Apuntes Cluster_Eolica

Consoli Quintana

2022-11-08

Práctica ANÁLISIS CLÚSTER JERÁRQUICO:

Consideramos *6 variables* que caracterizan a un grupo de 20 empresas de producción eléctrica mediante tecnología eólica.

OBJETIVO:

- Segmentar este conjunto de empresas.
- Hacer grupos homogéneos (**conglomerados**) y
- Caracterizar a estos grupos homogéneos.

¿Qué variables clasificadoras vamos a utilizar?:

1. SOLVENCIA
2. FPIOS (fondos propios).
3. MARGEN (margen de beneficio).
4. RES (resultado del ejercicio).
5. RENFIN (rentabilidad financiera).
6. APALANCA (grado de apalancamiento).

¿Qué método vamos a utilizar?

- Al haber pocos casos: –> método jerárquico de agrupación de casos –> **método de WARD.**

Preparación de los datos:

Como hemos hecho hasta ahora, utilizamos la base de datos del archivo Excel `eolica_20.xlsx` en concreto la 3ª hoja `Datos`.

Procedemos a eliminar todo lo que tengamos en el environment:

```
rm(list = ls())
```

Importamos los datos de la tercera hoja *Datos*: para ello volvemos a cargar el paquete necesario y comprobamos que la primera columna de los datos es una variable de tipo

CUALITATIVO y por tanto, tenemos que redefinir el *data frame* para que esa variable se convierta en los *nombres de los casos*:

```
library(readxl)

## Warning: package 'readxl' was built under R version 4.2.1

eolicos <- read_excel("eolica_20.xlsx", sheet = "Datos")
eolicos <- data.frame(eolicos, row.names = 1)
```

Le pedimos que nos haga un resumen de los datos para visualizarlos:

```
summary(eolicos)

##       RES              ACTIVO             FPIOS              RENECO       
##  Min.   : -5662   Min.   :  109024   Min.   : -77533   Min.   :-2.8130  
##  1st Qu.:  2865   1st Qu.:  187240   1st Qu.:  28867   1st Qu.: 0.8765  
##  Median :  7388   Median :  271636   Median :  85447   Median : 3.6150  
##  Mean   : 50754   Mean   : 1183599   Mean   : 593433   Mean   : 2.9399  
##  3rd Qu.: 21206   3rd Qu.:  813816   3rd Qu.: 252389   3rd Qu.: 4.7735  
##  Max.   :727548   Max.   :13492812   Max.   :6904824   Max.   : 8.5860  
##  NA's   :1                           NA's   :1         NA's   :1        
##      RENFIN           LIQUIDEZ          MARGEN          SOLVENCIA      
##  Min.   :-7.302   Min.   :0.0780   Min.   :-19.19   Min.   :-40.74  
##  1st Qu.: 2.442   1st Qu.:0.7342   1st Qu.: 14.42   1st Qu.: 11.26  
##  Median :11.338   Median :1.2345   Median : 22.40   Median : 23.68  
##  Mean   :15.304   Mean   :1.4200   Mean   : 33.16   Mean   : 32.68  
##  3rd Qu.:25.991   3rd Qu.:1.5615   3rd Qu.: 38.87   3rd Qu.: 52.62  
##  Max.   :52.261   Max.   :5.3300   Max.   :208.36   Max.   : 99.08  
##  NA's   :1                         NA's   :1                        
##     APALANCA           MATRIZ         
##  Min.   :-6265.50   Length:20          
##  1st Qu.:   16.13   Class :character   
##  Median :  145.93   Mode  :character   
##  Mean   :  -17.17                      
##  3rd Qu.:  504.74                      
##  Max.   : 1019.62                      
##
```

Observamos las variables y sus principales características descriptivas (como ya hemos hecho en otras prácticas).

Ahora vamos a realizar un nuevo *data frame* al que llamaremos `originales` y que estará compuesta por las VARIABLES CLASIFICADORAS a las que hacíamos referencia al principio de la práctica:

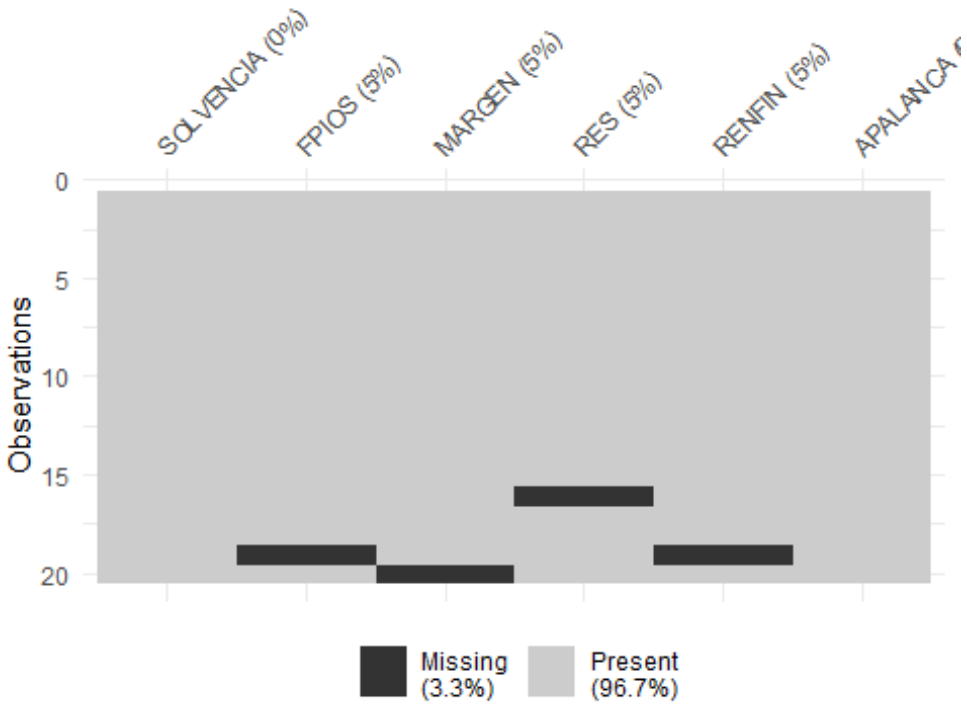
Para ello: seleccionamos dichas variables y las guardamos en el objeto (*data frame*) `originales` (previamente cargamos el paquete `dplyr`para poder usar el *pipe*:

```
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.1
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

originales <- eolicos %>% select(SOLVENCIA, FPIOS, MARGEN, RES, RENFIN,
APALANCA)
summary (originales)

##    SOLVENCIA          FPIOS             MARGEN            RES
##  Min.   :-40.74   Min.   : -77533   Min.   :-19.19   Min.   : -5662
##  1st Qu.: 11.26   1st Qu.:  28867   1st Qu.: 14.42   1st Qu.:  2865
##  Median : 23.68   Median :  85447   Median : 22.40   Median :  7388
##  Mean   : 32.68   Mean   : 593433   Mean   : 33.16   Mean   : 50754
##  3rd Qu.: 52.62   3rd Qu.: 252389   3rd Qu.: 38.87   3rd Qu.: 21206
##  Max.   : 99.08   Max.   :6904824   Max.   :208.36   Max.   :727548
##                   NA's   :1         NA's   :1        NA's   :1
##      RENFIN          APALANCA
##  Min.   :-7.302   Min.   :-6265.50
##  1st Qu.: 2.442   1st Qu.:   16.13
##  Median :11.338   Median :  145.93
##  Mean   :15.304   Mean   :  -17.17
##  3rd Qu.:25.991   3rd Qu.:  504.74
##  Max.   :52.261   Max.   : 1019.62
##  NA's   :1
```

Como hemos hecho hasta ahora, tenemos que localizar y tratar los VALORES FALTANTES o *missing values*:

```
library(visdat)

## Warning: package 'visdat' was built under R version 4.2.1

vis_miss(originales)

## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## Please use `gather()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning
was generated.
```




```
originales %>% filter(is.na(RES) | is.na(FPIOS) | is.na(RENFIN) |
is.na(MARGEN) | is.na(SOLVENCIA) | is.na(APALANCA)) %>%
  select(RES, FPIOS, RENFIN, MARGEN, SOLVENCIA, APALANCA)
```

```
##                                       RES    FPIOS RENFIN MARGEN
SOLVENCIA
## Biovent Energia SA                     NA  70033.0 11.952 22.792
38.082
## Parque Eolico Santa Catalina SL 3645.278       NA     NA 31.780    -
1.126
## WPD Wind Investment SL.         -850.068 108023.8 -1.049     NA
99.082
##                                  APALANCA
## Biovent Energia SA                141.163
## Parque Eolico Santa Catalina SL -6265.496
## WPD Wind Investment SL.             0.000
```

ANALIZAMOS EL GRÁFICO ANTERIOR:

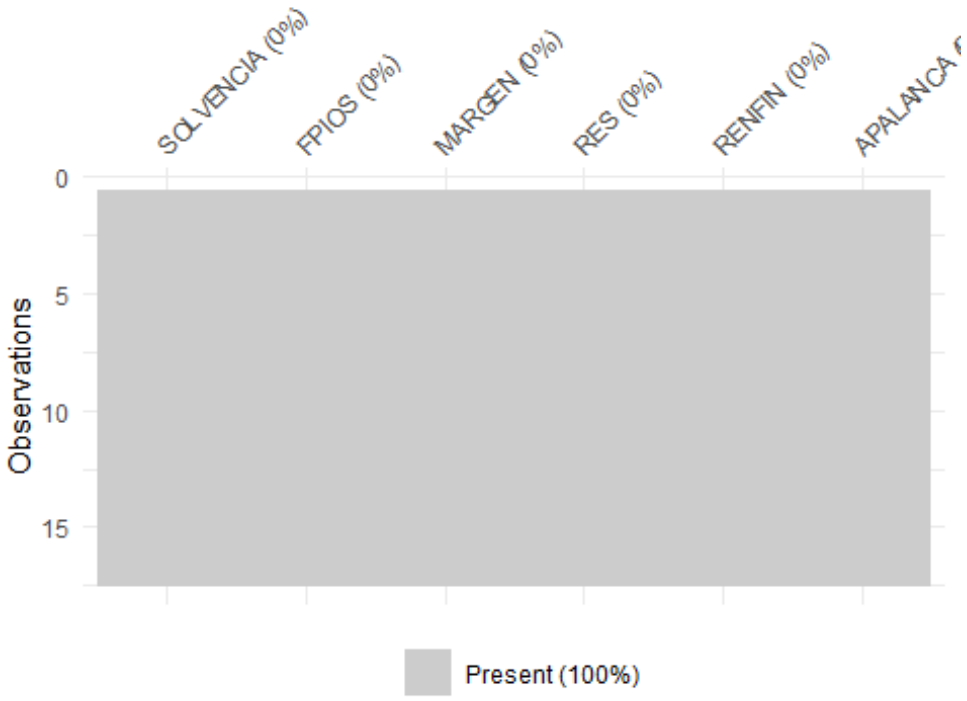
1. Existen 4 NA que afectan a 4 casos (variables) y a 3 empresas eólicas:
2. Estas empresas son: Biovent Energía SA, Pargue Eólico Santa Catalina SL y WPD Wind Investment SL.

ELIMINNAMOS LOS CASOS (sabemos que podíamos recurrir a buscar dichos datos en otras bases de datos o estimarlos, también).

```
originales <- originales %>%
  filter(! is.na(RES) & ! is.na(FPIOS) & ! is.na(RENFIN) & !
is.na(MARGEN) & ! is.na(SOLVENCIA) & ! is.na(APALANCA))
```

Comprobamos (que nunca lo hemos visualizado):

```
vis_miss(originales)
```



Podemos verificar en el Environment que el *data frame* `originales` ha pasado a tener 17 casos.

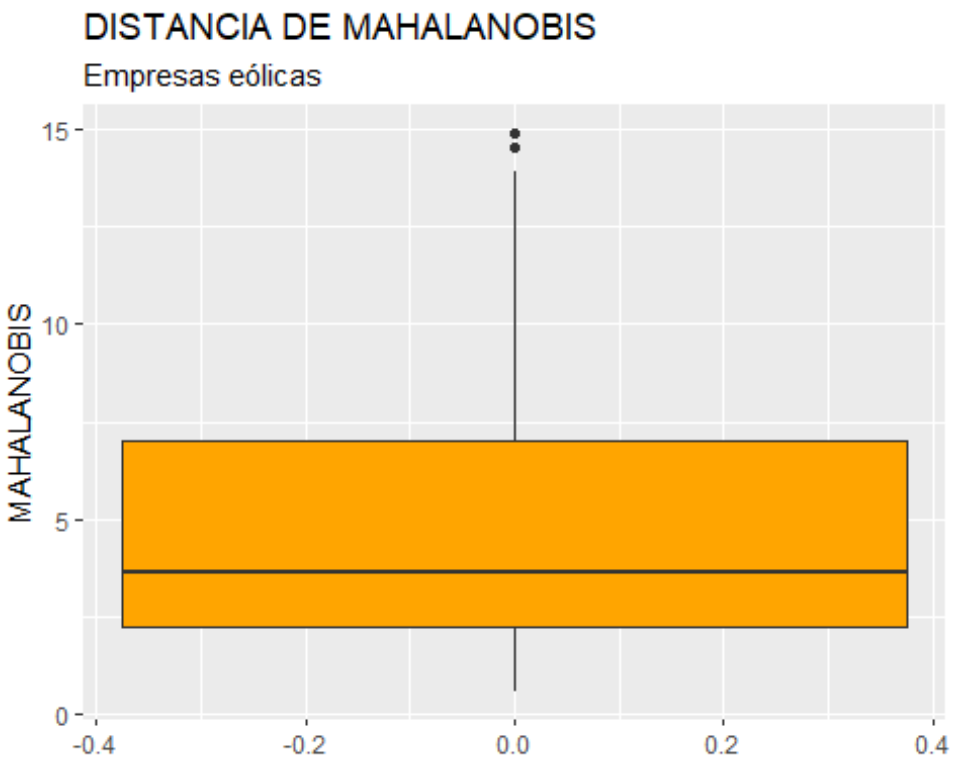
IDENTIFICAMOS CASOS ATÍTICOS O *OUTLIERS*:

1. Usamos la *distancia de Mahalanobis*
2. Recordamos que las distancias de los casos se van a almacenar en un *vector* al que llamaremos `MAHALANOBIS`.
3. Incluiremos este vector al *data frame* `originala`l que llamaremos `originales_maha`.
4. Usamos para ello la función de pegado de columnas: `cbin()`.

```
MAHALANOBIS <- mahalanobis(originales[],
                           center = colMeans(originales[]),
                           cov=cov(originales[]))
originales_maha <- cbind(originales, MAHALANOBIS)
```

5. Construimos el DIAGRAMA DE CAJA para `MAHALANOBIS` utilizando la función `ggplot()`del paquete `ggplot2`:

```
library (ggplot2)
ggplot(data = originales_maha, map = (aes(y = MAHALANOBIS))) +
  geom_boxplot(fill = "orange") +
  ggtitle("DISTANCIA DE MAHALANOBIS", subtitle = "Empresas eólicas") +
  ylab("MAHALANOBIS")
```



ANALIZAMOS EL GRÁFICO:

1. Existencia, por encima de la caja, de 2 OUTLIERS.

IDENTIFICAMOSLOS CASOS CONCRETOS, para ello:

1. Calculamos los cuartiles primero y tercero de *MAHALANOBIS* y
2. FILTRAMOS.

```
Q1M <- quantile (originales_maha$MAHALANOBIS, c(0.25))
Q3M <- quantile (originales_maha$MAHALANOBIS, c(0.75))

originales_maha %>% filter(MAHALANOBIS > Q3M + 1.5*IQR(MAHALANOBIS) |
MAHALANOBIS < Q1M - 1.5*IQR(MAHALANOBIS)) %>% select(MAHALANOBIS)

##                                MAHALANOBIS
## Holding De Negocios De GAS SL.    14.89967
## Elawan Energy SL.                 14.56272
```

Holding De Negocios de GAS SL y *Elawan Energy SL*, son las dos empresas que se comportan, según su *distancia de Mahalanobis* observada, como outliers.

RECORDATORIO: *Queremos realizar un análisis clúster* es por ello que NO queremos ELIMINAR LOS *OUTLIERS* puesto que pretendemos agrupar TODOS LOS CASOS que tenemos en el análisis. Es posible que si algún caso pertenece aislado, sin agruparse con otros, en el proces ode agrupación, quizá se trate de un candidato a *outlier*

TEORÍA VISTA:

Los métodos de agrupación usualmente se basan en la **distancia euclídea.**

La distancia euclídea es **sensible a las unidades de medida de las diferentes variables clasificadoras** –> por ello, trabajamos con las **variables tipificadas**.

PREGUNTA: ¿qué es tipificar? ¿Cómo lo hacemos con R? –> USAMOS LA FUNCIÓN `scale()`:

```
zoriginales <- data.frame(scale(originales))
summary (zoriginales)
```

```
##    SOLVENCIA           FPIOS              MARGEN             RES         
##  Min.   :-2.2409   Min.   :-0.4372   Min.   :-1.0273   Min.   :-0.3576  
##  1st Qu.:-0.5947   1st Qu.:-0.3738   1st Qu.:-0.4120   1st Qu.:-0.2998  
##  Median :-0.4413   Median :-0.3396   Median :-0.2597   Median :-0.2762  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.6523   3rd Qu.:-0.2001   3rd Qu.: 0.1089   3rd Qu.:-0.1550  
##  Max.   : 1.7774   Max.   : 3.7429   Max.   : 3.3793   Max.   : 3.8565  
##      RENFIN           APALANCA        
##  Min.   :-1.4212   Min.   :-1.7980  
##  1st Qu.:-0.7931   1st Qu.:-0.7181  
##  Median :-0.3065   Median :-0.3203  
##  Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.6511   3rd Qu.: 0.5121  
##  Max.   : 2.1408   Max.   : 1.7874
```

Hemos creado un nuevo *data frame* al que hemos llamado `zoriginales`compuesto por las variables clasificadoras tipificadas (¿CÒMO ES LA MEDIA DE LAS VARIABLES TIPIFICADAS AL RESTAR LA MEDIA DE SU VARIABLE ORIGINAL Y DIVIDIR ENTRE LA DESVIACIÓN TÍPICA?). Es con este *data frame* con el que vamos a trabajar para realizar nuestro *ANÁLISIS CLUSTER*:

ANTES DE APLICAR UN MÉTODO DE AGRUPACIÓN –> *CREAMOS LA MATRIZ DE DISTANCIAS* que en R será un objeto al que llamaremos `d`:

```
d <- dist(zoriginales)
```

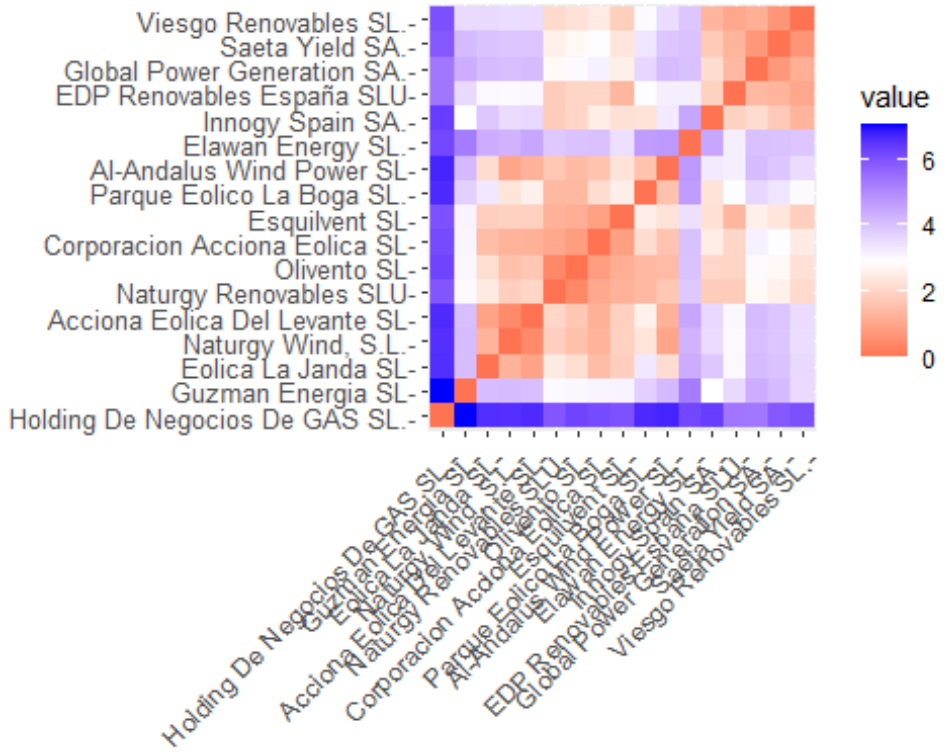
Como veis en el *Environment* aparece en el apartado de `values`como “dist”.

VAMOS A VISUALIZAR LA *MATRIZ DE DISTANCIAS* `D` :

Usamos un *gráfico de temperatura* con la función `fviz_dist()` del paquete `factoextra`:

```
library (factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.2

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

fviz_dist(d, lab_size = 10)
```



Como véis he modificado el argumento `lab_size = 10`. ¿Qué pensáis que significa?:

INTERPRETACIÓN DEL GRÁFICO DE TEMPERATURA:

1. COLOR AZUL: significa grandes distancias con el resto de empresas.
2. COLOR ROJO: significa poca distancia con el resto de empresas.
3. Los casos con intersecciones en **tonos más rojizos tenderán a agruparse con mayor facilidad** (o a agruparse antes).
4. Los casos cuya intersección está en un **tono azulado tenderán a pertenecer a grupos diferentes** (o a agruparse más tarde).
5. Observamos cómo las distancias de las dos empresas que fueron identificadas como outliers *(Holding de Negocios de Gas* y *Elawan Energy*) matienen grandes distancias (casillas azuladas) con el resto de empresas.

Probamos con la distancia de Mahalanobis.

ANÁLISIS CLÚSTER JERÁRQUICO MEDIANTE EL MÉTODO DE WARD:

USAMOS LA FUNCIÓN `hclust()` que guardaremos en un objeto al que llamaremos `cluster_j`.

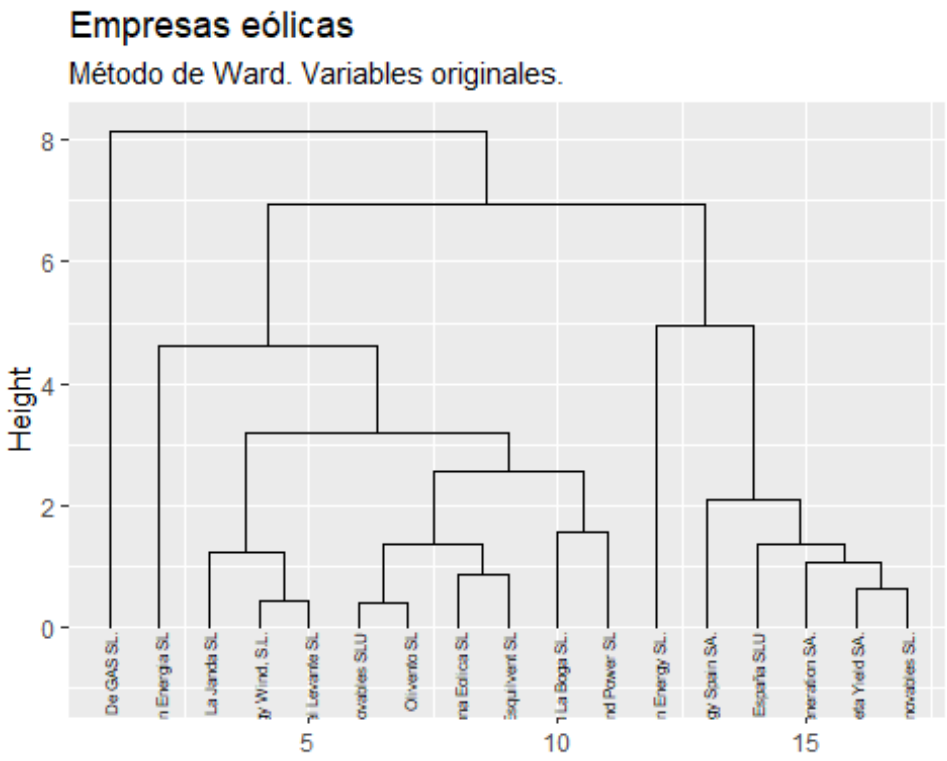
```
cluster_j<-hclust(d, method="ward.D2")
```

F1 para ayuda a la función `hclust`. "Whereas option"ward.D2" implements that criterion (Murtagh and Legendre 2014)".

VAMOS A REALIZAR EL *DENDOGRAMA* con el paquete `factoextra` y la función `fviz_dend()`:

```
fviz_dend(cluster_j, cex=0.4, rect = FALSE) +
  labs(title = "Empresas eólicas",
       subtitle = "Método de Ward. Variables originales.") +
  theme_grey()
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use
`guides(<scale> =
## "none")` instead.
```



ANÁLISIS DEL DENDOGRAMA:

1. EJE VERTICAL: recoge las distancias entre los casos y/o grupos previsos que se van agrupando.

2. Las dos empresas outliers (*Holding de Negocios de Gas* y *Elawan Energy*) se agrupan con el resto en una fase muy tardía del proceso de agrupamiento (muy cerca del único grupo).

DETERMINACIÓN DEL NÚMERO DE GRUPOS CON LOS QUE HEMOS DE QUEDARNOS:

Existen algoritmos y paquetes de R que aconsejan un número (por ejemplo, `NbClust()` del paquete `NbClust`), pero puede ser preferible que **el propio investigador decida el número de grupos a crear.**

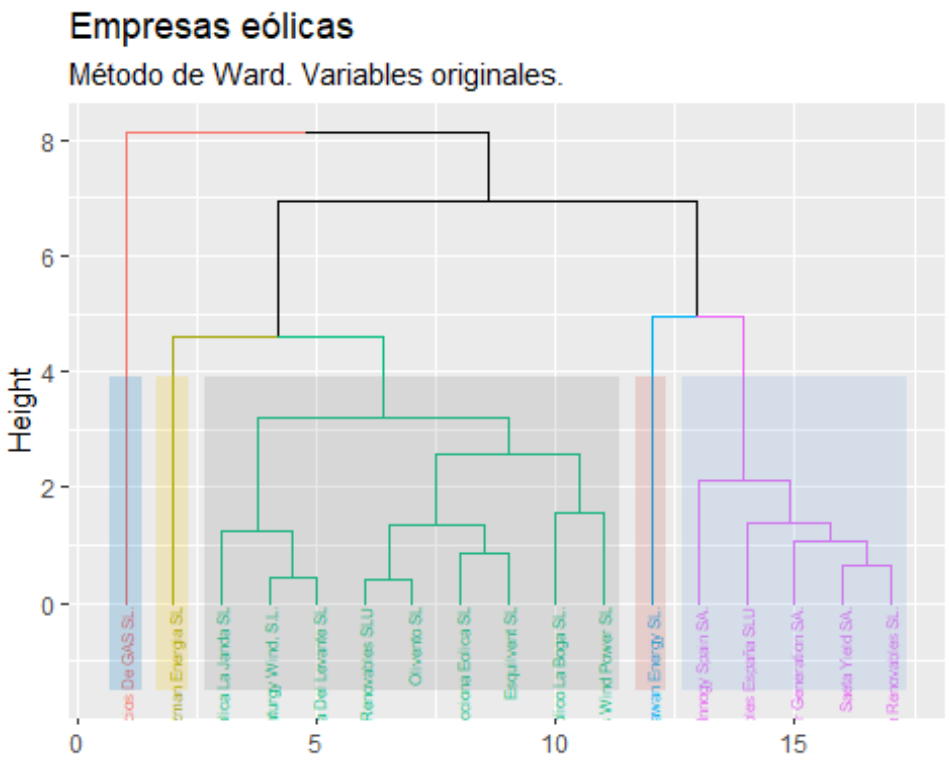
El dendograma informa de la sucesiva agrupación de casos y grupos precedentes. Además, ya que son muy pocos los casos a agrupar, las opciones (número de grupos) que se tienen son reducidas.

En este ejemplo, un número de grupos razonable podría ser 5, que contaría con el aval de mantener individualizados a las empresas outliers.

Si se acepta esta opción, se podrá visualizar de nuevo el dendograma coloreando los grupos formados, con el código siguiente:

```
fviz_dend(cluster_j, cex=0.4, k=5, kcolors = "jco", rect = TRUE,
rect_border = "jco", rect_fill = TRUE)+
  labs(title = "Empresas eólicas",
       subtitle = "Método de Ward. Variables originales.") +
  theme_grey()
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use
`guides(<scale> =
## "none")` instead.
```



Igualmente, si algún argumento no entendemos, nos vamos a la ayuda (presionamos F1):

ksignifica el número de grupos en el que cortamos el árbol. `rect`TRUE en este caso creamos los rectángulos para diferencias de los diferentes grupos creados, y tomarán los colores con el argumento `kcolor=al color que elijamos`.

RESULTADO:

1. Un grupo de 9 empresas (gris con texto en verde),.
2. Un grupo de 5 empresas (azul pálido con texto en morado),
3. Tenemos 3 empresas que no se han agrupado con ninguna otra:
    1. El outlier que mostraba una mayor distancia de Mahalanobis (y distancias euclídeas con el resto de empresas), *Holding de Negocios de Gas* (grupo azul con letra naranja),
    2. Otro outlier, *Elawan Energy* (grupo rosado con letra en azul) y,
    3. La empresa *Guzmán Energía* (grupo amarillo).

Puede afirmarse que el *análisis clúster, a partir de la observación de su dendograma, puede ser considerado un método de detección de outliers, por sí mismo.*

IDENTIFICACIÓN DE LOS CASOS CON UN MAYOR DETALLE Y CARATERIZACIÓN DE LOS GRUPOS EN FUNCIÓN DE LAS MEDIAS DE LAS VARIABLES ORIGINALES:

Pasos a seguir:

1. Creamos el *vector de valores enteros que indica el grupo al que pertenece cada empresa*.
2. El vector lo llamaremos `whatcluster_j`.
3. Usaremos la función `cuttree()`: compuesto de 2 argumentos (el primero = nombre del objeto que guarda la solución del análisis clúster `cluster_j`; el segundo = número de grupos que hemos decidido crear `k=5`).
4. Convertiremos esta variable en un factor (usamos la función `as.factor().)
5. Esta variable la incluimos en nuestro *data frame* original (no el de las variables tipificadas `zoriginales`. La incluimos con la función `cbind()).

```
whatcluster_j <- cutree(cluster_j, k=5)
whatcluster_j <- as.factor(whatcluster_j)
levels(whatcluster_j)

## [1] "1" "2" "3" "4" "5"
```

`Levels` nos proporciona acceso al atributo de niveles de una variable.

```
originales <- cbind(originales, whatcluster_j)
```

Nos vamos al *Environment* y observamos que se ha incluido esta nueva variable categórica.

OBTENCIÓN DE LAS MEDIAS DE CADA GRUPO:

1. Usamos las funciones `by_group()`y `summarise()` del paquete `dplyr`.
2. Redondeamos los decimales con la función `round()`.
3. Creamos el *data frame* `tablamedias`.

```
tablamedias <- originales %>%
  group_by(whatcluster_j) %>% summarise(obs = length(whatcluster_j),
                                        Solvencia =
round(mean(SOLVENCIA),2),
                                        Fondos_Propios =
round(mean(FPIOS),0),
                                        Margen = round(mean(MARGEN),2),
                                        Resultado = round(mean(RES),0),
                                        Rentabilidad_Financiera =
round(mean(RENFIN),2),
                                        Apalancamiento =
round(mean(APALANCA),2))
```

Representamos la tabla con los paquetes `knitr` y `kableExtra`:

```
library (knitr)
library (kableExtra)

## Warning: package 'kableExtra' was built under R version 4.2.2

## Warning in !is.null(rmarkdown::metadata$output) &&
rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows

knitr.table.format = "html"

tablamedias %>%
  kable(caption = "Método de Ward. 5 grupos. Medias de variables",
col.names = c("cluster", "observaciones", "solvencia", "FFPP", "Margen",
"Resultado", "Rentabilidad Financiera", "Apalancamiento")) %>%
  kable_styling(full_width = F, bootstrap_options = "striped",
"bordered", "condensed", position = "center", font_size = 12) %>%
  row_spec(0, bold= T, align = "c") %>%
  row_spec(1:5, bold= F, align = "c")
```

Método de Ward. 5 grupos. Medias de variables

cluster

observaciones

solvencia

FFPP

Margen

Resultado

Rentabilidad Financiera

Apalancamiento

1

1

51.17

6904824

91.15

727548

10.29

91.96

2

5

66.07

679149

15.93

21691

1.85

49.83

3

9

14.98

76425

23.95

13153

27.20

628.85

4

1

42.01

186302

208.36

12819

8.61

123.77

5

1

-40.74

-77533

-19.19

-5661

6.90

-343.54

F1 para ayuda: 1. `full_width`: toma VERDADERO o FALSO y controlará si la tabla HTML debe tener 100 el formato preferible para ancho_completo. Si no se especifica, una tabla HTML tendrá ancho completo de forma predeterminada, pero esta opción se establecerá en FALSO para una tabla LaTeX. 2. `bootstrap_options`(opición de arranque): Las opciones posibles incluyen básico, rayado, bordeado, flotante, condensado, receptivo y ninguno. En el enlace: https://www.w3schools.com/bootstrap/bootstrap_tables.asp se obtienen las disferentes posibilidades. En nuestro caso le hemos pedido: 1. *striped* : para sombrear las filas. 2. *bordered* : para incluir bordes en las filas. 3. *condensed* : para hacer una tabla más compacta.

3. `position` para determinar cómo colocar la tabla en una página. En nuestro caso, centrada.
4. `font_size = 12`: para el tamaño de la letra de las tablas.
5. `row_spec` para seleccionar una fila y especificar su apariencia: `row_spec(0, bold= T, align = "c")`. 5.1. Especificar el formato de la fila del encabezado cuando fila = 0. `row_spec(0, bold= T, align = "c")`

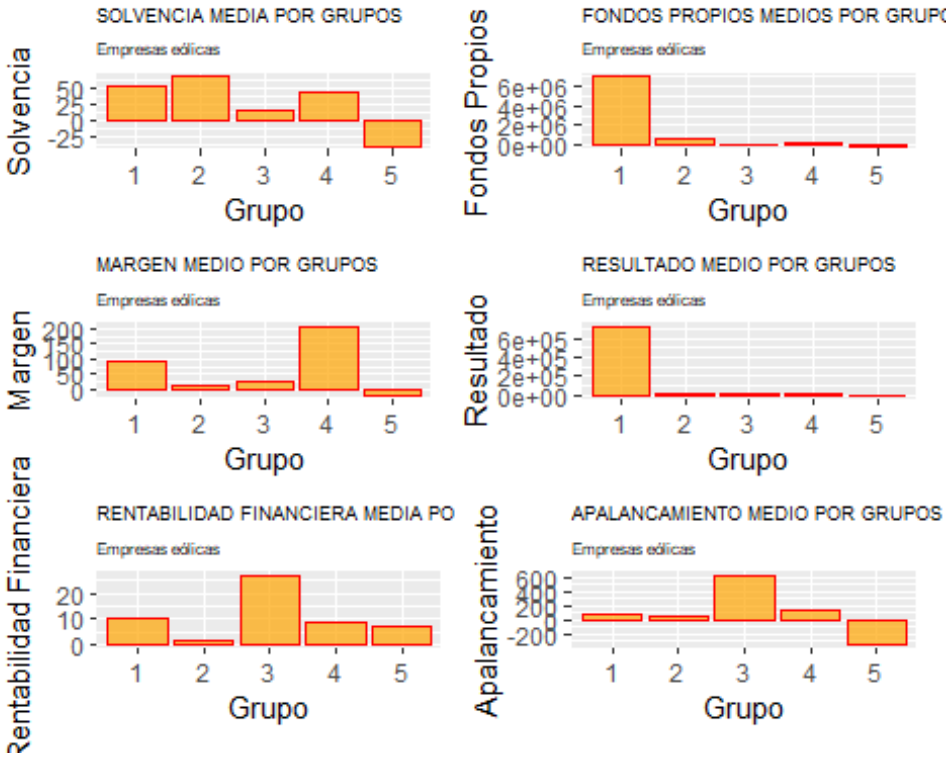
OTRA ALTERNATIVA: Comparación de las medias de los grupos, para cada variable con GRÁFICOS DE BARRAS:

Para ello vamos a crear cada gráfico de barras para cada una de las 6 variables clasificadoras y lo aportamos en un único gráfico con el paquete ya visto `patchwork`:

```
gsolve <- ggplot(data = tablamedias, map = (aes(y = Solvencia, x =
whatcluster_j))) +
           geom_bar(stat = "identity", colour = "red", fill = "orange",
alpha = 0.7) +
           ggtitle("SOLVENCIA MEDIA POR GRUPOS", subtitle = "Empresas
eólicas") +
           xlab ("Grupo") +
           ylab("Solvencia") +
           theme(plot.title= element_text(size=7), plot.subtitle =
element_text(size = 6))

gfpios <- ggplot(data = tablamedias, map = (aes(y = Fondos_Propios, x =
```

```
whatcluster_j))) +
          geom_bar(stat = "identity", colour = "red", fill = "orange",
alpha = 0.7) +
          ggtitle("FONDOS PROPIOS MEDIOS POR GRUPOS", subtitle =
"Empresas eólicas") +
          xlab ("Grupo") +
          ylab("Fondos Propios") +
          theme(plot.title= element_text(size=7), plot.subtitle =
element_text(size = 6))

gmargen <- ggplot(data = tablamedias, map = (aes(y = Margen, x =
whatcluster_j))) +
          geom_bar(stat = "identity", colour = "red", fill = "orange",
alpha = 0.7) +
          ggtitle("MARGEN MEDIO POR GRUPOS", subtitle = "Empresas
eólicas") +
          xlab ("Grupo") +
          ylab("Margen") +
          theme(plot.title= element_text(size=7), plot.subtitle =
element_text(size = 6))

gresul <- ggplot(data = tablamedias, map = (aes(y = Resultado, x =
whatcluster_j))) +
          geom_bar(stat = "identity", colour = "red", fill = "orange",
alpha = 0.7) +
          ggtitle("RESULTADO MEDIO POR GRUPOS", subtitle = "Empresas
eólicas") +
          xlab ("Grupo") +
          ylab("Resultado") +
          theme(plot.title= element_text(size=7), plot.subtitle =
element_text(size = 6))

grentf <- ggplot(data = tablamedias, map = (aes(y =
Rentabilidad_Financiera, x = whatcluster_j))) +
          geom_bar(stat = "identity", colour = "red", fill = "orange",
alpha = 0.7) +
          ggtitle("RENTABILIDAD FINANCIERA MEDIA POR GRUPOS", subtitle =
"Empresas eólicas") +
          xlab ("Grupo") +
          ylab("Rentabilidad Financiera") +
          theme(plot.title= element_text(size=7), plot.subtitle =
element_text(size = 6))

gapala <- ggplot(data = tablamedias, map = (aes(y = Apalancamiento, x =
whatcluster_j))) +
          geom_bar(stat = "identity", colour = "red", fill = "orange",
alpha = 0.7) +
          ggtitle("APALANCAMIENTO MEDIO POR GRUPOS", subtitle = "Empresas
eólicas") +
          xlab ("Grupo") +
```

```
          ylab("Apalancamiento") +
          theme(plot.title= element_text(size=7), plot.subtitle =
element_text(size = 6))

library (patchwork)

## Warning: package 'patchwork' was built under R version 4.2.1

(gsolve + gfpios) / (gmargen + gresul) / (grentf + gapala)
```



CONCLUSIONES DE LA TABLA Y DE LOS GRÁFICOS:

1. El **grupo 1** (única empresa outlier, *Holding de negocios de gas*) –> Se caracteriza sobre todo por **poseer unos fondos propios, y resultado muy superior al resto de los grupos** (a la media, si son grupos de más de una empresa).

2. E**l grupo 2** tiene, en media, una **mayor solvencia y menor apalancamiento; mientras que la rentabilidad financiera en la menor**.

3. El **grupo 3** destaca por poseer, en media, **la mayor rentabilidad financiera y grado de apalancamiento (y, por tanto, la menor solvencia).**

4. El **grupo 4** (caso atípico *Elawan Energy*), destaca por tener u**n elevado margen.**

5. El **grupo 5**, formado únicamente por la empresa *Guzmán Energía*, destaca por alcanzar v*alores negativos en solvencia, apalancamiento, y margen.*

PRESENTACIÓN EN DIFERENTES TABLAS DE LA INFORMACIÓN DE CADA GRUPO:

Para el *primer grupo*:

```
originales %>% filter(whatcluster_j == "1")%>%
  select ( SOLVENCIA, FPIOS, MARGEN, RES, RENFIN, APALANCA) %>%
  kable(caption = "Método de Ward. Grupo 1.") %>%
  kable_styling(full_width = F, bootstrap_options = "striped",
"bordered", "condensed", position = "center", font_size = 12) %>%
  row_spec(0, bold= T, align = "c")
```

Método de Ward. Grupo 1.

SOLVENCIA

FPIOS

MARGEN

RES

RENFIN

APALANCA

Holding De Negocios De GAS SL.

51.174

6904824

91.152

727548

10.287

91.964

Cambiamos el número de `whatcluster_j = X` para presentar cada una de las tablas de los siguientes grupos:

```
originales %>% filter(whatcluster_j == "2")%>%
  select ( SOLVENCIA, FPIOS, MARGEN, RES, RENFIN, APALANCA) %>%
  kable(caption = "Método de Ward. Grupo 2.") %>%
  kable_styling(full_width = F, bootstrap_options = "striped",
"bordered", "condensed", position = "center", font_size = 12) %>%
  row_spec(0, bold= T, align = "c")
```

Método de Ward. Grupo 2.

SOLVENCIA

FPIOS

MARGEN

RES

RENFIN

APALANCA

Global Power Generation SA.

86.917

1740487.00

22.403

39995.000

1.603

1.044

EDP Renovables España SLU

56.960

726783.00

47.193

67033.000

11.338

67.028

Saeta Yield SA.

83.489

665319.56

16.258

2084.476

0.432

17.067

Viesgo Renovables SL.

65.883

177707.00

11.818

4609.000

3.200

13.330

Innogy Spain SA.

37.096

85447.21

-18.025

-5268.573

-7.302

150.688

```
originales %>% filter(whatcluster_j == "3")%>%
  select ( SOLVENCIA, FPIOS, MARGEN, RES, RENFIN, APALANCA) %>%
  kable(caption = "Método de Ward. Grupo 3.") %>%
  kable_styling(full_width = F, bootstrap_options = "striped",
"bordered", "condensed", position = "center", font_size = 12) %>%
  row_spec(0, bold= T, align = "c")
```

Método de Ward. Grupo 3.

SOLVENCIA

FPIOS

MARGEN

RES

RENFIN

APALANCA

Naturgy Renovables SLU

16.274

318475.00

20.442

42737.000

12.043

494.729

Corporacion Acciona Eolica SL

15.737

136064.00

20.091

29592.000

28.990

422.263

Olivento SL

15.304

58341.00

16.629

7388.175

16.684

534.761

Parque Eolico La Boga SL.

9.646

29316.80

1.001

11.940

1.684

921.591

Naturgy Wind, S.L.

10.388

28418.00

39.575

8500.000

38.018

824.537

Al-Andalus Wind Power SL

8.591

21466.12

12.582

4403.214

27.350

1019.616

Acciona Eolica Del Levante SL

11.557

21769.00

27.520

6853.000

43.139

743.754

Esquilvent SL

30.938

48769.13

39.476

9010.214

24.633

218.275

Eolica La Janda SL

16.428

25206.75

38.256

9880.091

52.261

480.122

```
originales %>% filter(whatcluster_j == "4")%>%
  select ( SOLVENCIA, FPIOS, MARGEN, RES, RENFIN, APALANCA) %>%
  kable(caption = "Método de Ward. Grupo 4.") %>%
  kable_styling(full_width = F, bootstrap_options = "striped",
"bordered", "condensed", position = "center", font_size = 12) %>%
  row_spec(0, bold= T, align = "c")
```

Método de Ward. Grupo 4.

SOLVENCIA

FPIOS

MARGEN

RES

RENFIN

APALANCA

Elawan Energy SL.

42.01

186302

208.357

12818.98

8.605

123.771

```
originales %>% filter(whatcluster_j == "5")%>%
  select ( SOLVENCIA, FPIOS, MARGEN, RES, RENFIN, APALANCA) %>%
  kable(caption = "Método de Ward. Grupo 5.") %>%
  kable_styling(full_width = F, bootstrap_options = "striped",
"bordered", "condensed", position = "center", font_size = 12) %>%
  row_spec(0, bold= T, align = "c")
```

Método de Ward. Grupo 5.

SOLVENCIA

FPIOS

MARGEN

RES

RENFIN

APALANCA

Guzman Energia SL

-40.745

-77532.7

-19.193

-5661.463

6.904

-343.542

COMPONENTES PRINCIPALES PARA REALIZAR EL GRÁFICO DE DISPERSIÓN:

Problema del uso de las variables originales como clasificadoras cuando son más de 2: dificultad al trazar un gráfico de dispersión para dar una idea precisa de los grupos formados: su composición y su "Separación" de unos con otros.

SOLUCIÓN: realizar un gráfico de dispersión a partir de las *dos primeras componentes principales* de las variables originales.

REQUISITO: cuando se den las condiciones para aplicar esta técnica y ambas componentes recojan una elevada proporción de la comunalidad o varianza total de las variables originales.

RECORDATORIO: *condición básica para realizar componentes principales es la existencia de algunas variables con altas correlaciones*.

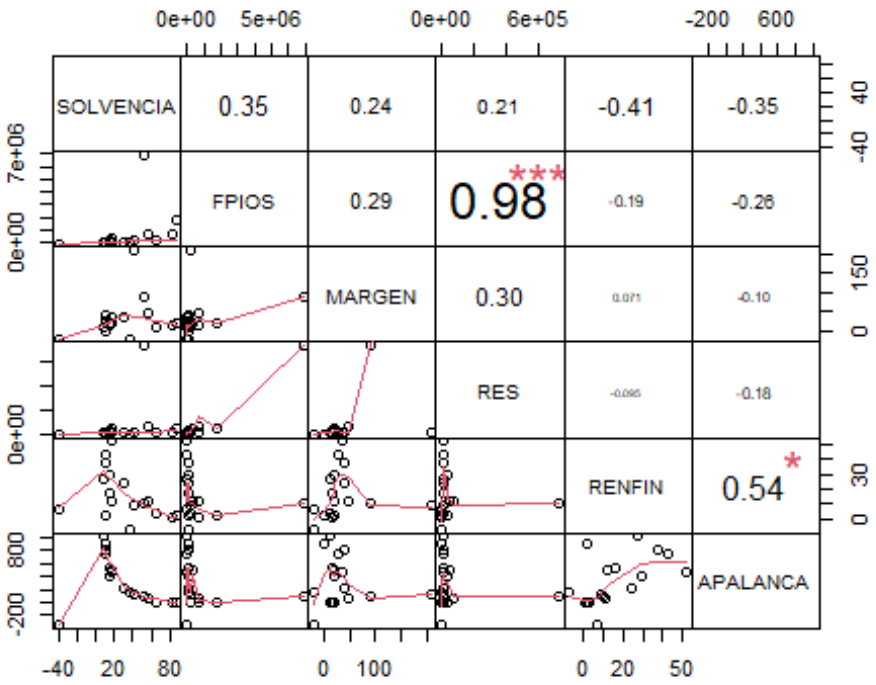
PROCEDIMIENTO:

1. Creamos la matriz de correlaciones: función `chart.Correlation ()` del paquete `PerformanceAnalytics`.
2. Previamente creación de un *data frame* que contenga solo las variables originales (sin incluir `whatcluster_j`).
3. Este *data frame* se guardará en el objeto `originales_cp`:

```
originales_cp <- originales %>% select(-whatcluster_j)
library(PerformanceAnalytics)

## Warning: package 'PerformanceAnalytics' was built under R version
4.2.1

## Loading required package: xts

## Warning: package 'xts' was built under R version 4.2.1
```

```
## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.2.1

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend

chart.Correlation(originales_cp, histogram = F, pch = 18)

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```



ANÁLISIS DEL GRÁFICO:

1. Existencia de elevadas correlaciones: FPIS con RES (0,98) y RENFIN con APALANCA (0,54).

OBTENCIÓN DE LAS COMPONENTES con la función `prcomp()`usando variables tipificadas (funcion `scale()`) y lo guardaremos en el objeto `componentes`:

```
componentes <- prcomp(originales_cp, scale=T)
summary (componentes)

## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6
## Standard deviation     1.6035 1.2188 0.9530 0.7869 0.63683 0.10069
## Proportion of Variance 0.4286 0.2476 0.1514 0.1032 0.06759 0.00169
## Cumulative Proportion  0.4286 0.6761 0.8275 0.9307 0.99831 1.00000
```

EXPLICACIÓN:

1. Las dos primeras componentes acumulan ya más de un 67% de la varianza total o comunalidad (información) que las variables originales guardan sobre el comportamiento de las empresas de la muestra (*Cumulative Proportion*).

2. La *Standard deviation* es la raíz cuadrada de los autovalores asociados a cada componente. Por tanto, podríamos utilizar esas dos componentes principales para clasificar a las empresas, en lugar de las 4 variables originales, con la ventaja de que dos variables pueden ser fácilmente representadas en un gráfico de dispersión.

OBTENCIÓN DE LAS CARGAS:

```
round(componentes$rotation, 4)
```

```
##                PC1     PC2     PC3     PC4     PC5     PC6
## SOLVENCIA -0.3914  0.2801  0.3789 -0.7186  0.3110 -0.1078
## FPIOS     -0.5460 -0.3179 -0.2863 -0.0746  0.0238  0.7161
## MARGEN    -0.2666 -0.2925  0.8058  0.3217 -0.2990  0.0326
## RES       -0.5022 -0.4142 -0.3158  0.0321 -0.0387 -0.6885
## RENFIN     0.3097 -0.5886  0.1592 -0.0006  0.7295  0.0142
## APALANCA   0.3599 -0.4658 -0.0031 -0.6111 -0.5288  0.0204
```

EXPLICACIÓN DEL RESULTADO:

1. COLUMNAS: las componentes (PC).
2. FILAS: las variables.
3. INTERSECCIÓNES: los coeficientes o *cargas*.

No vamos a emplear un método para obtener el número de componentes retenidas sugerido, ya que en este caso lo que queremos es que sean 2 dichas componentes, a fin de obtener una **representación gráfica bidimensional.** Además, sabemos que las dos componentes recogen ya una proporción apreciable del comportamiento (varianza total o comunalidad) de los individuos.

Para utilizar las puntuaciones de las dos primeras componentes, hemos de obtener sus puntuaciones de cada empresa. Estas puntuaciones están guardadas en la matriz "x" del objeto "prcomp" creado ("componentes").

1. Creamos un *data frame* con los siguientes elementos:
2. Variables originales.
3. Grupo de pertenencia.
4. Las puntuaciones de ambas componentes se denominará `Componentes` (con la C mayúscula) (a la primera componente la llamados `Componente.1` y a la segunda `Componente.2`).

```
Componente.1 <-componentes$x[,1]
Componente.2 <-componentes$x[,2]
Componentes <- cbind(originales, Componente.1, Componente.2)
summary (Componentes)
```

```
##    SOLVENCIA          FPIOS              MARGEN            RES          
##  Min.   :-40.74   Min.   : -77533   Min.   :-19.19   Min.   : -5662  
##  1st Qu.: 11.56   1st Qu.:  28418   1st Qu.: 12.58   1st Qu.:  4403  
##  Median : 16.43   Median :  85447   Median : 20.44   Median :  8500
```

```
##  Mean   : 30.45   Mean   : 652774   Mean   : 33.85   Mean   : 56561
##  3rd Qu.: 51.17   3rd Qu.: 318475   3rd Qu.: 39.48   3rd Qu.: 29592
##  Max.   : 86.92   Max.   :6904824   Max.   :208.36   Max.   :727548
##      RENFIN          APALANCA       whatcluster_j  Componente.1
##  Min.   :-7.302   Min.   :-343.54   1:1           Min.   :-4.8808
##  1st Qu.: 3.200   1st Qu.:  67.03   2:5           1st Qu.:-0.8032
##  Median :11.338   Median : 218.28   3:9           Median : 0.4576
##  Mean   :16.463   Mean   : 340.06   4:1           Mean   : 0.0000
##  3rd Qu.:27.350   3rd Qu.: 534.76   5:1           3rd Qu.: 1.0696
##  Max.   :52.261   Max.   :1019.62                 Max.   : 1.5808
##   Componente.2
##  Min.   :-2.40745
##  1st Qu.:-1.04376
##  Median : 0.01356
##  Mean   : 0.00000
##  3rd Qu.: 1.13383
##  Max.   : 1.67608
```

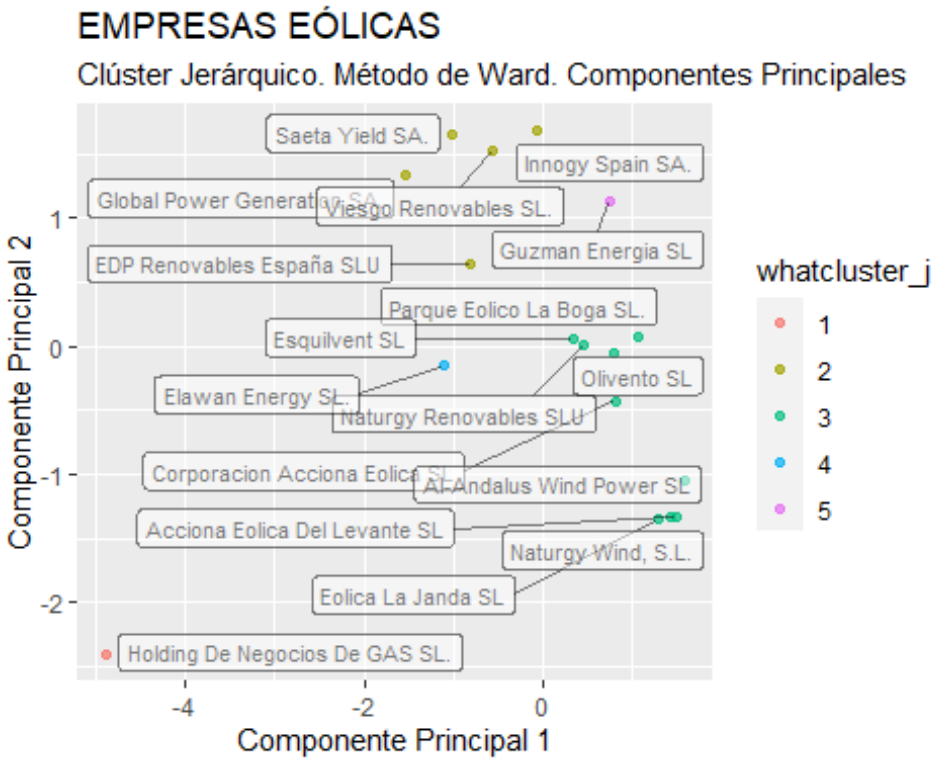
REALIZAMOS EL GRÁFICO DE DISPERSIÓN:

1. Función `ggplot()` del paquete `ggplot2()`.
2. Queremos etiquetar los casos sin etiquetas superpuestas –> para ello usamos la función `geom_label_repel()`del paquete `ggrepel`:

```
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.2.2
```

```
ggplot(data = Componentes, map = (aes(x = Componente.1, y = Componente.2,
col = whatcluster_j))) +
  geom_point(alpha = 0.7) +
  geom_label_repel(aes(label = row.names(Componentes)), size = 3, color =
"black", alpha = 0.5) +
  ggtitle("EMPRESAS EÓLICAS", subtitle = "Clúster Jerárquico. Método de
Ward. Componentes Principales") +
  xlab("Componente Principal 1") +
  ylab("Componente Principal 2")
```



EXPLICACIÓN DEL GRÁFICO:

1. Claro papel de outlier de la empresa *Holding de Negocios*, sobre todo en lo que respecta a la primera componente principal. Esto se debe a su especial comportamiento en las variables `FPIOS` y `RES`, que son, precisamente, las variables con mayor peso (con signo negativo) en tal componente.
2. En cambio, en el gráfico no existe, por ejemplo, una posición muy singular de la empresa *Guzmán Energía*, a pesar de constituir un grupo particular. Esto se puede deber a que su comportamiento especial se da en las variables `SOLVENCIA` y `APALANCA`; pero la primera de ellas no tiene un peso notable en ninguna de las dos primeras componentes.