Xavier Gonzalez Arriola

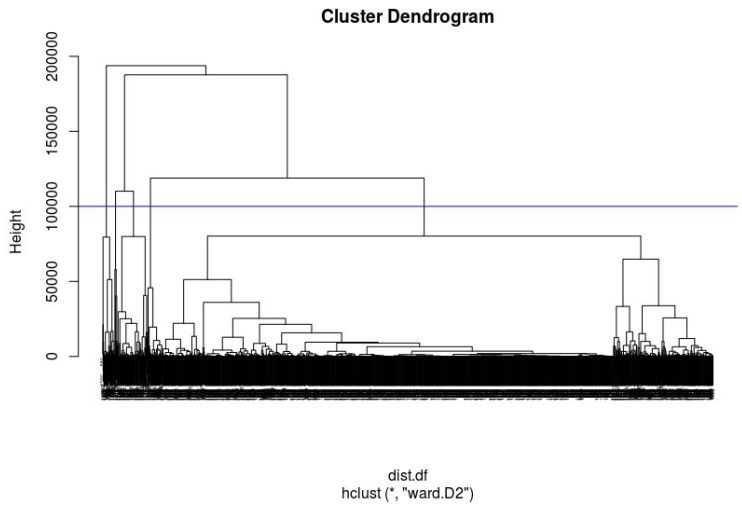
SESIÓN 5: CLUSTERING

Con los resultados obtenidos en el ACP de los datos “churn”, procedemos ahora a obtener una tipología de los clientes de la entidad bancaria según su posición en los diversos productos financieros. Para ello disponemos del Análisis de Componentes Principales realizado en la Sesión 4.

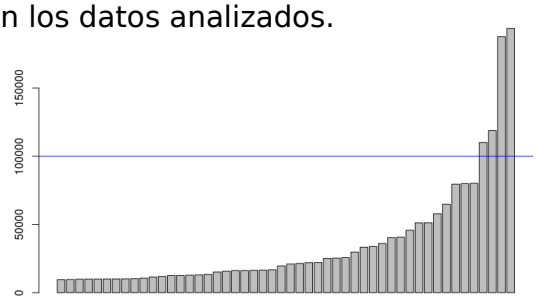
1. Con las componentes principales significativas efectuar una Clasificación Ascendente Jerárquica por el método de Ward. Explique en qué consiste el método de agregación de Ward?. Represente el dendrograma (o árbol jerárquico) obtenido.

El método de Ward consiste en agrupar en cada iteración, hasta obtener el árbol completo, el par de centros de gravedad (de objetos, puntos o grupos) tales que al agregarlos se pierde la mínima inercia/información.

```
> hclus.df <- hclust(dist.df,method="ward.D2")
```



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(hclus.df$height[(nrow(df)-50):(nrow(df)-1)])
```

Tomamos un corte que nos de 5 clases, cortando los saltos mayores en el dendograma cómo se indica en la figura del ejercicio anterior, correspondientes

al corte en las 4 últimas agregaciones indicado en el diagrama de barras del índice de nivel de las últimas agregaciones.

3. Obtenga la partición del árbol jerárquico en el número de clases finales deseado. Calcule la calidad de la partición obtenida mediante el cociente de la *Inercia entre clases* respecto de la *Inercia total*.

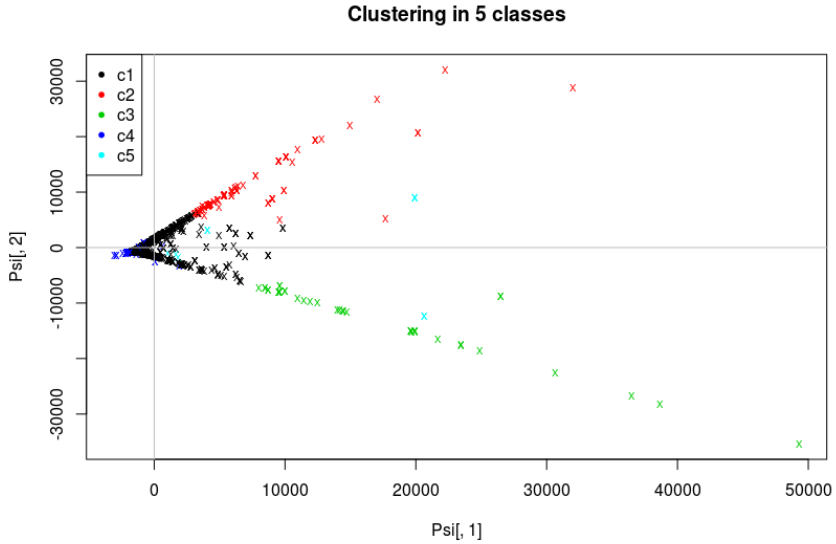
```
> nc=5
> cut5 <- cutree(hclus.df,nc)

> cdg <- aggregate(as.data.frame(Psi),list(cut5),mean)[,2:(nd+1)]
> cdg

> Bss <- sum(rowSums(cdg^2)*as.numeric(table(cut5))) # between cluster inertia
> Tss <- sum(Psi^2)                                  # total sum of squares

> 100*Bss/Tss

[1] 65.22459
```

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. Calcule de nuevo la calidad de la partición consolidada.

Consiste en mejorar las particiones eliminando la restricción de un árbol jerárquico de que las clases estén solapadas.

```
# CONSOLIDACION DE LA PARTICION
# CENTROS DE GRAVEDAD DE LAS nc CLASES OBTENIDAS POR CORTE DEL ARBOL JERARQUICO
cdg.nc <- aggregate(as.data.frame(Psi),list(cut5),mean)[,2:(nd+1)]
# ALGORITMO kmeans CON CENTROS INICIALES EN LOS CENTROIDES cdg.nc
kmeans <- kmeans(Psi,centers=cdg.nc)
# CALIDAD DE LA PARTICION FINAL EN 5 CLASSES
100*kmeans$betweenss/kmeans$totss
[1] 65.8196
```

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



6. Por último, 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.

Clase 1 – baja_NO:

```
$category$`1`
              Cla/Mod   Mod/Cla    Global      p.value    v.test
Baja=NO       94.25982 53.001133 49.824385 8.305290e-16  8.049618
edatcat=18-25 97.65625  7.078143  6.422479 1.358658e-04  3.815576
edatcat=56-65 81.53846 12.004530 13.045660 2.954332e-04 -3.619272
Baja=SI       83.00000 46.998867 50.175615 8.305290e-16 -8.049618

$quanti$`1`
                   v.test Mean in category Overall mean sd in category  Overall sd       p.value
antig            -3.643552        16.660388    16.804447       5.051839    4.921995  2.689010e-04
dif_Largo_plazo  -4.185951        -3.182203    26.150853     803.769451  872.348999  2.839748e-05
dif_Hipoteca     -5.803021       -68.610419     9.941295    1360.578324 1685.111379  6.513064e-09
dif_Plazo        -6.883936         9.819003   102.248836     826.305776 1671.484775  5.822101e-12
dif_Fondos_inv   -8.250363        83.827384   259.958354     741.419037 2657.600794  1.579138e-16
Total_Seguros    -8.588530       143.800680   279.960361     614.763948 1973.589325  8.808920e-18
Total_Vista      -9.910575       481.148924   565.386854     942.792360 1058.122407  3.744838e-23
Total_Inversion -16.843190       343.396376   853.314099    1266.745822 3768.797959  1.177166e-63
Total_activo    -17.150172       298.749151   616.766683     915.114772 2308.392562  6.266647e-66
Total_Plazo     -26.252397       545.862967  1321.739087    1316.530427 3679.175482 6.711039e-152
```

Clase 2 – baja_SI

```
$category$`2`
               Cla/Mod   Mod/Cla   Global      p.value    v.test
Baja=SI       9.600000 78.688525 50.17561 2.779055e-11  6.657827
Pension=SI    9.439124 56.557377 36.67837 4.483570e-06  4.587606
edatcat=66..  10.000000 48.360656 29.60361 7.245843e-06  4.486336
edatcat=26-35  2.325581  6.557377 17.26041 4.343472e-04 -3.518284
Pension=NO    4.199683 43.442623 63.32163 4.483570e-06 -4.587606
Baja=NO       2.618328 21.311475 49.82439 2.779055e-11 -6.657827

$quanti$`2`
                  v.test Mean in category Overall mean sd in category  Overall sd       p.value
Total_Plazo     35.581803       12808.3197   1321.73909    6668.808875 3679.175482 2.678635e-277
dif_Plazo       11.658824        1812.1440    102.24884    5159.583612 1671.484775  2.068796e-31
Total_Vista     10.749865        1563.4344    565.38685    1799.835074 1058.122407  5.934826e-27
dif_Largo_plazo  5.807020         470.6346     26.15085    1693.848828  872.348999  6.359467e-09
antig            3.356682          18.2541     16.80445       2.481279    4.921995  7.888381e-04
```

Clase 3 – inversores_mayores

```
$category$`3`
              Cla/Mod  Mod/Cla   Global      p.value   v.test
edatcat=56-65 6.153846 41.02564 13.04566 1.223891e-05 4.373288

$quanti$`3`
                  v.test Mean in category Overall mean sd in category Overall sd       p.value
Total_Inversion 36.305075       22553.1026     853.3141      11265.386   3768.798 1.345513e-288
dif_Fondos_inv  21.215064        9201.6438     259.9584      15552.737   2657.601 6.933307e-100
dif_Plazo       -3.819527        -910.2564     102.2488       3857.845   1671.485  1.337080e-04
```

Clase 4 – VISA_activo

```
$category$`4`
                    Cla/Mod  Mod/Cla    Global      p.value    v.test
VISA=SI            6.888889 55.35714 22.579027 8.196504e-08  5.362749
Debito_normal=SI 12.345679 17.85714  4.064225 6.299339e-05  4.001308
```

```
edatcat=46-55     6.557377 35.71429 15.303562 1.411959e-04  3.806064
Nomina=SI         4.833333 51.78571 30.105369 6.433271e-04  3.412658
Nomina=NO         1.938263 48.21429 69.894631 6.433271e-04 -3.412658
Debito_normal=NO  2.405858 82.14286 95.935775 6.299339e-05 -4.001308
VISA=NO           1.620220 44.64286 77.420973 8.196504e-08 -5.362749
edatcat=66..      0.000000  0.00000 29.603613 2.081832e-09 -5.991290

$quanti$`4`
                   v.test Mean in category Overall mean sd in category Overall sd       p.value
Total_activo     36.887794       11837.4286   616.766683      5336.5979  2308.3926 7.252104e-298
dif_Hipoteca     12.617481        2811.6786     9.941295      5612.8276  1685.1114  1.691451e-36
oper_ven_Libreta  3.482718         181.2652     2.690632       676.6577   389.1125  4.963512e-04
```

Clase 5 – asegurados

```
$quanti$`5`
                 v.test Mean in category Overall mean sd in category Overall sd      p.value
Total_Seguros   38.368935        24172.000    279.96036      11136.372  1973.5893 0.000000e+00
dif_Seguros     19.034027         4122.998     17.88439       6976.496   683.5619 8.913331e-81
Total_Inversion  5.411966         7288.700    853.31410       7648.453  3768.7980 6.233657e-08
Total_Plazo      4.673190         6746.500   1321.73909       8339.182  3679.1755 2.965576e-06
Total_Vista      4.493063         2065.400    565.38685       2206.461  1058.1224 7.020615e-06

$category$`5`
               Cla/Mod Mod/Cla   Global      p.value   v.test
edatcat=56-65 2.307692      60 13.04566 0.0006701135 3.401524
```

7. Evalúe la relación de la partición obtenida con la variable “Baja” mediante la tabla cruzando ambas variables.

La clase 2 tiene un número de bajas significativo.

Glosario para la descripción

A partir de las variables continuas

`Overall mean`: es la media global de la variable

`Mean in category`: es la media de la variable en la clase (=”cluster”) considerado

`v.test`: es ell valor del estadístico N(0,1) al comparar la “Mean in Category” con la “Overall mean”.

`p.value`: es el p.valor obtenido en la anterior comparación.

A partir de las variables categóricas

`Global`: es el porcentaje global de la modalidad (= categoría)

`Mod/Cla`: es el porcentaje de la modalidad en la clase (= cluster) considerado

`v.test`: es el valor del estadístico al comparar la proporción Global con la proporción en la clase (= Mod/Cla)

`p.value`: es el p.valor obtenido en la comparación de ambas proporciones

`Cla/Mod`: es el porcentaje de una modalidad en una clase, respecto del total de la modalidad. Da la especifidad de una clase respecto de una modalidad.