Distancia de Mahalanobis

Omar Sanchez Hernandez

4/6/2022

Ejercicio I

Ejercicio de Clase

Para desarrollar esta primera parte primero vamos utilizar datos propuestos.

```
# Cargar los datos
ventas= c( 1054, 1057, 1058, 1060, 1061, 1060, 1061,
           1062, 1062, 1064, 1062, 1062, 1064, 1056,
           1066, 1070)
clientes= c(63, 66, 68, 69, 68, 71, 70, 70, 71, 72, 72,
            73, 73, 75, 76, 78)
# Utilizamos la función data.frame() para crear
# un juego de datos en R
datos <- data.frame(ventas ,clientes)
```

Exploración de los datos

```
library(knitr)
dim(datos)
```

```
## [1] 16  2
```

```
str(datos)
```

```
## 'data.frame':    16 obs. of  2 variables:
##  $ ventas  : num  1054 1057 1058 1060 1061 ...
##  $ clientes: num  63 66 68 69 68 71 70 70 71 72 ...
```

```
kable(summary(datos))
```

ventas	clientes
Min. :1054	Min. :63.00
1st Qu.:1060	1st Qu.:68.75

ventas	clientes
Median :1062	Median :71.00
Mean :1061	Mean :70.94
3rd Qu.:1062	3rd Qu.:73.00
Max. :1070	Max. :78.00

La matriz de datos contiene 16 observaciones, 2 variables y todas las variables son numericas.

Calculo de la distancia de Mahalanobis

El método de distancia Mahalanobis mejora el método clásico de distancia de Gauss eliminando el efecto que pueden producir la correlación entre las variables a analizar

Determinar el número de outlier que queremos encontrar.

```
num.outliers <- 2
```

Ordenar los datos de mayor a menor distancia, según la métrica de Mahalanobis.

```
mah.ordenacion <- order(mahalanobis(datos, colMeans(datos), cov(datos)), decreasing=TRUE)
mah.ordenacion
```

```
##  [1] 14 16  1 15  2  5  3 10 13  8 12  4  6  7  9 11
```

Generar un vector boleano los dos valores más alejados segun la distancia Mahalanobis.

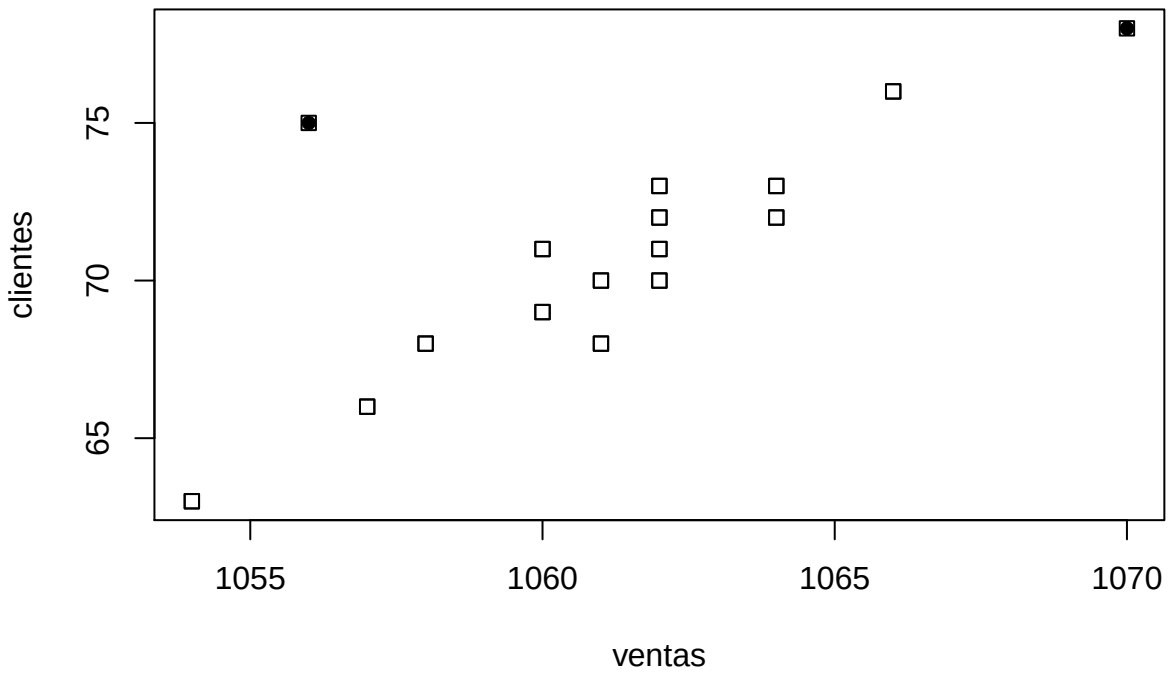
```
outlier2 <- rep(FALSE , nrow(datos))
outlier2[mah.ordenacion[1:num.outliers]] <- TRUE
```

Resaltar con un punto relleno los 2 valores outliers.

```
colorear.outlier <- outlier2 *16
```

Visualizar el gráfico con los datos destacando sus outlier.

```
plot(datos , pch=0)
points(datos , pch=colorear.outlier)
```



_____ Ejercicio II _____

Se generan datos, su matriz de varianzas y su distancia de mahalanbis

```
require(graphics)
ma <- cbind(1:6, 1:3)
(S <-  var(ma))
```

```
##      [,1] [,2]
## [1,]  3.5  0.8
## [2,]  0.8  0.8
```

```
mahalanobis(c(0, 0), 1:2, S)
```

```
## [1] 5.37037
```

```
x <- matrix(rnorm(100*3), ncol = 3)
stopifnot(mahalanobis(x, 0,
                      diag(ncol(x))) == rowSums(x*x))
```

Aqui se usa D^2 como la distancia Euclidea comun

```
Sx <- cov(x)
D2 <- mahalanobis(x, colMeans(x), Sx)
```

Gráfico de la densidad de las distancias de Mahalanobis

```
plot(density(D2, bw = 0.5),
     main="Squared Mahalanobis distances,
     n=100, p=3") ; rug(D2)
```

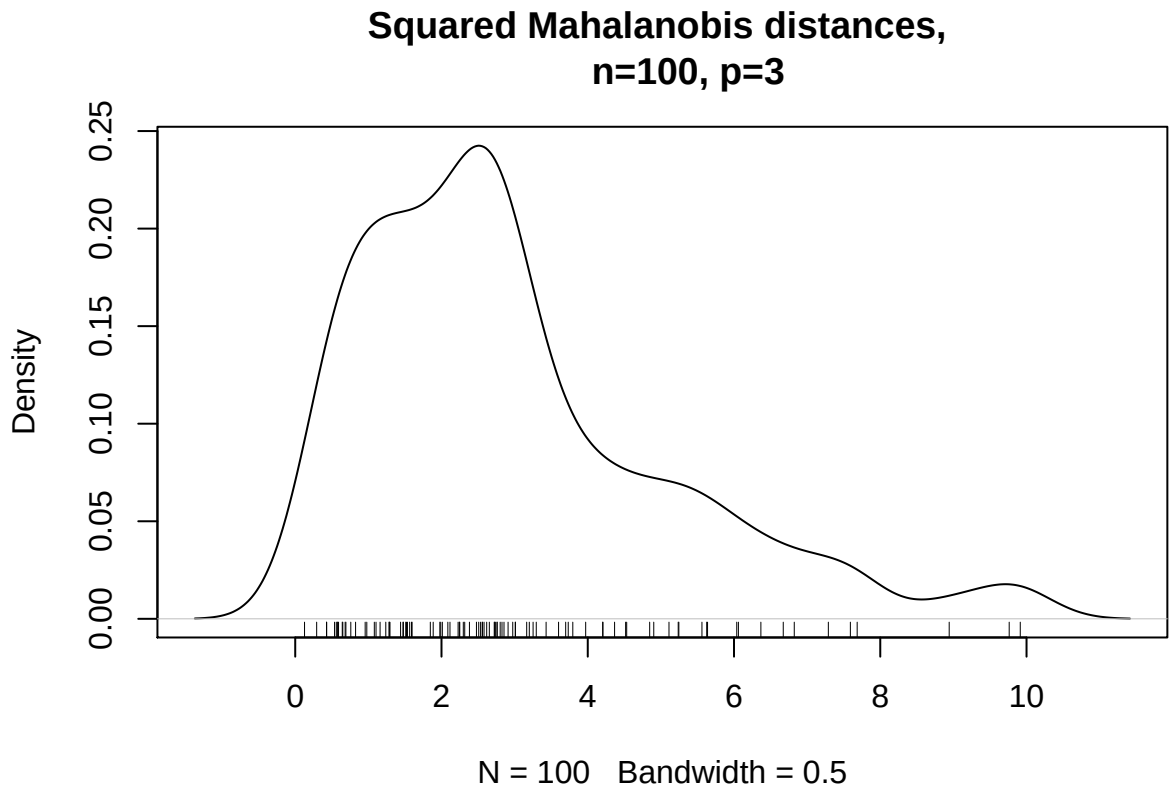
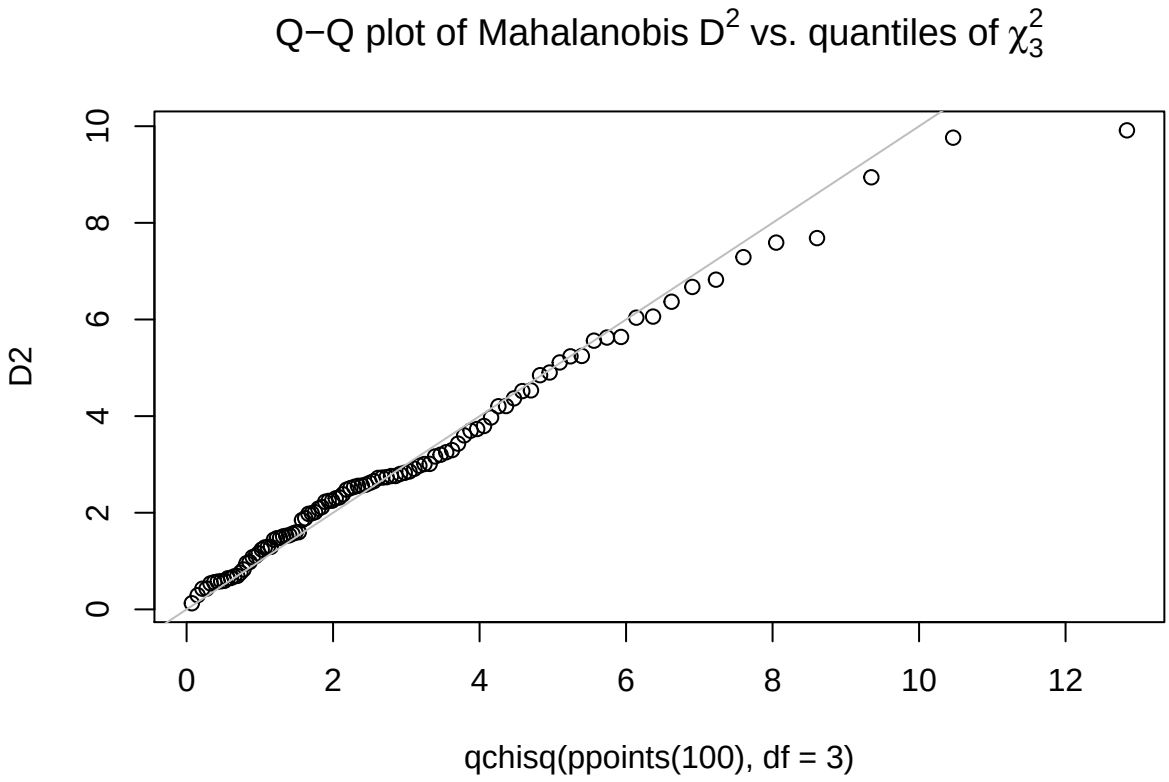


Gráfico qqplot sobre los datos

```
qqplot(qchisq(ppoints(100), df = 3), D2,
       main = expression("Q-Q plot of Mahalanobis" * ~D^2 *
                         " vs. quantiles of" * ~ chi[3]^2))
abline(0, 1, col = 'gray')
```



Ejercicio III

Planteamiento del problema.

Se desea encontrar si existen valores atípicos en la matriz de datos de la base de los palmer penguins a nivel multivariado, es por ello que se recurrirá a la distancia de Mahalanobis para encontrar los posibles valores atípicos y presentarlos en un gráfico 2d

Simular los datos o utilizar una matriz precargada en R.

Se carga la base de datos de los palmer penguins que esta incluida en el paquete ***"palmerpenguins"***

Se hace un proceso de limpieza para quedarse con las observaciones de las especies **"Adelie"** y **"Chinstrap"**.

```
library(palmerpenguins)
datos = as.data.frame(penguins)
#Quito valores faltantes
datos = datos[is.na(datos$bill_length_mm)==F,]
#Se extraen todos los pingüinos de la especie Gentoo debido a que los de la especie
#Chinstrap y Adelie presentan cierta homogeneidad entre si, mientras que los de Gentoo se
#Logran diferenciar muy bien de ellos.
datos = datos[datos$species!="Gentoo" ,]
```

```r
#Se cambian los nombres de las filas despues de remover a los pinguinos de la especie Gentoo
row.names(datos) = 1:nrow(datos)
#Extraigo unicamente las variables numericas continuas
datos = datos[,c(3,4,5,6)]
#Cambio el nombre de las columnas
names(datos) = c("L.Pico","P.Pico","Aleta","Masa")
kable(head(datos))
```

L.Pico	P.Pico	Aleta	Masa
39.1	18.7	181	3750
39.5	17.4	186	3800
40.3	18.0	195	3250
36.7	19.3	193	3450
39.3	20.6	190	3650
38.9	17.8	181	3625

Descripcion de la matriz de datos

```r
#Reviso la dimension de la matriz procesada
dim(datos)
```

```
## [1] 219   4
```

```r
names(datos)
```

```
## [1] "L.Pico" "P.Pico" "Aleta"  "Masa"
```

```r
str(datos)
```

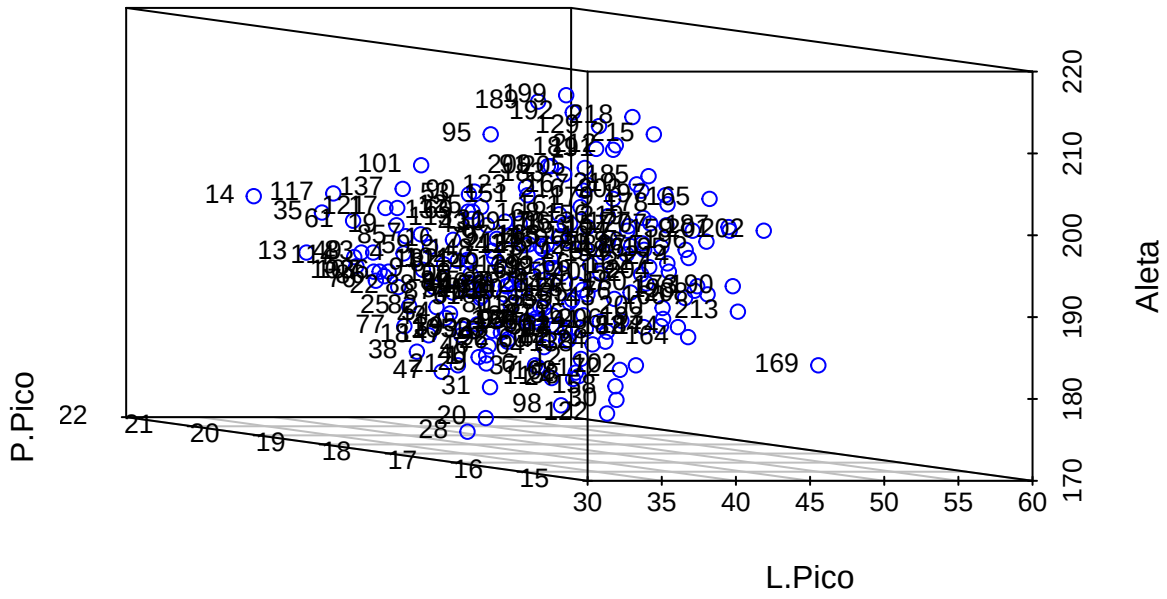
```
## 'data.frame':    219 obs. of  4 variables:
##  $ L.Pico: num  39.1 39.5 40.3 36.7 39.3 38.9 39.2 34.1 42 37.8 ...
##  $ P.Pico: num  18.7 17.4 18 19.3 20.6 17.8 19.6 18.1 20.2 17.1 ...
##  $ Aleta : int  181 186 195 193 190 181 195 193 190 186 ...
##  $ Masa  : int  3750 3800 3250 3450 3650 3625 4675 3475 4250 3300 ...
```

La matriz de datos esta compuesta por 219 observaciones y 4 variables numéricas de las cuales 2 son continuas y 2 son discretas.

Acercamiento Exploratorio en 3 dimensiones

Se hace un gráfico de dispersión en 3d para observar el comportamiento de los sujetos:

```
library(scatterplot3d)
zz <- scatterplot3d(x = datos[,1], y = datos[,2], z = datos[,3],
                    xlab = "L.Pico", ylab = "P.Pico", zlab = "Aleta",
                    pch = 1, color = "blue", grid =TRUE,
                    angle = 170)
zz.coords <- zz$xyz.convert(datos[,1], datos[,2], datos[,3])
text(zz.coords$x,
     zz.coords$y,
     labels = row.names(datos),
     cex = .8,
     pos = 2)
```



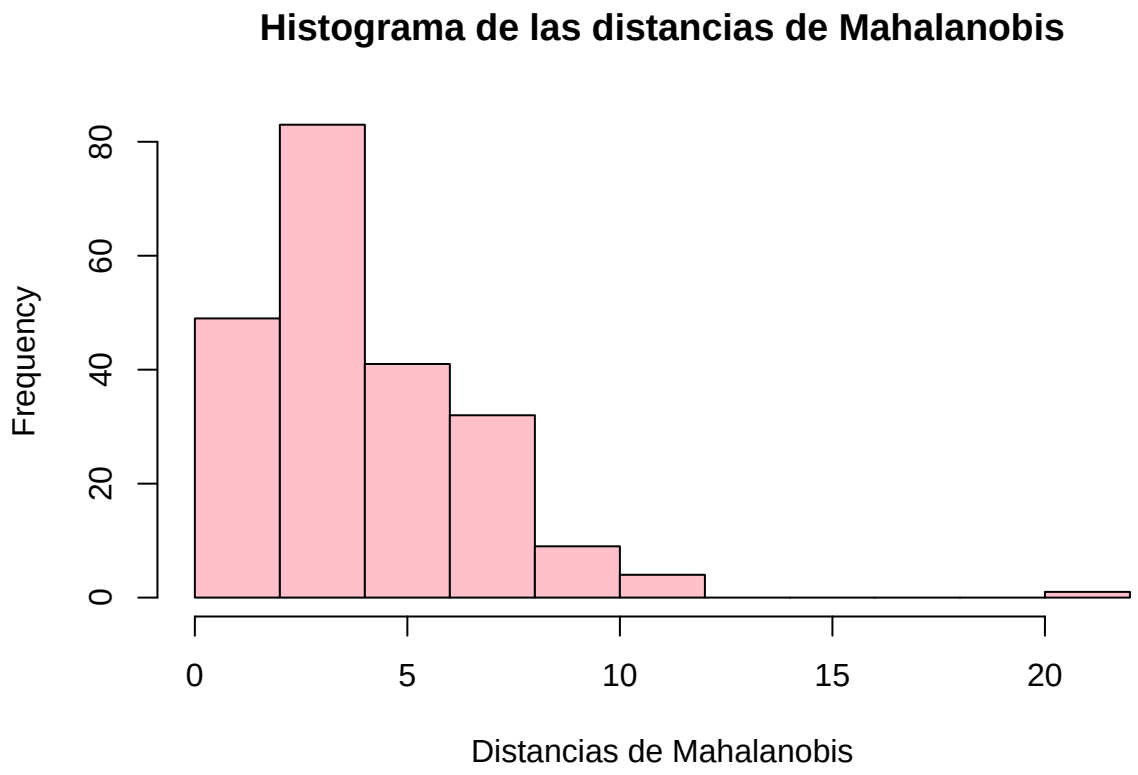
Se alcanza a identificar en el gráfico de dispersion 3D que existen valores muy alejados del resto como el 14, el 28 y el 169

Obtención de las distancias de mahalanobis

```
datosmahal = mahalanobis(datos,center = colMeans(datos), cov = cov(datos))
```

Observo la distribución de las distancias mediante un histograma

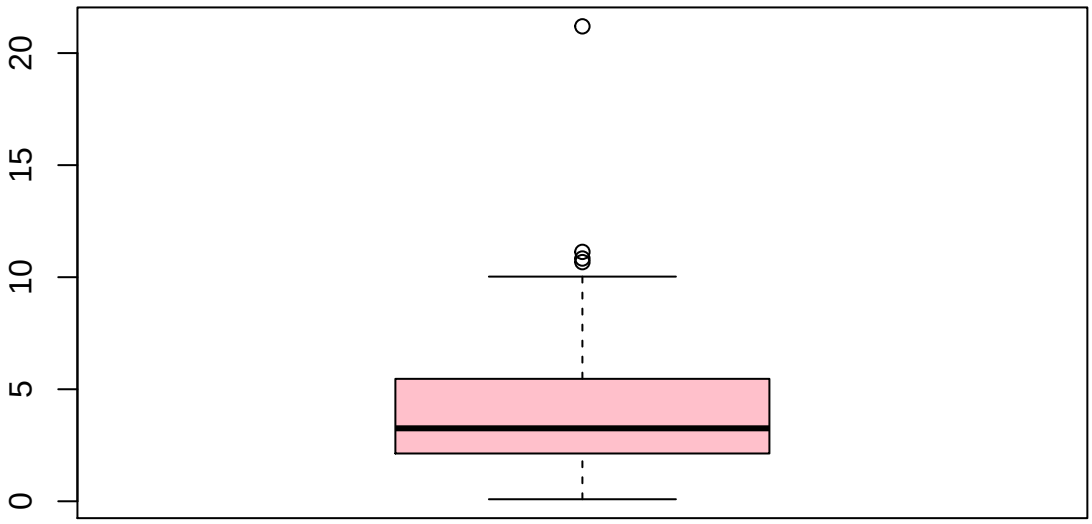
```
hist(datosmahal,col = "Pink",main = "Histograma de las distancias de Mahalanobis",
     xlab = "Distancias de Mahalanobis")
```



El histograma revela la presencia de valores muy alejado del resto.

Gráfico de cajas para detectar valores atipicos

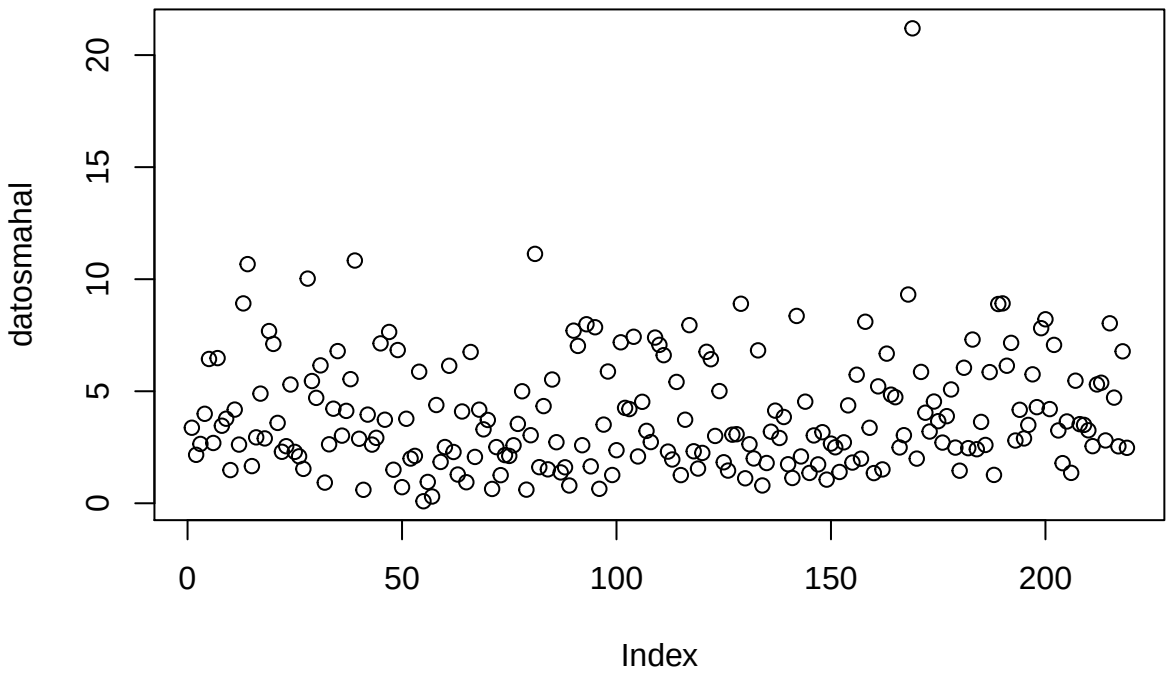
```
boxplot(datosmahal,col = "pink")
```



Se encuentra que existen 4 observaciones que se alejan mucho de las demás.

Gráfico de las distancias de Mahalanobis

```
plot(datosmahal)
```



Realizando un gráfico de dispersión de la variable se encuentra 1 dato muy alejado del resto y 3 mas que se alejan del resto.

Deteccion de los 4 valores mas alejados

```
dmsa = order(datosmahal,decreasing = T)[1:4]
dmsa
```

```
## [1] 169  81  39  14
```

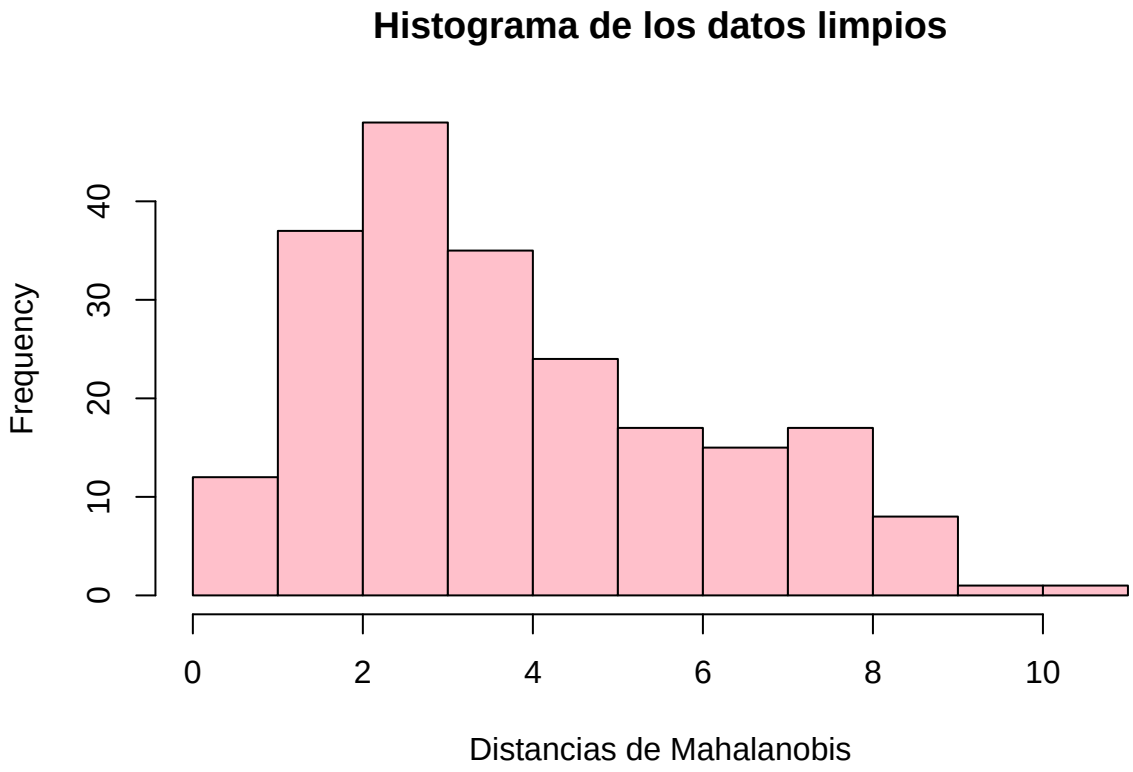
Se extraen los valores 169,81,39 y 14

```
limpios = datosmahal[-dmsa]
```

Una vez quitadas las observaciones que aparecían como valores muy alejados, se procede a observar el comportamiento de los gráficos.

Histograma de la nueva base

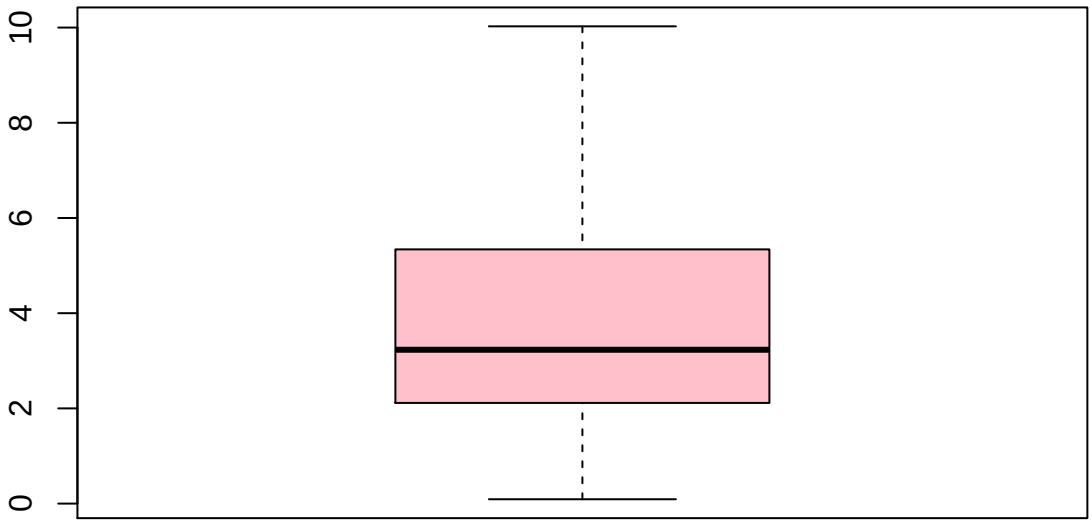
```
hist(limpios,col = "Pink", main = "Histograma de los datos limpios",xlab =  "Distancias de Mahalanobis")
```



El histograma ya no presenta valores muy alejados.

Gráfico de cajas de la nueva base

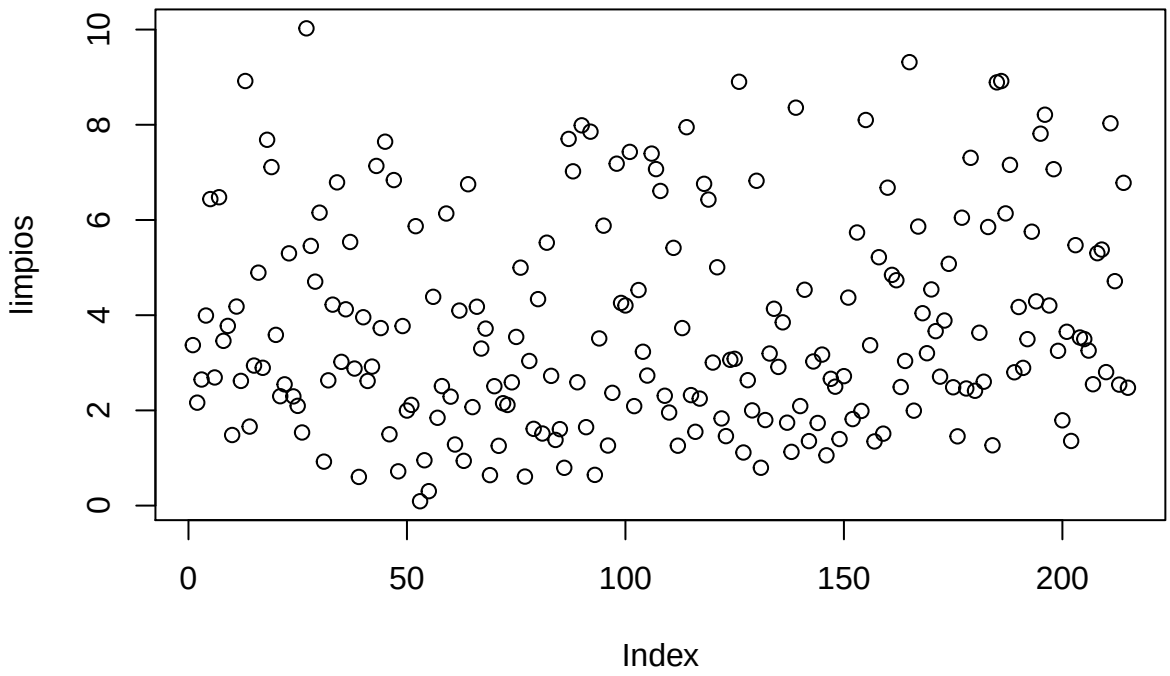
```
boxplot(limpios, col = "pink")
```



El gráfico de cajas ya no reporta ningún valor fuera de sus limites.

Revisamos el gráfico de dispersion de las distancias de Mahalanobis

```
plot(limpios)
```



Se puede observar que los datos ya tienen un comportamiento mas homogéneo.

Conclusion

Al gráfica los datos para poder observar su comportamiento nos topamos con la limitante de que por mucho podemos graficar 3 variables simultáneamente y suele suceder (como en este caso) que podemos llegar a tener 4 o mas variables que resultan de interés para nosotros y que no podemos alcanzar a representar al mismo tiempo por lo que la distancia de Mahalanobis resulta una opción viable para condensar esa información multivaridada en un solo valor que resulta mas manejable como se pudo observar en el transcurso del ejemplo con la base de pinguinos en donde se pudo pasar de tener 3 dimensiones en el espacio para representar su comportamiento a tener únicamente un solo valor que resumiera la información multivariante de los sujetos.