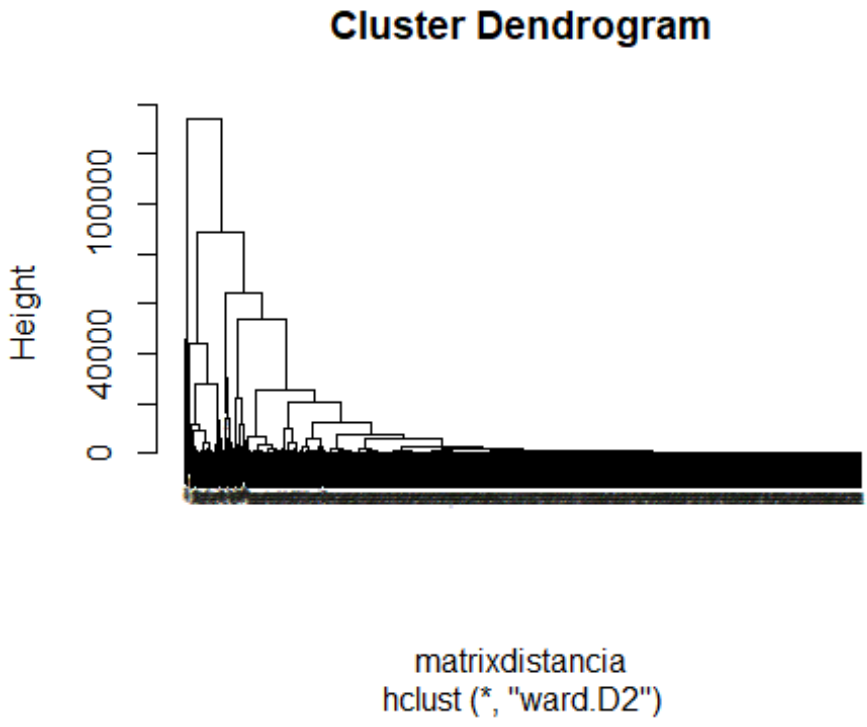
Exercise Clustering 15-01-2020

Daniel Ferreira Zanchetta and Lais Silva Almeida Zanchetta

1. Con las componentes principales halladas significativas, efectúe una Clasificación Ascendente Jerárquica por el método de Ward. Explique en qué consiste el método de agregación de Ward?. Represente el dendrograma (o árbol jerárquico) obtenido.

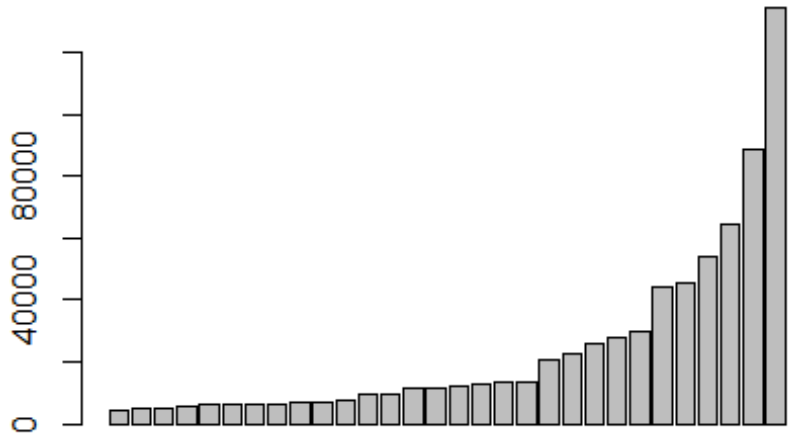
```
nd <- 3
Psi <- pca.churn$ind$coord[,1:nd]

matrixdistancia <- dist(Psi)

cluster.churn <- hclust(matrixdistancia,method="ward.D2")

plot(cluster.churn,cex=0.2)
```



Resp.: El método de agregación Ward consiste en calcular la distancia entre (dos) centros gravitaciones evitando la perdida de información que esta agregación pueda generar.

2. A la vista del diagrama de barras del índice de nivel de las últimas agregaciones efectuadas, decida el número de clases de clientes diferentes que existen en los datos analizados.

```
#Barplot con los individuos activos
barplot(cluster.churn$height[(nrow(Psi)-30):(nrow(Psi)-1)])
```



Resp.: Basados en el diagrama de barras, hemos decidido que el número de clases de cliente diferentes que existen en los datos analizados seria de 3 clases.

3. Obtenga la partición del árbol jerárquico en el número de clases finales deseado. Diga el número de clientes por "cluster", y calcule el centro de gravedad de los clusters obtenidos.

```
nc <- 3
# Corte del árbol considerando el número de clases obtenidas en el
ejercicio anterior
arbol.clas3 <- cutree(cluster.churn,nc)

# Número de clientes por clase ("cluster")
numclient.class <- table(arbol.clas3)
numclient.class
```

```
## arbol.clas3
##   1   2   3
## 939  54   7

# Centro de gravedad de los clusteres obtenidos
cdg <- aggregate(as.data.frame(Psi),list(arbol.clas3),mean)[,2:(nd+1)]
cdg

##        Dim.1     Dim.2      Dim.3
## 1  -242.7778 -476.9002   1.757148
## 2  -414.5576 8262.4762 -54.573789
## 3 35764.9272  233.6518 185.288991

# Calidad del árbol jerárquico

Bss <- sum(rowSums(cdg^2)*as.numeric(table(arbol.clas3)))
Tss <- sum(Psi^2)

100*Bss/Tss

## [1] 61.64111
```

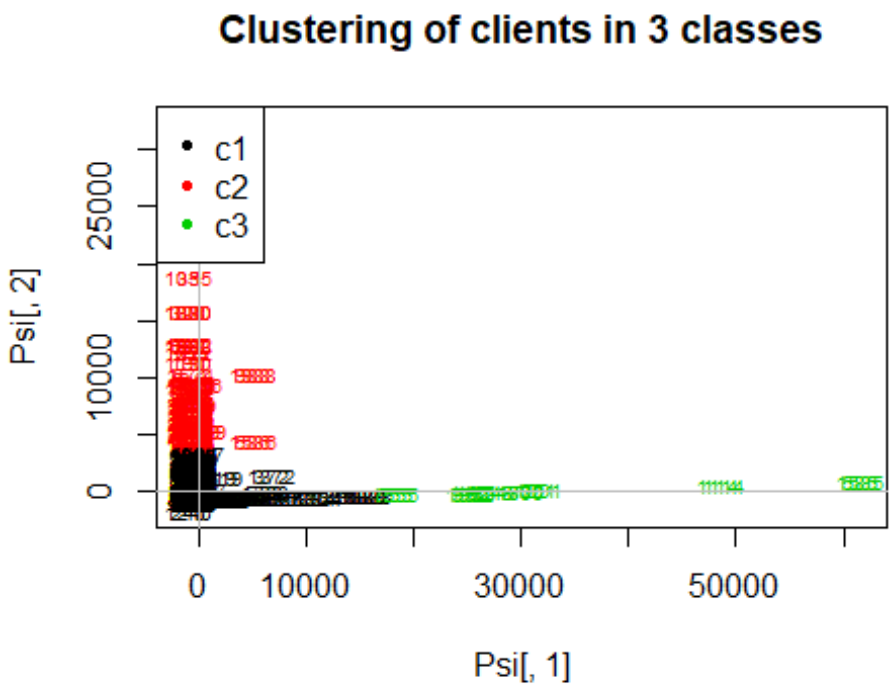
4. En qué consiste la operación de consolidación de una partición obtenida por corte del árbol jerárquico. Efectúe esta operación en la partición obtenida en el apartado 3 anterior. Diga el número de clientes en las clases finales obtenidas.

```
# Consolidación con centros iniciales en los centroides
consol.kmeans <- kmeans(Psi,centers=cdg)

# Número de clientes por clase final obtenida
consol.kmeans$size

## [1] 937  55   8
```

Resp.: Consiste en identificar/ reevaluar los individuos para agruparlos en el centro de gravedad que esté más cercano. Así se hace una consolidación. Este proceso es para solventar el problema de valores solapantes.

5. Represente la partición final obtenida en el primer gráfico factorial, distinguiendo con colores diferentes cada una de las clases de clientes detectados.

```
plot(Psi[,1],Psi[,2],type="n",main="Clustering of clients in 3 classes")
text(Psi[,1],Psi[,2],col=consol.kmeans$cluster,labels=row.names(churn_tid
y),cex = 0.6)
abline(h=0,v=0,col="gray")
legend("topleft",c("c1","c2","c3"),pch=20,col=c(1:nc))
```



6. Interpretamos las clases finales obtenidas. Para ello utilizamos la función "catdes" de R. Primero damos las características significativas de cada clase (identificada como "1" la primera clase por ejemplo) para las variables continuas (=quanti) (por ejemplo quanti$´1´ se refiere a las características significativas de las variables continuas en la primera clase). Después aparecen las modalidades (=category) significativas de las variables categóricas en cada una de las clases. Interprete y de un nombre a cada una de los tipos de cliente identificados.

```
# Descripción de las 3 clases finales obtenidas
library(FactoMineR)
result <-
catdes(cbind(as.factor(consol.kmeans$cluster),churn_tidy),num.var=1)

## Warning in data.frame(..., check.names = FALSE): row names were found
from
## a short variable and have been discarded

# Caracteristicas significativas de cada una de las 3 clases
result$quanti

## $`1`
##                      v.test Mean in category Overall mean sd in
```

```
category
## dif_Libreta      -2.059236       -53.755683    -41.93691
936.5669
## dif_Largo_plazo  -3.945021         6.185454     26.10933
810.0829
## dif_Fondos_inv   -5.831256       171.973458    261.78000
1683.4009
## Total_Vista      -6.081244       531.488260    569.17250
991.5231
## Total_Inversion  -6.636977       708.235326    853.11600
3006.7650
## dif_Plazo       -10.766110         7.353319    114.89096
1374.6009
## Total_Plazo     -12.538430      1063.713981   1332.66300
3344.2034
##                  Overall sd      p.value
## dif_Libreta       989.6283 3.947160e-02
## dif_Largo_plazo   870.8238 7.979286e-05
## dif_Fondos_inv   2655.5333 5.501159e-09
## Total_Vista      1068.4968 1.192534e-09
## Total_Inversion  3763.9694 3.201822e-11
## dif_Plazo        1722.2953 4.975814e-27
## Total_Plazo      3698.5604 4.600180e-36
##
## $`2`
##                     v.test Mean in category Overall mean sd in
category
## Total_Plazo      12.827446       5731.12727  1332.663000
5643.5097
## dif_Plazo        11.350575       1927.28918   114.890965
4256.9060
## Total_Vista       6.293389       1192.60000   569.172500
1881.3775
## dif_Largo_plazo   4.350000        377.30382    26.109325
1571.2084
## dif_Libreta       1.970540        138.85773   -41.936910
1664.3871
## oper_ven_Libreta -2.040635        -71.05982     2.540635
798.3778
##                  Overall sd      p.value
## Total_Plazo       3698.5604 1.150963e-37
## dif_Plazo         1722.2953 7.367698e-30
## Total_Vista       1068.4968 3.106086e-10
## dif_Largo_plazo    870.8238 1.361376e-05
## dif_Libreta        989.6283 4.877651e-02
## oper_ven_Libreta   389.0335 4.128712e-02
##
## $`3`
##                    v.test Mean in category Overall mean sd in category
## Total_Inversion 20.137438       19731.0625    853.11600     18262.3327
```

```
## dif_Fondos_inv  18.729188       12649.0369    261.78000     19831.3676
## dif_CC          -2.746121        -238.7475     26.92591       904.0399
##                 Overall sd      p.value
## Total_Inversion   3763.969 3.468165e-90
## dif_Fondos_inv    2655.533 2.862504e-78
## dif_CC             388.440 6.030455e-03

result$category

## $`1`
##                 Cla/Mod   Mod/Cla Global      p.value    v.test
## Pension=NO     95.48694 64.354322  63.15 2.579927e-05  4.207691
## edatcat=18-25  99.21875  6.776948   6.40 2.036871e-03  3.084803
## edatcat=26-35  97.10145 17.876201  17.25 2.222726e-03  3.058736
## Debito_aff= SI 96.11650 26.414088  25.75 6.670968e-03  2.712838
## edatcat=46-55  96.40523 15.741729  15.30 2.713318e-02  2.209597
## Debito_aff= NO 92.86195 73.585912  74.25 6.670968e-03 -2.712838
## Pension=SI     90.63772 35.645678  36.85 2.579927e-05 -4.207691
## edatcat=66     89.18919 28.175027  29.60 2.757108e-07 -5.139336
##
## $`2`
##                 Cla/Mod    Mod/Cla Global      p.value    v.test
## edatcat=66     9.628378 51.8181818  29.60 5.673304e-07  5.002019
## Pension=SI     8.548168 57.2727273  36.85 8.394347e-06  4.454871
## Debito_aff= NO 6.127946 82.7272727  74.25 3.207336e-02  2.143495
## Debito_aff= SI 3.689320 17.2727273  25.75 3.207336e-02 -2.143495
## edatcat=18-25  0.781250  0.9090909   6.40 5.587666e-03 -2.771045
## edatcat=26-35  2.318841  7.2727273  17.25 2.147315e-03 -3.069063
## Pension=NO     3.721298 42.7272727  63.15 8.394347e-06 -4.454871
##
## $`3`
##                Cla/Mod Mod/Cla Global    p.value   v.test
## VISA_aff= SI 5.7142857    12.5   1.75 0.03297481  2.13239
## VISA_aff= NO 0.7124682    87.5  98.25 0.03297481 -2.13239

# Clase 1
result$quanti$`1`

##                     v.test Mean in category Overall mean sd in
category
## dif_Libreta      -2.059236       -53.755683    -41.93691
936.5669
## dif_Largo_plazo  -3.945021         6.185454     26.10933
810.0829
## dif_Fondos_inv   -5.831256       171.973458    261.78000
1683.4009
## Total_Vista      -6.081244       531.488260    569.17250
991.5231
## Total_Inversion  -6.636977       708.235326    853.11600
3006.7650
## dif_Plazo       -10.766110         7.353319    114.89096
```

```
1374.6009
## Total_Plazo     -12.538430      1063.713981   1332.66300
3344.2034
##                  Overall sd      p.value
## dif_Libreta       989.6283 3.947160e-02
## dif_Largo_plazo   870.8238 7.979286e-05
## dif_Fondos_inv   2655.5333 5.501159e-09
## Total_Vista      1068.4968 1.192534e-09
## Total_Inversion  3763.9694 3.201822e-11
## dif_Plazo        1722.2953 4.975814e-27
## Total_Plazo      3698.5604 4.600180e-36

result$category$`1`

##                  Cla/Mod   Mod/Cla Global      p.value    v.test
## Pension=NO     95.48694 64.354322  63.15 2.579927e-05  4.207691
## edatcat=18-25  99.21875  6.776948   6.40 2.036871e-03  3.084803
## edatcat=26-35  97.10145 17.876201  17.25 2.222726e-03  3.058736
## Debito_aff= SI 96.11650 26.414088  25.75 6.670968e-03  2.712838
## edatcat=46-55  96.40523 15.741729  15.30 2.713318e-02  2.209597
## Debito_aff= NO 92.86195 73.585912  74.25 6.670968e-03 -2.712838
## Pension=SI     90.63772 35.645678  36.85 2.579927e-05 -4.207691
## edatcat=66     89.18919 28.175027  29.60 2.757108e-07 -5.139336

# Clase 2
result$quanti$`2`

##                     v.test Mean in category Overall mean sd in
category
## Total_Plazo      12.827446       5731.12727  1332.663000
5643.5097
## dif_Plazo        11.350575       1927.28918   114.890965
4256.9060
## Total_Vista       6.293389       1192.60000   569.172500
1881.3775
## dif_Largo_plazo   4.350000        377.30382    26.109325
1571.2084
## dif_Libreta       1.970540        138.85773   -41.936910
1664.3871
## oper_ven_Libreta -2.040635        -71.05982     2.540635
798.3778
##                   Overall sd      p.value
## Total_Plazo       3698.5604 1.150963e-37
## dif_Plazo         1722.2953 7.367698e-30
## Total_Vista       1068.4968 3.106086e-10
## dif_Largo_plazo    870.8238 1.361376e-05
## dif_Libreta        989.6283 4.877651e-02
## oper_ven_Libreta   389.0335 4.128712e-02

result$category$`2`
```

```
##                  Cla/Mod    Mod/Cla Global      p.value    v.test
## edatcat=66     9.628378 51.8181818  29.60 5.673304e-07  5.002019
## Pension=SI     8.548168 57.2727273  36.85 8.394347e-06  4.454871
## Debito_aff= NO 6.127946 82.7272727  74.25 3.207336e-02  2.143495
## Debito_aff= SI 3.689320 17.2727273  25.75 3.207336e-02 -2.143495
## edatcat=18-25  0.781250  0.9090909   6.40 5.587666e-03 -2.771045
## edatcat=26-35  2.318841  7.2727273  17.25 2.147315e-03 -3.069063
## Pension=NO     3.721298 42.7272727  63.15 8.394347e-06 -4.454871

# Clase 3
result$quanti$`3`

##                    v.test Mean in category Overall mean sd in category
## Total_Inversion 20.137438       19731.0625    853.11600     18262.3327
## dif_Fondos_inv  18.729188       12649.0369    261.78000     19831.3676
## dif_CC          -2.746121        -238.7475     26.92591       904.0399
##                 Overall sd      p.value
## Total_Inversion   3763.969 3.468165e-90
## dif_Fondos_inv    2655.533 2.862504e-78
## dif_CC             388.440 6.030455e-03

result$category$`3`

##                Cla/Mod Mod/Cla Global    p.value   v.test
## VISA_aff= SI 5.7142857    12.5   1.75 0.03297481  2.13239
## VISA_aff= NO 0.7124682    87.5  98.25 0.03297481 -2.13239
```

Resp.: Interpretación de las Clases obtenidas

- Cluster 1 (Jovenes): es de jovenes activos (grupo de personas no pensionistas);
- Cluster 2 (Mas riqueza/Mejores clientes): Personas con mas edad, y con signiticativa cantidad en Total Plazo, que pueden recibir invitación de otros bancos
- Cluster 3 (Inversores): Grupo de personas activa en el tema de inversiones

7. Efectúe ahora la asignación de los clientes que han sido baja en la tipología de clientes anterior (utilice para ello la función knn1 de la librería class).

```
library(class)
nd <- 3
client.result <- knn1(consol.kmeans$centers,
pca.churn$ind.sup$coord[,1:nd], cl=c("c1","c2","c3"))
client.result

##    [1] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c1 c1 c1
c2 c2
##   [24] c2 c1 c1 c1 c1 c1 c2 c2 c2 c2 c1 c1 c1 c1 c1 c1 c2 c1 c1 c1 c1
c1 c1
```

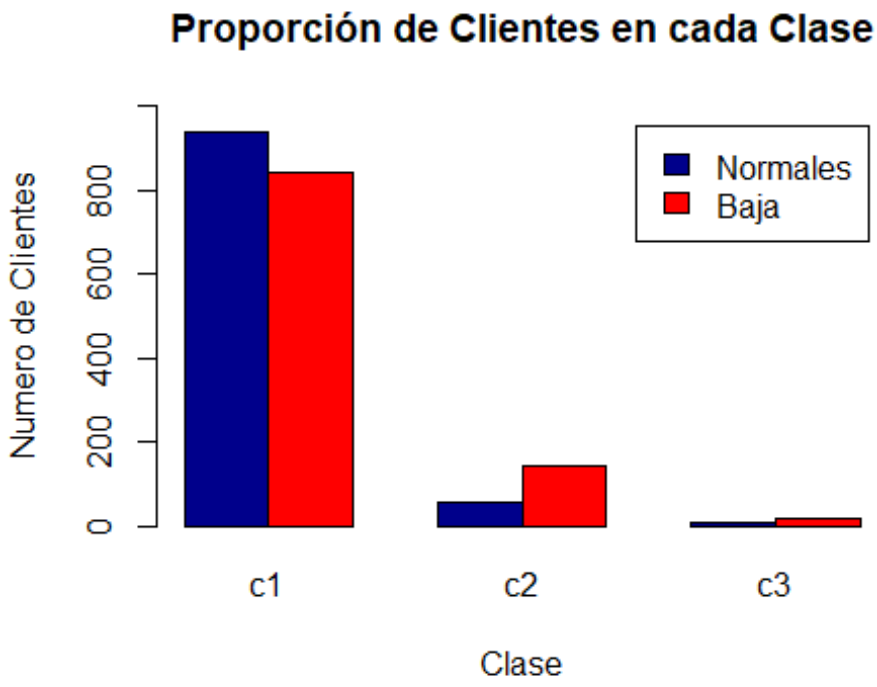
```
##   [47] c2 c2 c1 c1 c1 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2
c1 c1
##   [70] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##   [93] c1 c1 c2 c2 c3 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c3
c3 c1
##  [116] c2 c2 c1 c1 c2 c2 c2 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [139] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [162] c1 c2 c2 c2 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [185] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c1 c1 c1 c1 c1 c1 c1 c1
c2 c1
##  [208] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c1 c1 c1 c2 c2
c2 c1
##  [231] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c2 c2 c1 c1
c1 c1
##  [254] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [277] c1 c1 c2 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c3 c1 c1 c2
c1 c1
##  [300] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c2 c2
##  [323] c2 c2 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c1 c1 c1 c1 c1 c1
c1 c1
##  [346] c1 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [369] c3 c1 c1 c1 c1 c1 c1 c1 c2 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c1 c1
c1 c1
##  [392] c1 c1 c1 c2 c2 c3 c3 c1 c1 c1 c1 c1 c1 c1 c2 c1 c1 c1 c1 c2 c2
c2 c3
##  [415] c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c1 c1 c3 c3 c3
c2 c2
##  [438] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [461] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2
c1 c1
##  [484] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [507] c1 c1 c1 c1 c1 c1 c1 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [530] c1 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c2 c2 c2 c1 c2 c1 c1
c1 c1
##  [553] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c2 c1
c1 c2
##  [576] c2 c2 c1 c3 c3 c1 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c1 c1 c1 c1 c1
c3 c1
##  [599] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c3
c1 c1
```

```
##  [622] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [645] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2
c1 c1
##  [668] c1 c2 c1 c1 c1 c1 c1 c1 c3 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [691] c2 c1 c1 c1 c1 c2 c2 c2 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [714] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [737] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c2 c1
c1 c1
##  [760] c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c1 c1
c1 c2
##  [783] c2 c1 c2 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c2
c1 c1
##  [806] c1 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c2 c2 c1 c1 c1
c1 c1
##  [829] c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c2 c1 c1 c1 c1 c1 c2 c2 c2 c1
c1 c1
##  [852] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1
c1 c1
##  [875] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c2 c1
c2 c2
##  [898] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [921] c1 c1 c1 c1 c2 c2 c2 c2 c2 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1
c1 c1
##  [944] c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1
c1 c1
##  [967] c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c2 c1 c1 c1 c1 c1
c1 c1
##  [990] c1 c1 c1 c1 c1 c1 c1 c1 c1 c2 c1
## Levels: c1 c2 c3
```

8. Represente gráficamente (función barplot) la proporción de clientes en cada una de las clases, tanto los clientes “normales”, como los que han sido baja. ¿Podemos deducir que algunos clusters tienen un riesgo de baja mayor que otros?

```
#consol.mean$size es la cantidad de clientes "normales" por Cluster. El
client.result es el resultado del knn1 de los Individuos de Baja.

Xtot <- rbind(consol.kmeans$size,table(client.result))
library(dplyr)
barplot(Xtot,
        main = "Proporción de Clientes en cada Clase",
        xlab = "Clase",
        ylab = "Numero de Clientes",
```

```
        col = c("darkblue","red"),
        legend.text = c("Normales","Baja"),
        beside = TRUE,
        ylim = c(0,1000))
```



Resp.: Con el barplot dibujado podemos deducir que los clientes en los Clusters 2 y 3 tienen mas probabilidad de Baja, siendo es mas considerable cliente en el Cluster 2.