Actividad:

Reporte Final: Los peces y el mercurio pt 2

Modulo:

Modulo 5: Estadística e Inteligencia artificial avanzada para la ciencia de datos

Gpo 502

Nombre: Héctor Francisco Marin Garrido – A00827714

Maestra: Blanca R. Ruiz Hernández

Fecha de entrega: 2 de diciembre de 2022

Resumen

En los cuerpos acuíferos se presenta cada vez más comúnmente la problemática de la contaminación al punto que ya la población de peces de agua dulce esta contaminada, lo cual implica una amenaza a nuestra salud al ser consumidores de estos animales. Para esto se busca hacer un análisis estadístico acerca de las componentes que influyen a la contaminación de

esos lagos. Para esto se usarán distintos métodos estadísticos para determinar los componentes principales que contribuyen a esta catástrofe.

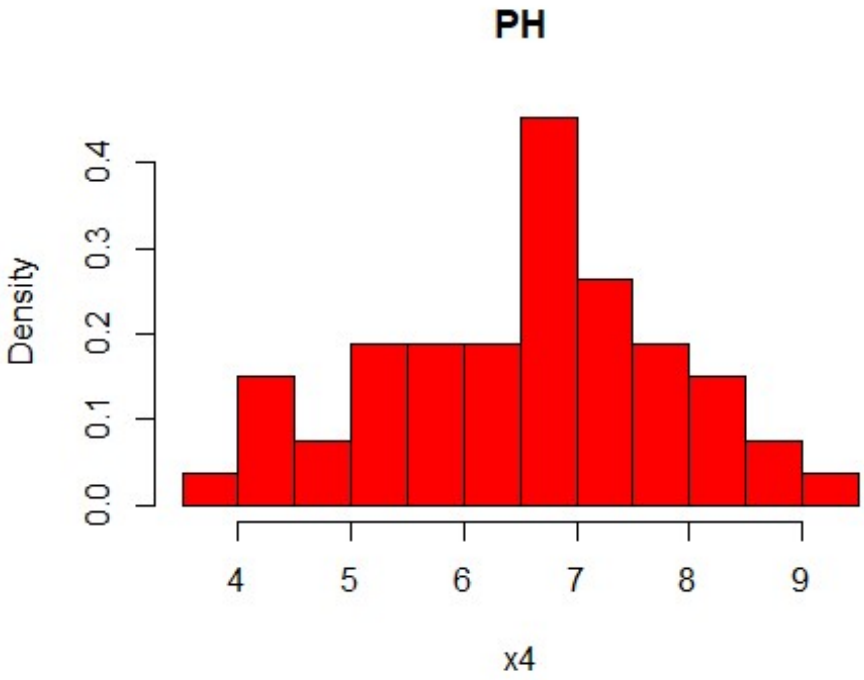
Introducción

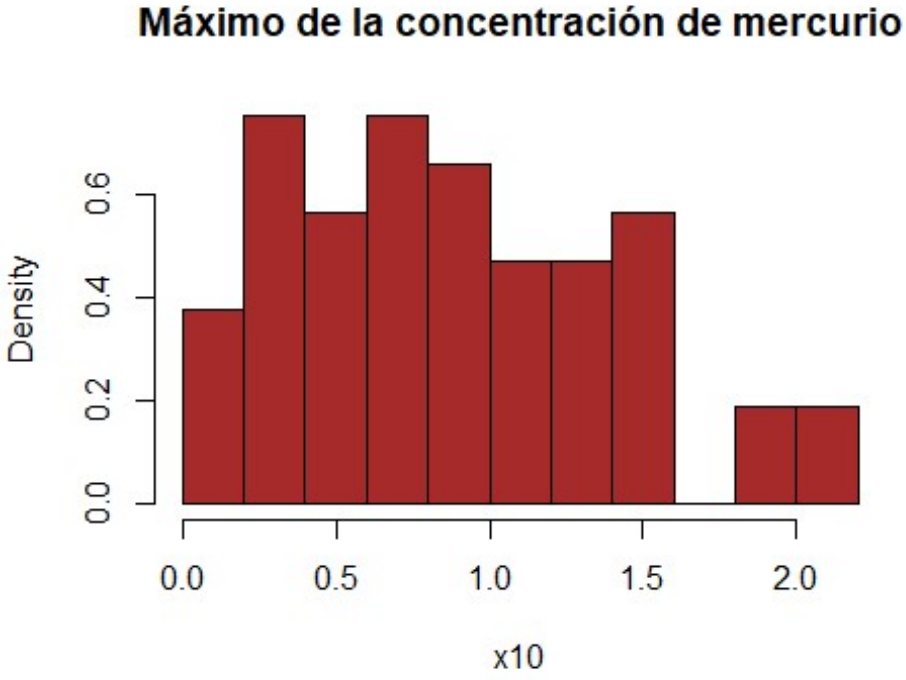
Dentro de la investigación realizada por (OCU, 2021) podemos observar que el mercurio suele liberarse al medio ambiente naturalmente por medio de procesos naturales. Por lo que este se puede encontrar en el suelo, el agua e incluso en la atmosfera. El problema que estamos enfrentando es que los humanos han estado aportando mayores cantidades de mercurio al medio ambiente de las que deberían de llegar de manera natural lo que nos lleva a mayor contaminación de todos estos medios. El enfoque que tomaremos nosotros será el daño realizado a los cuerpos acuíferos ms en especifico a los lagos. La base de datos otorgada es acerca de los lagos de florida por lo que tendremos que investigar cuales son los factores que mas influyen al alto nivel de contaminación de estos lagos. Para responder esto podemos responder a su vez otras preguntas que radican más en el área de la estadística, estas siendo: ¿En que facilita el estudio de la normalidad encontrada en un grupo de variables detectadas? Y ¿Cómo te ayudan los componentes principales a abordad este problema?

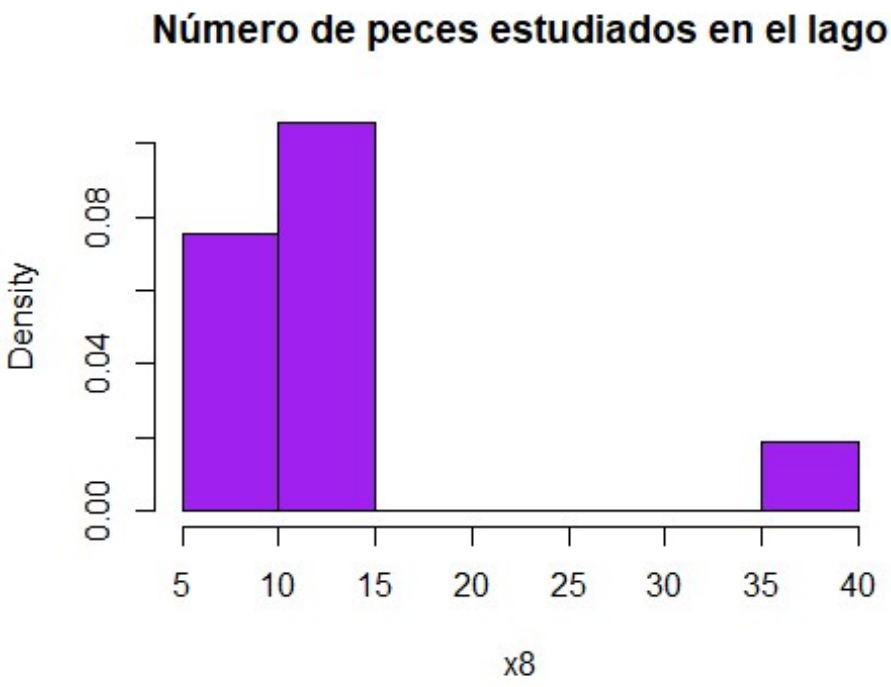
Análisis de Resultados

Análisis de normalidad

Para empezar con el análisis de normalidad visualizamos como se comportan los datos haciendo uso de histogramas, dentro de estos logramos observar que ninguno parece comportarse de manera normal. Cabe recalcar que algunas graficas como las de pH y la concentración máxima de mercurio parecen acercarse.


PH
Density
0.0
0.1
0.2
0.3
0.4
4
5
6
7
8
9
x4


Máximo de la concentración de mercurio
Density
0.0
0.2
0.4
0.6
0.0
0.5
1.0
1.5
2.0
x10



Para profundizar usaremos las pruebas de normalidad, tanto la de Mardia como la de Anderson-Darling, a continuación, se muestra la primera prueba.

Mardia Test

```
n_test = mvn(M2, mvnTest="mardia")
n_test$multivariateNormality
```

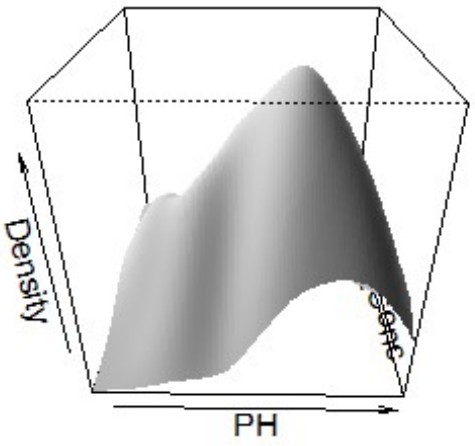
```
##              Test        Statistic              p value Result
## 1 Mardia Skewness 184.544953319842 1.65571079235445e-09     NO
## 2 Mardia Kurtosis  1.9860226693287   0.0470308068283103     NO
## 3             MVN             <NA>                 <NA>     NO
```

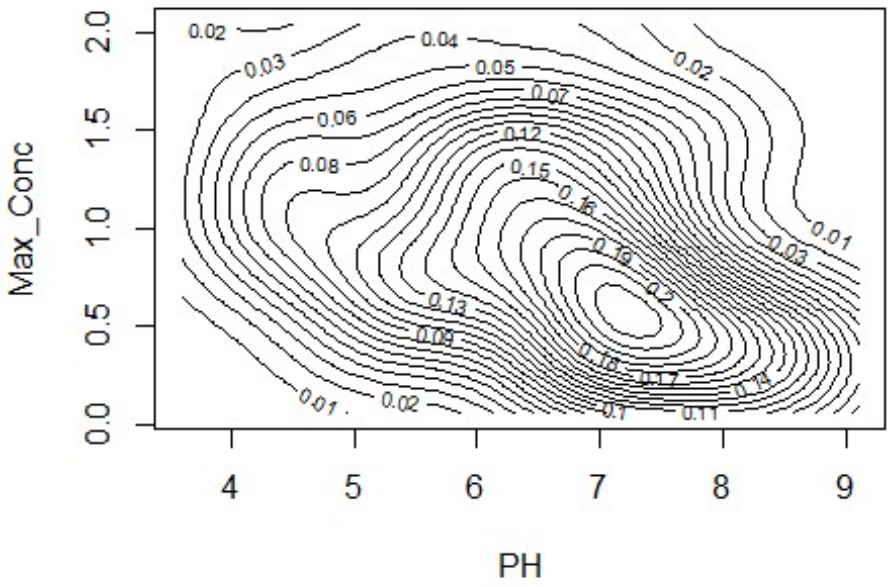
Se puede apreciar que no se logro pasar la prueba de normalidad debido a los valores de sesgo y de curtosis que se presentan en el análisis.

Procederemos a realizar la prueba de Anderson-Darling

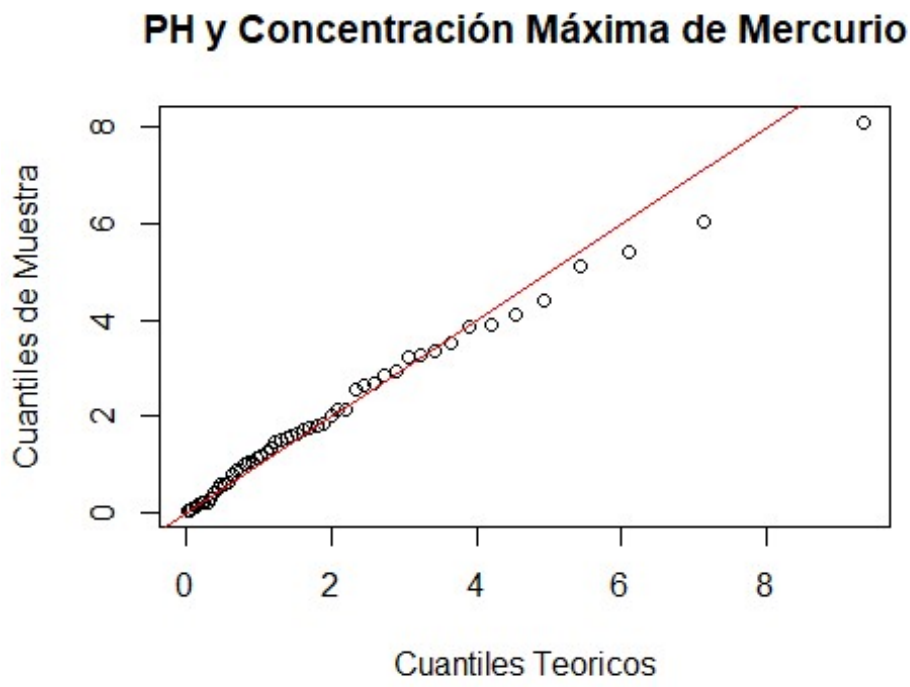
```
##              Test    Variable Statistic   p value Normality
## 1 Anderson-Darling Alcalinidad    3.6725  <0.001      NO
## 2 Anderson-Darling          PH    0.3496  0.4611      YES
## 3 Anderson-Darling       Calcio    3.9790  <0.001      NO
## 4 Anderson-Darling    Clorofila    4.7492  <0.001      NO
## 5 Anderson-Darling     Min_Conc    1.8380   1e-04      NO
## 6 Anderson-Darling     Max_Conc    0.6585   0.081      YES
## 7 Anderson-Darling     Est_Conc    0.8640  0.0248      NO
```

Podemos observar que nuestra corazonada del principio fue correcta ya que tanto el pH como la concentración máxima parecen comportarse de manera normal según la prueba de Anderson-Darling. Haciendo uso de estas dos variables que se distribuyen de manera normal podemos hacer una base de datas para poder realizar gráficos bivariados de ambas variables, los cuales se ven de la siguiente manera.

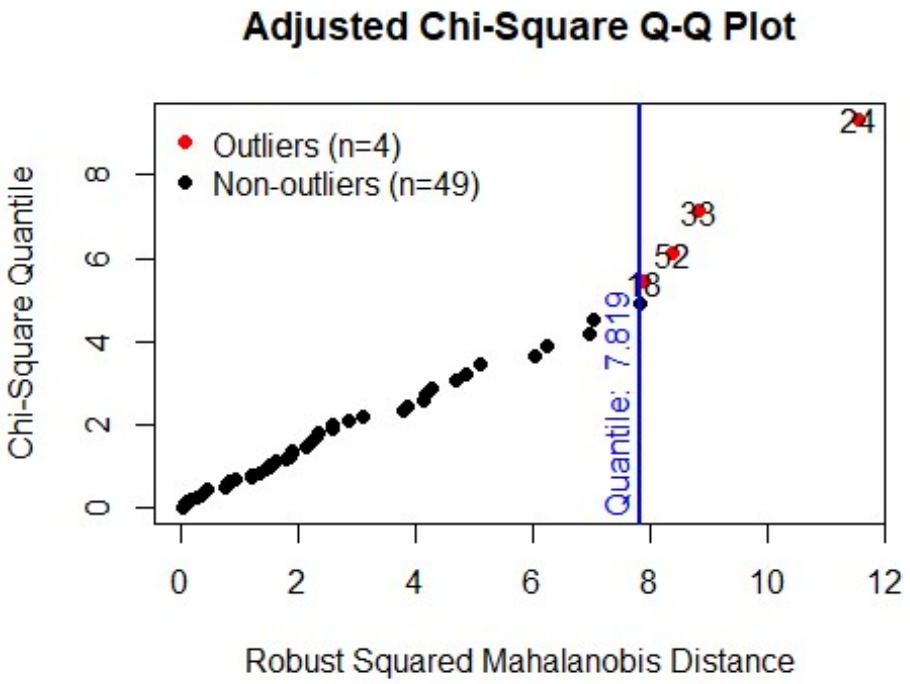




Podemos observar que los datos se centran dentro del área que nos indica que a mayor cantidad de pH en el agua la cantidad de concentración máxima de mercurio disminuirá, aunque es fácil observar que no se presenta un comportamiento muy homogéneo dentro de la grafica de contorno. Una vez observado esto procederemos a encontrar los componentes principales y los datos mas influyentes en nuestro modelo, para esto podemos usar una gráfica de plot bivariado.



En este grafico podemos observar que los datos se encuentran sesgados a la izquierda ya que presentan asimetría negativa por lo que tienden a ser normales, pero caen al final. Podemos checar datos atípicos con la gráfica de chi cuadrada y la distancia de mahalanobis.

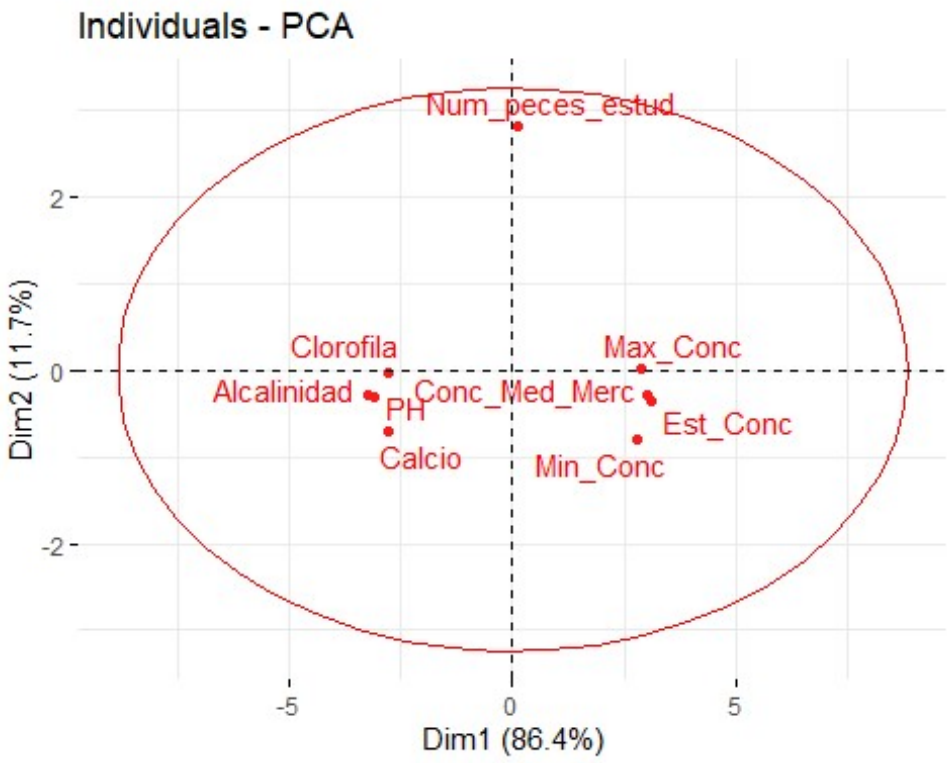


Análisis de componentes principales

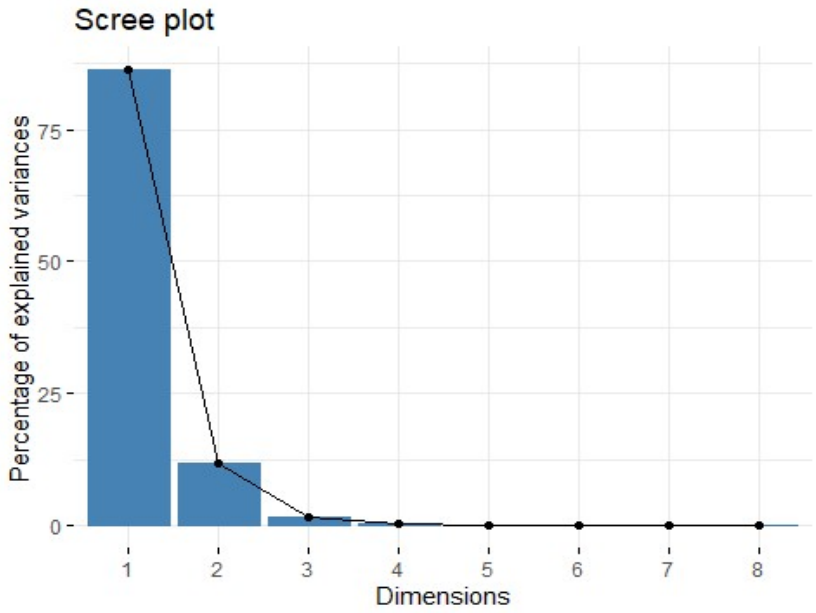
Empezamos con el análisis de componentes principales sacando la matriz de correlación de nuestras variables

```
##                Num_peces_estud    Min_Conc   Max_Conc   Est_Conc
## Alcalinidad         0.004950691 -0.55111301 -0.6047956 -0.6453385
## PH                 -0.029076933 -0.56603763 -0.5518152 -0.6397199
## Calcio             -0.097918189 -0.36699878 -0.4184350 -0.4905969
## Clorofila           0.051195243 -0.44069708 -0.4887723 -0.5455356
## Conc_Med_Merc       0.085676083  0.93036718  0.9158640  0.9672987
## Num_peces_estud     1.000000000 -0.07893672  0.1662619  0.0528902
## Min_Conc           -0.078936722  1.00000000  0.7766115  0.9158575
## Max_Conc            0.166261906  0.77661153  1.0000000  0.8851661
## Est_Conc            0.052890202  0.91585751  0.8851661  1.0000000
```

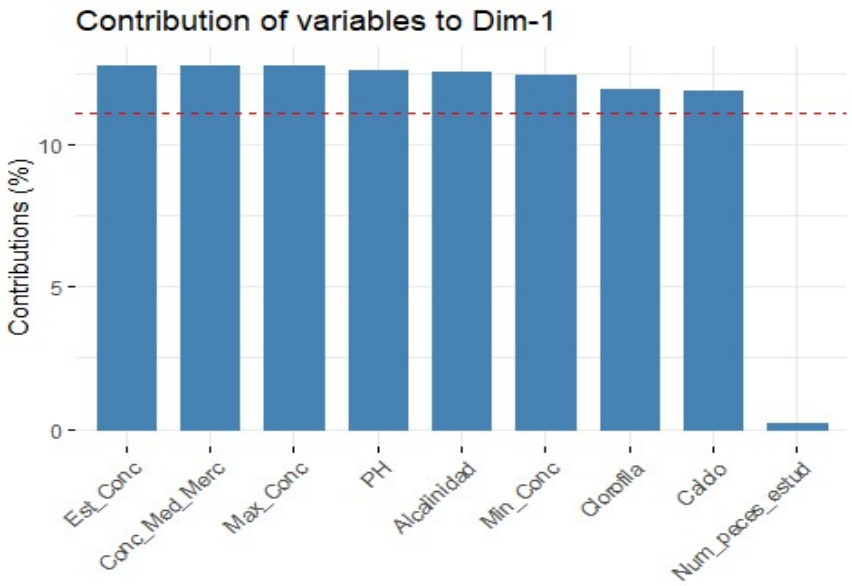
Con esta matriz podemos empezar a especular acerca de cuáles serán los componentes principales y cuáles serán las variables mas importantes en nuestro modelo, así como cuales no aportarán tanto. Pero para estar seguros procederemos con nuestro análisis de componentes principales.



Como se puede observar en las figuras pasadas podemos observar cómo se comportarán las variables según la dimensión en la que se encuentran, así como la aportación que le hace a esta. En la primera dimensión (PCA 1) podemos observar la concentración de la gran mayoría de nuestros datos, de igual manera podemos observar que las variables referentes al mercurio suelen incrementar la cantidad de mercurio en los lagos mientras que las variables como el pH, el calcio y la alcalinidad tienden a reducir la concentración de mercurio. Lo cual confirma nuestras sospechas que el pH, el calcio, la clorofila y la alcalinidad son benéficos para la salud de los lagos de florida y por ende los peces.



Con el plot de codo logramos ver que se alcanzan a reducir las dimensiones de 8 a 1 sola. Está teniendo una combinación lineal de las variables anteriores



En esta grafica nos podemos dar cuenta de la verdadera aportación que tienen las distintas variables, así como lo poco influyente que es la cantidad de peces estudiados para e estudio todo esto lo podemos observar de manera numérica en nuestra primera dimensión.

```
##                      Dim.1       Dim.2        Dim.3       Dim.4
Dim.5
## Alcalinidad     12.5702028  0.17197254 1.512505e+01  0.19885433
8.91674292
## PH              12.5945715  0.10195279 1.913838e+00 46.79736714
5.70314885
## Calcio          11.8816560  2.42957163 3.361792e+01 17.90502083
15.77803454
## Clorofila       11.9611991  0.05379898 4.709589e+01 29.08091520
3.22819069
## Conc_Med_Merc   12.7792349  0.55490106 3.029623e-01  0.18169160
1.15110041
## Num_peces_estud  0.1975586 93.49955239 1.663698e+00  0.61741085
0.02999291
## Min_Conc        12.4499927  2.66025508 4.489824e-04  4.82173110
10.17488150
## Max_Conc        12.7721368  0.01992908 2.301134e-01  0.08900537
54.87798873
## Est_Conc        12.7934475  0.50806646 5.007552e-02  0.30800358
0.13991946
```

Podemos observar que cada una de las componentes nos aporta aproximadamente el 12% de la información de nuestra componente principal dentro de la primera dimensión, lo que nos indica que nuestra segunda dimensión contiene la gran mayoría del componente ya que posee el 93% de la información.

Conclusiones

Podemos llegar a dos conclusiones para empezar podemos anotar que la prueba más útil para verificar comportamiento normal dentro de los datos fue la de Anderson-Darling ya que es más sensible para detectar desviaciones que suelen aparecer en colas de distribución.

En cuanto a los componentes principales observamos que este análisis nos ayudó a reducir la dimensión de nuestro modelo, ya que comenzamos con 8 variables lo que nos da 8 dimensiones, con ayuda del PCA se hizo una combinación lineal de las variables lo que nos ayudó a disminuir el tamaño y nos facilitó el procesamiento. Pudimos notar que el PH, calcio, clorofila y la alcalinidad nos ayudan a disminuir la cantidad de mercurio en los lagos. Mientras que la cantidad de peces que se estudiaron afectan al análisis, pero solo en aproximadamente el 12% de la información.

Referencias

OCU. (2021, 6 abril). Mercurio en el Pescado. www.ocu.org. Recuperado 1 de diciembre de 2022, de https://www.ocu.org/alimentacion/alimentos/noticias/mercurio-en-pescado-un-problema-serio522454