BANK MARKETING DATA: CASE STUDY

ANÀLISI DE DADES I EXPLOTACIÓ DE LA INFORMACIÓ

Montserrat Martinez
Aleix Costa

ÍNDEX

- Data Processing, Description, Validation and Profiling
- PCA & Clustering
- CA & Clustering
- Forecasting modeling of the numeric target
- Forecasting modeling of the categorical target

Data Processing, Description, Validation and Profiling

DATA DESCRIPTION

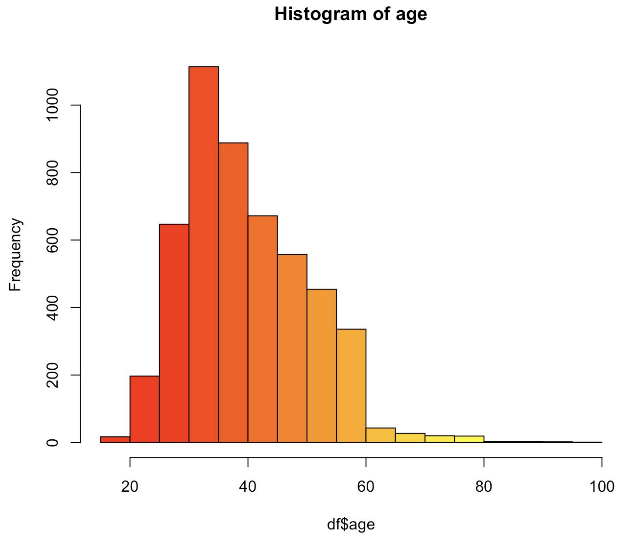
- Registre de trucades telefòniques d’un banc a diferents possibles clients
- Files de la mostra aleatòria: 5000 trucades
- Columnes de la mostra aleatòria: 21 variables
- 11 variables qualitatives
- 10 variables quantitatives
- Target numèric = variable “duration”
- Target categòric = variable “y”

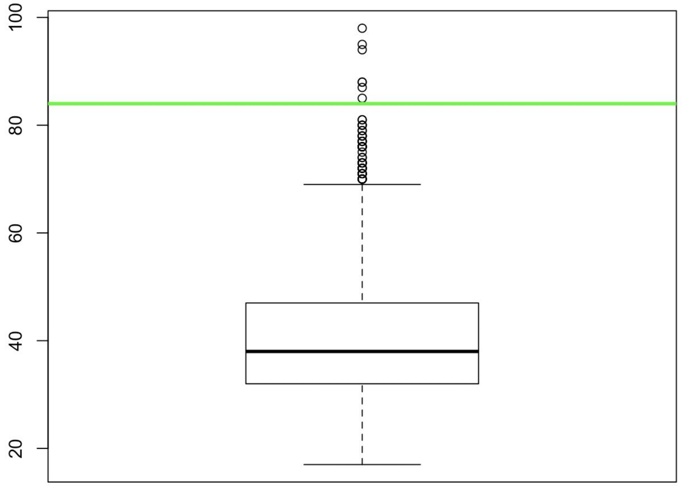
DATA DESCRIPTION

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

QUANTITATIVE VARIABLES

- S'ha d'analitzar una a una totes les variables numèriques
- Detectem els missing i els errors
- Detectem els outliers



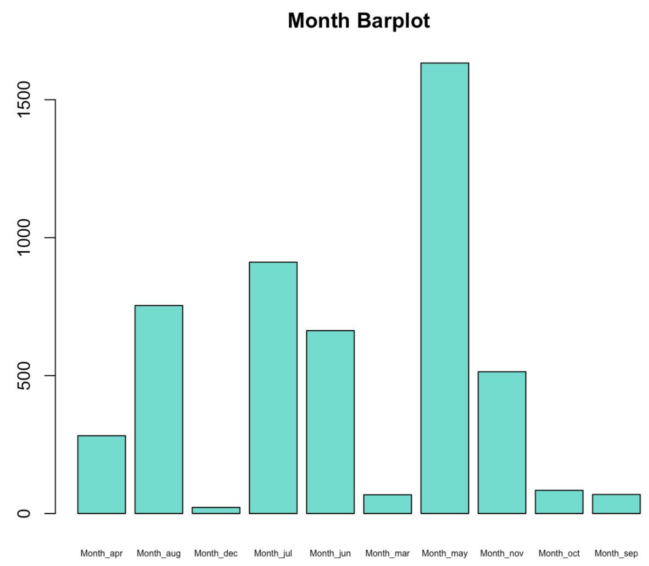


QUALITATIVE VARIABLES

- S’ha d’analitzar una a una totes les variables qualitatives
- Identificació dels missings (NA’s)
- Identificació dels errors

```
> summary(df$month)
Month_apr Month_aug Month_dec Month_jul Month_jun Month_mar Month_may Month_nov Month_oct Month_sep
      282       754        22       911       663        68      1633       514        84        69
```

```
> summary(df$job)
      Job_admin.   Job_blue-collar  Job_entrepreneur     Job_housemaid    Job_management       Job_retired Job_self-employed
            1315              1157               161               128               348               212               155
    Job_services       Job_student    Job_technician    Job_unemployed       Job_unknown
             477               105               789               108                45
```



IMPUTATION

Ara imputarem tots els NA's que tenim:

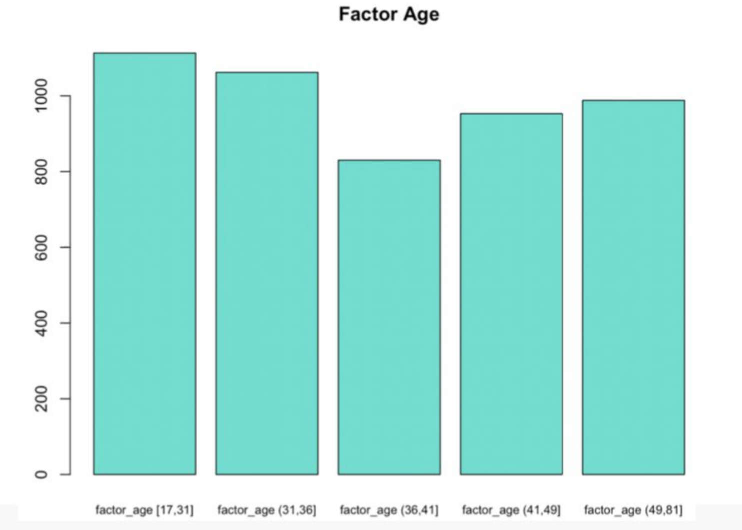
```
> summary(df$pdays)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   3.000   5.000   5.821   6.000  20.000    4816
```

```
> summary(df$pdays)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   16.00   16.00   15.63   16.00   16.00
```

```
table(df$pdays)
summary(df$pdays)
sel <- which(is.na(df$pdays))
sel
length(sel)
df[sel, "pdays"] <- 16
table(df$pdays)
summary(df$pdays)
hist(df$pdays, 10, main = "Pdays Histogram", col = "turquoise")
```

DISCRETIZATION

- Discretització de les variables numèriques
- Convertir a factors els diferents rangs de variables
- Tenir les dades ordenades segons intervals




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

PROFILING

Target "duration"

```
condes(df, which(names(df) == "duration"))

## $quanti
##                correlation      p.value
## pdays           0.52693895 0.000000e+00
## previous        0.02859224 4.435374e-02
## errors_indiv   -0.03476735 1.447588e-02
## nr.employed    -0.03619203 1.091224e-02
## campaign       -0.04179341 3.284450e-03
## missings_indiv -0.07328498 2.474678e-07
##
## $quali
##                                 R2       p.value
## factor_duration       0.8271873066  0.000000e+00
## factor_Pdays          0.4046346310  0.000000e+00
## y                     0.1863696068 9.891372e-224
## poutcome              0.0041874670  3.132625e-05
## month                 0.0073478185  3.327154e-05
## factor_cons.price.idx 0.0039803615  5.696640e-04
## factor_Previous       0.0019228074  2.038492e-03
## day_of_week           0.0029955473  5.075577e-03
## factor_cons.conf.idx  0.0026002247  1.194404e-02
## contact               0.0011105265  1.909343e-02
## default               0.0009897216  2.693284e-02
## factor_campaign       0.0013152237  3.866909e-02
```

Target "y"

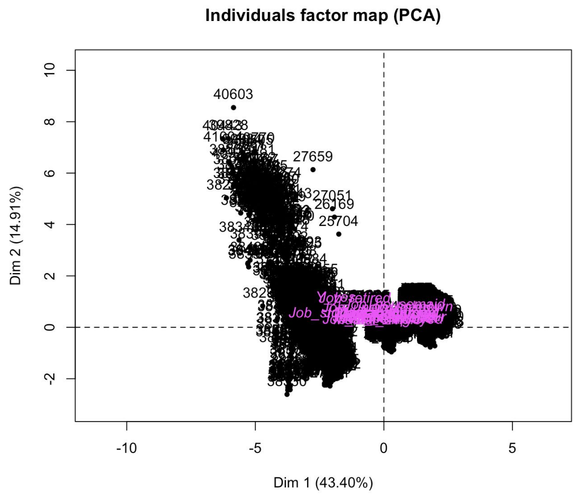
```
> catdes(df_catdes,21)

Link between the cluster variab
===============================
                 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
```

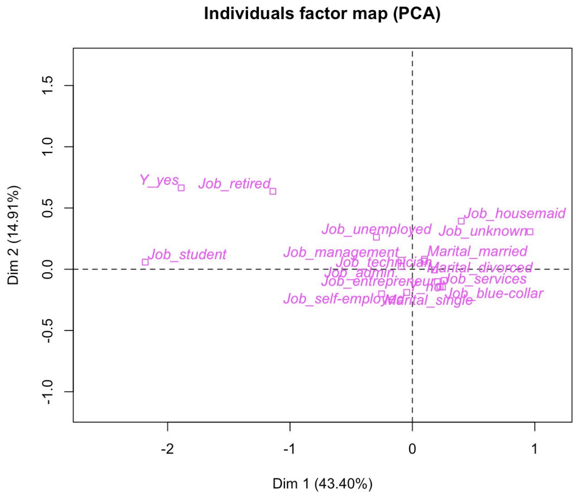
PCA & Clustering

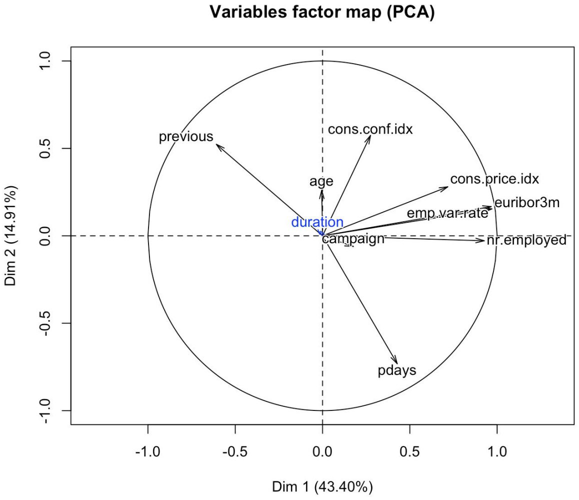
PCA

Creació PCA:

```r
res.pca<-PCA(df[,c("duration","y","marital","job",vars_conaux)],quanti.sup = 1,quali.sup = 2:4)
#LES VARIABLES ACTIVES NO PODEN SER FACTORS!
```







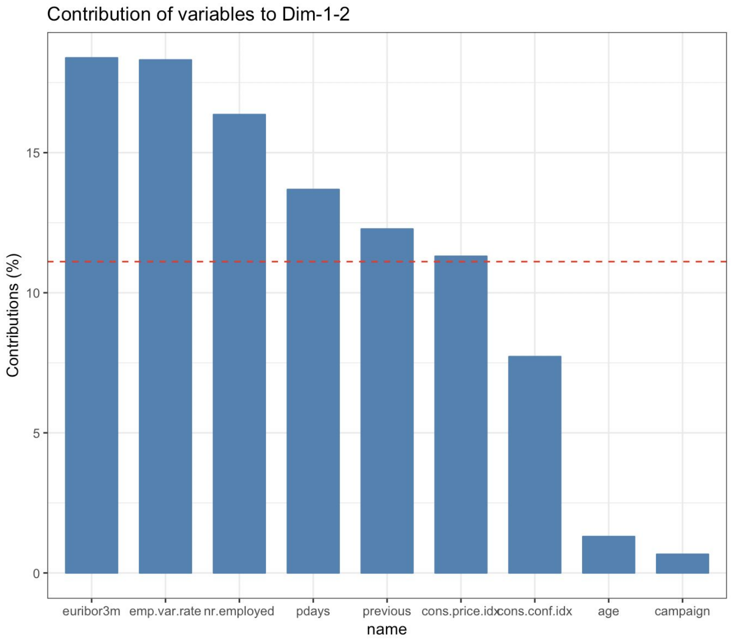
KAISER RULE

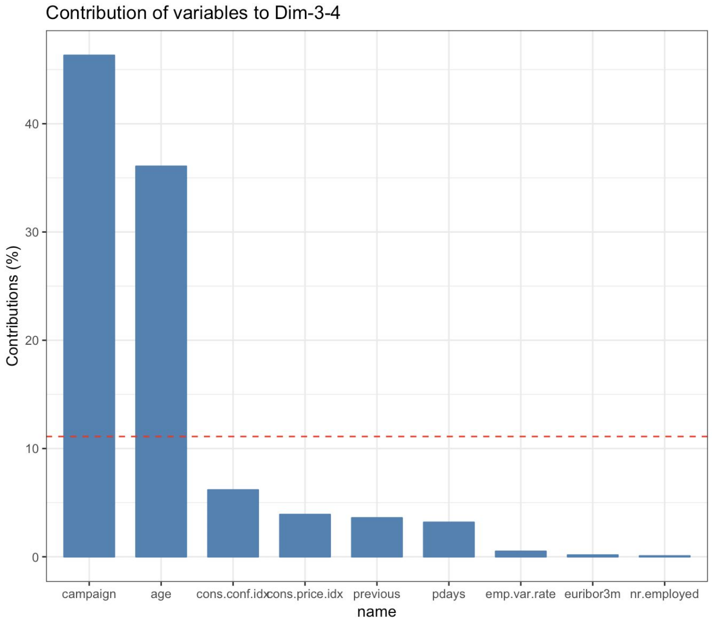
A partir de la taula de valors propis i seguint la regla de Kaiser hem decidit tenir en compte les 4 primeres components principals. Agafant les 4 components es representen més de tres quarts de les nostres dades (80.719).

```
> 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
```

```r
fviz_contrib(res.pca, choice = "var", axes = 1:2)+theme_bw()
fviz_contrib(res.pca, choice = "var", axes = 3:4)+theme_bw()
```

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

dimdesc(res.pca, axes = 1:4)
```





Clustering

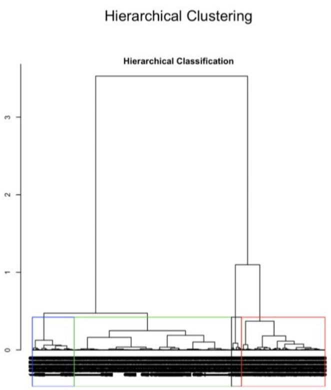
```
#Set clusters m'expliquen una mica mes d'un 80% de l'informacio, es la qualitat de la
representacio
info<-kcla$betweenss/kcla$totss
info

## [1] 0.8059886
```

Per regla general s'han de tenir més de 6 clusters i després de l'estudi, comprovem que amb 7 clusters tenim més d'un 80% d'informació representada.

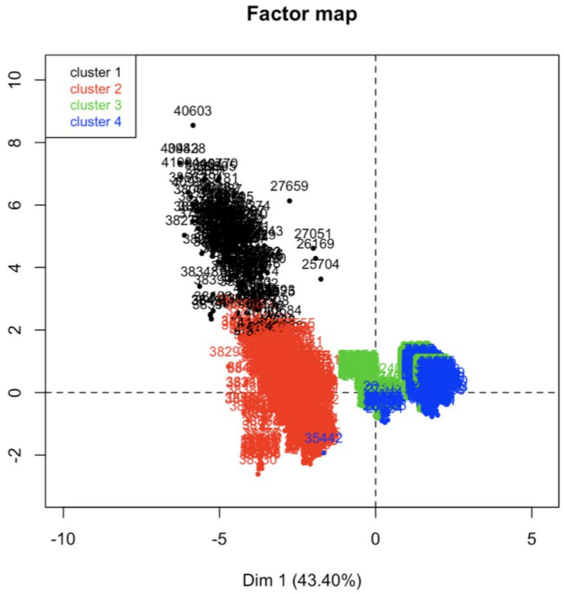
CA & Clustering

Hierarchical Clustering

```
# 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
```




```
## $`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
```

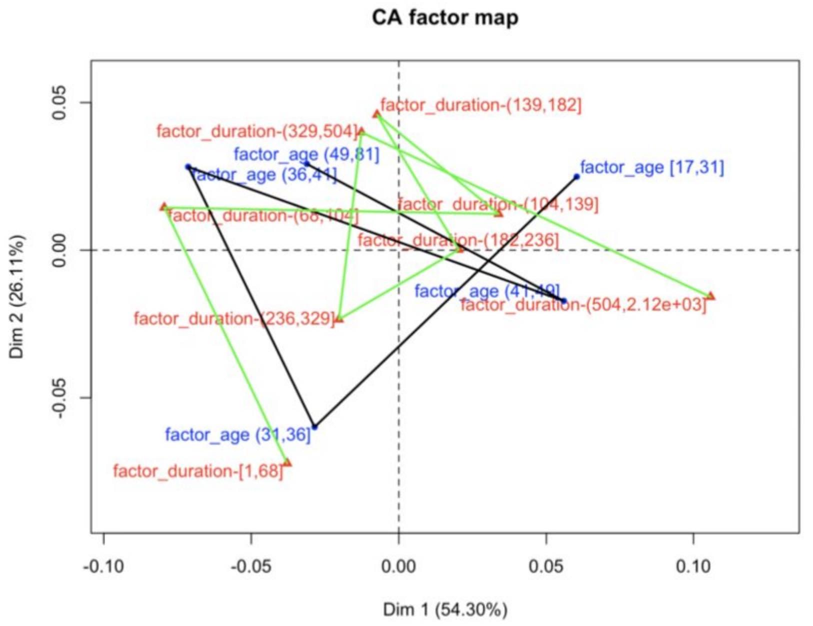


Correspondence Analysis (CA)

Analitza les relacions entre 2 factors de les dades de la nostra mostra.

- Factor_age & Factor_duration

```
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
```

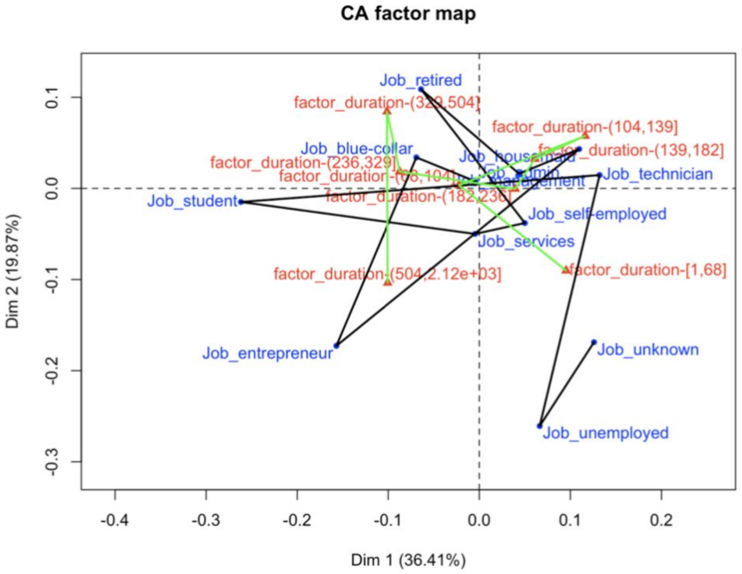


Podem veure que la durada de la trucada no depèn de l’edat del nostre individu.

Correspondence Analysis (CA)

- Job & Factor_duration

```
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
```



p-value molt proper al 5%, llavors es pot rebutjar la HO, llavors la durada de la trucada si que pot dependre del treball o a que es dediqui el nostre individu.

Forecasting modeling of numeric target

Model construction only with numeric as explanatory variables

```
> vars_model<-names(df)[c(1,11:14,16:20)]; vars_model
 [1] "age"           "duration"      "campaign"      "pdays"         "previous"      "emp.var.rate"   "cons.price.idx"
 [8] "cons.conf.idx"  "euribor3m"     "nr.employed"
> condes(df[,vars_model],which(vars_model == "duration"))
$quanti
             correlation      p.value
previous      0.02859224 4.435374e-02
nr.employed -0.03619203 1.091224e-02
campaign    -0.04179341 3.284450e-03
pdays       -0.06147234 1.516945e-05
```

```
m1<-lm(duration~previous+campaign+pdays+nr.employed,data=df)
```

- Anova (model) : Agafar variables significatives (*)
- Previous poc significativa
- nr.employed = vif > 3

```
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
```

TRANSFORMING VARIABLES

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

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.

```
> 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
```

Model construction only with factors as explanatory variables

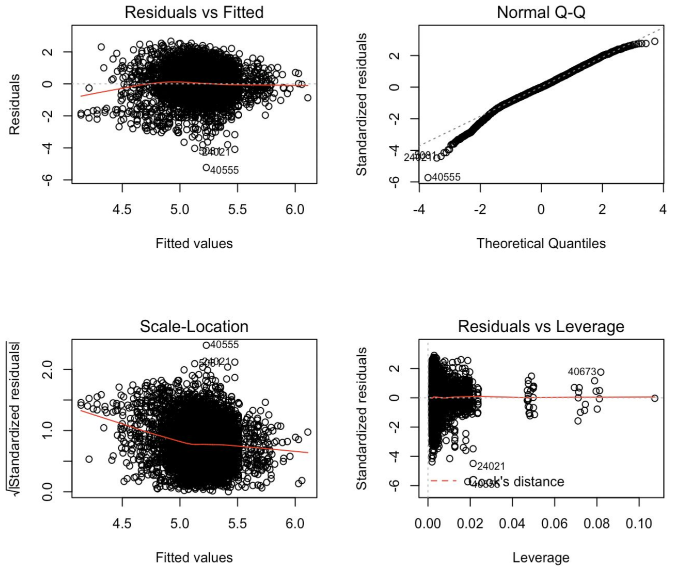
```
> vars_dis2<-names(df)[c(2:10,15,25,26:35)];vars_dis2
 [1] "job"               "marital"            "education"          "default"            "housing"
 [6] "loan"              "contact"            "month"              "day_of_week"        "poutcome"
[11] "season"            "factor_age"         "factor_duration"    "factor_campaign"    "factor_Pdays"
[16] "factor_Previous"   "factor_emp.var.rate" "factor_cons.price.idx" "factor_cons.conf.idx" "factor_euribor3m"
[21] "factor_nr.employed"
```

Anova -> Neteja efectes nets i variables significatives

```
#Our model
m12<-lm(log(duration)~campaign+pdays+poutcome+month+factor_cons.price.idx+day_of_week,data = df)
```

Estudi variables numèriques inicials si són millor com a factor o no : BIC (<) entre models

```
#Best solution:
m13<-lm(log(duration)~campaign+factor_Pdays+poutcome+month+factor_cons.price.idx+day_of_week,data = df)
```



Forecasting modeling of the categorical target

WORK AND TEST SAMPLES

```r
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
catdes(dfw[,c("y",vars_con)],1) #Numericas relacionadas
```

work sample -> treball amb les dades + creació de models

test sample -> predicció

Model construction only with numeric explanatory variables

```
> vars_con
 [1] "age"            "duration"       "campaign"       "pdays"          "previous"       "emp.var.rate"   "cons.price.idx"
 [8] "cons.conf.idx"  "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
```

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

MODEL MÉS CORRECTE

```
> 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
```

Model construction with factors as explanatory variables

```
gm10<-glm(y~pdays+poly(previous,2)+cons.price.idx+campaign,family=binomial,data = dfw)
```

Fem les comprovacions pertinents amb el BIC, per comprovar si explica més com a factor o com a numèrica, llavors obtenim:

```
## MILLOR MODEL FINS ARA:
gm11<-glm(y~factor_Pdays+factor_Previous+factor_cons.price.idx+factor_campaign,family=binomial,data = dfw)
```

S'afegeixen les noves variables factors que siguin més explicatives -> CATDES i aconseguim:

```
gm12<-glm(y~factor_Pdays+factor_Previous+factor_cons.price.idx+factor_campaign+poutcome+month+job+season+default+education,fam
ily=binomial,data = dfw)
```

VALIDACIÓ:

```
> 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
```

PREDICTIONS

```
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
```

In this section we have made the predictions to see the success rates of our model and we can see that we have a hit rate around 90%

TOTAL SUCCESS = 89.814%