# Reporte final: "Los peces y el mercurio"

Ariadna Jocelyn Guzmán Jiménez - A01749373

2022-12-03

Módulo 1: Estadística para ciencia de datos

Inteligencia artificial avanzada para la ciencia de datos II

Grupo 501

## Resumen

Mediante este trabajo, se presenta la implementación y el análisis de normalidad y componentes para la base de contaminación por mercurio en lagos. Esta problemática, es muy importante ya que además de afectar a los seres vivos del lago, puede afectar a los seres humanos si llegan a consumir alguno de ellos. Por eso, fue necesario poder realizar un entendimiento de todos los datos para poder enfatizar y lograr a interpretar cuales son las variables candidatas para los modelos. Inicialmente, en el módulo anterior, se había llegado a la conclusión de que **alcalinidad, calcio y clorofila** eran las variables que más afluencia tenían en esta problemática, por lo que con los análisis siguientes, verificaremos los resultados de la anterior entrega.

## Introducción

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.

En nuestra base de datos, encontramos los siguiente atributos:

- **X1** = número de indentificación
- **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 muscualar 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)

Dado el anterior análisis de los datos, vemos que las variables $X1, X2$ y $X12$ son variables de clasificación, por lo que para nuestros análisis, no haremos uso de ellas.

Por otra parte, con las descripciones de cada variable nos surgen interesantes las siguientes preguntas para poder resolver y poder ir sobre ellas para hacer hacer predicciones futuras, las cuales son:

- *¿Habrá direfencia significativa entre la concentración de mercurio por la edad de los peces?*
- *¿Hay evidencia para suponer que la concentración promedio de mercurio en los lagos es dañino para la salud humana?*
- *¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?*

En las siguientes líneas, se verá la implementación de modelos para la resolución de las preguntas anteriores y llegar a una conclusión concreta de esta problemática.

# Análisis de los resultados

En la parte de la lectura de datos, importamos nuestra base y por otra parte, hacemos una nueva variable que solo cuente con los datos númericos, ya que son los que nos servirán para nuestros análisis.

## Análisis de normalidad

Realice un análisis de normalidad de las variables continuas para identificar variables normales.

### Prueba de normalidad de Mardia y prueba de Anderson Darling

```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 410.214790601478 7.04198777815398e-23     NO
## 2 Mardia Kurtosis 4.59612555772731 4.30419392238868e-06     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    X3        3.6725  <0.001      NO
## 2 Anderson-Darling    X4        0.3496  0.4611      YES
## 3 Anderson-Darling    X5        4.0510  <0.001      NO
## 4 Anderson-Darling    X6        5.4286  <0.001      NO
## 5 Anderson-Darling    X7        0.9253  0.0174      NO
## 6 Anderson-Darling    X8        8.6943  <0.001      NO
## 7 Anderson-Darling    X9        1.9770  <0.001      NO
## 8 Anderson-Darling    X10       0.6585   0.081      YES
## 9 Anderson-Darling    X11       1.0469  0.0086      NO
##
## $Descriptives
##      n       Mean    Std.Dev Median  Min    Max  25th  75th       Skew
## X3  53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50  0.9679170
## X4  53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40 -0.2458771
## X5  53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60  1.3045868
## X6  53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70  2.4130571
## X7  53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77  0.5986343
```

```
## X8  53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00  2.5808773
## X9  53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33  1.0729099
## X10 53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33  0.4645925
## X11 53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70  0.9449951
##       Kurtosis
## X3  -0.4705349
## X4  -0.6239638
## X5   0.6130359
## X6   6.1042185
## X7  -0.6312607
## X8   6.0089455
## X9   0.4060828
## X10 -0.6692490
## X11  0.5733500
```

De acuerdo a los resultados anteriores de la prueba de normalidad multivariada, podemos observar que las únicas variables que cuentan con normalidad de nuestro conjunto de datos son $X4$ y $X10$, por lo que volveremos a realizar las pruebas para observar su comportamiento. Por otro lado, si analizamos la variabilidad descriptiva entre estas dos variables con su cociente entre su desviación estándar y su media, podemos ver que $X4$ cuenta con una menor variabilidad a comparación de $X10$, lo que verifica que su media es más confiable.

**Prueba de Mardia y Anderson Darling de las variables que sí tuvieron normalidad**

```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  6.17538668676458 0.186427564928852    YES
## 2 Mardia Kurtosis -1.12820795824432  0.25923210375991    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    x4        0.3496    0.4611    YES
## 2 Anderson-Darling    x10       0.6585    0.0810    YES
##
## $Descriptives
##      n      Mean   Std.Dev Median  Min  Max 25th 75th       Skew   Kurtosis
## x4  53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -0.2458771 -0.6239638
## x10 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925 -0.6692490
```

Ambas pruebas nos indican que las muestras proporcionadas provienen de una distribución normal.

Teniendo en cuenta las hipótesis:

$H_0$ : Las variables aleatorias en un estudio siguen una distribución normal.

$H_1$ : Las variables aleatorias en un estudio no siguen una distribución normal.

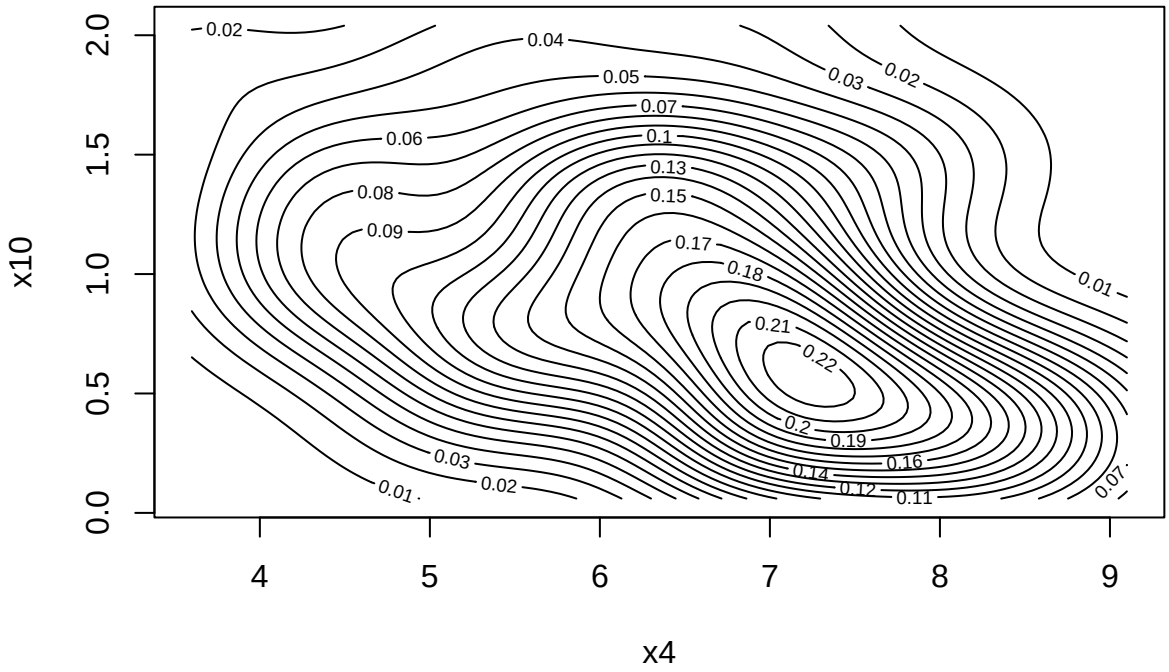
Y tomando en cuenta un valor de significancia de 0.05.

Podemos ver que los valores $p$ de la prueba de Anderson-Darling y Mardia, son mayores al valor de significancia, por lo que no podemos rechazar la hipótesis nula y de esta forma tenemos la evidencia suficiente de que los datos **si siguen una distribución normal.**

En el caso del sesgo, podemos observar que se cuenta con un valor de 0.18, indicándonos que la distribución es moderadamente simétrica con respecto a su media y mediana.

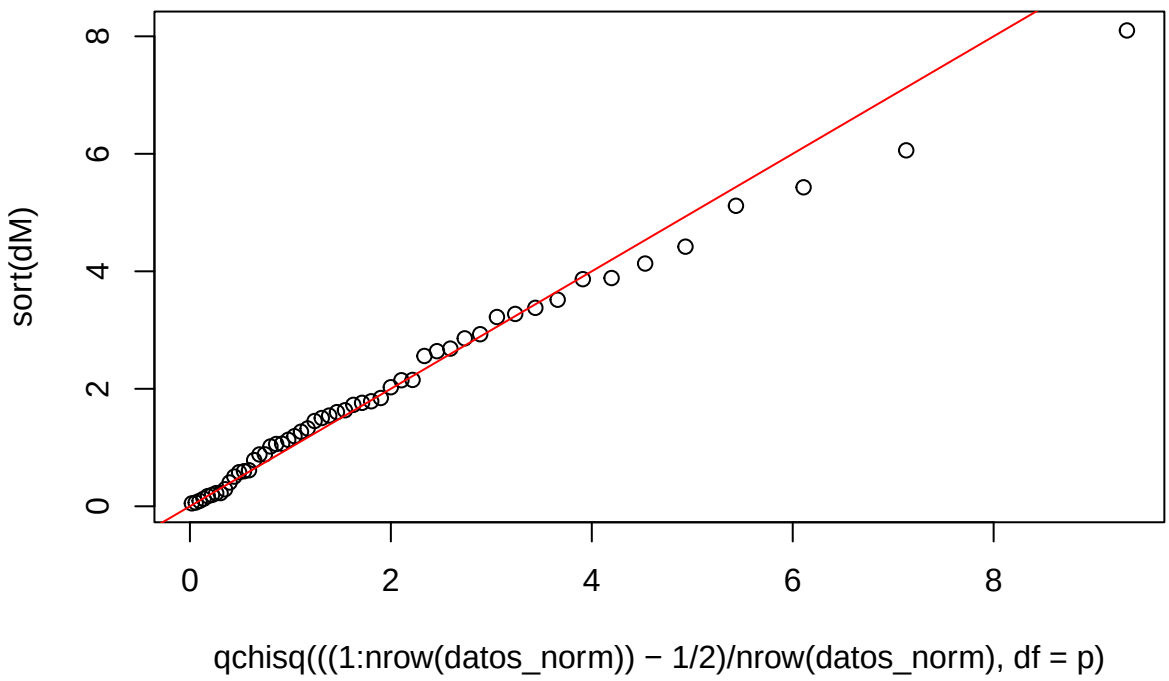
Por otro lado, para la curtosis se tiene un valor de 0.25, mostrando que la distribución

**Gráfica de contorno de la normal multivariada obtenida en el inciso B.**




```
## $multivariateNormality
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.7695729 0.1024763 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    x4        0.3496    0.4611    YES
## 2 Anderson-Darling    x10       0.6585    0.0810    YES
##
## $Descriptives
##      n      Mean   Std.Dev Median  Min  Max 25th 75th       Skew   Kurtosis
## x4  53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -0.2458771 -0.6239638
## x10 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925 -0.6692490
```

**Detección de datos atípicos o influyentes en la normal multivariada**




De acuerdo con la gráfica QQplot en base a la distancia de Mahalanobis entre los datos, podemos ver que estos siguen una asimetría negativa, con sesgo a la izquierda.

## Análisis de componentes principales

Realice un 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.

**¿Por qué es adecuado el uso de componentes principales para analizar la base?**

```
##             X3          X4          X5          X6          X7          X8
## X3   1.00000000  0.71916568  0.83260419  0.47753085 -0.59389671  0.01029074
## X4   0.71916568  1.00000000  0.57713272  0.60848276 -0.57540012 -0.01860607
## X5   0.83260419  0.57713272  1.00000000  0.40991385 -0.40067958 -0.08937901
## X6   0.47753085  0.60848276  0.40991385  1.00000000 -0.49137481 -0.01182027
## X7  -0.59389671 -0.57540012 -0.40067958 -0.49137481  1.00000000  0.07903426
## X8   0.01029074 -0.01860607 -0.08937901 -0.01182027  0.07903426  1.00000000
## X9  -0.52535654 -0.54196524 -0.33247623 -0.40045856  0.92720506 -0.08165278
## X10 -0.60479558 -0.55181523 -0.40791663 -0.48497215  0.91586397  0.16109174
## X11 -0.62795845 -0.61284905 -0.46440947 -0.50644193  0.95921481  0.02580046
##              X9        X10         X11
## X3  -0.52535654 -0.6047956 -0.62795845
## X4  -0.54196524 -0.5518152 -0.61284905
```

```
## X5  -0.33247623 -0.4079166 -0.46440947
## X6  -0.40045856 -0.4849721 -0.50644193
## X7   0.92720506  0.9158640  0.95921481
## X8  -0.08165278  0.1610917  0.02580046
## X9   1.00000000  0.7653532  0.91908939
## X10  0.76535319  1.0000000  0.85975810
## X11  0.91908939  0.8597581  1.00000000
```

El uso de componentes principales se obtiene a través de un proceso de cálculo de raíces y vectores, con el objetivo de contener la mayoría de la varianza observada y evitar la obtención de información no útil, de esta manera, se logra reducir la dimensionalidad del conjunto de datos. Para lograr esto, las variables tienen que ser correlacionadas, por lo que es importante visualizar una matriz de correlación de nuestro conjunto de datos, ya que nos proporciona una matriz cuadrada de dimensión y simétrica, ayudándonos para realizar el proceso de manera adecuada.

**Análisis de componentes principales y justificación del número de componentes principales apropiados para reducir la dimensión de la base**

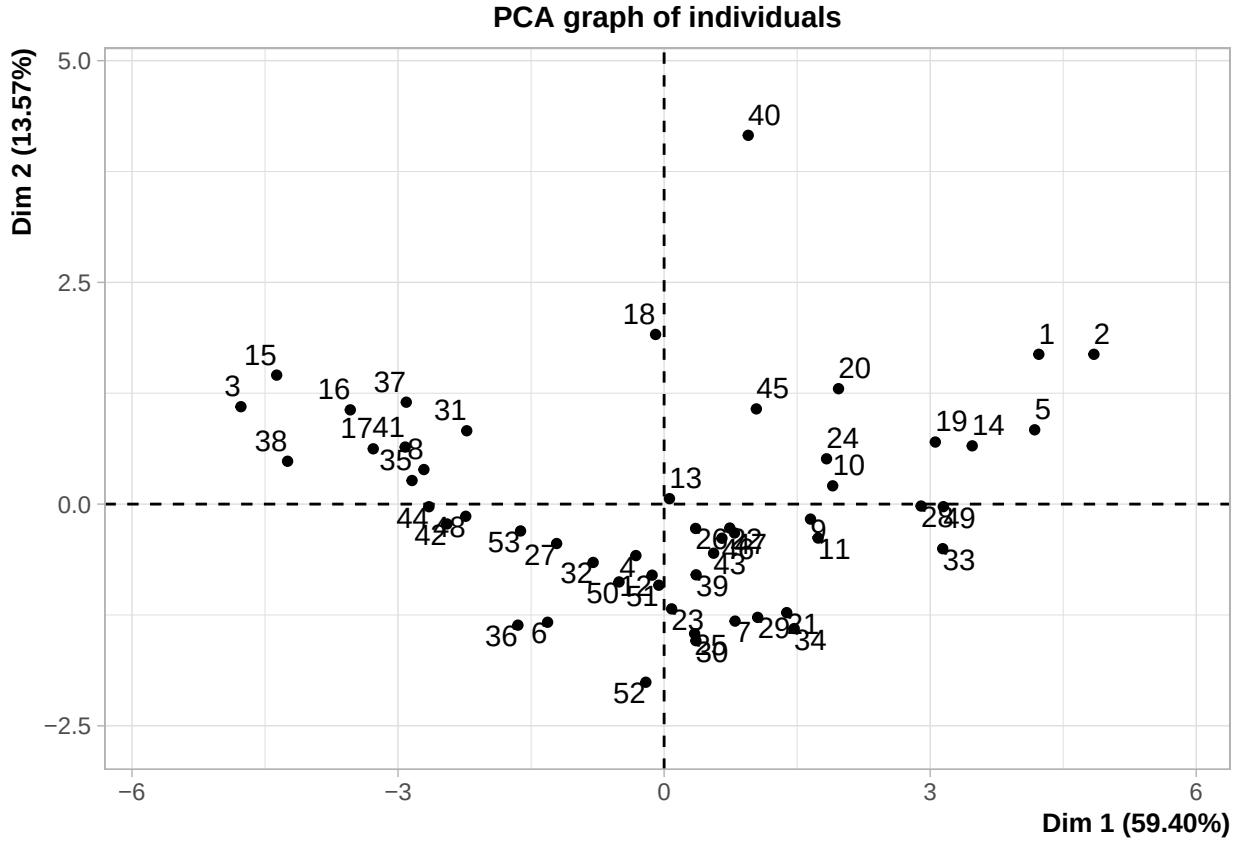
```
## Loading required package: ggplot2
```

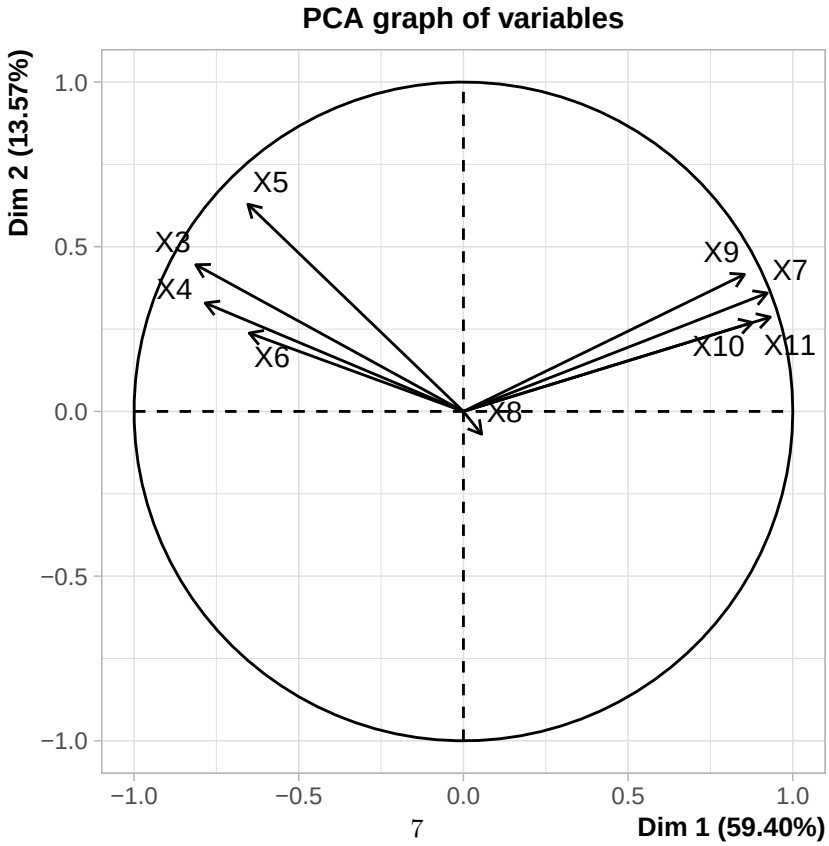
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

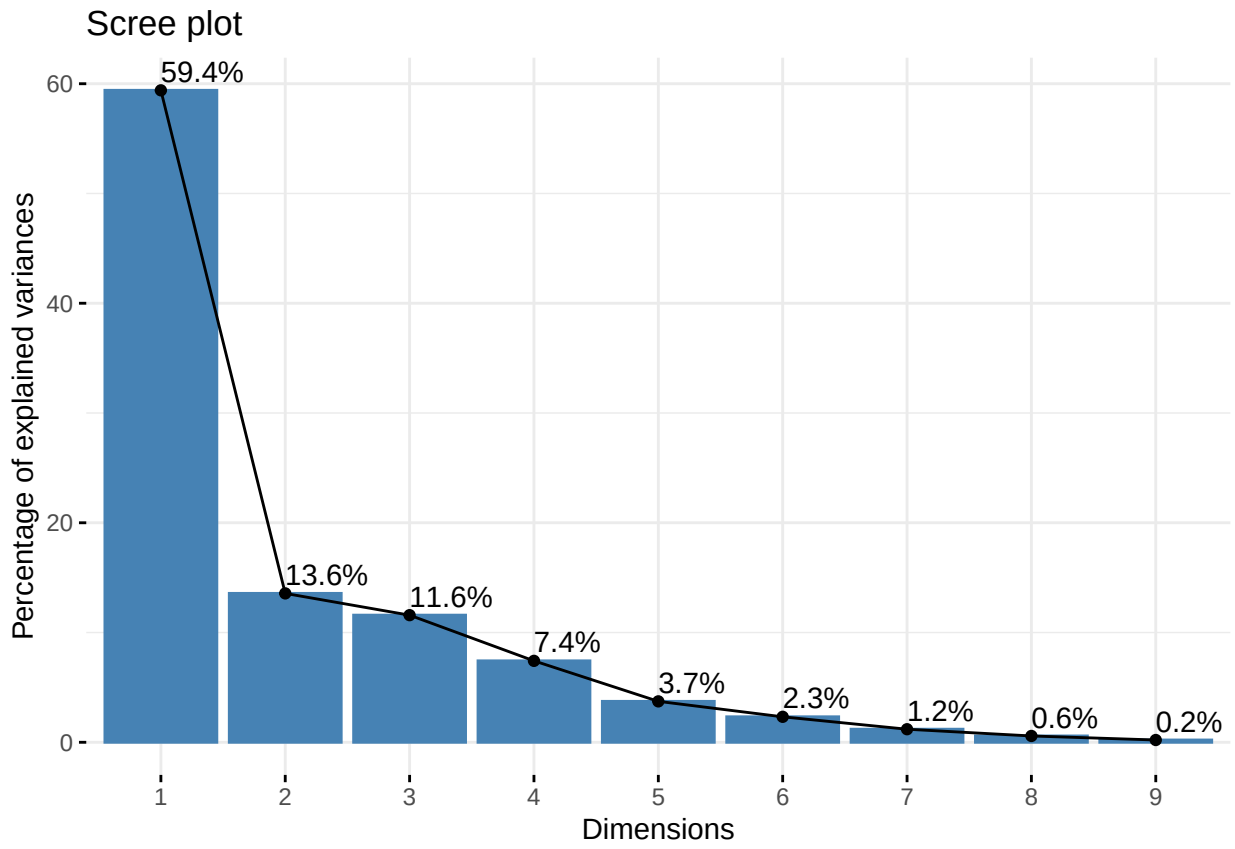
```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5     PC6     PC7
## Standard deviation     2.312 1.1049 1.0210 0.81723 0.5794 0.45710 0.32750
## Proportion of Variance 0.594 0.1357 0.1158 0.07421 0.0373 0.02322 0.01192
## Cumulative Proportion  0.594 0.7297 0.8455 0.91969 0.9570 0.98021 0.99212
##                           PC8     PC9
## Standard deviation     0.22810 0.13731
## Proportion of Variance 0.00578 0.00209
## Cumulative Proportion  0.99791 1.00000
```

De acuerdo con nuestro análisis, observamos que los componentes principales son aquellos no correlacionados con varianzas lo más grandes posibles. En este caso, para conocerlos aplicamos en nuestra fórmula una escala para poder utilizar la matriz de correlación de nuestro conjunto de datos. En este caso, se observa que el primer componente explica el 59.4% de la varianza de los datos, mientras que el segundo explica el 13.5%.

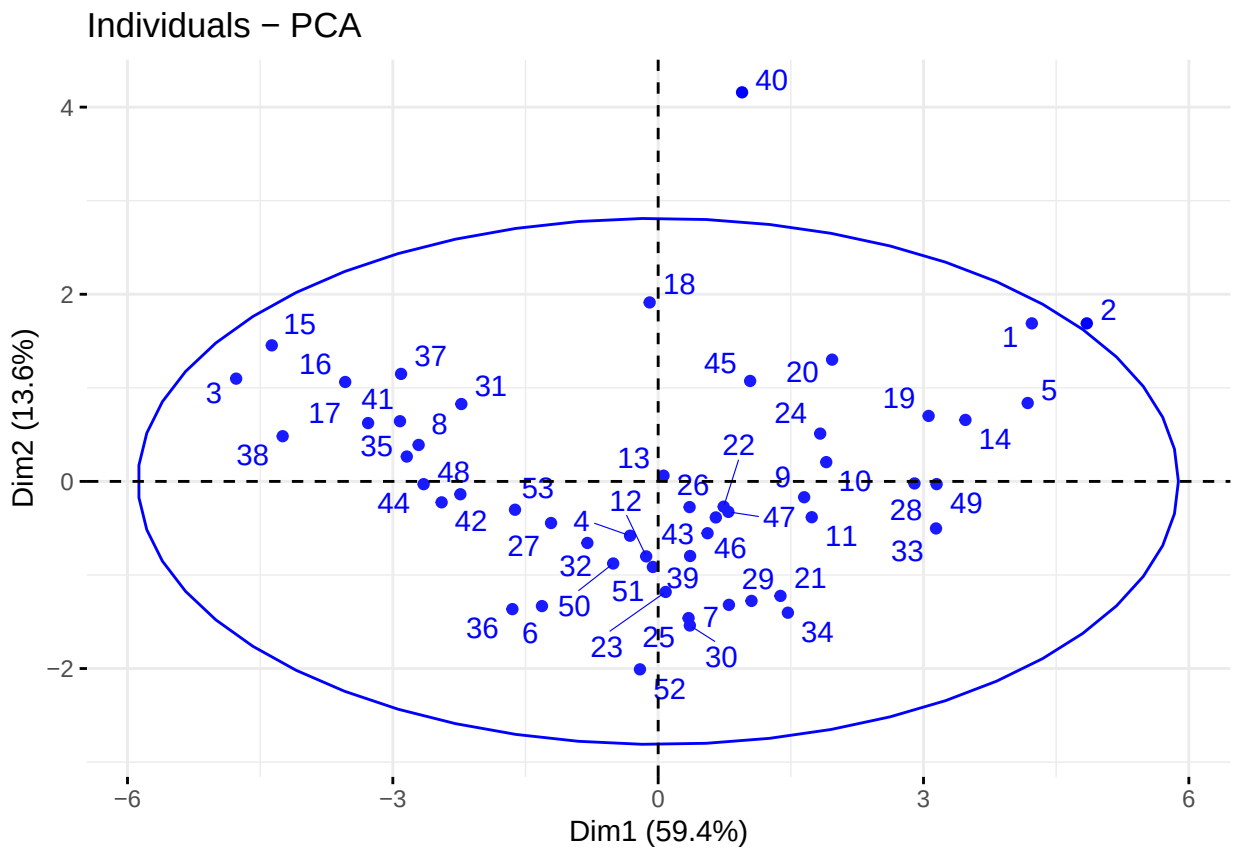
**Representación en gráficos de los vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes**

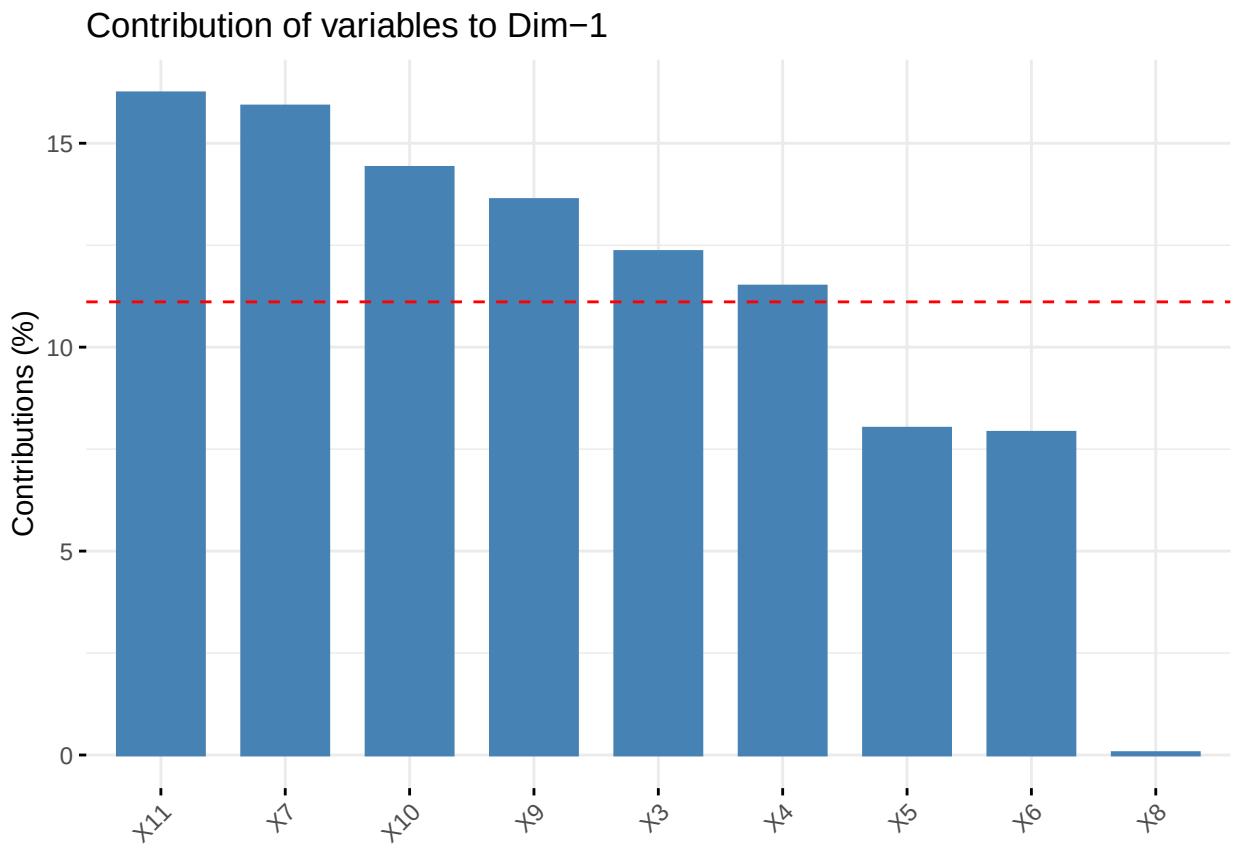





```
## Warning: Ignoring unknown parameters: level
```


Individuals − PCA
Dim2 (13.6%)
Dim1 (59.4%)
4
2
0
−2
−6
−3
0
3
6
40
18
15
16
37
3
41
31
17
8
38
35
48
44
42
53
27
4
32
51
50
36
6
12
13
26
22
9
10
47
43
46
11
39
29
21
7
34
23
25
30
52
45
20
24
19
14
1
2
5
28
49
33




**Interpretación de resultados. ¿A qué conclusiones se llega con el análisis y qué significado tienen los componentes seleccionados en el contexto del problema?**

En el gráfico de variables, podemos observar una interpretación de los 2 primeros componentes principales, donde el primero, como se mencionó anteriormente, aporta un 59.40%, mientras que el segundo un 13.57%. De misma forma, se visualiza que 5 de las variables se agrupan de manera positiva, mientras que 4 lo hacen de forma negativa.

Por otra parte, en screeplot se muestran los porcentajes de las explicaciones de variaciones totales de cada componente en el conjunto de datos, una representación gráfica de nuestro resultado del comando **prcomp**.

Mediante la gráfica de contribución de variables, podemos observar que las variables que más aporte realizan y que superan el valor medio de la contribución son:

- X11 - Estimación de la concentración de mercurio en el pez
- X7 - Concentración media de mercurio
- X10 - Máximo de la concentración de mercurio
- X9 - Mínimo de la concentración de mercurio
- X3 - Alcalinidad
- X4 - PH

# Conclusión

Gracias a el análisis de normalidad en los datos, se pudo conocer cuanto difería la distribución de los datos observados con respecto a lo esperado, esto gracias a las representaciones gráficas y test de hipótesis, como los fueron mardia y anderson-darling. De esta manera, pudimos comprobar que nuestro conjunto de datos, no tuviera una falta de normalidad y evitar la ineficienia en nuestros resultados al contar solo con datos aproximados y no totalmente exactos. Con ello,ya tenemos conciencia de que si existe una normalidad multivariada, que nos permitió dar paso al análisis de componentes principales.

Por otro lado, el análisis de componentes principales, nos permitió realizar una especie de minería de datos, ya que se logró extraer fácilmente información de nuestro conjunto de datos, de esta forma, se logró observar una proyección sobre los datos que son mejor representados en términos de mínimos cuadrados y poder realizar una predicción sobre la resolución de nuestro problema.

Gracias a dicho análisis se pudieron identificar los siguientes componentes como los más importantes:

- X11 - Estimación de la concentración de mercurio en el pez
- X7 - Concentración media de mercurio
- X10 - Máximo de la concentración de mercurio
- X9 - Mínimo de la concentración de mercurio
- X3 - Alcalinidad
- X4 - PH

Finalmente, de acuerdo con los análisis previamente realizados sobre este problema y los actuales, se logra apreciar que las variables clasificadas como las más importantes en contribución de componentes cpinciden con las que nos dieron un mejor funcionamiento para la implementación de la regresión múltiple. Sin embargo, con ayuda de estos últimos resultados, vemos que la **alcalinidad y el ph** son las que más afluencia tienen en la contaminación por mercurio, ya que son las variables que coincidieron en el análisis de esta entrega y la del módulo anterior. Esto, nos da una nueva perspectiva de entendimiento de los datos, donde podemos enfocarnos sobre dichas variables para obtener nuevos estudios y resultados para la resolución del problema.

# Anexos

- https://github.com/A01749373/Portafolio-de-Implementacion-A01749373