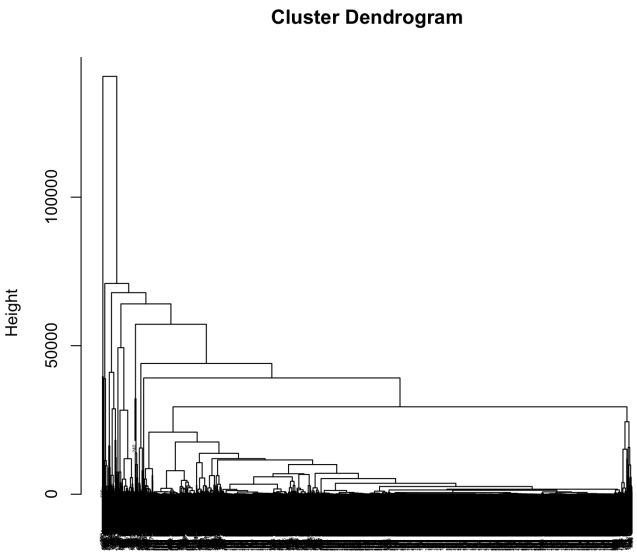
SESIÓN 5: CLUSTERING

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.

Partiendo del ejercicio de la s4, con datos limpios

```
> Psi = pca.churn_NoNA$ind$coord
> dist.churn <- dist(Psi)
> hclus.churn <- hclust(dist.churn,method="ward.D2")
> plot(hclus.churn,cex=0.3)
```

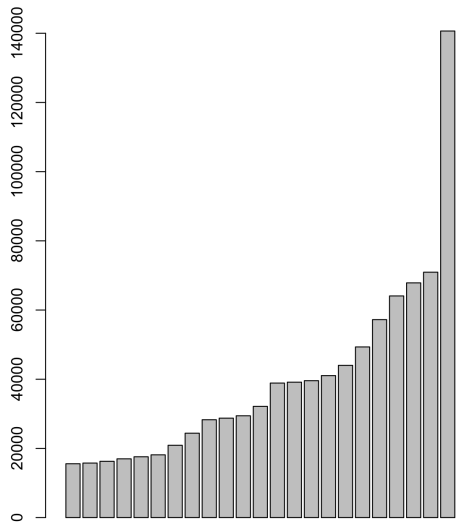


El árbol obtenido es demasiado grande para poder observar nada a nivel de hojas. Aunque si se aprecian unos posibles cortes/grupos que se pueden hacer en los niveles superiores.

El método de Ward agrupa los individuos más cercanos, y dibuja un arco entre ellos que equivale a la distancia en el espacio euclideo dónde están representados. A mayor diferencia entre los arcos, mayor distancia entre los grupos/elementos en la representación.

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.churn$height[(nrow(churn)-30):(nrow(churn)-1)])
```



Se podría hacer un corte a partir de la quinta barra, obteniendo así 6 grupos. Aunque cierto que, fuera de la distancia entre el primer y segunda barra, no hay mucha diferencia entre los primeros grupos. Lo que indica que los elementos entre ellos serán muy similares y que posiblemente cueste hacer una clusterización eficaz y bien definida.

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*.

Esta es la partición y la representación gráfica de los diferentes grupos

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

Y la calidad la calculamos con:

```
> cdg <- aggregate(as.data.frame(Psi),list(cut5),mean)[,2:(nc+1)]
> cdg
       Dim.1          Dim.2      Dim.3
1  -259.5451       5.210966  -190.5791
2  -535.0277    2925.563509  4713.6581
3 40497.1979    1272.086881  1237.5173
4  7635.8741    -930.841072 -1869.0733
5   581.6085 -10291.718730  6352.2072
> Bss <- sum(rowSums(cdg^2)*as.numeric(table(cut5)))
> Tss <- sum(Psi^2)
> 100*Bss/Tss
[1] 58.28222
```

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.

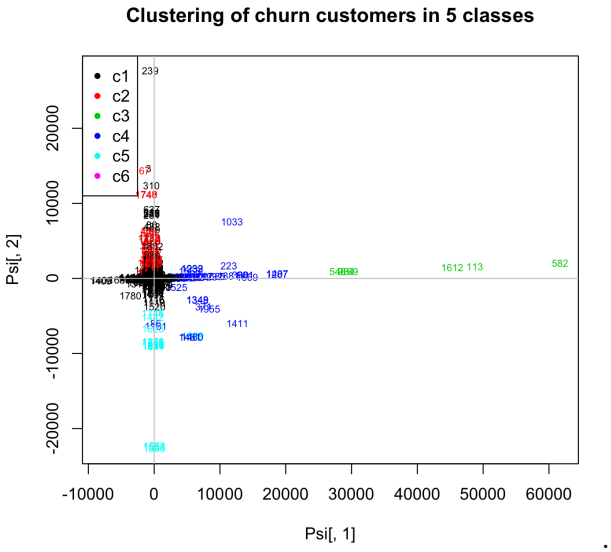
La operación de consolidación mejora el clustering teniendo en cuenta el centro de las zonas creadas durante el proceso anterior, el cual se utiliza para hacer un kmeans.

```
> k_def <- kmeans(Psi,centers=cdg)
> 100*k_def$betweenss/k_def$totss
[1] 61.54643
```

La calidad después de la consolidación es unos dos puntos superior.

5. Represente la partición 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 churn customers in
5 classes")
> text(Psi[,1],Psi[,2],col=cut5,cex = 0.6)
> abline(h=0,v=0,col="gray")
> legend("topleft",c("c1","c2","c3","c4","c5","c6"),pch=20,col=c(1:6))
```



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.

```
>catdes(cbind(as.factor(k_def$cluster),churn_NoNA),num.var=1,proba=0.0
01)
$quanti$`1`
              v.test Mean in category Overall mean sd in category Overall sd  p.value
dif_Hipoteca      12.641712   99.648831    9.941295 1100.6668 1685.1114  1.243170e-36
oper_caj_Libreta   3.353601   -6.042665   -7.765429  114.7920  121.9888  7.976715e-04
Total_Plazo      -10.508808 1158.922597 1321.739087 3221.7095 3679.1755  7.868164e-26
dif_Plazo        -13.637409    6.258234  102.248836  757.6144 1671.4848  2.399346e-42
Total_Inversion  -15.977112  599.745455  853.314099 2740.3282 3768.7980  1.844989e-57
dif_Fondos_inv   -22.396912    9.305694  259.958354  517.3371 2657.6008 4.218181e-111

$quanti$`2`
v.test Mean in category Overall mean sd in category Overall sd p.value
dif_Plazo   32.4273 11341.51  102.2488 5624.005 1671.485 1.131798e-230
Total_Plazo 18.4892 15427.39 1321.7391 9572.922 3679.175  2.522650e-76
```

```
$quanti$`3`
    v.test Mean in category Overall mean sd in category Overall sd p.value
dif_Fondos_inv  37.41022 40797.62 259.9584 12489.77 2657.601 2.665201e-306
Total_Inversion 26.37062 41376.33 853.3141 12175.61 3768.798 2.978181e-153

$quanti$`4`
    v.test Mean in category Overall mean sd in category Overall sd p.value
dif_Fondos_inv  16.97712  9229.114 259.9584 4047.299 2657.601 1.213018e-64
Total_Inversion 12.49907 10217.680 853.3141 4236.783 3768.798 7.553125e-36
dif_Plazo       -8.58105 -2749.037 102.2488 4933.739 1671.485 9.401112e-18

$quanti$`5`
             v.test Mean in category Overall mean sd in category Overall sd p.value
dif_Ahorro         4.034215    667.2871  7.076187 1283.4870  614.3409  5.478508e-05
oper_caj_Libreta  -4.236159   -145.4250 -7.765429  188.4488  121.9888  2.273756e-05
oper_ven_Libreta  -4.495126   -463.2507  2.690632  575.1106  389.1125  6.952865e-06
dif_Hipoteca     -29.935734 -13428.0000  9.941295 6837.9981 1685.1114 6.747763e-197
```

De los 5 grupos que hay, son:

Grupo1 (*familias*): Gente con hipotecas y con transacciones en caja. Se caracterizan también por poseer nada productos a plazos ni inversiones.

Grupo2 (*mileuristas*): Gente con los productos a plazos.

Grupo3 (*inversores*): Grupo con inversiones y con productos a plazo.

Grupo4 (*especuladores*): Grupo con inversiones pero sin productos a plazo.

Grupo 5 (*ahorradores*): Tienen ahorros, pocas operativas ni en ventanilla ni en caja y no tienen hipoteca.

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

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.