Reporte Final Los Peces y el Mercurio: Procesamiento de Datos Multivariados

Módulo 5 (Portafolio Implementación)

Jorge Chávez Badillo A01749448

2022-12-03

Contaminación por Mercurio

Resumen

Para este reporte final del portafolio de implementación fue necesario utilizar diferentes modelos estadísticos para tratar el problema de contaminación de mercurio en lagos, ya que este es un tema sumamente importante, pues además de afectar la vida de los peces, también puede llegar a afectar de una forma fuerte la salud de los seres humanos si se consume un pescado contaminado por mercurio, por esta razón, fue necesario hacer un entendimiento de datos riguroso para poder decidir de qué manera implementar los modelos y que así se llegara a una conclusión sobre qué factores son los que tienen mayor efeto en la contaminación de los lagos.

En este trabajo se implementaron diferentes herramientas estadísticas, siendo las principales la prueba de normalidad para encontrar las variables normales y posteriormente la implementación del análisis de componentes principales para determinar los factores que más intervienen en el problema y se encontró que las variables que cuentan con un mayor nivel de contribución en los componentes son: $X11$ Estimación de la concentración de mercurio, $X7$ Concentración media de mercurio, $X10$ Máximo de concentración de mercurio, $X9$ Mínimo de concentración de mercurio, $X3$ Alcalinida y $X4$ PH, y además de ello, se encontró que las variables encontradas coinciden con las utilizadas en el bloque anterior.

Introducción

Descripción del 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.csv Descargar mercurio.csv y su descripción es la siguiente:

- *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 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)

Alrededor de la principal pregunta de investigación que surge en este estudio: ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida? pueden surgir preguntas paralelas que desglosan esta pregunta general:

1. ¿Hay evidencia para suponer que la concentración promedio de mercurio en los lagos es dañino para la salud humana? Considera que las normativas de referencia para evaluar los niveles máximos de Hg (Reglamento 34687-MAG y los reglamentos internacionales CE 1881/2006 y Codex Standard 193-1995) establecen que la concentración promedio de mercurio en productos de la pesca no debe superar los 0.5 mg de Hg/kg.
2. ¿Habrá diferencia significativa entre la concentración de mercurio por la edad de los peces?
3. Si el muestreo se realizó lanzando una red y analizando los peces que la red encontraba ¿Habrá influencia del número de peces encontrados en la concentración de mercurio en los peces?
4. ¿Las concentraciones de alcalinidad, clorofila, calcio en el agua del lago influyen en la concentración de mercurio de los peces?

Es muy importante el poder analizar estos datos, pues de alguna manera nos permite conocer y entender el comportamiento de la contaminación de lagos por mercurio, lo que en algún futuro puede ser de ayuda para evitar o disminuir esta problemática, pues esta tiene un nivel daño bastante elevado ya que es posible tener consecuencias negativas en la salud de los peces y la de los seres humanos.

Análisis de Resultados

Implementación de Herramientas Estadísticas

Para la solución de este problema, fue necesario utilizar como herramientas estadísticas el análisis de normalidad de las variables continuas para poder identificar variables normales, por otro lado, la segunda herramienta estadística implementada fue el análisis de componentes principales para poder identificar los componentes principales que intervienen en el problema de la contaminación por mercurio.

1. Análisis de Normalidad

Para la solución de este reporte es necesario hacer un análisis se normalidad de las variables continuas para identificar variables normales.

a) Como primer paso es necesario realizar las pruebas de Mardia y Anderson Darling para poder identificar las variables que son normales y detectar posible normalidad multivariada de grupos de variables.

```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 410.214790601476  7.0419877781575e-23     NO
## 2 Mardia Kurtosis  4.5961255577272 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
```

b) Como paso siguiente se procede a hacer la prueba de Mardia y Anderson Darling de las variables que si tuvieron normalidad en el inciso anterior, en adición a ello es importante tener una interpretación de los resultados obtenidos con base a ambas pruebas, así como en la interpretación del sesgo y la curtosis de cada una de ellas.

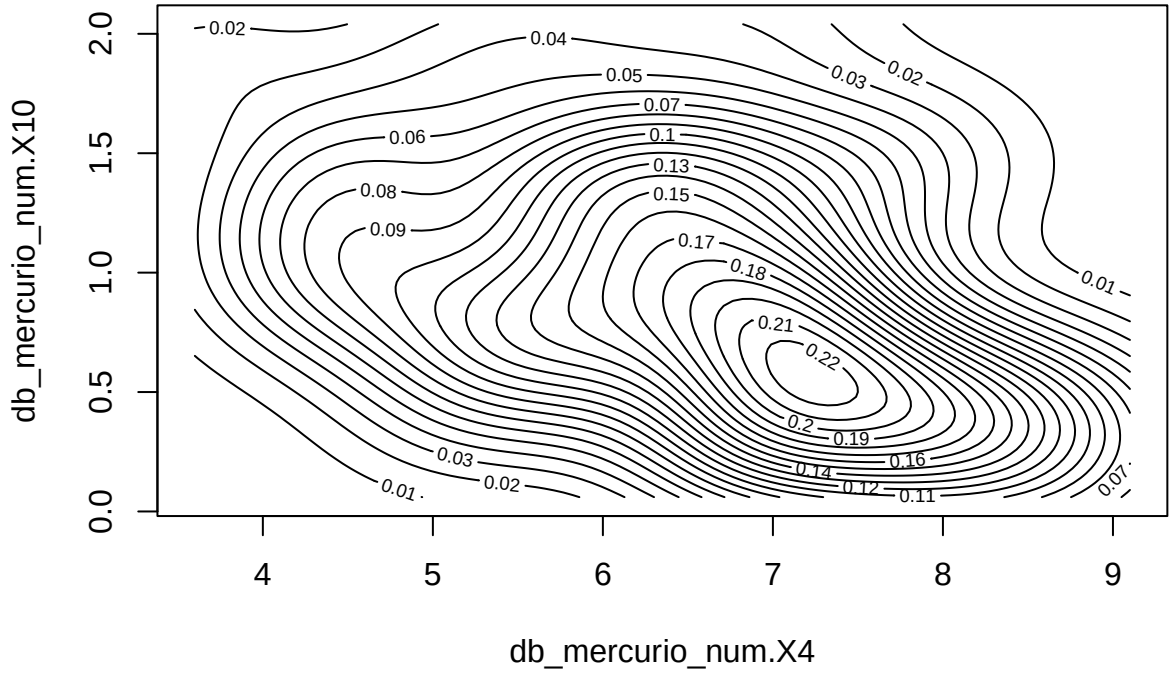
```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  6.17538668676459 0.186427564928851    YES
## 2 Mardia Kurtosis -1.12820795824432 0.259232103759911    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##               Test            Variable Statistic   p value Normality
## 1 Anderson-Darling db_mercurio_num.X4     0.3496    0.4611    YES
## 2 Anderson-Darling db_mercurio_num.X10    0.6585    0.0810    YES
##
## $Descriptives
##                       n      Mean   Std.Dev Median  Min  Max 25th 75th
## db_mercurio_num.X4  53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40
```

```
## db_mercurio_num.X10 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33
##                             Skew   Kurtosis
## db_mercurio_num.X4  -0.2458771 -0.6239638
## db_mercurio_num.X10  0.4645925 -0.6692490
```

Tanto la prueba de Mardia y Anderson Darling, arrojan como resultado que las variables, en efecto, provienen de una distribución normal y al observar el valor p de las estadísticas de asimetría y curtosis, tenemos que estos son mayores a el grado de significancia $\alpha = 0.05$ y por lo tanto se concluye que hay si existe normalidad multivariable.

Por otro lado, al enfocarse en los valores de la curtosis y el sesgo, tenemos que la curtosis cuenta con un valor de 0.25, lo que indica que la forma de la distribución normal es casi mesocúrtica ya que el coeficiente de curtosis se encuentra cercano al 0 y finalmente, el valor del sesgo es de 0.18, se puede decir que este es moderadamente simétrico debido a que el coeficiente se encuentra dentro del rango de 0 a 0.5.

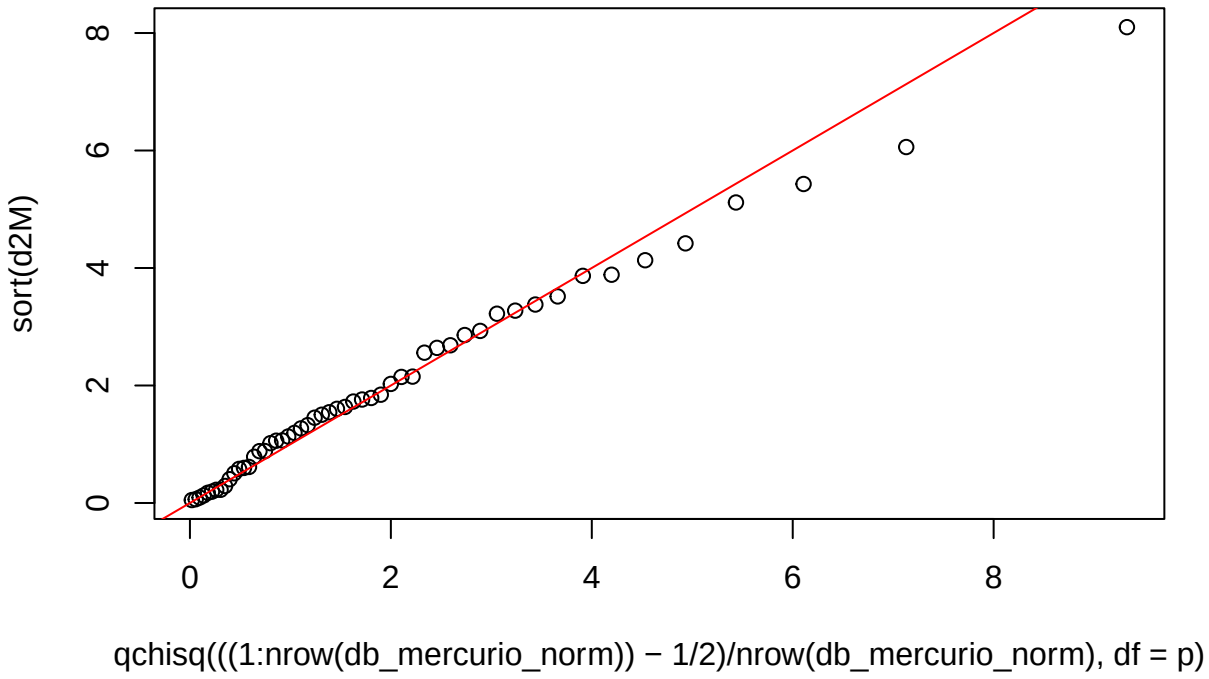
c) Gráfica de contorno de la normal multivariada obtenida en el inciso B.




```
## $multivariateNormality
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.7695729 0.1024763 YES
##
## $univariateNormality
##               Test            Variable Statistic   p value Normality
## 1 Anderson-Darling db_mercurio_num.X4     0.3496    0.4611    YES
## 2 Anderson-Darling db_mercurio_num.X10    0.6585    0.0810    YES
```

```
##
## $Descriptives
##                         n      Mean   Std.Dev Median  Min  Max 25th 75th
## db_mercurio_num.X4  53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40
## db_mercurio_num.X10 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33
##                            Skew   Kurtosis
## db_mercurio_num.X4  -0.2458771 -0.6239638
## db_mercurio_num.X10  0.4645925 -0.6692490
```

d) Detección de datos atípicos o influyentes en la normal multivariada encontrada en el inciso B (con aydua de la distancia de Mahalanobis y del gráfico QQplot multivariado).



De acuerdo con el test de multinormalidad de Q-Q Plot y utilizando al distancia de Mahalanobis, podemos observar que los datos tienen una asimetría negativa con sesgo a la izquierda.

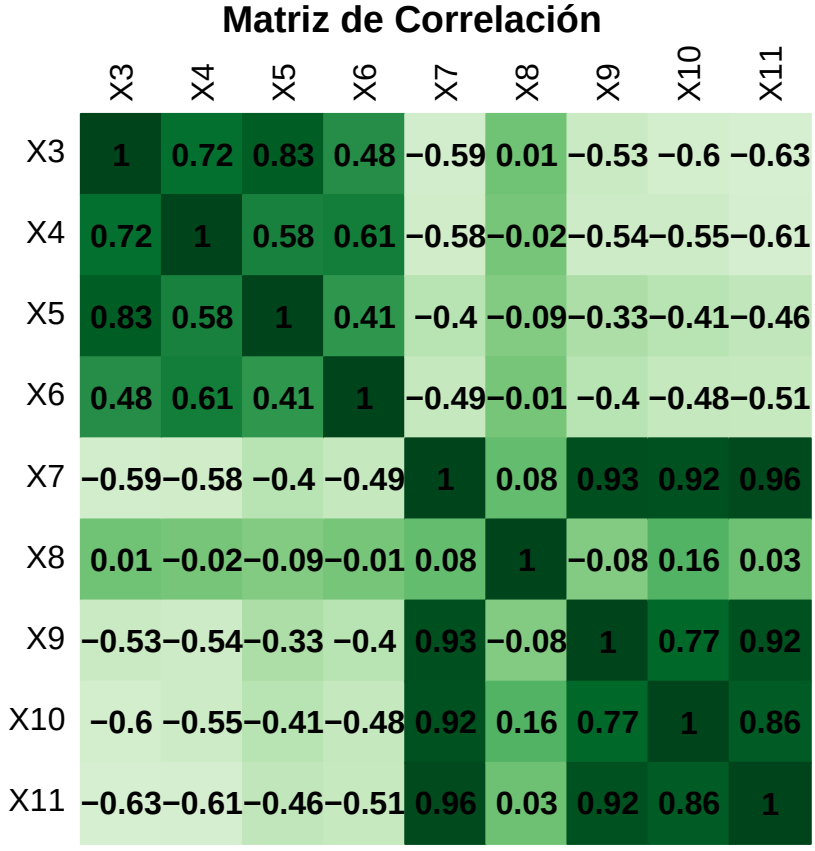
2. Análisis de Componentes Principales

Ahora, se procede a realizar el análisis de componentes principales con la base de datos completa para poder identificar los factores principales que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce.

a) Antes de comenzar con el análisis de componentes principales es necesario justificar por qué es adecuado el uso de esta herramienta estadística para el análisis de esta base de datos usando la matriz de correlaciones.

```
## corrplot 0.92 loaded
```

```
##              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 método de componentes principales busca representar una matriz de datos mediante variables con poca pérdida de información, donde las variables son combinaciones lineales de las originales, es decir, de los componentes, dando como resultado componentes con un alto porcentaje de la información y se busca

principalmente reducir la dimensionalidad del conjunto de los datos y además, como podemos observar en la matriz de correlacion, la mayoría de las variables tienen un coeficiente aceptable de correlación, lo cual es muy bueno ya que con ello se cumplen los supuestos de los componentes principales, los cuales mencionan que las variables tienen que estar correlacionadas y deben de ser numéricas.

b) Ahora, se procede a hacer el análisis de componentes principales justificando el número de componentes principales apropiados para reducir la dimensión de la base.

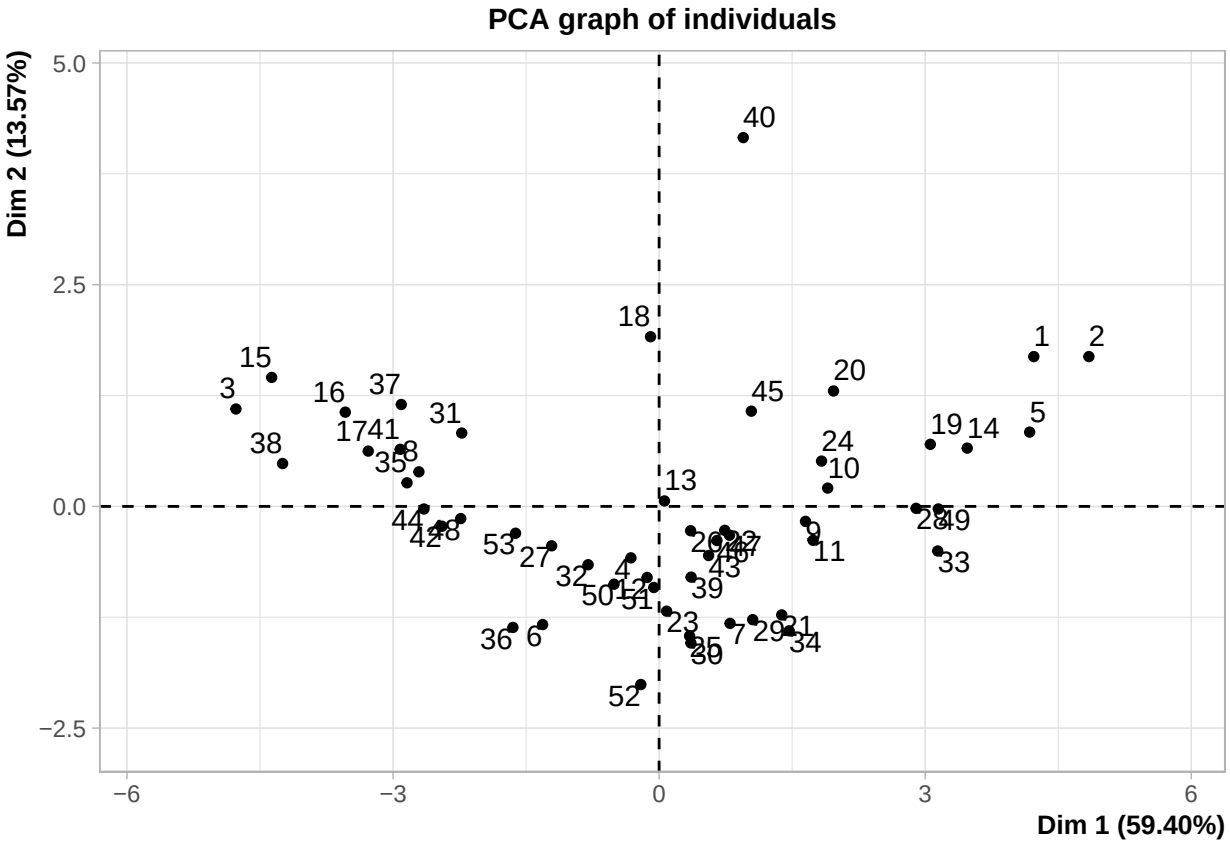
```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5     PC6     PC7
## Standard deviation     2.312 1.1049 1.0210 0.81723 0.5794 0.45710 0.32750
## Proportion of Variance 0.594 0.1357 0.1158 0.07421 0.0373 0.02322 0.01192
## Cumulative Proportion  0.594 0.7297 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
```

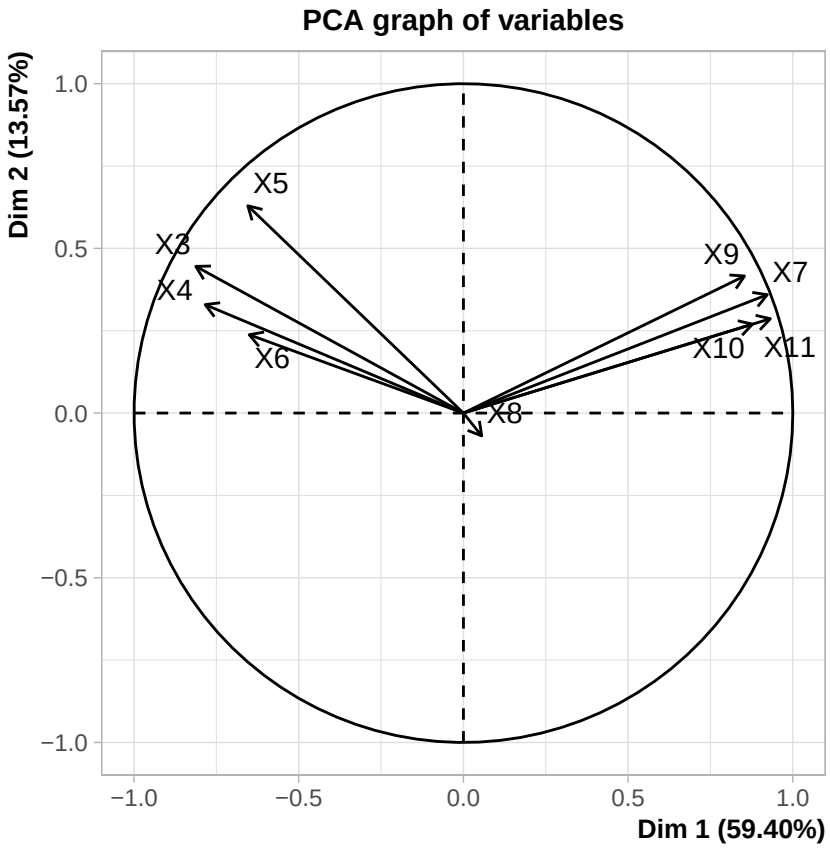
En el caso de este problema, el número apropiado de componentes principales sería dos ya que los dos primeros componentes son los que explican mayormente la información de la base de datos y los componentes restantes son menos significantes para la explicación del porcentaje de varianza; observando los resultados se tiene que el componente número 1 explica el 59% de la información, mientras que el segundo explica solo el 13% de la misma, sin embargo, se tiene que el porcentaje de varianza explicado por ambos componentes principales es del 72%.

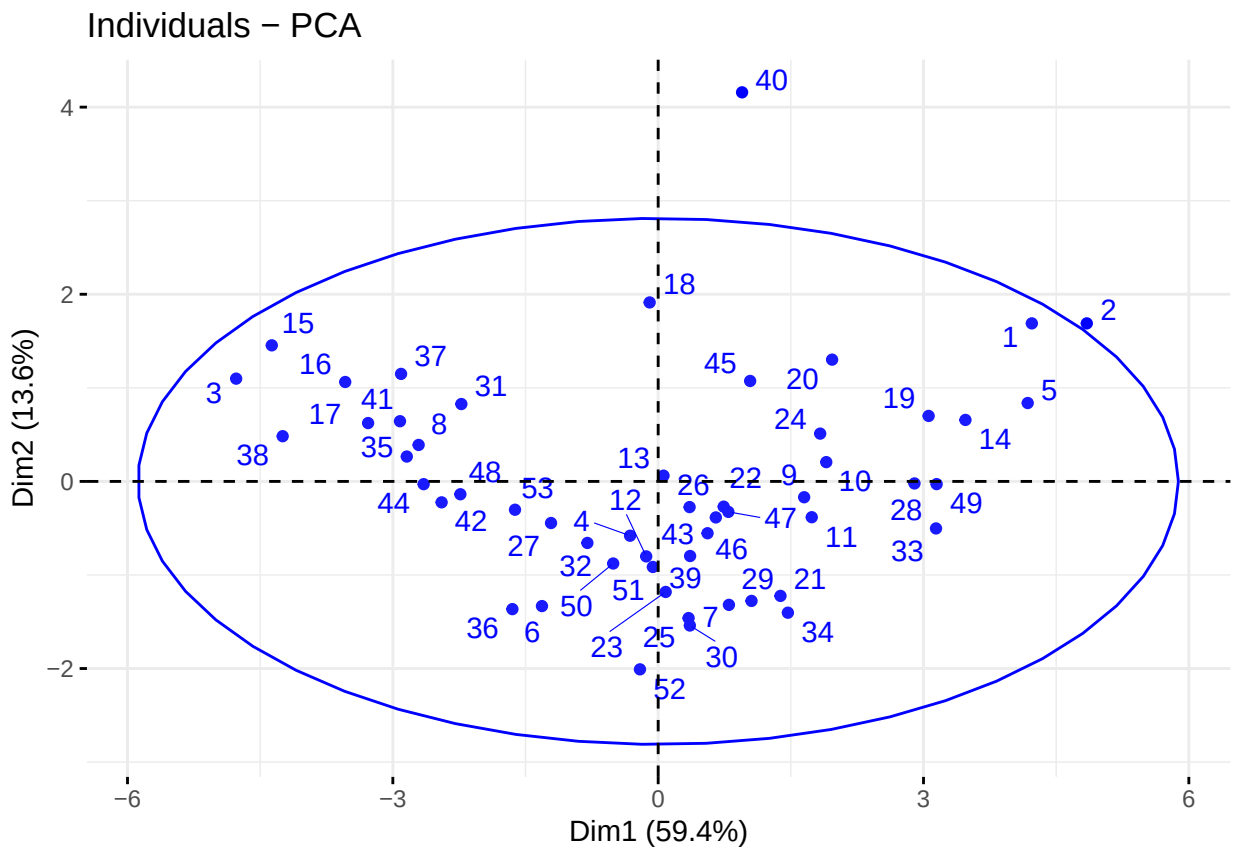
c) Representación en un gráfico sobre los vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes.

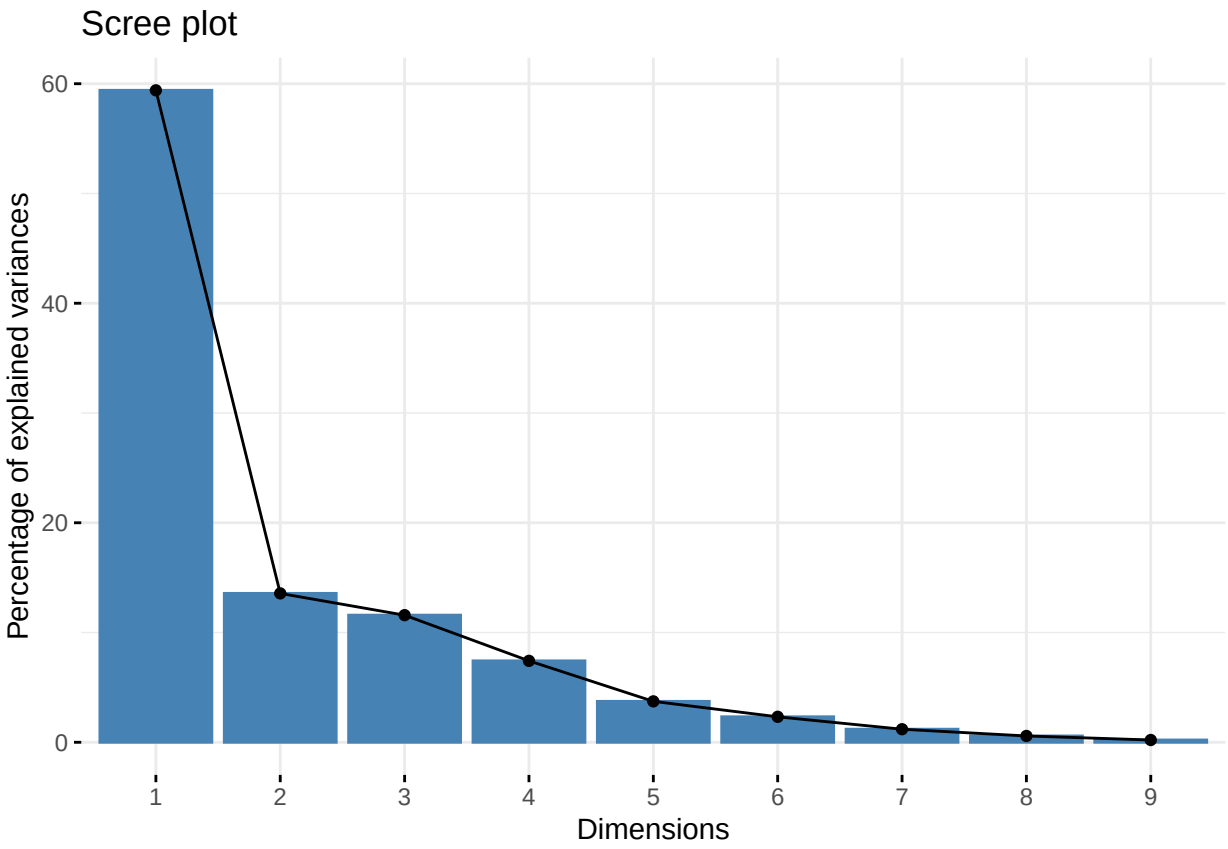
```
## Loading required package: ggplot2
```

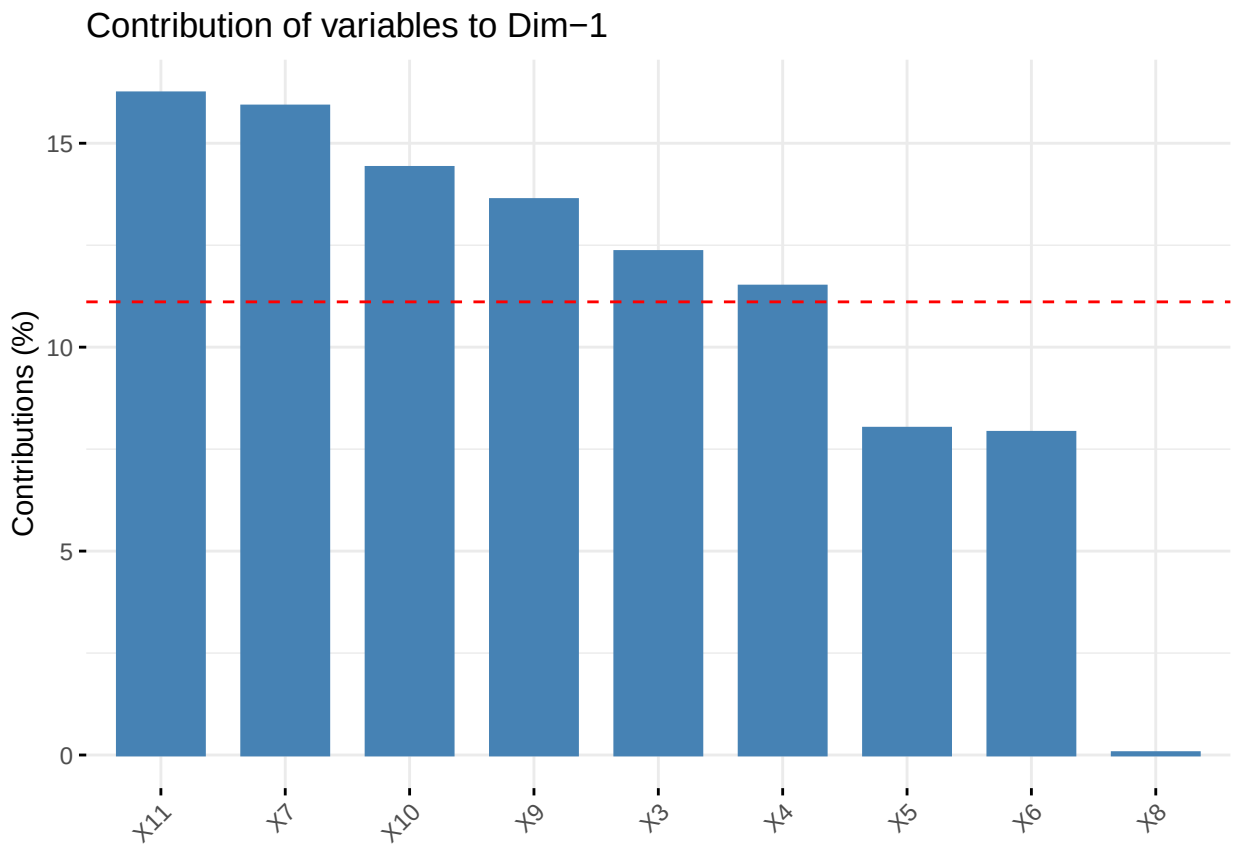
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```


PCA graph of individuals
Dim 2 (13.57%)
Dim 1 (59.40%)
5.0
2.5
0.0
−2.5
−6
−3
0
3
6


PCA graph of variables
Dim 2 (13.57%)
Dim 1 (59.40%)
1.0
0.5
0.0
−0.5
−1.0
X5
X3
X4
X6
X9
X7
X10
X11
X8


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
4
27
32
51
50
36
6
23
25
52
30
7
39
29
21
34
12
43
46
47
11
13
26
22
9
10
45
20
24
19
14
5
1
2
28
49
33


Scree plot
Percentage of explained variances
60
40
20
0
1
2
3
4
5
6
7
8
9
Dimensions



d) Interpretación de los resultados. Explique brevemente a qué conclusiones llega con su análisis y qué significado tienen los componentes seleccionados en el contexto del problema.

Con respecto a las gráficas anteriores podemos observar que los primeros dos compontentes son los que mayor importancia tienen y de acuerdo con el último gráfico de barras podemos observar que las variables $X11, X7, X10, X9, X3 y X4$ son las que cuentan con una mayor contribución en los componentes ya que estas se encuentran por encima de la media de los coeficientes de contribución.

Conclusión

Emite una conclusión general: Une las conclusiones aquí obtenidas con las ya obtenidas en el análisis que ya habías realizada anteriormente, ¿de qué forma te ayuda este nuevo análisis a contestar la pregunta principal del estudio: ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida? ¿en qué puede facilitar el estudio la normalidad encontrada en un grupo de variables detectadas? ¿cómo te ayudan los componentes principales a abordar este problema?

Gracias a la implementación del método de componentes principales, fue posible identificar qué componentes son los más importantes y que brindan una mayor información para poder hacer los análisis necesarios para poder responder a la pregunta principal de estudio, en este análisis de concluyó que las variables de la base de datos que cuentan con un mayor nivel de contribución en los componentes son:

- $X11$ Estimación de la concentración de mercurio.
- $X7$ Concentración media de mercurio.
- $X10$ Máximo de concentración de mercurio.
- $X9$ Mínimo de concentración de mercurio.

- $X3$ Alcalinidad
- $X4$ PH

Respondiendo a la pregunta sobre la normalidad en un grupo de variables, es posible afirmar que aunque al utilizar el método de componentes principales no se requiere el cumplimiento del supuesto de normalidad, es importante destacar que cuando las variables cuentan con normalidad, los componentes tienen interpretaciones mucho más detalladas y brindan mayor información.

Dado que los componentes principales buscan las direcciones de las proyecciones que conserven más las propiedades de los datos, donde generalmente se tiene una mayor variabilidad, se puede concluir que los componentes principales son una herramienta de gran ayuda para este problema pues al tener una cantidad significante de diferentes variables numéricas este método nos permite encontrar los componentes con las combinaciones lineales que brinden mayor información para el análisis de los datos, logrando así disminuir la dimensionalidad del conjunto de datos, así como la evaluación de la semejanza entre los individuos a través de las semejanzas entre ellos o la topología entre individuos.

De acuerdo con el análisis que se realizó en el bloque anterior y comparándolo directamente con lo que se obtuvo en este análisis, tenemos que coinciden las variables que son más importantes en el nivel de contribución de los componentes con las variables que mejor funcionaron para la regresión múltiple de la entrega anterior; concluyendo así que los factores que tienen un mayor efecto en la contaminación por mercurio en los peces de los lagos de Florida son la alcalinidad, el ph, ya que estos dos fueron los que coincidieron al utilizar componentes principales y el análisis de la entrega anterior.

Anexos

Liga de Github: https://github.com/A01749448/portafolio-implementacion