Momento de Retroalimentación

Jose Pablo Cruz Ramos

2022-10-25

Estudiante: "José Pablo Cruz Ramos"

Matricula: A01138740

Modulo 1: Estadística para ciencia de datos y nombre de la concentración.

Grupo: 502

Resumen

A) Problematica

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.

El objetivo esta en que con base en los datos presentados se debe investigar cuales de estos factores son aquellos que impactan de mayor manera al nivel de contaminacion por mercurio en los lagos.

B) Abordamiento

Como parte de la manera en la que se abordo la problemática, reutilizamos el reporte pasado en el cual ya contabamos el analisis de los datos y algunas conclusiones sobre aquellos factores que afectaban mas a la contaminacion de mercurio en los lago. En esta parte daremos uso de los conceptos de **componentes principales** y **normalidad de variables**, esto con la finalidad de poder visualizar en menos dimensiones la relevancia de cada variable con la variable objetivo y obtener la meta principal la cual es cuales afectan mayormente a la contaminación de mercurio.

Introduccion

Los pescados conforme pasa el tiempo se contaminan por el mercurio que absorben en las aguas ya sea dulces o saladas, el mercurio de dichos peces afectan directamente a la salud de las personas que los consumen. Se presenta un set de datos el cual demuestra registros de los lagos acerca de sus caracteristicas y a la vez se presenta informacion de los peces dentro. El objetivo es encontrar los factores que influyen en la contaminacion por mercurio del lago.

Pregunta a analizar: *¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?*

La importancia de la solución a esta problematica recae en el impacto que tiene el mercurio en la salud de los seres humanos justo como se menciona en la problematica planteada, el encontrar aquellos factores por

los que los lagos se contaminan de mercurio puede ayudar a encontrar maneras de reducir y mitigar dicha contaminacion. De esta manera se cuidan las vidas de las personas que consuman los pescados de estos lagos.

Factores y su descripcion

A continuacion se observan las variables presentes dentro del conjunto de datos del estudio realizado en los lagos de Florida.

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)

Analisis de Resultados

Para poder encontrar los factores que mayor impactan a la contaminación por mercurio en los lagos, haremos un analisis profundo sobre las variables, mas en especifico sobre su relevancia con la variable objetivo utilizando los conceptos de componentes principales.

Analisis de normalidad

Información relevante acerca de las variables, la cual se obtiene con la funcion de summary y apply la cual nos brinda la desviacion estandar de cada variable:

```
##  alcalinidad_agua_mgl       pH          calcio_mgl   clorofila_mgl
##  Min.   :  1.20       Min.   :3.600   Min.   : 1.1   Min.   :  0.70
##  1st Qu.:  6.60       1st Qu.:5.800   1st Qu.: 3.3   1st Qu.:  4.60
##  Median : 19.60       Median :6.800   Median :12.6   Median : 12.80
##  Mean   : 37.53       Mean   :6.591   Mean   :22.2   Mean   : 23.12
##  3rd Qu.: 66.50       3rd Qu.:7.400   3rd Qu.:35.6   3rd Qu.: 24.70
##  Max.   :128.00       Max.   :9.100   Max.   :90.7   Max.   :152.40
##  mercurio_promedio   num_peces      min_mercurio     max_mercurio
##  Min.   :0.0400    Min.   : 4.00   Min.   :0.0400   Min.   :0.0600
##  1st Qu.:0.2700    1st Qu.:10.00   1st Qu.:0.0900   1st Qu.:0.4800
##  Median :0.4800    Median :12.00   Median :0.2500   Median :0.8400
##  Mean   :0.5272    Mean   :13.06   Mean   :0.2798   Mean   :0.8745
##  3rd Qu.:0.7700    3rd Qu.:12.00   3rd Qu.:0.3300   3rd Qu.:1.3300
##  Max.   :1.3300    Max.   :44.00   Max.   :0.9200   Max.   :2.0400
##    edad_peces
##  Min.   :0.0000
##  1st Qu.:1.0000
##  Median :1.0000
##  Mean   :0.8113
##  3rd Qu.:1.0000
```

```
##  Max.   :1.0000
```

```
## [1] "Standard Deviation"
```

```
## alcalinidad_agua_mgl                   pH           calcio_mgl
##           38.2035267            1.2884493           24.9325744
##        clorofila_mgl    mercurio_promedio            num_peces
##           30.8163214            0.3410356            8.5606773
##         min_mercurio         max_mercurio           edad_peces
##            0.2264058            0.5220469            0.3949977
```

A) Realice la prueba de normalidad de Mardia y la prueba de Anderson Darling

Utilizaremos estas pruebas para identificar las variables que son normales y detectar posible normalidad multivariada de grupos de variables; Para esto utilizaremos la librería de MVN la cual se encarga de realizar las pruebas y nos arroja una serie de valores sobre las variables como la normalidad de cada una, su valor p, su sesgo, curtosis etc.

```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 368.078419447431 1.56422833730708e-17     NO
## 2 Mardia Kurtosis 2.87355952668305  0.00405874583752697     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##               Test             Variable Statistic   p value Normality
## 1 Anderson-Darling alcalinidad_agua_mgl    3.6725  <0.001      NO
## 2 Anderson-Darling          pH             0.3496  0.4611      YES
## 3 Anderson-Darling      calcio_mgl         4.0510  <0.001      NO
## 4 Anderson-Darling    clorofila_mgl        5.4286  <0.001      NO
## 5 Anderson-Darling  mercurio_promedio      0.9253  0.0174      NO
## 6 Anderson-Darling      num_peces          8.6943  <0.001      NO
## 7 Anderson-Darling     min_mercurio        1.9770  <0.001      NO
## 8 Anderson-Darling     max_mercurio        0.6585   0.081      YES
## 9 Anderson-Darling      edad_peces        14.3350  <0.001      NO
##
## $Descriptives
##                         n       Mean    Std.Dev Median  Min    Max  25th  75th
## alcalinidad_agua_mgl 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_mgl           53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60
## clorofila_mgl        53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70
## mercurio_promedio    53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77
## num_peces            53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00
## min_mercurio         53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33
## max_mercurio         53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33
## edad_peces           53  0.8113208  0.3949977   1.00 0.00   1.00  1.00  1.00
##                            Skew   Kurtosis
## alcalinidad_agua_mgl  0.9679170 -0.4705349
## pH                   -0.2458771 -0.6239638
## calcio_mgl            1.3045868  0.6130359
## clorofila_mgl         2.4130571  6.1042185
## mercurio_promedio     0.5986343 -0.6312607
## num_peces             2.5808773  6.0089455
## min_mercurio          1.0729099  0.4060828
```

```
## max_mercurio          0.4645925 -0.6692490
## edad_peces           -1.5465748  0.4005116
```

Como podemos observar con la prueba de Anderson Darling, la unica variable normal es el pH y la variable de max_mercurio. Ahora realizaremos la misma prueba pero solo para aquellas variables que si tuvieron normalidad.

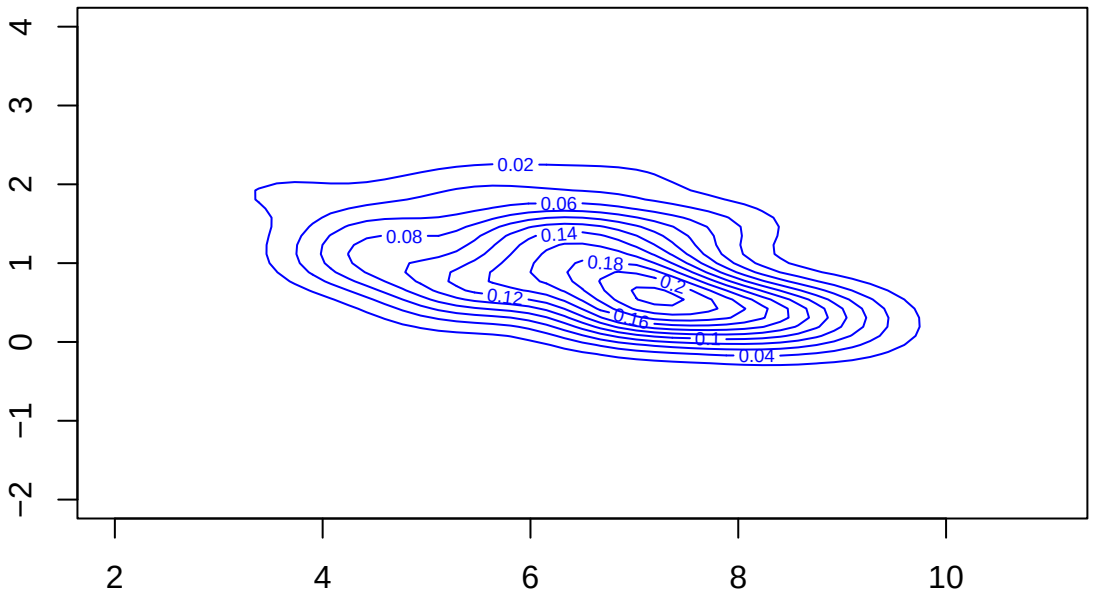
B) Realice la prueba de normalidad de Mardia y la prueba de Anderson Darling con variables que mostraron normalidad

```
## $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      pH         0.3496    0.4611    YES
## 2 Anderson-Darling max_mercurio    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
## max_mercurio 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925
##                 Kurtosis
## pH           -0.6239638
## max_mercurio -0.6692490
```

Podemos observar que al solo utilizar variables con normalidad, los resultados de las pruebas de Mardia y Anderson Darling nos arrojan resultados distintos. En este caso ambas pruebas de normalidad fueron pasadas, lo cual indica que estas variables son las que mejor distribucion normal tienen. En cuanto a la Kurtosis, en ambas variables es negativa esto indica una distribucion de sus valores relativamente planos comparada con la distribucion normal.

C) Grafica de contorno multivariado

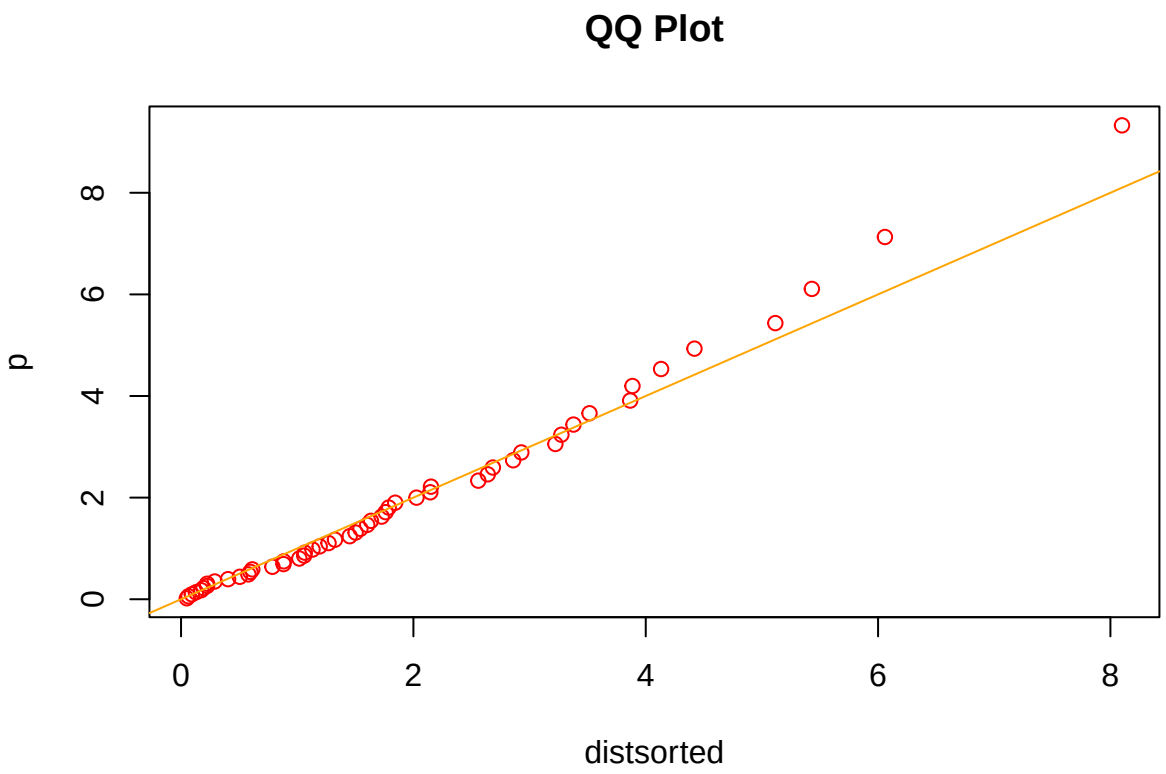
Realiza la prueba de Mardia y Anderson Darling de las variables que sí tuvieron normalidad en los incisos anteriores. Interpreta los resultados obtenidos con base en ambas pruebas y en la interpretación del sesgo y la curtosis de cada una de ellas.



D) Detecta datos atípicos o influyentes en la normal multivariada encontrada en el inciso B (auxíliate de la distancia de Mahalanobis y del gráfico QQplot multivariado)

Daremos uso de la distancia de Mahalanobis y el grafico de QQplot multivariado para encontrar aquellos "outliers" o datos atipicos dentro de la normal multivariada que se encontro previamente.Podemos utilizar la función de mahalanobis ya hecha en R para poder obtener las distancias, para esto necesitamos la covarianza del mismo dataset, y el vector de distribucion de la media.

```
##          1          2          3          4          5          6          7
## 1.19340732 3.86623312 4.13223576 0.06401297 2.55854258 0.59679099 1.78765285
##          8          9         10         11         12         13         14
## 1.50256004 1.01802464 1.63445643 1.26957081 0.22408796 0.61327299 5.43003539
##         15         16         17         18         19         20         21
## 2.14558317 2.15128332 2.68426029 5.11541618 0.78631278 1.84332204 3.88561874
##         22         23         24         25         26         27         28
## 0.04988963 0.88374242 8.09988683 2.02671417 0.58001802 0.50566324 2.64029211
##         29         30         31         32         33         34         35
## 2.92897033 1.06510339 0.88216107 0.40467236 6.05908470 3.51592769 1.45168260
##         36         37         38         39         40         41         42
## 1.32485980 1.13164718 2.85893264 0.22445752 3.37787450 1.54397421 1.72672287
##         43         44         45         46         47         48         49
## 0.28960341 1.76103444 1.60277372 0.09262808 0.19257450 3.22197239 3.27393627
##         50         51         52         53
## 0.17321589 0.12786732 4.41942776 1.06000856
```



A simple vista podemos observar que los datos estan bastante concentrados del lado izquierdo, posteriormente los datos se comienzan a disparar y tener distancias mas lejandas, podríamos decir que si existen unos cuantos datos atípicos de los cuales podemos deshacernos. A continuación demostraremos los resultados de normalidad con anderson para ver si cambian los resultados al eliminar aquellos datos atipicos.

```
## $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 max_mercurio    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
## max_mercurio 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925
##                 Kurtosis
## pH           -0.6239638
## max_mercurio -0.6692490
```

Podemos observar que bajó el valor p, igualmente, ahora pasaremos al analisis de componentes principales.

Análisis de componentes principales

Es adecuado utilizar el analisis de componentes principales para identificar los factores que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce, de una manera mas sencilla. Ya que

este analisis simplifica el labor de las dimensiones, al reducirlas y manteniendo las tendencias y patrones, de esta manera se identifican mas facilemente aquellas variables con mayor relevancia a la contaminación de mercurio en los lagos.

```
cat("La suma de la diagonal de covar y la suma de los valores es el mismo valor:",sum(vvp$values), sum(
```

```
## La suma de la diagonal de covar y la suma de los valores es el mismo valor: 9 9
```

```
cat("La suma de la diagonal de covar y la suma de los valores es el mismo valor:",sum(vvp_2$values), su
```

```
## La suma de la diagonal de covar y la suma de los valores es el mismo valor: 3106.33 3106.33
```

Comparacion entre componentes necesarios para mayor variabilidad explicada entre correlacion y covarianza

```
## Componentes con Covar: 0.5027982 0.6420323 0.7664771 0.8672251 0.9326721 0.966304 0.9879928 0.997635
```

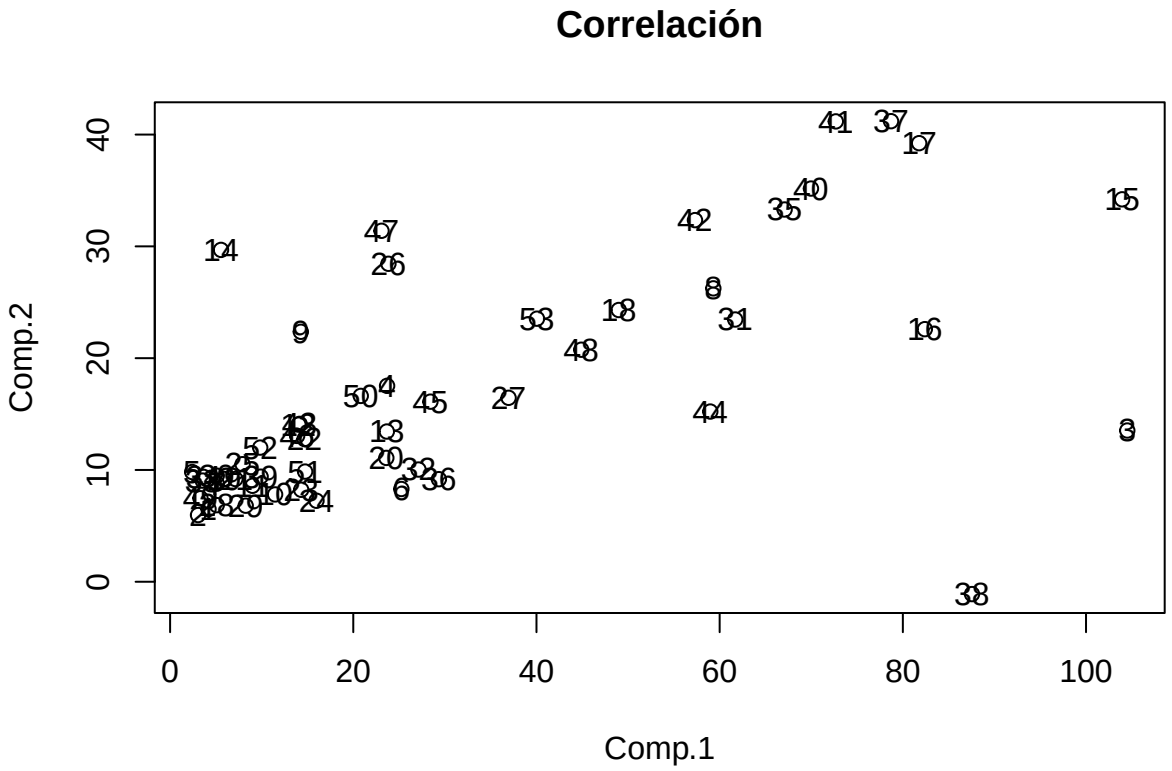
```
## Componentes con Correlacion: 0.5027982 0.6420323 0.7664771 0.8672251 0.9326721 0.966304 0.9879928 0.9
```

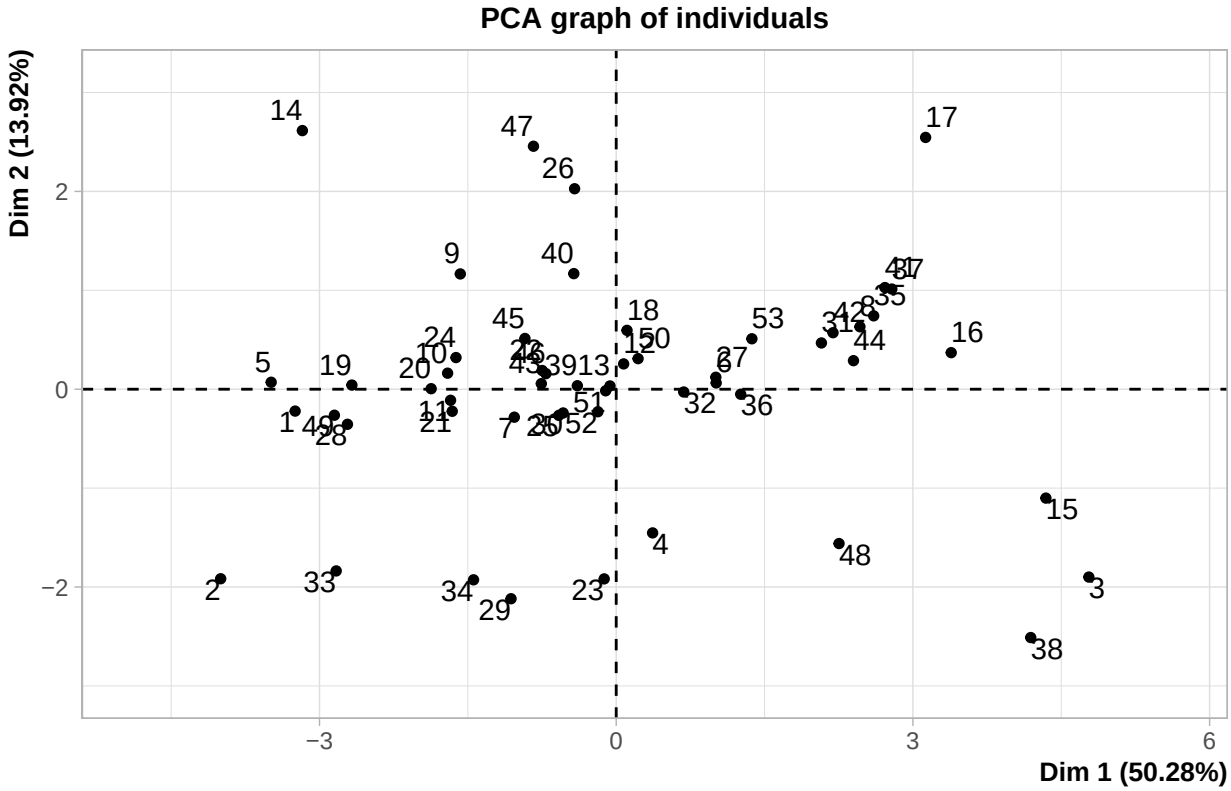
Como podemos observar al utilizar la covarianza en lugar de la correlacion, se necesitan de mas componentes para encontrar mayor variabilidad explicada.Esto se debe a que la matriz de correlacion es similar a realizar una estandarizacion por lo que los datos ya cobran mas sentido entre los mismos.

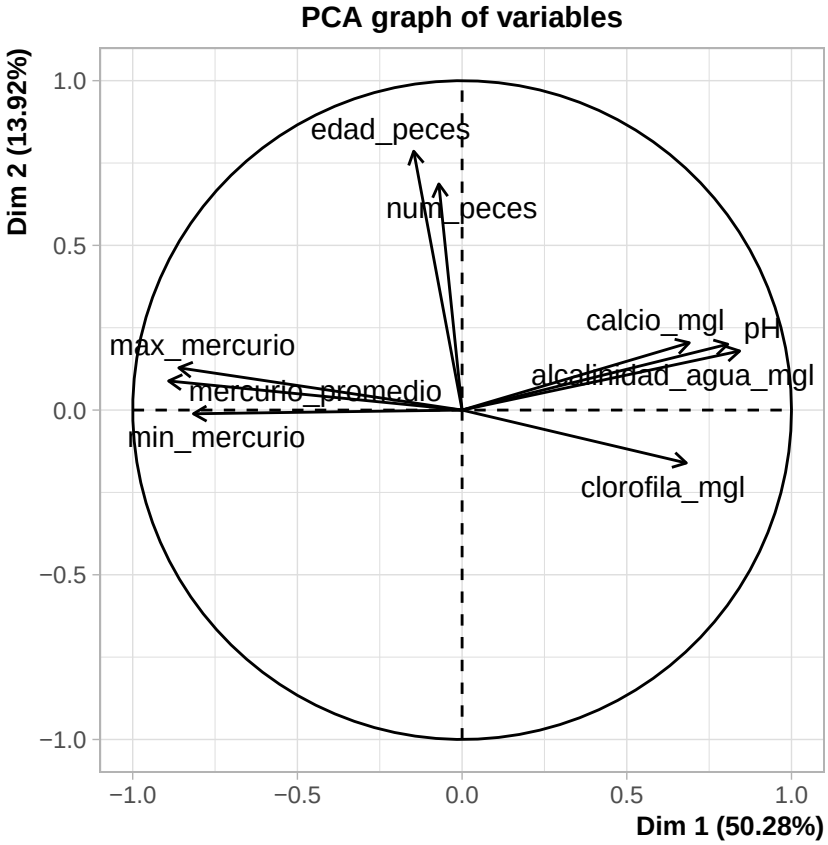
Por esta razón nos quedaremos con el uso de la correlación.Con los datos que se obtienen podemos observar que ya se puede observar la mayor parte de la variabilidad con 5 componentes. Se alcanza alrededor del 92% con estos cinco. Ahora pasaremos a realizar un diagrama sobre los vectores de cada factor.

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

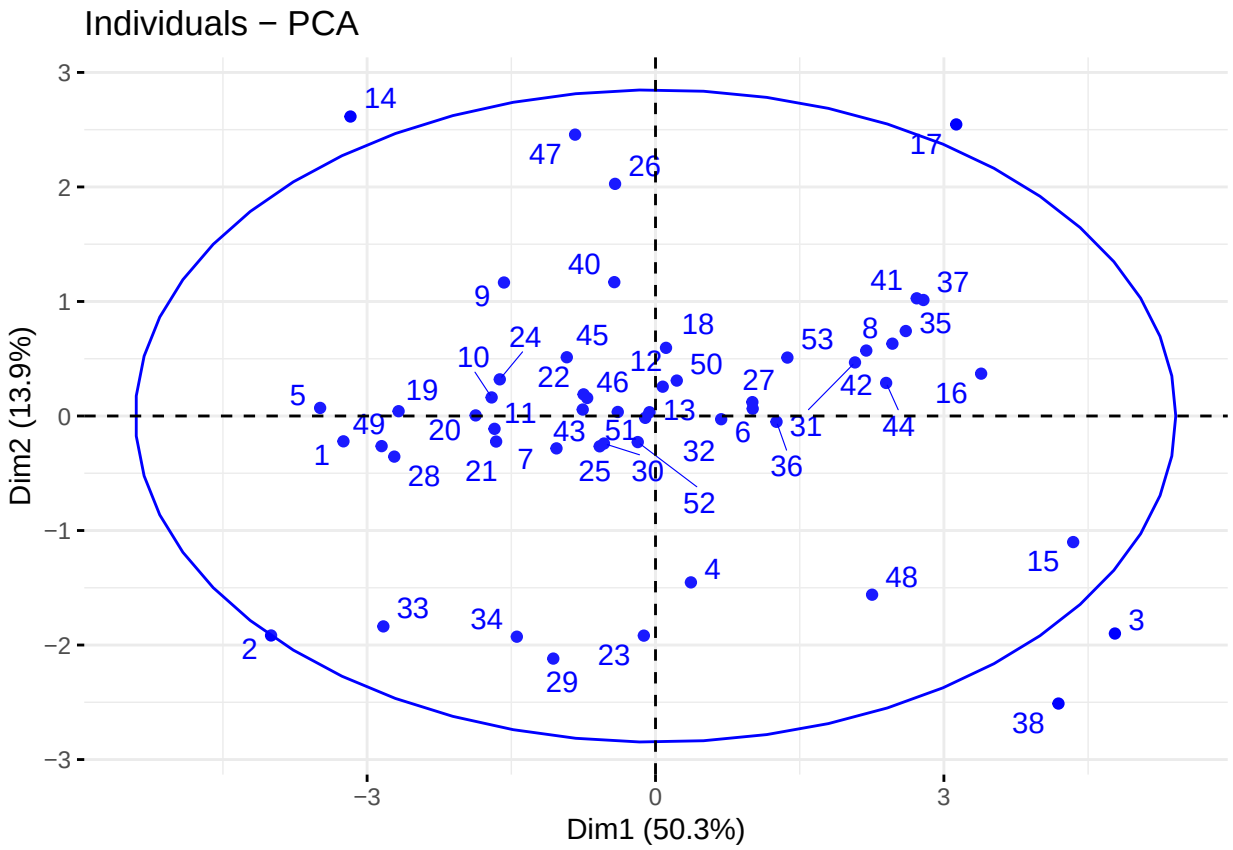
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```



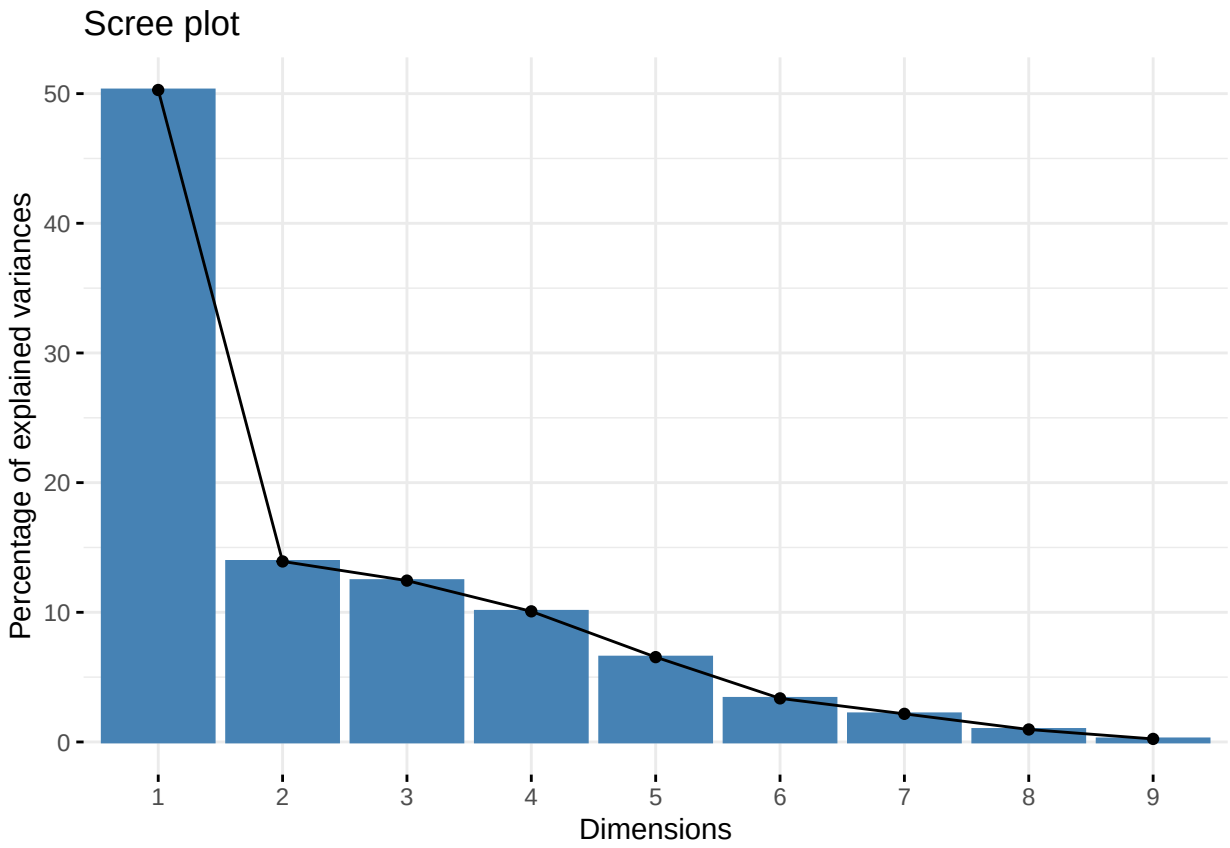




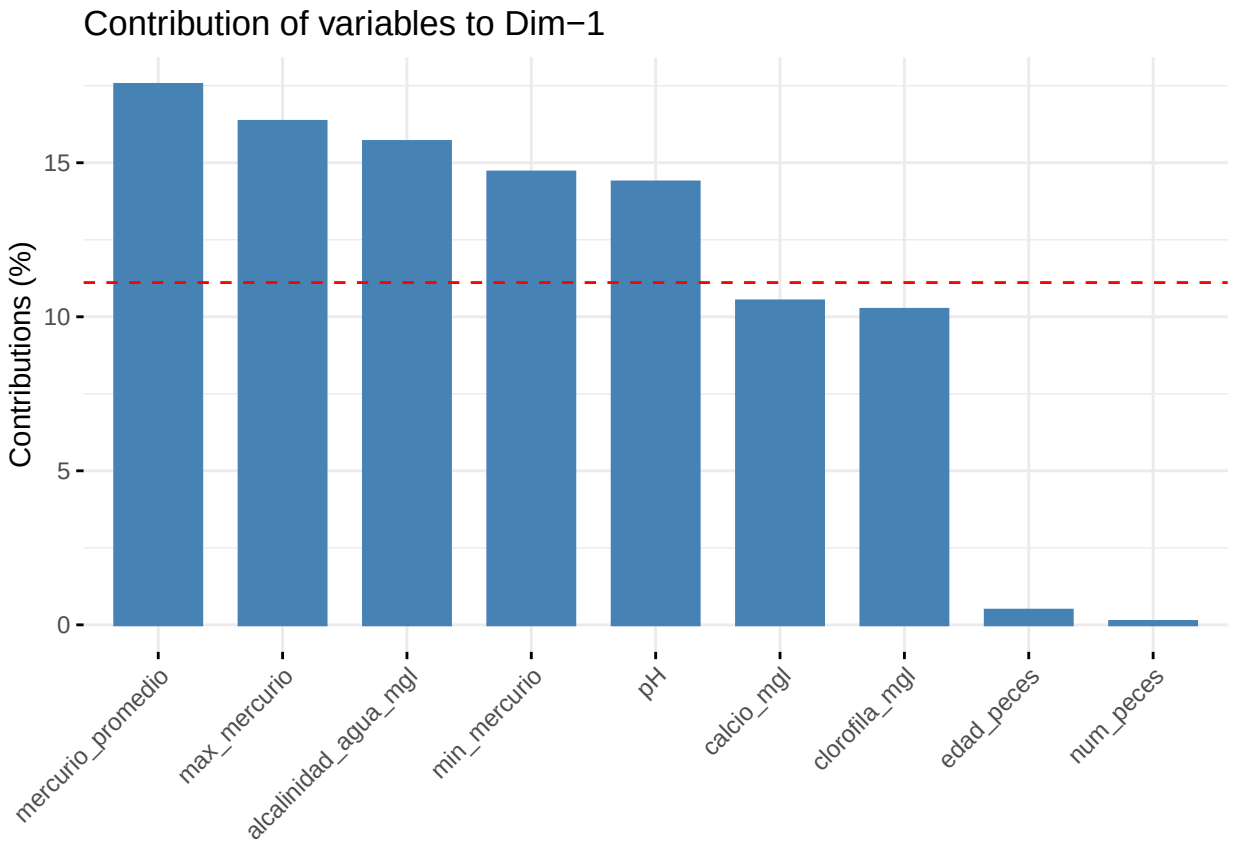

```
## Warning: ggrepel: 1 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
fviz_screeplot(cp3)
```




```
fviz_contrib(cp3, choice = c("var"))
```



Como podemos observar tanto del grafico de PCA y de los histogramas sobre el porcentaje de variabilidad explicada, esque es notable que hay ciertos factores que tienen mayor peso que otros, en este caso el primer componente es el que mayor variabilidad explicada nos provee, seguido de una gran disminución hacia el segundo. Sus pesos son principalmente alcalinidad, ph, clorofila y calcio. Esto embona con el analisis previo que habíamos realizado en el cual comentamos que estas variables eran las que mayor efecto tenian en la contaminación de mercurio.

Con los histogramas confirmamos que la variabilidad explicada se encuentra en un 99% aproximadamente cuando alcanza al quinto componente, otra manera de ver el analisis es que en el PCA aquellas variables mas alejadas son las que mas se pueden observar y que asi mismo mayor efecto tienen.

Conclusiones

Con el uso de los componentes principales se logró profundizar los hallazgos de aquellos factores que afectaban la contaminación de los lagos a través de otro acercamiento. Finalmente lo que se logró fue reconfirmar que los factores ya antes encontrados en el reporte pasado por lo que llegamos a conclusiones con gran similitud, lo cual es que el ph, la alcalinidad y el calcio son los que mayor relación tienen con el nivel de mercurio con los lagos. Esta vez incluso agregaremos la clorofila la cual con base en los graficos tambien demostro contribución.

Todo esto hace sentido debido a que el nivel de alcalinidad se ve afectado por el nivel de ph en el agua, lo cual esta comprobado cientificamente. La alcalinidad esta ligada al pH de manera que si un cuerpo liquido es mas alcalino entonces tiende a tener un ph mas alto (un ph base no acido), por lo que el bajar el nivel de alcalindiad involucra un mayor nivel de pH esto simboliza que se puede aumentar la cantidad de mercurio presente como se menciono en el articulo anexado en la parte inferior, esto nos ayuda a concluir que las variable que se encontraron que afectan al mercurio en este analisis satisfacen el comportamiento real del mercurio en los lagos.

ANEXO

LINK DEL DRIVE CON ARCHIVO R Y PDF: https://drive.google.com/drive/folders/1AJyoZmRN490iRqjAVUtIFI9pgyGktbLr?usp=sharing

Articulo mencionado *Articulo:* https://www.waterboards.ca.gov/water_issues/programs/swamp/docs/cwt/guidance/3140sp.pdf