Los peces y el mercurio (Momento de Retroalimentación: Módulo 5 Procesamiento de datos multivariados. Portafolio Implementación)

Andres Piñones Besnier - A01570150

2022-12-1

LINK AL DRIVE: https://drive.google.com/drive/folders/1g5PrEp-furYdtaoqstIQTF9tVkJo5vtL?usp=sharing

EL PROBLEMA

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio. Las variables que se midieron se encuentran en mercurio.csvDescargar mercurio.csv y su descripción es la siguiente:

Para este reporte se realizará un análisis de datos multivariados haciendo uso de pruebas de normalidad, análisis de componentes principales matrices y vectores aleatorios.

- X1 = número de identificación
- X2 = nombre del lago
- X3 = alcalinidad (mg/l de carbonato de calcio)
- X4 = PH
- X5 = calcio (mg/l)
- X6 = clorofila (mg/l)
- X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago
- X8 = número de peces estudiados en el lago
- X9 = mínimo de la concentración de mercurio en cada grupo de peces
- X10 = máximo de la concentración de mercurio en cada grupo de peces
- X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
- X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Análisis de normalidad de las variables continuas para identificar variables normales.

Realizamos la prueba de normalidad de Mardia y la prueba de Anderson Darling para identificar las variables que son normales y detectar posible normalidad multivariada de grupos de variables.

```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 502.667343452414  3.6277693977554e-24     NO
## 2 Mardia Kurtosis 4.83254138772002 1.34801075923896e-06     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##                Test      Variable Statistic   p value Normality
```

```
## 1  Anderson-Darling  alcalinidad     3.6725  <0.001      NO
## 2  Anderson-Darling       ph         0.3496  0.4611      YES
## 3  Anderson-Darling    calcio        4.0510  <0.001      NO
## 4  Anderson-Darling   clorofila      5.4286  <0.001      NO
## 5  Anderson-Darling mediaMercurio    0.9253  0.0174      NO
## 6  Anderson-Darling   numPeces       8.6943  <0.001      NO
## 7  Anderson-Darling  minMercurio     1.9770  <0.001      NO
## 8  Anderson-Darling  maxMercurio     0.6585   0.081      YES
## 9  Anderson-Darling  estimacion      1.0469  0.0086      NO
## 10 Anderson-Darling    madurez      14.3350  <0.001      NO
##
## $Descriptives
##                 n       Mean    Std.Dev Median  Min    Max  25th  75th
## alcalinidad    53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50
## ph             53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40
## calcio         53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60
## clorofila      53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70
## mediaMercurio  53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77
## numPeces       53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00
## minMercurio    53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33
## maxMercurio    53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33
## estimacion     53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70
## madurez        53  0.8113208  0.3949977   1.00 0.00   1.00  1.00  1.00
##                      Skew   Kurtosis
## alcalinidad     0.9679170 -0.4705349
## ph             -0.2458771 -0.6239638
## calcio          1.3045868  0.6130359
## clorofila       2.4130571  6.1042185
## mediaMercurio   0.5986343 -0.6312607
## numPeces        2.5808773  6.0089455
## minMercurio     1.0729099  0.4060828
## maxMercurio     0.4645925 -0.6692490
## estimacion      0.9449951  0.5733500
## madurez        -1.5465748  0.4005116
```

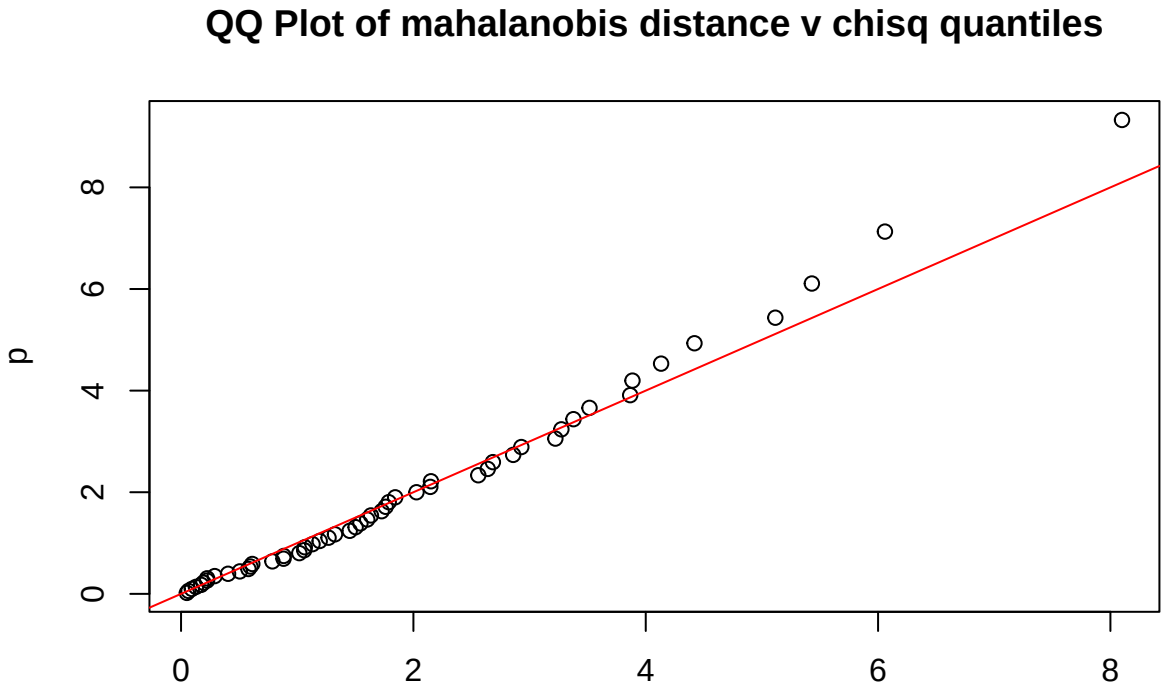
Ahora realizamos la prueba de Mardia y Anderson Darling de las variables que sí tuvieron normalidad en los incisos anteriores. En este caso las variables que presentaron normalidad fueron ph y maxMercurio.

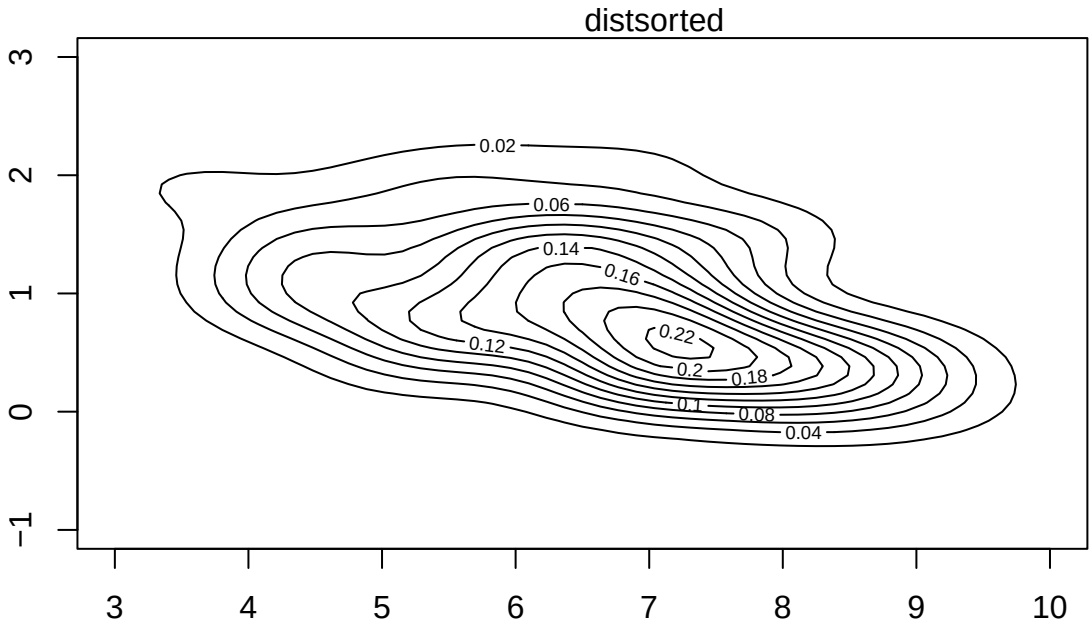
```
## $multivariateNormality
##              Test          Statistic           p value Result
## 1 Mardia Skewness   6.53855430534145 0.162377302354508    YES
## 2 Mardia Kurtosis -0.889321233851276 0.373830462900113    YES
## 3             MVN               <NA>              <NA>    YES
##
## $univariateNormality
##               Test    Variable Statistic   p value Normality
## 1 Anderson-Darling     ph         0.3496    0.4611    YES
## 2 Anderson-Darling maxMercurio    0.6585    0.0810    YES
##
## $Descriptives
##              n      Mean   Std.Dev Median  Min  Max 25th 75th       Skew
## ph          53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -0.2458771
## maxMercurio 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925
##               Kurtosis
## ph          -0.6239638
```

```
## maxMercurio -0.6692490
```

Respecto a los resultados obtenidos en Mardia Skewness y Mardia Kurtosis tenemos un valor de p mayor a 0.05 que usamos como nivel de significancia, por lo tanto no se rechaza H_0 y se puede asumir los datos tienen normal multivariada.

Buscaremos datos atípicos o influyentes en la normal multivariada encontrada en el inciso B por medio de la distancia de Mahalanobis y del gráfico QQplot multivariado.





Como se puede ver en el gráfico existen algunos valores atípicos. Volveremos a realizar la prueba pero ahora sin datos atípicos.

```
## $multivariateNormality
```

```
##              Test          Statistic           p value Result
## 1 Mardia Skewness   6.53855430534145 0.162377302354508    YES
## 2 Mardia Kurtosis -0.889321233851276 0.373830462900113    YES
## 3             MVN               <NA>              <NA>    YES
##
## $univariateNormality
##               Test    Variable Statistic   p value Normality
## 1 Anderson-Darling     ph         0.3496    0.4611    YES
## 2 Anderson-Darling maxMercurio    0.6585    0.0810    YES
##
## $Descriptives
##               n      Mean   Std.Dev Median  Min  Max 25th 75th       Skew
## ph           53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -0.2458771
## maxMercurio 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925
##                Kurtosis
## ph           -0.6239638
## maxMercurio -0.6692490
```

Análisis de componentes principales

Realizamos el análisis de componentes principales con la base de datos completa para identificar los factores principales que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce.

El analisis de componentes principales simplifica la complejidad de datos de varias dimensiones manteniendo tendencias y patrones. Con el análisis transformamos los datos y reducimos la dimensión lo que permite sumarizar las variables. En este caso nos permite identificar que variabes estan relacionadas y analizar de mejor manera su comportamiento y de esta forma acercarnos más a identificar que variables intervienen en la contaminación por mercurio.

Realizamos el análisis de componentes principales

```
##  [1] 0.5361226 0.6615488 0.7832169 0.8741602 0.9333019 0.9636166 0.9842903
##  [8] 0.9929724 0.9981363 1.0000000
```

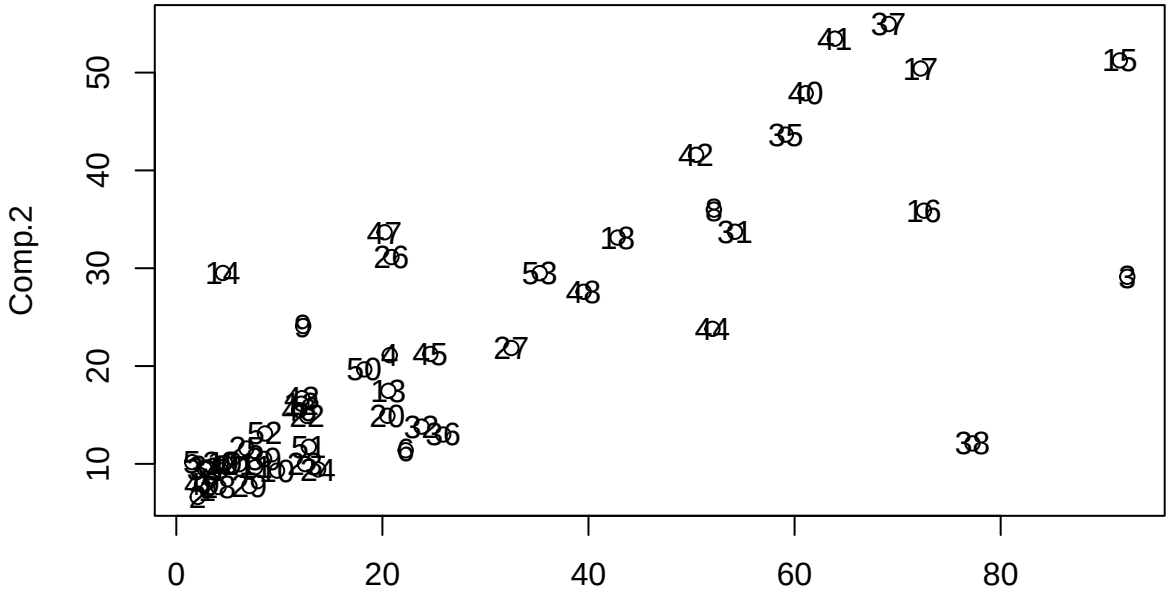
De acuerdo a los resultados se puede optar por emplear los primeros 5 componentes ya que con ellos es posible explicar un poco más del 92% de la varianza observada. De ahí en adelante los demás tienen diferencia no muy significativa.

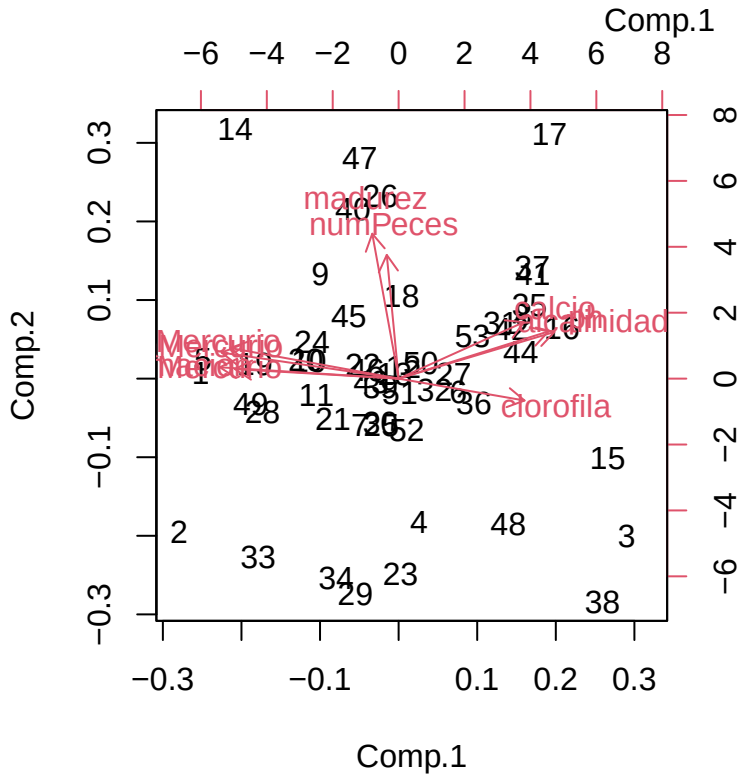
Gráfico de los vectores asociados a las variables y puntuaciones de las observaciones de las dos primeras componentes

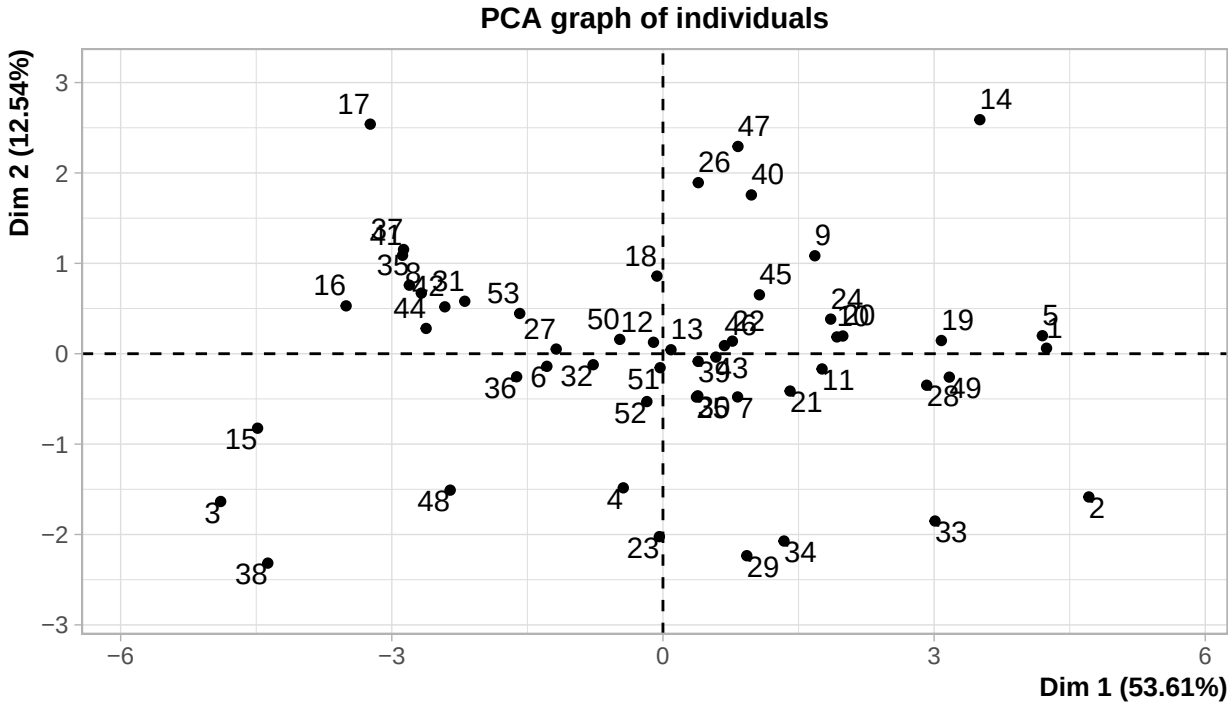
```
## Loading required package: ggplot2
```

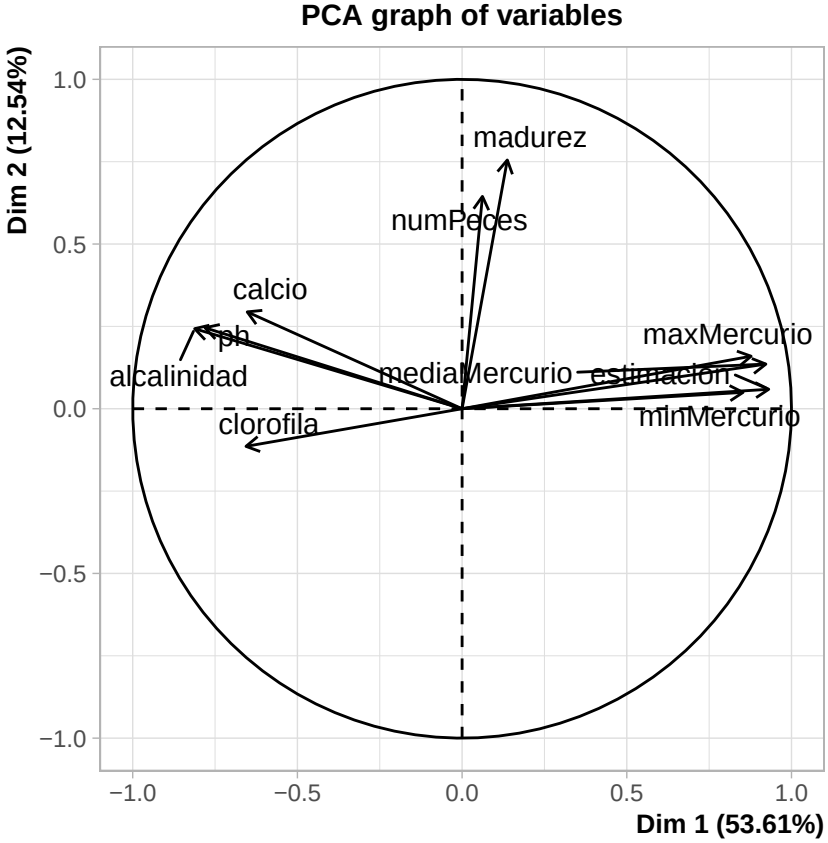
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

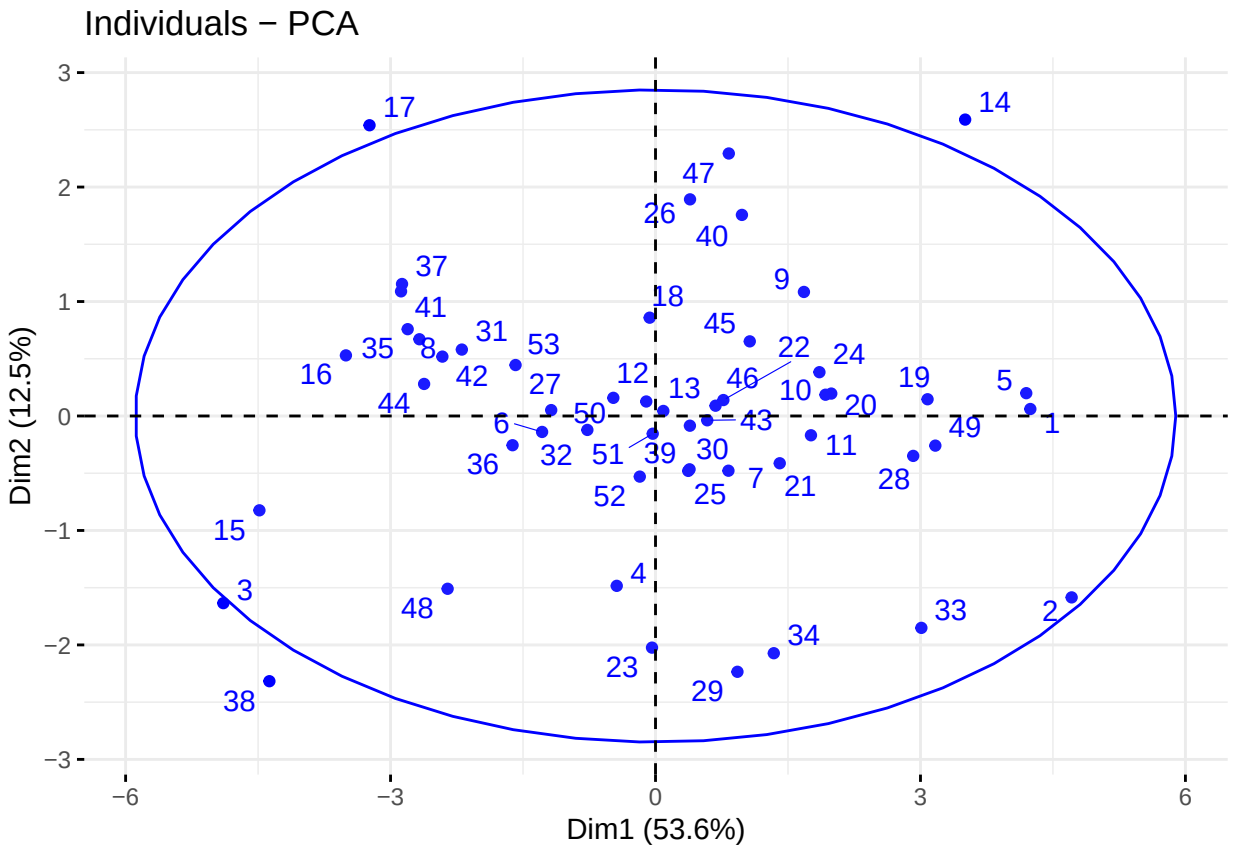
Mat de Correlación

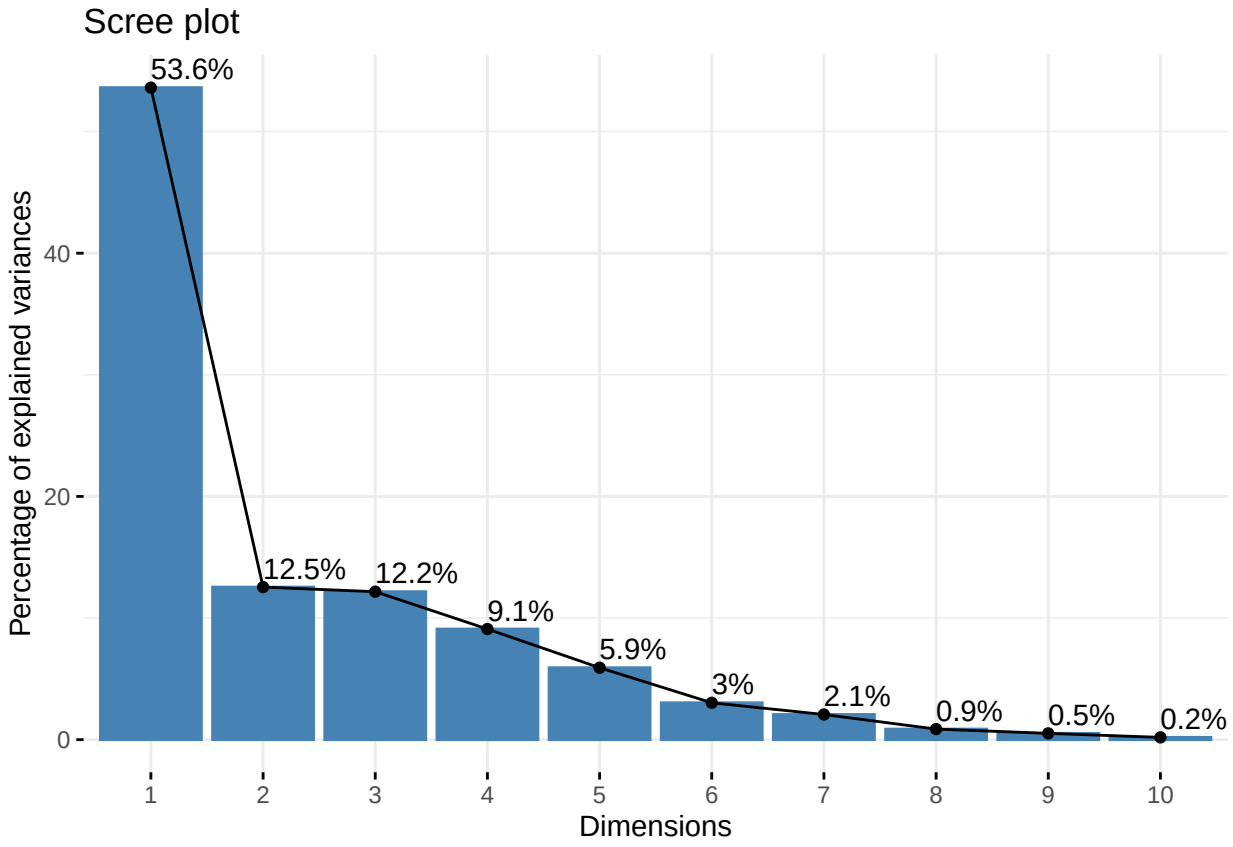


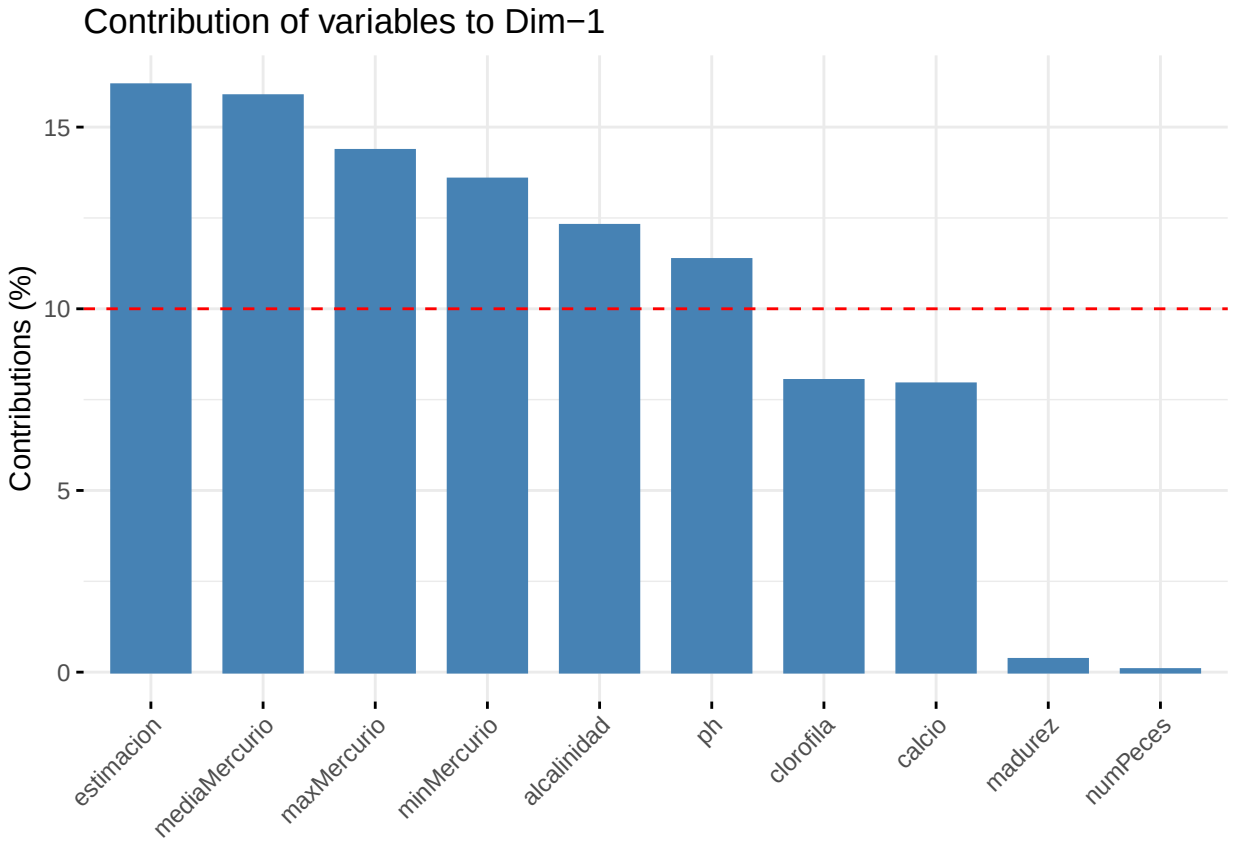



PCA graph of individuals
Dim 2 (12.54%)
Dim 1 (53.61%)
3
2
1
0
−1
−2
−3
−6
−3
0
3
6
17
14
47
26
40
9
18
45
16
53
24
44
27
50
12
13
19
5
36
6
32
51
39
11
28
49
52
21
15
48
4
2
3
33
23
34
29
38


PCA graph of variables
Dim 2 (12.54%)
Dim 1 (53.61%)
1.0
0.5
0.0
−0.5
−1.0
madurez
numPeces
calcio
ph
alcalinidad
mediaMercurio
maxMercurio
estimacion
minMercurio
clorofila


Individuals − PCA
Dim2 (12.5%)
Dim1 (53.6%)
3
2
1
0
−1
−2
−3
−6
−3
0
3
6
17
14
47
26
40
37
41
9
18
45
22
24
31
53
35
8
42
16
27
12
13
46
10
19
5
44
50
20
6
43
49
1
36
32
51
39
30
11
28
52
25
7
21
15
4
3
48
33
2
34
23
29
38


Scree plot
Percentage of explained variances
Dimensions
40
20
0
1
2
3
4
5
6
7
8
9
10
53.6%
12.5%
12.2%
9.1%
5.9%
3%
2.1%
0.9%
0.5%
0.2%



Resultados

En el scree plot es posible visualizar que utilizar hasta el 5to componente principal como se mencionó anteriormente nos permite explicar aproximadamente el 93.3% de la varianza observada. Sin embargo el primer componente tiene una gran diferencia de peso comparado con los demás.

En el otro gráfico circular de las variables podemos interpretar la correlación entre las variables. Variables correlacionadas positivamente estan agrupadas. Variables con correlación negativa se posicionan en cuadrantes opuestos. Las distancias entre las variables y el origen miden la calidad de la variable en el mapa de factores. Las más alejadas del origen estan mejor representadas.

Emite una conclusión general:

Por medio de estas nuevas herramientas de análisis pudimos hacer un acercamiento con otra metodología para intentar nuevamente encontrar las relaciones entre las variables de esta base de datos y determinar los factores que afectan el nivel de mercurio en el agua. Se pudo robustecer el análisis realizado anteriormente y llegar a conclusiones similares. Este análisis y sus resultados combinado con lo obtenido en el anterior nos permite confirmar que hay una relación entre el ph, alcalinidad y calcio con el nivel de mercurio. Sin embargo por medio del analisis de normalidad multivariada pudimos observar que existe una relación entre el ph y el mercurio, relacionada con la alcalinidad.