Análisis de datos y explotación de la información

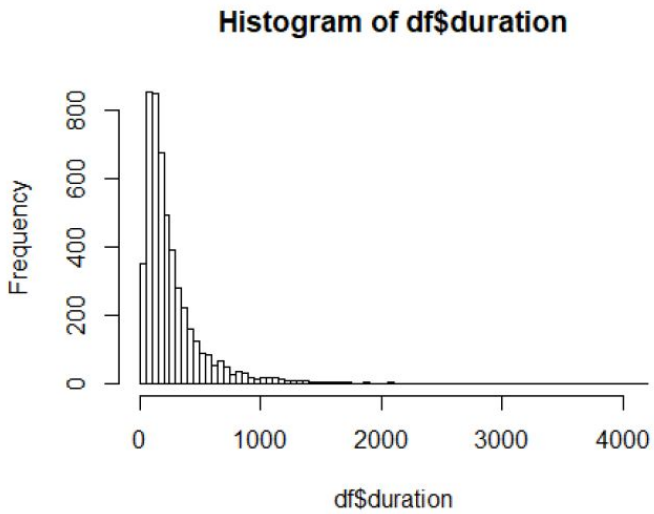
Guillem Valls
Sergio Mazzariol

Preparación de la muestra

Deliverable 1

First steps

- ➔ Preparation of the sample.
- ➔ Initialize data and functions. Like vars_cat, vars_num, dqr, dqri, calcQ.
- ➔ Analysis and data exploration. With summary we see the amount of uknows.
- ➔ Treatment of the target variables.
  - ◆ “y” all
  - ◆ 5 < “duration” < 2079
    - 6 outliers



Tratamiento de variables no-Target Categóricas

Si son menos de 300 uknows, los pasamos a NA's. Luego realizamos la imputación con imputeMCA

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

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

Creación de nuevos niveles de los factores.

Agrupamos subcategorías en menos categorías.

➔ **Job:** La agrupamos por posible ingreso monetario.

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

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

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

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

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

Creación de nuevos niveles de los factores.

Agrupamos subcategorías en menos categorías.

➔ **Month**: En función de las temporadas aunque no tan estricto.

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

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

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

summary(df$f.season)
df$f.season<-factor(df$f.season,levels=1:3,labels=c("Mar-May","Jun-Aug","Sep-Dec"))
```

Creación de nuevos niveles de los factores.

Agrupamos subcategorías en menos categorías.

➔ **Education**: Nivel de estudio.

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

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

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

df$f.education<-factor(df$f.education,levels=1:3,labels=c("Basic","High School","Professional"))
```

Tratamiento de variable no-Target Numéricas.

- ➔ **Age:** edades entre 18-92
- ➔ **Campaing**: En 10 meses que dura la campaña 20 contactos, implica un contacto cada 15 días. No hay valores mayores a esto.
- ➔ **Pdays:** tomamos como missings los 999 y los ponemos al valor de nuestro máximo más 1
- ➔ **previous**: Nuestro máximo de veces contactado en campañas previas es 6.

Inconsistencias

- ➔ **pdays/previous/poutcome**. debería existir la relación directa entre **previous=0**, **outcome=nonexistent** y **pdays=999** por lo que podemos detectar errores. Al ver el resultado podemos decir que hay inconsistencias entre el pdays y previous, ya que todos los que son **pdays = 999**, deberían ser **previous = nonexistent**, lo que en este caso nos dan **526 individuos** que no cumplen esta condición. Suponen más de un 10% de la muestra.
- ➔ En los índices **trimestrales/mensuales**. Vemos que cada individuo tiene valores diferentes, esto puede ser porque los datos se han podido tomar en diferentes años, pero como no tenemos más información no podemos hacer nada más.

Imputación de variables númericas

Usamos el imputePCA, el cual nos da valores con decimales, los cuales hemos tenido que redondear. Al hacer un summary y comprobar los datos antes y después de la imputación todo parece estar bien.

Resumen del Data Quality Report y Ranking

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

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

Resumen del Data Quality Report y Ranking

Vemos un máximo de 3 missings en un individuo.

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

prop.table(table(dqri$missings))

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

Factorització de variables quantitatives

Age

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

Duration

```
##      [5,120]     (120,180]     (180,300] (300,2.1e+03]
##      1255           1240          1249           1246
```

Campaign

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

Pdays

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

Previous

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

Resultat del CONDES

```
## $quanti
##             correlation      p.value
## campaign -0.05940135 2.683764e-05
##
## $quali
##                        R2       p.value
## f.duration 0.621168787  0.000000e+00
## y          0.177066645 2.228224e-213
## f.campaign 0.003783221  7.858324e-05
## month      0.004450289  8.185248e-03
```

```
## $category
##                                Estimate       p.value
## f.duration-(300,2.1e+03]  310.35106  0.000000e+00
## y.yes                     170.13318 2.228224e-213
## f.campaign-(1,2]           23.01041  3.895001e-05
## month.apr                  35.25783  4.865526e-03
## f.season.Mar-May           13.19170  6.782891e-03
## month.aug                 -25.22225  7.943838e-03
```

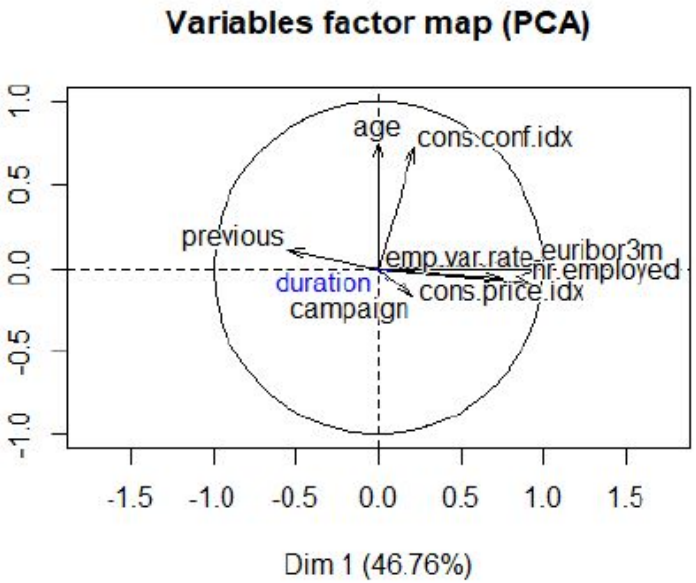
Resultat del CATDES

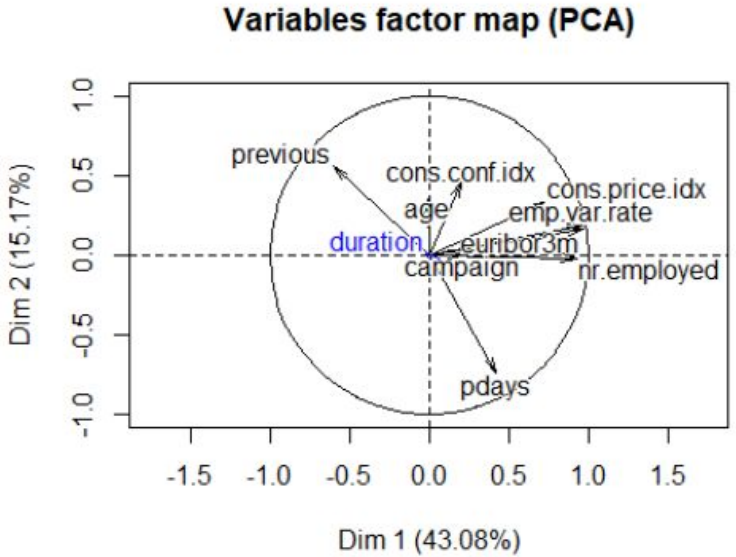
```
## $y.yes
##                                              Cla/Mod   Mod/Cla    Global
## f.duration=f.duration-(300,2.1e+03]        26.797386 68.081181 27.595190
## poutcome=poutcome.success                  66.025641 19.003690  3.126253
## f.pdays=f.pdays-[0,998]                    62.285714 20.110701  3.507014
## f.previous=f.previous-(1,6]                46.715328 11.808118  2.745491
## contact=contact.cellular                   13.768342 81.365314 64.188377
## default=default.no                         12.436548 90.405904 78.957916
## job=job.retired                            28.378378 11.623616  4.448898
## month=month.oct                            42.500000  6.273063  1.603206
## month=month.mar                            44.444444  5.166052  1.262525
## f.age=f.age-(50,92]                        16.780822 27.121771 17.555110
```

Deliverable 2

Valores propios y ejes dominantes

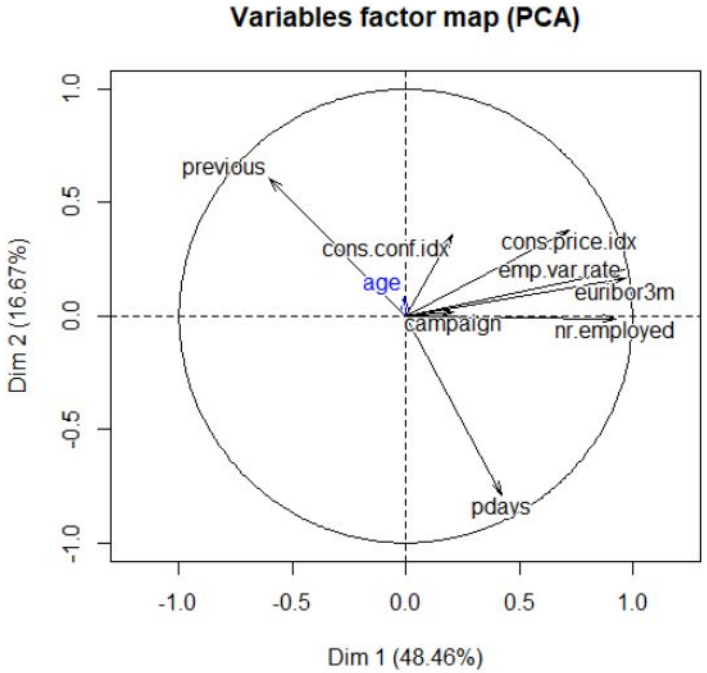
Primero decidimos probar como se ve el PCA con y son la variable pdays y duration.





Valores propios y ejes dominantes

Primero decidimos probar como se ve el PCA con y son la variable pdays y duration.



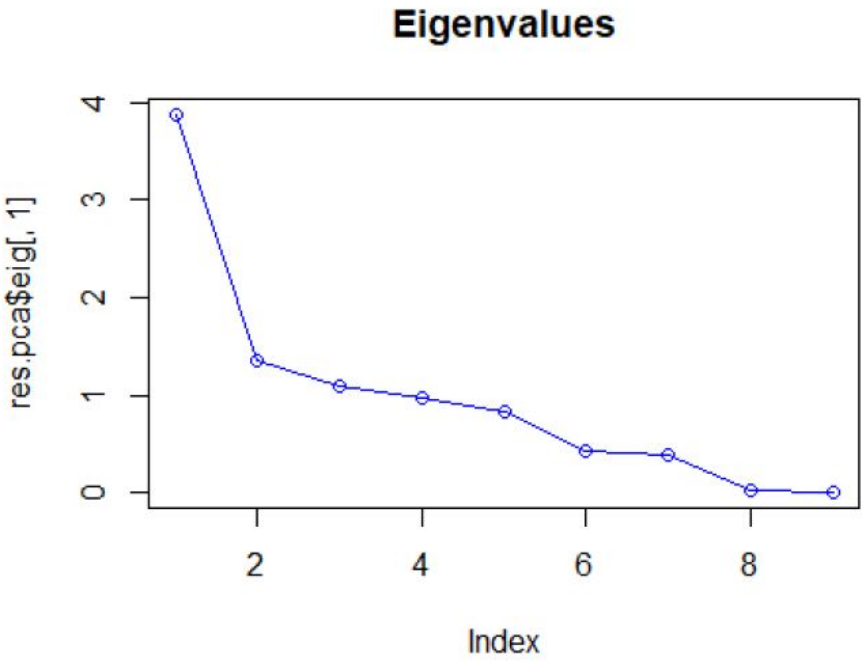
¿Cuántos ejes debemos interpretar de acuerdo con la regla de Kaiser y Elbow?

Por la ley de Kaiser, deberíamos utilizar los 3 primeros ejes factoriales, los cuales son mayores a 1.
Si tomamos en cuenta el criterio del 80% se deberían coger las 4 primeras dimensiones.
Para realizar el futuro análisis, conviene utilizar dimensiones pares, por lo que decidimos solo usar 4.

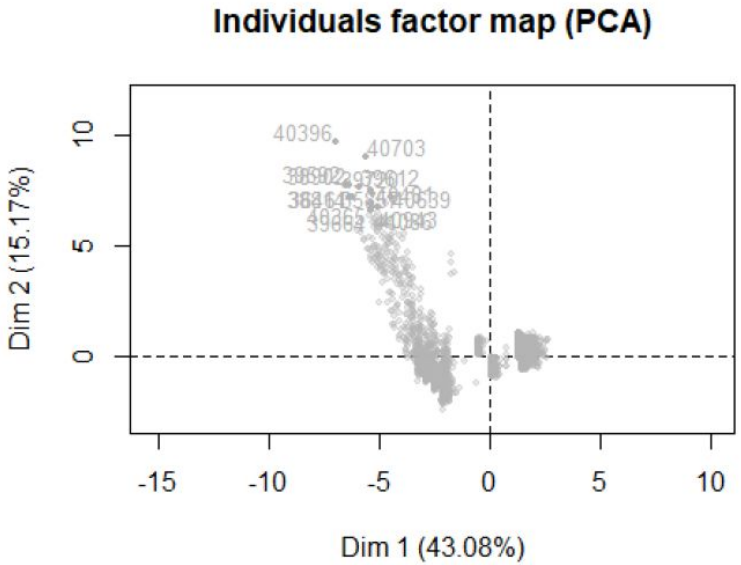
Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
Variance	3.88	1.36	1.10	0.97	0.83	0.43	0.39
% of var.	43.08	15.17	12.28	10.74	9.25	4.82	4.29
Cumulative % of var.	43.08	58.24	70.52	81.26	90.50	95.33	99.62

ElBow



¿Son algunos individuos "demasiado contributivos"?

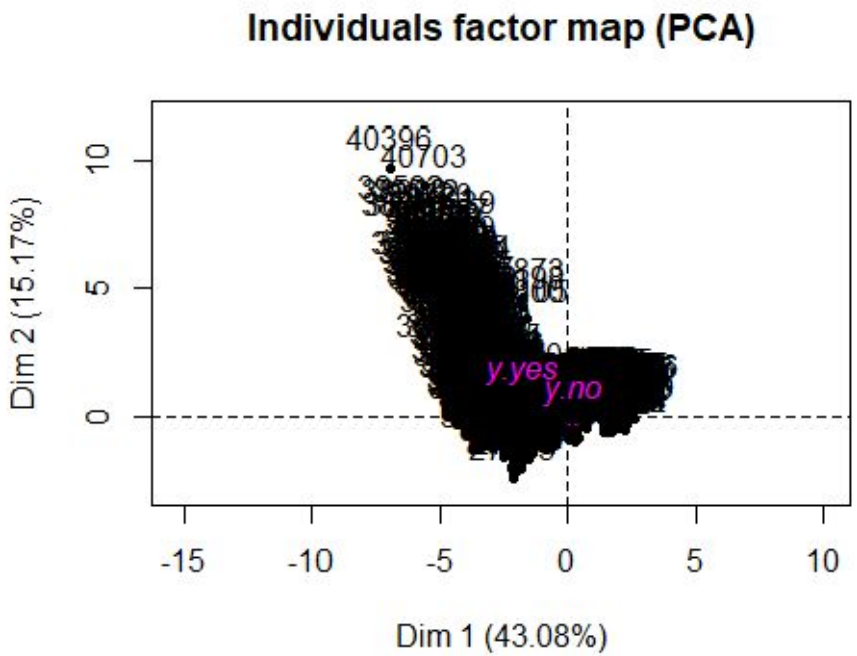


➔ Para los 5 individuos de la dimensión 1, vemos que principalmente son gente mayor de 45 años, todos han comprado el producto, han sido contactados mediante el móvil, han sido contactados previamente, comprado un producto en una campaña anterior y la duración de la llamada ha sido mayor a los 300s.

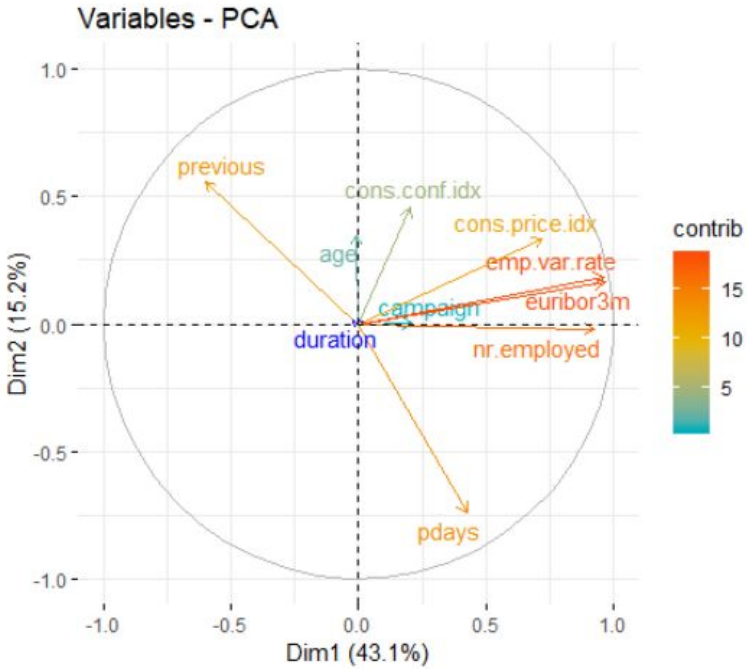
➔ Para la dimensión 2 podemos ver prácticamente las mismas características menos la duración que ha sido menor.

Interpretando los ejes

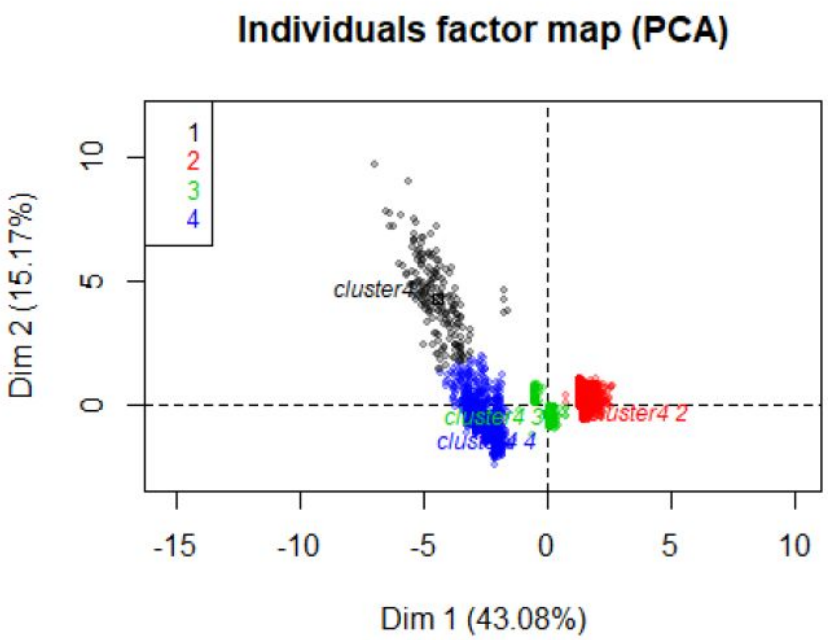
Al hacer el PCA con la variable target Y como suplementaria, podemos ver que en el gráfico de rp, el factor NO, está muy cerca del centro, por lo que no se ve representada en estos ejes factoriales. En cambio el factor SI, está a una distancia mayor del centro, aunque poco significativa.

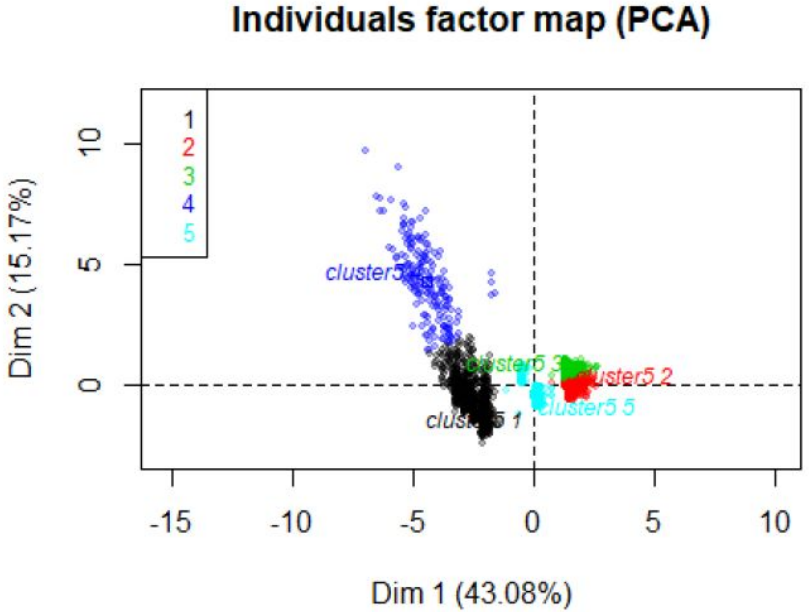


Contribución

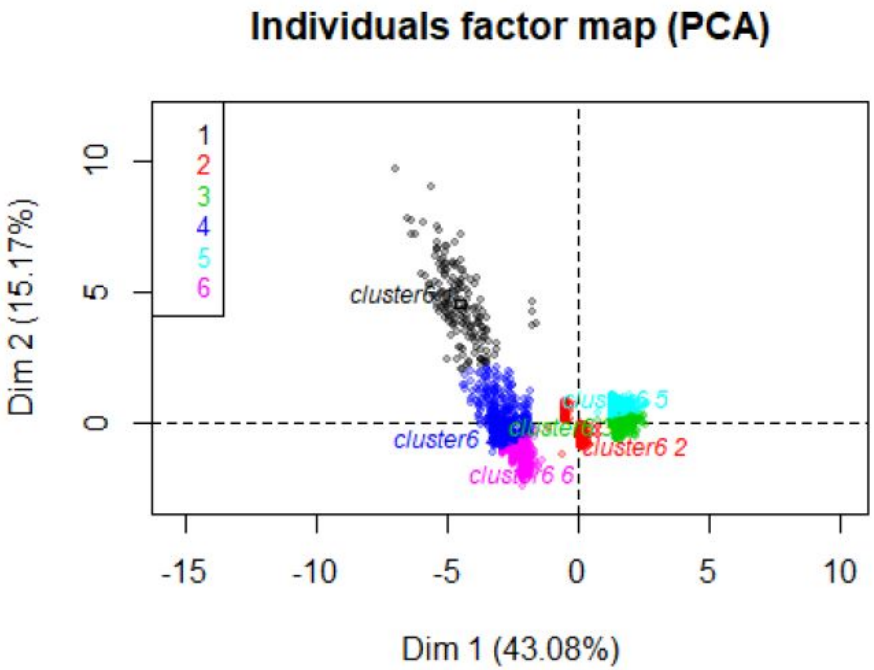


Clasificación de K-means





Clasificación de K-means



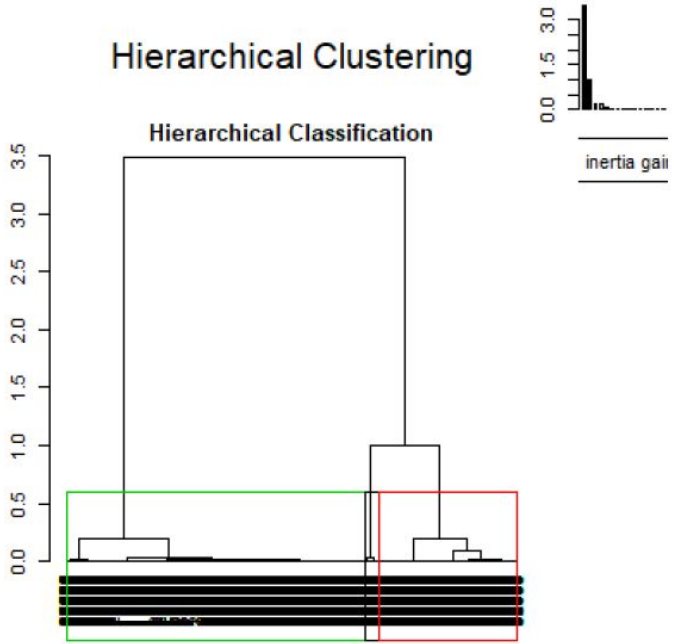
Descripción de los clusters

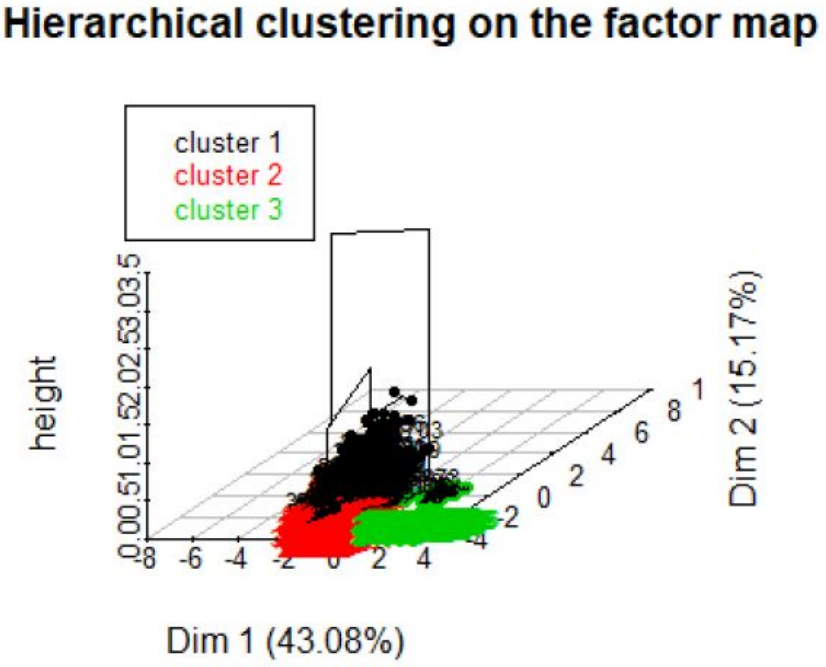
- **Cluster 1:** Meses de jun-Ago, No han sido contactados previamente, Han sido contactados más de una vez en la campaña actual, Contactados por teléfono fijo, Trabajo normalmente es, servicio, técnicos o blue collar, No compraron el producto en su mayoría.
- **Cluster 2:** La temporada de mar-may están sobrerrepresentadas en este cluster, Han sido contactados en su mayoría por teléfono móvil, No han comprado el producto en campañas anteriores, Han sido contactados en campañas previas, La categoría student está sobrerrepresentada, La aceptación del producto está sobrerrepresentada.

Descripción de los clusters

- ➔ **Cluster 3:** Temporada de Sep-Dec, Contactados por móvil en su mayoría, La categoría de job.management está sobrerepresentada, No han adquirido el producto, Una gran cantidad de individuos rechazó el producto (y.no)
- ➔ **Cluster 4:** Han sido contactados previamente f.pdays[0,22], Han comprado el producto en una campaña previa, Han comprado el producto y.yes, Temporada de Sep-Dec, Han sido contactados por móvil, Una parte importante son job.retired, Una edad de f.age-(50,92]

Agrupación jerárquica





Agrupación jerárquica

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

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

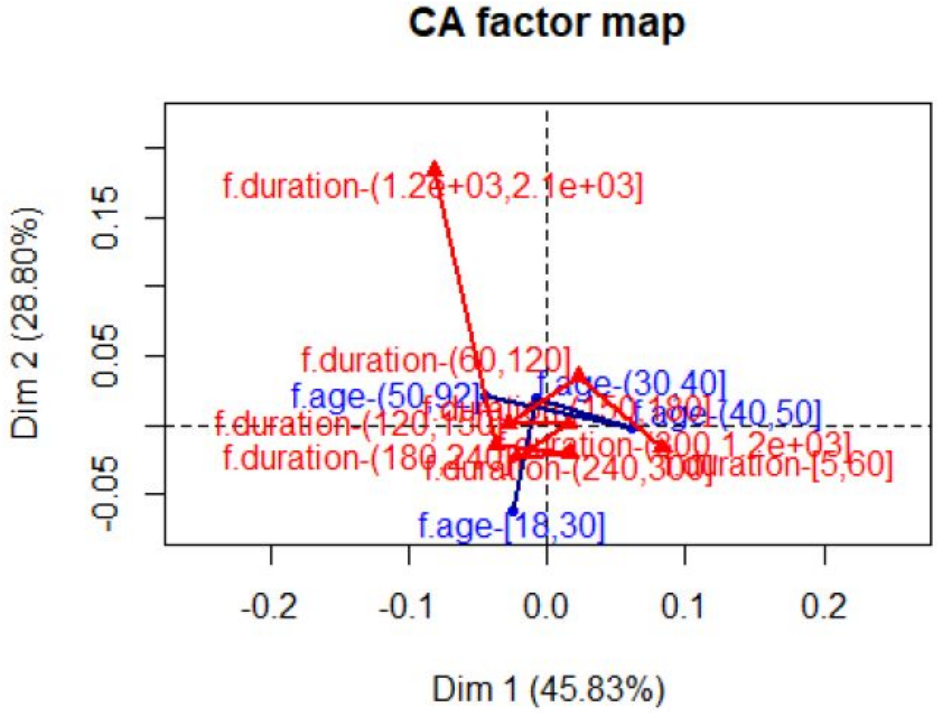
Cluster 3: Han sido contactados previamente, No han sido contactados en campañas previas, Han sido contactados en la temporada de f.season.Jun-Aug, Han rechazado el producto, Tienen una leve representación de f.job.Serv-Tech-BlueC.

Análisis de CA

Para experimentar y para que tenga más sentido el análisis de correspondencias, refactorizaremos a 8 niveles la variable duration. Con esta nueva variable factorizada y f.age hacemos análisis de correspondencias.

Para saber cuantas dimensiones debemos considerar, obtenemos la media de los eigenvalues. Vemos que solamente tiene sentido considerar el primer eje, ya que este es el único valor mayor a la media (kaiser).

Análisis de CA

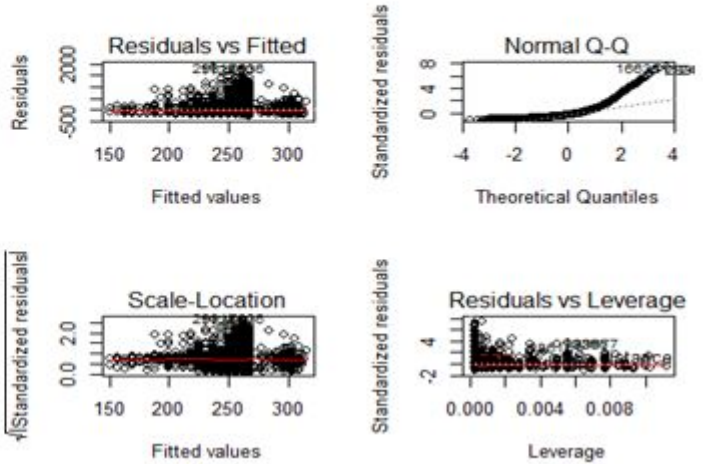


Deliverable 3

Modelització amb target numèric

Model simple

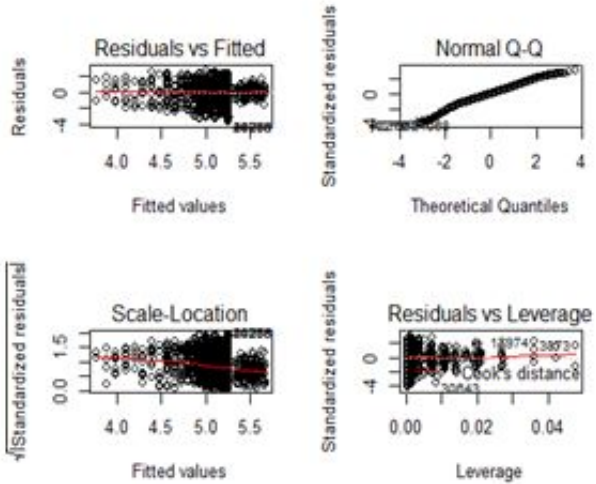
```
m3<-lm(duration~campaign+pdays,data=df)

Anova(m3)
##              Sum Sq   Df F value    Pr(>F)
## campaign    1058016    1 16.7722 4.281e-05 ***
## pdays        208524    1  3.3056    0.0691 .

vif(m3)
## campaign    pdays
## 1.003138 1.003138
```



Modelització amb target numèric

Model transformat

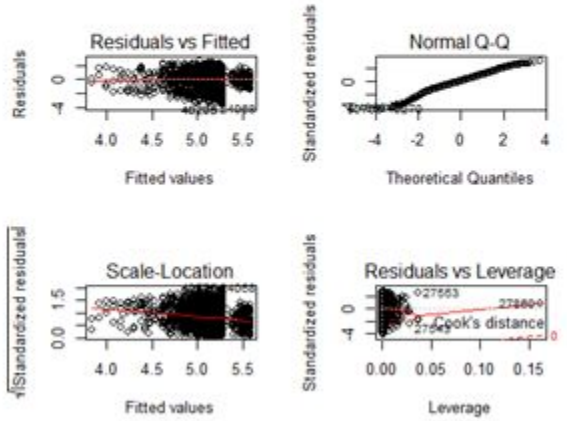
```
m20<-lm(log(duration)~poly(campaign,2)+poly(pdays,2),data=df)
summary(m20)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.17868    0.01274 406.451  < 2e-16 ***
## poly(campaign, 2)1 -10.03807    0.90154 -11.134  < 2e-16 ***
## poly(campaign, 2)2  -1.79572    0.90036  -1.994 0.046158 *
## poly(pdays, 2)1     -3.34605    0.90176  -3.711 0.000209 ***
## poly(pdays, 2)2     -1.90923    0.90014  -2.121 0.033968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9 on 4985 degrees of freedom
## Multiple R-squared:  0.02951,    Adjusted R-squared:  0.02873
## F-statistic: 37.89 on 4 and 4985 DF,  p-value: < 2.2e-16
```



Modelització amb target numèric

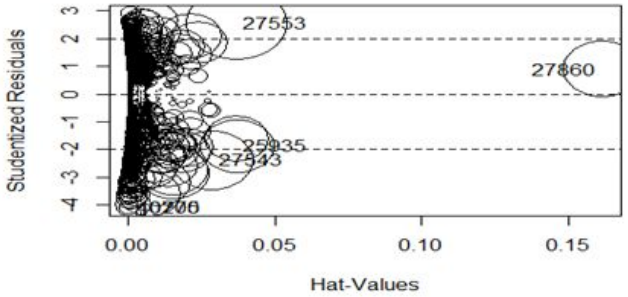
Model amb factors

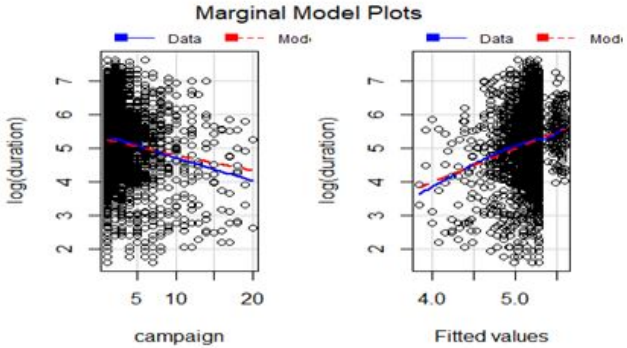
```
m31<-lm(log(duration)~(f.influentMonth*campaign+f.pdays),data=df)
Anova(m31)
## Anova Table (Type II tests)
##
## Response: log(duration)
##                          Sum Sq   Df  F value    Pr(>F)
## f.influentMonth             8.4    2   5.1981  0.005557 **
## campaign                  102.8    1 127.0831 < 2.2e-16 ***
## f.pdays                    15.0    1  18.5535 1.684e-05 ***
## f.influentMonth:campaign    5.0    2   3.0728  0.046377 *
## Residuals                4028.9 4983
```

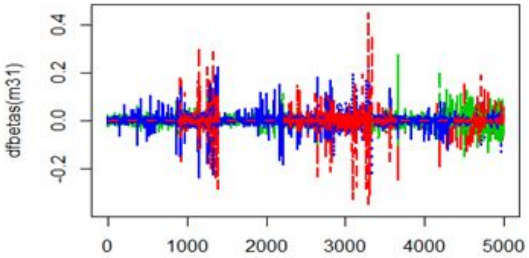


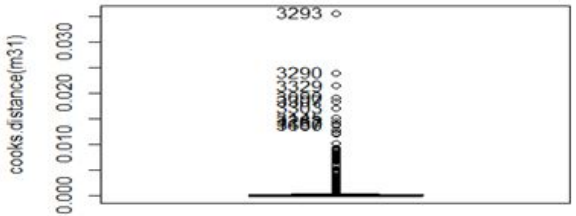
Modelització amb target numèric

Validació





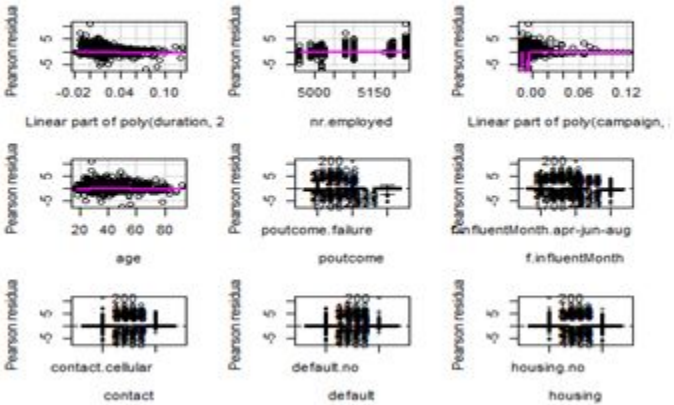




Modelització amb target binari

Model final

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

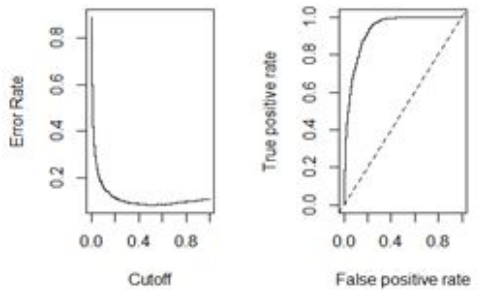


Modelització amb target binari

Capacitat predictiva

```
p<-factor(ifelse(
      predict(gm21, dft, type = "response") < 0.4, 0, 1 ))
tabConfusion<-table(p, dft[,"y"])
p    y.no y.yes
  0 1062    51
  1   56    79

capacidadPredictiva     capacidadPredictivaNull
[1] 0.9142628            [1] 0.8958333
```

```
MejoraModelo <- capacidadPredictiva -
                                capacidadPredictivaNull
MejoraModelo*100
## [1] 1.842949
```



FI