-03 3.037415  
## education=Education\_basic.9y 4.834718e-03 2.817845  
## day\_of\_week=Day\_of\_week\_thu 3.473033e-02 2.111489  
## marital=NA 4.755159e-02 -1.981354  
## month=Month\_aug 4.458324e-02 -2.008565  
## day\_of\_week=Day\_of\_week\_mon 3.396930e-02 -2.120436  
## job=Job\_management 1.036131e-03 -3.280528  
## marital=Marital\_divorced 1.985520e-04 -3.720852  
## month=Month\_dec 6.832667e-05 -3.982039  
## job=Job\_student 3.773708e-07 -5.080032  
## education=Education\_basic.4y 1.595910e-10 -6.395913  
## month=Month\_oct 3.086689e-11 -6.642375  
## marital=Marital\_married 4.856257e-12 -6.909716  
## month=Month\_mar 1.742771e-13 -7.367178  
## month=Month\_sep 7.194562e-14 -7.484271  
## y=Y\_yes 5.068766e-24 -10.108435  
## month=Month\_may 6.732938e-32 -11.754030  
## poutcome=Poutcome\_success 3.848356e-33 -11.993388  
## contact=Contact\_cellular 6.123716e-34 -12.144659  
## job=Job\_retired 3.356116e-35 -12.379936  
## poutcome=Poutcome\_failure 5.197520e-44 -13.914149  
## month=Month\_apr 9.136003e-56 -15.731946  
##   
## $`CLUSTER-3`  
## Cla/Mod Mod/Cla Global p.value  
## month=Month\_apr 71.5302491 24.2753623 5.6813587 9.058650e-100  
## contact=Contact\_cellular 23.8171355 89.9758454 63.2430247 4.766994e-80  
## month=Month\_may 28.7654321 56.2801932 32.7537404 3.770219e-53  
## default=Default\_no 18.8298416 89.0096618 79.1346543 4.585452e-16  
## job=Job\_student 49.5238095 6.2801932 2.1229276 5.110953e-15  
## marital=Marital\_single 23.3841685 38.8888889 27.8406793 3.735231e-14  
## month=Month\_mar 52.2388060 4.2270531 1.3546300 2.449652e-11  
## y=Y\_yes 23.9530988 17.2705314 12.0703599 1.377677e-06  
## job=Job\_blue-collar 19.8426573 27.4154589 23.1298019 1.594152e-03  
## poutcome=Poutcome\_failure 21.7131474 13.1642512 10.1496159 2.243483e-03  
## month=Month\_oct 30.1204819 3.0193237 1.6781237 2.366941e-03  
## day\_of\_week=Day\_of\_week\_fri 19.3381593 22.5845411 19.5511524 1.728303e-02  
## job=Job\_unknown 4.6511628 0.2415459 0.8693894 2.069975e-02  
## marital=Marital\_divorced 13.3451957 9.0579710 11.3627173 1.957459e-02  
## education=NA 10.9523810 2.7777778 4.2458552 1.717205e-02  
## job=Job\_housemaid 6.3492063 0.9661836 2.5475131 5.317520e-04  
## y=Y\_no 15.7507473 82.7294686 87.9296401 1.377677e-06  
## job=Job\_retired 4.3689320 1.0869565 4.1649818 3.200463e-08  
## marital=Marital\_married 14.2666667 51.6908213 60.6550748 9.627802e-09  
## month=Month\_jun 8.3333333 6.5217391 13.1014962 4.302173e-11  
## poutcome=Poutcome\_success 0.0000000 0.0000000 3.3966842 2.391195e-14  
## default=Default\_unknown 8.8178295 10.9903382 20.8653457 4.585452e-16  
## month=Month\_nov 0.9727626 0.6038647 10.3922362 7.650679e-36  
## month=Month\_aug 1.6021362 1.4492754 15.1435503 2.014127e-47  
## month=Month\_jul 2.0156775 2.1739130 18.0549939 2.454945e-53  
## contact=Contact\_telephone 4.5654565 10.0241546 36.7569753 4.766994e-80  
## v.test  
## month=Month\_apr 21.202484  
## contact=Contact\_cellular 18.945973  
## month=Month\_may 15.345946  
## default=Default\_no 8.122005  
## job=Job\_student 7.824151  
## marital=Marital\_single 7.569896  
## month=Month\_mar 6.676351  
## y=Y\_yes 4.828207  
## job=Job\_blue-collar 3.156975  
## poutcome=Poutcome\_failure 3.055950  
## month=Month\_oct 3.039852  
## day\_of\_week=Day\_of\_week\_fri 2.380631  
## job=Job\_unknown -2.313416  
## marital=Marital\_divorced -2.334404  
## education=NA -2.383003  
## job=Job\_housemaid -3.464230  
## y=Y\_no -4.828207  
## job=Job\_retired -5.530101  
## marital=Marital\_married -5.737159  
## month=Month\_jun -6.593279  
## poutcome=Poutcome\_success -7.627624  
## default=Default\_unknown -8.122005  
## month=Month\_nov -12.498048  
## month=Month\_aug -14.465066  
## month=Month\_jul -15.373761  
## contact=Contact\_telephone -18.945973  
##   
## $`CLUSTER-4`  
## Cla/Mod Mod/Cla Global  
## poutcome=Poutcome\_success 88.0952381 89.1566265 3.3966842  
## y=Y\_yes 20.1005025 72.2891566 12.0703599  
## month=Month\_sep 31.8840580 13.2530120 1.3950667  
## contact=Contact\_cellular 4.7953964 90.3614458 63.2430247  
## month=Month\_oct 22.8915663 11.4457831 1.6781237  
## job=Job\_student 16.1904762 10.2409639 2.1229276  
## default=Default\_no 3.9856924 93.9759036 79.1346543  
## month=Month\_dec 31.8181818 4.2168675 0.4448039  
## month=Month\_mar 13.4328358 5.4216867 1.3546300  
## job=Job\_retired 7.7669903 9.6385542 4.1649818  
## education=Education\_professional.course 5.3781513 19.2771084 12.0299232  
## job=Job\_admin. 4.5349731 35.5421687 26.3040841  
## education=Education\_university.degree 4.3741588 39.1566265 30.0444804  
## job=Job\_unemployed 8.4112150 5.4216867 2.1633643  
## job=Job\_self-employed 0.6578947 0.6024096 3.0731905  
## job=Job\_services 1.6913319 4.8192771 9.5632835  
## education=Education\_basic.6y 1.1194030 1.8072289 5.4185200  
## education=Education\_basic.9y 1.5810277 7.2289157 15.3457339  
## month=Month\_jul 1.2318029 6.6265060 18.0549939  
## job=Job\_blue-collar 1.1363636 7.8313253 23.1298019  
## default=Default\_unknown 0.9689922 6.0240964 20.8653457  
## month=Month\_may 0.8641975 8.4337349 32.7537404  
## contact=Contact\_telephone 0.8800880 9.6385542 36.7569753  
## y=Y\_no 1.0577144 27.7108434 87.9296401  
## poutcome=Poutcome\_nonexistent 0.0000000 0.0000000 86.4537000  
## p.value v.test  
## poutcome=Poutcome\_success 8.703859e-239 32.997907  
## y=Y\_yes 1.563077e-76 18.514996  
## month=Month\_sep 1.485214e-16 8.257688  
## contact=Contact\_cellular 6.857204e-16 8.073035  
## month=Month\_oct 1.540670e-11 6.744017  
## job=Job\_student 5.601823e-08 5.431067  
## default=Default\_no 7.923004e-08 5.368873  
## month=Month\_dec 5.003774e-06 4.564629  
## month=Month\_mar 4.144742e-04 3.530692  
## job=Job\_retired 1.826039e-03 3.117158  
## education=Education\_professional.course 6.245827e-03 2.734589  
## job=Job\_admin. 7.620775e-03 2.668425  
## education=Education\_university.degree 1.092935e-02 2.544950  
## job=Job\_unemployed 1.223931e-02 2.505168  
## job=Job\_self-employed 3.827209e-02 -2.071929  
## job=Job\_services 2.484116e-02 -2.243864  
## education=Education\_basic.6y 2.234725e-02 -2.284413  
## education=Education\_basic.9y 1.505217e-03 -3.173676  
## month=Month\_jul 1.791942e-05 -4.289353  
## job=Job\_blue-collar 1.542134e-07 -5.247457  
## default=Default\_unknown 7.923004e-08 -5.368873  
## month=Month\_may 5.514070e-14 -7.519133  
## contact=Contact\_telephone 6.857204e-16 -8.073035  
## y=Y\_no 1.563077e-76 -18.514996  
## poutcome=Poutcome\_nonexistent 2.421178e-153 -26.378457  
##   
## $`CLUSTER-5`  
## Cla/Mod Mod/Cla Global  
## poutcome=Poutcome\_nonexistent 28.531338 97.99196787 86.4537000  
## default=Default\_unknown 40.794574 33.81526104 20.8653457  
## marital=Marital\_married 30.833333 74.29718876 60.6550748  
## month=Month\_aug 42.723632 25.70281124 15.1435503  
## contact=Contact\_telephone 34.103410 49.79919679 36.7569753  
## y=Y\_no 27.408600 95.74297189 87.9296401  
## education=Education\_basic.4y 38.247012 15.42168675 10.1496159  
## marital=Marital\_divorced 34.875445 15.74297189 11.3627173  
## month=Month\_nov 34.435798 14.21686747 10.3922362  
## job=Job\_management 34.782609 9.63855422 6.9753336  
## job=Job\_housemaid 39.682540 4.01606426 2.5475131  
## month=Month\_may 28.024691 36.46586345 32.7537404  
## job=Job\_retired 34.951456 5.78313253 4.1649818  
## job=Job\_unknown 46.511628 1.60642570 0.8693894  
## education=Education\_university.degree 23.216689 27.71084337 30.0444804  
## job=Job\_services 20.084567 7.63052209 9.5632835  
## education=Education\_high.school 21.785714 19.59839357 22.6445613  
## job=Job\_admin. 21.983090 22.97188755 26.3040841  
## month=Month\_dec 0.000000 0.00000000 0.4448039  
## month=Month\_jul 20.156775 14.45783133 18.0549939  
## month=Month\_jun 17.129630 8.91566265 13.1014962  
## month=Month\_oct 3.614458 0.24096386 1.6781237  
## month=Month\_mar 0.000000 0.00000000 1.3546300  
## month=Month\_sep 0.000000 0.00000000 1.3950667  
## job=Job\_student 0.952381 0.08032129 2.1229276  
## poutcome=Poutcome\_success 0.000000 0.00000000 3.3966842  
## y=Y\_yes 8.877722 4.25702811 12.0703599  
## contact=Contact\_cellular 19.980818 50.20080321 63.2430247  
## default=Default\_no 21.052632 66.18473896 79.1346543  
## poutcome=Poutcome\_failure 4.980080 2.00803213 10.1496159  
## month=Month\_apr 0.000000 0.00000000 5.6813587  
## marital=Marital\_single 8.932462 9.87951807 27.8406793  
## p.value v.test  
## poutcome=Poutcome\_nonexistent 5.574912e-57 15.908020  
## default=Default\_unknown 5.398614e-36 12.525739  
## marital=Marital\_married 3.498103e-31 11.614012  
## month=Month\_aug 1.826830e-30 11.471863  
## contact=Contact\_telephone 1.065106e-27 10.907179  
## y=Y\_no 9.633353e-27 10.705093  
## education=Education\_basic.4y 8.065390e-12 6.837381  
## marital=Marital\_divorced 4.841989e-08 5.457017  
## month=Month\_nov 7.075034e-07 4.959293  
## job=Job\_management 3.671864e-05 4.127213  
## job=Job\_housemaid 2.952524e-04 3.619430  
## month=Month\_may 1.339529e-03 3.207374  
## job=Job\_retired 1.408222e-03 3.192961  
## job=Job\_unknown 2.534928e-03 3.019141  
## education=Education\_university.degree 3.722773e-02 -2.083258  
## job=Job\_services 6.367842e-03 -2.728213  
## education=Education\_high.school 2.729912e-03 -2.996619  
## job=Job\_admin. 1.857985e-03 -3.112041  
## month=Month\_dec 1.669487e-03 -3.143486  
## month=Month\_jul 1.045039e-04 -3.879889  
## month=Month\_jun 1.688691e-07 -5.230700  
## month=Month\_oct 1.281132e-07 -5.281525  
## month=Month\_mar 3.135831e-09 -5.924325  
## month=Month\_sep 1.739524e-09 -6.020432  
## job=Job\_student 1.572260e-12 -7.067962  
## poutcome=Poutcome\_success 2.613913e-22 -9.714554  
## y=Y\_yes 9.633353e-27 -10.705093  
## contact=Contact\_cellular 1.065106e-27 -10.907179  
## default=Default\_no 5.398614e-36 -12.525739  
## poutcome=Poutcome\_failure 3.398883e-36 -12.562395  
## month=Month\_apr 2.489983e-37 -12.767508  
## marital=Marital\_single 6.416022e-69 -17.545698  
##   
## $`CLUSTER-6`  
## Cla/Mod Mod/Cla Global p.value  
## poutcome=Poutcome\_nonexistent 8.793265 100.0000000 86.453700 1.600430e-25  
## month=Month\_jul 15.117581 35.9042553 18.054994 8.156780e-18  
## contact=Contact\_telephone 10.561056 51.0638298 36.756975 4.194641e-09  
## y=Y\_no 8.254771 95.4787234 87.929640 2.174139e-07  
## month=Month\_jun 12.500000 21.5425532 13.101496 2.371417e-06  
## default=Default\_unknown 10.174419 27.9255319 20.865346 6.895477e-04  
## loan=Loan\_no 8.059701 86.1702128 81.277800 9.349510e-03  
## day\_of\_week=Day\_of\_week\_thu 9.246902 25.7978723 21.209058 2.638068e-02  
## job=Job\_student 2.857143 0.7978723 2.122928 4.699588e-02  
## marital=Marital\_single 6.390704 23.4042553 27.840679 4.362357e-02  
## loan=Loan\_yes 5.625000 11.9680851 16.174687 1.794441e-02  
## month=Month\_mar 0.000000 0.0000000 1.354630 4.822099e-03  
## month=Month\_sep 0.000000 0.0000000 1.395067 4.107437e-03  
## month=Month\_oct 0.000000 0.0000000 1.678124 1.333768e-03  
## default=Default\_no 6.923863 72.0744681 79.134654 6.895477e-04  
## month=Month\_nov 3.307393 4.5212766 10.392236 2.222579e-05  
## poutcome=Poutcome\_success 0.000000 0.0000000 3.396684 1.341034e-06  
## y=Y\_yes 2.847571 4.5212766 12.070360 2.174139e-07  
## month=Month\_may 4.876543 21.0106383 32.753740 1.827153e-07  
## contact=Contact\_cellular 5.882353 48.9361702 63.243025 4.194641e-09  
## month=Month\_apr 0.000000 0.0000000 5.681359 1.135353e-10  
## poutcome=Poutcome\_failure 0.000000 0.0000000 10.149616 6.087260e-19  
## v.test  
## poutcome=Poutcome\_nonexistent 10.441628  
## month=Month\_jul 8.597364  
## contact=Contact\_telephone 5.876329  
## y=Y\_no 5.183797  
## month=Month\_jun 4.718884  
## default=Default\_unknown 3.393702  
## loan=Loan\_no 2.599002  
## day\_of\_week=Day\_of\_week\_thu 2.220561  
## job=Job\_student -1.986337  
## marital=Marital\_single -2.017690  
## loan=Loan\_yes -2.366763  
## month=Month\_mar -2.818684  
## month=Month\_sep -2.869791  
## month=Month\_oct -3.208613  
## default=Default\_no -3.393702  
## month=Month\_nov -4.241271  
## poutcome=Poutcome\_success -4.833574  
## y=Y\_yes -5.183797  
## month=Month\_may -5.216114  
## contact=Contact\_cellular -5.876329  
## month=Month\_apr -6.447733  
## poutcome=Poutcome\_failure -8.890430  
##   
## $`CLUSTER-7`  
## Cla/Mod Mod/Cla Global  
## poutcome=Poutcome\_failure 39.0438247 71.2727273 10.149616  
## contact=Contact\_cellular 7.9283887 90.1818182 63.243025  
## month=Month\_may 9.8148148 57.8181818 32.753740  
## marital=Marital\_single 7.9883805 40.0000000 27.840679  
## default=Default\_no 6.2595810 89.0909091 79.134654  
## month=Month\_apr 12.0996441 12.3636364 5.681359  
## job=Job\_student 16.1904762 6.1818182 2.122928  
## y=Y\_yes 9.2127303 20.0000000 12.070360  
## month=Month\_oct 14.4578313 4.3636364 1.678124  
## poutcome=Poutcome\_success 10.7142857 6.5454545 3.396684  
## month=Month\_sep 13.0434783 3.2727273 1.395067  
## day\_of\_week=Day\_of\_week\_fri 7.1354705 25.0909091 19.551152  
## month=Month\_jun 3.0864198 7.2727273 13.101496  
## marital=Marital\_married 4.5666667 49.8181818 60.655075  
## y=Y\_no 5.0586342 80.0000000 87.929640  
## job=Job\_retired 0.4854369 0.3636364 4.164982  
## education=Education\_basic.4y 1.9920319 3.6363636 10.149616  
## default=Default\_unknown 2.9069767 10.9090909 20.865346  
## month=Month\_nov 1.5564202 2.9090909 10.392236  
## month=Month\_aug 1.6021362 4.3636364 15.143550  
## month=Month\_jul 1.4557671 4.7272727 18.054994  
## contact=Contact\_telephone 1.4851485 9.8181818 36.756975  
## poutcome=Poutcome\_nonexistent 1.4265669 22.1818182 86.453700  
## p.value v.test  
## poutcome=Poutcome\_failure 6.338765e-145 25.634232  
## contact=Contact\_cellular 1.723673e-25 10.434584  
## month=Month\_may 1.748450e-18 8.772434  
## marital=Marital\_single 7.703282e-06 4.473269  
## default=Default\_no 8.466879e-06 4.453025  
## month=Month\_apr 1.392197e-05 4.345088  
## job=Job\_student 6.909637e-05 3.979376  
## y=Y\_yes 1.089450e-04 3.869755  
## month=Month\_oct 2.487692e-03 3.024835  
## poutcome=Poutcome\_success 7.460265e-03 2.675568  
## month=Month\_sep 1.778430e-02 2.370079  
## day\_of\_week=Day\_of\_week\_fri 2.036235e-02 2.319603  
## month=Month\_jun 1.754083e-03 -3.128990  
## marital=Marital\_married 1.843993e-04 -3.739483  
## y=Y\_no 1.089450e-04 -3.869755  
## job=Job\_retired 8.641328e-05 -3.925880  
## education=Education\_basic.4y 4.226314e-05 -4.094746  
## default=Default\_unknown 8.466879e-06 -4.453025  
## month=Month\_nov 1.900719e-06 -4.763703  
## month=Month\_aug 6.200589e-09 -5.811256  
## month=Month\_jul 1.745944e-11 -6.725830  
## contact=Contact\_telephone 1.723673e-25 -10.434584  
## poutcome=Poutcome\_nonexistent 7.515621e-142 -25.357039  
##   
##   
## Link between the cluster variable and the quantitative variables  
## ================================================================  
## Eta2 P-value  
## age 0.474040466 0.000000e+00  
## campaign 0.558436885 0.000000e+00  
## pdays 0.892215906 0.000000e+00  
## previous 0.560755628 0.000000e+00  
## emp.var.rate 0.894046500 0.000000e+00  
## cons.price.idx 0.453861592 0.000000e+00  
## cons.conf.idx 0.352386993 0.000000e+00  
## euribor3m 0.973955527 0.000000e+00  
## nr.employed 0.869891520 0.000000e+00  
## duration 0.006155359 3.146859e-05  
##   
## Description of each cluster by quantitative variables  
## =====================================================  
## $`CLUSTER-1`  
## v.test Mean in category Overall mean sd in category  
## age 24.992876 54.1040262 40.0525729 12.9633587  
## cons.conf.idx 12.408521 -37.6166907 -40.6182329 6.8636111  
## previous 8.392611 0.3942308 0.1708451 0.5787631  
## pdays 3.182317 15.9807692 15.6263647 0.1951710  
## campaign -5.053737 1.8397436 2.3891187 1.2785047  
## cons.price.idx -24.309345 92.8144647 93.5857345 0.5526930  
## euribor3m -28.297286 0.9678942 3.6487535 0.2725778  
## nr.employed -28.712276 5053.1480769 5167.8073595 40.2045371  
## emp.var.rate -30.784863 -2.5365385 0.1073999 0.7128929  
## Overall sd p.value  
## age 10.2585844 7.307003e-138  
## cons.conf.idx 4.4137411 2.349529e-35  
## previous 0.4856692 4.754639e-17  
## pdays 2.0320681 1.461020e-03  
## campaign 1.9835304 4.332492e-07  
## cons.price.idx 0.5789159 1.561416e-130  
## euribor3m 1.7286683 3.732084e-176  
## nr.employed 72.8658491 2.680973e-181  
## emp.var.rate 1.5670994 4.178487e-208  
##   
## $`CLUSTER-2`  
## v.test Mean in category Overall mean sd in category  
## euribor3m 34.964558 4.81339966 3.6487535 0.2864047  
## nr.employed 33.750774 5215.19466743 5167.8073595 17.0298403  
## emp.var.rate 33.469955 1.11806193 0.1073999 0.5129521  
## cons.price.idx 26.054971 93.87637787 93.5857345 0.4030072  
## pdays 9.542349 16.00000000 15.6263647 0.0000000  
## cons.conf.idx 6.040870 -40.10447248 -40.6182329 2.8899983  
## campaign -13.047872 1.89042626 2.3891187 1.0282696  
## previous -15.315055 0.02752294 0.1708451 0.1636014  
## age -31.388217 33.84805046 40.0525729 5.1452934  
## Overall sd p.value  
## euribor3m 1.7286683 7.781109e-268  
## nr.employed 72.8658491 1.041496e-249  
## emp.var.rate 1.5670994 1.319276e-245  
## cons.price.idx 0.5789159 1.181688e-149  
## pdays 2.0320681 1.396325e-21  
## cons.conf.idx 4.4137411 1.532856e-09  
## campaign 1.9835304 6.534699e-39  
## previous 0.4856692 6.066061e-53  
## age 10.2585844 2.930161e-216  
##   
## $`CLUSTER-3`  
## v.test Mean in category Overall mean sd in category  
## pdays 5.797820 16.0000000 15.6263647 0.0000000  
## previous -2.545245 0.1316425 0.1708451 0.3381017  
## campaign -9.699313 1.7789855 2.3891187 1.1102043  
## age -13.699771 35.5955424 40.0525729 7.7075184  
## cons.price.idx -31.605862 93.0054674 93.5857345 0.3555162  
## nr.employed -34.897408 5087.1652174 5167.8073595 31.8350959  
## cons.conf.idx -35.509624 -45.5887066 -40.6182329 3.1682232  
## emp.var.rate -40.362573 -1.8985507 0.1073999 0.3905253  
## euribor3m -43.244710 1.2779831 3.6487535 0.1943923  
## Overall sd p.value  
## pdays 2.0320681 6.718246e-09  
## previous 0.4856692 1.092011e-02  
## campaign 1.9835304 3.035350e-22  
## age 10.2585844 1.018447e-42  
## cons.price.idx 0.5789159 3.067183e-219  
## nr.employed 72.8658491 8.138904e-267  
## cons.conf.idx 4.4137411 3.491769e-276  
## emp.var.rate 1.5670994 0.000000e+00  
## euribor3m 1.7286683 0.000000e+00  
##   
## $`CLUSTER-4`  
## v.test Mean in category Overall mean sd in category  
## previous 42.528318 1.7469880 0.1708451 0.9228475  
## cons.conf.idx 7.824800 -37.9827704 -40.6182329 6.0515896  
## duration 4.547735 351.6385542 262.7672867 274.7841904  
## campaign -4.202564 1.7530120 2.3891187 1.0553178  
## cons.price.idx -5.128640 93.3591687 93.5857345 0.8261510  
## emp.var.rate -18.831759 -2.1445783 0.1073999 0.8798621  
## euribor3m -20.520883 0.9417771 3.6487535 0.5259618  
## nr.employed -26.293197 5021.6084337 5167.8073595 49.4738746  
## pdays -66.391579 5.3313253 15.6263647 3.3588376  
## Overall sd p.value  
## previous 0.4856692 0.000000e+00  
## cons.conf.idx 4.4137411 5.084663e-15  
## duration 256.0881160 5.422624e-06  
## campaign 1.9835304 2.639083e-05  
## cons.price.idx 0.5789159 2.918428e-07  
## emp.var.rate 1.5670994 4.147513e-79  
## euribor3m 1.7286683 1.401424e-93  
## nr.employed 72.8658491 2.293978e-152  
## pdays 2.0320681 0.000000e+00  
##   
## $`CLUSTER-5`  
## v.test Mean in category Overall mean sd in category  
## age 34.592509 48.75341365 40.0525729 6.0606902  
## euribor3m 27.183291 4.80089398 3.6487535 0.2850300  
## emp.var.rate 25.150085 1.07373494 0.1073999 0.5016688  
## nr.employed 23.561540 5209.90128514 5167.8073595 17.9321967  
## cons.conf.idx 19.380080 -38.52096386 -40.6182329 2.8913196  
## cons.price.idx 13.001943 93.77028514 93.5857345 0.3715335  
## pdays 7.499251 16.00000000 15.6263647 0.0000000  
## duration -2.122793 249.43855422 262.7672867 242.1298277  
## campaign -8.964372 1.95315496 2.3891187 1.0831782  
## previous -12.660989 0.02008032 0.1708451 0.1402751  
## Overall sd p.value  
## age 10.2585844 3.274542e-262  
## euribor3m 1.7286683 1.023658e-162  
## emp.var.rate 1.5670994 1.410167e-139  
## nr.employed 72.8658491 9.560810e-123  
## cons.conf.idx 4.4137411 1.136703e-83  
## cons.price.idx 0.5789159 1.192733e-38  
## pdays 2.0320681 6.418366e-14  
## duration 256.0881160 3.377118e-02  
## campaign 1.9835304 3.120553e-19  
## previous 0.4856692 9.726559e-37  
##   
## $`CLUSTER-6`  
## v.test Mean in category Overall mean sd in category  
## campaign 50.728690 7.377660 2.3891187 2.2493334  
## emp.var.rate 14.853947 1.261436 0.1073999 0.3583756  
## euribor3m 14.426375 4.885128 3.6487535 0.2632639  
## nr.employed 14.060411 5218.600266 5167.8073595 16.9326584  
## cons.price.idx 12.194694 93.935734 93.5857345 0.3601197  
## pdays 3.708759 16.000000 15.6263647 0.0000000  
## cons.conf.idx 2.580999 -40.053457 -40.6182329 2.9787704  
## age 2.265504 41.204787 40.0525729 8.8773006  
## previous -7.095467 0.000000 0.1708451 0.0000000  
## Overall sd p.value  
## campaign 1.9835304 0.000000e+00  
## emp.var.rate 1.5670994 6.559004e-50  
## euribor3m 1.7286683 3.531615e-47  
## nr.employed 72.8658491 6.649921e-45  
## cons.price.idx 0.5789159 3.317468e-34  
## pdays 2.0320681 2.082779e-04  
## cons.conf.idx 4.4137411 9.851477e-03  
## age 10.2585844 2.348177e-02  
## previous 0.4856692 1.289158e-12  
##   
## $`CLUSTER-7`  
## v.test Mean in category Overall mean sd in category  
## previous 25.936115 0.9090909 0.1708451 0.6052115  
## campaign 9.978225 3.5490909 2.3891187 2.4643548  
## duration -2.048501 232.0218182 262.7672867 238.7423097  
## age -8.119420 35.1709091 40.0525729 7.8079138  
## cons.price.idx -11.740027 93.1874073 93.5857345 0.5662071  
## cons.conf.idx -13.746299 -44.1741210 -40.6182329 4.2645317  
## emp.var.rate -21.377494 -1.8560000 0.1073999 0.4394112  
## nr.employed -23.275177 5068.4105455 5167.8073595 48.7286468  
## euribor3m -24.465933 1.1700255 3.6487535 0.2279239  
## Overall sd p.value  
## previous 0.4856692 2.608160e-148  
## campaign 1.9835304 1.898329e-23  
## duration 256.0881160 4.051090e-02  
## age 10.2585844 4.684180e-16  
## cons.price.idx 0.5789159 7.946097e-32  
## cons.conf.idx 4.4137411 5.360204e-43  
## emp.var.rate 1.5670994 2.164450e-101  
## nr.employed 72.8658491 7.911039e-120  
## euribor3m 1.7286683 3.406047e-132

Ara procedirem a l’explicació de cada cluster:

Cluster 1: En aquest cluster veiem que el nombre de cops que s’ha contactat anteriorment es superior a la mitjana i tambe es pot observar que es caracteritza perque s’ha contactat durant els mesos d’hivern, sobretot desembre.

Cluster 2: En aquest segon cluster veiem que no hi ha hagut cap mena de campanya de marqueting anteriorment i que principalment es caracteritza pels mesos d’estiu, ja que son els que tenen un v.test major, també podem dir que destaquen els individus que estan solters.

Cluster 3: Aquest cluster es caracteritza perque s’ha contactat durant els mesos de la primavera (abril, maig) a la majoria d’individus i les persones d’aquest cluster son la majoria estudiants.

Cluster 4: Aquest cluster es caracteritza perque s’ha contactat durant els mesos de septembre i octubre a la majoria d’individus i veiem que hi ha hagut una campanya de marqueting exitosa anteriorment.

Cluster 5: Aquest cluster es caracteritza perque s’ha contactat durant el mes d’agost principalment i la major part estan casats i a molts els han contactat a traves del mobil.

Cluster 6: Aquest cluster es caracteritza per un tipus d’individu el qual s’ha contactat a traves del mobil i el nombre de contactes realitzats durant aquesta campanya i per a aquest individus es superior a la mitjana.

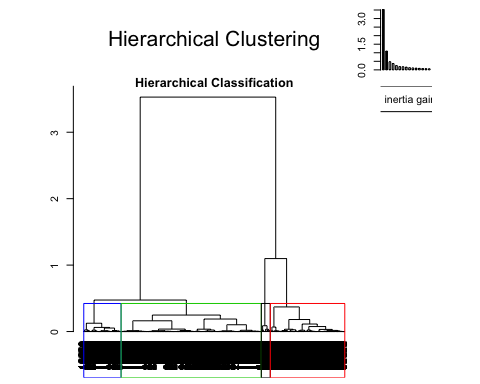
Cluster 7: En aquest cluster veiem que el nombre de cops que s’ha contactat anteriorment es superior a la mitjana i la majoria d’aquest individus estan solters.

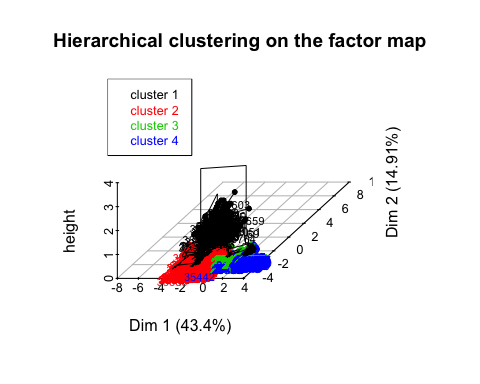
### Hierarchical Clustering

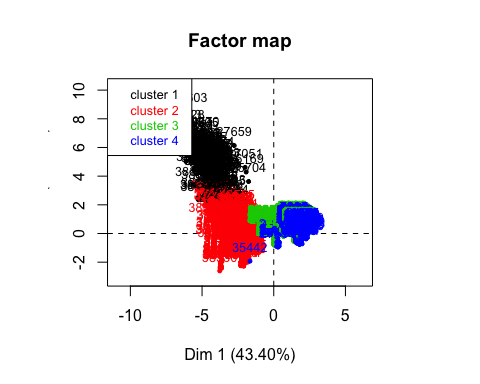
Ara el que farem sera aplicar la classificacio jerarquica de clustering.

Seguidament executem una comanda especifica per poder veure quin es el nombre de clusters mes adequat, ja que aixi podrem veure un grafic on podrem seleccionar com volem agrupar els clusters.

res.hcpc <- **HCPC**(res.pca, nb.clust = 4, order =TRUE) *#Hay que cortar en un punto que no haya muchos saltos apartir de ahi cerca del cero, a primera vista podemos ver que, deberiamos ver grupos uniformes, pero no salen limpias las particiones.*







**attributes**(res.hcpc) *#Tiene esas listas*

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

**summary**(res.hcpc**$**data.clust) *#Nos dice el tamaño de cada cluster*

## duration y marital   
## Min. : 1.0 Y\_no :4349 Marital\_divorced: 562   
## 1st Qu.: 104.0 Y\_yes: 597 Marital\_married :3000   
## Median : 182.0 Marital\_single :1377   
## Mean : 262.8 NA's : 7   
## 3rd Qu.: 329.0   
## Max. :2122.0   
##   
## job age campaign pdays   
## Job\_admin. :1301 Min. :17.00 Min. : 1.000 Min. : 1.00   
## Job\_blue-collar:1144 1st Qu.:32.00 1st Qu.: 1.000 1st Qu.:16.00   
## Job\_technician : 784 Median :38.00 Median : 2.000 Median :16.00   
## Job\_services : 473 Mean :40.05 Mean : 2.389 Mean :15.63   
## Job\_management : 345 3rd Qu.:47.00 3rd Qu.: 3.000 3rd Qu.:16.00   
## Job\_retired : 206 Max. :81.00 Max. :14.000 Max. :16.00   
## (Other) : 693   
## previous emp.var.rate cons.price.idx cons.conf.idx   
## Min. :0.0000 Min. :-3.4000 Min. :92.20 Min. :-50.80   
## 1st Qu.:0.0000 1st Qu.:-1.8000 1st Qu.:93.08 1st Qu.:-42.70   
## Median :0.0000 Median : 1.1000 Median :93.92 Median :-41.80   
## Mean :0.1708 Mean : 0.1074 Mean :93.59 Mean :-40.62   
## 3rd Qu.:0.0000 3rd Qu.: 1.4000 3rd Qu.:93.99 3rd Qu.:-36.40   
## Max. :5.0000 Max. : 1.4000 Max. :94.77 Max. :-29.80   
##   
## euribor3m nr.employed clust   
## Min. :0.634 Min. :4964 1: 180   
## 1st Qu.:1.344 1st Qu.:5099 2:1401   
## Median :4.857 Median :5191 3:2713   
## Mean :3.649 Mean :5168 4: 652   
## 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :5.045 Max. :5228   
##

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

## p.value df  
## y 5.654668e-177 3  
## job 6.528644e-45 33  
## marital 6.394260e-06 9

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

## Eta2 P-value  
## campaign 0.523277769 0.000000e+00  
## pdays 0.844684307 0.000000e+00  
## previous 0.483134285 0.000000e+00  
## emp.var.rate 0.886188857 0.000000e+00  
## cons.price.idx 0.443420176 0.000000e+00  
## euribor3m 0.972728324 0.000000e+00  
## nr.employed 0.862075267 0.000000e+00  
## cons.conf.idx 0.178928568 6.471519e-211  
## duration 0.004360341 7.907623e-05  
## age 0.001841458 2.786614e-02

res.hcpc**$**desc.var**$**quanti

## $`1`  
## v.test Mean in category Overall mean sd in category  
## previous 44.747785 1.7611111 0.1708451 0.9089914  
## cons.conf.idx 7.085967 -38.3296660 -40.6182329 6.1337026  
## duration 4.545145 347.9388889 262.7672867 273.7414263  
## cons.price.idx -3.937550 93.4189333 93.5857345 0.8322883  
## campaign -4.173711 1.7833333 2.3891187 1.1891874  
## emp.var.rate -19.056038 -2.0777778 0.1073999 0.8795552  
## euribor3m -21.375733 0.9448556 3.6487535 0.5073431  
## nr.employed -27.925157 5018.9133333 5167.8073595 50.1367856  
## pdays -64.626980 6.0166667 15.6263647 4.0599329  
## Overall sd p.value  
## previous 0.4856692 0.000000e+00  
## cons.conf.idx 4.4137411 1.380767e-12  
## duration 256.0881160 5.489737e-06  
## cons.price.idx 0.5789159 8.231785e-05  
## campaign 1.9835304 2.996782e-05  
## emp.var.rate 1.5670994 5.854424e-81  
## euribor3m 1.7286683 2.247650e-101  
## nr.employed 72.8658491 1.320822e-171  
## pdays 2.0320681 0.000000e+00  
##   
## $`2`  
## v.test Mean in category Overall mean sd in category  
## previous 14.010599 0.324768 0.1708451 0.5230112  
## pdays 7.336442 15.963597 15.6263647 0.3826571  
## age -2.098021 39.565714 40.0525729 11.9152285  
## campaign -5.633923 2.136331 2.3891187 1.6501597  
## cons.conf.idx -29.535606 -43.567123 -40.6182329 5.4810779  
## cons.price.idx -45.687783 92.987431 93.5857345 0.4524177  
## nr.employed -55.147107 5076.909707 5167.8073595 39.0431169  
## emp.var.rate -60.532277 -2.038401 0.1073999 0.5550071  
## euribor3m -62.859907 1.190700 3.6487535 0.2529861  
## Overall sd p.value  
## previous 0.4856692 1.342670e-44  
## pdays 2.0320681 2.193465e-13  
## age 10.2585844 3.590331e-02  
## campaign 1.9835304 1.761553e-08  
## cons.conf.idx 4.4137411 1.005224e-191  
## cons.price.idx 0.5789159 0.000000e+00  
## nr.employed 72.8658491 0.000000e+00  
## emp.var.rate 1.5670994 0.000000e+00  
## euribor3m 1.7286683 0.000000e+00  
##   
## $`3`  
## v.test Mean in category Overall mean sd in category  
## euribor3m 51.50287 4.79738150 3.6487535 0.2961639  
## emp.var.rate 48.14380 1.08075931 0.1073999 0.5264601  
## nr.employed 47.78941 5212.73276815 5167.8073595 17.5946210  
## cons.price.idx 32.15316 93.82588058 93.5857345 0.3977156  
## cons.conf.idx 19.13856 -39.52841872 -40.6182329 2.9848752  
## pdays 14.25192 16.00000000 15.6263647 0.0000000  
## previous -22.97193 0.02690748 0.1708451 0.1618131  
## campaign -27.58456 1.68322202 2.3891187 0.7827272  
## Overall sd p.value  
## euribor3m 1.7286683 0.000000e+00  
## emp.var.rate 1.5670994 0.000000e+00  
## nr.employed 72.8658491 0.000000e+00  
## cons.price.idx 0.5789159 7.978769e-227  
## cons.conf.idx 4.4137411 1.205556e-81  
## pdays 2.0320681 4.361442e-46  
## previous 0.4856692 8.897223e-117  
## campaign 1.9835304 1.704849e-167  
##   
## $`4`  
## v.test Mean in category Overall mean sd in category  
## campaign 50.391249 6.036810 2.3891187 2.3318009  
## emp.var.rate 20.351770 1.271319 0.1073999 0.3024787  
## euribor3m 19.794810 4.897537 3.6487535 0.2113150  
## nr.employed 18.610074 5217.294939 5167.8073595 17.2495005  
## cons.price.idx 15.733815 93.918144 93.5857345 0.3546711  
## cons.conf.idx 7.263154 -39.448313 -40.6182329 3.0538591  
## pdays 5.038317 16.000000 15.6263647 0.0000000  
## previous -9.639132 0.000000 0.1708451 0.0000000  
## Overall sd p.value  
## campaign 1.9835304 0.000000e+00  
## emp.var.rate 1.5670994 4.478120e-92  
## euribor3m 1.7286683 3.299949e-87  
## nr.employed 72.8658491 2.662485e-77  
## cons.price.idx 0.5789159 8.870154e-56  
## cons.conf.idx 4.4137411 3.781682e-13  
## pdays 2.0320681 4.696423e-07  
## previous 0.4856692 5.464823e-22

Amb la comanda del “chi2” podem observar que les variables “y”, “job” i “marital” son les que mes caracteritzen la particio en els quatre clusters que utilitzarem en el nostre analisi i tambe es podria fer amb 5 clusters, pero com no canviava molt hem vist mes convenient agafar o fer la particio en 4 clusters pel nostre estudi.

## Descripcio dels clusters

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

## $`1`  
## Cla/Mod Mod/Cla Global p.value  
## y=Y\_yes 20.9380235 69.4444444 12.070360 2.066147e-76  
## job=Job\_student 17.1428571 10.0000000 2.122928 3.138518e-08  
## job=Job\_retired 9.2233010 10.5555556 4.164982 1.951846e-04  
## job=Job\_admin. 4.9961568 36.1111111 26.304084 3.190110e-03  
## job=Job\_unemployed 9.3457944 5.5555556 2.163364 6.838252e-03  
## job=Job\_self-employed 0.6578947 0.5555556 3.073190 2.588876e-02  
## job=Job\_services 1.6913319 4.4444444 9.563283 1.061796e-02  
## job=Job\_blue-collar 1.1363636 7.2222222 23.129802 9.936978e-09  
## y=Y\_no 1.2646585 30.5555556 87.929640 2.066147e-76  
## v.test  
## y=Y\_yes 18.499963  
## job=Job\_student 5.533529  
## job=Job\_retired 3.725169  
## job=Job\_admin. 2.948799  
## job=Job\_unemployed 2.704620  
## job=Job\_self-employed -2.227876  
## job=Job\_services -2.555027  
## job=Job\_blue-collar -5.731801  
## y=Y\_no -18.499963  
##   
## $`2`  
## Cla/Mod Mod/Cla Global p.value  
## y=Y\_yes 49.74874 21.1991435 12.0703599 2.352298e-32  
## job=Job\_student 65.71429 4.9250535 2.1229276 1.147770e-15  
## job=Job\_retired 48.05825 7.0663812 4.1649818 9.670365e-10  
## marital=Marital\_single 33.69644 33.1192006 27.8406793 2.569686e-07  
## job=Job\_unknown 11.62791 0.3568879 0.8693894 1.005915e-02  
## job=Job\_housemaid 17.46032 1.5703069 2.5475131 4.501339e-03  
## job=Job\_technician 23.85204 13.3476089 15.8511929 2.166634e-03  
## marital=Marital\_married 26.13333 55.9600286 60.6550748 2.309388e-05  
## y=Y\_no 25.38515 78.8008565 87.9296401 2.352298e-32  
## v.test  
## y=Y\_yes 11.842536  
## job=Job\_student 8.009926  
## job=Job\_retired 6.114758  
## marital=Marital\_single 5.152550  
## job=Job\_unknown -2.573790  
## job=Job\_housemaid -2.840709  
## job=Job\_technician -3.066386  
## marital=Marital\_married -4.232665  
## y=Y\_no -11.842536  
##   
## $`3`  
## Cla/Mod Mod/Cla Global p.value  
## y=Y\_no 59.14003 94.80280133 87.9296401 1.640791e-61  
## marital=Marital\_married 56.56667 62.55068190 60.6550748 2.650603e-03  
## job=Job\_entrepreneur 64.37500 3.79653520 3.2349373 1.346392e-02  
## job=Job\_services 59.83087 10.43125691 9.5632835 2.190374e-02  
## job=Job\_blue-collar 57.69231 24.32731294 23.1298019 2.760138e-02  
## job=Job\_technician 58.29082 16.84482123 15.8511929 3.477799e-02  
## job=Job\_unknown 69.76744 1.10578695 0.8693894 4.818052e-02  
## marital=NA 14.28571 0.03685957 0.1415285 4.004320e-02  
## marital=Marital\_single 50.10893 25.43309989 27.8406793 3.229271e-05  
## job=Job\_retired 32.52427 2.46959086 4.1649818 4.817695e-11  
## job=Job\_student 12.38095 0.47917435 2.1229276 5.264065e-20  
## y=Y\_yes 23.61809 5.19719867 12.0703599 1.640791e-61  
## v.test  
## y=Y\_no 16.548523  
## marital=Marital\_married 3.005597  
## job=Job\_entrepreneur 2.471257  
## job=Job\_services 2.292033  
## job=Job\_blue-collar 2.202906  
## job=Job\_technician 2.110934  
## job=Job\_unknown 1.975773  
## marital=NA -2.053303  
## marital=Marital\_single -4.156665  
## job=Job\_retired -6.576463  
## job=Job\_student -9.158465  
## y=Y\_yes -16.548523  
##   
## $`4`  
## Cla/Mod Mod/Cla Global p.value v.test  
## y=Y\_no 14.210163 94.7852761 87.929640 3.150217e-10 6.291200  
## job=Job\_student 4.761905 0.7668712 2.122928 4.802179e-03 -2.820012  
## y=Y\_yes 5.695142 5.2147239 12.070360 3.150217e-10 -6.291200

Cluster 1: Els individus que pertanyen al cluster numero 1 es detaquen perque tenen la variable “y = yes”, per tant, aixo vol dir que son individus que SI que contracten el producte i a mes tambe podem observar que la majoria d’aquests individus son estudiants.

Cluster 2: Els individus que pertanyen al cluster numero 2 es detaquen perque tenen la variable “y = yes”, per tant, aixo vol dir que son individus que SI que contracten el producte i a mes tambe podem observar que la majoria d’aquests individus son estudiants i estan solters.

Cluster 3: Els individus que pertanyen al cluster numero 3 es detaquen perque tenen la variable “y = no”, per tant, aixo vol dir que son individus que NO contracten el producte i a mes tambe podem observar que la majoria d’aquests individus treballen com empresaris o en el sector de serveis i que estan casats.

Cluster 4: Els individus que pertanyen al cluster numero 4 es detaquen perque tenen la variable “y = no”, per tant, aixo vol dir que son individus que NO contracten el producte i a mes tambe podem observar que la majoria d’aquests individus son estudiants.

### 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  
## 36910 40420 40457 40031 39208   
## 0.8996255 0.9520736 1.0182792 1.0842884 1.1687768   
## --------------------------------------------------------   
## Cluster: 2  
## 34135 31328 31002 32850 32962   
## 0.7368927 0.7400291 0.7406566 0.7427179 0.7427179   
## --------------------------------------------------------   
## Cluster: 3  
## 24034 4467 4473 726 5358   
## 0.6391974 0.6502367 0.6502367 0.6503246 0.6503246   
## --------------------------------------------------------   
## Cluster: 4  
## 5296 7006 3322 6693 1049   
## 0.6445766 0.6572942 0.6627406 0.6627473 0.6627498

res.hcpc**$**desc.ind**$**dist

## Cluster: 1  
## 41004 40603 40930 40443 39828   
## 11.14194 10.75528 10.61921 10.42103 10.07574   
## --------------------------------------------------------   
## Cluster: 2  
## 37819 38061 38985 38677 38583   
## 6.455196 6.447230 6.406478 6.351079 6.344856   
## --------------------------------------------------------   
## Cluster: 3  
## 18895 23309 22214 14894 19305   
## 3.303387 3.303373 3.303371 3.265879 3.249192   
## --------------------------------------------------------   
## Cluster: 4  
## 18491 11713 11630 23559 35442   
## 6.349686 6.335066 6.315248 6.301241 6.048853

*# NO ES NECESSARI!*  
  
#### Characteristic individuals  
para1<-**which**(**rownames**(res.pca**$**ind**$**coord)**%in%names**(res.hcpc**$**desc.ind**$**para[[1]]))  
para2<-**which**(**rownames**(res.pca**$**ind**$**coord)**%in%names**(res.hcpc**$**desc.ind**$**para[[2]]))  
para3<-**which**(**rownames**(res.pca**$**ind**$**coord)**%in%names**(res.hcpc**$**desc.ind**$**para[[3]]))  
para4<-**which**(**rownames**(res.pca**$**ind**$**coord)**%in%names**(res.hcpc**$**desc.ind**$**para[[4]]))  
*# to be completed... as many as cluster you choose*  
  
dist1<-**which**(**rownames**(res.pca**$**ind**$**coord)**%in%names**(res.hcpc**$**desc.ind**$**dist[[1]]))  
dist2<-**which**(**rownames**(res.pca**$**ind**$**coord)**%in%names**(res.hcpc**$**desc.ind**$**dist[[2]]))  
dist3<-**which**(**rownames**(res.pca**$**ind**$**coord)**%in%names**(res.hcpc**$**desc.ind**$**dist[[3]]))  
dist4<-**which**(**rownames**(res.pca**$**ind**$**coord)**%in%names**(res.hcpc**$**desc.ind**$**dist[[4]]))

### Correspondence Analysis (CA)

En la part final del nostre estudi el que farem sera l’analisi de correspondencies simples (CA) per poder analitzar les relacions entre 2 factors de les dades de la nostra mostra.

Per fer l’analisi de correspondencies simples utilitzarem com a target el factor discretitzat factor\_duration i realitzarem dues taules de contingencia per comparar aquest target amb 2 variables qualitatives mes. Aquestes dues variables seran “job” i “factor\_age”.

### Factor\_age i Factor\_duration

*# Contingency tables - Complex : solo cuentan con los target discretizados*  
**names**(df)

## [1] "age" "job"   
## [3] "marital" "education"   
## [5] "default" "housing"   
## [7] "loan" "contact"   
## [9] "month" "day\_of\_week"   
## [11] "duration" "campaign"   
## [13] "pdays" "previous"   
## [15] "poutcome" "emp.var.rate"   
## [17] "cons.price.idx" "cons.conf.idx"   
## [19] "euribor3m" "nr.employed"   
## [21] "y" "missings\_indiv"   
## [23] "errors\_indiv" "outliers\_indiv"   
## [25] "season" "factor\_age"   
## [27] "factor\_duration" "factor\_campaign"   
## [29] "factor\_Pdays" "factor\_Previous"   
## [31] "factor\_emp.var.rate" "factor\_cons.price.idx"  
## [33] "factor\_cons.conf.idx" "factor\_euribor3m"   
## [35] "factor\_nr.employed" "CLUSTER"   
## [37] "f.CLUSTER"

*# Target factor\_duration vs job*  
*# Podemos elegir la variable que queramos con la de f\_duration y en este caso hemos elegido job para este ejemplo*  
  
**table**(df**$**factor\_age, df**$**factor\_duration)

##   
## factor\_duration-[1,68] factor\_duration-(68,104]  
## factor\_age [17,31] 129 127  
## factor\_age (31,36] 155 137  
## factor\_age (36,41] 104 112  
## factor\_age (41,49] 119 108  
## factor\_age (49,81] 122 139  
##   
## factor\_duration-(104,139] factor\_duration-(139,182]  
## factor\_age [17,31] 143 140  
## factor\_age (31,36] 125 123  
## factor\_age (36,41] 101 105  
## factor\_age (41,49] 124 117  
## factor\_age (49,81] 119 135  
##   
## factor\_duration-(182,236] factor\_duration-(236,329]  
## factor\_age [17,31] 135 135  
## factor\_age (31,36] 126 139  
## factor\_age (36,41] 101 110  
## factor\_age (41,49] 126 119  
## factor\_age (49,81] 120 116  
##   
## factor\_duration-(329,504]  
## factor\_age [17,31] 148  
## factor\_age (31,36] 127  
## factor\_age (36,41] 114  
## factor\_age (41,49] 110  
## factor\_age (49,81] 119  
##   
## factor\_duration-(504,2.12e+03]  
## factor\_age [17,31] 156  
## factor\_age (31,36] 130  
## factor\_age (36,41] 83  
## factor\_age (41,49] 130  
## factor\_age (49,81] 118

*#Le digo que calcule unas probabilidades en la dimension 1, calculo los perfiles por fila que tenemos*  
*#Calculo los perfiles de fila y la suma tendria que dar mas o menos 1 y tenemos que ver si es equivalente al perfil marginal fila*  
**prop.table**(**table**(df**$**factor\_age, df**$**factor\_duration), 1) *# Por filas*

##   
## factor\_duration-[1,68] factor\_duration-(68,104]  
## factor\_age [17,31] 0.1159030 0.1141060  
## factor\_age (31,36] 0.1459510 0.1290019  
## factor\_age (36,41] 0.1253012 0.1349398  
## factor\_age (41,49] 0.1248688 0.1133263  
## factor\_age (49,81] 0.1234818 0.1406883  
##   
## factor\_duration-(104,139] factor\_duration-(139,182]  
## factor\_age [17,31] 0.1284816 0.1257862  
## factor\_age (31,36] 0.1177024 0.1158192  
## factor\_age (36,41] 0.1216867 0.1265060  
## factor\_age (41,49] 0.1301154 0.1227702  
## factor\_age (49,81] 0.1204453 0.1366397  
##   
## factor\_duration-(182,236] factor\_duration-(236,329]  
## factor\_age [17,31] 0.1212938 0.1212938  
## factor\_age (31,36] 0.1186441 0.1308851  
## factor\_age (36,41] 0.1216867 0.1325301  
## factor\_age (41,49] 0.1322141 0.1248688  
## factor\_age (49,81] 0.1214575 0.1174089  
##   
## factor\_duration-(329,504]  
## factor\_age [17,31] 0.1329739  
## factor\_age (31,36] 0.1195857  
## factor\_age (36,41] 0.1373494  
## factor\_age (41,49] 0.1154250  
## factor\_age (49,81] 0.1204453  
##   
## factor\_duration-(504,2.12e+03]  
## factor\_age [17,31] 0.1401617  
## factor\_age (31,36] 0.1224105  
## factor\_age (36,41] 0.1000000  
## factor\_age (41,49] 0.1364113  
## factor\_age (49,81] 0.1194332

*#Marginal row profile*  
**prop.table**(**table**(df**$**factor\_duration))

##   
## factor\_duration-[1,68] factor\_duration-(68,104]   
## 0.1271735 0.1259604   
## factor\_duration-(104,139] factor\_duration-(139,182]   
## 0.1237364 0.1253538   
## factor\_duration-(182,236] factor\_duration-(236,329]   
## 0.1229276 0.1251516   
## factor\_duration-(329,504] factor\_duration-(504,2.12e+03]   
## 0.1249495 0.1247473

*#Esta proporcion se mantiene en cualquiera de los colectivos mirados anteriormente? Se tiene que hacer la comparacion*  
  
*#Podemos comprobar ahora los perfiles columna*  
*#Column profile*  
**prop.table**(**table**(df**$**factor\_age, df**$**factor\_duration), 2) *# dim 2*

##   
## factor\_duration-[1,68] factor\_duration-(68,104]  
## factor\_age [17,31] 0.2050874 0.2038523  
## factor\_age (31,36] 0.2464229 0.2199037  
## factor\_age (36,41] 0.1653418 0.1797753  
## factor\_age (41,49] 0.1891892 0.1733547  
## factor\_age (49,81] 0.1939587 0.2231140  
##   
## factor\_duration-(104,139] factor\_duration-(139,182]  
## factor\_age [17,31] 0.2336601 0.2258065  
## factor\_age (31,36] 0.2042484 0.1983871  
## factor\_age (36,41] 0.1650327 0.1693548  
## factor\_age (41,49] 0.2026144 0.1887097  
## factor\_age (49,81] 0.1944444 0.2177419  
##   
## factor\_duration-(182,236] factor\_duration-(236,329]  
## factor\_age [17,31] 0.2220395 0.2180937  
## factor\_age (31,36] 0.2072368 0.2245557  
## factor\_age (36,41] 0.1661184 0.1777060  
## factor\_age (41,49] 0.2072368 0.1922456  
## factor\_age (49,81] 0.1973684 0.1873990  
##   
## factor\_duration-(329,504]  
## factor\_age [17,31] 0.2394822  
## factor\_age (31,36] 0.2055016  
## factor\_age (36,41] 0.1844660  
## factor\_age (41,49] 0.1779935  
## factor\_age (49,81] 0.1925566  
##   
## factor\_duration-(504,2.12e+03]  
## factor\_age [17,31] 0.2528363  
## factor\_age (31,36] 0.2106969  
## factor\_age (36,41] 0.1345219  
## factor\_age (41,49] 0.2106969  
## factor\_age (49,81] 0.1912480

*#Marginal colum profile*  
**prop.table**(**table**(df**$**factor\_age))

##   
## factor\_age [17,31] factor\_age (31,36] factor\_age (36,41]   
## 0.2250303 0.2147190 0.1678124   
## factor\_age (41,49] factor\_age (49,81]   
## 0.1926810 0.1997574

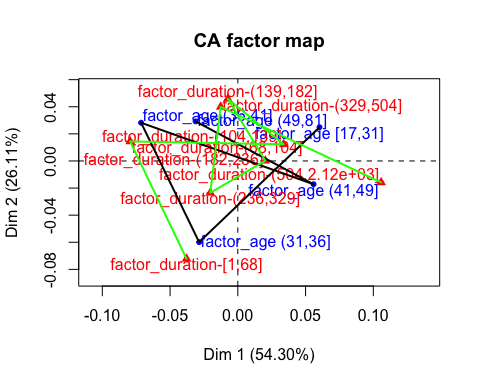
*#El perfil columna de les diferents columnes es pot considerar diferent que el marginal? Evidentment SI*  
  
*# HO: factor\_duration -factor\_age independency*  
**chisq.test**(**table**(df**$**factor\_age, df**$**factor\_duration))

##   
## Pearson's Chi-squared test  
##   
## data: table(df$factor\_age, df$factor\_duration)  
## X-squared = 24.084, df = 28, p-value = 0.6771

*# Accepto la hipotesi nula porque el pvalor es 0.6771*

En aquesta part de la nostra investigacio podem veure que la hipotesi nula s’accepta perque el pvalor es 0.6771, es mes gran que un 5%. Llavors, podem dir que la durada de la trucada no depen de l’edat del nostre individu.

*# CA - factor\_duration vs factor\_age*  
res.ca <- **CA**(**table**(df**$**factor\_age, df**$**factor\_duration))  
*# Interpretacio numerica: Com mes lluny estigui la rodona (blau) hace referencia al factor que esta en las filas y el rojo a las columnas, entonces como mas lejos este del centro de gravedad, quiere decir que es mas remarcables, es decir, mas raro es, los que estan mas cerca no me aporta nada*  
  
*#Link levels in row*  
*#plot.CA(res.ca)*  
**lines**(res.ca**$**row**$**coord[,1], res.ca**$**row**$**coord[,2],lwd=2)  
*#No tenemos que ver nada porque hemos visto que no tienen nada que ver*  
  
*#Link levels in columns*  
**lines**(res.ca**$**col**$**coord[,1], res.ca**$**col**$**coord[,2],lwd=2, col = "green")



Com podem veure a l’hora de l’execució tenim que el factor\_duration-(182,236] es el que mes destaca en que no ens aporta cap mena d’informacio ja que es troba mes a prop del centre de gravetat, A partir de les taules de contingencia i els seus diferents perfils intentem observar si hi pot haver alguna relacio de dependencia entre els dos factors, tot i aixi visualtment ens resulta complicat.

## Eigenvalues and dominant axes analysis

En aquest subapartat realitzarem un estudi dels valors propis i dels eixos dominants per tal de determinar quantes dimensions tindrem en compte.

*#attributes(res.ca)*  
res.ca**$**eig

## eigenvalue percentage of variance  
## dim 1 0.0026443419 54.304636  
## dim 2 0.0012712615 26.106835  
## dim 3 0.0006783276 13.930247  
## dim 4 0.0002755277 5.658282  
## cumulative percentage of variance  
## dim 1 54.30464  
## dim 2 80.41147  
## dim 3 94.34172  
## dim 4 100.00000

*#No es extraño que los eigenvalues sean pequeños, cojemos tantas dimensiones como las que tengan un valor propio > mitjana de este valor*  
**mean**(res.ca**$**eig[,1]) *#Mean eigenvalue*

## [1] 0.001217365

*#KAISER: take as many as dimensions as eigenvalue > mean eig*  
**sum**(res.ca**$**eig[,1]) *#Total inertia, contra mas grande hay mas realcion entre las variables*

## [1] 0.004869459

*#Rows*  
res.ca**$**row

## $coord  
## Dim 1 Dim 2 Dim 3 Dim 4  
## factor\_age [17,31] 0.06028947 0.02489821 0.016824123 -0.017911266  
## factor\_age (31,36] -0.02855663 -0.05990896 0.001849387 -0.011976354  
## factor\_age (36,41] -0.07137954 0.02821861 0.035431074 0.012368421  
## factor\_age (41,49] 0.05590097 -0.01718799 -0.006454706 0.027347518  
## factor\_age (49,81] -0.03117779 0.02922096 -0.044479508 -0.003718468  
##   
## $contrib  
## Dim 1 Dim 2 Dim 3 Dim 4  
## factor\_age [17,31] 30.931887 10.973430 9.3900174 26.201633  
## factor\_age (31,36] 6.621655 60.620433 0.1082646 11.177750  
## factor\_age (36,41] 32.333583 10.511397 31.0565438 9.317238  
## factor\_age (41,49] 22.769835 4.477691 1.1834564 52.300922  
## factor\_age (49,81] 7.343039 13.417048 58.2617178 1.002457  
##   
## $cos2  
## Dim 1 Dim 2 Dim 3 Dim 4  
## factor\_age [17,31] 0.7481200 0.12759233 0.0582576789 0.066029941  
## factor\_age (31,36] 0.1791707 0.78856386 0.0007514654 0.031513932  
## factor\_age (36,41] 0.6979823 0.10908582 0.1719751003 0.020956822  
## factor\_age (41,49] 0.7422793 0.07017444 0.0098965007 0.177649715  
## factor\_age (49,81] 0.2545863 0.22363176 0.5181605763 0.003621366  
##   
## $inertia  
## [1] 0.0010933337 0.0009772756 0.0012249745 0.0008111667 0.0007627082

*#Tenemos las coordenadas, las contribuciones, el cos2 (indica la calidad de la representacion de cada una de las categorias en el eje que toca), inertia*  
  
*#Cols*  
res.ca**$**col

## $coord  
## Dim 1 Dim 2 Dim 3  
## factor\_duration-[1,68] -0.037838157 -0.0722213624 -0.003223726  
## factor\_duration-(68,104] -0.079484346 0.0144258906 -0.032136966  
## factor\_duration-(104,139] 0.033810408 0.0122767778 0.007661304  
## factor\_duration-(139,182] -0.007382708 0.0458581152 -0.028290026  
## factor\_duration-(182,236] 0.020271631 0.0001678354 -0.004295253  
## factor\_duration-(236,329] -0.020308835 -0.0234604264 0.030890479  
## factor\_duration-(329,504] -0.012654910 0.0399387089 0.047274023  
## factor\_duration-(504,2.12e+03] 0.105787697 -0.0158309148 -0.017544387  
## Dim 4  
## factor\_duration-[1,68] -0.007649896  
## factor\_duration-(68,104] -0.009047632  
## factor\_duration-(104,139] 0.013727248  
## factor\_duration-(139,182] 0.001523774  
## factor\_duration-(182,236] 0.031880042  
## factor\_duration-(236,329] 0.009810846  
## factor\_duration-(329,504] -0.019119842  
## factor\_duration-(504,2.12e+03] -0.020319730  
##   
## $contrib  
## Dim 1 Dim 2 Dim 3  
## factor\_duration-[1,68] 6.8855540 5.217867e+01 0.1948379  
## factor\_duration-(68,104] 30.0939744 2.061979e+00 19.1780380  
## factor\_duration-(104,139] 5.3490977 1.467004e+00 1.0706886  
## factor\_duration-(139,182] 0.2583755 2.073648e+01 14.7898845  
## factor\_duration-(182,236] 1.9103338 2.723842e-04 0.3343393  
## factor\_duration-(236,329] 1.9520412 5.418430e+00 17.6054166  
## factor\_duration-(329,504] 0.7567194 1.567789e+01 41.1661239  
## factor\_duration-(504,2.12e+03] 52.7939041 2.459281e+00 5.6606712  
## Dim 4  
## factor\_duration-[1,68] 2.7011103  
## factor\_duration-(68,104] 3.7422993  
## factor\_duration-(104,139] 8.4625062  
## factor\_duration-(139,182] 0.1056364  
## factor\_duration-(182,236] 45.3442221  
## factor\_duration-(236,329] 4.3720403  
## factor\_duration-(329,504] 16.5782111  
## factor\_duration-(504,2.12e+03] 18.6939743  
##   
## $cos2  
## Dim 1 Dim 2 Dim 3  
## factor\_duration-[1,68] 0.2131635 7.765763e-01 0.001547281  
## factor\_duration-(68,104] 0.8268767 2.723722e-02 0.135172172  
## factor\_duration-(104,139] 0.7418214 9.780641e-02 0.038089380  
## factor\_duration-(139,182] 0.0184129 7.104333e-01 0.270369437  
## factor\_duration-(182,236] 0.2842387 1.948377e-05 0.012760955  
## factor\_duration-(236,329] 0.2048606 2.733758e-01 0.473955531  
## factor\_duration-(329,504] 0.0367676 3.662142e-01 0.513088482  
## factor\_duration-(504,2.12e+03] 0.9201376 2.060604e-02 0.025308058  
## Dim 4  
## factor\_duration-[1,68] 0.0087129229  
## factor\_duration-(68,104] 0.0107138956  
## factor\_duration-(104,139] 0.1222828325  
## factor\_duration-(139,182] 0.0007843904  
## factor\_duration-(182,236] 0.7029808969  
## factor\_duration-(236,329] 0.0478080704  
## factor\_duration-(329,504] 0.0839297108  
## factor\_duration-(504,2.12e+03] 0.0339483217  
##   
## $inertia  
## [1] 0.0008541689 0.0009624017 0.0001906772 0.0003710622 0.0001777230  
## [6] 0.0002519696 0.0005442359 0.0015172202

*#Durada mes curta es la que te mes contribucio!*  
  
*#Phi2 = Intensity of the association Chisq/nobservations*  
**sum**(res.ca**$**eig[,1]) *#Total inertia = Phi2*

## [1] 0.004869459

**chisq.test**(**table**(df**$**factor\_age, df**$**factor\_duration))

##   
## Pearson's Chi-squared test  
##   
## data: table(df$factor\_age, df$factor\_duration)  
## X-squared = 24.084, df = 28, p-value = 0.6771

*#24.084/4946 porque son las observaciones*

### Job i Factor\_duration

*# Contingency tables - Complex : solo cuentan con los target discretizados*  
**names**(df)

## [1] "age" "job"   
## [3] "marital" "education"   
## [5] "default" "housing"   
## [7] "loan" "contact"   
## [9] "month" "day\_of\_week"   
## [11] "duration" "campaign"   
## [13] "pdays" "previous"   
## [15] "poutcome" "emp.var.rate"   
## [17] "cons.price.idx" "cons.conf.idx"   
## [19] "euribor3m" "nr.employed"   
## [21] "y" "missings\_indiv"   
## [23] "errors\_indiv" "outliers\_indiv"   
## [25] "season" "factor\_age"   
## [27] "factor\_duration" "factor\_campaign"   
## [29] "factor\_Pdays" "factor\_Previous"   
## [31] "factor\_emp.var.rate" "factor\_cons.price.idx"  
## [33] "factor\_cons.conf.idx" "factor\_euribor3m"   
## [35] "factor\_nr.employed" "CLUSTER"   
## [37] "f.CLUSTER"

*# Target factor\_duration vs job*  
*# Podemos elegir la variable que queramos con la de f\_duration y en este caso hemos elegido job para este ejemplo*  
  
**table**(df**$**job, df**$**factor\_duration)

##   
## factor\_duration-[1,68] factor\_duration-(68,104]  
## Job\_admin. 162 169  
## Job\_blue-collar 131 141  
## Job\_entrepreneur 18 17  
## Job\_housemaid 14 14  
## Job\_management 47 35  
## Job\_retired 18 29  
## Job\_self-employed 20 25  
## Job\_services 75 61  
## Job\_student 8 17  
## Job\_technician 109 96  
## Job\_unemployed 20 14  
## Job\_unknown 7 5  
##   
## factor\_duration-(104,139] factor\_duration-(139,182]  
## Job\_admin. 164 167  
## Job\_blue-collar 133 135  
## Job\_entrepreneur 12 18  
## Job\_housemaid 22 17  
## Job\_management 39 47  
## Job\_retired 24 33  
## Job\_self-employed 23 20  
## Job\_services 52 52  
## Job\_student 10 7  
## Job\_technician 116 105  
## Job\_unemployed 10 16  
## Job\_unknown 7 3  
##   
## factor\_duration-(182,236] factor\_duration-(236,329]  
## Job\_admin. 150 157  
## Job\_blue-collar 137 157  
## Job\_entrepreneur 24 21  
## Job\_housemaid 16 19  
## Job\_management 53 45  
## Job\_retired 21 28  
## Job\_self-employed 12 13  
## Job\_services 54 57  
## Job\_student 13 19  
## Job\_technician 111 85  
## Job\_unemployed 10 15  
## Job\_unknown 7 3  
##   
## factor\_duration-(329,504]  
## Job\_admin. 167  
## Job\_blue-collar 165  
## Job\_entrepreneur 18  
## Job\_housemaid 10  
## Job\_management 43  
## Job\_retired 29  
## Job\_self-employed 17  
## Job\_services 64  
## Job\_student 14  
## Job\_technician 82  
## Job\_unemployed 5  
## Job\_unknown 4  
##   
## factor\_duration-(504,2.12e+03]  
## Job\_admin. 165  
## Job\_blue-collar 145  
## Job\_entrepreneur 32  
## Job\_housemaid 14  
## Job\_management 36  
## Job\_retired 24  
## Job\_self-employed 22  
## Job\_services 58  
## Job\_student 17  
## Job\_technician 80  
## Job\_unemployed 17  
## Job\_unknown 7

*#Le digo que calcule unas probabilidades en la dimension 1, calculo los perfiles por fila que tenemos*  
*#Calculo los perfiles de fila y la suma tendria que dar mas o menos 1 y tenemos que ver si es equivalente al perfil marginal fila*  
**prop.table**(**table**(df**$**job, df**$**factor\_duration), 1) *# Por filas*

##   
## factor\_duration-[1,68] factor\_duration-(68,104]  
## Job\_admin. 0.12451960 0.12990008  
## Job\_blue-collar 0.11451049 0.12325175  
## Job\_entrepreneur 0.11250000 0.10625000  
## Job\_housemaid 0.11111111 0.11111111  
## Job\_management 0.13623188 0.10144928  
## Job\_retired 0.08737864 0.14077670  
## Job\_self-employed 0.13157895 0.16447368  
## Job\_services 0.15856237 0.12896406  
## Job\_student 0.07619048 0.16190476  
## Job\_technician 0.13903061 0.12244898  
## Job\_unemployed 0.18691589 0.13084112  
## Job\_unknown 0.16279070 0.11627907  
##   
## factor\_duration-(104,139] factor\_duration-(139,182]  
## Job\_admin. 0.12605688 0.12836280  
## Job\_blue-collar 0.11625874 0.11800699  
## Job\_entrepreneur 0.07500000 0.11250000  
## Job\_housemaid 0.17460317 0.13492063  
## Job\_management 0.11304348 0.13623188  
## Job\_retired 0.11650485 0.16019417  
## Job\_self-employed 0.15131579 0.13157895  
## Job\_services 0.10993658 0.10993658  
## Job\_student 0.09523810 0.06666667  
## Job\_technician 0.14795918 0.13392857  
## Job\_unemployed 0.09345794 0.14953271  
## Job\_unknown 0.16279070 0.06976744  
##   
## factor\_duration-(182,236] factor\_duration-(236,329]  
## Job\_admin. 0.11529593 0.12067640  
## Job\_blue-collar 0.11975524 0.13723776  
## Job\_entrepreneur 0.15000000 0.13125000  
## Job\_housemaid 0.12698413 0.15079365  
## Job\_management 0.15362319 0.13043478  
## Job\_retired 0.10194175 0.13592233  
## Job\_self-employed 0.07894737 0.08552632  
## Job\_services 0.11416490 0.12050740  
## Job\_student 0.12380952 0.18095238  
## Job\_technician 0.14158163 0.10841837  
## Job\_unemployed 0.09345794 0.14018692  
## Job\_unknown 0.16279070 0.06976744  
##   
## factor\_duration-(329,504]  
## Job\_admin. 0.12836280  
## Job\_blue-collar 0.14423077  
## Job\_entrepreneur 0.11250000  
## Job\_housemaid 0.07936508  
## Job\_management 0.12463768  
## Job\_retired 0.14077670  
## Job\_self-employed 0.11184211  
## Job\_services 0.13530655  
## Job\_student 0.13333333  
## Job\_technician 0.10459184  
## Job\_unemployed 0.04672897  
## Job\_unknown 0.09302326  
##   
## factor\_duration-(504,2.12e+03]  
## Job\_admin. 0.12682552  
## Job\_blue-collar 0.12674825  
## Job\_entrepreneur 0.20000000  
## Job\_housemaid 0.11111111  
## Job\_management 0.10434783  
## Job\_retired 0.11650485  
## Job\_self-employed 0.14473684  
## Job\_services 0.12262156  
## Job\_student 0.16190476  
## Job\_technician 0.10204082  
## Job\_unemployed 0.15887850  
## Job\_unknown 0.16279070

*#Marginal row profile*  
**prop.table**(**table**(df**$**factor\_duration))

##   
## factor\_duration-[1,68] factor\_duration-(68,104]   
## 0.1271735 0.1259604   
## factor\_duration-(104,139] factor\_duration-(139,182]   
## 0.1237364 0.1253538   
## factor\_duration-(182,236] factor\_duration-(236,329]   
## 0.1229276 0.1251516   
## factor\_duration-(329,504] factor\_duration-(504,2.12e+03]   
## 0.1249495 0.1247473

*#Esta proporcion se mantiene en cualquiera de los colectivos mirados anteriormente? Se tiene que hacer la comparacion*  
  
*#Podemos comprobar ahora los perfiles columna*  
*#Column profile*  
**prop.table**(**table**(df**$**job, df**$**factor\_duration), 2) *# dim 2*

##   
## factor\_duration-[1,68] factor\_duration-(68,104]  
## Job\_admin. 0.257551669 0.271268058  
## Job\_blue-collar 0.208267091 0.226324238  
## Job\_entrepreneur 0.028616852 0.027287319  
## Job\_housemaid 0.022257552 0.022471910  
## Job\_management 0.074721781 0.056179775  
## Job\_retired 0.028616852 0.046548957  
## Job\_self-employed 0.031796502 0.040128411  
## Job\_services 0.119236884 0.097913323  
## Job\_student 0.012718601 0.027287319  
## Job\_technician 0.173290938 0.154093098  
## Job\_unemployed 0.031796502 0.022471910  
## Job\_unknown 0.011128776 0.008025682  
##   
## factor\_duration-(104,139] factor\_duration-(139,182]  
## Job\_admin. 0.267973856 0.269354839  
## Job\_blue-collar 0.217320261 0.217741935  
## Job\_entrepreneur 0.019607843 0.029032258  
## Job\_housemaid 0.035947712 0.027419355  
## Job\_management 0.063725490 0.075806452  
## Job\_retired 0.039215686 0.053225806  
## Job\_self-employed 0.037581699 0.032258065  
## Job\_services 0.084967320 0.083870968  
## Job\_student 0.016339869 0.011290323  
## Job\_technician 0.189542484 0.169354839  
## Job\_unemployed 0.016339869 0.025806452  
## Job\_unknown 0.011437908 0.004838710  
##   
## factor\_duration-(182,236] factor\_duration-(236,329]  
## Job\_admin. 0.246710526 0.253634895  
## Job\_blue-collar 0.225328947 0.253634895  
## Job\_entrepreneur 0.039473684 0.033925687  
## Job\_housemaid 0.026315789 0.030694669  
## Job\_management 0.087171053 0.072697900  
## Job\_retired 0.034539474 0.045234249  
## Job\_self-employed 0.019736842 0.021001616  
## Job\_services 0.088815789 0.092084006  
## Job\_student 0.021381579 0.030694669  
## Job\_technician 0.182565789 0.137318255  
## Job\_unemployed 0.016447368 0.024232633  
## Job\_unknown 0.011513158 0.004846527  
##   
## factor\_duration-(329,504]  
## Job\_admin. 0.270226537  
## Job\_blue-collar 0.266990291  
## Job\_entrepreneur 0.029126214  
## Job\_housemaid 0.016181230  
## Job\_management 0.069579288  
## Job\_retired 0.046925566  
## Job\_self-employed 0.027508091  
## Job\_services 0.103559871  
## Job\_student 0.022653722  
## Job\_technician 0.132686084  
## Job\_unemployed 0.008090615  
## Job\_unknown 0.006472492  
##   
## factor\_duration-(504,2.12e+03]  
## Job\_admin. 0.267423015  
## Job\_blue-collar 0.235008104  
## Job\_entrepreneur 0.051863857  
## Job\_housemaid 0.022690438  
## Job\_management 0.058346840  
## Job\_retired 0.038897893  
## Job\_self-employed 0.035656402  
## Job\_services 0.094003241  
## Job\_student 0.027552674  
## Job\_technician 0.129659643  
## Job\_unemployed 0.027552674  
## Job\_unknown 0.011345219

*#Marginal colum profile*  
**prop.table**(**table**(df**$**job))

##   
## Job\_admin. Job\_blue-collar Job\_entrepreneur Job\_housemaid   
## 0.263040841 0.231298019 0.032349373 0.025475131   
## Job\_management Job\_retired Job\_self-employed Job\_services   
## 0.069753336 0.041649818 0.030731905 0.095632835   
## Job\_student Job\_technician Job\_unemployed Job\_unknown   
## 0.021229276 0.158511929 0.021633643 0.008693894

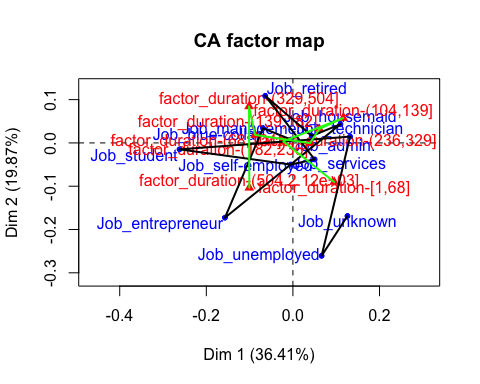
*#El perfil columna de les diferents columnes es pot considerar diferent que el marginal? Evidentment SI*  
  
*# HO: factor\_duration -factor\_age independency*  
**chisq.test**(**table**(df**$**job, df**$**factor\_duration))

##   
## Pearson's Chi-squared test  
##   
## data: table(df$job, df$factor\_duration)  
## X-squared = 95.774, df = 77, p-value = 0.07247

*# Accepto la hipotesi nula porque el pvalor es 0.07247*

En aquesta part de la nostra investigacio podem veure que rebutgem la hipotesi nula perque el pvalor es 0.07247, encara que sigui una mica mes gran que un 5%. Llavors, podem dir que la durada de la trucada podria dependre del treball o a que es dediqui el nostre individu.

*# CA - factor\_duration vs factor\_age*  
res.ca <- **CA**(**table**(df**$**job, df**$**factor\_duration))  
  
*#Link levels in row*  
*#plot.CA(res.ca)*  
**lines**(res.ca**$**row**$**coord[,1], res.ca**$**row**$**coord[,2],lwd=2)  
*#No tenemos que ver nada porque hemos visto que no tienen nada que ver*  
  
*#Link levels in columns*  
**lines**(res.ca**$**col**$**coord[,1], res.ca**$**col**$**coord[,2],lwd=2, col = "green")



Com podem veure a l’hora de l’execució tenim que el Job\_admin, Job\_management i factor\_duration-(68,104] son els que mes destaquen en que no ens aporta cap mena d’informacio ja que es troba mes a prop del centre de gravetat, A partir de les taules de contingencia i els seus diferents perfils intentem observar si hi pot haver alguna relacio de dependencia entre els dos factors.

## Eigenvalues and dominant axes analysis

En aquest subapartat realitzarem un estudi dels valors propis i dels eixos dominants per tal de determinar quantes dimensions tindrem en compte.

res.ca**$**eig

## eigenvalue percentage of variance cumulative percentage of variance  
## dim 1 0.007050333 36.409534 36.40953  
## dim 2 0.003847258 19.868124 56.27766  
## dim 3 0.003249026 16.778713 73.05637  
## dim 4 0.002161419 11.162064 84.21844  
## dim 5 0.001718252 8.873445 93.09188  
## dim 6 0.001041702 5.379590 98.47147  
## dim 7 0.000295984 1.528529 100.00000

*#No es extraño que los eigenvalues sean pequeños, cojemos tantas dimensiones como las que tengan un valor propio > mitjana de este valor*  
**mean**(res.ca**$**eig[,1]) *#Mean eigenvalue*

## [1] 0.002766282

*#KAISER: take as many as dimensions as eigenvalue > mean eig*  
**sum**(res.ca**$**eig[,1]) *#Total inertia, contra mas grande hay mas relacion entre las variables*

## [1] 0.01936397

*#Rows*  
res.ca**$**row

## $coord  
## Dim 1 Dim 2 Dim 3 Dim 4  
## Job\_admin. -0.004186212 0.007701902 -0.028673409 -0.002854957  
## Job\_blue-collar -0.069096677 0.033980063 0.012824475 -0.012061634  
## Job\_entrepreneur -0.157069320 -0.172777310 0.095319234 0.073797366  
## Job\_housemaid 0.109639754 0.043116843 0.011457101 0.159644423  
## Job\_management 0.043596504 0.016129874 0.118496133 -0.013785385  
## Job\_retired -0.064052069 0.108786121 -0.065965837 0.053120289  
## Job\_self-employed 0.050005969 -0.037996065 -0.208308255 0.003606401  
## Job\_services -0.004690390 -0.049982823 -0.007662952 -0.096663163  
## Job\_student -0.261854900 -0.014852099 0.006879286 0.082853328  
## Job\_technician 0.132101028 0.014640019 0.023566514 0.012316894  
## Job\_unemployed 0.066308261 -0.260754891 -0.065825305 0.068094425  
## Job\_unknown 0.125895275 -0.168760946 0.041355718 -0.024777522  
## Dim 5  
## Job\_admin. 0.000286884  
## Job\_blue-collar 0.001586189  
## Job\_entrepreneur 0.021690042  
## Job\_housemaid 0.017201430  
## Job\_management -0.044076180  
## Job\_retired -0.077827955  
## Job\_self-employed 0.049353299  
## Job\_services -0.007967764  
## Job\_student 0.094554874  
## Job\_technician 0.019815606  
## Job\_unemployed -0.149509980  
## Job\_unknown 0.237538144  
##   
## $contrib  
## Dim 1 Dim 2 Dim 3 Dim 4  
## Job\_admin. 0.06538166 0.4055718 6.65623555 0.09919351  
## Job\_blue-collar 15.66306072 6.9417492 1.17084112 1.55684451  
## Job\_entrepreneur 11.31980610 25.1008208 9.04635891 8.15095688  
## Job\_housemaid 4.34353103 1.2310026 0.10292308 30.03896341  
## Job\_management 1.88043659 0.4717106 30.14534050 0.61328707  
## Job\_retired 2.42364957 12.8117579 5.57825188 5.43744611  
## Job\_self-employed 1.08999261 1.1532285 41.04396339 0.01849262  
## Job\_services 0.02984114 6.2100805 0.17284073 41.34186280  
## Job\_student 20.64652601 0.1217193 0.03092208 6.74242469  
## Job\_technician 39.23419417 0.8830675 2.70956455 1.11256497  
## Job\_unemployed 1.34913477 38.2334270 2.88510940 4.64102364  
## Job\_unknown 1.95444563 6.4358644 0.45764881 0.24693979  
## Dim 5  
## Job\_admin. 0.001259937  
## Job\_blue-collar 0.033868409  
## Job\_entrepreneur 0.885727032  
## Job\_housemaid 0.438691050  
## Job\_management 7.886532880  
## Job\_retired 14.682417960  
## Job\_self-employed 4.356473625  
## Job\_services 0.353340345  
## Job\_student 11.046285370  
## Job\_technician 3.622345822  
## Job\_unemployed 28.143834431  
## Job\_unknown 28.549223139  
##   
## $cos2  
## Dim 1 Dim 2 Dim 3 Dim 4  
## Job\_admin. 0.016594367 0.056171271 0.7785329044 0.0077182245  
## Job\_blue-collar 0.739280414 0.178790006 0.0254667819 0.0225271988  
## Job\_entrepreneur 0.315537130 0.381804603 0.1162060750 0.0696545455  
## Job\_housemaid 0.245681834 0.037995385 0.0026827883 0.5208881086  
## Job\_management 0.102696948 0.014057762 0.7586868817 0.0102681461  
## Job\_retired 0.130823433 0.377368975 0.1387577892 0.0899787729  
## Job\_self-employed 0.048798014 0.028173169 0.8467816202 0.0002538087  
## Job\_services 0.001699359 0.192978417 0.0045358573 0.7217539703  
## Job\_student 0.677204963 0.002178584 0.0004673965 0.0677982733  
## Job\_technician 0.916061394 0.011251113 0.0291543176 0.0079636948  
## Job\_unemployed 0.040898751 0.632469657 0.0403051494 0.0431318294  
## Job\_unknown 0.151106062 0.271523191 0.0163055017 0.0058530031  
## Dim 5  
## Job\_admin. 7.793471e-05  
## Job\_blue-collar 3.895871e-04  
## Job\_entrepreneur 6.017118e-03  
## Job\_housemaid 6.047363e-03  
## Job\_management 1.049693e-01  
## Job\_retired 1.931481e-01  
## Job\_self-employed 4.753252e-02  
## Job\_services 4.903883e-03  
## Job\_student 8.830119e-02  
## Job\_technician 2.061232e-02  
## Job\_unemployed 2.079290e-01  
## Job\_unknown 5.379349e-01  
##   
## $inertia  
## [1] 0.0002777825 0.0014937470 0.0025292871 0.0012464633 0.0012909540  
## [6] 0.0013061525 0.0015748203 0.0012380548 0.0021494951 0.0030196024  
## [11] 0.0023257064 0.0009119086

*#Tenemos las coordenadas, las contribuciones, el cos2 (indica la calidad de la representacion de cada una de las categorias en el eje que toca), inertia*  
  
*#Cols*  
res.ca**$**col

## $coord  
## Dim 1 Dim 2 Dim 3  
## factor\_duration-[1,68] 0.09531274 -0.0902001831 0.004152345  
## factor\_duration-(68,104] -0.02226995 0.0035790782 -0.085408752  
## factor\_duration-(104,139] 0.11616899 0.0575727078 -0.037797854  
## factor\_duration-(139,182] 0.06103176 0.0322551944 -0.020524250  
## factor\_duration-(182,236] 0.03929675 -0.0004978809 0.122488603  
## factor\_duration-(236,329] -0.08742549 0.0184521454 0.038039160  
## factor\_duration-(329,504] -0.10141238 0.0842994678 0.004928502  
## factor\_duration-(504,2.12e+03] -0.10067398 -0.1036350174 -0.023679053  
## Dim 4 Dim 5  
## factor\_duration-[1,68] -0.0736844595 -0.023341450  
## factor\_duration-(68,104] -0.0007433834 0.016809274  
## factor\_duration-(104,139] 0.0310640066 0.056862266  
## factor\_duration-(139,182] 0.0338717427 -0.080127776  
## factor\_duration-(182,236] 0.0143242316 0.033492169  
## factor\_duration-(236,329] 0.0421650872 -0.036944879  
## factor\_duration-(329,504] -0.0803395322 0.007811121  
## factor\_duration-(504,2.12e+03] 0.0350721387 0.027175814  
##   
## $contrib  
## Dim 1 Dim 2 Dim 3  
## factor\_duration-[1,68] 16.386601 2.689429e+01 0.06748859  
## factor\_duration-(68,104] 0.886059 4.193966e-02 28.28039945  
## factor\_duration-(104,139] 23.684714 1.066054e+01 5.44099668  
## factor\_duration-(139,182] 6.622772 3.389889e+00 1.62524571  
## factor\_duration-(182,236] 2.692484 7.920435e-04 56.76592066  
## factor\_duration-(236,329] 13.567602 1.107590e+00 5.57372088  
## factor\_duration-(329,504] 18.226645 2.307983e+01 0.09341381  
## factor\_duration-(504,2.12e+03] 17.933124 3.482513e+01 2.15281423  
## Dim 4 Dim 5  
## factor\_duration-[1,68] 31.94547451 4.0324168  
## factor\_duration-(68,104] 0.00322048 2.0713099  
## factor\_duration-(104,139] 5.52424913 23.2840689  
## factor\_duration-(139,182] 6.65386009 46.8400112  
## factor\_duration-(182,236] 1.16695240 8.0250777  
## factor\_duration-(236,329] 10.29445944 9.9416456  
## factor\_duration-(329,504] 37.31246714 0.4436845  
## factor\_duration-(504,2.12e+03] 7.09931680 5.3617852  
##   
## $cos2  
## Dim 1 Dim 2 Dim 3  
## factor\_duration-[1,68] 0.38193087 3.420563e-01 0.0007248861  
## factor\_duration-(68,104] 0.05170872 1.335574e-03 0.7605543115  
## factor\_duration-(104,139] 0.58668609 1.440982e-01 0.0621097313  
## factor\_duration-(139,182] 0.25497352 7.121682e-02 0.0288348615  
## factor\_duration-(182,236] 0.08321936 1.335863e-05 0.8085417008  
## factor\_duration-(236,329] 0.46736782 2.081979e-02 0.0884798704  
## factor\_duration-(329,504] 0.42099659 2.909017e-01 0.0009943206  
## factor\_duration-(504,2.12e+03] 0.41139957 4.359557e-01 0.0227592517  
## Dim 4 Dim 5  
## factor\_duration-[1,68] 2.282625e-01 0.022905431  
## factor\_duration-(68,104] 5.761707e-05 0.029459366  
## factor\_duration-(104,139] 4.195080e-02 0.140563863  
## factor\_duration-(139,182] 7.853412e-02 0.439490460  
## factor\_duration-(182,236] 1.105742e-02 0.060450181  
## factor\_duration-(236,329] 1.087148e-01 0.083462456  
## factor\_duration-(329,504] 2.642136e-01 0.002497602  
## factor\_duration-(504,2.12e+03] 4.992911e-02 0.029977434  
##   
## $inertia  
## [1] 0.003024919 0.001208115 0.002846243 0.001831278 0.002281069 0.002046699  
## [7] 0.003052374 0.003073277

*#Durada mes curta es la que te mes contribucio!*  
  
*#Phi2 = Intensity of the association Chisq/nobservations*  
**sum**(res.ca**$**eig[,1]) *#Total inertia = Phi2*

## [1] 0.01936397

**chisq.test**(**table**(df**$**job, df**$**factor\_duration))

##   
## Pearson's Chi-squared test  
##   
## data: table(df$job, df$factor\_duration)  
## X-squared = 95.774, df = 77, p-value = 0.07247

*#95.774/4946 porque son las observaciones*

————– DELIVERABLE 3 —————-

## Model construction only with numeric explanatory variables

### Multivariant Data Analysis

Ara el que farem serà analitzar quines són les variables numèriques més relacionades amb el nostre target duration, per tal de decidir quines d’aquestes utilitzarem en la construcció dels diferents models fins trobar l’òptim.

*#En vars\_model també tenim la variable "duration" perquè és necessari per poder veure les més relacionades amb aquesta*  
vars\_model<-**names**(df)[**c**(1,11**:**14,16**:**20)]; vars\_model

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

*# condes(df[,vars\_model],which(vars\_model == "duration"))*

A partir d’executar la comanda “condes” podem veure que les variables més relacionades són previous, nr.employed, campaign i pdays, tot i que la correlació que presenten és molt baixa i poc significativa. Tot i així les podem considerar com a candidates a formar part de la construcció del nostre model.

### Model Construction

A partir de tot l’anàlisi realitzat fins ara, començarem la construcció dels models, partint d’un model més complexe de totes les variables numèriques. Realitzarem diferents anàlisis per a cada model fins a trobar el model més adient o òptim a la nostra situació o joc de dades.

## Initial modelling

**names**(df)

## [1] "age" "job"   
## [3] "marital" "education"   
## [5] "default" "housing"   
## [7] "loan" "contact"   
## [9] "month" "day\_of\_week"   
## [11] "duration" "campaign"   
## [13] "pdays" "previous"   
## [15] "poutcome" "emp.var.rate"   
## [17] "cons.price.idx" "cons.conf.idx"   
## [19] "euribor3m" "nr.employed"   
## [21] "y" "missings\_indiv"   
## [23] "errors\_indiv" "outliers\_indiv"   
## [25] "season" "factor\_age"   
## [27] "factor\_duration" "factor\_campaign"   
## [29] "factor\_Pdays" "factor\_Previous"   
## [31] "factor\_emp.var.rate" "factor\_cons.price.idx"  
## [33] "factor\_cons.conf.idx" "factor\_euribor3m"   
## [35] "factor\_nr.employed" "CLUSTER"   
## [37] "f.CLUSTER"

*#Las variables socioeconomicas estan relacionadas entre ellas, pero no tienen nada que ver con el target*  
*#vars\_exp<-names(df)[c(1,12:14,16:20)]; vars\_exp*  
  
vars\_conaux *#numèriques = vars\_exp*

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

*#vars\_con\_aux2 #numeriques (sense age) que es la que utilitzem!*  
**condes**(df,11)

## $quanti  
## correlation p.value  
## previous 0.02859224 4.435374e-02  
## errors\_indiv -0.03476735 1.447588e-02  
## nr.employed -0.03619203 1.091224e-02  
## CLUSTER -0.04004368 4.853468e-03  
## campaign -0.04179341 3.284450e-03  
## pdays -0.06147234 1.516945e-05  
## missings\_indiv -0.07328498 2.474678e-07  
##   
## $quali  
## R2 p.value  
## factor\_duration 0.8271873066 0.000000e+00  
## y 0.1863696068 9.891372e-224  
## factor\_Pdays 0.0051824450 4.017238e-07  
## poutcome 0.0041874670 3.132625e-05  
## f.CLUSTER 0.0061553592 3.146859e-05  
## month 0.0073478185 3.327154e-05  
## factor\_cons.price.idx 0.0039803615 5.696640e-04  
## factor\_Previous 0.0019228074 2.038492e-03  
## day\_of\_week 0.0029955473 5.075577e-03  
## factor\_cons.conf.idx 0.0026002247 1.194404e-02  
## contact 0.0011105265 1.909343e-02  
## default 0.0009897216 2.693284e-02  
## factor\_campaign 0.0013152237 3.866909e-02  
##   
## $category  
## Estimate p.value  
## factor\_duration-(504,2.12e+03] 547.162252 0.000000e+00  
## Y\_yes 169.675531 9.891372e-224  
## factor\_duration-(329,504] 138.462468 3.985182e-48  
## factor\_Pdays-[0,15] 49.355073 4.017238e-07  
## CLUSTER-4 82.017790 5.318613e-06  
## Poutcome\_success 62.641078 7.933875e-06  
## factor\_cons.price.idx-(93.4,93.9] 27.117765 2.010384e-04  
## Month\_jul 12.946601 2.986551e-04  
## factor\_Previous-(1,5] 34.966136 2.038492e-03  
## Contact\_cellular 8.850090 1.909343e-02  
## Default\_no 9.913335 2.693284e-02  
## Month\_dec 104.090396 2.868142e-02  
## Day\_of\_week\_tue 14.917687 4.872420e-02  
## Education\_illiterate 178.585152 4.932974e-02  
## CLUSTER-7 -37.598946 4.049876e-02  
## Education\_university.degree -38.308971 3.857651e-02  
## factor\_cons.conf.idx-(-36.4,-29.8] -13.574401 3.768483e-02  
## CLUSTER-5 -20.182210 3.375761e-02  
## factor\_cons.conf.idx-(-42,-40.3] -17.926886 2.695593e-02  
## Default\_unknown -9.913335 2.693284e-02  
## Contact\_telephone -8.850090 1.909343e-02  
## Month\_jun -37.404273 1.736971e-02  
## factor\_campaign-(3,14] -16.741883 1.148865e-02  
## Job\_technician -25.341033 1.106827e-02  
## Day\_of\_week\_mon -19.239047 7.577039e-03  
## Month\_aug -39.248662 5.073298e-03  
## factor\_cons.price.idx-(93,93.4] -19.809889 2.312144e-03  
## factor\_Previous-[0,1] -34.966136 2.038492e-03  
## factor\_Pdays-(15,17] -49.355073 4.017238e-07  
## factor\_duration-(182,236] -56.414720 8.764699e-09  
## factor\_duration-(139,182] -103.067426 8.297196e-27  
## factor\_duration-(104,139] -141.910732 3.245807e-49  
## factor\_duration-(68,104] -177.221056 2.195363e-78  
## factor\_duration-[1,68] -222.636796 8.250905e-127  
## Y\_no -169.675531 9.891372e-224

m1<-**lm**(duration**~**previous**+**euribor3m**+**campaign**+**pdays**+**nr.employed,data=df)  
*#summary(m1)*  
**Anova**(m1)

## Anova Table (Type II tests)  
##   
## Response: duration  
## Sum Sq Df F value Pr(>F)   
## previous 69540 1 1.0663 0.3018273   
## euribor3m 393980 1 6.0413 0.0140094 \*   
## campaign 441217 1 6.7656 0.0093209 \*\*   
## pdays 726966 1 11.1473 0.0008478 \*\*\*  
## nr.employed 478090 1 7.3310 0.0068008 \*\*   
## Residuals 322161286 4940   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*# Inferential criteria o Bayesian info criteria*  
*# Remove non significant variables*  
  
*#Les variables que sobran son las que tienen un pvalor por encima del 5%*  
*#Aqui se ponen las que tengan un p valor menor que 5*

Veiem que aquest model i segurament tots els que realitzarem amb el target numèric tenen una explicabilitat molt baixa (menys del 0.005 del % de les dades),i per tant serà díficil obtenir dades rellevants. Tot i així procedirem a fer un procés metadològic de “Modeling” del target numèric.

Ara el que farem és fer un segon model i només posaré les variables que tenen un p-valor per sota d’un 5%, llavors em queda el mateix model que m1 però sense les variables previous.

m2<-**lm**(duration**~**euribor3m**+**campaign**+**pdays**+**nr.employed,data=df)  
**summary**(m2)

##   
## Call:  
## lm(formula = duration ~ euribor3m + campaign + pdays + nr.employed,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -305.24 -158.09 -83.76 65.34 1858.59   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2419.0524 788.7005 3.067 0.00217 \*\*  
## euribor3m 15.9367 6.5155 2.446 0.01448 \*   
## campaign -4.7524 1.8455 -2.575 0.01005 \*   
## pdays -6.2056 1.9320 -3.212 0.00133 \*\*  
## nr.employed -0.4075 0.1584 -2.573 0.01012 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 255.4 on 4941 degrees of freedom  
## Multiple R-squared: 0.006577, Adjusted R-squared: 0.005773   
## F-statistic: 8.178 on 4 and 4941 DF, p-value: 1.434e-06

**Anova**(m2)

## Anova Table (Type II tests)  
##   
## Response: duration  
## Sum Sq Df F value Pr(>F)   
## euribor3m 390168 1 5.9827 0.014481 \*   
## campaign 432446 1 6.6310 0.010051 \*   
## pdays 672831 1 10.3170 0.001327 \*\*  
## nr.employed 431626 1 6.6184 0.010122 \*   
## Residuals 322230826 4941   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*#I ara el que farem serà un vif (variance inflation factor) per veure les variables explicatives del model que estan correlacionades*  
**vif**(m2)

## euribor3m campaign pdays nr.employed   
## 9.620996 1.016309 1.168931 10.105172

Ara en el nostre tercer model el que farem és que quan executem el vif veiem que tenim les variables nr.employed i euribor3m amb un vif > 3, llavors això no és vàlid, perquè inflarà la variança de la nostra mostra. Llavors primer el que fem és eliminar nr.employed y després en el model número 4 eliminarem euribor3m també per veure quin és el que té una millor explicabilitat.

m3<-**lm**(duration**~**campaign**+**pdays**+**euribor3m,data=df)  
**summary**(m3)

##   
## Call:  
## lm(formula = duration ~ campaign + pdays + euribor3m, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -319.98 -159.03 -83.08 67.50 1854.92   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 391.323 28.316 13.820 < 2e-16 \*\*\*  
## campaign -4.967 1.845 -2.692 0.00712 \*\*   
## pdays -7.505 1.866 -4.023 5.84e-05 \*\*\*  
## euribor3m 0.162 2.204 0.074 0.94141   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 255.5 on 4942 degrees of freedom  
## Multiple R-squared: 0.005247, Adjusted R-squared: 0.004643   
## F-statistic: 8.688 on 3 and 4942 DF, p-value: 9.541e-06

m4<-**lm**(duration**~**campaign**+**pdays**+**nr.employed,data=df)  
**summary**(m4)

##   
## Call:  
## lm(formula = duration ~ campaign + pdays + nr.employed, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -316.92 -158.62 -83.03 66.73 1857.76   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 604.23452 267.59919 2.258 0.023990 \*   
## campaign -4.78623 1.84642 -2.592 0.009565 \*\*   
## pdays -6.93604 1.90973 -3.632 0.000284 \*\*\*  
## nr.employed -0.04289 0.05359 -0.800 0.423582   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 255.5 on 4942 degrees of freedom  
## Multiple R-squared: 0.005374, Adjusted R-squared: 0.004771   
## F-statistic: 8.901 on 3 and 4942 DF, p-value: 7.024e-06

m5<-**step**(m1, k=**log**(**nrow**(df)))

## Start: AIC=54873.63  
## duration ~ previous + euribor3m + campaign + pdays + nr.employed  
##   
## Df Sum of Sq RSS AIC  
## - previous 1 69540 322230826 54866  
## - euribor3m 1 393980 322555266 54871  
## - campaign 1 441217 322602503 54872  
## - nr.employed 1 478090 322639376 54872  
## <none> 322161286 54874  
## - pdays 1 726966 322888252 54876  
##   
## Step: AIC=54866.19  
## duration ~ euribor3m + campaign + pdays + nr.employed  
##   
## Df Sum of Sq RSS AIC  
## - euribor3m 1 390168 322620995 54864  
## - nr.employed 1 431626 322662452 54864  
## - campaign 1 432446 322663273 54864  
## <none> 322230826 54866  
## - pdays 1 672831 322903657 54868  
##   
## Step: AIC=54863.67  
## duration ~ campaign + pdays + nr.employed  
##   
## Df Sum of Sq RSS AIC  
## - nr.employed 1 41810 322662805 54856  
## - campaign 1 438650 323059645 54862  
## <none> 322620995 54864  
## - pdays 1 861130 323482124 54868  
##   
## Step: AIC=54855.81  
## duration ~ campaign + pdays  
##   
## Df Sum of Sq RSS AIC  
## - campaign 1 475707 323138512 54855  
## <none> 322662805 54856  
## - pdays 1 1134867 323797672 54865  
##   
## Step: AIC=54854.59  
## duration ~ pdays  
##   
## Df Sum of Sq RSS AIC  
## <none> 323138512 54855  
## - pdays 1 1225723 324364235 54865

*#vif(m5) # Dóna error perquè tenim menys de dos variables!*  
  
m6<-**lm**(duration**~**campaign**+**pdays,data=df)  
**summary**(m6)

##   
## Call:  
## lm(formula = duration ~ campaign + pdays, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -319.93 -158.86 -82.90 67.12 1855.14   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 391.279 28.307 13.82 < 2e-16 \*\*\*  
## campaign -4.953 1.835 -2.70 0.00697 \*\*   
## pdays -7.467 1.791 -4.17 3.1e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 255.5 on 4943 degrees of freedom  
## Multiple R-squared: 0.005245, Adjusted R-squared: 0.004843   
## F-statistic: 13.03 on 2 and 4943 DF, p-value: 2.264e-06

**vif**(m6)

## campaign pdays   
## 1.003368 1.003368

Amb aquesta sortida el que podem comprobar és que les variables que són més significatives són campaign i pdays, però si fem el step veiem que la millor és pdays, però un model amb només una variable és molt poc i no explicaria el suficient, llavors agafem campaign i pdays.

Quan executem el vif en el nostre model definitiu veiem que les dos variables que tenim tenen un vif < 3, llavors això vol dir que el nostre model és correcte i que anem en bona direcció.

## Transforming variables

Ara el que farem serà una transformació de les nostres variables per veure si podem explicar més en el nostre model.

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

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## previous 0.1 1 0.0688 0.7931   
## campaign 93.7 1 108.1953 < 2e-16 \*\*\*  
## nr.employed 0.1 1 0.1424 0.7060   
## pdays 17.0 1 19.5908 9.8e-06 \*\*\*  
## Residuals 4277.7 4941   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

m8<-**lm** (**log**(duration)**~**campaign**+**pdays,data=df)  
**summary**(m8)

##   
## Call:  
## lm(formula = log(duration) ~ campaign + pdays, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2586 -0.5401 -0.0011 0.6236 2.7295   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.88173 0.10307 57.066 < 2e-16 \*\*\*  
## campaign -0.06979 0.00668 -10.447 < 2e-16 \*\*\*  
## pdays -0.03458 0.00652 -5.303 1.19e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9303 on 4943 degrees of freedom  
## Multiple R-squared: 0.02834, Adjusted R-squared: 0.02795   
## F-statistic: 72.09 on 2 and 4943 DF, p-value: < 2.2e-16

*#Polinomic regression*  
m9 <- **lm**(**log**(duration)**~poly**(campaign,2)**+poly**(pdays,2), data=df)  
**summary**(m9)

##   
## Call:  
## lm(formula = log(duration) ~ poly(campaign, 2) + poly(pdays,   
## 2), data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2184 -0.5456 0.0019 0.6134 2.8100   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.17462 0.01319 392.362 < 2e-16 \*\*\*  
## poly(campaign, 2)1 -9.69878 0.92913 -10.439 < 2e-16 \*\*\*  
## poly(campaign, 2)2 -4.30252 0.92758 -4.638 3.6e-06 \*\*\*  
## poly(pdays, 2)1 -4.99650 0.92914 -5.378 7.9e-08 \*\*\*  
## poly(pdays, 2)2 -2.94158 0.92757 -3.171 0.00153 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9275 on 4941 degrees of freedom  
## Multiple R-squared: 0.03452, Adjusted R-squared: 0.03374   
## F-statistic: 44.16 on 4 and 4941 DF, p-value: < 2.2e-16

**Anova**(m9)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## poly(campaign, 2) 112.3 2 65.273 < 2.2e-16 \*\*\*  
## poly(pdays, 2) 33.5 2 19.477 3.755e-09 \*\*\*  
## Residuals 4250.6 4941   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*# marginalModelPlots(m9)*

Com podem observar les nostres variables més significatives del nostre model són campaign y pdays, llavors com a conclusió el nostre model será m8 que té una explicabilitat d’un 2,8%. Però quan fem la transformació logarítmica veiem que té una mica més d’explicabilitat el nostre model, perquè el Multiple R-squared és major, veiem que tenim una explicabilitat d’un 3,4%.

La diferència és sumament petita i fent les diferents execucions que venen a continuació hem vist que no hi ha cap tipus de diferència, com era correcte agafar un d’aquests dos models vam optar per agafar el m8 en comptes del m9. Si que és veritat que hauriem de treballar amb la cuadràtica, però vam seguir el nostre estudi sense ell, ja que no hi havia molta diferència.

CONCLUSIÓ: El Multiple R-squared (variabilitat de les dades) és molt petit i això vol dir que el nostre target és complicat d’interpretar, és a dir, no podem explicar el nostre target (duration, en aquest cas) amb les variables que tenim.

## Adding factors as explanatory variables

Ara el que farem és afegir variables factors com a variables explicatives, llavors hem de trobar les que poden ser més significatives i ara a continuació farem aquest estudi.

vars\_dis2<-**names**(df)[**c**(2**:**10,15,25,26**:**35)];vars\_dis2

## [1] "job" "marital"   
## [3] "education" "default"   
## [5] "housing" "loan"   
## [7] "contact" "month"   
## [9] "day\_of\_week" "poutcome"   
## [11] "season" "factor\_age"   
## [13] "factor\_duration" "factor\_campaign"   
## [15] "factor\_Pdays" "factor\_Previous"   
## [17] "factor\_emp.var.rate" "factor\_cons.price.idx"  
## [19] "factor\_cons.conf.idx" "factor\_euribor3m"   
## [21] "factor\_nr.employed"

*# Agafem el nostre millor model que tenim fins ara*  
m10<-**step**(m8,k=**log**(**nrow**(df)))

## Start: AIC=-692.34  
## log(duration) ~ campaign + pdays  
##   
## Df Sum of Sq RSS AIC  
## <none> 4277.8 -692.34  
## - pdays 1 24.342 4302.1 -672.78  
## - campaign 1 94.458 4372.3 -592.82

*# maux4<-step(m9,k=log(nrow(df))) Con el modelo que usa poly!*  
  
**condes**(df[,**c**("duration",vars\_dis2)],1,proba = 0.01)

## $quali  
## R2 p.value  
## factor\_duration 0.827187307 0.000000e+00  
## factor\_Pdays 0.005182445 4.017238e-07  
## poutcome 0.004187467 3.132625e-05  
## month 0.007347818 3.327154e-05  
## factor\_cons.price.idx 0.003980361 5.696640e-04  
## factor\_Previous 0.001922807 2.038492e-03  
## day\_of\_week 0.002995547 5.075577e-03  
##   
## $category  
## Estimate p.value  
## factor\_duration-(504,2.12e+03] 547.16225 0.000000e+00  
## factor\_duration-(329,504] 138.46247 3.985182e-48  
## factor\_Pdays-[0,15] 49.35507 4.017238e-07  
## Poutcome\_success 62.64108 7.933875e-06  
## factor\_cons.price.idx-(93.4,93.9] 27.11777 2.010384e-04  
## Month\_jul 12.94660 2.986551e-04  
## factor\_Previous-(1,5] 34.96614 2.038492e-03  
## Day\_of\_week\_mon -19.23905 7.577039e-03  
## Month\_aug -39.24866 5.073298e-03  
## factor\_cons.price.idx-(93,93.4] -19.80989 2.312144e-03  
## factor\_Previous-[0,1] -34.96614 2.038492e-03  
## factor\_Pdays-(15,17] -49.35507 4.017238e-07  
## factor\_duration-(182,236] -56.41472 8.764699e-09  
## factor\_duration-(139,182] -103.06743 8.297196e-27  
## factor\_duration-(104,139] -141.91073 3.245807e-49  
## factor\_duration-(68,104] -177.22106 2.195363e-78  
## factor\_duration-[1,68] -222.63680 8.250905e-127

Després de l’execució anterior el que hem vist són les variables més correlacionades amb el nostre model que són aquelles que tenen un p-valor << 0.01. Aquestes variables són: factor\_Pdays+ poutcome+month+factor\_cons.price.idx+ factor\_Previous+day\_of\_week

Llavors ara estudiarem el cas, és a dir, al nostre model li afegim aquests factors.

*#Avoid numeric and factors simultaneously for the same concept*  
m11<-**lm**(**log**(duration)**~**campaign**+**pdays**+**poutcome**+**month**+**factor\_cons.price.idx**+** factor\_Previous**+**day\_of\_week,data = df)  
**summary**(m11) *#Take a look to NA estimates*

##   
## Call:  
## lm(formula = log(duration) ~ campaign + pdays + poutcome + month +   
## factor\_cons.price.idx + factor\_Previous + day\_of\_week, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1845 -0.5552 -0.0061 0.6031 2.6685   
##   
## Coefficients:  
## Estimate  
## (Intercept) 5.406988  
## campaign -0.069743  
## pdays 0.002901  
## poutcomePoutcome\_nonexistent 0.009651  
## poutcomePoutcome\_success 0.378327  
## monthMonth\_aug -0.212340  
## monthMonth\_dec 0.141391  
## monthMonth\_jul -0.187828  
## monthMonth\_jun -0.351201  
## monthMonth\_mar -0.185593  
## monthMonth\_may -0.345035  
## monthMonth\_nov -0.269914  
## monthMonth\_oct -0.228642  
## monthMonth\_sep -0.352472  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] -0.110456  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] 0.088951  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] 0.219283  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] 0.002831  
## factor\_Previousfactor\_Previous-(1,5] 0.188940  
## day\_of\_weekDay\_of\_week\_mon 0.060226  
## day\_of\_weekDay\_of\_week\_thu 0.085789  
## day\_of\_weekDay\_of\_week\_tue 0.211005  
## day\_of\_weekDay\_of\_week\_wed 0.150490  
## Std. Error t value  
## (Intercept) 0.306736 17.627  
## campaign 0.006710 -10.393  
## pdays 0.018666 0.155  
## poutcomePoutcome\_nonexistent 0.049726 0.194  
## poutcomePoutcome\_success 0.207580 1.823  
## monthMonth\_aug 0.066472 -3.194  
## monthMonth\_dec 0.214603 0.659  
## monthMonth\_jul 0.114380 -1.642  
## monthMonth\_jun 0.105853 -3.318  
## monthMonth\_mar 0.130310 -1.424  
## monthMonth\_may 0.092767 -3.719  
## monthMonth\_nov 0.069135 -3.904  
## monthMonth\_oct 0.130712 -1.749  
## monthMonth\_sep 0.140611 -2.507  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] 0.070455 -1.568  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] 0.096588 0.921  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] 0.049133 4.463  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] 0.074668 0.038  
## factor\_Previousfactor\_Previous-(1,5] 0.098283 1.922  
## day\_of\_weekDay\_of\_week\_mon 0.041383 1.455  
## day\_of\_weekDay\_of\_week\_thu 0.041253 2.080  
## day\_of\_weekDay\_of\_week\_tue 0.042899 4.919  
## day\_of\_weekDay\_of\_week\_wed 0.041820 3.598  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## campaign < 2e-16 \*\*\*  
## pdays 0.876480   
## poutcomePoutcome\_nonexistent 0.846126   
## poutcomePoutcome\_success 0.068431 .   
## monthMonth\_aug 0.001410 \*\*   
## monthMonth\_dec 0.510022   
## monthMonth\_jul 0.100625   
## monthMonth\_jun 0.000914 \*\*\*  
## monthMonth\_mar 0.154438   
## monthMonth\_may 0.000202 \*\*\*  
## monthMonth\_nov 9.58e-05 \*\*\*  
## monthMonth\_oct 0.080316 .   
## monthMonth\_sep 0.012218 \*   
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] 0.117007   
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] 0.357133   
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] 8.26e-06 \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] 0.969754   
## factor\_Previousfactor\_Previous-(1,5] 0.054612 .   
## day\_of\_weekDay\_of\_week\_mon 0.145640   
## day\_of\_weekDay\_of\_week\_thu 0.037615 \*   
## day\_of\_weekDay\_of\_week\_tue 9.00e-07 \*\*\*  
## day\_of\_weekDay\_of\_week\_wed 0.000323 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9212 on 4923 degrees of freedom  
## Multiple R-squared: 0.05104, Adjusted R-squared: 0.0468   
## F-statistic: 12.03 on 22 and 4923 DF, p-value: < 2.2e-16

*#Com no ha sortit cap NA, de moment no tenim cap variable problemàtica!*  
  
  
**Anova** (m11)

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## campaign 91.7 1 108.0209 < 2.2e-16 \*\*\*  
## pdays 0.0 1 0.0242 0.876480   
## poutcome 2.8 2 1.6624 0.189794   
## month 22.6 9 2.9525 0.001679 \*\*   
## factor\_cons.price.idx 20.6 4 6.0598 7.335e-05 \*\*\*  
## factor\_Previous 3.1 1 3.6957 0.054612 .   
## day\_of\_week 24.8 4 7.3018 7.367e-06 \*\*\*  
## Residuals 4177.9 4923   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*#Para limpiar! Efectes nets*  
*#Poutcome és problemàtica perquè es 0.1 i les demés veiem que si que són significatives!*

A partir d’executar Anova(m11) podem veure quines són les variables significatives llavors agafem el nou model, que el que li hem tret és la variables poutcome i factor\_Previous(encara que aquesta última es podria agafar també com a significativa, perquè hi ha un . ).

Ara quan tenim el nostre model m8 amb els factors significatius corresponents el que hem de fer és veure si les nostres variables numèriques inicials del nostre model són més explicatives com a numèriques o com a factors.

*#Our model*  
m12<-**lm**(**log**(duration)**~**campaign**+**pdays**+**poutcome**+**month**+**factor\_cons.price.idx**+**day\_of\_week,data = df)  
**summary**(m12)

##   
## Call:  
## lm(formula = log(duration) ~ campaign + pdays + poutcome + month +   
## factor\_cons.price.idx + day\_of\_week, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2483 -0.5570 -0.0058 0.6015 2.6707   
##   
## Coefficients:  
## Estimate  
## (Intercept) 5.531569  
## campaign -0.069960  
## pdays -0.003735  
## poutcomePoutcome\_nonexistent -0.013441  
## poutcomePoutcome\_success 0.350904  
## monthMonth\_aug -0.208718  
## monthMonth\_dec 0.163868  
## monthMonth\_jul -0.193449  
## monthMonth\_jun -0.370057  
## monthMonth\_mar -0.185277  
## monthMonth\_may -0.343337  
## monthMonth\_nov -0.268959  
## monthMonth\_oct -0.219786  
## monthMonth\_sep -0.336518  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] -0.110291  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] 0.099605  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] 0.221876  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] 0.030606  
## day\_of\_weekDay\_of\_week\_mon 0.060586  
## day\_of\_weekDay\_of\_week\_thu 0.086819  
## day\_of\_weekDay\_of\_week\_tue 0.212060  
## day\_of\_weekDay\_of\_week\_wed 0.152392  
## Std. Error t value  
## (Intercept) 0.299894 18.445  
## campaign 0.006711 -10.424  
## pdays 0.018349 -0.204  
## poutcomePoutcome\_nonexistent 0.048267 -0.278  
## poutcomePoutcome\_success 0.207146 1.694  
## monthMonth\_aug 0.066463 -3.140  
## monthMonth\_dec 0.214343 0.765  
## monthMonth\_jul 0.114374 -1.691  
## monthMonth\_jun 0.105427 -3.510  
## monthMonth\_mar 0.130345 -1.421  
## monthMonth\_may 0.092788 -3.700  
## monthMonth\_nov 0.069152 -3.889  
## monthMonth\_oct 0.130666 -1.682  
## monthMonth\_sep 0.140404 -2.397  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] 0.070475 -1.565  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] 0.096455 1.033  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] 0.049128 4.516  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] 0.073276 0.418  
## day\_of\_weekDay\_of\_week\_mon 0.041394 1.464  
## day\_of\_weekDay\_of\_week\_thu 0.041260 2.104  
## day\_of\_weekDay\_of\_week\_tue 0.042907 4.942  
## day\_of\_weekDay\_of\_week\_wed 0.041820 3.644  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## campaign < 2e-16 \*\*\*  
## pdays 0.838711   
## poutcomePoutcome\_nonexistent 0.780664   
## poutcomePoutcome\_success 0.090330 .   
## monthMonth\_aug 0.001697 \*\*   
## monthMonth\_dec 0.444597   
## monthMonth\_jul 0.090828 .   
## monthMonth\_jun 0.000452 \*\*\*  
## monthMonth\_mar 0.155254   
## monthMonth\_may 0.000218 \*\*\*  
## monthMonth\_nov 0.000102 \*\*\*  
## monthMonth\_oct 0.092623 .   
## monthMonth\_sep 0.016577 \*   
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] 0.117652   
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] 0.301815   
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] 6.44e-06 \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] 0.676204   
## day\_of\_weekDay\_of\_week\_mon 0.143350   
## day\_of\_weekDay\_of\_week\_thu 0.035414 \*   
## day\_of\_weekDay\_of\_week\_tue 7.98e-07 \*\*\*  
## day\_of\_weekDay\_of\_week\_wed 0.000271 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9215 on 4924 degrees of freedom  
## Multiple R-squared: 0.05032, Adjusted R-squared: 0.04627   
## F-statistic: 12.43 on 21 and 4924 DF, p-value: < 2.2e-16

*#marginalModelPlots(m12)*  
  
*#par(mfrow=c(2,2))*  
*#plot(m13)*  
  
*#Estudi de campaign*  
  
*# Decide wether campaign should be considered either numeric, or factor (never both)*  
maux<-**lm**(**log**(duration)**~**factor\_campaign**+**pdays**+**month**+**factor\_cons.price.idx**+**day\_of\_week,data = df)

**BIC**(m12,maux) *#Choose option with minimum BIC*

## df BIC  
## m12 23 13400.74  
## maux 22 13420.62

*#El BIC més petit es el recomanable*  
*#La variable campaign numèrica m'explica més que factor\_campaign perquè el BIC de m12 és més petit que el de maux*  
  
*# Estudi de pdays*  
  
maux2<-**lm**(**log**(duration)**~**campaign**+**factor\_Pdays**+**poutcome**+**month**+**factor\_cons.price.idx**+**day\_of\_week,data = df)  
**BIC**(m12,maux2) *#Choose option with minimum BIC, for me pdays as numeric is not an option*

## df BIC  
## m12 23 13400.74  
## maux2 23 13395.80

*#El factor\_Pdays m'explica més que la variable numèrica pdays perquè el BIC de maux2 és més petir que el de m12*  
  
maux3<-**lm**(**log**(duration)**~**factor\_campaign**+**factor\_Pdays**+**poutcome**+**month**+**factor\_cons.price.idx**+**day\_of\_week,data = df)  
**BIC**(m12,maux3)

## df BIC  
## m12 23 13400.74  
## maux3 24 13429.43

*#Hi ha una millor explicabilitat en el maux2!*  
  
*#Best solution:*  
m13<-**lm**(**log**(duration)**~**campaign**+**factor\_Pdays**+**poutcome**+**month**+**factor\_cons.price.idx**+**day\_of\_week,data = df)

Després del nostre estudi, el que podem veure o les conclusions que podem treure és que les nostres variables numèriques del model incial, campaign i pdays, és que campaign és més explicativa sent numèrica mentre que la variable pdays és més explicativa quan s’utilitza com a factor i això es pot comprovar amb la comanda “BIC”.

És pot veure com en maux3 tenim un BIC més petit que en el nostre model m12, però si comprobem tots els models auxiliar veiem que el BIC més petit és el que ens dóna el model maux2.

*#Try to combine both criteria*  
**Anova**(m13) *#Check significant variables*

## Anova Table (Type II tests)  
##   
## Response: log(duration)  
## Sum Sq Df F value Pr(>F)   
## campaign 91.8 1 108.2467 < 2.2e-16 \*\*\*  
## factor\_Pdays 4.2 1 4.9628 0.025943 \*   
## poutcome 0.2 2 0.1296 0.878431   
## month 22.5 9 2.9462 0.001715 \*\*   
## factor\_cons.price.idx 20.6 4 6.0794 7.075e-05 \*\*\*  
## day\_of\_week 25.6 4 7.5441 4.692e-06 \*\*\*  
## Residuals 4176.8 4924   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

m14<-**step**(m13,k=**log**(**nrow**(df))) *#I priorize BIC criteria*

## Start: AIC=-648.84  
## log(duration) ~ campaign + factor\_Pdays + poutcome + month +   
## factor\_cons.price.idx + day\_of\_week  
##   
## Df Sum of Sq RSS AIC  
## - month 9 22.492 4199.3 -698.84  
## - poutcome 2 0.220 4177.1 -665.60  
## - factor\_cons.price.idx 4 20.628 4197.5 -658.50  
## - day\_of\_week 4 25.597 4202.4 -652.65  
## - factor\_Pdays 1 4.210 4181.1 -652.37  
## <none> 4176.8 -648.84  
## - campaign 1 91.822 4268.7 -549.80  
##   
## Step: AIC=-698.84  
## log(duration) ~ campaign + factor\_Pdays + poutcome + factor\_cons.price.idx +   
## day\_of\_week  
##   
## Df Sum of Sq RSS AIC  
## - poutcome 2 0.401 4199.7 -715.38  
## - day\_of\_week 4 22.889 4222.2 -705.98  
## - factor\_Pdays 1 5.071 4204.4 -701.38  
## <none> 4199.3 -698.84  
## - factor\_cons.price.idx 4 43.631 4243.0 -681.74  
## - campaign 1 94.896 4294.2 -596.82  
##   
## Step: AIC=-715.38  
## log(duration) ~ campaign + factor\_Pdays + factor\_cons.price.idx +   
## day\_of\_week  
##   
## Df Sum of Sq RSS AIC  
## - day\_of\_week 4 22.803 4222.5 -722.62  
## <none> 4199.7 -715.38  
## - factor\_cons.price.idx 4 45.083 4244.8 -696.59  
## - factor\_Pdays 1 39.056 4238.8 -678.10  
## - campaign 1 95.751 4295.5 -612.39  
##   
## Step: AIC=-722.62  
## log(duration) ~ campaign + factor\_Pdays + factor\_cons.price.idx  
##   
## Df Sum of Sq RSS AIC  
## <none> 4222.5 -722.62  
## - factor\_cons.price.idx 4 48.066 4270.6 -700.66  
## - factor\_Pdays 1 40.106 4262.7 -684.37  
## - campaign 1 100.169 4322.7 -615.17

**summary**(m14)

##   
## Call:  
## lm(formula = log(duration) ~ campaign + factor\_Pdays + factor\_cons.price.idx,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1686 -0.5522 -0.0012 0.6094 2.6940   
##   
## Coefficients:  
## Estimate  
## (Intercept) 5.746773  
## campaign -0.072224  
## factor\_Pdaysfactor\_Pdays-(15,17] -0.491280  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] 0.004904  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] 0.219195  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] 0.189446  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] -0.014655  
## Std. Error t value  
## (Intercept) 0.072690 79.059  
## campaign 0.006672 -10.824  
## factor\_Pdaysfactor\_Pdays-(15,17] 0.071729 -6.849  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] 0.038153 0.129  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] 0.042427 5.166  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] 0.042045 4.506  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] 0.044780 -0.327  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## campaign < 2e-16 \*\*\*  
## factor\_Pdaysfactor\_Pdays-(15,17] 8.34e-12 \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] 0.898   
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] 2.48e-07 \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] 6.76e-06 \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] 0.743   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9246 on 4939 degrees of freedom  
## Multiple R-squared: 0.04089, Adjusted R-squared: 0.03973   
## F-statistic: 35.1 on 6 and 4939 DF, p-value: < 2.2e-16

*#No tenim NA! -> PERFECTE!*  
  
*#Anova(m13)*  
  
*#m15<-lm(log(duration)~campaign+factor\_Pdays+factor\_cons.price.idx+day\_of\_week,data = df)*  
*#summary(m15)*  
*#Anova(m15)*  
  
*#Ara volem saber els nivells que tenim*  
**summary**(df[,**c**("campaign", "factor\_Pdays","factor\_cons.price.idx")])

## campaign factor\_Pdays   
## Min. : 1.000 factor\_Pdays-[0,15] : 179   
## 1st Qu.: 1.000 factor\_Pdays-(15,17]:4767   
## Median : 2.000   
## Mean : 2.389   
## 3rd Qu.: 3.000   
## Max. :14.000   
## factor\_cons.price.idx  
## factor\_cons.price.idx-[92.2,93] :1059   
## factor\_cons.price.idx-(93,93.4] :1359   
## factor\_cons.price.idx-(93.4,93.9]: 889   
## factor\_cons.price.idx-(93.9,94] : 921   
## factor\_cons.price.idx-(94,94.8] : 718   
##

*#model.matrix(m14)*

Per aconseguir la nostra matriu he agafat les variables més significatives que m’ha donat la comanda “step”, podiem agafar també a partir de fer l’Anova del nostre model final que teníem fins el moment, però hem decidit agafar el model m14 per averiguar els nivells que tenim. Fent l’Anova tenim el model m15 que també posaria en el summary les variables “month” i “day\_of\_week”, mentre que el model m14 ens dóna les variables que tenim en el summary. (Era correcte agafar qualsevol de les dues opcions).

Després de tot l’estudi hem vist que nosaltres hem fet un model i un estudi Variable Numèrica VS. Factor Mai es pot donar una interacció entre dos variables numèriques!

##Interaction: order 2 no more  
  
m15<-**lm**(**log**(duration)**~**(campaign**+**factor\_Pdays**+**factor\_cons.price.idx)**^**2,data = df)  
*#summary(m15)*  
*#coef(m15)*  
  
m16<-**step**(m15,k=**log**(**nrow**(df)))

## Start: AIC=-726.41  
## log(duration) ~ (campaign + factor\_Pdays + factor\_cons.price.idx)^2  
##   
## Df Sum of Sq RSS AIC  
## - factor\_Pdays:factor\_cons.price.idx 3 2.215 4163.9 -749.30  
## - campaign:factor\_Pdays 1 0.356 4162.0 -734.50  
## <none> 4161.7 -726.41  
## - campaign:factor\_cons.price.idx 4 58.796 4220.5 -691.05  
##   
## Step: AIC=-749.3  
## log(duration) ~ campaign + factor\_Pdays + factor\_cons.price.idx +   
## campaign:factor\_Pdays + campaign:factor\_cons.price.idx  
##   
## Df Sum of Sq RSS AIC  
## - campaign:factor\_Pdays 1 0.454 4164.3 -757.27  
## <none> 4163.9 -749.30  
## - campaign:factor\_cons.price.idx 4 58.630 4222.5 -714.17  
##   
## Step: AIC=-757.27  
## log(duration) ~ campaign + factor\_Pdays + factor\_cons.price.idx +   
## campaign:factor\_cons.price.idx  
##   
## Df Sum of Sq RSS AIC  
## <none> 4164.3 -757.27  
## - campaign:factor\_cons.price.idx 4 58.222 4222.5 -722.62  
## - factor\_Pdays 1 36.552 4200.9 -722.55

*#Anova(m16)*  
**anova**(m16,m15) *#Fisher test - Priority to BIC criteria*

## Analysis of Variance Table  
##   
## Model 1: log(duration) ~ campaign + factor\_Pdays + factor\_cons.price.idx +   
## campaign:factor\_cons.price.idx  
## Model 2: log(duration) ~ (campaign + factor\_Pdays + factor\_cons.price.idx)^2  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 4935 4164.3   
## 2 4931 4161.7 4 2.6684 0.7904 0.5312

*#Prioritzo el criteri step per agafar les redundants*

Després d’aquesta execució podem veure segons el criteri de Fisher que els dos models no són equivalents, i això ho podem saber mirant el p-valor i és molt petit!

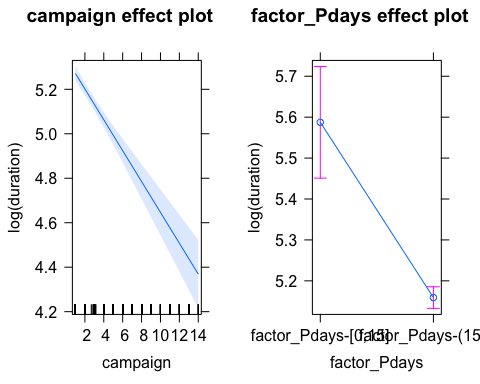
# Interactions between numeric variables and factors

## Model Additiu

*#Exemple adhoc: Y ~ X+A*   
m17<-**lm**(**log**(duration)**~**campaign**+**factor\_Pdays,data = df)  
**summary**(m17)

##   
## Call:  
## lm(formula = log(duration) ~ campaign + factor\_Pdays, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2555 -0.5417 0.0013 0.6222 2.7306   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.753204 0.070467 81.644 < 2e-16 \*\*\*  
## campaign -0.069384 0.006676 -10.394 < 2e-16 \*\*\*  
## factor\_Pdaysfactor\_Pdays-(15,17] -0.428324 0.070898 -6.041 1.64e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9295 on 4943 degrees of freedom  
## Multiple R-squared: 0.02997, Adjusted R-squared: 0.02958   
## F-statistic: 76.37 on 2 and 4943 DF, p-value: < 2.2e-16

*#Suport visual*  
*# scatterplot(log(duration)~campaign|factor\_Pdays,data=df)*  
  
*#Interpretation of models through effects library*  
**library**(effects)  
**plot**(**allEffects**(m17))



A l’eix de les ordenades tenim el logaritme de “duration” (eix vertical), campaign en aquest cas augmenta, és a dir, el número de campanyes implica una disminució en el logaritme de la durada = efecte negatiu Però el factor\_Pdays calcula un valor de confianza segons els intervals que tenim i d’aquesta manera ens ayuda a interpretar el que tenim com a sortida

Llavors ara és hora de interpretar el nostre model: Y ~ X+A i = 1 (que és equivalent al factor\_Pdays[0,15]) Yi = Y1 = 5.75-0.069X i = 2 (que és equivalent al factor\_Pdays[15,17]) Yi = Y2 = (5.75-0,428)-0.069X

## Model Interaccions

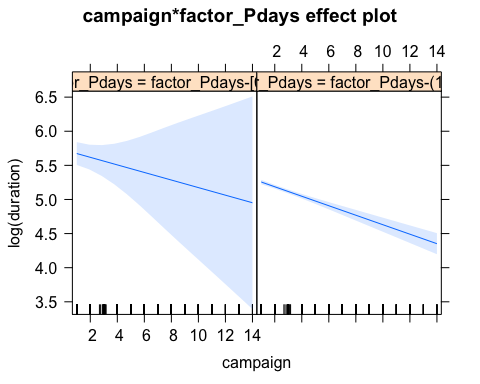
*# Y ~ X\*A (que és equivalent a X+A+A:X)*   
m18<-**lm**(**log**(duration)**~**campaign**\***factor\_Pdays,data = df) *#Concepte d'interacció ara*  
**summary**(m18)

##   
## Call:  
## lm(formula = log(duration) ~ campaign \* factor\_Pdays, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2557 -0.5418 0.0014 0.6220 2.7311   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 5.72867 0.13376 42.828  
## campaign -0.05549 0.06474 -0.857  
## factor\_Pdaysfactor\_Pdays-(15,17] -0.40343 0.13541 -2.979  
## campaign:factor\_Pdaysfactor\_Pdays-(15,17] -0.01405 0.06509 -0.216  
## Pr(>|t|)   
## (Intercept) <2e-16 \*\*\*  
## campaign 0.3915   
## factor\_Pdaysfactor\_Pdays-(15,17] 0.0029 \*\*   
## campaign:factor\_Pdaysfactor\_Pdays-(15,17] 0.8291   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9296 on 4942 degrees of freedom  
## Multiple R-squared: 0.02998, Adjusted R-squared: 0.0294   
## F-statistic: 50.92 on 3 and 4942 DF, p-value: < 2.2e-16

*# Las interaccions son rellevants?*  
**anova**(m17,m18)

## Analysis of Variance Table  
##   
## Model 1: log(duration) ~ campaign + factor\_Pdays  
## Model 2: log(duration) ~ campaign \* factor\_Pdays  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 4943 4270.6   
## 2 4942 4270.6 1 0.040249 0.0466 0.8291

*#pvalue << 0.05 -> HO Rejected -> m18 X\*A*  
*#anova(petit, gran)*  
  
**plot**(**allEffects**(m18))



*# Hi han moltes observacions influents per això hi ha tanta zona blau clar, per l'interval de confiança que tenim!*

També el que hem pogut comprobar és si les nostres interaccions són rellevants i amb la comanda “anova” fem com unaména de comparació per veure els dos models que tenim i poder treure com a conclusió que haure d’acceptar la hipòtesi nula, perquè el pvalor que surt és més gran que 0.05 (5%).

Ara és hora d’interpretar el nostre model: Y ~ X\*A i = 1 (que és equivalent al factor\_Pdays[0,15]) Yi = Y1 = 5.73-0.055X i = 2 (que és equivalent al factor\_Pdays[15,17]) Yi = Y2 = (5.73-0.403)+(-0.055-0.014)X

### Binary Regression

## Explanatory numeric variables

# Initial modelling

El que farem al començament de tot és dividir la modelització inicial (que tenim fins ara) en mostres de treball i mostres per testejar. En aquest apartat trobarem el “Eta2”, que no el podem interpretar del tot bé ja que s’utilitza més en el MCA i no l’hem pogut fer a classe, però és com un coeficient de determinació quan tenim variables involucrades que són factors. A l’hora d’escollir el nostre millor model, és bona tècnica agafar com a referència també el “Estimate” que ens dóna el pes que se li dóna a cada variable en el model, llavors veiem quines són les més explicatives. I finalment, el “z value” és una aproximació del “Estimate/Std.Error”, valors de la normal estàndard.

*# Divide into work and test samples*  
  
**set.seed**(123)  
sam<-**sample**(1**:nrow**(df),0.75**\*nrow**(df)) *#Random sample without replacement*  
  
dfw<-df[sam,]  
dft<-df[**-**sam,]  
  
*# Numeric variables*  
vars\_con

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

**catdes**(dfw[,**c**("y",vars\_con)],1) *#Numericas relacionadas*

##   
## Link between the cluster variable and the quantitative variables  
## ================================================================  
## Eta2 P-value  
## duration 0.17671414 9.254637e-159  
## nr.employed 0.14477732 4.417482e-128  
## pdays 0.13675760 1.481722e-120  
## euribor3m 0.10793163 4.600661e-94  
## emp.var.rate 0.09974083 1.089368e-86  
## previous 0.07808778 1.666707e-67  
## cons.price.idx 0.01621864 6.967791e-15  
## campaign 0.00438049 5.487012e-05  
##   
## Description of each cluster by quantitative variables  
## =====================================================  
## $Y\_no  
## v.test Mean in category Overall mean sd in category  
## nr.employed 23.169685 5177.7302797 5.167214e+03 64.7069872  
## pdays 22.518818 15.8902551 1.559935e+01 1.1196236  
## euribor3m 20.005261 3.8549862 3.641860e+00 1.6193552  
## emp.var.rate 19.231198 0.2851214 9.937989e-02 1.4698800  
## cons.price.idx 7.754916 93.6098528 9.358235e+01 0.5538129  
## campaign 4.030243 2.4041326 2.356065e+00 1.9968564  
## previous -17.016154 0.1251153 1.763279e-01 0.4006136  
## duration -25.597969 223.6446357 2.640345e+02 203.6701199  
## Overall sd p.value  
## nr.employed 73.8222624 9.207180e-119  
## pdays 2.1010235 2.715126e-112  
## euribor3m 1.7326984 4.955848e-89  
## emp.var.rate 1.5708408 2.028852e-82  
## cons.price.idx 0.5767261 8.840227e-15  
## campaign 1.9397909 5.571924e-05  
## previous 0.4894910 6.233339e-65  
## duration 256.6235243 1.607064e-144  
##   
## $Y\_yes  
## v.test Mean in category Overall mean sd in category  
## duration 25.597969 552.1666667 2.640345e+02 380.8900798  
## previous 17.016154 0.5416667 1.763279e-01 0.8073244  
## campaign -4.030243 2.0131579 2.356065e+00 1.4234264  
## cons.price.idx -7.754916 93.3861820 9.358235e+01 0.6881347  
## emp.var.rate -19.231198 -1.2256579 9.937989e-02 1.6296390  
## euribor3m -20.005261 2.1214627 3.641860e+00 1.7541244  
## pdays -22.518818 13.5241228 1.559935e+01 4.6959610  
## nr.employed -23.169685 5092.1901316 5.167214e+03 89.6674427  
## Overall sd p.value  
## duration 256.6235243 1.607064e-144  
## previous 0.4894910 6.233339e-65  
## campaign 1.9397909 5.571924e-05  
## cons.price.idx 0.5767261 8.840227e-15  
## emp.var.rate 1.5708408 2.028852e-82  
## euribor3m 1.7326984 4.955848e-89  
## pdays 2.1010235 2.715126e-112  
## nr.employed 73.8222624 9.207180e-119

*# EXEMPLE!*  
*# Model NULL, només tenim una constant*  
*# gm0<-glm(y~1,family=binomial,data = dfw)*  
*# summary(gm0)*  
  
*# binomial = distribucion que le damos a la variable de respuesta*  
*# Si volem podem utilitzar duration, sino no, si es posa és com fer una mica de trampa, no té sentit utilitzar-la com a variable explicativa, però si volem és pot utilitzar.*  
gm1<-**glm**(y**~**nr.employed**+**pdays**+**euribor3m**+**emp.var.rate**+**previous**+**cons.price.idx**+**campaign,family=binomial,data = dfw)  
*# summary(gm1)*  
*# Anova(gm1) #Test efectes nets*  
**vif**(gm1)

## nr.employed pdays euribor3m emp.var.rate previous   
## 16.957527 1.416024 24.098435 31.623083 1.692257   
## cons.price.idx campaign   
## 7.702834 1.027985

*#Saca los problemas de col·linealitat!*  
*#Més gran que 3 SON DOLENTES!*  
  
*#Remove colinear variables*  
*#Es treuran per separat i la que canviï menys el model s'agafa fins que siguin quasi totes significatives*  
gm2<-**glm**(y**~**nr.employed**+**pdays**+**euribor3m**+**previous**+**cons.price.idx**+**campaign,family=binomial,data = dfw)  
*# summary(gm2)*  
**vif**(gm2)

## nr.employed pdays euribor3m previous cons.price.idx   
## 14.181816 1.417321 18.347138 1.684602 2.968792   
## campaign   
## 1.022954

*# Anova(gm2)*  
  
*# gm3<-glm(y~nr.employed+pdays+previous+cons.price.idx+campaign,family=binomial,data = dfw)*  
*# summary(gm3)*  
*# vif(gm3)*  
*# Anova(gm3)*  
  
gm4<-**glm**(y**~**pdays**+**previous**+**cons.price.idx**+**campaign,family=binomial,data = dfw)  
**summary**(gm4)

##   
## Call:  
## glm(formula = y ~ pdays + previous + cons.price.idx + campaign,   
## family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2876 -0.4763 -0.4141 -0.3734 2.5103   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 44.22567 8.67333 5.099 3.41e-07 \*\*\*  
## pdays -0.23029 0.02344 -9.824 < 2e-16 \*\*\*  
## previous 0.49007 0.10292 4.762 1.92e-06 \*\*\*  
## cons.price.idx -0.45626 0.09254 -4.930 8.21e-07 \*\*\*  
## campaign -0.06844 0.03318 -2.063 0.0391 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2765.1 on 3708 degrees of freedom  
## Residual deviance: 2406.1 on 3704 degrees of freedom  
## AIC: 2416.1  
##   
## Number of Fisher Scoring iterations: 5

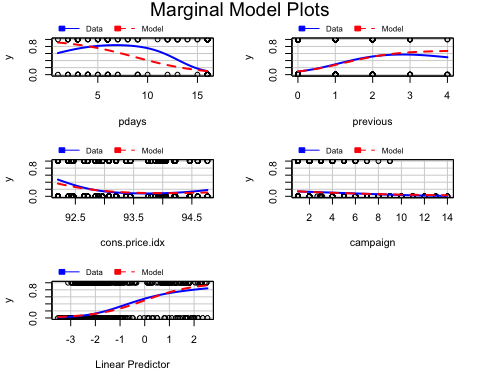
**vif**(gm4)

## pdays previous cons.price.idx campaign   
## 1.366062 1.394791 1.023703 1.015790

**Anova**(gm4)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## pdays 120.636 1 < 2.2e-16 \*\*\*  
## previous 20.643 1 5.535e-06 \*\*\*  
## cons.price.idx 24.457 1 7.600e-07 \*\*\*  
## campaign 4.603 1 0.03192 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**marginalModelPlots**(gm4) *# Some missfit data vs model*



Ara el que hem fet ha sigut trobar el nostre millor model lineas generalitzat i el que hem fet per aconseguir-ho ha sigut que a partir d’una mostra aleatòria hem anat elaborant els nostres models i amb la comanda “vif” hem anat treient els problemes de col·linealitat, és a dir, les variables que tenien un vif > 3 s’han de treure i anar probant diferents models amb les variables corresponents fins arribar a tenir un model on totes les nostres variables són significatives, però no hi ha cap estratègia òptima per dur a terme aquestes comprovacions.

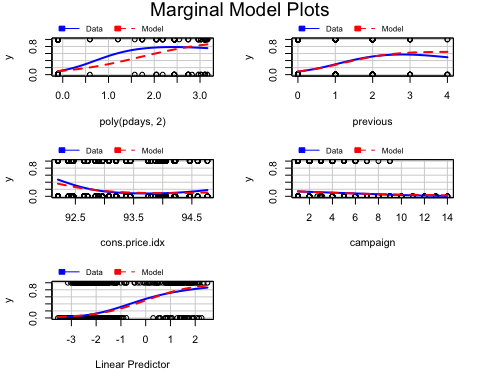
Hem aconseguit disminuir la discrepancia amb el nostre últim model (Residual deviance < Null deviance) i també es pot considerar correcte ja que Grau de llibertat = Num. observacions (3709) - Num. variables (5) = 3704 i una altra manera de veure que anem bé és que la Residual deviance és igual o inferior als graus de llibertat (2232.7 < 3704).

Com podem veure en les nostres transformacions, al model gm3 li hem tret la variable “euribor3m” respecte al model gm2 perquè segons el vif era una variable que afectava molt a la variança, però quan executàvem Anova hem vist que hi havien dos variables que no eren significatives, llavors hem optat per treure la variable “nr.employed” (Que en el model gm2 també sortia amb el vif elevat) que és el nostre model gm4 i ara quan executem Anova(gm4) podem veure que totes les variables implicades en el model són significatives, que és el que buscàvem.

# Transforming variables

El que farem a continuació és a partir del marginalPlots podem veure on hi ha un desajust entre les observacions i la predicció, llavors hem de trobar la manera d’arreglar-ho:

gm5<-**glm**(y**~poly**(pdays,2)**+**previous**+**cons.price.idx**+**campaign,family=binomial,data = dfw)  
*# summary(gm5)*  
*# Anova(gm5)*  
**marginalModelPlots**(gm5)

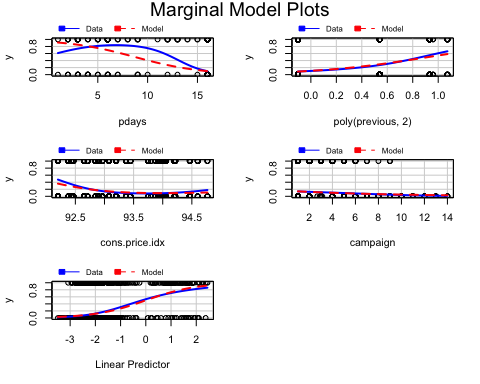


gm6<-**glm**(y**~**pdays**+poly**(previous,2)**+**cons.price.idx**+**campaign,family=binomial,data = dfw)  
**vif**(gm6)

## GVIF Df GVIF^(1/(2\*Df))  
## pdays 1.411412 1 1.188028  
## poly(previous, 2) 1.616349 2 1.127545  
## cons.price.idx 1.151112 1 1.072899  
## campaign 1.016208 1 1.008072

**marginalModelPlots**(gm6)

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



Després de fer les comprovacions aplicant el cuadràtic, veiem que en la variable pdays no canvia, sino que provoca un desajust més gran, després era hora de provar-ho amb previous i amb aquesta variable si que hi ha hagut una mica de millora, amb les variables que no són numèriques no fa falta fer-ho perquè mai sortirà res al marginalModelPlots. Llavors el model que ens quedarem serà el gm6 que és el que té menor desajust entre les observacions i les prediccions fetes.

# Adding Factors

Seguidament el que hem de fer és agafar el nostre millor model des del punt anterior i introduim els factors. El que s’ha de fer és anar probant totes les variables numèriques del nostre model fins ara com a factors i llavors ens quedem amb la que més t’expliqui segons ens indiqui el BIC.

gm10<-**glm**(y**~**pdays**+poly**(previous,2)**+**cons.price.idx**+**campaign,family=binomial,data = dfw)  
  
*# First step: Choose between numeric explanatory variable or factor*  
*# Check for all numerical variables: one by one*  
  
*# Pdays: covariate or factor??*  
gm10a<-**glm**(y**~**factor\_Pdays**+**previous**+**cons.price.idx**+**campaign,family=binomial,data = dfw)  
**BIC**(gm10,gm10a)

## df BIC  
## gm10 6 2453.155  
## gm10a 5 2421.241

*# Explica més com a factor que com a numèrica! (BIC gm10a < BIC gm10)*  
*# L'ordre pot modificar els resultats pero no es pot fer res*  
  
*# Previous?*  
gm10b<-**glm**(y**~**factor\_Pdays**+**factor\_Previous**+**cons.price.idx**+**campaign,family=binomial,data = dfw)  
**BIC**(gm10,gm10b)

## df BIC  
## gm10 6 2453.155  
## gm10b 5 2418.271

*# Explica més com a factor que com a numèrica! (BIC gm10b < BIC gm10)*  
  
*# Cons.price.idx?*  
gm10c<-**glm**(y**~**factor\_Pdays**+**factor\_Previous**+**factor\_cons.price.idx**+**campaign,family=binomial,data = dfw)  
**BIC**(gm10,gm10c)

## df BIC  
## gm10 6 2453.155  
## gm10c 8 2394.856

*# Explica més com a factor que com a numèrica! (BIC gm10c < BIC gm10)*  
  
*# Campaign?*  
gm10d<-**glm**(y**~**factor\_Pdays**+**factor\_Previous**+**factor\_cons.price.idx**+**factor\_campaign,family=binomial,data = dfw)  
**BIC**(gm10,gm10d)

## df BIC  
## gm10 6 2453.155  
## gm10d 9 2406.311

*# Explica més com a factor que com a numèrica! (BIC gm10d < BIC gm10)*  
  
## MILLOR MODEL FINS ARA:  
gm11<-**glm**(y**~**factor\_Pdays**+**factor\_Previous**+**factor\_cons.price.idx**+**factor\_campaign,family=binomial,data = dfw)

Podem veure o arribar a la conclusió després dels resultats que totes les variables del nostre model ideal fins ara que és el gm10 expliquen més com a factors que com a variables numèriques.

# Adding new factors

Ara a continuació el que farem serà després del nostre model elaborat fins ara (gm11), li afegirem les variables factors que surtin que són més explicatives al nostre model.

*# Add to your best model all new factors that are significantly related "y" according to catdes(). I assume gm10 as the best model at this point*  
vars\_dis2

## [1] "job" "marital"   
## [3] "education" "default"   
## [5] "housing" "loan"   
## [7] "contact" "month"   
## [9] "day\_of\_week" "poutcome"   
## [11] "season" "factor\_age"   
## [13] "factor\_duration" "factor\_campaign"   
## [15] "factor\_Pdays" "factor\_Previous"   
## [17] "factor\_emp.var.rate" "factor\_cons.price.idx"  
## [19] "factor\_cons.conf.idx" "factor\_euribor3m"   
## [21] "factor\_nr.employed"

**catdes**(dfw[,**c**("y",vars\_dis2)],1)

##   
## Link between the cluster variable and the categorical variables (chi-square test)  
## =================================================================================  
## p.value df  
## poutcome 2.712647e-126 2  
## factor\_Pdays 3.806493e-126 1  
## factor\_duration 2.092643e-122 7  
## factor\_euribor3m 1.068403e-109 6  
## factor\_nr.employed 1.791399e-80 1  
## month 6.985212e-66 9  
## factor\_emp.var.rate 6.316792e-57 2  
## factor\_Previous 1.141518e-51 1  
## factor\_cons.price.idx 3.525616e-33 4  
## contact 1.649866e-19 1  
## job 6.448891e-15 11  
## season 2.880483e-11 2  
## factor\_cons.conf.idx 2.949610e-11 4  
## factor\_age 2.089730e-10 4  
## default 1.153536e-09 1  
## education 1.675919e-03 7  
## factor\_campaign 2.204092e-03 2  
##   
## Description of each cluster by the categories  
## =============================================  
## $Y\_no  
## Cla/Mod  
## factor\_Pdays=factor\_Pdays-(15,17] 90.29453  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] 94.72850  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] 94.98158  
## poutcome=Poutcome\_nonexistent 91.06583  
## factor\_Previous=factor\_Previous-[0,1] 89.07680  
## factor\_duration=factor\_duration-[1,68] 99.78947  
## contact=Contact\_telephone 94.25113  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] 96.43917  
## factor\_duration=factor\_duration-(68,104] 97.60349  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] 95.78755  
## month=Month\_may 92.91139  
## factor\_duration=factor\_duration-(104,139] 96.19687  
## default=Default\_unknown 94.13299  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] 92.84294  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] 94.43535  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] 94.69835  
## factor\_duration=factor\_duration-(139,182] 94.88273  
## job=Job\_blue-collar 91.91439  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] 91.28751  
## factor\_age=factor\_age (36,41] 92.40924  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] 92.22973  
## factor\_campaign=factor\_campaign-(3,14] 91.42857  
## factor\_euribor3m=NA 93.24324  
## factor\_age=factor\_age (41,49] 90.90909  
## season=Summer 89.17346  
## season=Spring 89.36464  
## education=Education\_basic.6y 92.85714  
## job=Job\_services 91.54930  
## education=Education\_basic.9y 90.69767  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] 90.24745  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] 90.05168  
## month=Month\_jul 90.20468  
## education=NA 82.09877  
## education=Education\_professional.course 84.26966  
## marital=Marital\_single 85.56505  
## job=Job\_admin. 85.19270  
## poutcome=Poutcome\_failure 82.64249  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] 84.28005  
## factor\_campaign=factor\_campaign-[1,2] 86.54147  
## month=Month\_apr 78.57143  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] 82.53275  
## factor\_duration=factor\_duration-(329,504] 80.84211  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] 81.17871  
## job=Job\_retired 74.02597  
## month=Month\_dec 38.88889  
## job=Job\_student 64.93506  
## factor\_age=factor\_age [17,31] 81.89252  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] 78.90467  
## month=Month\_mar 52.83019  
## season=Aut-Win 78.63720  
## default=Default\_no 86.02991  
## month=Month\_sep 49.12281  
## month=Month\_oct 48.57143  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] 77.69328  
## contact=Contact\_cellular 84.08044  
## factor\_Previous=factor\_Previous-(1,5] 39.21569  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] 76.72414  
## poutcome=Poutcome\_success 21.80451  
## factor\_duration=factor\_duration-(504,2.12e+03] 57.20430  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] 72.76560  
## factor\_Pdays=factor\_Pdays-[0,15] 23.61111  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] 60.89030  
## Mod/Cla  
## factor\_Pdays=factor\_Pdays-(15,17] 98.9548109  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] 73.4706425  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] 63.4183830  
## poutcome=Poutcome\_nonexistent 89.3021826  
## factor\_Previous=factor\_Previous-[0,1] 98.7703658  
## factor\_duration=factor\_duration-[1,68] 14.5711651  
## contact=Contact\_telephone 38.3031048  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] 19.9815555  
## factor\_duration=factor\_duration-(68,104] 13.7719029  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] 16.0774670  
## month=Month\_may 33.8456809  
## factor\_duration=factor\_duration-(104,139] 13.2185675  
## default=Default\_unknown 22.1948970  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] 28.7119582  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] 17.7374731  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] 15.9237627  
## factor\_duration=factor\_duration-(139,182] 13.6796803  
## job=Job\_blue-collar 23.7626806  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] 28.9886259  
## factor\_age=factor\_age (36,41] 17.2148786  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] 16.7845066  
## factor\_campaign=factor\_campaign-(3,14] 18.6904396  
## factor\_euribor3m=NA 8.4844759  
## factor\_age=factor\_age (41,49] 19.6741469  
## season=Summer 47.0949892  
## season=Spring 39.7786658  
## education=Education\_basic.6y 5.9944666  
## job=Job\_services 9.9907777  
## education=Education\_basic.9y 15.5856133  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] 19.0593298  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] 21.4263757  
## month=Month\_jul 18.9671073  
## education=NA 4.0885337  
## education=Education\_professional.course 11.5278205  
## marital=Marital\_single 27.6975100  
## job=Job\_admin. 25.8223179  
## poutcome=Poutcome\_failure 9.8063326  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] 19.6126652  
## factor\_campaign=factor\_campaign-[1,2] 67.9987704  
## month=Month\_apr 5.0722410  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] 17.4300646  
## factor\_duration=factor\_duration-(329,504] 11.8044882  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] 13.1263449  
## job=Job\_retired 3.5044574  
## month=Month\_dec 0.2151860  
## job=Job\_student 1.5370427  
## factor\_age=factor\_age [17,31] 21.5493391  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] 11.9581924  
## month=Month\_mar 0.8607439  
## season=Aut-Win 13.1263449  
## default=Default\_no 77.8051030  
## month=Month\_sep 0.8607439  
## month=Month\_oct 1.0451891  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] 18.8441439  
## contact=Contact\_cellular 61.6968952  
## factor\_Previous=factor\_Previous-(1,5] 1.2296342  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] 24.6234245  
## poutcome=Poutcome\_success 0.8914848  
## factor\_duration=factor\_duration-(504,2.12e+03] 8.1770673  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] 26.5293575  
## factor\_Pdays=factor\_Pdays-[0,15] 1.0451891  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] 11.7737473  
## Global  
## factor\_Pdays=factor\_Pdays-(15,17] 96.117552  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] 68.023726  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] 58.560259  
## poutcome=Poutcome\_nonexistent 86.007010  
## factor\_Previous=factor\_Previous-[0,1] 97.249933  
## factor\_duration=factor\_duration-[1,68] 12.806686  
## contact=Contact\_telephone 35.643030  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] 18.172014  
## factor\_duration=factor\_duration-(68,104] 12.375303  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] 14.720949  
## month=Month\_may 31.949312  
## factor\_duration=factor\_duration-(104,139] 12.051766  
## default=Default\_unknown 20.679428  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] 27.123214  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] 16.473443  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] 14.747910  
## factor\_duration=factor\_duration-(139,182] 12.644918  
## job=Job\_blue-collar 22.674575  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] 27.851173  
## factor\_age=factor\_age (36,41] 16.338636  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] 15.961176  
## factor\_campaign=factor\_campaign-(3,14] 17.929361  
## factor\_euribor3m=NA 7.980588  
## factor\_age=factor\_age (41,49] 18.980857  
## season=Summer 46.319763  
## season=Spring 39.040173  
## education=Education\_basic.6y 5.661903  
## job=Job\_services 9.571313  
## education=Education\_basic.9y 15.071448  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] 18.522513  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] 20.868159  
## month=Month\_jul 18.441628  
## education=NA 4.367754  
## education=Education\_professional.course 11.997843  
## marital=Marital\_single 28.390402  
## job=Job\_admin. 26.583985  
## poutcome=Poutcome\_failure 10.407118  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] 20.409814  
## factor\_campaign=factor\_campaign-[1,2] 68.913454  
## month=Month\_apr 5.661903  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] 18.522513  
## factor\_duration=factor\_duration-(329,504] 12.806686  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] 14.181720  
## job=Job\_retired 4.152063  
## month=Month\_dec 0.485306  
## job=Job\_student 2.076031  
## factor\_age=factor\_age [17,31] 23.078997  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] 13.291992  
## month=Month\_mar 1.428957  
## season=Aut-Win 14.640065  
## default=Default\_no 79.320572  
## month=Month\_sep 1.536802  
## month=Month\_oct 1.887301  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] 21.272580  
## contact=Contact\_cellular 64.356970  
## factor\_Previous=factor\_Previous-(1,5] 2.750067  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] 28.147749  
## poutcome=Poutcome\_success 3.585872  
## factor\_duration=factor\_duration-(504,2.12e+03] 12.537072  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] 31.976274  
## factor\_Pdays=factor\_Pdays-[0,15] 3.882448  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] 16.958749  
## p.value  
## factor\_Pdays=factor\_Pdays-(15,17] 8.869751e-75  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] 1.507798e-74  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] 5.042204e-58  
## poutcome=Poutcome\_nonexistent 1.973670e-42  
## factor\_Previous=factor\_Previous-[0,1] 3.468405e-32  
## factor\_duration=factor\_duration-[1,68] 6.379655e-28  
## contact=Contact\_telephone 1.980375e-21  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] 1.103609e-17  
## factor\_duration=factor\_duration-(68,104] 9.799768e-16  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] 4.572876e-12  
## month=Month\_may 5.588679e-12  
## factor\_duration=factor\_duration-(104,139] 5.778351e-11  
## default=Default\_unknown 6.864912e-11  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] 1.199456e-09  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] 2.049229e-09  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] 4.409414e-09  
## factor\_duration=factor\_duration-(139,182] 3.093764e-08  
## job=Job\_blue-collar 1.147472e-05  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] 2.232259e-05  
## factor\_age=factor\_age (36,41] 5.330682e-05  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] 1.315903e-04  
## factor\_campaign=factor\_campaign-(3,14] 8.490243e-04  
## factor\_euribor3m=NA 1.321112e-03  
## factor\_age=factor\_age (41,49] 3.150623e-03  
## season=Summer 1.132004e-02  
## season=Spring 1.330307e-02  
## education=Education\_basic.6y 1.391403e-02  
## job=Job\_services 1.644228e-02  
## education=Education\_basic.9y 1.667389e-02  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] 2.208263e-02  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] 2.320824e-02  
## month=Month\_jul 2.488580e-02  
## education=NA 3.411171e-02  
## education=Education\_professional.course 2.211982e-02  
## marital=Marital\_single 1.360164e-02  
## job=Job\_admin. 5.787768e-03  
## poutcome=Poutcome\_failure 2.174750e-03  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] 1.710649e-03  
## factor\_campaign=factor\_campaign-[1,2] 1.094741e-03  
## month=Month\_apr 1.240920e-04  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] 1.090858e-05  
## factor\_duration=factor\_duration-(329,504] 3.954639e-06  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] 3.004724e-06  
## job=Job\_retired 2.283305e-06  
## month=Month\_dec 1.322665e-06  
## job=Job\_student 1.832741e-07  
## factor\_age=factor\_age [17,31] 1.347689e-08  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] 2.386878e-09  
## month=Month\_mar 3.535525e-10  
## season=Aut-Win 7.720262e-11  
## default=Default\_no 6.864912e-11  
## month=Month\_sep 1.077496e-12  
## month=Month\_oct 1.064412e-15  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] 1.225232e-19  
## contact=Contact\_cellular 1.980375e-21  
## factor\_Previous=factor\_Previous-(1,5] 3.468405e-32  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] 7.827988e-34  
## poutcome=Poutcome\_success 2.315983e-72  
## factor\_duration=factor\_duration-(504,2.12e+03] 2.002945e-74  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] 1.507798e-74  
## factor\_Pdays=factor\_Pdays-[0,15] 8.869751e-75  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] 1.278016e-86  
## v.test  
## factor\_Pdays=factor\_Pdays-(15,17] 18.296217  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] 18.267281  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] 16.057787  
## poutcome=Poutcome\_nonexistent 13.651647  
## factor\_Previous=factor\_Previous-[0,1] 11.809932  
## factor\_duration=factor\_duration-[1,68] 10.953687  
## contact=Contact\_telephone 9.506051  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] 8.562589  
## factor\_duration=factor\_duration-(68,104] 8.029341  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] 6.918240  
## month=Month\_may 6.889759  
## factor\_duration=factor\_duration-(104,139] 6.549362  
## default=Default\_unknown 6.523579  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] 6.080316  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] 5.993856  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] 5.868053  
## factor\_duration=factor\_duration-(139,182] 5.536046  
## job=Job\_blue-collar 4.387337  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] 4.240295  
## factor\_age=factor\_age (36,41] 4.040634  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] 3.823463  
## factor\_campaign=factor\_campaign-(3,14] 3.336297  
## factor\_euribor3m=NA 3.211354  
## factor\_age=factor\_age (41,49] 2.952647  
## season=Summer 2.532661  
## season=Spring 2.475551  
## education=Education\_basic.6y 2.459475  
## job=Job\_services 2.398947  
## education=Education\_basic.9y 2.393821  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] 2.288944  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] 2.269989  
## month=Month\_jul 2.243171  
## education=NA -2.118749  
## education=Education\_professional.course -2.288304  
## marital=Marital\_single -2.467615  
## job=Job\_admin. -2.759569  
## poutcome=Poutcome\_failure -3.065268  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] -3.136350  
## factor\_campaign=factor\_campaign-[1,2] -3.264974  
## month=Month\_apr -3.837898  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] -4.398332  
## factor\_duration=factor\_duration-(329,504] -4.613752  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] -4.670497  
## job=Job\_retired -4.726582  
## month=Month\_dec -4.836318  
## job=Job\_student -5.215548  
## factor\_age=factor\_age [17,31] -5.679906  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] -5.969017  
## month=Month\_mar -6.273266  
## season=Aut-Win -6.505952  
## default=Default\_no -6.523579  
## month=Month\_sep -7.120227  
## month=Month\_oct -8.019194  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] -9.066836  
## contact=Contact\_cellular -9.506051  
## factor\_Previous=factor\_Previous-(1,5] -11.809932  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] -12.124560  
## poutcome=Poutcome\_success -17.990419  
## factor\_duration=factor\_duration-(504,2.12e+03] -18.251775  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] -18.267281  
## factor\_Pdays=factor\_Pdays-[0,15] -18.296217  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] -19.726465  
##   
## $Y\_yes  
## Cla/Mod  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] 39.1096979  
## factor\_Pdays=factor\_Pdays-[0,15] 76.3888889  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] 27.2344013  
## factor\_duration=factor\_duration-(504,2.12e+03] 42.7956989  
## poutcome=Poutcome\_success 78.1954887  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] 23.2758621  
## factor\_Previous=factor\_Previous-(1,5] 60.7843137  
## contact=Contact\_cellular 15.9195643  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] 22.3067174  
## month=Month\_oct 51.4285714  
## month=Month\_sep 50.8771930  
## default=Default\_no 13.9700884  
## season=Aut-Win 21.3627993  
## month=Month\_mar 47.1698113  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] 21.0953347  
## factor\_age=factor\_age [17,31] 18.1074766  
## job=Job\_student 35.0649351  
## month=Month\_dec 61.1111111  
## job=Job\_retired 25.9740260  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] 18.8212928  
## factor\_duration=factor\_duration-(329,504] 19.1578947  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] 17.4672489  
## month=Month\_apr 21.4285714  
## factor\_campaign=factor\_campaign-[1,2] 13.4585290  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] 15.7199472  
## poutcome=Poutcome\_failure 17.3575130  
## job=Job\_admin. 14.8073022  
## marital=Marital\_single 14.4349478  
## education=Education\_professional.course 15.7303371  
## education=NA 17.9012346  
## month=Month\_jul 9.7953216  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] 9.9483204  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] 9.7525473  
## education=Education\_basic.9y 9.3023256  
## job=Job\_services 8.4507042  
## education=Education\_basic.6y 7.1428571  
## season=Spring 10.6353591  
## season=Summer 10.8265425  
## factor\_age=factor\_age (41,49] 9.0909091  
## factor\_euribor3m=NA 6.7567568  
## factor\_campaign=factor\_campaign-(3,14] 8.5714286  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] 7.7702703  
## factor\_age=factor\_age (36,41] 7.5907591  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] 8.7124879  
## job=Job\_blue-collar 8.0856124  
## factor\_duration=factor\_duration-(139,182] 5.1172708  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] 5.3016453  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] 5.5646481  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] 7.1570577  
## default=Default\_unknown 5.8670143  
## factor\_duration=factor\_duration-(104,139] 3.8031320  
## month=Month\_may 7.0886076  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] 4.2124542  
## factor\_duration=factor\_duration-(68,104] 2.3965142  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] 3.5608309  
## contact=Contact\_telephone 5.7488654  
## factor\_duration=factor\_duration-[1,68] 0.2105263  
## factor\_Previous=factor\_Previous-[0,1] 10.9232049  
## poutcome=Poutcome\_nonexistent 8.9341693  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] 5.0184162  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] 5.2715022  
## factor\_Pdays=factor\_Pdays-(15,17] 9.7054698  
## Mod/Cla  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] 53.9473684  
## factor\_Pdays=factor\_Pdays-[0,15] 24.1228070  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] 70.8333333  
## factor\_duration=factor\_duration-(504,2.12e+03] 43.6403509  
## poutcome=Poutcome\_success 22.8070175  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] 53.2894737  
## factor\_Previous=factor\_Previous-(1,5] 13.5964912  
## contact=Contact\_cellular 83.3333333  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] 38.5964912  
## month=Month\_oct 7.8947368  
## month=Month\_sep 6.3596491  
## default=Default\_no 90.1315789  
## season=Aut-Win 25.4385965  
## month=Month\_mar 5.4824561  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] 22.8070175  
## factor\_age=factor\_age [17,31] 33.9912281  
## job=Job\_student 5.9210526  
## month=Month\_dec 2.4122807  
## job=Job\_retired 8.7719298  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] 21.7105263  
## factor\_duration=factor\_duration-(329,504] 19.9561404  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] 26.3157895  
## month=Month\_apr 9.8684211  
## factor\_campaign=factor\_campaign-[1,2] 75.4385965  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] 26.0964912  
## poutcome=Poutcome\_failure 14.6929825  
## job=Job\_admin. 32.0175439  
## marital=Marital\_single 33.3333333  
## education=Education\_professional.course 15.3508772  
## education=NA 6.3596491  
## month=Month\_jul 14.6929825  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] 16.8859649  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] 14.6929825  
## education=Education\_basic.9y 11.4035088  
## job=Job\_services 6.5789474  
## education=Education\_basic.6y 3.2894737  
## season=Spring 33.7719298  
## season=Summer 40.7894737  
## factor\_age=factor\_age (41,49] 14.0350877  
## factor\_euribor3m=NA 4.3859649  
## factor\_campaign=factor\_campaign-(3,14] 12.5000000  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] 10.0877193  
## factor\_age=factor\_age (36,41] 10.0877193  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] 19.7368421  
## job=Job\_blue-collar 14.9122807  
## factor\_duration=factor\_duration-(139,182] 5.2631579  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] 6.3596491  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] 7.4561404  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] 15.7894737  
## default=Default\_unknown 9.8684211  
## factor\_duration=factor\_duration-(104,139] 3.7280702  
## month=Month\_may 18.4210526  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] 5.0438596  
## factor\_duration=factor\_duration-(68,104] 2.4122807  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] 5.2631579  
## contact=Contact\_telephone 16.6666667  
## factor\_duration=factor\_duration-[1,68] 0.2192982  
## factor\_Previous=factor\_Previous-[0,1] 86.4035088  
## poutcome=Poutcome\_nonexistent 62.5000000  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] 23.9035088  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] 29.1666667  
## factor\_Pdays=factor\_Pdays-(15,17] 75.8771930  
## Global  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] 16.958749  
## factor\_Pdays=factor\_Pdays-[0,15] 3.882448  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] 31.976274  
## factor\_duration=factor\_duration-(504,2.12e+03] 12.537072  
## poutcome=Poutcome\_success 3.585872  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] 28.147749  
## factor\_Previous=factor\_Previous-(1,5] 2.750067  
## contact=Contact\_cellular 64.356970  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] 21.272580  
## month=Month\_oct 1.887301  
## month=Month\_sep 1.536802  
## default=Default\_no 79.320572  
## season=Aut-Win 14.640065  
## month=Month\_mar 1.428957  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] 13.291992  
## factor\_age=factor\_age [17,31] 23.078997  
## job=Job\_student 2.076031  
## month=Month\_dec 0.485306  
## job=Job\_retired 4.152063  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] 14.181720  
## factor\_duration=factor\_duration-(329,504] 12.806686  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] 18.522513  
## month=Month\_apr 5.661903  
## factor\_campaign=factor\_campaign-[1,2] 68.913454  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] 20.409814  
## poutcome=Poutcome\_failure 10.407118  
## job=Job\_admin. 26.583985  
## marital=Marital\_single 28.390402  
## education=Education\_professional.course 11.997843  
## education=NA 4.367754  
## month=Month\_jul 18.441628  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] 20.868159  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] 18.522513  
## education=Education\_basic.9y 15.071448  
## job=Job\_services 9.571313  
## education=Education\_basic.6y 5.661903  
## season=Spring 39.040173  
## season=Summer 46.319763  
## factor\_age=factor\_age (41,49] 18.980857  
## factor\_euribor3m=NA 7.980588  
## factor\_campaign=factor\_campaign-(3,14] 17.929361  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] 15.961176  
## factor\_age=factor\_age (36,41] 16.338636  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] 27.851173  
## job=Job\_blue-collar 22.674575  
## factor\_duration=factor\_duration-(139,182] 12.644918  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] 14.747910  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] 16.473443  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] 27.123214  
## default=Default\_unknown 20.679428  
## factor\_duration=factor\_duration-(104,139] 12.051766  
## month=Month\_may 31.949312  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] 14.720949  
## factor\_duration=factor\_duration-(68,104] 12.375303  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] 18.172014  
## contact=Contact\_telephone 35.643030  
## factor\_duration=factor\_duration-[1,68] 12.806686  
## factor\_Previous=factor\_Previous-[0,1] 97.249933  
## poutcome=Poutcome\_nonexistent 86.007010  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] 58.560259  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] 68.023726  
## factor\_Pdays=factor\_Pdays-(15,17] 96.117552  
## p.value  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] 1.278016e-86  
## factor\_Pdays=factor\_Pdays-[0,15] 8.869751e-75  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] 1.507798e-74  
## factor\_duration=factor\_duration-(504,2.12e+03] 2.002945e-74  
## poutcome=Poutcome\_success 2.315983e-72  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] 7.827988e-34  
## factor\_Previous=factor\_Previous-(1,5] 3.468405e-32  
## contact=Contact\_cellular 1.980375e-21  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] 1.225232e-19  
## month=Month\_oct 1.064412e-15  
## month=Month\_sep 1.077496e-12  
## default=Default\_no 6.864912e-11  
## season=Aut-Win 7.720262e-11  
## month=Month\_mar 3.535525e-10  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] 2.386878e-09  
## factor\_age=factor\_age [17,31] 1.347689e-08  
## job=Job\_student 1.832741e-07  
## month=Month\_dec 1.322665e-06  
## job=Job\_retired 2.283305e-06  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] 3.004724e-06  
## factor\_duration=factor\_duration-(329,504] 3.954639e-06  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] 1.090858e-05  
## month=Month\_apr 1.240920e-04  
## factor\_campaign=factor\_campaign-[1,2] 1.094741e-03  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] 1.710649e-03  
## poutcome=Poutcome\_failure 2.174750e-03  
## job=Job\_admin. 5.787768e-03  
## marital=Marital\_single 1.360164e-02  
## education=Education\_professional.course 2.211982e-02  
## education=NA 3.411171e-02  
## month=Month\_jul 2.488580e-02  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] 2.320824e-02  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] 2.208263e-02  
## education=Education\_basic.9y 1.667389e-02  
## job=Job\_services 1.644228e-02  
## education=Education\_basic.6y 1.391403e-02  
## season=Spring 1.330307e-02  
## season=Summer 1.132004e-02  
## factor\_age=factor\_age (41,49] 3.150623e-03  
## factor\_euribor3m=NA 1.321112e-03  
## factor\_campaign=factor\_campaign-(3,14] 8.490243e-04  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] 1.315903e-04  
## factor\_age=factor\_age (36,41] 5.330682e-05  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] 2.232259e-05  
## job=Job\_blue-collar 1.147472e-05  
## factor\_duration=factor\_duration-(139,182] 3.093764e-08  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] 4.409414e-09  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] 2.049229e-09  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] 1.199456e-09  
## default=Default\_unknown 6.864912e-11  
## factor\_duration=factor\_duration-(104,139] 5.778351e-11  
## month=Month\_may 5.588679e-12  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] 4.572876e-12  
## factor\_duration=factor\_duration-(68,104] 9.799768e-16  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] 1.103609e-17  
## contact=Contact\_telephone 1.980375e-21  
## factor\_duration=factor\_duration-[1,68] 6.379655e-28  
## factor\_Previous=factor\_Previous-[0,1] 3.468405e-32  
## poutcome=Poutcome\_nonexistent 1.973670e-42  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] 5.042204e-58  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] 1.507798e-74  
## factor\_Pdays=factor\_Pdays-(15,17] 8.869751e-75  
## v.test  
## factor\_euribor3m=factor\_euribor3m-[0.634,1.266] 19.726465  
## factor\_Pdays=factor\_Pdays-[0,15] 18.296217  
## factor\_nr.employed=factor\_nr.employed-[4.96e+03,5.1e+03] 18.267281  
## factor\_duration=factor\_duration-(504,2.12e+03] 18.251775  
## poutcome=Poutcome\_success 17.990419  
## factor\_emp.var.rate=factor\_emp.var.rate-[-3.4,-1.8] 12.124560  
## factor\_Previous=factor\_Previous-(1,5] 11.809932  
## contact=Contact\_cellular 9.506051  
## factor\_cons.price.idx=factor\_cons.price.idx-[92.2,93] 9.066836  
## month=Month\_oct 8.019194  
## month=Month\_sep 7.120227  
## default=Default\_no 6.523579  
## season=Aut-Win 6.505952  
## month=Month\_mar 6.273266  
## factor\_emp.var.rate=factor\_emp.var.rate-(-1.8,-0.1] 5.969017  
## factor\_age=factor\_age [17,31] 5.679906  
## job=Job\_student 5.215548  
## month=Month\_dec 4.836318  
## job=Job\_retired 4.726582  
## factor\_cons.price.idx=factor\_cons.price.idx-(94,94.8] 4.670497  
## factor\_duration=factor\_duration-(329,504] 4.613752  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-36.4,-29.8] 4.398332  
## month=Month\_apr 3.837898  
## factor\_campaign=factor\_campaign-[1,2] 3.264974  
## factor\_cons.conf.idx=factor\_cons.conf.idx-[-50.8,-46.2] 3.136350  
## poutcome=Poutcome\_failure 3.065268  
## job=Job\_admin. 2.759569  
## marital=Marital\_single 2.467615  
## education=Education\_professional.course 2.288304  
## education=NA 2.118749  
## month=Month\_jul -2.243171  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-40.3,-36.4] -2.269989  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.4,93.9] -2.288944  
## education=Education\_basic.9y -2.393821  
## job=Job\_services -2.398947  
## education=Education\_basic.6y -2.459475  
## season=Spring -2.475551  
## season=Summer -2.532661  
## factor\_age=factor\_age (41,49] -2.952647  
## factor\_euribor3m=NA -3.211354  
## factor\_campaign=factor\_campaign-(3,14] -3.336297  
## factor\_euribor3m=factor\_euribor3m-(1.415,4.856] -3.823463  
## factor\_age=factor\_age (36,41] -4.040634  
## factor\_cons.price.idx=factor\_cons.price.idx-(93,93.4] -4.240295  
## job=Job\_blue-collar -4.387337  
## factor\_duration=factor\_duration-(139,182] -5.536046  
## factor\_euribor3m=factor\_euribor3m-(4.864,4.961] -5.868053  
## factor\_euribor3m=factor\_euribor3m-(4.961,4.964] -5.993856  
## factor\_cons.conf.idx=factor\_cons.conf.idx-(-46.2,-42] -6.080316  
## default=Default\_unknown -6.523579  
## factor\_duration=factor\_duration-(104,139] -6.549362  
## month=Month\_may -6.889759  
## factor\_euribor3m=factor\_euribor3m-(4.856,4.864] -6.918240  
## factor\_duration=factor\_duration-(68,104] -8.029341  
## factor\_cons.price.idx=factor\_cons.price.idx-(93.9,94] -8.562589  
## contact=Contact\_telephone -9.506051  
## factor\_duration=factor\_duration-[1,68] -10.953687  
## factor\_Previous=factor\_Previous-[0,1] -11.809932  
## poutcome=Poutcome\_nonexistent -13.651647  
## factor\_emp.var.rate=factor\_emp.var.rate-(-0.1,1.4] -16.057787  
## factor\_nr.employed=factor\_nr.employed-(5.1e+03,5.23e+03] -18.267281  
## factor\_Pdays=factor\_Pdays-(15,17] -18.296217

*# No hem de repetir els factors que ja tenim fins al moment comprovats i això s'ha de fer agafant el model estudiat anteriorment*  
gm12<-**glm**(y**~**factor\_Pdays**+**factor\_Previous**+**factor\_cons.price.idx**+**factor\_campaign**+**poutcome**+**month**+**job**+**season**+**default**+**education,family=binomial,data = dfw)  
*# Anova(gm12)*  
*# summary(gm12)*  
  
*#Amb el summary(gm12) he vist que tinc NA a la meva vostra en la variable factor "season" i per això també em surt error en l'execució del vif, perquè tenia aquesta variable que no era molt redundant, llavors:*  
  
gm12a<-**glm**(y**~**factor\_Pdays**+**factor\_Previous**+**factor\_cons.price.idx**+**factor\_campaign**+**poutcome**+**month**+**job**+**default**+**education,family=binomial,data = dfw)  
**Anova**(gm12a) *# Mirem les que ens interessen i les que no!*

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## factor\_Pdays 1.112 1 0.29164   
## factor\_Previous 4.045 1 0.04430 \*   
## factor\_cons.price.idx 57.732 4 8.686e-12 \*\*\*  
## factor\_campaign 1.580 2 0.45392   
## poutcome 6.035 2 0.04892 \*   
## month 87.675 9 4.762e-15 \*\*\*  
## job 12.743 11 0.31047   
## default 6.003 1 0.01428 \*   
## education 7.193 6 0.30338   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**vif**(gm12a)

## GVIF Df GVIF^(1/(2\*Df))  
## factor\_Pdays 9.527644 1 3.086688  
## factor\_Previous 1.560871 1 1.249348  
## factor\_cons.price.idx 31.904305 4 1.541634  
## factor\_campaign 1.055823 2 1.013673  
## poutcome 11.555512 2 1.843730  
## month 36.559308 9 1.221331  
## job 3.689568 11 1.061137  
## default 1.089252 1 1.043672  
## education 3.182190 6 1.101270

*#A partir de l'Anova veiem que hi han variables factors no significatives, que no ens aporten res al model, llavors les treiem:*  
  
gm12b<-**glm**(y**~**factor\_Previous**+**factor\_cons.price.idx**+**poutcome**+**month**+**default,family=binomial,data = dfw)  
**Anova**(gm12b)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## factor\_Previous 7.266 1 0.007027 \*\*   
## factor\_cons.price.idx 65.835 4 1.716e-13 \*\*\*  
## poutcome 120.651 2 < 2.2e-16 \*\*\*  
## month 109.822 9 < 2.2e-16 \*\*\*  
## default 8.504 1 0.003543 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**vif**(gm12b)

## GVIF Df GVIF^(1/(2\*Df))  
## factor\_Previous 1.351135 1 1.162383  
## factor\_cons.price.idx 28.887284 4 1.522609  
## poutcome 1.521054 2 1.110545  
## month 28.115574 9 1.203641  
## default 1.035864 1 1.017774

gm13<-**step**(gm12b,k=**log**(**nrow**(dfw)))

## Start: AIC=2354.17  
## y ~ factor\_Previous + factor\_cons.price.idx + poutcome + month +   
## default  
##   
## Df Deviance AIC  
## - factor\_Previous 1 2213.5 2353.2  
## <none> 2206.2 2354.2  
## - default 1 2214.7 2354.5  
## - factor\_cons.price.idx 4 2272.1 2387.1  
## - month 9 2316.1 2390.0  
## - poutcome 2 2326.9 2458.4  
##   
## Step: AIC=2353.22  
## y ~ factor\_cons.price.idx + poutcome + month + default  
##   
## Df Deviance AIC  
## <none> 2213.5 2353.2  
## - default 1 2222.1 2353.6  
## - factor\_cons.price.idx 4 2278.1 2384.9  
## - month 9 2327.5 2393.3  
## - poutcome 2 2374.7 2498.0

*#vif(gm13)*  
  
*#* END *POINT: No colinearity, all net effects for factors and numeric variables should be significant*  
*# colinearity: Se mira con el vig, el apartado GVIF que sean < 3*

Després de fer el procés de modelització introduint les millores pas a pas, hem pogut observar que el nostre millor model completat amb els factors que faltaven és el model gm12b, i també ho podem comprovar executant la comanda Anova i veiem com totes les variables factors són significatives. Un model també òptim i correcte seria el gm13, ja que aquest surt després d’executar la comanda “step” al model gm12b.

*# Check your final model at this point: all coefficients should be available in the summary(model)*  
**summary**(gm12b)

##   
## Call:  
## glm(formula = y ~ factor\_Previous + factor\_cons.price.idx + poutcome +   
## month + default, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3646 -0.4763 -0.3483 -0.2866 2.7158   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 0.20017 0.29558  
## factor\_Previousfactor\_Previous-(1,5] 0.79436 0.29289  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] -1.65895 0.23230  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] -1.13814 0.31381  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] -1.08805 0.26039  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] -0.40926 0.23599  
## poutcomePoutcome\_nonexistent -0.03669 0.17995  
## poutcomePoutcome\_success 2.47038 0.27019  
## monthMonth\_aug -1.32216 0.25693  
## monthMonth\_dec -0.30063 0.60409  
## monthMonth\_jul -1.29686 0.35683  
## monthMonth\_jun -1.87335 0.34855  
## monthMonth\_mar 0.07422 0.37630  
## monthMonth\_may -2.24742 0.31011  
## monthMonth\_nov -1.31315 0.26964  
## monthMonth\_oct -0.47742 0.38193  
## monthMonth\_sep -0.73219 0.41880  
## defaultDefault\_unknown -0.49048 0.17571  
## z value Pr(>|z|)  
## (Intercept) 0.677 0.498265  
## factor\_Previousfactor\_Previous-(1,5] 2.712 0.006684  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] -7.142 9.23e-13  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] -3.627 0.000287  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] -4.179 2.93e-05  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] -1.734 0.082873  
## poutcomePoutcome\_nonexistent -0.204 0.838448  
## poutcomePoutcome\_success 9.143 < 2e-16  
## monthMonth\_aug -5.146 2.66e-07  
## monthMonth\_dec -0.498 0.618717  
## monthMonth\_jul -3.634 0.000279  
## monthMonth\_jun -5.375 7.67e-08  
## monthMonth\_mar 0.197 0.843652  
## monthMonth\_may -7.247 4.25e-13  
## monthMonth\_nov -4.870 1.12e-06  
## monthMonth\_oct -1.250 0.211293  
## monthMonth\_sep -1.748 0.080411  
## defaultDefault\_unknown -2.791 0.005248  
##   
## (Intercept)   
## factor\_Previousfactor\_Previous-(1,5] \*\*   
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] .   
## poutcomePoutcome\_nonexistent   
## poutcomePoutcome\_success \*\*\*  
## monthMonth\_aug \*\*\*  
## monthMonth\_dec   
## monthMonth\_jul \*\*\*  
## monthMonth\_jun \*\*\*  
## monthMonth\_mar   
## monthMonth\_may \*\*\*  
## monthMonth\_nov \*\*\*  
## monthMonth\_oct   
## monthMonth\_sep .   
## defaultDefault\_unknown \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2765.1 on 3708 degrees of freedom  
## Residual deviance: 2206.2 on 3691 degrees of freedom  
## AIC: 2242.2  
##   
## Number of Fisher Scoring iterations: 6

*# Month too many levels. Try to use season*  
gm14<-**glm**(y**~**factor\_Previous**+**factor\_cons.price.idx**+**poutcome**+**season**+**default,family=binomial,data = dfw)  
**Anova**(gm14)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## factor\_Previous 8.978 1 0.0027321 \*\*   
## factor\_cons.price.idx 68.010 4 5.969e-14 \*\*\*  
## poutcome 160.529 2 < 2.2e-16 \*\*\*  
## season 9.555 2 0.0084162 \*\*   
## default 13.495 1 0.0002392 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**vif**(gm14)

## GVIF Df GVIF^(1/(2\*Df))  
## factor\_Previous 1.302512 1 1.141277  
## factor\_cons.price.idx 2.507984 4 1.121800  
## poutcome 1.457428 2 1.098745  
## season 2.328777 2 1.235327  
## default 1.022145 1 1.011012

*#Ahora no nos aparecen NA!*  
  
*#anova(gm12b,gm12) #Test for nested models not equivalent*  
**Anova**(gm12b, test="LR")

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## factor\_Previous 7.266 1 0.007027 \*\*   
## factor\_cons.price.idx 65.835 4 1.716e-13 \*\*\*  
## poutcome 120.651 2 < 2.2e-16 \*\*\*  
## month 109.822 9 < 2.2e-16 \*\*\*  
## default 8.504 1 0.003543 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*#cooks.distance(gm12b)*

# Add to the best model: INTERACTIONS

Un cop utilitzades variables numèriques i factors en la construcció del model, en aquest apartat utilitzarem les interaccions per tal de veure si aquesta eina millora el nostre model. I el model que tenim fins ara és el model gm12b i si surten NA agafem el model gm14, llavors farem les interaccions sobre aquest.

En el primer cas provarem de utilitzar factor\_Previous com a interacció:

mf1<-**glm**(y **~** (factor\_cons.price.idx**+**poutcome**+**month**+**default)**\***(factor\_Previous), family = binomial, data = dfw)  
  
**Anova**(mf1,test="LR")

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## factor\_cons.price.idx 58.580 4 5.765e-12 \*\*\*  
## poutcome 112.230 2 < 2.2e-16 \*\*\*  
## month 116.016 9 < 2.2e-16 \*\*\*  
## default 7.624 1 0.005759 \*\*   
## factor\_Previous 7.266 1 0.007027 \*\*   
## factor\_cons.price.idx:factor\_Previous 2.694 3 0.441214   
## poutcome:factor\_Previous 1.244 1 0.264685   
## month:factor\_Previous 7.044 9 0.632521   
## default:factor\_Previous 0.880 1 0.348089   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

A partir del test d’efectes nets veiem que la interacció amb factor\_Previous no aporta res rellevant al model. Continuem amb el model anterior gm12b.

A continuació intentarem una interacció amb poutcome:

mf2<-**glm**(y **~** (factor\_Previous**+**factor\_cons.price.idx**+**month**+**default)**\***(poutcome), family = binomial, data = dfw)  
  
  
**Anova**(mf2,test="LR")

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## factor\_Previous 2.484 1 0.114983   
## factor\_cons.price.idx 57.032 4 1.218e-11 \*\*\*  
## month 115.339 9 < 2.2e-16 \*\*\*  
## default 5.134 1 0.023460 \*   
## poutcome 120.651 2 < 2.2e-16 \*\*\*  
## factor\_Previous:poutcome 0.391 1 0.531576   
## factor\_cons.price.idx:poutcome 10.417 6 0.108173   
## month:poutcome 41.408 18 0.001337 \*\*   
## default:poutcome 1.727 2 0.421763   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**BIC**(mf2, gm12b)

## df BIC  
## mf2 45 2513.298  
## gm12b 18 2354.171

Es pot veure que hi ha una interacció que si que és rellevant, que és la month:poutcome

mf3<-**step**(mf2, k=**log**(**nrow**(dfw)))

## Start: AIC=2513.3  
## y ~ (factor\_Previous + factor\_cons.price.idx + month + default) \*   
## (poutcome)  
##   
## Df Deviance AIC  
## - month:poutcome 18 2184.9 2406.8  
## - factor\_cons.price.idx:poutcome 6 2153.9 2474.4  
## - default:poutcome 2 2145.2 2498.6  
## - factor\_Previous:poutcome 1 2143.9 2505.5  
## <none> 2143.5 2513.3  
##   
## Step: AIC=2406.77  
## y ~ factor\_Previous + factor\_cons.price.idx + month + default +   
## poutcome + factor\_Previous:poutcome + factor\_cons.price.idx:poutcome +   
## default:poutcome  
##   
## Df Deviance AIC  
## - factor\_cons.price.idx:poutcome 6 2203.2 2375.8  
## - default:poutcome 2 2186.8 2392.3  
## - factor\_Previous:poutcome 1 2185.0 2398.6  
## <none> 2184.9 2406.8  
## - month 9 2300.2 2448.2  
##   
## Step: AIC=2375.77  
## y ~ factor\_Previous + factor\_cons.price.idx + month + default +   
## poutcome + factor\_Previous:poutcome + default:poutcome  
##   
## Df Deviance AIC  
## - default:poutcome 2 2205.4 2361.5  
## - factor\_Previous:poutcome 1 2204.1 2368.4  
## <none> 2203.2 2375.8  
## - factor\_cons.price.idx 4 2269.2 2408.9  
## - month 9 2315.2 2413.8  
##   
## Step: AIC=2361.53  
## y ~ factor\_Previous + factor\_cons.price.idx + month + default +   
## poutcome + factor\_Previous:poutcome  
##   
## Df Deviance AIC  
## - factor\_Previous:poutcome 1 2206.2 2354.2  
## <none> 2205.4 2361.5  
## - default 1 2213.8 2361.7  
## - factor\_cons.price.idx 4 2272.0 2395.3  
## - month 9 2316.1 2398.2  
##   
## Step: AIC=2354.17  
## y ~ factor\_Previous + factor\_cons.price.idx + month + default +   
## poutcome  
##   
## Df Deviance AIC  
## - factor\_Previous 1 2213.5 2353.2  
## <none> 2206.2 2354.2  
## - default 1 2214.7 2354.5  
## - factor\_cons.price.idx 4 2272.1 2387.1  
## - month 9 2316.1 2390.0  
## - poutcome 2 2326.9 2458.4  
##   
## Step: AIC=2353.22  
## y ~ factor\_cons.price.idx + month + default + poutcome  
##   
## Df Deviance AIC  
## <none> 2213.5 2353.2  
## - default 1 2222.1 2353.6  
## - factor\_cons.price.idx 4 2278.1 2384.9  
## - month 9 2327.5 2393.3  
## - poutcome 2 2374.7 2498.0

**Anova**(mf3,test="LR")

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## factor\_cons.price.idx 64.601 4 3.122e-13 \*\*\*  
## month 114.026 9 < 2.2e-16 \*\*\*  
## default 8.582 1 0.003396 \*\*   
## poutcome 161.220 2 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**BIC**(mf3, gm12b)

## df BIC  
## mf3 17 2353.218  
## gm12b 18 2354.171

Un cop realitzades les interaccions realitzem una comparació del model de partida sense interaccions (gm12b) i el millor model obtingut a partir de les interaccions. Per poca diferència, però veiem que el model sense interaccions és millor. Per tant continuarem amb el model gm12b.

## Model final

Un cop realitzat l’anterior estudi, proposem el model gm14 com a model final, ja que és el mateix que el model gm12b, l’única cosa que agrupa els mesos segons les estacions.

*#summary(gm12b)*  
**summary**(gm14)

##   
## Call:  
## glm(formula = y ~ factor\_Previous + factor\_cons.price.idx + poutcome +   
## season + default, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2327 -0.4963 -0.3845 -0.2898 2.7465   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) -1.47527 0.15932  
## factor\_Previousfactor\_Previous-(1,5] 0.86497 0.28652  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] -0.85528 0.16264  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] -0.46882 0.20044  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] -1.60689 0.24339  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] -0.18375 0.19862  
## poutcomePoutcome\_nonexistent -0.06729 0.17421  
## poutcomePoutcome\_success 2.71804 0.26050  
## seasonSummer -0.24255 0.17346  
## seasonAut-Win 0.29833 0.17494  
## defaultDefault\_unknown -0.59889 0.17241  
## z value Pr(>|z|)  
## (Intercept) -9.260 < 2e-16  
## factor\_Previousfactor\_Previous-(1,5] 3.019 0.002537  
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] -5.259 1.45e-07  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] -2.339 0.019337  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] -6.602 4.05e-11  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8] -0.925 0.354887  
## poutcomePoutcome\_nonexistent -0.386 0.699287  
## poutcomePoutcome\_success 10.434 < 2e-16  
## seasonSummer -1.398 0.162027  
## seasonAut-Win 1.705 0.088134  
## defaultDefault\_unknown -3.474 0.000514  
##   
## (Intercept) \*\*\*  
## factor\_Previousfactor\_Previous-(1,5] \*\*   
## factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] \*   
## factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] \*\*\*  
## factor\_cons.price.idxfactor\_cons.price.idx-(94,94.8]   
## poutcomePoutcome\_nonexistent   
## poutcomePoutcome\_success \*\*\*  
## seasonSummer   
## seasonAut-Win .   
## defaultDefault\_unknown \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2765.1 on 3708 degrees of freedom  
## Residual deviance: 2306.5 on 3698 degrees of freedom  
## AIC: 2328.5  
##   
## Number of Fisher Scoring iterations: 6

## Interpretació del model final

Y = -1.475 + 0.863factor\_Previousfactor\_Previous-(1,5] - 0.855factor\_cons.price.idxfactor\_cons.price.idx-(93,93.4] - 0.469factor\_cons.price.idxfactor\_cons.price.idx-(93.4,93.9] - 1.607factor\_cons.price.idxfactor\_cons.price.idx-(93.9,94] + 2.712poutcomePoutcome\_success + 0.298seasonAut-Win - 0.598defaultDefault\_unknown

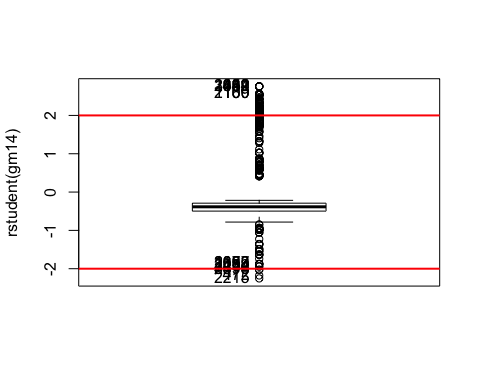
# Validació del model

## Anàlisi dels residus

**Boxplot**(**rstudent**(gm14), id.n=2)

## [1] 2215 472 2899 3378 2252 2434 1053 1373 2167 2690 144 460 612 932  
## [15] 1491 2359 3432 100 1180 2109

**abline**(h=**c**(2,**-**2),col="red",lwd=2)

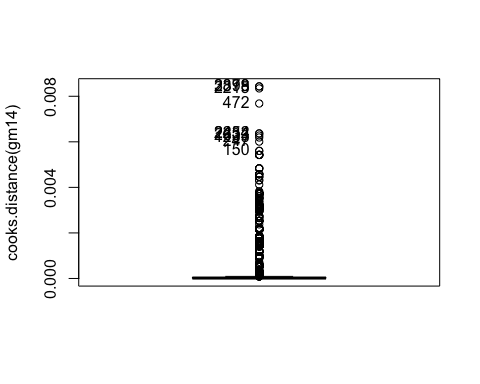


out2 <- **which**(**rstudent**(gm14) **>=** 3 **|** **rstudent**(gm14) **<=** **-**3);**length**(out2)

## [1] 0

A partir de l’anàlisi de residus veiem que no hi han quasi possibles outliers. Però ens centrarem en buscar si existeix alguna dada influent entre aquests:

infl<-**Boxplot**(**cooks.distance**(gm14), id.n=4)



llinfl<-**which**(**abs**(**cooks.distance**(gm14))**>**3);**length**(llinfl)

## [1] 0

dfw[llinfl,]

## [1] age job marital   
## [4] education default housing   
## [7] loan contact month   
## [10] day\_of\_week duration campaign   
## [13] pdays previous poutcome   
## [16] emp.var.rate cons.price.idx cons.conf.idx   
## [19] euribor3m nr.employed y   
## [22] missings\_indiv errors\_indiv outliers\_indiv   
## [25] season factor\_age factor\_duration   
## [28] factor\_campaign factor\_Pdays factor\_Previous   
## [31] factor\_emp.var.rate factor\_cons.price.idx factor\_cons.conf.idx   
## [34] factor\_euribor3m factor\_nr.employed CLUSTER   
## [37] f.CLUSTER   
## <0 rows> (or 0-length row.names)

**influencePlot**(gm14,id.n=3)

## Warning in plot.window(...): "id.n" is not a graphical parameter

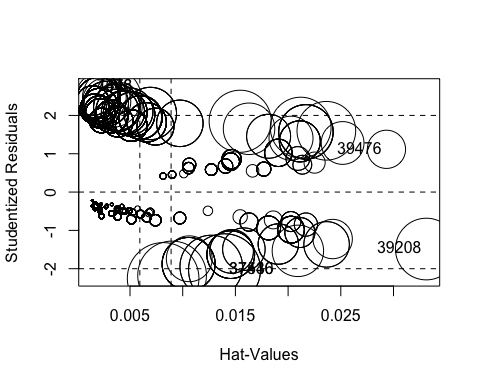
## Warning in plot.xy(xy, type, ...): "id.n" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not  
## a graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not  
## a graphical parameter

## Warning in box(...): "id.n" is not a graphical parameter

## Warning in title(...): "id.n" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a  
## graphical parameter



## StudRes Hat CookD  
## 7446 2.757323 0.001401683 0.005424479  
## 4678 2.757323 0.001401683 0.005424479  
## 39208 -1.486823 0.033113887 0.006190260  
## 37440 -2.027270 0.013967882 0.008433304  
## 39476 1.115459 0.029331514 0.002375588  
## 37536 -2.027270 0.013967882 0.008433304

A partir del gràfic observat a priori es pot veure que les dades més influents són les “39208” i “39476” observant el leverage que hi ha en el plot corresponent.

# Predicció

## WORK

pre1<-**predict**(gm14,type="response")  
pn<- **as.numeric**(pre1)  
**summary**(df**$**y)

## Y\_no Y\_yes   
## 4349 597

pre.y <- **factor**(**ifelse**(pn**<**0.5,0,1),labels=**c**("pre.Success?-no","pre.Success?-yes"))  
  
tt<-**table**(pre.y,dfw**$**y);tt

##   
## pre.y Y\_no Y\_yes  
## pre.Success?-no 3224 353  
## pre.Success?-yes 29 103

100**\*sum**(**diag**(tt))**/sum**(tt)

## [1] 89.70073

## TEST

pre<-**predict**(gm14,type="response",newdata=dft)  
pn<- **as.numeric**(pre)  
**summary**(df**$**y)

## Y\_no Y\_yes   
## 4349 597

pre.y <- **factor**(**ifelse**(pn**<**0.5,0,1),labels=**c**("pre.Success?-no","pre.Success?-yes"))  
  
tt<-**table**(pre.y,dft**$**y);tt

##   
## pre.y Y\_no Y\_yes  
## pre.Success?-no 1086 116  
## pre.Success?-yes 10 25

100**\*sum**(**diag**(tt))**/sum**(tt)

## [1] 89.81407

En aquest apartat hem realitzat les prediccions per tal de veure les taxes d’encert del nostre model. Tenim una taxa d’encert del 89.814%.

Ara tenim una altra manera de calcular la predicció:

**library**("ROCR")

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

dadesroc<-**prediction**(**predict**(gm14,type="response"),dfw**$**y)   
**par**(mfrow=**c**(1,2))  
**plot**(**performance**(dadesroc,"err"))  
**plot**(**performance**(dadesroc,"tpr","fpr")) **>** **abline**(0,1,lty=2)

