Práctica Tema 1. TMAAS.

Consoli Quintana

2022-10-15

TEMA 1. Introducción al análisis multivariante de datos económicos: aplicación al análisis de los sectores económicos.

Análisis descriptivo inicial.

1. Preparación de los datos para que sean procesados correctamente.
2. Visión inicial de la información básica

Buenas prácticas:

- Vamos a crear un proyecto de R con el nombre `explora`:
- En él guardamos el *script* propuesto para la práctica con nombre `explora_describe.R`
- Guardamos el *excel* propuesto: `eolica_100_mv.xlsx`
- Vamos a trabajar con *la tercera hoja* `Datos`: variables económico-fras de las 100 empresas estudiadas.
- Abrimos el *script*: File –> Open File.

Vamos a *limpiar el environment*:

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

Vamos a *importar los datos del excel*:

Para ello:

1. Cargamos el paquete `readxl`:

```
library(readxl)
```

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

2. Vamos a leerlo:

```
eolica_100 <- read_excel("eolica_100_mv.xlsx", sheet = "Datos")
```

3. Resumen de los datos:

```
summary(eolica_100)
```

```
##     NOMBRE               RES                ACTIVO             FPIOS       
##  Length:100         Min.   : -5661.5   Min.   :   24944   Min.   : -
77533
##  Class :character   1st Qu.:   669.5   1st Qu.:   34547   1st Qu.:
```

```
2305
##  Mode  :character   Median :  2084.5   Median :   46950   Median :
11936
##                     Mean   : 11529.8   Mean   :  277270   Mean   :
123743
##                     3rd Qu.:  3806.7   3rd Qu.:   85610   3rd Qu.:
28292
##                     Max.   :727548.0   Max.   :13492812   Max.
:6904824
##                     NA's   :1          NA's   :1
##      RENECO           RENFIN            LIQUIDEZ           ENDEUDA
##  Min.   :-2.813   Min.   :-359.773   Min.   :  0.0140   Min.   :
0.917
##  1st Qu.: 1.558   1st Qu.:   2.556   1st Qu.:  0.6567   1st Qu.:
50.852
##  Median : 4.236   Median :  15.326   Median :  1.0650   Median :
83.346
##  Mean   : 5.416   Mean   :  17.243   Mean   :  2.7214   Mean   :
72.227
##  3rd Qu.: 7.970   3rd Qu.:  31.307   3rd Qu.:  1.6078   3rd Qu.:
95.388
##  Max.   :35.262   Max.   : 588.190   Max.   :128.4330   Max.
:140.745
##  NA's   :2                                             NA's   :2
##      MARGEN            SOLVENCIA         APALANCA           MATRIZ
##  Min.   :-2248.157   Min.   :-40.74   Min.   :-8254.11   Length:100
##  1st Qu.:   12.316   1st Qu.:  4.71   1st Qu.:   16.13   Class
:character
##  Median :   26.618   Median : 16.65   Median :  161.97   Mode
:character
##  Mean   :    3.228   Mean   : 27.57   Mean   :  345.03
##  3rd Qu.:   39.590   3rd Qu.: 45.59   3rd Qu.:  623.13
##  Max.   :  400.899   Max.   : 99.08   Max.   :12244.35
##  NA's   :2
##   DIMENSION
##  Length:100
##  Class :character
##  Mode  :character
##
##
##
##
```

Vamos a observar los datos y ver qué nos quieren decir:

- Primera columna: no es una variable, sino que son los nombres de los *casos u observaciones* de Para evitar que R lo considere como una variable:

1. Definir el *data frame* para que esa primera columna sea el conjunto de *nombres de los individuos* de nuestra muestra.

```
eolica_100 <- data.frame(eolica_100, row.names = 1)
```

row.names NULL or a single integer or character string specifying a column to be used as row names, or a character or integer vector giving the row names for the data frame.

Traducción: *NULL o un entero único o cadena de caracteres que especifica una columna que se usará como nombres de fila, o un vector de carácter o entero que proporciona los nombres de fila para el marco de datos.*

Vamos a realizar un `summary`:

```
summary(eolica_100)
```

```
##       RES                ACTIVO              FPIOS              RENECO
##  Min.   : -5661.5   Min.   :   24944   Min.   : -77533   Min.   :-
2.813
##  1st Qu.:   669.5   1st Qu.:   34547   1st Qu.:   2305   1st Qu.:
1.558
##  Median :  2084.5   Median :   46950   Median :  11936   Median :
4.236
##  Mean   : 11529.8   Mean   :  277270   Mean   : 123743   Mean   :
5.416
##  3rd Qu.:  3806.7   3rd Qu.:   85610   3rd Qu.:  28292   3rd Qu.:
7.970
##  Max.   :727548.0   Max.   :13492812   Max.   :6904824   Max.
:35.262
##  NA's   :1          NA's   :1                            NA's   :2
##      RENFIN            LIQUIDEZ           ENDEUDA            MARGEN
##  Min.   :-359.773   Min.   :  0.0140   Min.   :  0.917   Min.   :-
2248.157
##  1st Qu.:   2.556   1st Qu.:  0.6567   1st Qu.: 50.852   1st Qu.:
12.316
##  Median :  15.326   Median :  1.0650   Median : 83.346   Median :
26.618
##  Mean   :  17.243   Mean   :  2.7214   Mean   : 72.227   Mean   :
3.228
##  3rd Qu.:  31.307   3rd Qu.:  1.6078   3rd Qu.: 95.388   3rd Qu.:
39.590
##  Max.   : 588.190   Max.   :128.4330   Max.   :140.745   Max.   :
400.899
##                                        NA's   :2         NA's   :2
##    SOLVENCIA         APALANCA           MATRIZ           DIMENSION
##  Min.   :-40.74   Min.   :-8254.11   Length:100         Length:100
##  1st Qu.:  4.71   1st Qu.:   16.13   Class :character   Class
:character
##  Median : 16.65   Median :  161.97   Mode  :character   Mode
:character
##  Mean   : 27.57   Mean   :  345.03
##  3rd Qu.: 45.59   3rd Qu.:  623.13
##  Max.   : 99.08   Max.   :12244.35
##
```

Ya no aparece la primera columna “NOMBRE”, puesto que ya no es considera como una variable.

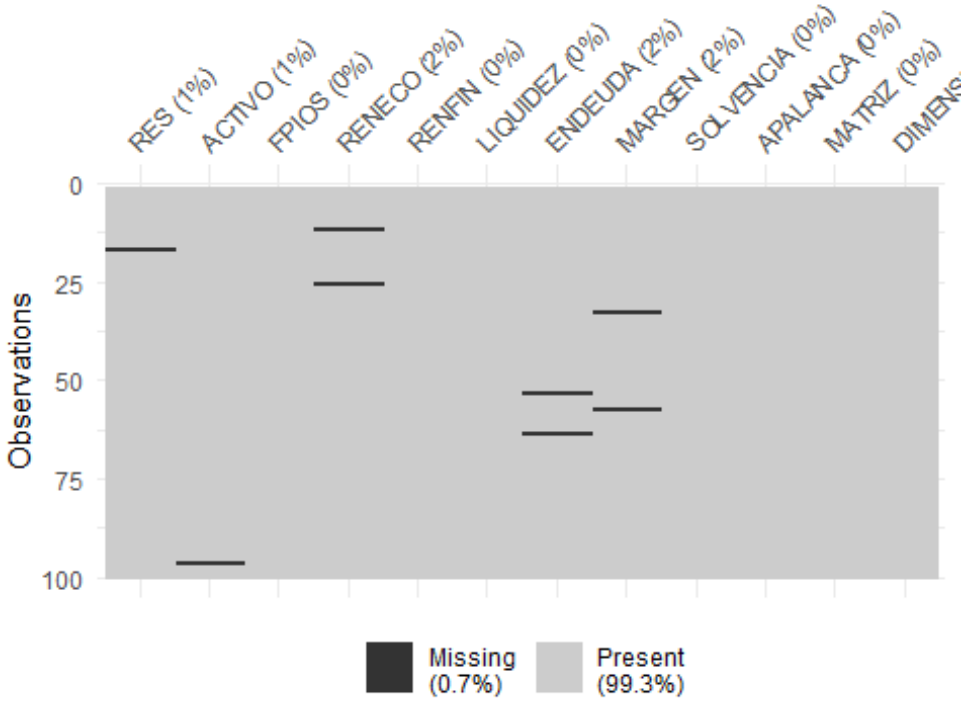
Comenzamos el análisis:

ANÁLISIS DE UNA SOLA VARIABLE. BÚSQUEDA DE *VALORES PERDIDOS* O *MISSING VALUES*:

Comenzamos analizando la variable *Rentabilidad Económica (RENECO)*:

Usamos *el paquete* `visdat`, para ello cargamos el paquete. La *función* que usaremos es `vis_miss`

```
library(visdat)

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

vis_miss(eolica_100)

## 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.
```



Analizamos el *gráfico*; en concreto la variable `RENECO`:

El gráfico nos muestra que *una total del 2% de los casos no tienen dato* esto es: si tenemos 100 empresas, no disponemos datos de 2 de ellas.

PREGUNTA: ¿Qué significan las líneas negras?

Localización de los datos concretos de valores faltantes:

- Usamos el paquete `dplyr` para ello lo cargamos.
- *Buenas prácticas*: hacemos una *copia del data frame original* y la llamaremos `muestra`:
- Filtramos los casos para *detectar aquellos que carecen de valor* de la variable `RENECO`:

```
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

muestra<- select(eolica_100, everything())
muestra %>% filter(is.na(RENECO)) %>% select(RENECO)

##                       RENECO
## Viesgo Renovables SL.     NA
## Sargon Energias SLU       NA

muestra <- muestra %>% filter(! is.na(RENECO))
```

Vamos a ver qué significa cada línea del script:

Línea 105: `muestra<- select(eolica_100, everything())`: la función `select`. Si pulsamos sobre `select`F1 nos lleva al visor a la ayuda `Help`. Bajamos a ***`Arguments`***:

.data A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from dbplyr or dtplyr). See Methods, below, for more details.

One or more unquoted expressions separated by commas. Variable names can be used as if they were positions in the data frame, so expressions like x:y can be used to select a range of variables.

Es decir, escogemos toda la base de datos `eolica_100`y la incluimos en el objeto que se llamará `muestra`. Es importante incluir el segundo argumento que es la que nos incluye también las variables.

Línea 106: `muestra %>% filter(is.na(RENECO)) %>% select(RENECO)`

Significado: La base de datos `muestra` vamos a introducirla con el `pipe` en la función `is.na()`cuya utilidad es: comprobar si en la posición correspondiente a una fila o variable (en nuestro caso `RENECO`) hay o no un dato o valor. Como resultado: las dos empresas son: *Viesgo Renovables SL* y *Sargon Energias SLU*

¿Cómo actuar ante la existencia de *missing values*?

Varias formas:

1. *Intentar obtener por otro canal de información el conjunto de valores de `RENECO` que no están disponibles*
2. *Recurrir a alguna estimación para los mismos y asignarlos*
3. En caso de imposibilidad o gran dificultad: *eliminar los casos* (especialmente si representan un % muy reducido respecto al total de los casos)

Vamos a optar por el 3er caso: ELIMINACIÓN DE MISSING VALUES:

Usando el operador "`!`" le pedimos a R que *no* lo incluya:

```
muestra <- muestra %>% filter(! is.na(RENECO))
```

Ahora si observamos en el *Environment* el *data frame* `muestra` aparece con 2 variables menos: pasamos de 100 empresas a 98.

Siguiente paso: *detección de presencia de outliers*:

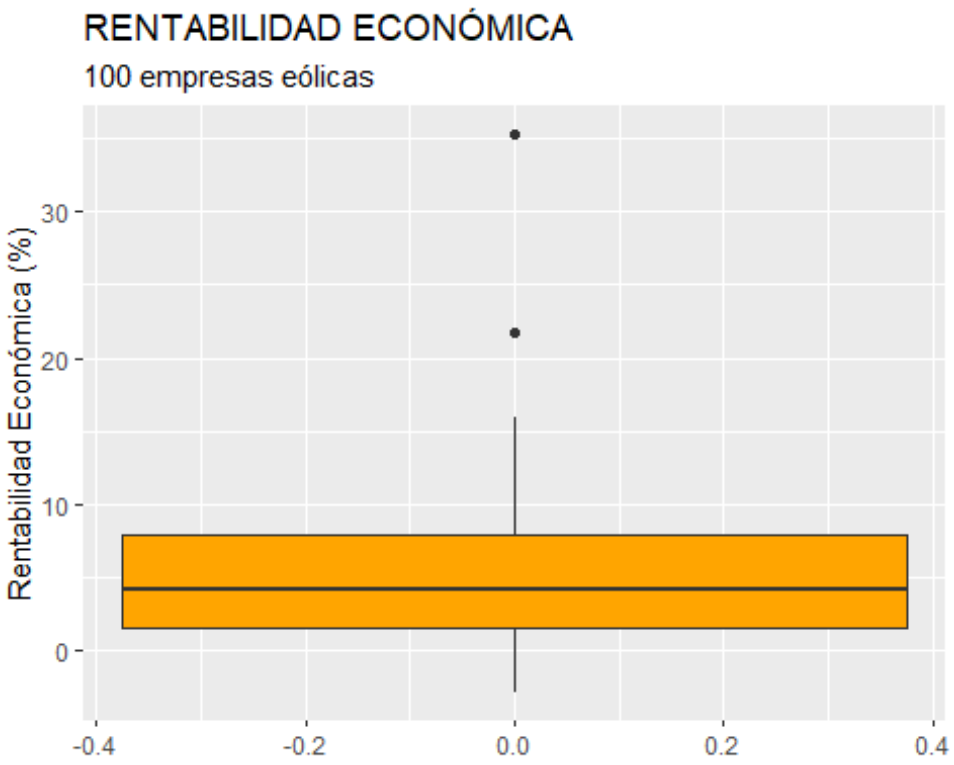
OUTLIERS O CASOS ATÍPICOS EN LA MUESTRA:

Como ya vimos en la asignatura *TAEDE* podemos *representar gráficamente* la variable para detectar la presencia de *outliers*

Gráfico usado: *boxplot* o *gráfico de caja*.

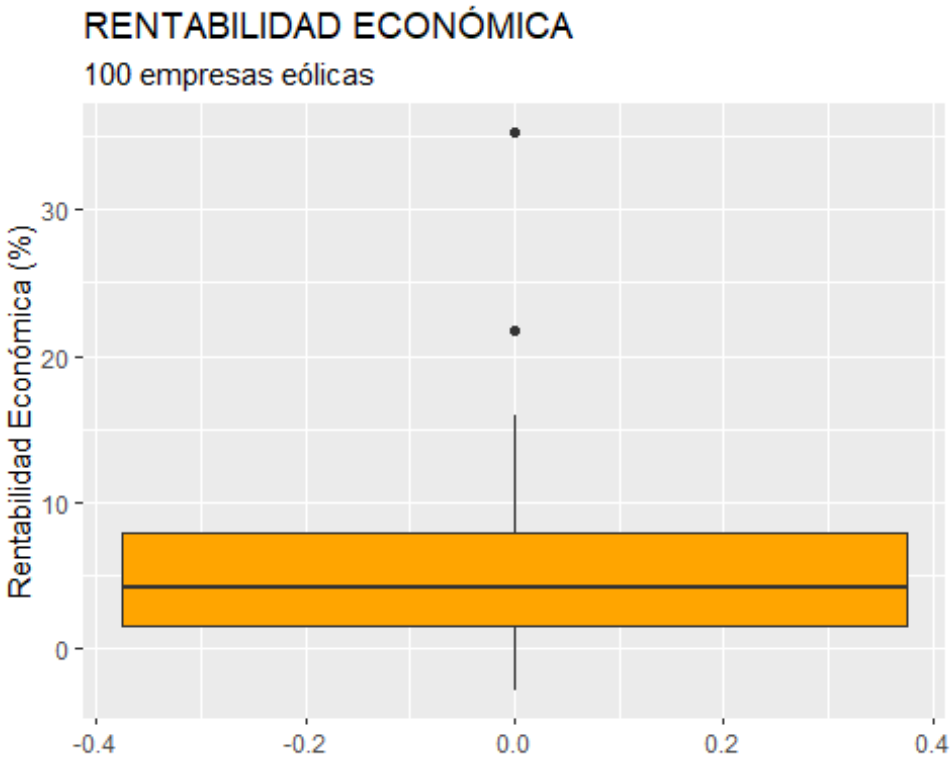
Recordamos que hay que cargar el paquete `ggplot2`:

```
library (ggplot2)
ggplot(data = muestra, map = (aes(y = RENECO))) +
    geom_boxplot(fill = "orange") +
    ggtitle("RENTABILIDAD ECONÓMICA", subtitle = "100 empresas eólicas")
+
    ylab("Rentabilidad Económica (%)")
```



Recordad que R considera un gráfico como un lienzo en el que vamos añadiendo capas con el operador +: Podemos escribir de diferentes formas para que me realice la misma operación, en *TAEDE* usábamos:

```
muestra %>% ggplot(aes(y = RENECO)) +
    geom_boxplot(fill = "orange") +
    ggtitle("RENTABILIDAD ECONÓMICA", subtitle = "100 empresas eólicas")
+
    ylab("Rentabilidad Económica (%)")
```



Podéis usar la expresión con la que más cómodos estéis.

_*¿Analizamos cada una de las líneas del código?:

Realizo un *gráfico* del *data frame* `muestra`y escojo la variable `RENECO`: `ggplot(data = muestra, map = (aes(y = RENECO))) +`

El gráfico escogido es *gráfico de cajas o boxplot* y quiero que me lo coloree en naranja con el *argumento* `orange`. `geom_boxplot(fill = "orange") +`

Quiero que me ponga un título que sea *RENTABILIDAD ECONÓMICA* con la función `ggtitle` y el argumento `subtitle`para poner un subtítulo (*100 empresas eólicas*). (En realidad hemos eliminado 2 empresas pues no disponíamos datos):
`ggtitle("RENTABILIDAD ECONÓMICA", subtitle = "100 empresas eólicas") +`

Por último, le pido a R que me ponga un título o *etiqueta* a la variable *y*: *Rentabilidad Económica (%)*: `ylab("Rentabilidad Económica (%)"`

Recordatorio: La caja contiene el 50% de los casos centrales: del primer cuartil al tercero, pasando por la línea horizontal que es la mediana. Las líneas verticales son cada uno de los datos que sobresalen del *rango intercuartílico*. Por arriba de la caja, llega hasta el valor más grande (no atípico). Por debajo de la caja, el mínimo. Los puntos significan `valor atípico`: _el valor que se aleja más de 1,5 veces del rango intercuartílico del tercer cuartil (por arriba) o del primer cuartil (por abajo).

PREGUNTA: ¿Cuántos *outliers* o valores atípicos tenemos de la variable `RENECO`?:

DOS.

¿Cómo podemos identificar los casos exactos?

IDENTIFICACIÓN DE CASOS ATÍPICOS EN *R* MEDIANTE FILTROS:

Usamos el siguiente código:

```
Q1 <- quantile (muestra$RENECO, c(0.25))
Q3 <- quantile (muestra$RENECO, c(0.75))
Q1; Q3
```

```
##     25%
## 1.55775
```

```
##  75%
## 7.97
```

He pedido a R que cree los `objetos Q1 y Q3` es decir los *cuartiles primero (Q1) y tercero (Q3)* mediante la función `quantile`:

`Q1 <- quantile (muestra$RENECO, c(0.25))` y `Q3 <- quantile (muestra$RENECO, c(0.75))`

En realidad esta función calcula los percentiles.

Para calcular los *outliers*:

Tenemos que introducir la siguiente expresión para filtrar (usando la función `filter`) por aquellos casos con valores de `RENECO` mayores que Q3 más 1,5 veces el rango intercuartílico de la variable; o menores que Q1 menos 1,5 veces dicho rango intercuartílico.

Para calcular el *rango intercuartílico*: `función IQR` Función `select`, para mostrar los datos.

Vamos a ello:

```
muestra %>% filter(RENECO > Q3 + 1.5*IQR(RENECO) | RENECO < Q1 -
1.5*IQR(RENECO)) %>% select(RENECO)
```

```
##                      RENECO
## Molinos Del Ebro SA 35.262
## Sierra De Selva SL  21.761
```

Las dos *empresas atípicas* en función a la variable *Rentabilidad Económica* son *Molinos del Ebro SA* y *Sierra De Selva SL*

¿Cómo actuar ante la existencia de *outliers*?

Al igual que con los *missing values*, su tratamiento dependerá de la información que tengamos (corrección, estimación,…), pero si no tenemos información fiable, y *los*

outliers no representan un gran % respecto al total de casos (en nuestro caso 2/100) podemos optar por su eliminación.

Optamos por la ELIMINACIÓN DE OUTLIERS:

Finalidad de la eliminación de los dos casos: *no distorsionar los resultados en la aplicación posterior de ciertas técnicas*.

¿Cómo lo hacemos? Seguimos el siguiente código:

```
muestra_so <- muestra %>% filter(RENECO <= Q3 + 1.5*IQR(RENECO) & RENECO
>= Q1 - 1.5*IQR(RENECO))
```

Hemos creado un nuevo *data frame* al que llamamos `muestra_so`(muestra sin outliers). Ese objeto en realidad es: el *data frame* `muestra`es filtrada de nuevo, pero *nòtese* que: *las desigualdades deben cambiar, así como el operador "|" por el operador "&".*

Se observa cómo en el *Global Environment* pasamos de 98 empresas (`muestra`) a 96 empresas (`muestra_so`).

Comenzamos ahora la *Descripción de una variable*:

Queremos hacernos una *idea inicial de la estructura del sector para la variable o variables analizadas* –> usamos para ello: 1- *Medidas descriptivas* y 2- *Gráficos básicos*.

Recordatorio de la asignatura de TAEDE se ha visto (unidades 8, 9 y 10) *usamos medias y/o gráficos de posición, dispersión y forma (asimetría y curtosis)*.

Vistazo rápido a las medidas descriptivas básicas de la variable usando la función `descr`:

Pasos a seguir: 1. Cargamos el paquete `summarytools` para utilizar la función `descr`.

```
library (summarytools)

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

descr(muestra_so$RENECO,
      stats = c("mean", "sd", "min", "q1", "med", "q3", "max", "iqr",
"cv"),
      transpose = FALSE,
      style = "simple",
      justify = "center",
      headings = T)

## Descriptive Statistics
## muestra_so$RENECO
## N: 96
##
##                RENECO
## ------------- --------
```

```
##     Mean         4.94
##    Std.Dev      4.31
##      Min       -2.81
##      Q1         1.42
##    Median       4.14
##      Q3         7.84
##      Max       15.88
##      IQR        6.40
##      CV         0.87
```

En TAEDE usábamos la siguiente expresión. Ver la diferencia cada uno de forma individual.

```
library(summarytools)
muestra_so %>%
  select(RENECO) %>%
  descr()
```

```
## Descriptive Statistics
## muestra_so$RENECO
## N: 96
##
##                      RENECO
## ----------------- --------
##              Mean     4.94
##           Std.Dev     4.31
##               Min    -2.81
##                Q1     1.42
##            Median     4.14
##                Q3     7.84
##               Max    15.88
##               MAD     4.12
##               IQR     6.40
##                CV     0.87
##          Skewness     0.55
##       SE.Skewness     0.25
##          Kurtosis    -0.43
##           N.Valid    96.00
##         Pct.Valid   100.00
```

En este caso hemos pedido a *R*: desviación típica, valor mínimo, primer cuartil, mediana, tercer cuartil, valor máximo, rango intercuartílico, y coeficiente de variación o CV de la variable `RENECO`.

Recordatorio: CV mide el grado de *apuntamiento o curtosis* de la distribución, suponiendo que es acampanada y unimodal, de modo que un valor sensiblemente menor que cero indica que la distribución es “achatada” o platicúrtica, un valor sensiblemente mayor a uno evidencia una distribución muy “apuntada” o leptocúrtica, y un valor próximo a uno indica una distribución de apuntamiento normal, es decir,

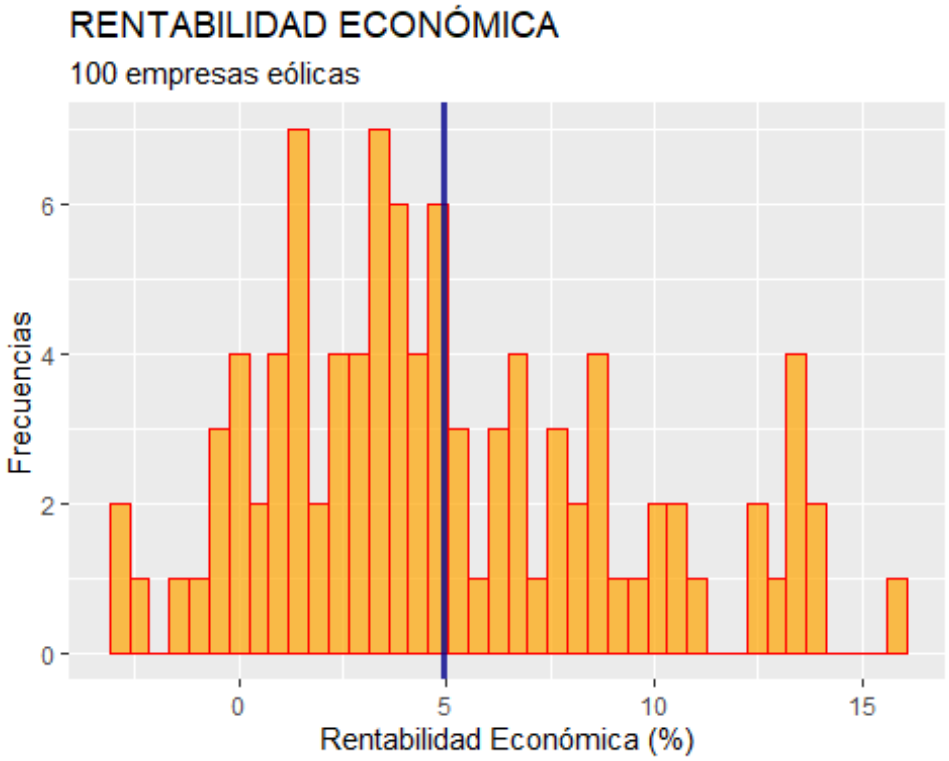
una distribución mesocúrtica. Vimos en TAEDE que podíamos representar gráficamente también su comportamiento.

Realizar un análisis gráfico suele dar una idea más atractiva de la estructura de la distribución de frecuencias en relación con la variable a analizar

PREGUNTA: ¿Cuál era el gráfico usado para representar la estructura de la distribución de frecuenccias de esta variable?

HISTOGRAMA mediante GGPLOT2:

```
ggplot(data = muestra_so, map = aes(x = RENECO)) +
geom_histogram(bins = 40, colour = "red", fill = "orange", alpha = 0.7) +
geom_vline(xintercept = mean(muestra_so$RENECO), color = "dark blue",
size = 1.2, alpha = 0.8) +
ggtitle("RENTABILIDAD ECONÓMICA", subtitle = "100 empresas eólicas")+
xlab("Rentabilidad Económica (%)") +
ylab("Frecuencias")
```

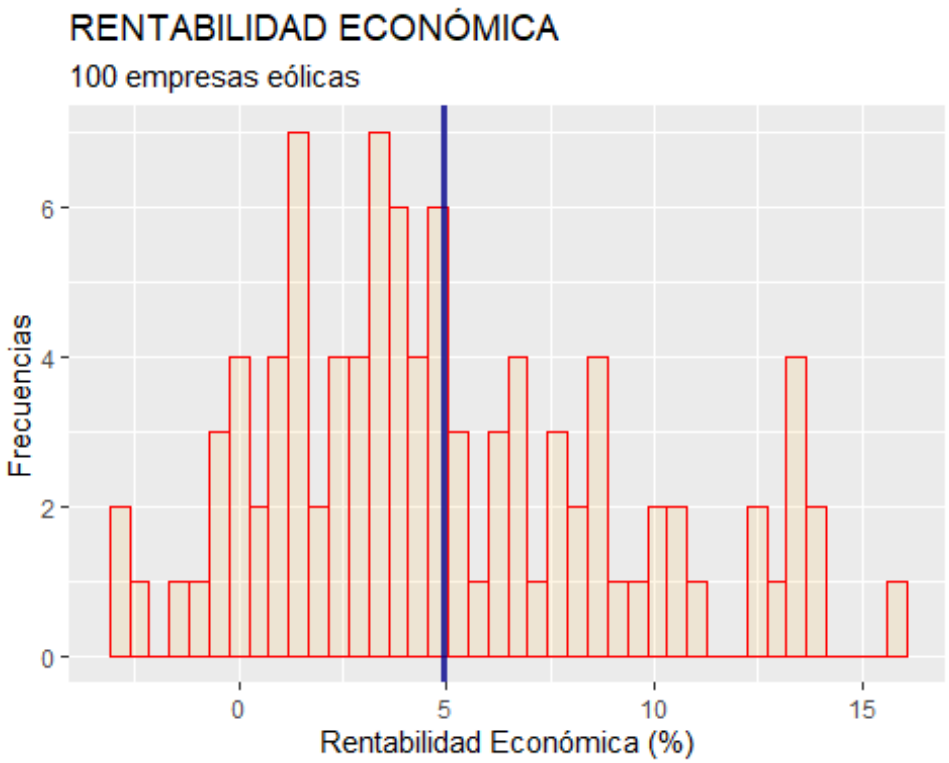


Recordad que el argumento `alpha`nos medía el *nivel de transparencia*. Imaginad que queremos un sombreado naranja muy tenue. Pasamos de un *alpha = 0.7 a un alpha = 0.1*.

Representamos con ese cambio en la segunda línea del código:

```
ggplot(data = muestra_so, map = aes(x = RENECO)) +
geom_histogram(bins = 40, colour = "red", fill = "orange", alpha = 0.1) +
geom_vline(xintercept = mean(muestra_so$RENECO), color = "dark blue",
```

```
size = 1.2, alpha = 0.8) +
ggtitle("RENTABILIDAD ECONÓMICA", subtitle = "100 empresas eólicas")+
xlab("Rentabilidad Económica (%)") +
ylab("Frecuencias")
```



En este gráfico se ha pedido introducir una *línea vertical azul* para indicar la rentabilidad media, con la función `geom_vline`. (Recordad que podéis revisar la ayuda de *R* con F1)

`geom_vline(): xintercept`

`geom_hline(): yintercept`

En nuestro caso, para la X con el siguiente código:

```
geom_vline(xintercept = mean(muestra_so$RENECO), color = "dark blue",
size = 1.2, alpha = 0.8)
```

ESTUDIAMOS LA NORMALIDAD:

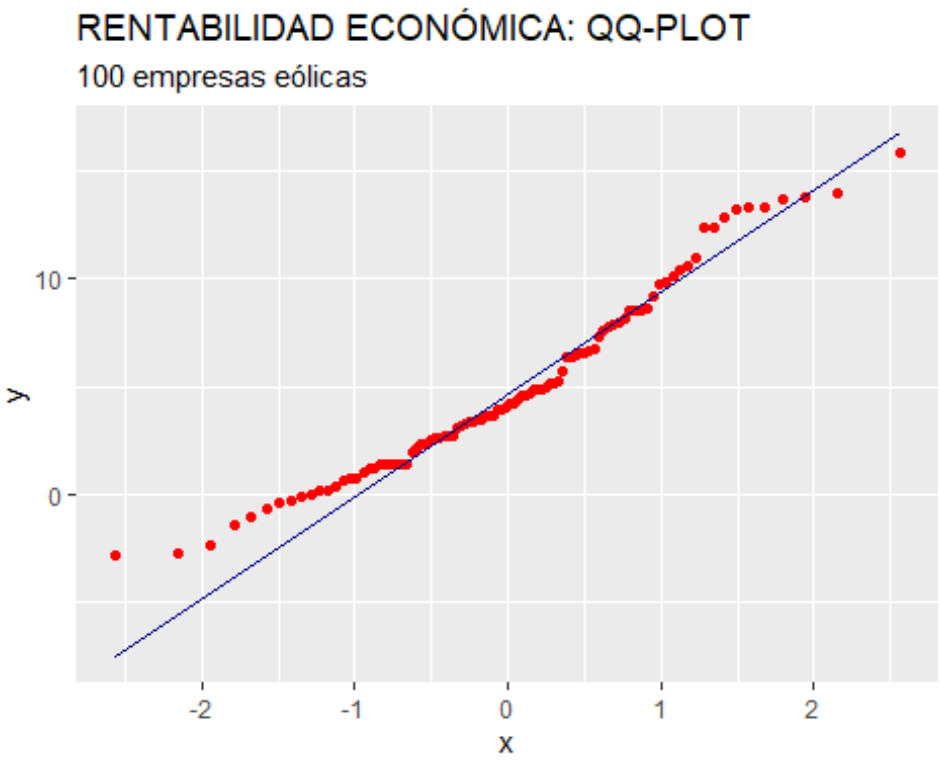
Podemos comprobar si las variables siguen una *distribución* mediante un análisis gráfico. También podemos recurrir a análisis formales basados en contrastes de normalidad.

Comprobación de la normalidad mediante método *gráfico*:

GRÁFICO QQ (cuantil-cuantil):

Para comparar los cuantiles de nuestra muestra con los de una distribución normal teórica usamos el *gráfico QQ* con la función `stat_qq`y `stat_qq_line`. Si los puntos se sitúan cercanos a la diagonal, entonces se asumirá un comportamiento (aproximadamente) normal:

```
ggplot(data = muestra_so, aes(sample = RENECO)) +
  stat_qq(colour = "red") +
  stat_qq_line(colour = "dark blue") +
  ggtitle("RENTABILIDAD ECONÓMICA: QQ-PLOT", subtitle = "100 empresas
eólicas")
```



También se podía representar el histograma y sacábamos conclusiones (TAEDE) no definitivas.

CONCLUSIÓN DEL GRÁFICO QQ:

Es difícil obtener una conclusión sólida. Se aprecia, una separación notable de los puntos primeros con respecto a la línea, lo que induce a pensar en que podría ***no* seguirse una distribución normal.**

Realizamos la **prueba de normalidad de Shapiro-Wilk:** con un buen comportamiento en muestras relativamente pequeñas.

Recordad:

- Hipótesis nula: normalildad.
- Para un nivel de significación del 5%, un p-valor superior a 0,05 implica el **no rechazo** de normalidad.

```
shapiro.test(x = muestra_so$RENECO)

##
##  Shapiro-Wilk normality test
##
## data:  muestra_so$RENECO
## W = 0.9605, p-value = 0.005523
```

Como el p-valor es (muy) inferior a 0.05, se rechaza la hipótesis nula. Para una significación estadística del 5%, admitimos que `RENECO` no sigue, para nuestra muestra, un comportamiento normal, como ya se anticipó con el gráfico qq.

ANÁLISIS DE DOS VARIABLES. BÚSQUEDA DE *VALORES PERDIDOS* O *MISSING VALUES*:

Es la aplicación de lo estudiado en las unidades 9 y 10 de TAEDE.

Nos centramos en el caso de *variables métricas*, cuantitativas.

Algunas técnicas multivariantes son el *análisis de componentes principales, el análisis de regresión, el análisis clúster…*

Todos estos análisis requieren de una fase inicial de: 1. Puesta a punto de la base de datos 2. Que ofrezca una fotografía inicial de cómo es el sector en cuanto a las variables en estudio. 3. Conveniente aplicar, para cada variable, algunos de los análisis gráficos básicos vistos. 4. Comprobación del *grado de intensidad en la relación estadística* entre las variables implicadas.

Incorporación de la variable `ACTIVO`: *volumen de activos de la empresa en miles de euros*.

Detección de valores perdidos o missing values:

Igualmente, trabajamos con una copia del *data frame* original (`eolica_100`) –> `muestra2`.

```
muestra2 <- select(eolica_100, everything())
muestra2 %>% filter(is.na(RENECO) | is.na(ACTIVO)) %>% select(RENECO,
ACTIVO)

##                               RENECO ACTIVO
## Viesgo Renovables SL.             NA 269730
## Sargon Energias SLU               NA  85745
## La Caldera Energia Burgos SL   2.643     NA
```

```
muestra2 <- muestra2 %>% filter(! is.na(RENECO) & ! is.na(ACTIVO) )
```

Recordatorio: el operador "|" significa "o" y el operador "&" significa "y".

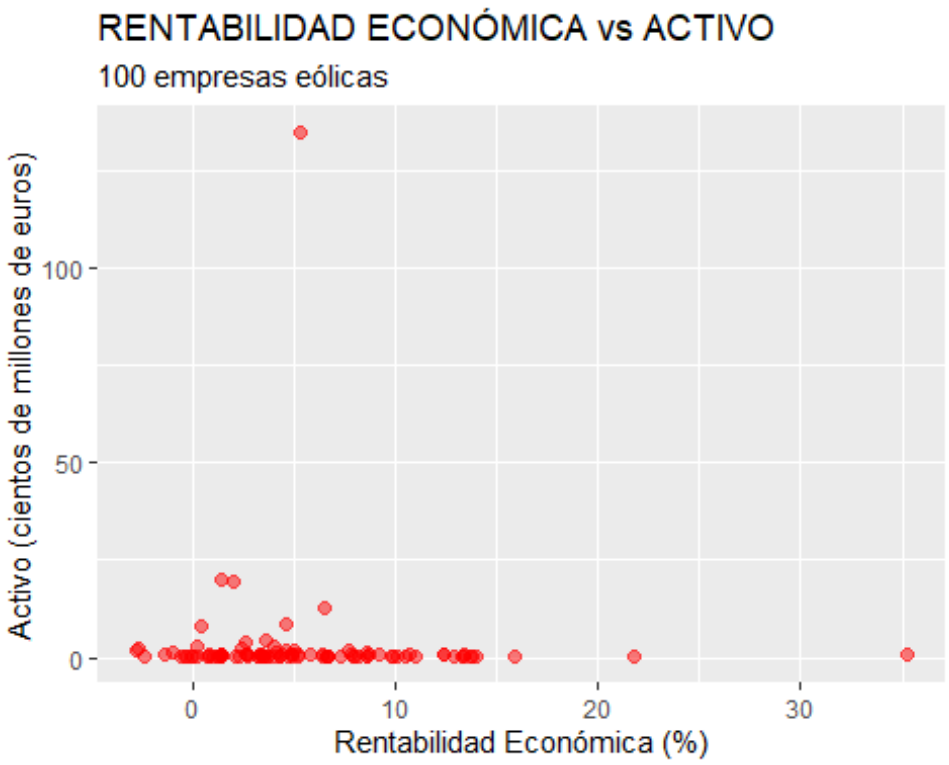
En la *línea 474* proponemos el tratamiento de los *missing values* eliminándolos.

Si observamos el *Global Environment* vemos que la *muestra2* está compuesta por 97 observaciones (100 empresas, menos las 3 de las que no obtenemos datos: 2 para la variable `RENECO`y 1 para la variable `ACTIVO`).

Detección de valores atípicos o outliers:

Una posibilidad: generar un gráfico o diagrama de dispersión.

```
dispersion <- ggplot(data = muestra2, map = (aes(x = RENECO, y =
ACTIVO/100000))) +
              geom_point(colour = "red", size = 2, alpha = 0.5) +
              ggtitle("RENTABILIDAD ECONÓMICA vs ACTIVO", subtitle = "100
empresas eólicas") +
              xlab("Rentabilidad Económica (%)") +
              ylab("Activo (cientos de millones de euros)")
dispersion
```



Los puntos muy alejados del resto serían los *outliers*

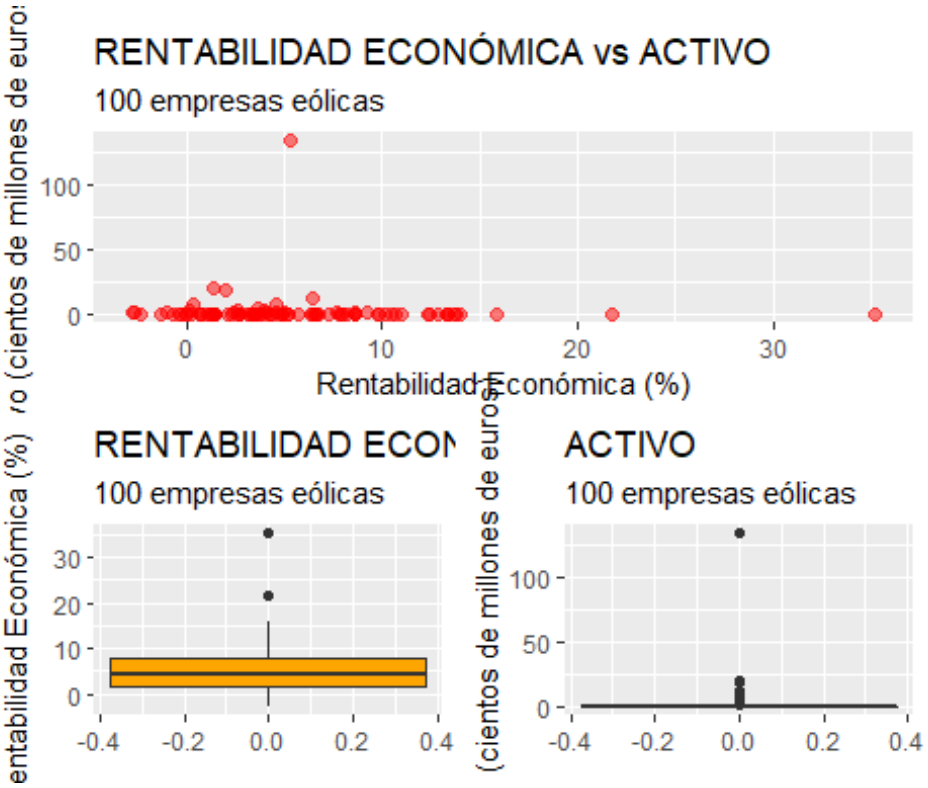
Otra posibilidad: generar gráficos de caja para cada una de las variables:

Estos gráficos los vamos a asignar a dos objetos, "caja_RENECO" y "caja_ACTIVO".

```
caja_RENECO <- ggplot(data = muestra2, map = (aes(y = RENECO))) +
               geom_boxplot(fill = "orange") +
               ggtitle("RENTABILIDAD ECONÓMICA", subtitle = "100 empresas
eólicas") +
               ylab("Rentabilidad Económica (%)")

caja_ACTIVO <- ggplot(data = muestra2, map = (aes(y = ACTIVO/100000))) +
               geom_boxplot(fill = "orange") +
               ggtitle("ACTIVO", subtitle = "100 empresas eólicas") +
               ylab("Activo (cientos de millones de euros)")
```

Además, usamos el paquete `patchwork`para realizar gráficos más atractivos, combinando los 3 gráficos anteriores:

```
library (patchwork)

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

dispersion / (caja_RENECO | caja_ACTIVO)
```



Línea 522

El operador "/" indica que los gráficos siguientes se dispondrán inmediatamente debajo. El operados "|" indica que el gráfico siguiente se dispone al lado del anterior.

Para el gráfico de la variable `ACTIVO`tenemos un caso muy destacado de *outlier*.

Proponemos el tratamiento de los *outliers* eliminándolos.

Aplicamos la misma técnica: detección y eliminación:

1. *Creamos los objetos Q1 y Q3 para cada una de las variables*:

```
Q1_RENECO <- quantile (muestra2$RENECO, c(0.25))
Q3_RENECO <- quantile (muestra2$RENECO, c(0.75))
Q1_ACTIVO <- quantile (muestra2$ACTIVO, c(0.25))
Q3_ACTIVO <- quantile (muestra2$ACTIVO, c(0.75))

muestra2 %>% filter(RENECO > Q3_RENECO + 1.5*IQR(RENECO) | RENECO <
Q1_RENECO - 1.5*IQR(RENECO) | ACTIVO > Q3_ACTIVO + 1.5*IQR(ACTIVO) |
ACTIVO < Q1_ACTIVO - 1.5*IQR(ACTIVO)) %>% select(RENECO, ACTIVO)

##                                 RENECO      ACTIVO
## Holding De Negocios De GAS SL.  5.264 13492812.00
## Global Power Generation SA.     1.393  2002458.00
## Naturgy Renovables SLU          1.959  1956869.00
## EDP Renovables España SLU       6.458  1275939.00
## Corporacion Acciona Eolica SL   4.562   864606.00
## Saeta Yield SA.                 0.360   796886.38
## Elawan Energy SL.               3.615   443467.00
## Olivento SL                     2.553   381206.98
## Parque Eolico La Boga SL.       0.162   303904.36
## Naturgy Wind, S.L.              3.949   273542.00
## Al-Andalus Wind Power SL        2.349   249853.83
## Innogy Spain SA.               -2.708   230338.51
## Guzman Energia SL              -2.813   190286.98
## Acciona Eolica Del Levante SL   4.985   188354.00
## Biovent Energia SA              4.551   183899.00
## Esquilvent SL                   7.621   157630.62
## Molinos Del Ebro SA            35.262    62114.37
## Sierra De Selva SL             21.761    27728.00
```

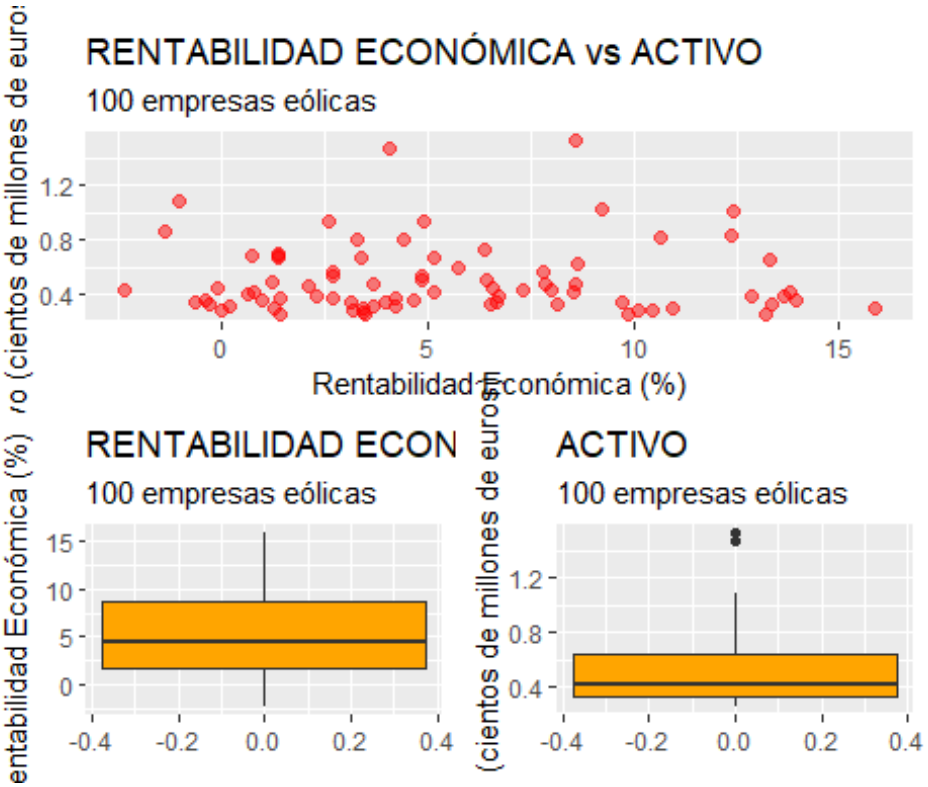
Son 18 las empresas atípicas que se ven en la tabla. Procedemos a su eliminación, creando un nuevo *data framee* que llamaremos `muestra2_so`:

```
muestra2_so <- muestra2 %>% filter(RENECO <= Q3_RENECO + 1.5*IQR(RENECO)
& RENECO >= Q1_RENECO - 1.5*IQR(RENECO) & ACTIVO <= Q3_ACTIVO +
1.5*IQR(ACTIVO) & ACTIVO >= Q1_ACTIVO - 1.5*IQR(ACTIVO))
```

Comprobación de los gráficos con la nueva muestra, sin las 18 empresas "atípicas":

```
dispersion_so <- ggplot(data = muestra2_so, map = (aes(x = RENECO, y =
ACTIVO/100000))) +
  geom_point(colour = "red", size = 2, alpha = 0.5) +
  ggtitle("RENTABILIDAD ECONÓMICA vs ACTIVO", subtitle = "100 empresas
eólicas") +
  xlab("Rentabilidad Económica (%)") +
  ylab("Activo (cientos de millones de euros)")

caja_RENECO_so <- ggplot(data = muestra2_so, map = (aes(y = RENECO))) +
  geom_boxplot(fill = "orange") +
```

```
  ggtitle("RENTABILIDAD ECONÓMICA", subtitle = "100 empresas eólicas") +
  ylab("Rentabilidad Económica (%)")

caja_ACTIVO_so <- ggplot(data = muestra2_so, map = (aes(y =
ACTIVO/100000))) +
  geom_boxplot(fill = "orange") +
  ggtitle("ACTIVO", subtitle = "100 empresas eólicas") +
  ylab("Activo (cientos de millones de euros)")

dispersion_so / (caja_RENECO_so | caja_ACTIVO_so)
```



En el gráfico de dispersión se aprecia una nube de puntos más homogénea que antes de haber eliminado los casos u observaciones anteriores.

Vuelven a aparecer dos outliers en la variable ACTIVO –> *Pregunta* ¿por qué, si ya los habíamos eliminado?

ANÁLISIS DE MÁS DE DOS VARIABLES (todas cuantitativas). BÚSQUEDA DE *VALORES PERDIDOS* O *MISSING VALUES*:

Procedimiento básico: detección y eliminación de *missing values*:

Trabajamos con 4 variables: `RENECO`(Rentabilidad Económica) `ACTIVO`(Volumen de activos de la empresa) `MARGEN`(Margen de beneficio) `RES`(Resultado del ejercicio).

1. Creación de una copia del *data frame* al que llamaremos `muestra3`:

2. Que muestra los NA de cada una de las variables.

```
muestra3<- select(eolica_100, everything())
muestra3 %>% filter(is.na(RENECO) | is.na(ACTIVO) | is.na(MARGEN) |
is.na(RES)) %>% select(RENECO, ACTIVO, MARGEN, RES)

##                              RENECO    ACTIVO   MARGEN       RES
## Viesgo Renovables SL.            NA 269730.00   11.818  4609.000
## Biovent Energia SA            4.551 183899.00   22.792        NA
## Sargon Energias SLU              NA  85745.00 -615.625 -2216.000
## Parc Eolic Sant Antoni SL     1.361  69654.00       NA   668.000
## Eolica La Brujula SA          7.295  42146.98       NA  2306.062
## La Caldera Energia Burgos SL  2.643        NA   14.448   511.304

muestra3 <- muestra3 %>% filter(! is.na(RENECO) & ! is.na(ACTIVO) & !
is.na(MARGEN) & ! is.na(RES))
```

Con esta última línea, eliminamos los 6 valores perdidos (*NA*).

Detección de valores atípicos o outliers:

Al haber más de 2 variables, no puede utilizarse un gráfico de dispersión (más de 2 ejes). Si las variables que entran en el análisis son numerosas, podría ser poco operativo estudiar las variables una a una.

Distancia de Mahalanobis:

Proporciona un resumen del comportamiento de cada caso en todas las variables del análisis. Procedimiento:

1. Calculamos un vector con los *valores de la distancia de Mahalanobis* para las 4 variables en cada uno de los casos (empresas eólicas). Al vector lo llamaremos `muestra3.maha` que recoge las distancias.

2. Lo añadiremos al data frame `muestra3` mediante la función de pegado de columnas `cbind`, pasando a ser `muestra3.mahauna` columna (o variable) más.

```
muestra3.variables <- muestra3 %>% select(RENECO, ACTIVO, MARGEN, RES)
muestra3.maha <-mahalanobis(muestra3.variables,
                            center = colMeans(muestra3.variables),
                            cov = cov(muestra3.variables))
muestra3 <- cbind(muestra3, muestra3.maha)
```

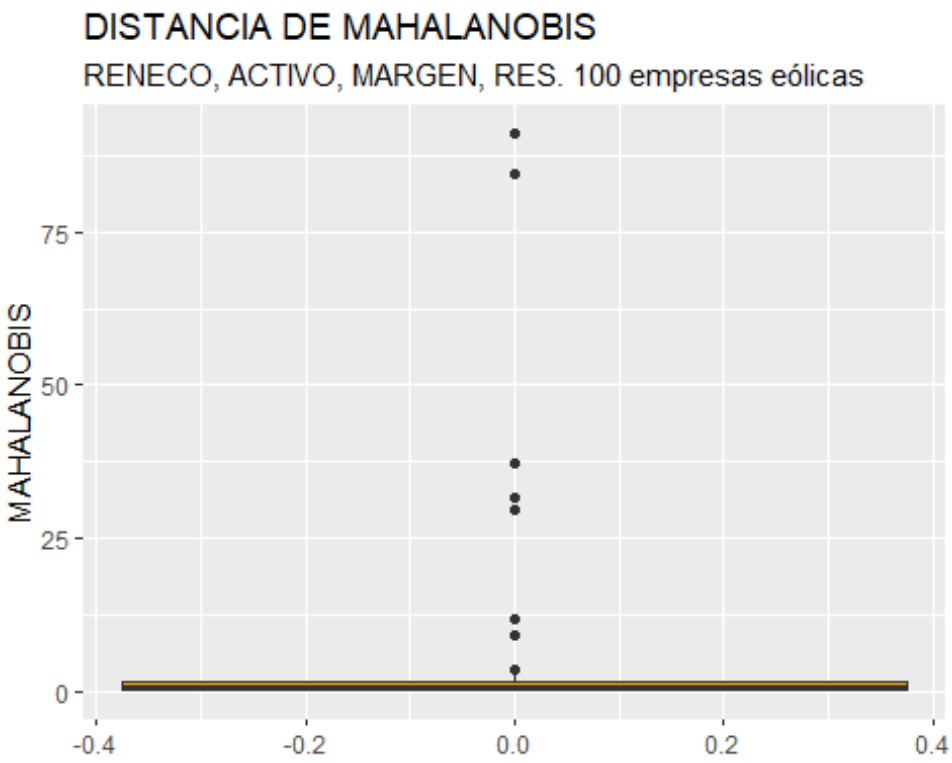
3. Creamos un diagrama de caja de la variable `muestra3.maha`. Para simplificar el nombre, vamos a cambiarlo por, `MAHALANOBIS`, usando la función `rename` del paquete `dplyr`.

```
muestra3 <- rename(muestra3, c(MAHALANOBIS = muestra3.maha))
```

4. Representamos el diagrama de caja:

```
ggplot(data = muestra3, map = (aes(y = MAHALANOBIS))) +
    geom_boxplot(fill = "orange") +
    ggtitle("DISTANCIA DE MAHALANOBIS", subtitle = "RENECO, ACTIVO,
```

```
MARGEN, RES. 100 empresas eólicas ") +
    ylab("MAHALANOBIS")
```



5. Obtenemos los casos concretos:

```
Q1M <- quantile (muestra3$MAHALANOBIS, c(0.25))
Q3M <- quantile (muestra3$MAHALANOBIS, c(0.75))

muestra3 %>% filter(MAHALANOBIS > Q3M + 1.5*IQR(MAHALANOBIS) |
MAHALANOBIS < Q1M - 1.5*IQR(MAHALANOBIS)) %>% select(MAHALANOBIS, RENECO,
ACTIVO, MARGEN, RES)

##                                MAHALANOBIS RENECO      ACTIVO
MARGEN
## Holding De Negocios De GAS SL.   91.041690  5.264 13492812.00
91.152
## Global Power Generation SA.      37.255573  1.393  2002458.00
22.403
## Naturgy Renovables SLU           31.675561  1.959  1956869.00
20.442
## Saeta Yield SA.                  11.891027  0.360   796886.38
16.258
## Molinos Del Ebro SA              29.589696 35.262    62114.37
41.821
## Tarraco Eolica SA                 3.600426 12.868    38102.00
400.899
## WPD Parque Eolico Navillas SL.   84.589929 -0.416    35511.45 -
2248.157
```

```
## Brulles Eolica SL                  3.599069 15.882    29722.58
47.227
## Sierra De Selva SL                 9.055155 21.761    27728.00
47.045
##                                          RES
## Holding De Negocios De GAS SL. 727548.0000
## Global Power Generation SA.     39995.0000
## Naturgy Renovables SLU          42737.0000
## Saeta Yield SA.                  2084.4760
## Molinos Del Ebro SA             17026.2569
## Tarraco Eolica SA                4953.0000
## WPD Parque Eolico Navillas SL.   -110.9293
## Brulles Eolica SL                3540.5693
## Sierra De Selva SL               4525.0000
```

6. Obtamos por eliminarlos: creamos un nuevo *data frame* con nombre `muestra3_so`

```
muestra3_so <- muestra3 %>% filter(MAHALANOBIS <= Q3M +
1.5*IQR(MAHALANOBIS) & MAHALANOBIS >= Q1M - 1.5*IQR(MAHALANOBIS))
```

La `muestra3_so`será una réplica de `muestra3`, sin incluir los casos detectados como atípicos (85 empresas).

ANÁLISIS CON MÁS DE DOS VARIABLES. CORRELACIÓN ENTRE VARIABLES:

Una característica muy importante es la *intensidad con que tales variables están relacionadas entre sí* o, el estudio de las *correlaciones*.

Visualización de la *matriz de correlaciones*:

Una forma es usando del paquete `PerformanceAnalytics`, la función `chart.Correlation()`:

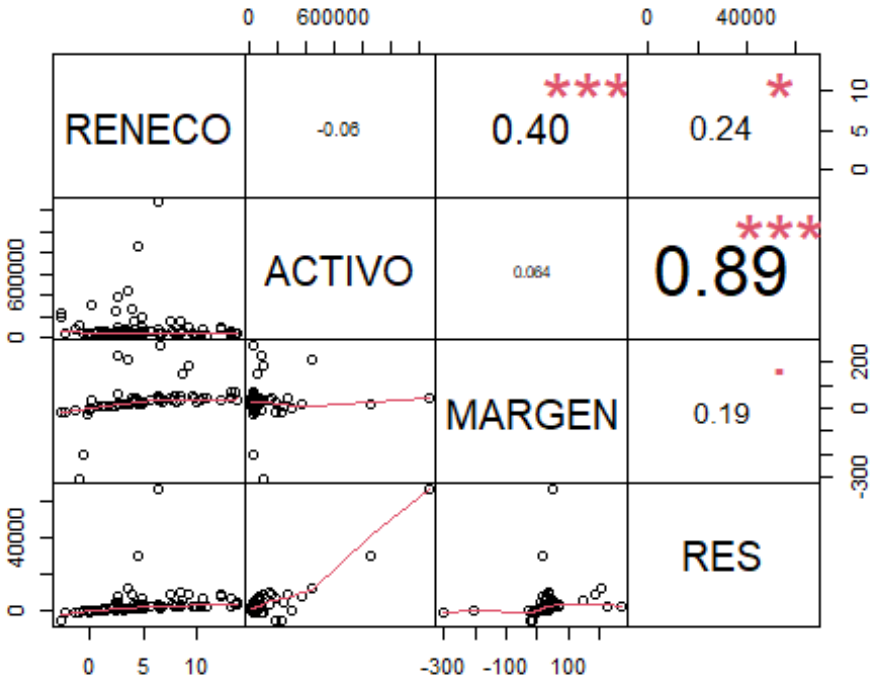
1. Creamos un nuevo *data frame* con nombre `muestra3_so_variables`:

```
muestra3_so_variables <- muestra3_so %>% select(RENECO, ACTIVO, MARGEN,
RES)
```

2. Cargamos el nuevo paquete y le pedimos que nos dibuje la matriz de correlaciones:

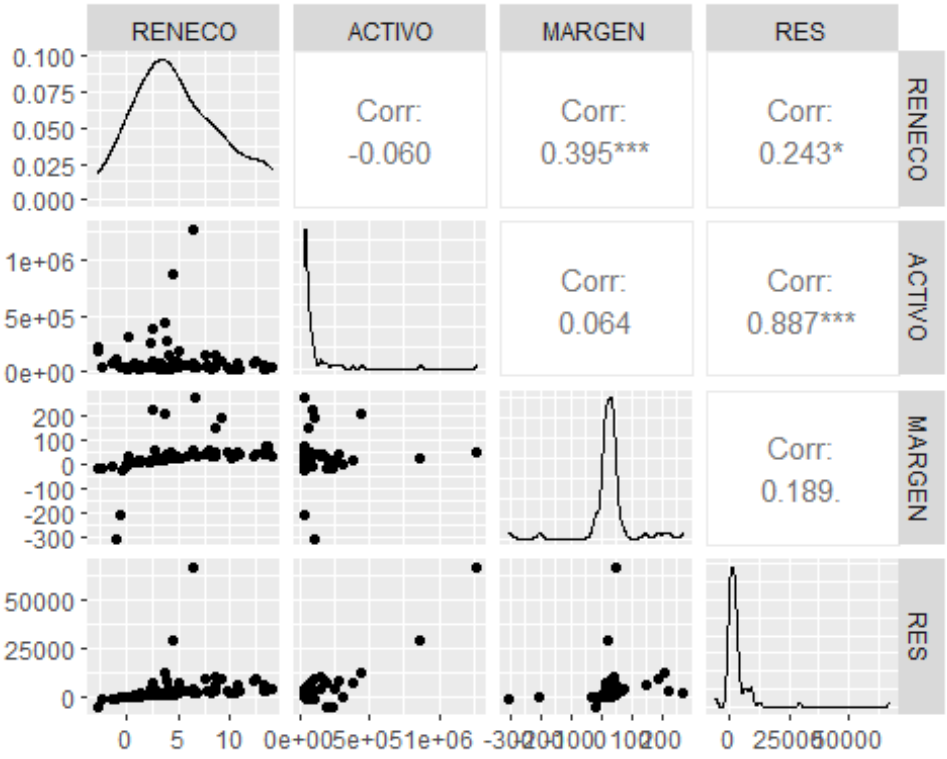
```
library(PerformanceAnalytics)

## 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(muestra3_so_variables, 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
```



En TAEDE vimos que podíamos crear la matriz de correlaciones con el siguiente código (menos visual)

```
library(GGally)

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

muestra3_so_variables %>%
  select(RENECO, ACTIVO, MARGEN, RES) %>%
  ggpairs()
```



Recordatorio: Un coeficiente de correlación puede tomar un valor entre -1 (fuerte relación, en sentido opuesto) a 1 (fuerte relación, en el mismo sentido).

Conclusiones:

- Las variables `ACTIVO` y `RES` mantienen una relación muy intensa y en sentido positivo.
- Entre `MARGEN` y `RENECO` existe también una relación de intensidad destacable.
- Entre `ACTIVO` y `MARGEN`; y `RENECO` y `ACTIVO` apenas están estadísticamente relacionadas.

HEMOS LLEGADO AL FIN DE LA PRÁCTICA. MUCHAS GRACIAS.