Reporte Final de "los peces y el mercurio

Amy Murakami Tsutsumi - A01750185

2022-10-20

Inteligencia Artifical para la Ciencia de Datos II (Grupo 501)

Módulo 5: Estadística Avanzada para ciencia de datos

1. Resumen

La problemática a resolver es identificar los principales factores que influyen en el nivel de contaminación por mercurio en los peces que se encuentran en el lago de Florida. Los métodos y técnicas estadísticas utilizadas son la prueba de normalidad de Mardia, la prueba de Anderson Darling, gráfico de contorno de la normal multivariada, la distancia de Mahalanobis, gráfico QQplot multivariado y análisis de componentes principales (matriz de correlaciones, gráfico de scree plot y contribución de variables). Después de todo el proceso de análisis se obtuvo que los factores más influyentes son la alcalinidad, PH, concentración media de mercurio, mínimo de concentración de mercurio, máximo de concentración de mercurio y estimación de la concentración de mercurio en el pez de 3 años.

2. Introducción

Este portafolio de implementación tiene el propósito de utilizar herramientas estadísticas vistas en el módulo cinco para poder analizar la información y de esta manera se pueda contestar la pregunta de investigación establecida. El problema consiste en la contaminación por mercurio de los peces que se encuentran en agua dulce, por lo tanto, se utilizará un dataset con información de 53 lagos de Florida. El dataset contiene las siguientes variables:

- X1: Número de identificación
- X2: Nombre del lago
- X3: Alcalinidad (mg/l de carbonatode 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)

La pregunta que se debe contestar a lo largo de la implementación es la siguiente: ¿cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida? Este análisis es de suma importancia para determinar cuáles son las principales causas de la contaminación en los peces. Esta información no sólo afecta a los peces, sino que a la salud humana, ya que los humanos consumimos pescado y puede resultar en algo dañino.

Por lo tanto, se realizará un análisis de normalidad para identificar las variables que son normales y también detectar posible normalidad multivariada de grupos de variables. Al igual que un análisis de componentes principales con toda la base de datos para poder identificar los factores principales que generan la contaminación por mercurio de los peces en agua dulce.

3. Análisis de resultados

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

A. Realice 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 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 con la prueba de Mardia y Anderson Darling, las variables que tienen normalidad son la X4 y X10. Además, al observar las medidas descriptivas se puede notar la variabilidad de las variables utilizando el cociente entre la desviación estándar y la media. De esta forma, se pueden distinguir las variables en las que la media es más confiable ya que tienen una menor variabilidad, por ejemplo, la X4 (0.19549904), X10 (0.59694683), X7 (0.64691794), X8 (0.65565881) y X11 (0.66002426), a diferencia de las otras variables que tienen una variablidad entre 0.8 y 1.33.

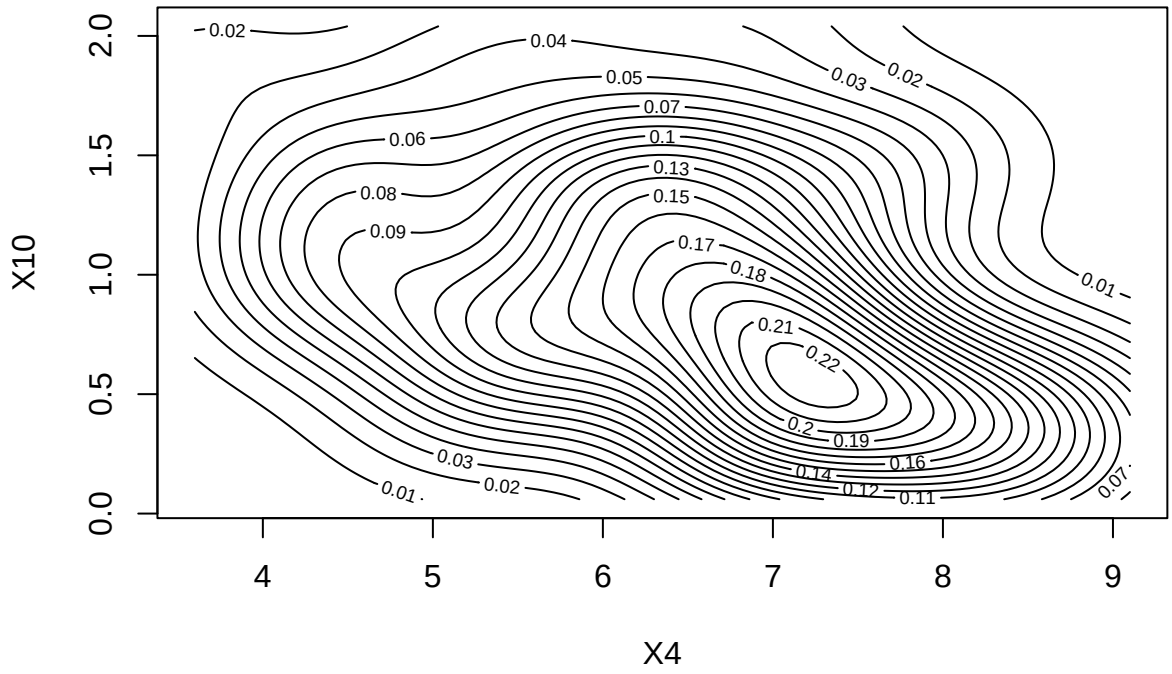
B. Realiza la prueba de Mardia y Anderson Darling de las variables que sí tuvieron normalidad en los incisos anteriores. Interpreta los resultados obtenidos con base en ambas pruebas y en la interpretación del sesgo y la curtosis de cada una de ellas.

```
## $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
```

En la prueba de Mardia indica si un grupo de variables sigue o no una distribución normal multivariante. Dado que el valor de p de la skewness (0.18642) y kurtosis (0.2592) no son menores a 0.05, entonces no se rechaza la hipótesis H0 (los datos se distribuyen normalmente). De la misma manera, en la prueba de Anderson-Darling no se rechaza la hipótesis H0 ya que el valor p de la variable X4 es 0.4611 y de la variable X10 es 0.0810 ambos valores no son menores a 0.05.

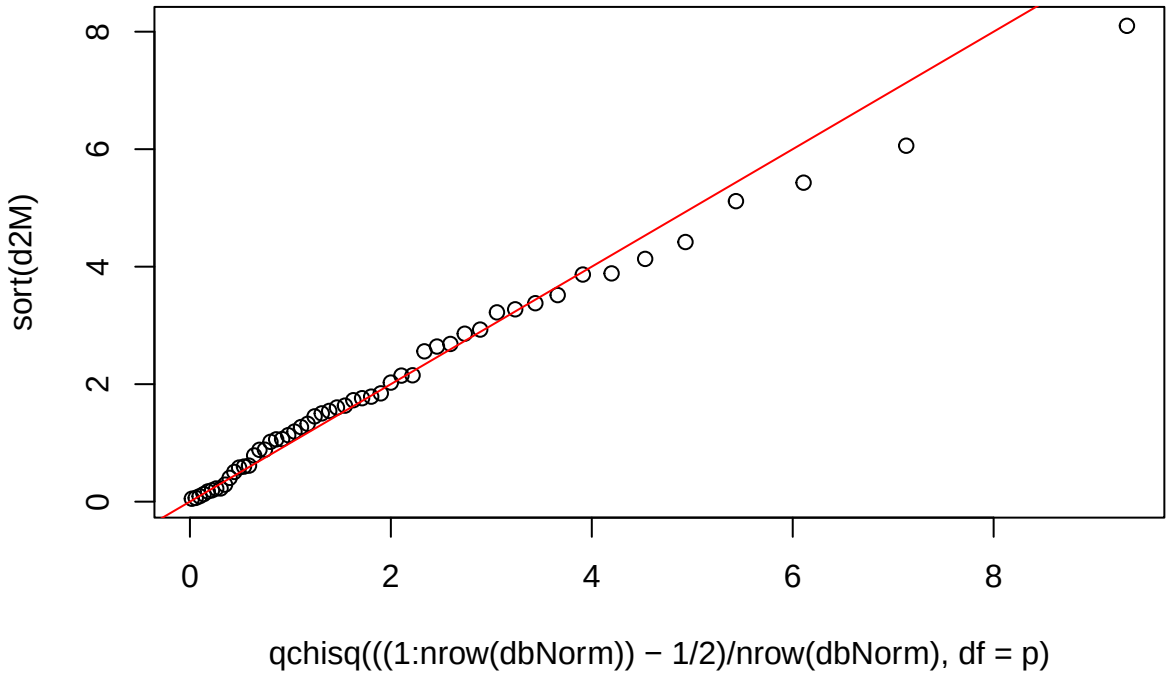
C. Haz la gráfica de contorno de la normal multivariada obtenida en el inciso B.

A partir de este gráfico, podemos decir que existe una correlación negativa entre las variables ya que en caso contrario, si la correlación fuera nula, las curvas de nivel serían de forma circular y no elípticas. Además, al observar los contornos se puede notar un sesgo en la gráfica, sin embargo, las pruebas realizadas indican que si es multivariado.




```
## $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    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
```

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)



De acuerdo con la gráfica de QQ-Plot se puede notar que la gráfica de probabilidad normal tiene una asimetría negativa con sesgo a la izquierda ya que no se aproxima en su totalidad a la línea y = x. Además, se puede observar que este conjunto de datos no parece seguir una distribución normal específicamente en la cola que se encuentra a la derecha.

3.2 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.

A. Justifique por qué es adecuado el uso de componentes principales para analizar la base (haz uso de la matriz de correlaciones)

```
## [1] "Matriz de correlaciones:"
```

```
##             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
```

Es esencial el uso de componentes principales para analizar la base dado que se puede investigar el conjunto de datos multidimensionales para visualizar y analizar las correlaciones entre las variables para limitar el número de variables. Además, el uso de la matriz de correlaciones asegura que se le dé la misma importancia a todas las variables, ya que en la matriz de correlaciones todos los elementos de la diagonal son iguales a 1. De esta manera, se realizan los cálculos de los componentes sobre variables originales estandarizadas, es decir, que tienen media 0 y varianza 1.

B. Realiza el análisis de componentes principales y justifica el número de componentes principales apropiados para reducir la dimensión de la base

```
## [1] "Valores propios:"
```

```
## [1] 5.34590819 1.22090789 1.04253153 0.66786333 0.33571266 0.20893778 0.10725403
## [8] 0.05203127 0.01885332
```

```
## [1] "Vectores propios:"
```

```
##              [,1]        [,2]        [,3]        [,4]        [,5]         [,6]
##  [1,] -0.35136146 -0.40301855 -0.07586402  0.30359419  0.03194121  0.284360283
##  [2,] -0.33907420 -0.29786166 -0.07470140 -0.23236707 -0.82623084  0.054271109
##  [3,] -0.28306469 -0.56943030  0.02991336  0.37427137  0.32816132 -0.298278080
##  [4,] -0.28126962 -0.21524882 -0.06147214 -0.83056128  0.39488490 -0.099142969
##  [5,]  0.39890941 -0.32518645 -0.05648045 -0.04980219 -0.06539303  0.004765464
##  [6,]  0.02398876  0.06261499 -0.96994179  0.05149024  0.09004998  0.149954574
##  [7,]  0.36905050 -0.37647100  0.11743644 -0.11401063  0.10565624  0.489107573
##  [8,]  0.37957032 -0.24428857 -0.16175615 -0.02767633 -0.16523448 -0.711214479
##  [9,]  0.40293860 -0.25922456  0.00756517 -0.07091614 -0.04298253  0.223233955
##              [,7]         [,8]         [,9]
##  [1,]  0.72620919 -0.082971700  0.007161703
##  [2,] -0.22348526  0.009782475 -0.032988603
##  [3,] -0.48766992  0.140957430 -0.017292418
##  [4,]  0.11144724  0.043959526  0.028777382
##  [5,]  0.01398475 -0.053416125  0.849768758
##  [6,] -0.14013431 -0.011952152 -0.041106334
##  [7,] -0.22360542 -0.528271290 -0.340326567
##  [8,]  0.30736177 -0.211913074 -0.311145559
##  [9,]  0.09015694  0.802648566 -0.247594211
```

```
## [1] "Lambdas: "
```

```
## $values
## [1] 5.34590819 1.22090789 1.04253153 0.66786333 0.33571266 0.20893778 0.10725403
## [8] 0.05203127 0.01885332

## [1] "Varianza total:"

## [1] 9

## [1] "Proporción de varianza explicada por cada componente: "

## [1] 0.593989799 0.135656432 0.115836836 0.074207036 0.037301407 0.023215309
## [7] 0.011917115 0.005781252 0.002094814

## [1] "Cumsum: "

## [1] 0.5939898 0.7296462 0.8454831 0.9196901 0.9569915 0.9802068 0.9921239
## [8] 0.9979052 1.0000000

## 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 los valores de la proporción de varianza (PC1 = 0.592 y PC2 = 0.7297) y la proporción acumulada (0.7297) se utilizarán los primeros dos componentes principales.

El primer componente principal de la matriz de correlaciones sería:

Y1 = - 0.35136146*X1* - *0.40301855*X2 - 0.07586402*X3* + *0.30359419*X4 + 0.03194121*X5* + *0.284360283*X6 + 0.72620919*X7* - *0.082971700*X8 + 0.007161703*X9

Las variables que más contribuyen al primer componente principal son X7 (mínimo de la concentración de mercurio en cada grupo de peces), X2 (PH), X1 (alcalinidad), X4 (clorofila) y X6 (número de peces estudiados en el lago).

El segundo componente principal de la matriz de correlaciones sería:

Y2 = - 0.33907420*X1* - *0.29786166*X2 - 0.07470140*X3* - *0.23236707*X4 - 0.82623084*X5* + *0.054271109*X6 - 0.22348526*X7* + *0.009782475*X8 - 0.032988603*X9

Las variables que más contribuyen al segundo componente principal son: X5 (concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago), X1 (alcalinidad), X2 (PH), X4 (clorofila) y X7 (mínimo de la concentración de mercurio en cada grupo de peces).

Las variables más importantes de cada componente se obtienen observando los coeficientes de cada variable y en valor absoluto los que sean mayores son los que contribuyen más a la determinación de dicho componente principal.

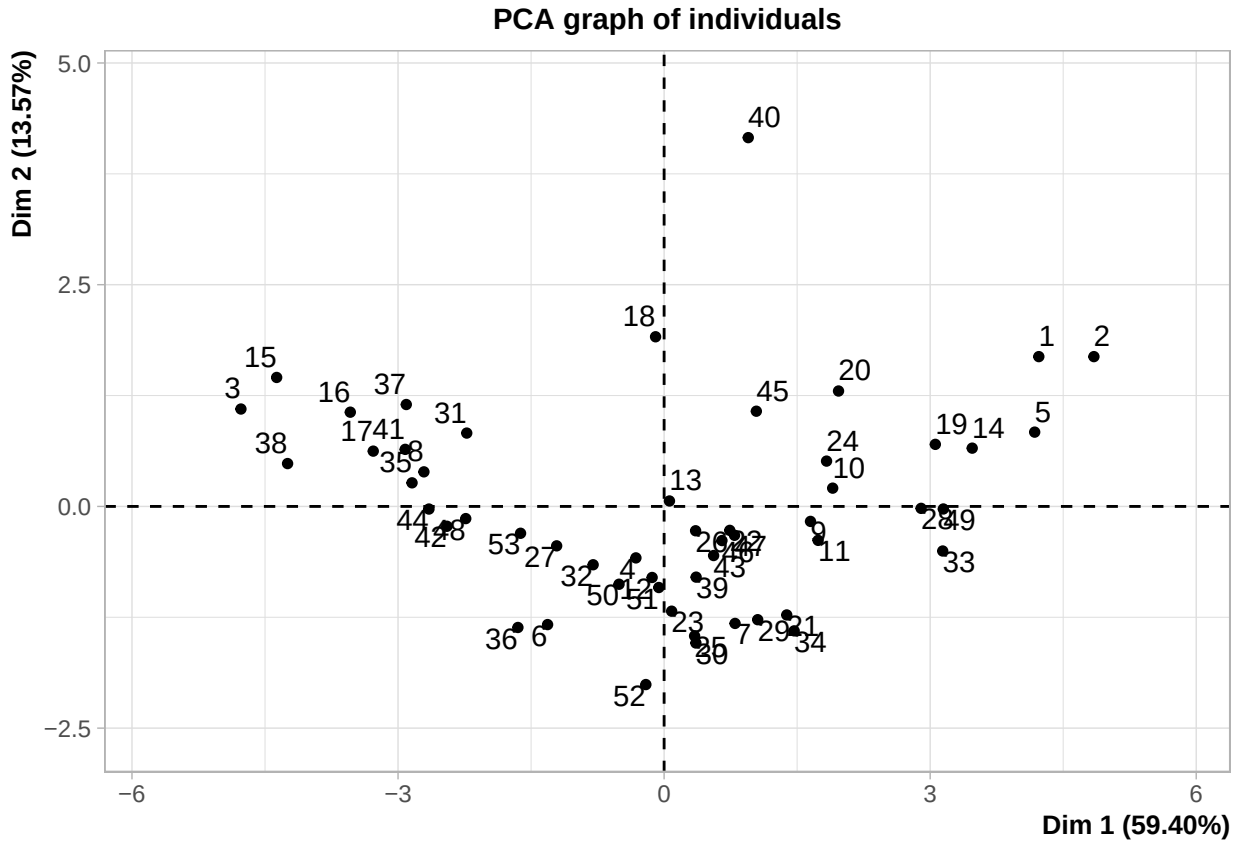
Además, se puede observar que la proporción de la varianza del primer componente es de 0.593989799 y del segundo es 0.135656432 lo que indica el porcentaje de contribución de cada componente principal. También, utilizando los valores obtenidos para la proporción acumulada se tiene que los primeros dos componentes agrupan un 72.96% de la variación.

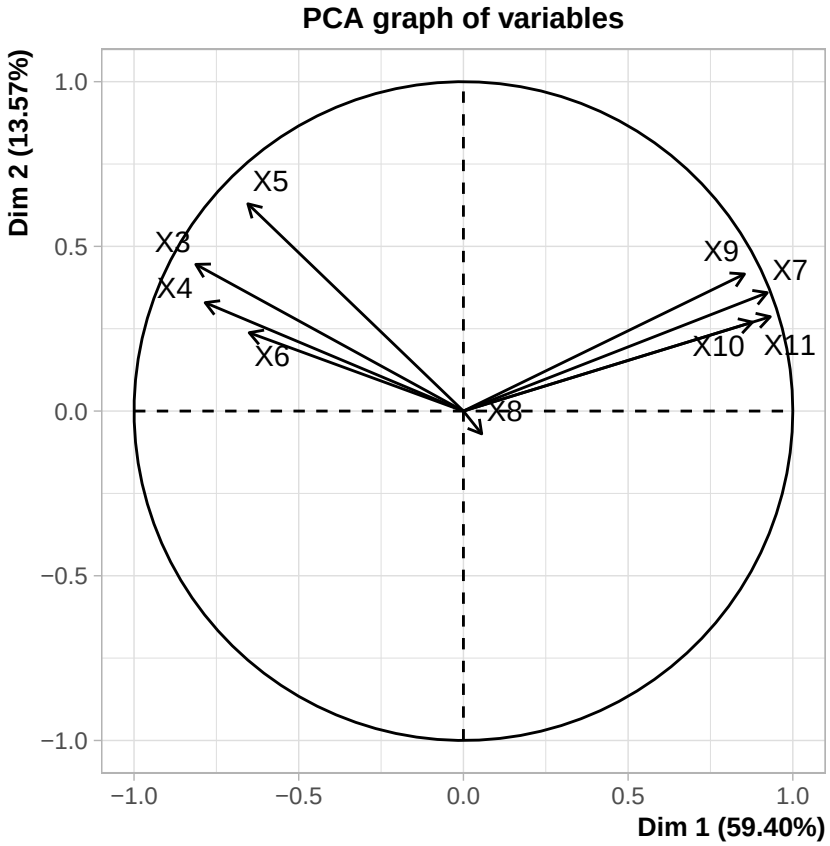
C. Representa en un gráfico los vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes

El análisis de PCA se realizará al grupo de las variables X3 a X11 ya que son las variables numéricas para realizar un análisis exploratorio de los datos. De esta manera podremos visualizar de mejor manera la variación que existe entre las diferentes variables.

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

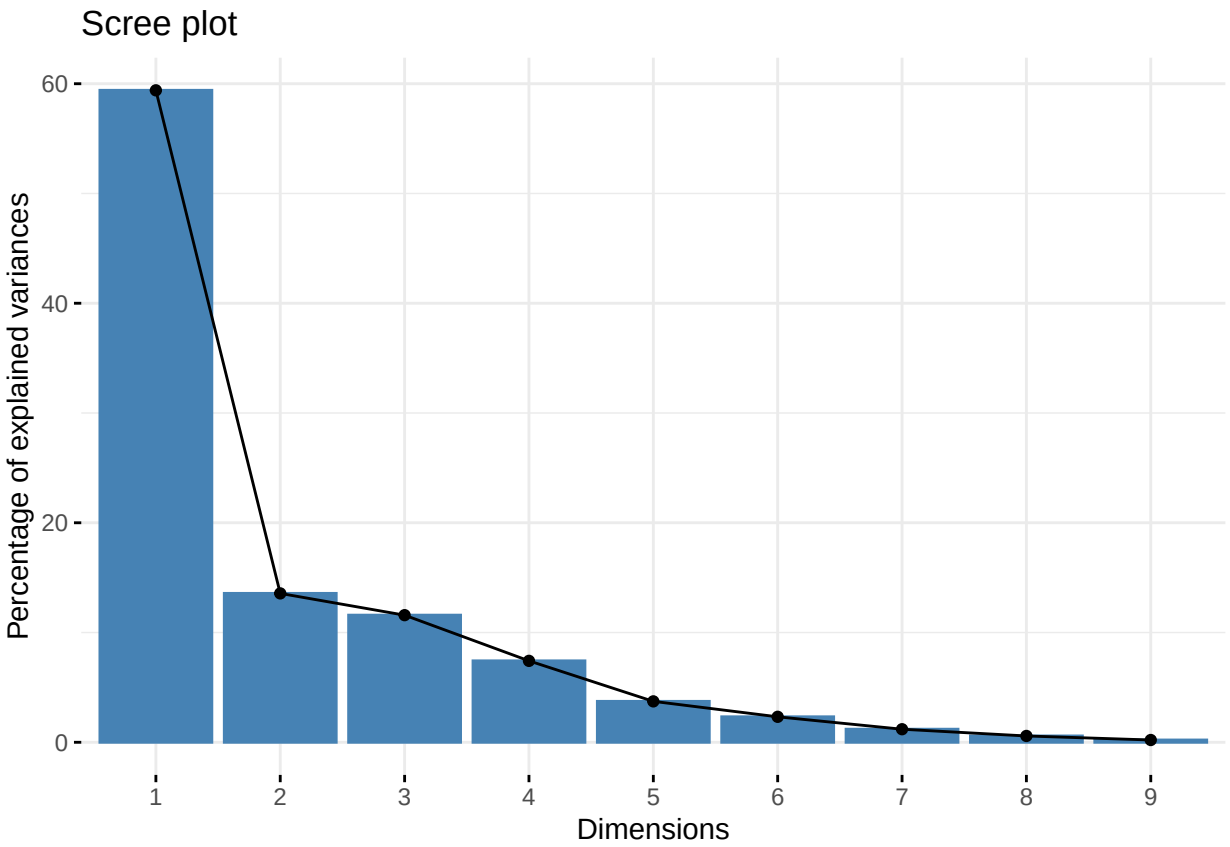
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```



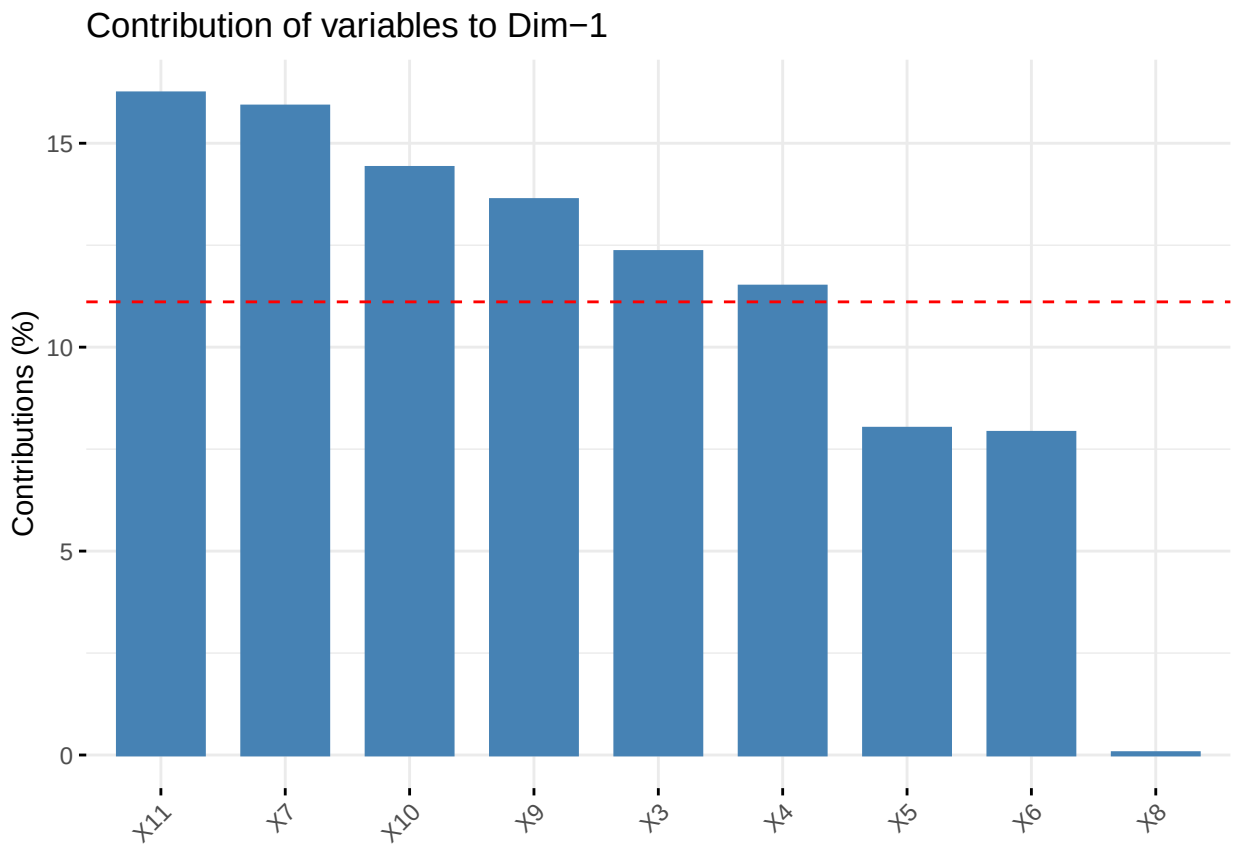


En la primera gráfica (PCA de los individuales) se muestran las observaciones en relación de los primeros dos componentes principales. Por lo tanto, se muestra que el componente principal 1 logra explicar una mayor cantidad de la variación ya que las observaciones se esparsen a lo largo del eje x; mientras que en el eje y que representa la dimensión del componente principal 2 no se nota lo mismo.

En la segunda gráfica (PCA por variables) se muestra de mejor manera que para el componente principal 1 se logra explicar el 59.40 % de la variación mientras que en el componente principal 2 solo se explica el 13.57%. Además, las variables de importancia para el primer componente son X7, X9, X10 y X11 con valores positivos y X3, X4, X5 y X6 con valores negativos; mientras que la variables más importantes para el componente 2 son X3, X4, X5, X6, X7, X9, X10 y X11 con valores positivos y X8 con valor negativo.



Asimismo, en el screeplot, se observa que el porcentaje de las varianzas explicadas para el componente 1 es del 59.40%, el porcentaje en el componente 2 es de 13.57% y se muestra que los componentes principales 3 y 4 también explican gran cantidad de la variabilidad.



Por último, en la gráfica de contribución de variables del componente principal 1 se muestra una línea roja que indica el valor medio de contribución. Por lo tanto, las variables con una contribución mayor a este límite son importantes para este componente. Entonces las variables X11, X7, X10, X9, X3 y X4 son las que más contribuyen al componente principal 1.

D. Explique brevemente a qué conclusiones llega con su análisis y qué significado tienen los componentes seleccionados en el contexto del problema

Después de haber realizado el análisis de PCA se pued concluir que las variables más significativas en el análisis de normalidad son el X4 y el X10. Mientras que en el análisis de componentes principales son las variables X3, X4, X7, X9, X10 y X11.

4. Conclusión

En el análisis realizado anteriormente se obtuvo que las variables más relevantes para contestar las preguntas seleccionadas son X3 (alcalinidad), X5 (calcio), X6 (clorofila) y X7 (concentración media de mercurio). Sin embargo, el análisis realizado para esta entrega difiere un poco con los resultados anteriores dado que se realizaron para toda la base de datos mientras que en el análisis pasado nosotros seleccionamos cuáles variables utilizar para los modelos. Considero que este nuevo análisis fue más completo ya que se estandarizaron los datos y se le dio la misma importancia a todas las variables de modo que generaron resultados más precisos. Teniendo en cuenta la pregunta principal, los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida fueron X3 (alcalinidad), X4 (PH), X7 (concentración media de mercurio), X9 (mínimo de concentración de mercurio), X10 (máximo de concentración de mercurio) y

X11 (estimación de la concentración de mercurio en el pez de 3 años). Por un lado, el estudio de normalidad nos permitió visualizar cuánto difiere la distribución de los datos respecto a lo que esperamos, de esta forma se logra observar qué variables tienen mejor precisión y fiabilidad. Además, para este análisis fue de suma importancia la creación de componentes principales ya que fue un método que ayudó a simplificar la complejidad del espacio muestral con varias dimensiones para poder seleccionar las variables que son más significativas.

5. Referencias bibliográficas

Korkmaz, S. (2022, October 12). Package 'MVN.' MVN: Multivariate Normality Tests. Retrieved December 1, 2022, from https://cran.r-project.org/web/packages/MVN/MVN.pdf

Amat, J. (2017, June). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE: RPubs. Retrieved December 1, 2022, from https://rpubs.com/Joaquin_AR/287787

6. Anexos

https://github.com/A01750185/E2-PortafolioImplementacion.git