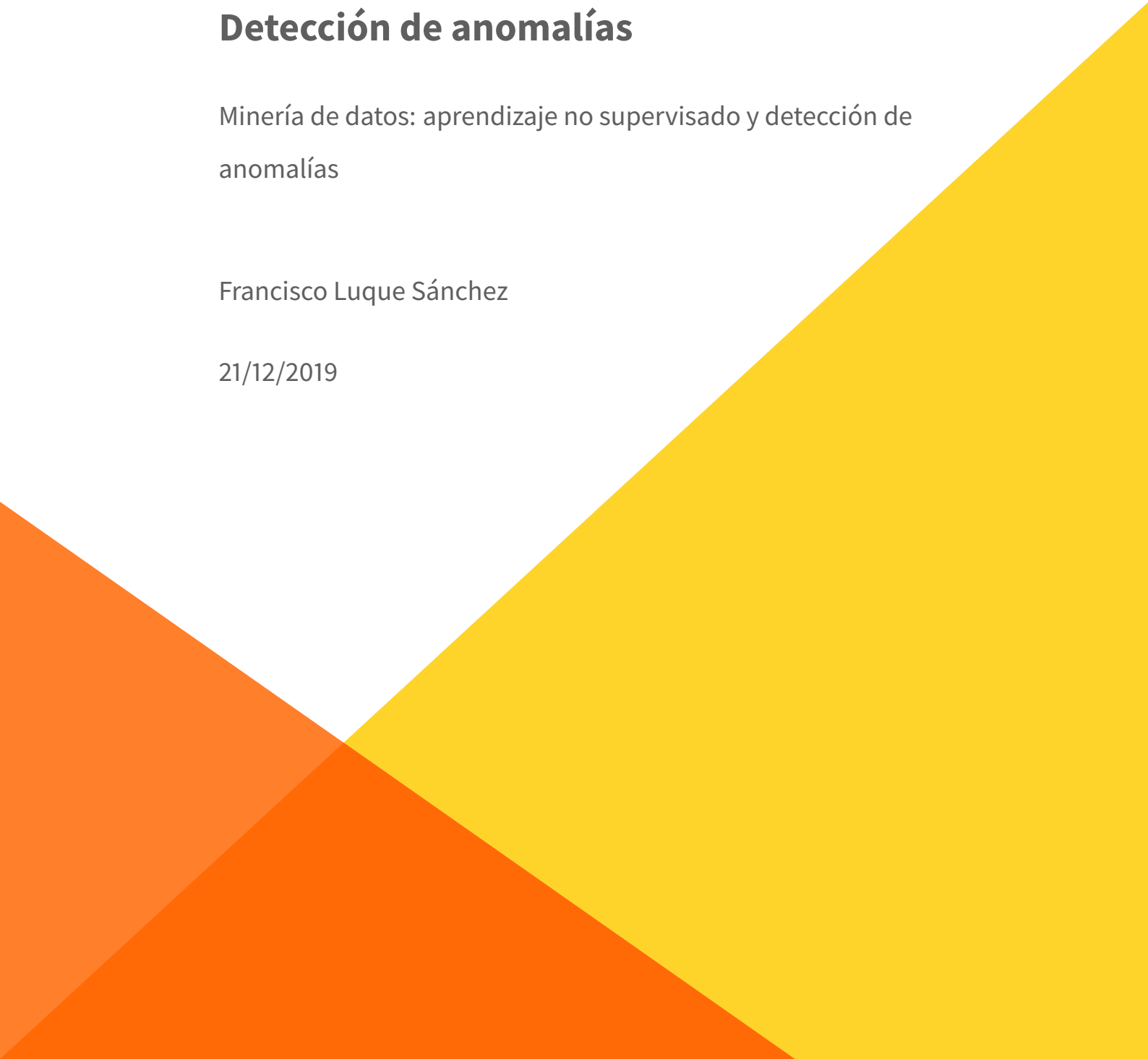

Detección de anomalías

Minería de datos: aprendizaje no supervisado y detección de anomalías

Francisco Luque Sánchez

21/12/2019



1 Introducción

En este trabajo se va a desarrollar una práctica sobre detección de ejemplos anómalos en un conjunto de datos. Se define como dato anómalo aquel elemento de un conjunto de datos cuyo comportamiento difiere del comportamiento esperable. Este comportamiento se traduce en que los valores registrados para alguna de las variables se sale de los valores esperados. En función de la naturaleza de la desviación, podremos tener distintos tipos de comportamientos anómalos:

- Ejemplos que presentan un valor extremo para alguno de los atributos medidos.
- Ejemplos que presentan una combinación de valores anormal para varias columnas estudiadas en conjunto, pero con valores comunes si son estudiados individualmente.

Estudiaremos distintas técnicas de detección de datos anómalos, tanto del primer como del segundo tipo.

1.1 Conjunto de datos de trabajo

El conjunto de datos con el que trabajaremos contiene información sobre enfermedades de la columna vertebral. Se nos proveen seis variables numéricas correspondientes a seis características biomecánicas, y se nos pide reconocer elementos anómalos dentro del conjunto. En total, se nos proporcionan 240 ejemplos, de los cuales 210 son ejemplos etiquetados como normales y 30 como anómalos. Resulta curioso, en la descripción del dataset, el hecho de que los pacientes etiquetados como anómalos son los pacientes sanos, mientras que los pacientes normales son aquellos que presentan la enfermedad. En el campo médico suele presentarse justo la casuística contraria, en la que la clase con pocos ejemplos suele ser la que contiene a los individuos enfermos. Esta diferencia puede deberse a que las enfermedades de espalda son relativamente comunes, o a que el conjunto de datos ha sido preparado específicamente.

En cualquier caso, en el enfoque que daremos a la solución del problema, no utilizaremos las etiquetas que se nos han provisto, ya que vamos a utilizar técnicas no supervisadas para la resolución del problema. Utilizaremos las etiquetas al final, para ver si los métodos que hemos empleado marcan como anómalos los ejemplos que realmente pertenecen a dicha clase, o si por el contrario los métodos que estamos utilizando no tienen un buen comportamiento para el problema que nos ocupa.

Comenzamos viendo los métodos de detección de outliers en una sola dimensión.

2 IQR

El primer método que estudiaremos está basado en el rango intercuartílico. Este método trabaja de forma univariante, tratando todas las variables del conjunto por separado. Por tanto, nos permitirá detectar valores extremos en cada una de las variables, pero no nos servirá para detectar combinaciones atípicas de valores. El funcionamiento del test se basa en el estudio de los cuartiles de una distribución normal. Cuando se trabaja con variables normalmente distribuidas, un enfoque típico para la detección de anomalías consiste en considerar como anómalos aquellos datos que se salen del intervalo $(\mu - k\sigma, \mu + k\sigma)$, donde μ y σ son la media y la desviación típica de la distribución, respectivamente. En función del valor de k que tomemos, cubriremos a un mayor número de puntos de la distribución normal. En concreto, suele tomarse el intervalo con $k = 2$, de forma que aproximadamente el 95 % de los valores de la distribución caen en el intervalo, o $k = 3$, intervalo en el que se encuentran el 99.7 % de los valores. Los valores de la distribución que quedan fuera de estos límites se consideran datos atípicos

No obstante, la asunción de normalidad en los datos es una suposición fuerte, que usualmente no se cumple. Además la media y la desviación típica de los datos son dos estadísticos muy sensibles a la existencia de valores anómalos. De esta forma, en lugar de utilizar el método descrito anteriormente, resulta más útil comparar el comportamiento del rango intercuartílico con el de la desviación, y utilizar el primer estadístico, que es más robusto ante la existencia de valores atípicos. Para la distribución normal, el primer cuartil está en el valor -0.67 , y el tercer cuartil en el valor 0.67 . El rango intercuartílico es, por tanto 1.34 . Tomando entonces el valor $k' = 1.5$, tenemos que el intervalo $(-1.5 * IQR, 1.5 * IQR)$ contiene aproximadamente los mismos valores que el intervalo $(-2\sigma, 2\sigma)$.

Utilizando esta justificación, el método IQR etiqueta como datos anómalos normales para una variable aquellos que se desvían de la media más de 1.5 por el valor del rango intercuartílico. Por un razonamiento similar, se marcan como datos anómalos extremos aquellos que se desvían de la media más de $3 * IQR$.

2.1 Aplicación del algoritmo

Pasamos a aplicar este método para detectar outliers en nuestro conjunto de datos. Mostraremos en primer lugar el proceso completo para una variable, y lo aplicaremos después para todas las demás:

```
## Leemos el dataset y seleccionamos una columna
# dataset <- read.csv("dataset/ecoli.data")
dataset <- as.data.frame(read.mat("dataset/vertebral.mat")$X)
columna.scaled <- scale(dataset$V5)
```

```
## Calculamos los cuartiles y el IQR
cuartil.primerio <- quantile(columna.scaled, .25)
cuartil.tercero <- quantile(columna.scaled, .75)
iqr              <- cuartil.tercero - cuartil.primerio

## Calculamos los valores a partir de los cuales se consideran los outliers
extremo.superior.outlier.normal <- cuartil.tercero + 1.5*iqr
extremo.inferior.outlier.normal <- cuartil.primerio - 1.5*iqr
extremo.superior.outlier.extremo <- cuartil.tercero + 3*iqr
extremo.inferior.outlier.extremo <- cuartil.primerio - 3*iqr

## Construimos los vectores que nos determinan los outliers
vector.es.outlier.normal <- (
  (columna.scaled > extremo.superior.outlier.normal &
   columna.scaled < extremo.superior.outlier.extremo) |
  (columna.scaled < extremo.inferior.outlier.normal &
   columna.scaled > extremo.inferior.outlier.extremo)
)

vector.es.outlier.extremo <- (
  columna.scaled < extremo.inferior.outlier.extremo |
  columna.scaled > extremo.superior.outlier.extremo
)

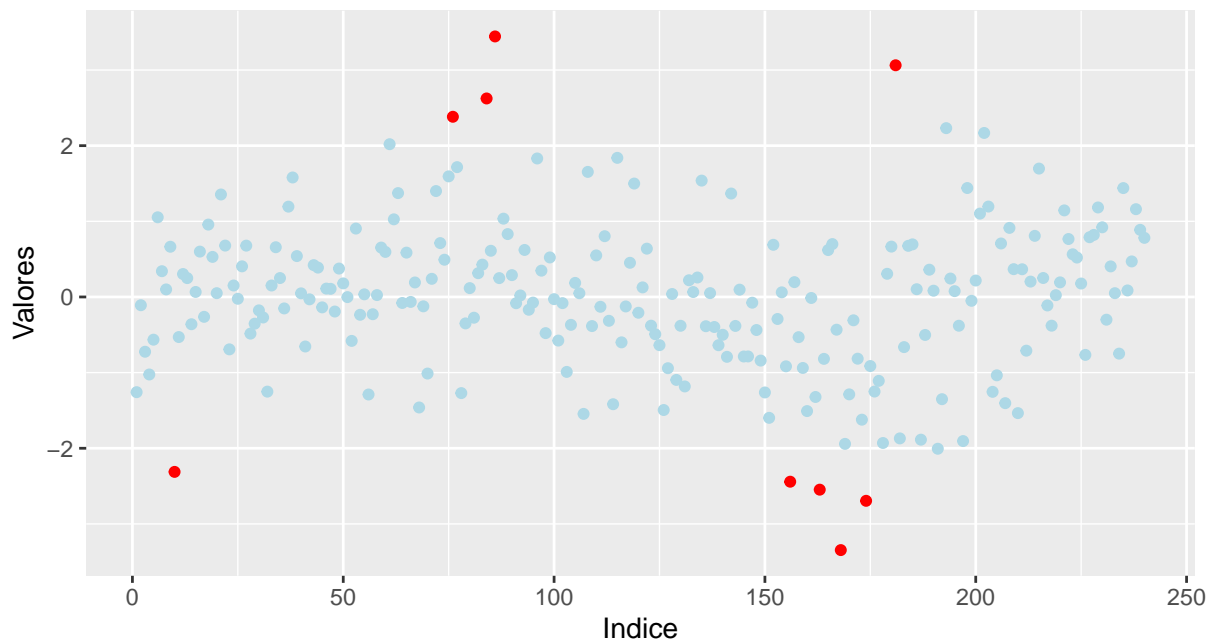
## Calculamos los valores normales y extremos
claves.outliers.normales <- which(vector.es.outlier.normal)
data.frame.outliers.normales <- dataset[claves.outliers.normales,]
valores.outliers.normales <- data.frame.outliers.normales["V5"]

claves.outliers.extremos <- which(vector.es.outlier.extremo)
data.frame.outliers.extremos <- dataset[claves.outliers.extremos,]
valores.outliers.extremos <- data.frame.outliers.extremos["V5"]

valores.normalizados.outliers.normales <-
  ↪ columna.scaled[claves.outliers.normales]

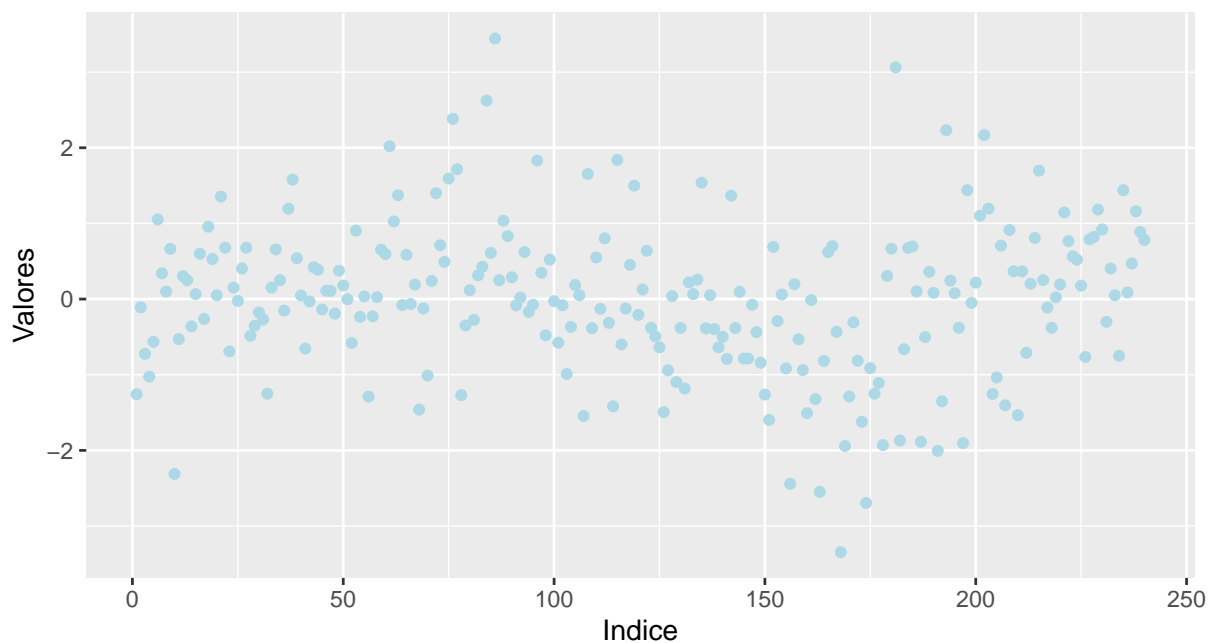
par(mfrow=c(1,2))
## Mostramos gráficamente los outliers
MiPlot_Univariate_Outliers(columna.scaled, claves.outliers.normales,
                             "Outliers normales")
```

Outliers normales



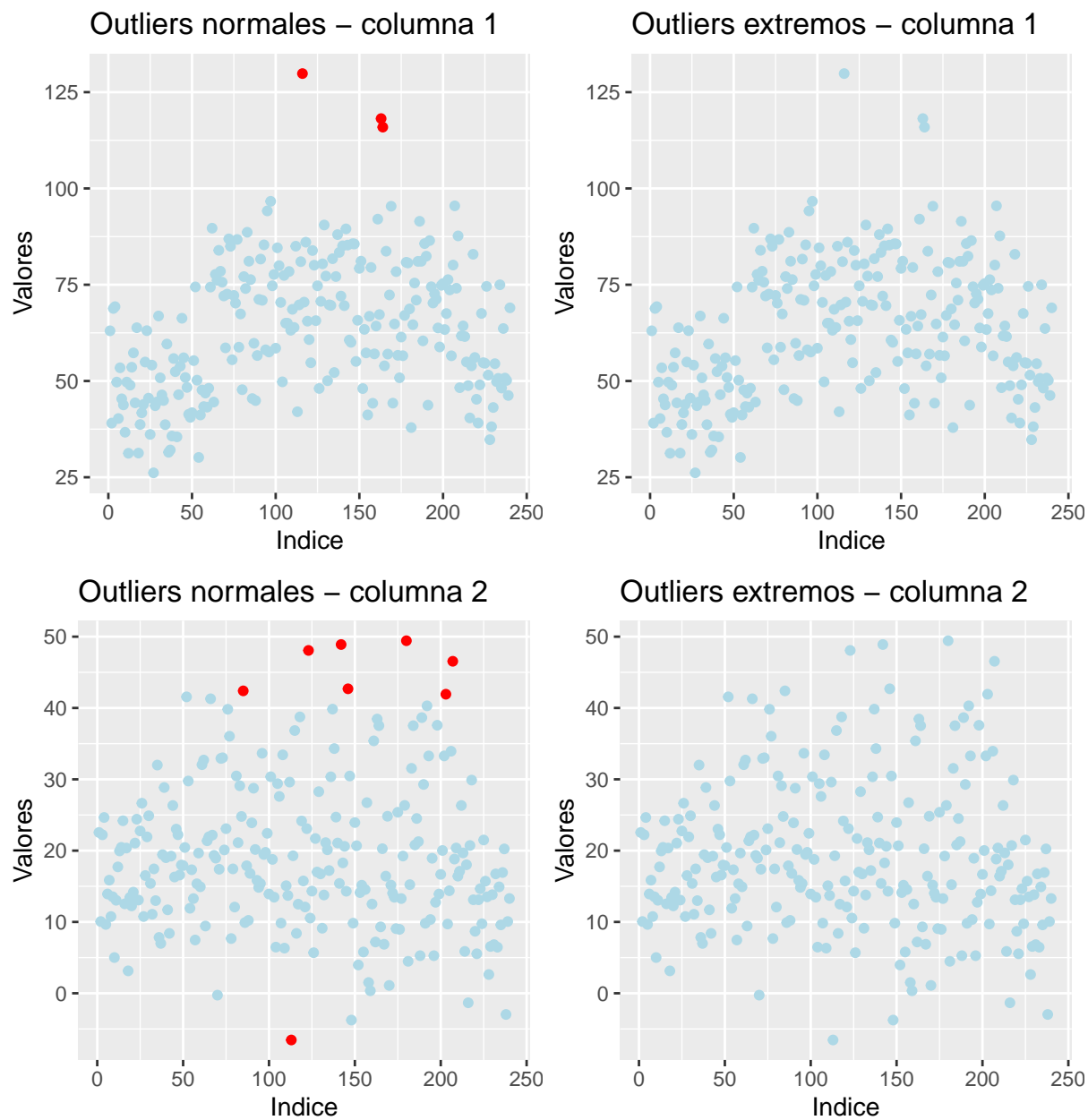
```
MiPlot_Univariate_Outliers(columna.scaled, claves.outliers.extremos,  
                             "Outliers extremos")
```

Outliers extremos

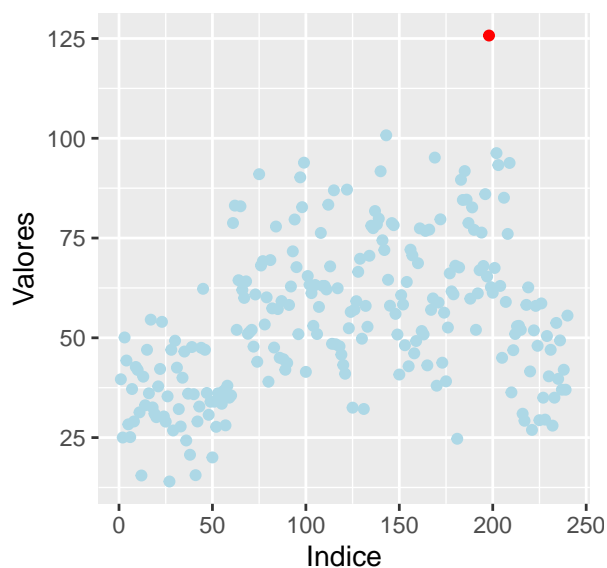


Una vez hemos visto cómo se aplica el método a una determinada columna, nos interesará aplicarlo a

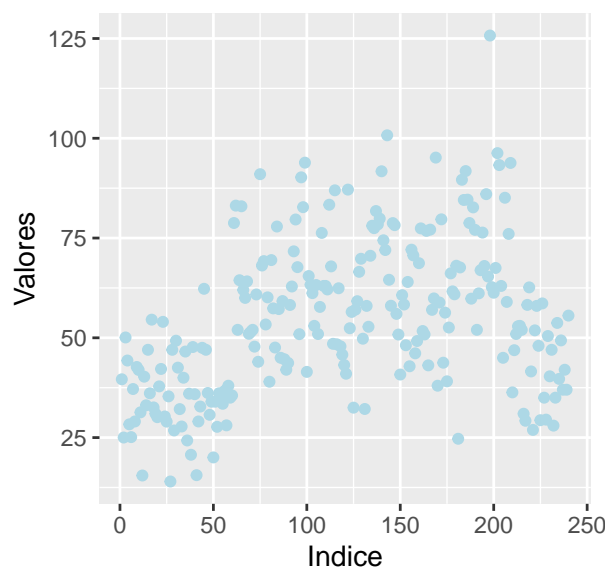
todas las variables de nuestro conjunto (ocultamos el código porque no aporta nueva información).
Mostramos gráficamente los resultados:



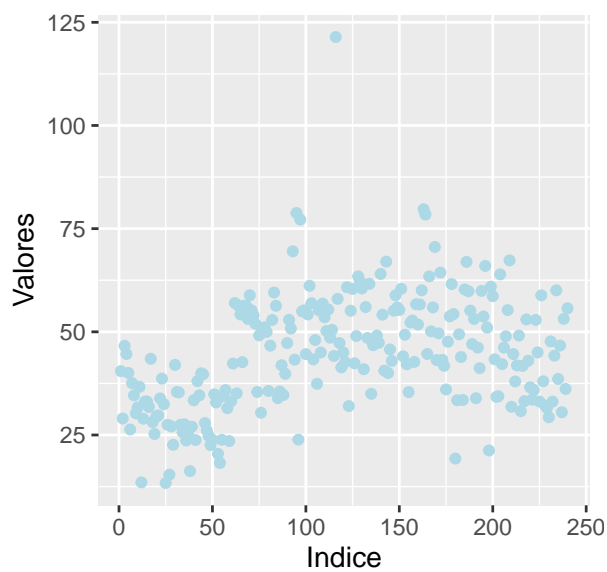
Outliers normales – columna 3



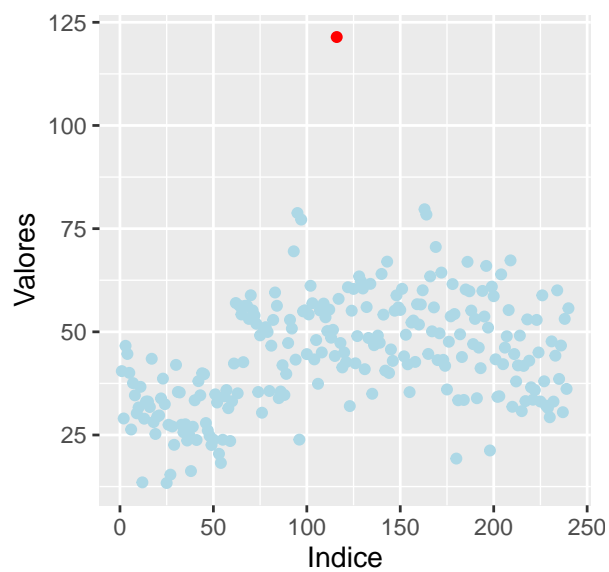
Outliers extremos – columna 3

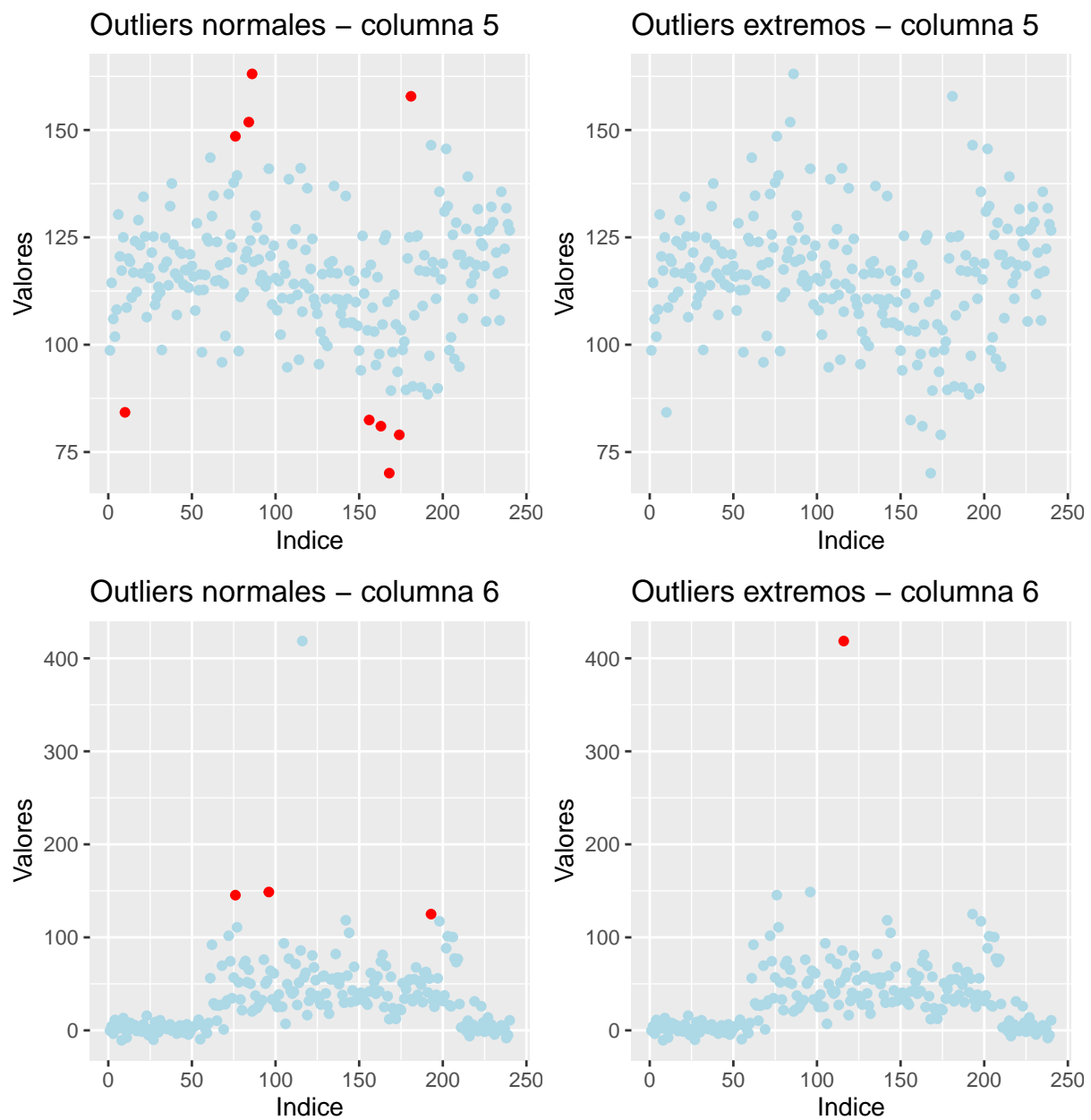


Outliers normales – columna 4



Outliers extremos – columna 4

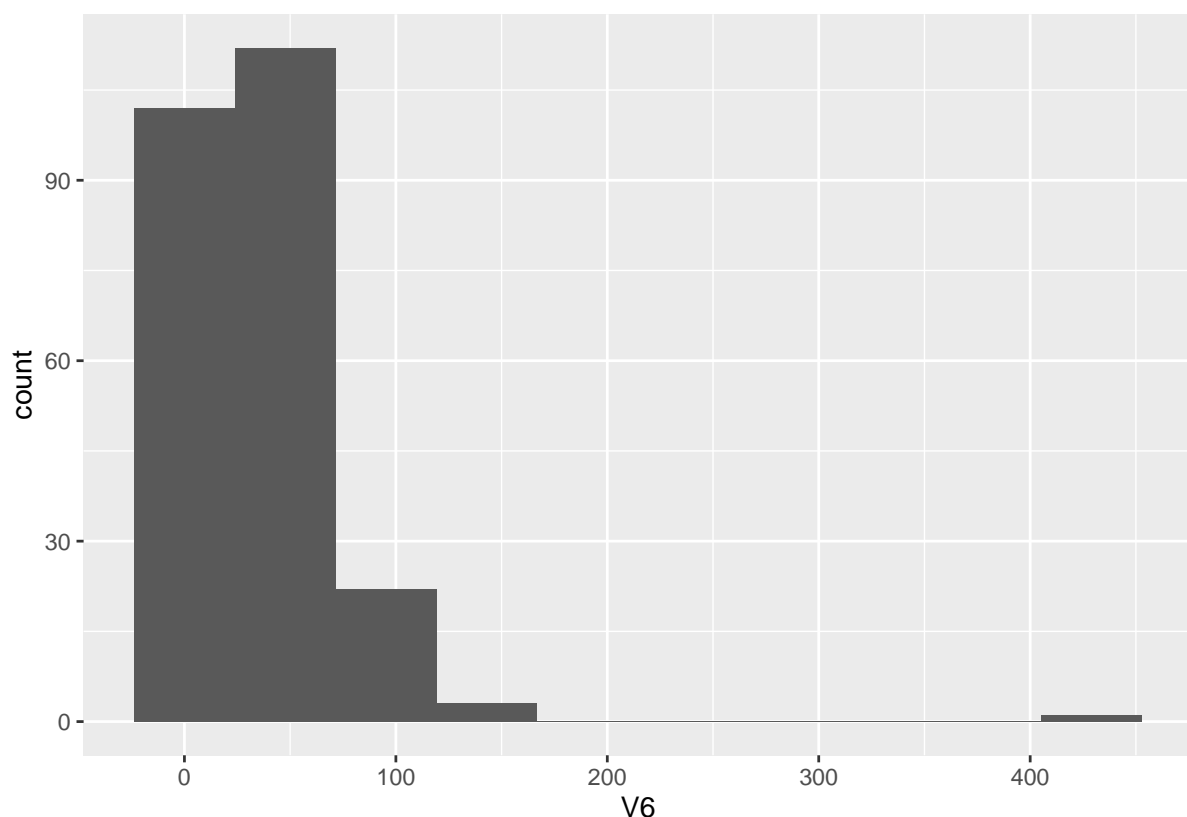


**Table 1:** Número de outliers en cada columna del conjunto de datos

Columna 1	Columna 2	Columna 3	Columna 4	Columna 5	Columna 6
3	8	1	1	9	4

Se pueden extraer varias conclusiones de las gráficas y la tabla anteriores. En primer lugar, tenemos

que ninguna de las variables posee un número muy grande de elementos anómalos, siendo la columna 5 la que más outliers tiene. A pesar de esto, solamente 9 de las 240 medidas son anómalas, lo cual corresponde a menos de un 4 % de los valores, lo cual es relativamente poco. Otro fenómeno a remarcar es la cantidad tan pequeña de outliers extremos que tienen los datos. Cuatro de las columnas no presentan ningún outlier extremo, y las otras dos solamente presentan uno. A pesar de esto, es curioso cómo la columna 4 sólo presenta un outlier, pero el mismo es extremo. Esto remarca la gran distancia que hay entre este valor y el resto de valores de la variable. Otro punto a remarcar es el sesgo a la derecha que presentan las medidas. Para todas las columnas (a excepción de la quinta) la mayoría de los outliers se agrupan en la parte superior del gráfico, en los valores altos. Para la columna 6, en particular, se puede observar el fuerte sesgo que presenta en el histograma siguiente:



Donde tenemos el grueso de los datos agrupado en la parte izquierda del gráfico (cerca de 0) y un pequeño grupo de valores dispersos en la parte derecha. En la descripción del conjunto de datos, se nos dice que esta variable corresponde con el grado de espondilolistesis. Esta patología consiste en el desplazamiento de una vértebra de su posición original al deslizarse sobre el disco intervertebral que la sujeta. Parece lógico pensar que esta variable va a tener un sesgo hacia la derecha, ya que al ser positiva y tener su valor normal (sano) en el cero, lo más probable es que los valores se agrupen en torno a un valor bajo, y se den pocos casos en los que el valor crezca significativamente.

Una vez hemos visto cómo podemos detectar valores anómalos utilizando el método del rango intercuartílico, vamos a ver cómo podemos dotar de más robustez a la detección de anomalías univariantes.

3 Tests estadísticos para la detección de anomalías univariantes

En este apartado veremos cómo podemos aplicar tests estadísticos para la detección de datos anómalos univariantes. En primer lugar, aplicaremos el test de Grubbs, el cual nos permite saber si existe un único outlier en una determinada columna del conjunto de datos, y posteriormente utilizaremos el test de Rosner, el cual nos permitirá deducir si hay hasta k outliers, para k conocido de antemano. Comenzamos comentando el test de Grubbs.

3.1 Test de Grubbs

Este test estadístico permite encontrar outliers en una distribución univariante, asumiendo que existe normalidad en la misma. Existen dos versiones del test, dependiendo si hacemos el test con una o dos colas. La versión del test de una cola nos permite saber si el máximo o el mínimo de los valores es anómalo, pero no trabaja sobre ambos valores simultáneamente. Nosotros utilizaremos el test de dos colas, que trabaja sobre toda la muestra. Para ambas versiones del test, las hipótesis que plantea el mismo son las siguientes.

- H_0 : No hay anomalías en el conjunto de datos
- H_1 : Hay exactamente una anomalía en el conjunto de datos

El estadístico del test de Grubbs se define como:

$$G = \frac{\max_{i=1,\dots,N} |Y_i - \bar{Y}|}{S}$$

Donde \bar{Y} y S representan la media y la desviación típica muestrales, simultáneamente. Para el test de dos colas, la hipótesis nula se rechaza para un nivel de significación α si el estadístico G cumple:

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t_{\alpha/(2N), N-2}^2}{N-2 + t_{\alpha/(2N), N-2}^2}}$$

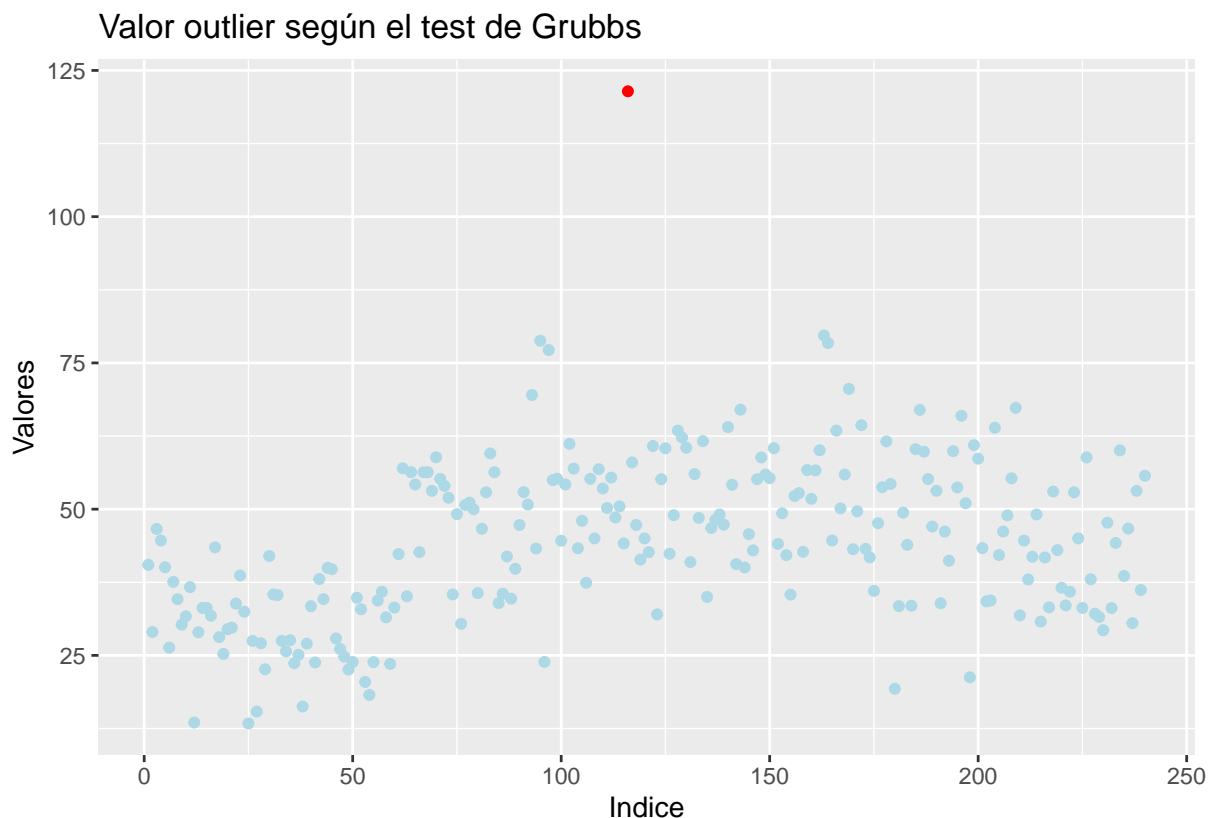
Donde $t_{\alpha/(2N), N-2}$ es el valor crítico para el nivel de significación $\alpha/2N$ de una distribución T de Student con $N-2$ grados de libertad. Vamos a ver cómo podemos aplicar este test a una columna de nuestro conjunto de datos.

```
selected.col <- dataset$V4

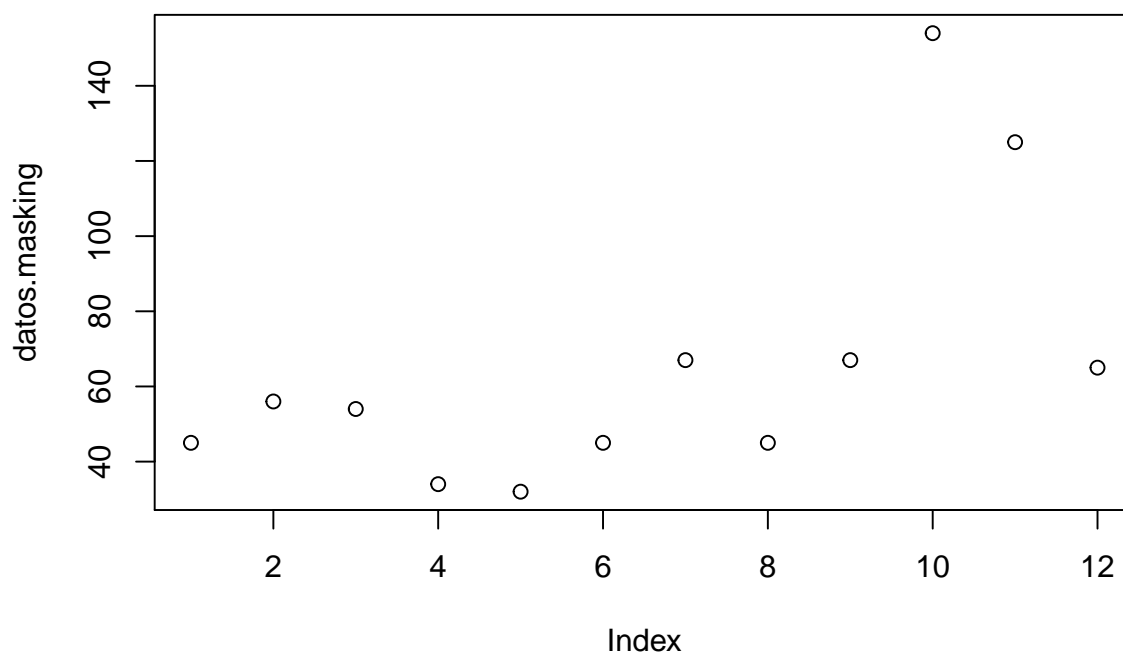
## Dado que el p-valor es inferior a 0.05, tenemos que el valor de mayor
## dispersión respecto de la media es un valor anómalo
grubbs.test(selected.col, two.sided = T)
```

```
##
## Grubbs test for one outlier
##
## data: selected.col
## G = 5.50547, U = 0.87265, p-value = 3.084e-06
## alternative hypothesis: highest value 121.43 is an outlier
```

```
indice.de.outlier.Grubbs <- order(abs(selected.col - mean(selected.col)),
                                decreasing = T)[1]
valor.de.outlier.Grubbs <- selected.col[indice.de.outlier.Grubbs]
MiPlot_Univariate_Outliers(selected.col, indice.de.outlier.Grubbs,
                             "Valor outlier según el test de Grubbs")
```



Podemos observar que, para esta variable, se detecta que el valor de máxima varianza es un valor anómalo. Este test nos permite averiguar, precisamente, si el valor de máxima varianza se desvía anormalmente del resto del conjunto de datos. No obstante, ofrece varios problemas. En primer lugar, sólo nos permite averiguar si hay un valor anómalo, pero no nos aporta información del resto de puntos. Por ejemplo, en el caso anterior, el punto que se encuentra cercano al marcado como outlier, y que presenta el mismo valor de y , es claramente otro punto anómalo. Además de este, probablemente el punto aislado que se muestra a la derecha también tenga un valor anormalmente alto. Para estos dos puntos, el test de Grubbs no aporta información. Una posible forma de solventar esta problemática consiste en ejecutar iterativamente este algoritmo mientras se sigan encontrando anomalías, quitando cada vez el punto marcado como anomalía. No obstante, el hecho de trabajar con un único punto puede producir ciertos problemas. Por ejemplo existe una problemática conocida como masking, en la cual la presencia de varios valores anómalos hacen que este test no permita rechazar la hipótesis nula. Ponemos a continuación un ejemplo de masking:



```
##  
## Grubbs test for one outlier  
##  
## data:  datos.masking  
## G = 2.39188, U = 0.43262, p-value = 0.05614  
## alternative hypothesis: highest value 154 is an outlier
```

Obtenemos un p-valor superior a 0.05, por lo que con los niveles de significación estándar, no podemos rechazar la hipótesis nula y por tanto no tendríamos ningún outlier. No obstante, podemos observar que existen dos valores que se desvían significativamente de los valores normales, pero su presencia hace que la desviación típica del conjunto de datos aumente, y por tanto este método no detecte su presencia. En el siguiente apartado veremos otro test estadístico, que nos permitirá afrontar esta problemática.

3.2 Test de Rosner

Para solucionar el problema anterior, suelen utilizarse otros tests estadísticos, que permiten averiguar si hay un número k (fijo de antemano) de outliers en la muestra. No obstante, estos métodos carecen de mucho interés, ya que existen tests más informativos, que nos permiten discernir si existen un número de outliers menor o igual al número k fijado de antemano, en lugar de necesitar el número exacto de outliers. Entre estos tests se encuentra el test de Rosner. Este test trabaja de la siguiente manera. Se calcula la serie de estadísticos siguiente:

$$R_{i+1} = \frac{|x^{(i)} - \bar{x}^{(i)}|}{s^{(i)}}$$

Donde $\bar{x}^{(i)}$ y $s^{(i)}$ son la media y la desviación típica del conjunto tras eliminar los i valores más extremos, y $x^{(i)}$ es la observación de dicho subconjunto que más se desvía de la media. Una vez los valores R_1, \dots, R_k se han calculado, se realizan una serie de tests de hipótesis consistentes en comparar los valores R_i con una serie de valores λ_i , los cuales se calculan como

$$\lambda_{i+1} = \frac{t_{p,n-i-2}(n-i-1)}{\sqrt{(n-i-2 + t_{p,n-i-2})(n-i)}}$$

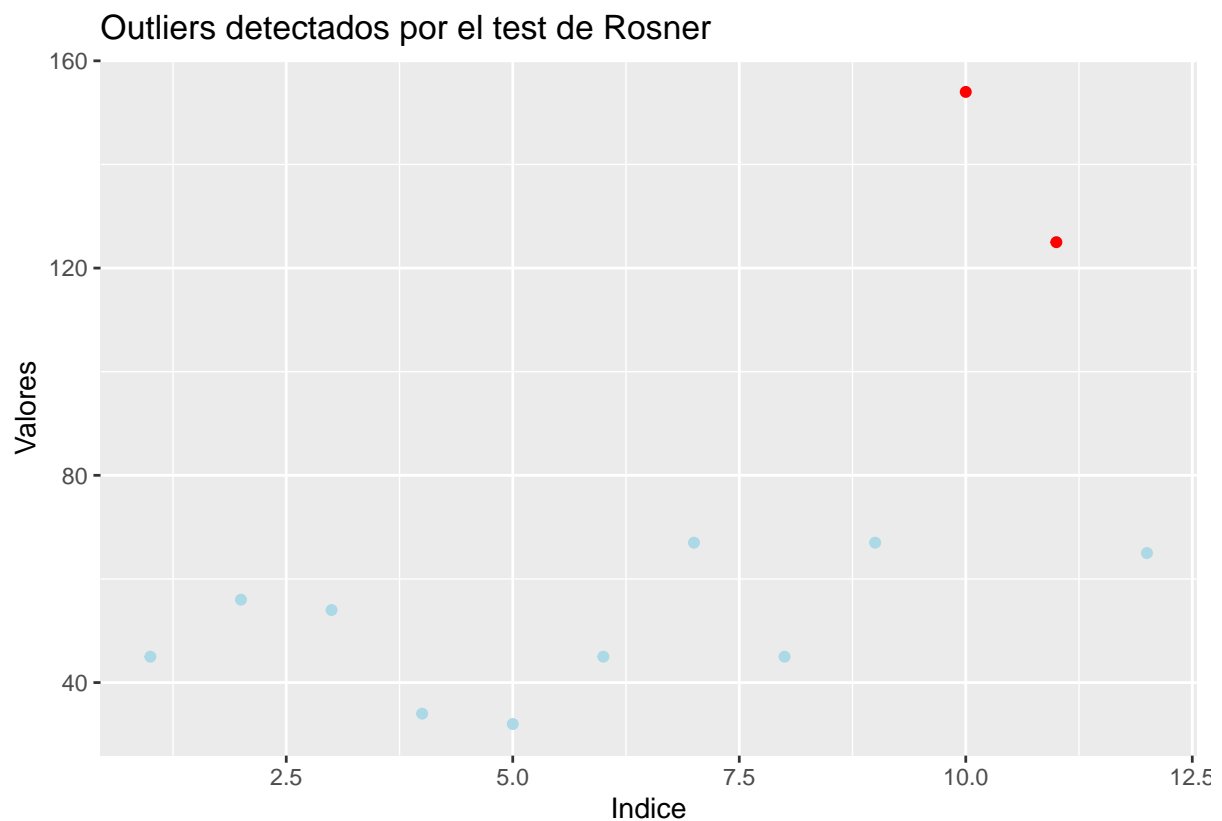
Donde de nuevo, $t_{p,n}$ corresponde al percentil p extraído de una distribución t de Student con n grados de libertad.

Para ver cómo funciona el algoritmo sobre nuestros datos, trataremos de encontrar los outliers por este método sobre los datos sintéticos que hemos generado previamente:

```
## Warning in rosnerTest(datos.masking, 3): The true Type I error may be larger
## than assumed. See the help file for 'rosnerTest' for a table with information on
## the estimated Type I error level.

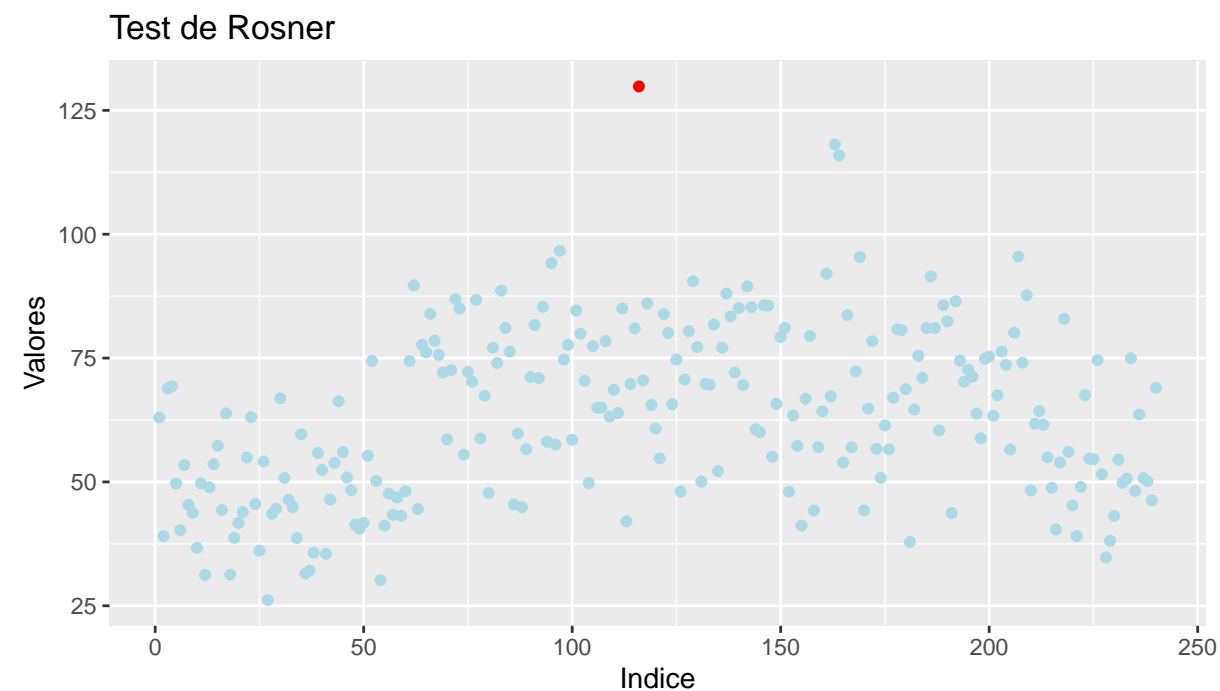
## [1] TRUE TRUE FALSE

## [1] 10 11 5
```

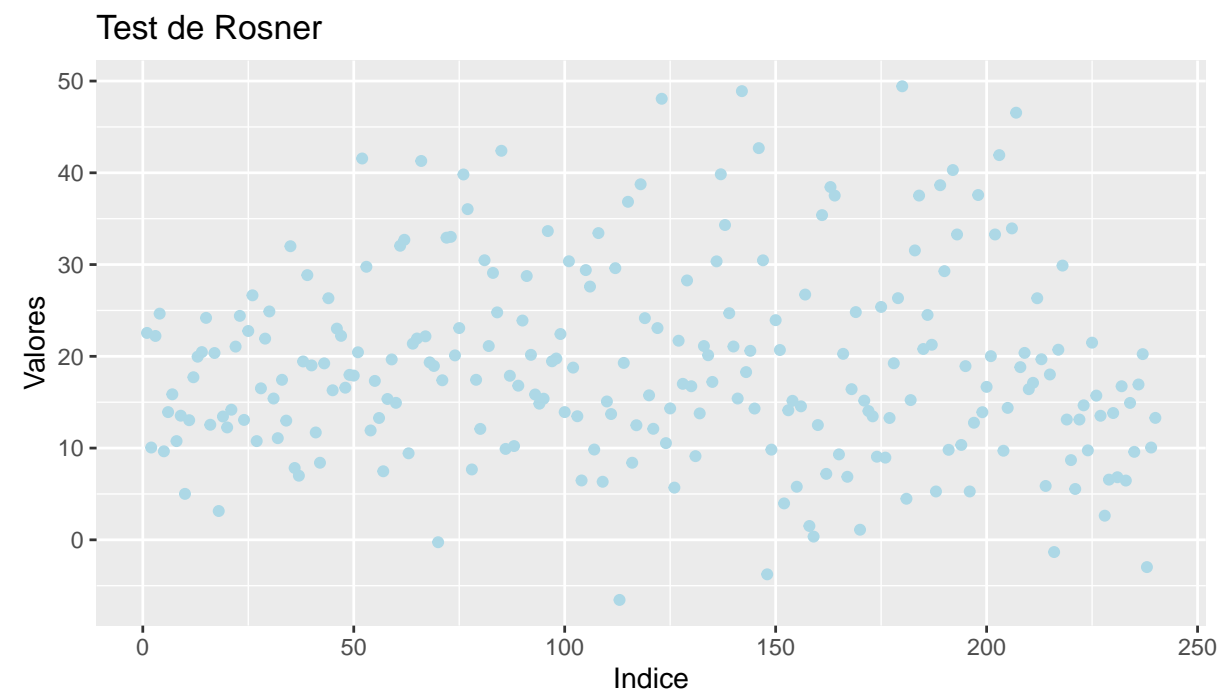


En nuestro caso, hemos ejecutado el test para buscar un máximo de 3 outliers, pero sólo han dado positivo dos de ellos (los cuales habíamos observado a simple vista a priori). La ventaja que aporta este test frente al de Grubbs es la capacidad de detectar más de un elemento anómalo al mismo tiempo. Aplicamos a continuación este test a nuestro conjunto de datos, para tratar de detectar valores anómalos sobre nuestras variables.

```
## $V1
```

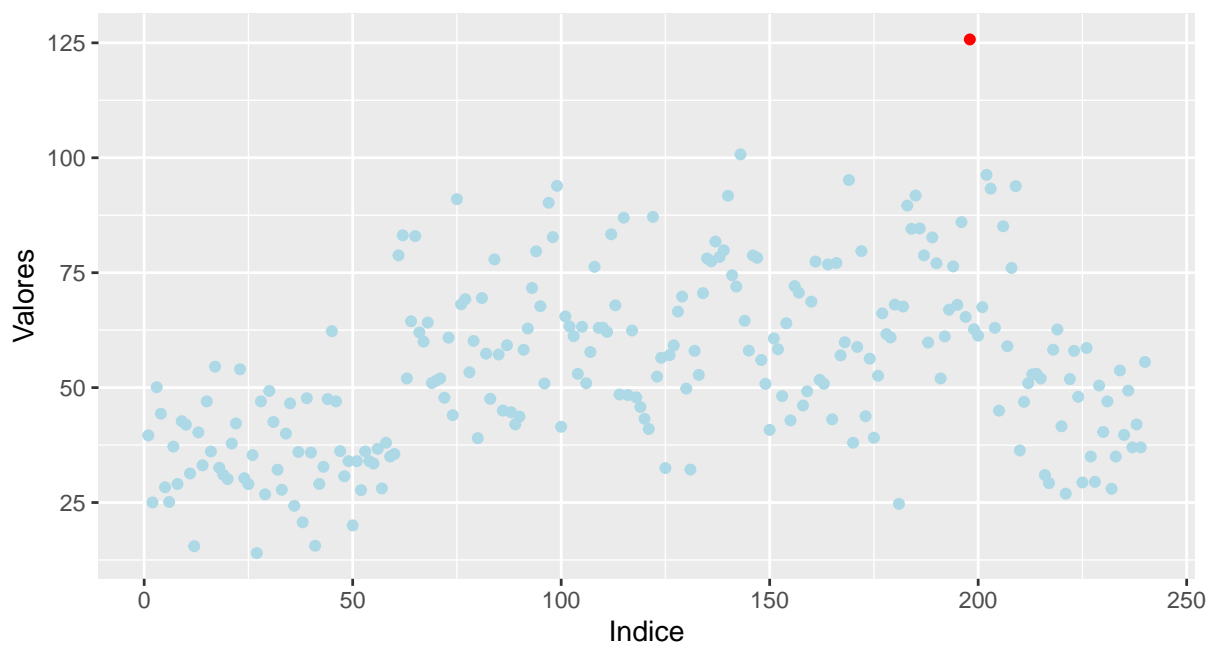


```
##
## $V2
```



```
##
## $V3
```

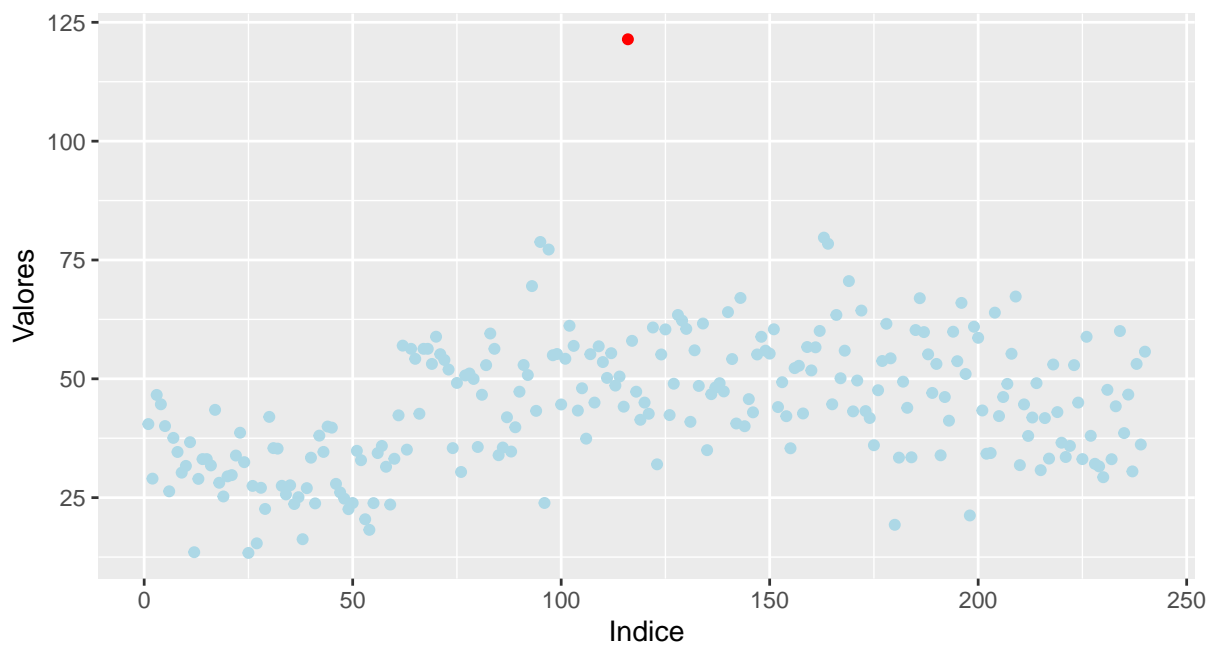
Test de Rosner



##

\$V4

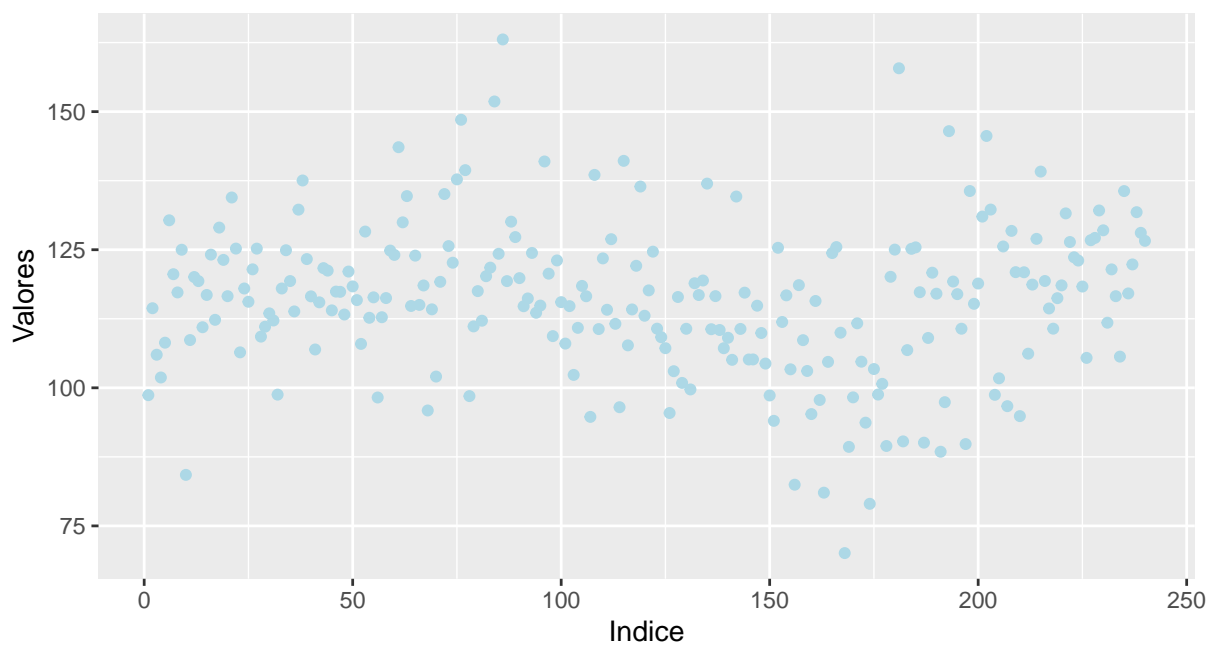
Test de Rosner



##

\$V5

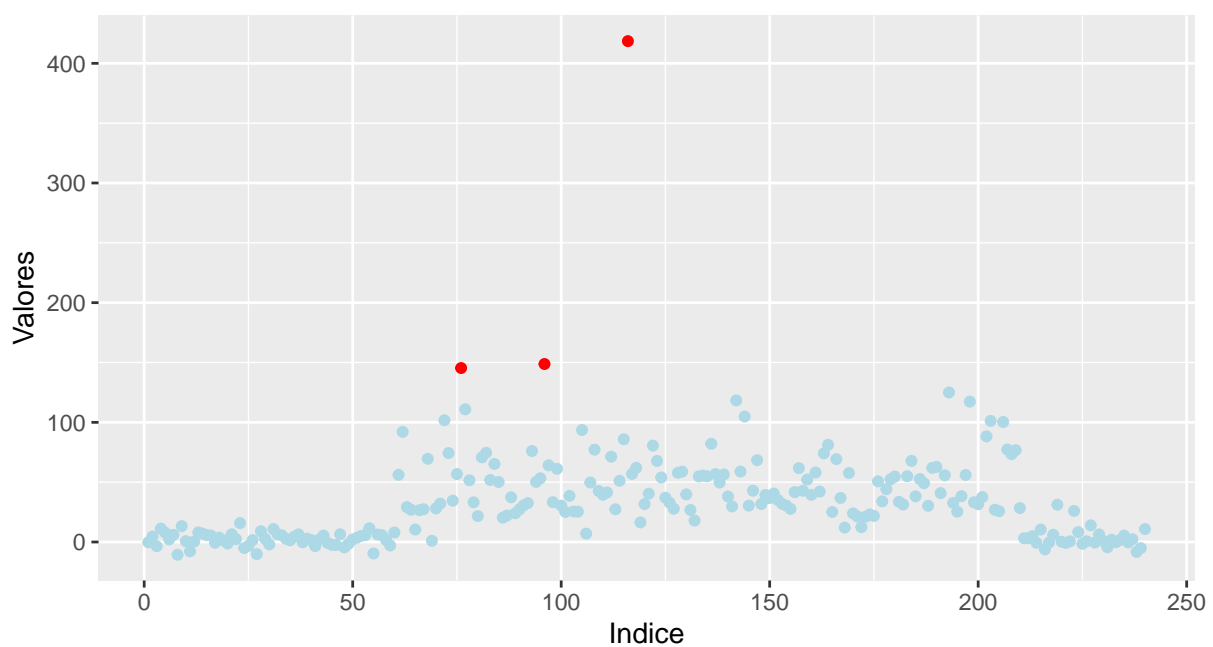
Test de Rosner



##

\$V6

Test de Rosner



Podemos observar que este test nos ha permitido encontrar varios outliers en las distintas variables. No obstante, al ser un test estadístico y utilizar un nivel de significación $\alpha = 0.05$, algunos de los

valores anómalos que detectamos con el método IQR ahora no son marcados como tales. En particular, para las variables 2 y 5 no llega a detectarse ningún dato anómalo. Por otra parte, este test nos permite detectar de forma simultánea varios outliers, como ocurre con la variable 6, en la que se detectan 3 outliers (la anomalía extrema y dos de las anomalías normales).

Una vez hemos visto cómo encontrar anomalías univariantes, veremos cómo podemos encontrar outliers cuya problemática viene por una combinación extraña de los valores de las variables, y no sólo por un valor extremo en alguna de ellas.

4 Test de hipótesis para detección de anomalías multivariantes

Una vez que hemos visto tests de hipótesis que nos permiten detectar outliers en variables unidimensionales, lo que nos será realmente interesante es encontrar valores anómalos en distribuciones de mayor dimensionalidad. En este caso no buscaremos valores anormales dentro de una única variable, si no que la idea será encontrar combinaciones atípicas de valores en varias variables. De esta forma, aunque los valores que toma cada variable por separado puedan ser valores normales, la anomalía proviene de la combinación de los mismos.

Vamos a aplicar este test sobre nuestro conjunto de datos. Para tratar de dar una información lo más visual posible, trabajaremos en primer lugar sobre dos columnas del dataset, en lugar del conjunto de datos completo. De esta forma, podremos observar gráficamente si somos capaces de detectar outliers bivariantes que no son considerados como tales cuando trabajamos con las variables de forma independiente. Este test tiene dos posibles formas de aplicación, para detectar un único outlier, o para detectar todos los posibles outliers del conjunto de datos. En el primero de los casos, las hipótesis que se imponen son

- H_0 : en el conjunto de datos no hay ningún outlier.
- H_1 : en el conjunto de datos hay exactamente un outlier.

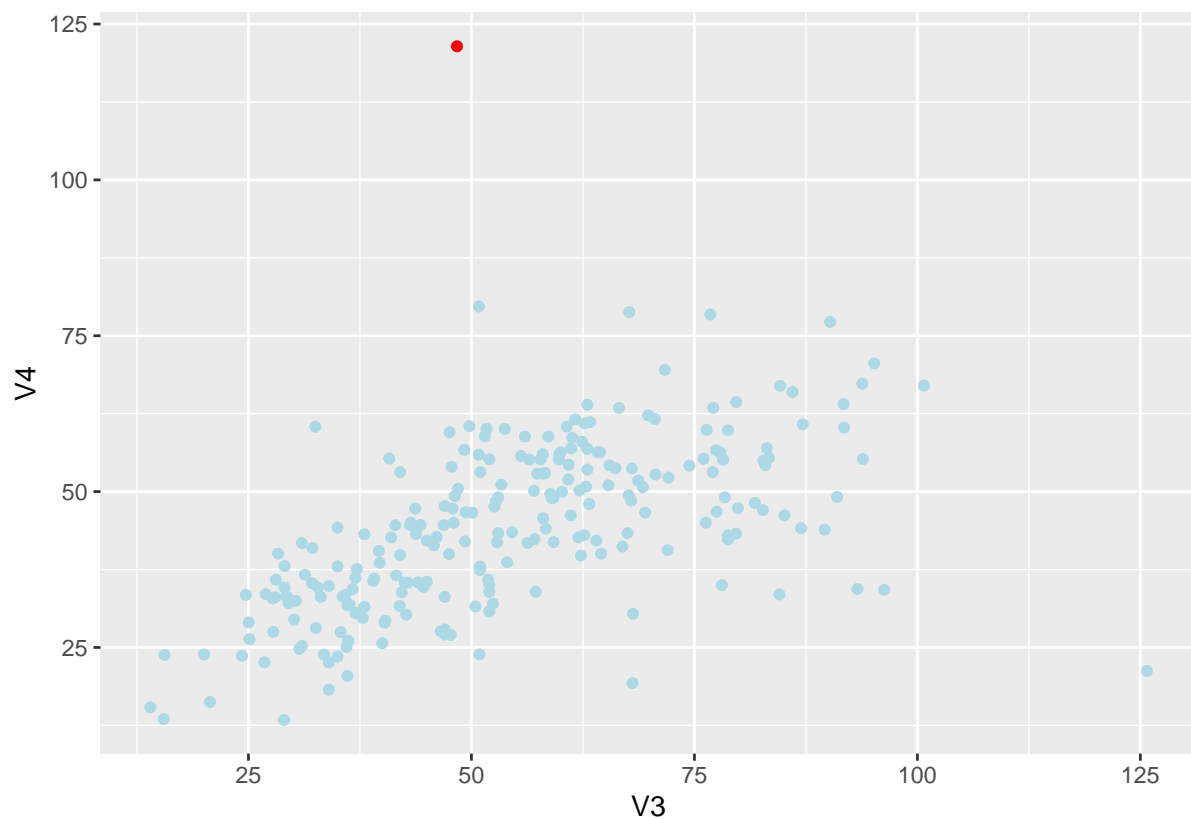
Usando este test, sólo podremos sacar conclusiones sobre el valor más desviado de nuestra distribución. En el segundo de los casos, se relaja la hipótesis alternativa, lo cual nos obliga a aplicar una corrección sobre el nivel de significación, para evitar incurrir en un error FWER. La nueva hipótesis alternativa es la siguiente:

- H_1 : en el conjunto de datos hay al menos un outlier

Aplicamos ambos métodos sobre las columnas 3 y 4 de nuestro dataset. Comenzamos por el primero de los métodos:

Table 2: Valores normalizados para el outlier encontrado con el test de Cerioli de tipo a

V3	V4
-0.3185214	5.505469

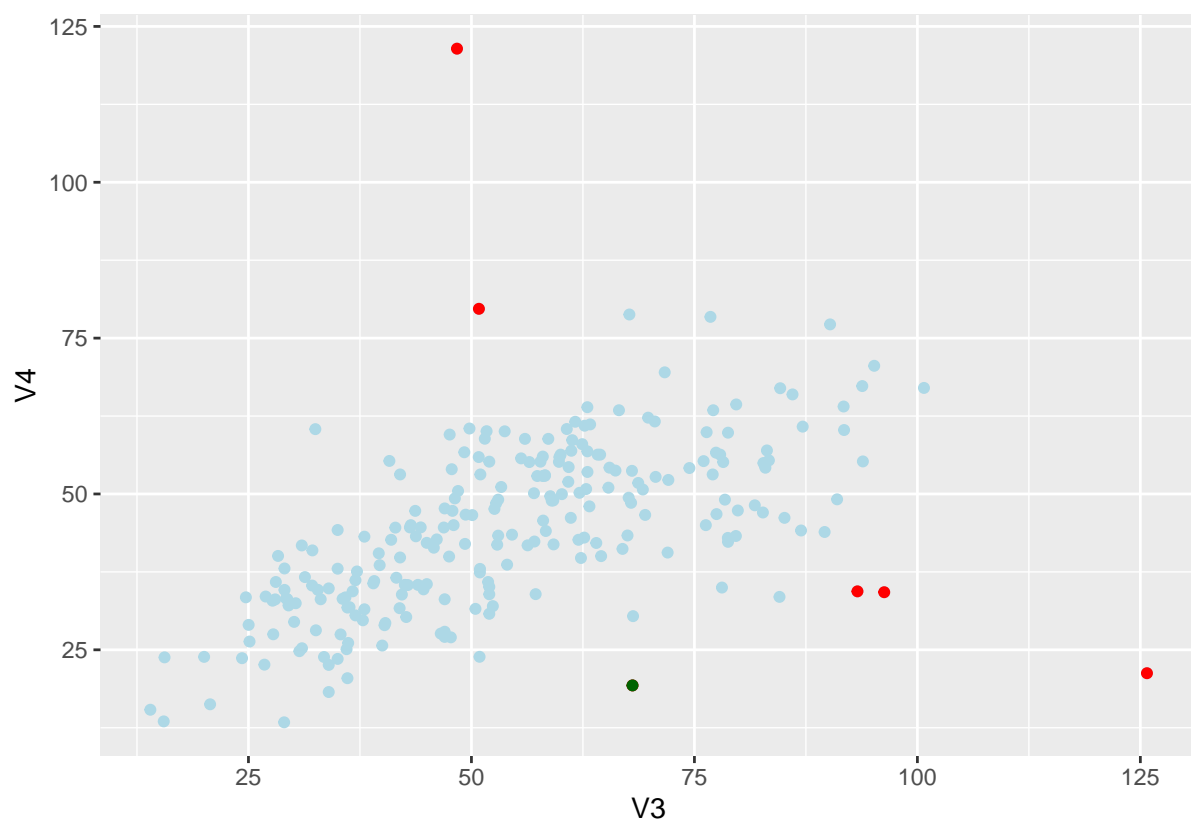


En el caso del primer outlier, podemos comprobar que tiene un valor anormalmente alto para la variable V4, por lo que en realidad no es uno de esos puntos que estamos buscando, que tienen valores normales en ambas variables pero su combinación es anormal. No obstante, si en lugar de utilizar el test como un test de tipo a lo utilizamos como un test de tipo b, corrigiendo adecuadamente el nivel de significación, lo que obtenemos es lo siguiente:

Table 3: Valores normalizados para los outliers encontrados con el test de Cerioli de tipo b

	V3	V4
116	-0.3185214	5.5054685
163	-0.1900641	2.5195172

	V3	V4
180	0.7091368	-1.8030642
198	3.7210944	-1.6628182
202	2.1827402	-0.7333308
203	2.0260850	-0.7233133



Podemos observar que ahora sí obtenemos resultados más interesantes. Aunque los valores para las variables del punto con índice 180, marcado en verde en el gráfico, no son especialmente anómalos (para la variable V3 el valor está a menos de 1 unidad de la media, y para la variable V4 a menos de 2), la combinación de valores sí que es atípica, y por esto el algoritmo nos marca el punto como anómalo. Algo similar podemos decir de los puntos con índice 163, 202 y 203, aunque para éstos el valor de una de las variables es ligeramente alto.

Podemos extraer conclusiones interesantes observando los resultados obtenidos por el test de Rosner y los resultados obtenidos por este método. El test de Rosner sólo nos arrojó resultados positivos para un punto en cada una de las dos variables que estamos utilizando. Probablemente, estos dos puntos

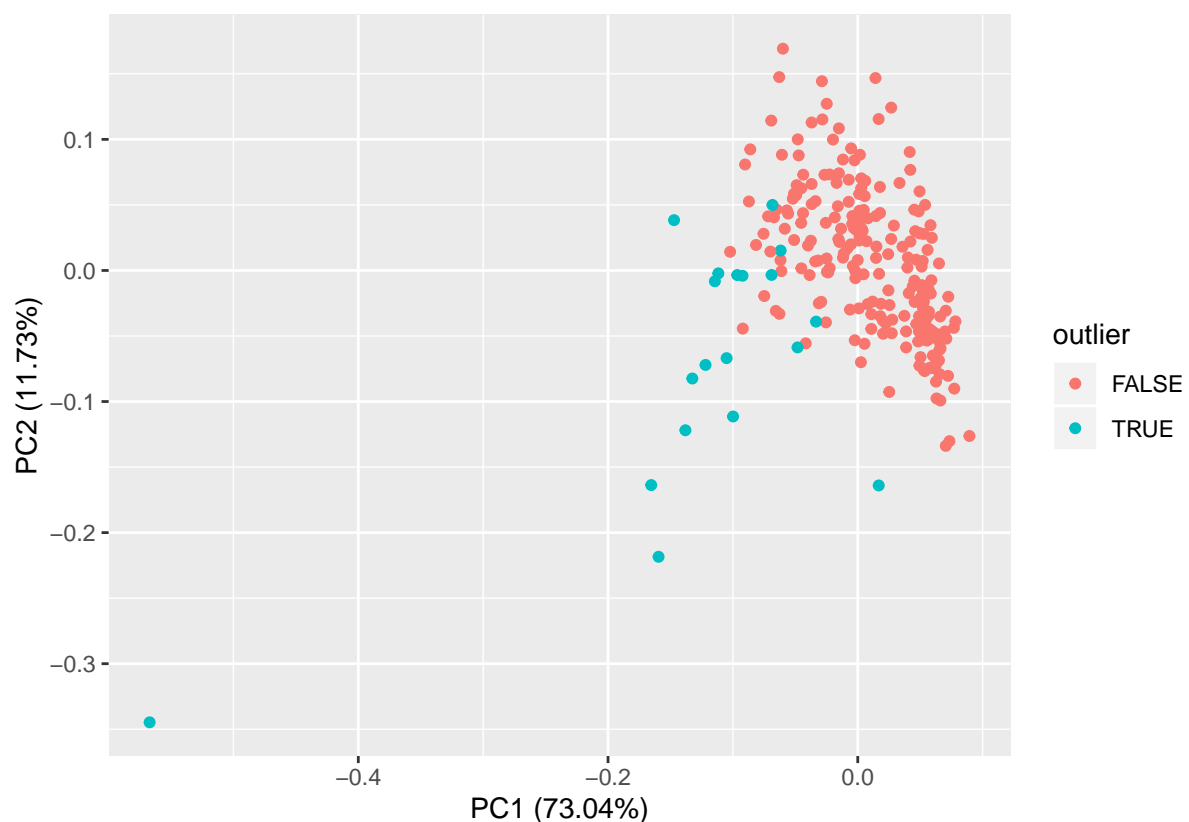
correspondan a los dos valores más extremos en cada uno de los ejes de la nube de puntos anterior (los que corresponden a los índices 116 y 198 en la tabla). Ahora, al combinar las dos variables y trabajar con un test multidimensional, tenemos muchos más registros con valores atípicos, los cuales resultan más interesantes de estudiar que las anomalías univariadas.

Este test puede ser aplicado a distribuciones de mayor dimensionalidad, no sólo para distribuciones bidimensionales. Podemos utilizar el test sobre el conjunto de datos al completo. No obstante, existe una dependencia lineal muy alta entre algunas de nuestras variables, y si intentamos ejecutar el algoritmo sobre las 6 columnas se nos produce un error porque la matriz de covarianzas con la que trabaja el test se convierte en una matriz singular y no es diagonalizable. Esto nos ha obligado a descartar la primera variable y trabajar con las otras cinco. El resultado que obtenemos es el siguiente (la representación gráfica se realiza sobre las dos primeras componentes principales del conjunto de datos:

Table 4: Valores normalizados para los outliers encontrados con el test de Cerioli de tipo b sobre el conjunto de datos completo

	V2	V3	V4	V5	V6
72	1.3427818	-0.3493302	0.6784310	1.4000975	1.7133342
76	2.0036006	0.7122699	-1.0080985	2.3821844	2.8100161
77	1.6410614	0.7697102	0.4451647	1.7162638	1.9429190
96	1.4118368	-0.1864088	-1.4746310	1.8309014	2.8946661
115	1.7177892	1.6960647	-0.0249455	1.8389334	1.3152030
116	-1.0098867	-0.3185214	5.5054685	-0.5998549	9.6714369
123	2.7948568	-0.1086034	-0.8928964	-0.3793417	0.8595500
142	2.8744619	0.9148773	-0.2782469	1.3672396	2.1310580
144	0.1602134	0.5253280	-0.3190328	0.0967331	1.7922069
163	1.8722041	-0.1900641	2.5195172	-2.5472346	1.0180489
180	2.9252942	0.7091368	-1.8030642	0.6655403	0.5320023
181	-1.3858533	-1.5545310	-0.7920052	3.0627086	0.0025004
184	1.7830079	1.5696961	-0.7869964	0.6757628	0.8605547
193	1.3763502	0.6506522	-0.2360300	2.2317681	2.2975949
198	1.7887625	3.7210944	-1.6628182	1.4402572	2.1049345

	V2	V3	V4	V5	V6
202	1.3763502	2.1827402	-0.7333308	2.1682428	1.3762414
203	2.2059703	2.0260850	-0.7233133	1.1949180	1.7007748
206	1.4396507	1.5989384	0.1203092	0.7071604	1.6774145
207	2.6490738	0.2360381	0.3177985	-1.4037788	1.0994335



Donde de nuevo podemos observar que tenemos algunos outliers multidimensionales sobre el conjunto completo. Si quitamos aquellos outliers en los que al menos una variable se desvía más de 2σ de su media, los outliers que nos quedan son los siguientes:

Table 5: Valores outliers multivariados puros

	V2	V3	V4	V5	V6
72	1.3427818	-0.3493302	0.6784310	1.4000975	1.7133342
77	1.6410614	0.7697102	0.4451647	1.7162638	1.9429190

	V2	V3	V4	V5	V6
115	1.7177892	1.6960647	-0.0249455	1.8389334	1.3152030
144	0.1602134	0.5253280	-0.3190328	0.0967331	1.7922069
184	1.7830079	1.5696961	-0.7869964	0.6757628	0.8605547
206	1.4396507	1.5989384	0.1203092	0.7071604	1.6774145

Una vez hemos visto cómo detectar outliers utilizando tests estadísticos, en el siguiente apartado trataremos de utilizar modelos basados en distancias para la detección de anomalías.

5 Modelos basados en distancias para detección de anomalías

En este apartado, utilizaremos métodos no paramétricos basados en distancias para la detección de puntos anómalos en nuestro conjunto de datos. La principal ventaja que tienen estos métodos respecto a los anteriores es que no necesitan suponer una distribución sobre los datos a priori. Los test estadísticos que hemos utilizado previamente asumían que los datos seguían una distribución normal, lo cual es una asunción bastante fuerte, que no siempre se cumple cuando se trabaja en un problema real. Los métodos que vamos a utilizar a continuación no necesitan hacer estas suposiciones, dado que son métodos no paramétricos.

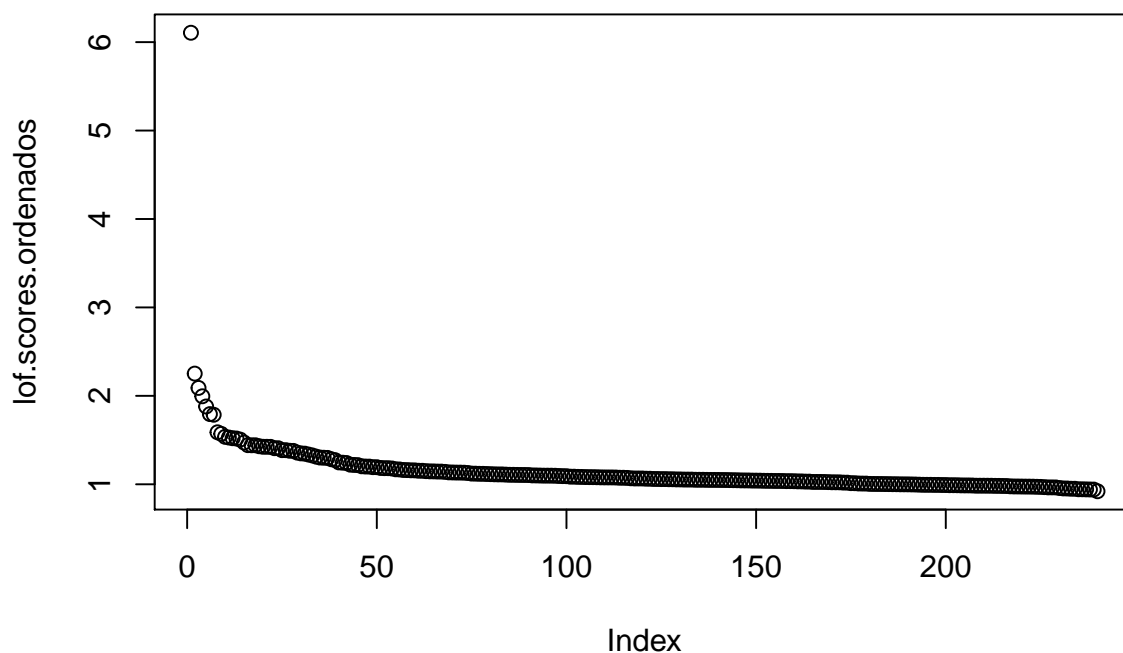
Vamos a estudiar tres métodos distintos basados en distancias. El primero de ellos, conocido como Local Outlier Factor, se basa en la idea del vecino más cercano. Los dos métodos restantes son modificaciones de algoritmos clásicos de clustering, concretamente de k-means y PAM, los cuales marcan los puntos como anómalos en función de la distancia que haya entre cada punto y el centro del cluster al que pertenece.

Comenzamos estudiando el método basado en vecindad.

5.1 LOF

En este apartado, veremos un método basado en distancias para la detección de datos anómalos. Concretamente, este algoritmo, conocido como Local Outlier Factor (LOF), se basa en la idea de vecinos más cercanos. Para cada punto del conjunto de datos, se calcula una puntuación, la cual se calcula como una función dependiente de la distancia de dicho punto a sus k vecinos más cercanos, cuanto mayor sea este valor, más aislado del resto de puntos del conjunto se encuentra el dato en cuestión. Para este algoritmo nos hace falta, por tanto, establecer la función de distancia con la que trabajaremos,

que en nuestro caso será la distancia euclídea porque todas las variables con las que trabajamos son numéricas y carecemos de información para poder utilizar una distancia más apropiada, y el valor de k . En principio, estableceremos dicho valor a 5, ya que no tenemos mucha información sobre el conjunto de datos.



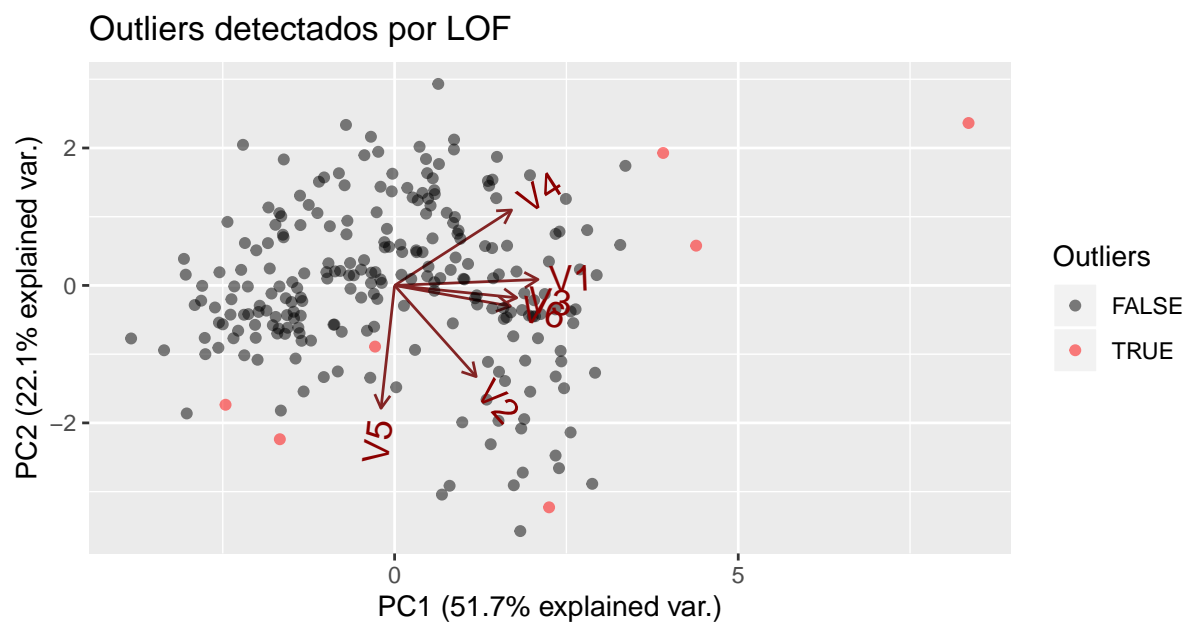
En el plot anterior podemos ver como hay un primer grupo de valores anómalos detectados por este método, y tras un pequeño salto, se produce una cierta estabilización de dichos valores, que empiezan a decrecer a un ritmo más lento. Concretamente, en el primer grupo hay siete puntos, contando con el outlier con un valor superior a 6. Este método no nos marca el número de outliers, si no que nos da un determinado coeficiente para cada punto. Tras calcular ese coeficiente, tenemos que decidir nosotros cuál es el valor a partir del cual se considera que un punto es anómalo. Esta es una de las problemáticas del algoritmo. El índice de anomalía que se nos da no es fácilmente interpretable. Por la construcción del mismo, es evidente que un valor menor o igual que 1 representa un punto no anómalo, pero para cualquier valor superior a 1, es difícil establecer el punto en el que los ejemplos empiezan a considerarse anómalos. El valor óptimo es muy dependiente del dataset, y por lo general difícil de establecer. Como ya hemos comentado, inspeccionando los valores ordenados, parece sensato establecer el número de outliers en 7:

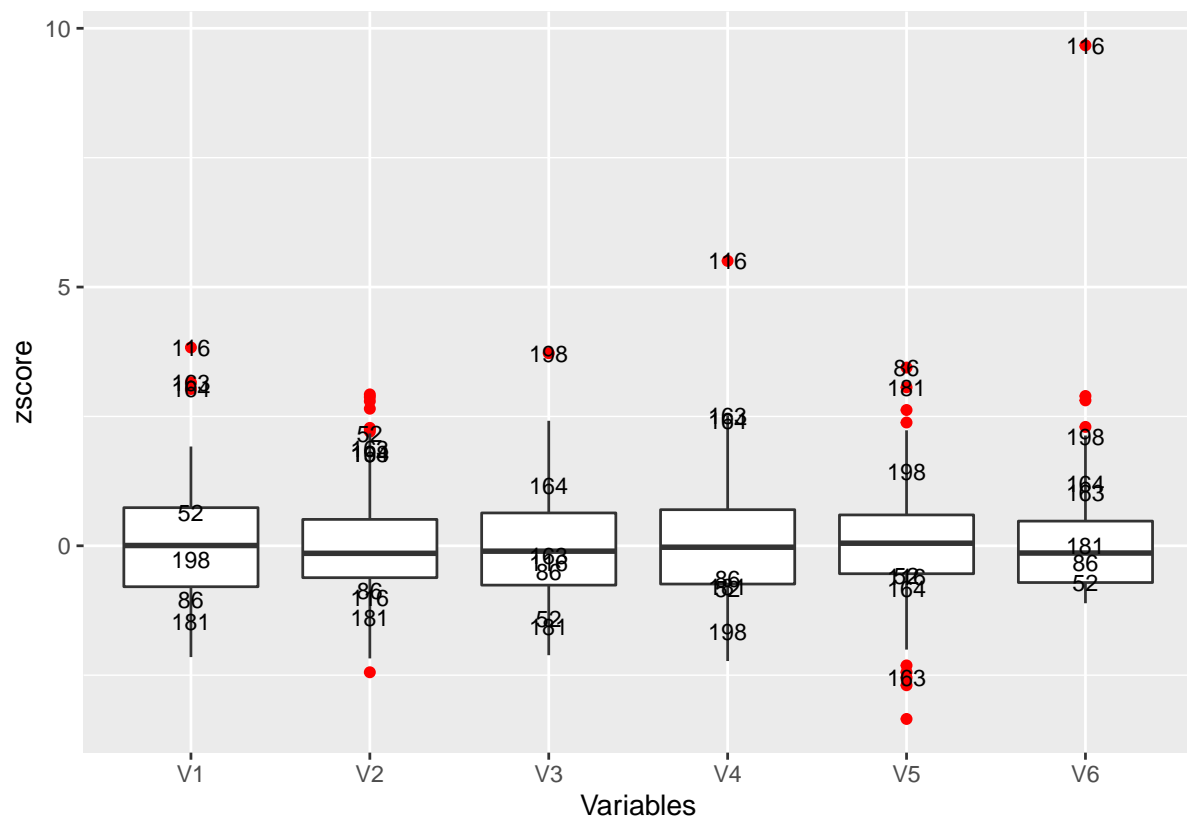
##

NA

NA


```
## Warning: Removed 7 rows containing missing values (geom_text).
```





En el boxplot anterior podemos comprobar que muchos de los puntos que hemos detectado tienen algún outlier univariante. Como lo que nos interesa con este método es buscar outliers multivariantes, vamos a tratar de limpiar los elementos anómalos univariantes de las anomalías que hemos obtenido. Una vez hemos etiquetado los puntos anómalos con LOF, vamos a eliminar de este conjunto aquellos puntos que tengan algún outlier univariante. Para esto, extraeremos, utilizando el método IQR, los registros que tengan algún outlier para alguna variable.

##

NA

```
## Warning: Removed 1 rows containing missing values (geom_text).
```

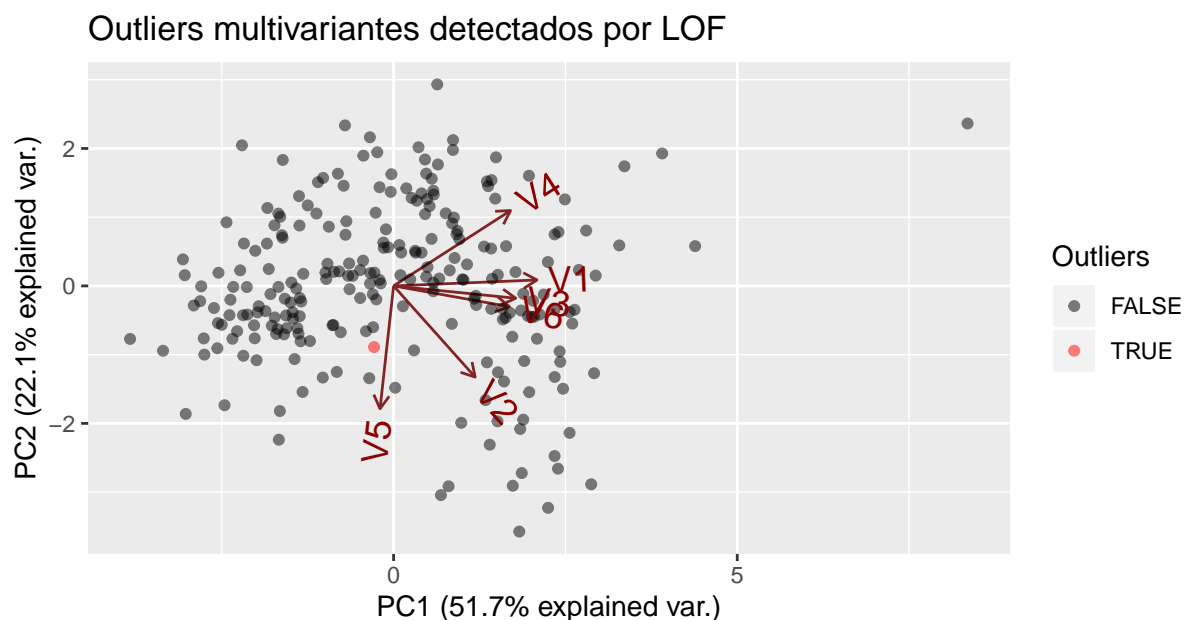


Table 6: Valores de las variables normalizadas para el outlier multivariante detectado con LOF

V1	V2	V3	V4	V5	V6
0.6354491	2.170484	-1.398398	-0.8306444	-0.5808703	-0.7161453

Como podemos observar en la gráfica anterior, casi todos los outliers que se han detectado por este método son outliers univariantes. Tras limpiar los outliers detectados por el método IQR, sólo nos ha quedado un punto relevante. Además, este punto tiene un valor ligeramente alto para la segunda variable, como se puede ver en la tabla anterior. A pesar de esto, no es un punto con valores muy extremos, como podemos comprobar en la nube de puntos previa. El hecho de que aparezca rodeado de otros puntos nos indica que para ninguna de las variables se tiene un valor excesivamente alto, y que por tanto es posible que hayamos encontrado realmente un outlier multivariante.

Una vez hemos visto el funcionamiento del algoritmo basado en vecindad, pasamos a comentar los métodos basados en clustering.

5.2 Métodos basados en clustering

Los métodos de agrupamiento basados en clustering que vamos a estudiar se basan todos en la misma idea. Cada cluster del conjunto estará representado por un elemento característico o centroide, y cada ejemplo será considerado como más o menos anómalo en función de la distancia a la que se encuentre del centroide de su grupo. La diferencia entre los algoritmos que veremos reside en la forma que tienen de buscar los centroides. Comenzamos con el método conocido como k-means

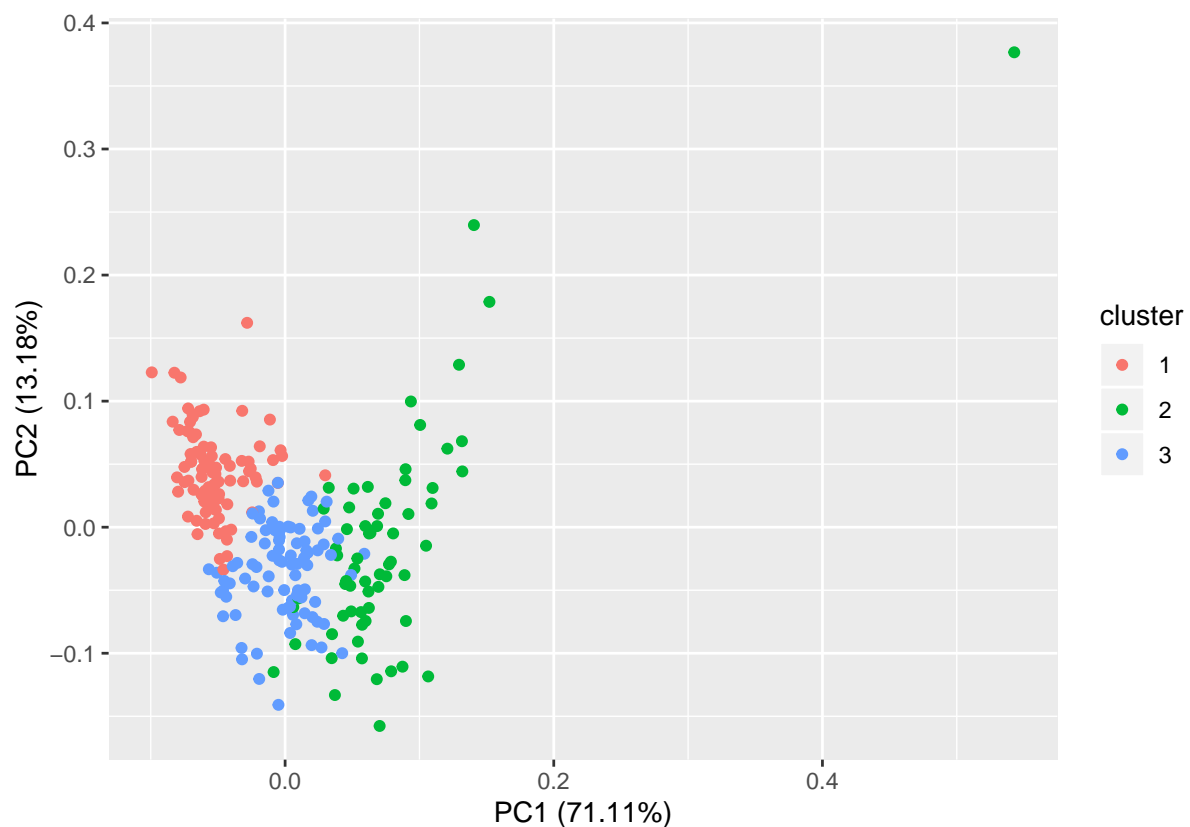
5.2.1 K-means

El algoritmo de k-means o algoritmo de las k-medias es un algoritmo de agrupamiento cuyo funcionamiento se basa en la construcción de instancias artificiales, que funcionan como centros de sus clusters. El funcionamiento del algoritmo es el que sigue. Se debe establecer a priori el número k de grupos que van a formarse. Una vez especificado dicho valor:

- Se inicializan aleatoriamente k elementos artificiales en el espacio de características (centroides)
- Se asigna cada ejemplo del conjunto al centroide más cercano
- Mientras no se alcance un punto de equilibrio (la asignación de los puntos a los centroides no cambie en una iteración)
 - Se recalcula la posición de los centroides como la media aritmética de los puntos que pertenecen al cluster
 - Se reasignan los puntos a su cluster más cercano.

Como se puede intuir de la descripción del algoritmo, este método es dependiente de la función de distancia que se utilice. Usualmente, para atributos numéricos, suelen usarse distancias de Minkowski, y más concretamente la distancia de Minkowski para $k = 2$ (distancia euclídea). No obstante, para otro tipo de dato, o con el fin de aportar información al problema, se puede utilizar otra función de distancia. No estudiaremos cómo afecta el cambio de la distancia al resultado del algoritmo, ya que carece de interés por el momento. En primera instancia, nos limitaremos al uso de la distancia euclídea, la cual está bien definida en nuestro problema, ya que todos los atributos de nuestro conjunto de datos son atributos numéricos continuos. Más adelante, trataremos de mejorar los resultados utilizando otras funciones de distancia.

Mostramos una ejecución del algoritmo. Trabajaremos con tres clusters en primer lugar. En la gráfica puede observarse el resultado de la ejecución del algoritmo. Es recomendable remarcar que, aunque la visualización la hemos hecho sobre los datos originales, el algoritmo se ha ejecutado sobre los datos normalizados, ya que las diferencias en la varianza de las distintas variables afectan notablemente en las distancias calculadas, por lo que se debe trabajar con los datos normalizados para evitar introducir un sesgo.



Una vez tenemos los resultados del modelo, se nos proporcionan también las coordenadas de los tres centroides:

Table 7: Coordenadas normalizadas de los centroides calculados por el algoritmo

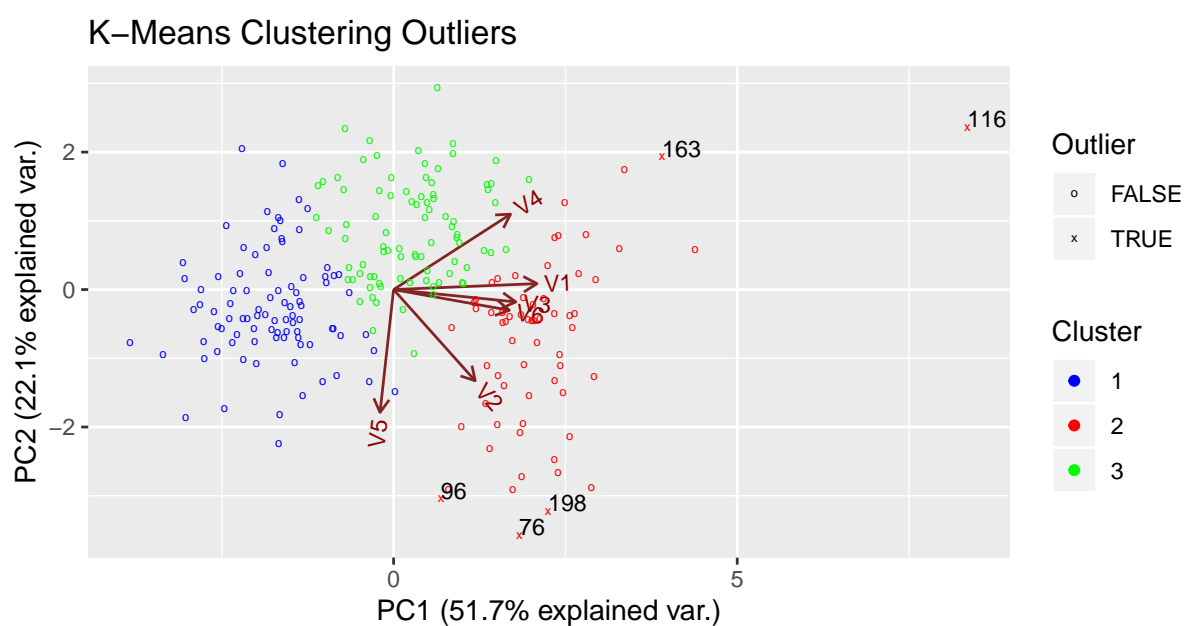
	V1	V2	V3	V4	V5	V6
1	-0.9613104	-0.4430808	-0.9239822	-0.8614206	0.3343162	-0.6513279
2	1.1265992	1.1233280	1.0597100	0.5589021	0.3724088	1.0640711
3	0.1676231	-0.3722758	0.1783367	0.4855595	-0.6270073	-0.1102446

Con esta información, podemos calcular las distancias de cada punto a su centroide. Tras esto, ordenando las distancias obtenidas, tenemos los puntos que más se alejan de sus respectivos centroides, y por tanto los datos peor representados por su conjunto. Mostramos a continuación los primeros outliers de nuestro conjunto de datos, ordenados en función de la distancia a su centroide:

Table 8: Índice y distancia de los 15 puntos más alejados de sus respectivos centroides

116	163	198	96	76	52	86	180	181	169	164	10	168	135	202
10.64	4.31	4.08	3.66	3.31	3.23	3.19	3.16	3.07	3.04	3	2.87	2.8	2.78	2.67

La problemática principal de este método, al igual que ocurría con LOF, es el hecho de tener que identificar los outliers a partir de los valores de distancia anteriores, cosa que puede no ser sencilla en todos los casos. Observando los valores, tenemos un punto significativamente alejado de su centroide, tras este otros cuatro puntos en los que la distancia aún se decrementa rápidamente entre uno y otro, y después los valores comienzan a estabilizarse. Consideramos, por tanto, que existen 5 outliers en nuestro conjunto de datos. Mostramos dichos outliers en una gráfica sobre el conjunto de datos completo:



En el biplot se muestran los 5 puntos marcados como outliers. Se puede observar que todos los puntos se encuentran en la parte más externa del gráfico, lo cual suele indicar que los outliers son univariantes. En efecto, en la siguiente tabla podemos observar que, en todos los casos, al menos una de las columnas presenta un valor alto:

Table 9: Valores de las características de los primeros 5 outliers encontrados por el algoritmo k-means

	V1	V2	V3	V4	V5	V6
116	3.8323022	-1.009887	-0.3185214	5.505469	-0.5998549	9.671437
163	3.1577315	1.872204	-0.1900641	2.519517	-2.5472346	1.018049
198	-0.2647478	1.788763	3.7210944	-1.662818	1.4402572	2.104935
96	-0.3403413	1.411837	-0.1864088	-1.474631	1.8309014	2.894666
76	0.3925113	2.003601	0.7122699	-1.008099	2.3821844	2.810016

Esto nos indica que este método no parece una buena alternativa para la búsqueda de outliers multi-variantes.

// TODO: aumentar el número de outliers

5.2.2 PAM

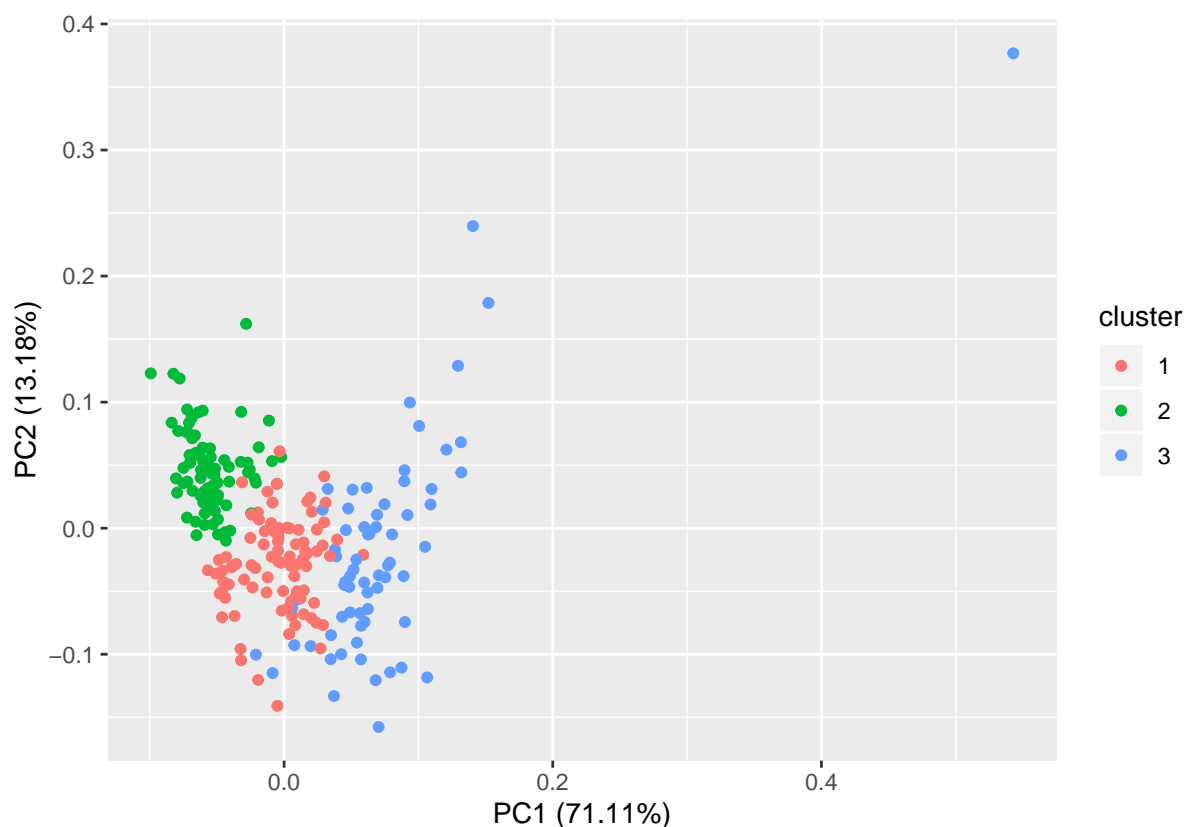
En este apartado, utilizaremos una política distinta para la selección de los centroides de los clusters. Este método de agrupamiento, conocido como particionado alrededor de medioides (*partitioning around medoids*, PAM por sus siglas en inglés), selecciona los centros entre los puntos que hay en el conjunto, en lugar de generar instancias artificiales.

Existen diversos algoritmos que computan soluciones de PAM. El más típico es una búsqueda voraz, que no encuentra necesariamente el óptimo, pero aporta buenas soluciones. Su funcionamiento es parecido al que describimos para el método de las k-medias:

1. Se inicializan aleatoriamente los k medioides
2. Se asocia cada punto al cluster definido por el medioide más cercano
3. Mientras el coste final de la configuración decrezca:
 - Para cada medioide m_i y cada punto no medioide o
 1. Se intercambian los papeles de dichos puntos y se recalcula el coste
 2. Si el coste es mejor que el resultado actual, se guarda el intercambio
 - Se realiza el mejor intercambio encontrado. Si no se ha encontrado un intercambio mejor que el resultado actual, el algoritmo termina

El coste de la configuración se define como la suma de las distancias de cada punto al medioide de su *cluster*, de la misma manera que ocurre en el algoritmo de las k-medias.

A continuación vamos a estudiar cómo podemos utilizar este algoritmo para la detección de elementos anómalos en nuestro conjunto de datos. Mostramos primero el resultado del algoritmo de clustering sobre el conjunto de datos:



Como podemos observar, el resultado apenas difiere al obtenido por el algoritmo k-means, aunque sí que es posible encontrar algunos puntos cuyo cluster ha cambiado. Esto se debe a que el funcionamiento de ambos algoritmos es muy similar, y se basan en concepciones similares, pero el hecho de forzar que los centros de los clusters sean ejemplos del conjunto de datos hace que algunos puntos en las fronteras pasen de unos grupos a otros. Al igual que hicimos con el algoritmo anterior, vamos a ver las distancias de los puntos más alejados a sus respectivos centroides.

Table 10: Índice y distancia de los 15 puntos más alejados de sus respectivos medioides

116	198	163	96	76	180	52	10	168	86	164	202	193	142	181
10.86	4.23	4.18	4.03	3.74	3.29	3.22	3.15	3.11	2.96	2.94	2.91	2.9	2.86	2.86

Podemos ver cómo el resultado obtenido es similar al obtenido por el método k-means. Esto se debe a

que ambos algoritmos tienen un funcionamiento muy similar. Existen pequeñas diferencias entre los mismos (por ejemplo, los outliers 2 y 3 han intercambiado sus posiciones), pero en general el resultado obtenido es muy parecido. Además, las distancias se han modificado muy poco, y siguiendo la misma política que antes, seleccionaríamos 5 datos anómalos, que como hemos estudiado anteriormente, son outliers univariantes, por lo que no son especialmente interesantes.

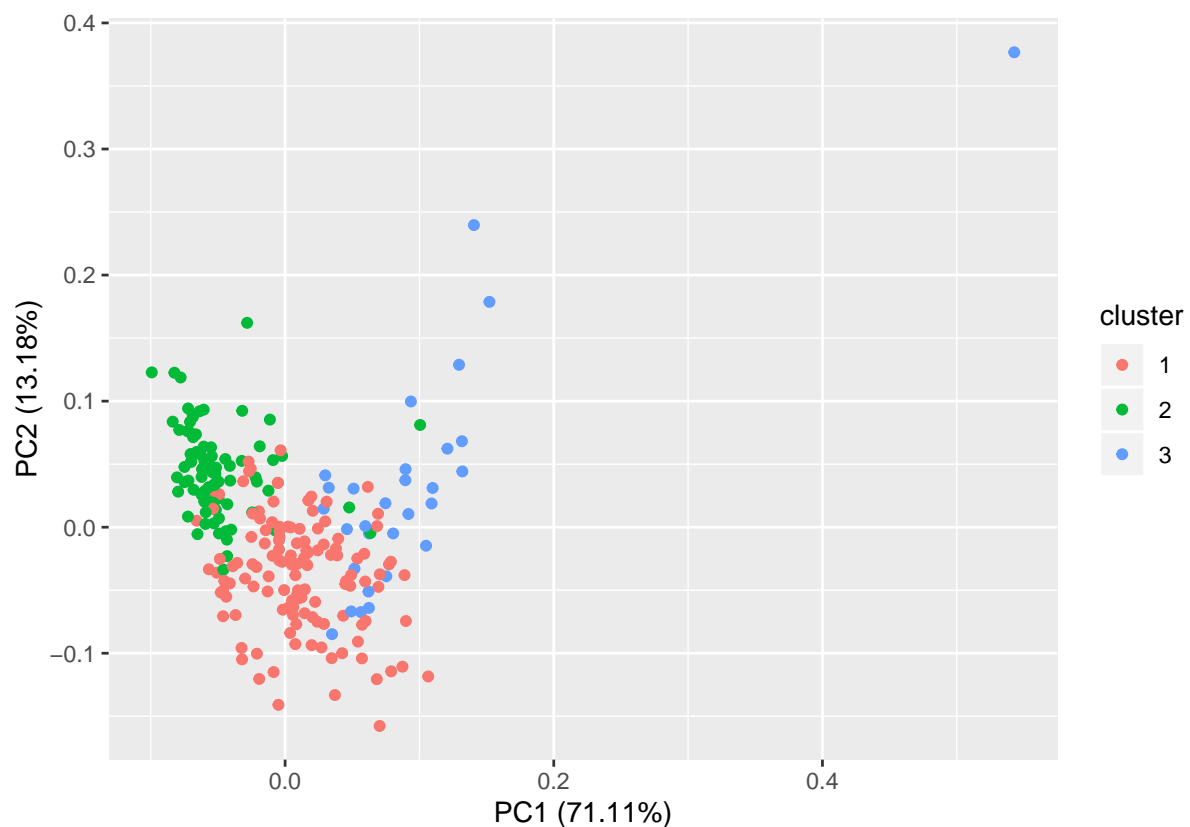
Estos métodos de clustering, al estar basados en distancias, nos permiten cambiar la función de distancia entre dos puntos y obtener así distintos resultados, posiblemente más interesantes que los obtenidos con la distancia euclídea.

5.2.3 Otras métricas de distancia

En este apartado, trataremos de estudiar cómo afecta el uso de una determinada medida de distancia en el resultado. En lugar de utilizar la distancia euclídea, que es menos informativa, utilizaremos la distancia de Mahalanobis, que tiene en cuenta las varianzas de las variables y sus covarianzas, de forma que la distancia en una variable que tiene una varianza pequeña se considera más relevante que una en la que la covarianza es mayor. Matemáticamente, la distancia de Mahalanobis se define como sigue. Sean x e y dos vectores aleatorios provenientes de una misma distribución con matriz de covarianzas S , se define la distancia de Mahalanobis entre x e y como

$$d_M(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

Podemos observar que, si todas las componentes del vector aleatorio son independientes (la covarianza entre ellas es 0), e idénticamente distribuidas (sus varianzas coinciden), nos reducimos a la distancia euclídea. Vamos a ver a continuación los resultados obtenidos por el método PAM al cambiar la función de distancia entre nuestros puntos. En primer lugar, mostramos el resultado obtenido por el algoritmo de agrupamiento al cambiar la distancia:



Ahora sí que podemos detectar grandes diferencias entre el resultado obtenido por los otros dos algoritmos y el resultado actual. El cambio en la función de distancia ha hecho que los agrupamientos sean significativamente distintos en este caso. Al igual que hicimos anteriormente, vamos a mostrar los puntos del conjunto ordenados en función de la distancia de Mahalanobis al centroide de su cluster.

Table 11: Índice y distancia de los 15 puntos más alejados de sus respectivos medioides

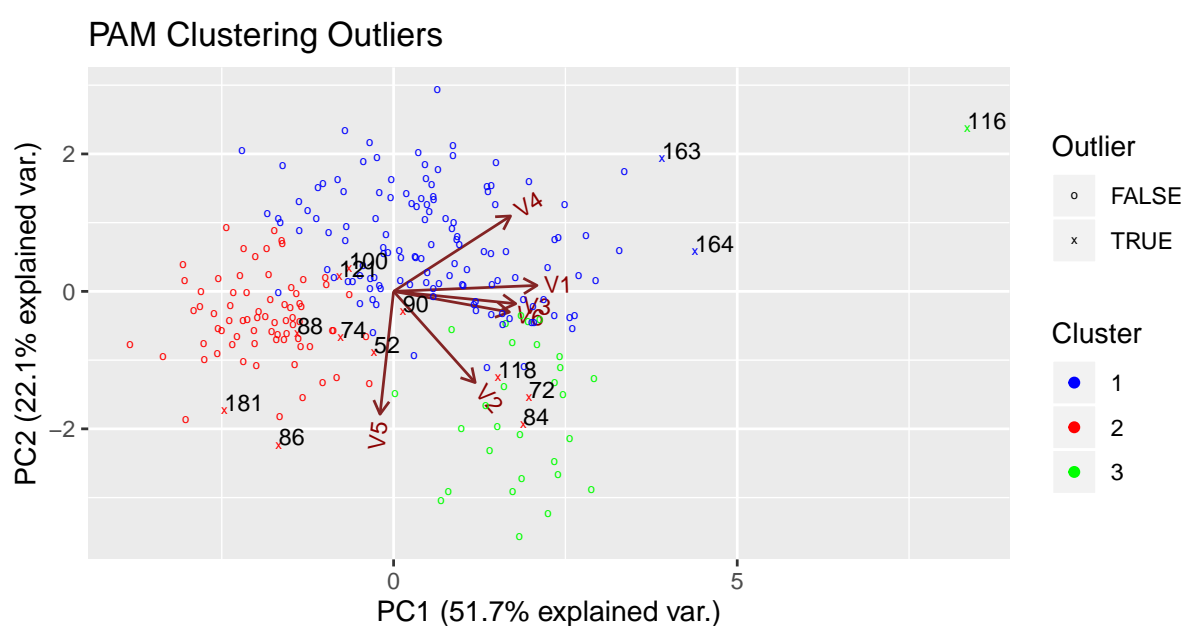
72	116	118	84	181	163	121	88
374.71	369.95	144.23	135.49	72.27	66.33	61.06	50.27

86	74	164	100	52	90	63
49.66	40.87	35.15	32.02	29.4	28.63	27.65

Lo primero que podemos observar es cómo ha variado la magnitud de las distancias al cambiar la función que hemos utilizado. Otro detalle significativo es que el mayor outlier ha cambiado, pasando

a ser un punto que ni siquiera estaba entre los 15 mayores outliers de ninguna de las ejecuciones anteriores. Este hecho pone de manifiesto las grandes diferencias que pueden surgir en un mismo algoritmo cambiando simplemente alguno de sus parámetros, en este caso, la función de distancia utilizada.

En cuanto al punto en el que dejamos de considerar como outliers a los puntos del conjunto, vamos a coger los 9 primeros, ya que consideramos que a partir del 10 la distancia parece estabilizarse y desciende más lentamente. Dichos puntos, representados en un biplot, son los siguientes:



Podemos observar cómo ahora sí que tenemos outliers más interesantes, y no outliers univariantes exclusivamente. De hecho, si mostramos los valores de las variables para estos outliers:

Table 13: Valores de las características de los primeros 14 outliers encontrados por el algoritmo PAM

	V1	V2	V3	V4	V5	V6
72	1.355030	1.3427818	-0.3493302	0.6784310	1.4000975	1.7133342
116	3.832302	-1.0098867	-0.3185214	5.5054685	-0.5998549	9.6714369
118	1.305403	1.9009770	-0.3451528	0.2004498	0.4515987	0.7153687

	V1	V2	V3	V4	V5	V6
84	1.020341	0.5620756	1.2224437	0.8458675	2.6238727	0.7962509
181	-1.472512	-1.3858533	-1.5545310	-0.7920052	3.0627086	0.0025004
163	3.157732	1.8722041	-0.1900641	2.5195172	-2.5472346	1.0180489
121	-0.500761	-0.6550203	-0.7038932	-0.1315611	0.1266703	0.1725539
88	-1.068000	-0.8353308	-0.5143404	-0.7004160	1.0350095	0.0966954
86	-1.037417	-0.8650628	-0.4950196	-0.6403105	3.4438605	-0.3313270

Podemos observar que los resultados que obtenemos aquí son mucho más interesantes que los obtenidos por los anteriores métodos. En primer lugar, el outlier de mayor distancia es claramente un outlier univariante. Si recordamos los tests estadísticos, era una de las anomalías detectadas por el test de Cerioli que no se eliminaban con IQR. Para el resto de outliers, tenemos algunos que son claramente univariantes (por ejemplo, el punto 116, que ha aparecido en varias ocasiones con los métodos anteriores, o el 163, que era el tercer outlier más alto para el método PAM con la distancia euclídea), pero también hemos conseguido detectar varios outliers con combinaciones extrañas de variables, como el elemento 121, que tiene todos sus valores en el intervalo $[-1, 1]$ (hay que tener en cuenta que estamos mostrando los valores normalizados), o el elemento 88, al que le ocurre algo similar, aunque en algunas variables el valor es ligeramente mayor que 1 o menor que -1.

Parece más interesante, por tanto, el uso de estas distancias a la hora de encontrar outliers multivariantes. El uso de la distancia euclídea tiene la problemática de que no pondera bien la dispersión que existe entre los datos. Aunque trabajemos con las variables normalizadas, un valor muy alto para una variable provocado por una anomalía muy fuerte en la misma sigue siendo demasiado representativa. El uso de la distancia de Mahalanobis reduce aún más esta problemática, provocando el fenómeno que hemos visto, en el que tenemos variables que, a pesar de tener unos valores más o menos normales en todas las variables, tienen una distancia muy grande, dado que la combinación de dichos valores es extraña.

Finalmente, vamos a utilizar la distancia relativa como medida de distancia. En este caso, lo que se hace es dividir la distancia euclídea de cada punto por la mediana de las distancias de su cluster. De esta manera, se tiene en cuenta la densidad de los clusters, y no se ven penalizados puntos que están en zonas del espacio poco densas, mientras que aquellos que se encuentran en zonas con gran densidad serán considerados anómalos con una desviación menor. El problema que presenta esta métrica es que no podemos calcular el resultado del algoritmo de agrupamiento con esta función de distancia, ya que los métodos de los que disponemos no permiten especificar la función de distancia de la forma en la que pretendemos calcularla. Lo que haremos en este caso será utilizar la distancia que hemos

especificado sobre el resultado obtenido por el algoritmo de clustering que utiliza la distancia euclídea estándar. En la siguiente tabla podemos observar los nuevos resultados obtenidos:

Table 14: Índice y distancia de los 15 puntos más alejados de sus respectivos medioides

116	198	163	96	76	180	52	10	164	202	193	86	142	181	123
9.41	3.66	3.63	3.49	3.24	2.85	2.74	2.68	2.55	2.52	2.52	2.51	2.48	2.43	2.36

Podemos ver cómo ha cambiado el resultado respecto al obtenido con la distancia euclídea. Las distancias han variado ligeramente respecto a las que obtuvimos previamente. A pesar de las diferencias, los primeros 8 outliers no han variado, por lo que este método no parece el más adecuado para el conjunto de datos con el que estamos trabajando. Resulta mucho más interesante el enfoque anterior, que nos permitió detectar outliers multivariantes.

6 Resultados de la detección de anomalías

En este apartado, veremos si los métodos que hemos utilizado para la detección de anomalías nos han servido para identificar los verdaderos valores marcados como anómalos en nuestro conjunto de datos. Como ya dijimos anteriormente, el conjunto de datos viene etiquetado con los valores que son anómalos. En concreto, si tomamos el orden de los elementos que trae el dataset de partida, los últimos 30 elementos son los que corresponden a ejemplos anómalos. Vamos a ver a continuación si los métodos que hemos estudiado nos habrían permitido identificar correctamente los elementos que se consideran anomalías. Comenzamos por los outliers detectados por el método IQR.

Table 15: Outliers detectados por IQR frente a los outliers reales

	FALSE	TRUE
0	188	22
1	30	0

Podemos comprobar que los resultados obtenidos son bastante malos. No hemos conseguido detectar ninguna de las anomalías presentes en el conjunto de datos. Esto se debe probablemente a que los elementos marcados como anómalos son realmente los normales, los cuales no presentarán valores extremos para estas características, ya que las pruebas médicas suelen presentar valores anormales

en los pacientes enfermos y no en los sanos.

Lo mismo nos ocurrirá con los tests estadísticos, por tanto, así que no mostraremos los resultados obtenidos por el test de Rosner (todos los puntos marcados como outliers por este test habían sido marcados previamente por IQR). Pasamos por tanto al test de Cerioli. No estudiaremos el outlier encontrado por el test de Cerioli de tipo A, ya que es un único punto. Estudiamos por tanto los outliers encontrados por el test de Cerioli de tipo B:

Table 16: Outliers detectados por el test de Cerioli frente a los outliers reales

	0	1
FALSE	191	30
TRUE	19	0

De nuevo, el test de cerioli no ha sido capaz de identificar ninguno de los elementos anómalos de nuestro conjunto de datos. Pasamos ahora a estudiar los resultados obtenidos por los métodos no paramétricos. Comenzamos viendo los outliers detectados por LOF. Aunque estemos utilizando conocimiento del conjunto de datos, dado que nuestra intención es comprobar si los algoritmos que estamos utilizando son capaces de marcar como outliers los ejemplos en los que estamos interesados, consideraremos como outliers los 30 puntos con un valor de outlier mayor, tantos como outliers sabemos que existen en nuestro conjunto. Esta aproximación no es del todo correcta, ya que estamos utilizando información de la que no dispondríamos a priori, pero aceptaremos la aproximación para ver si estos métodos funcionan correctamente con nuestro conjunto de datos:

Table 17: 30 primeros outliers detectados por LOF frente a los outliers reales

	0	1
FALSE	183	27
TRUE	27	3

Tenemos que, de los 30 elementos con una puntuación de LOF más elevada, sólomente 3 son realmente valores anómalos. Esto indica que este método no funciona tampoco especialmente bien para detectar los outliers de nuestro conjunto de datos. Pasamos a ver los métodos basados en clustering. Comenzamos con el método de k-means:

Table 18: 30 primeros outliers detectados por k-means frente a los outliers reales

	0	1
FALSE	181	29
TRUE	29	1

En este caso, sólomente 1 de los outliers detectados es realmente una anomalía. Pasamos al método PAM con la distancia euclídea:

Table 19: 30 primeros outliers detectados por PAM frente a los outliers reales

	0	1
FALSE	181	29
TRUE	29	1

Como era esperable, los resultados obtenidos son los mismos que por k-means. Como ya dijimos anteriormente, los algoritmos tienen un comportamiento muy parecido, por lo que era de esperar que los resultados que obtuviesen en este caso fueran muy similares. Finalmente, observamos los resultados obtenidos con las otras distancias:

Table 20: 30 primeros outliers detectados con la distancia de Mahalanobis frente a los outliers reales

	0	1
FALSE	181	29
TRUE	29	1

Table 21: 30 primeros outliers detectados con la distancia relativa frente a los outliers reales

	0	1
FALSE	181	29
TRUE	29	1

De nuevo, los dos resultados coinciden con los obtenidos anteriormente. Era natural pensar que iba a darse un resultado similar con estos algoritmos también, debido a que, aunque la función de distancia fuese distinta, y los resultados obtenidos por los mismos fueran diferentes, los puntos que estaban siendo marcados por estos algoritmos eran puntos con carácter anómalo. Como ya dijimos anteriormente, en este conjunto de datos los puntos marcados como anomalías no son aquellos que presentan la enfermedad, sino los individuos sanos. Este tipo de ejemplos suelen mostrar valores normales en las características, siendo los sujetos que padecen la enfermedad los que presentan mediciones anómalas. Por tanto, este dataset sería más conveniente abordarlo utilizando técnicas de clasificación no balanceada, en lugar de técnicas de detección de anomalías.