# Procesamiento de datos multivariados

Módulo 1: Estadística Avanzada para la ciencia de datos. Inteligencia Artificial Avanzada para la Ciencia de Datos II. Grupo 502

Andrea Piñeiro

2022-10-23

## Resumen

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 este reporte se realizó un análisis de normalidad multivariada en las variables, así como un análisis univariado para verificar que variables son normales.

Posteriormente se realizó un análisis de componentes principales para encontrar los factores con mayor influencia en la contaminación de mercurio y de esta forma reducir la dimensionalidad del problema.

## Introducción

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud y hay límites en cuanto a los niveles máximos de Hg de mercurio.

En este reporte se realizará: - Un análisis de normalidad de las variables continuas para identificar variables normales con el objetivo de ver la diferencia que hay entre la distribución esperada y la observada. - Un análisis de componentes principales con la base de datos completa para identificar los factores principales que intervienen en el problema de la contaminación de mercurio. - Emitir una conclusión general unida a la realizada previamente.

## Análisis de los resultados

### Análisis de normalidad

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

El objetivo del análisis de normalidad es análizar la diferencia que hay entre los datos y una distribución normal.

#### A. Prueba de normalidad de Mardia y la prueba de Anderson Darling

La prueba de Mardia determina si un grupo de variables sigue una distribución normal multivariante.

$H_0$ : Las variables siguen una distribución normal multivariante. $H_1$ : Las variables no siguen una distribución normal multivariante. $\alpha$ : 0.05

Tenemos un total de 10 variables (excluyendo la variable de id ya que esta no es necesaria y la variable de edad de peces ya que es una variable categórica).

```
##    alcalinidad  ph calcio clorofila concentración_mercurio peces
## 1          5.9 6.1    3.0       0.7                   1.23     5
## 2          3.5 5.1    1.9       3.2                   1.33     7
## 3        116.0 9.1   44.1     128.3                   0.04     6
## 4         39.4 6.9   16.4       3.5                   0.44    12
## 5          2.5 4.6    2.9       1.8                   1.20    12
## 6         19.6 7.3    4.5      44.1                   0.27    14
## 7          5.2 5.4    2.8       3.4                   0.48    10
## 8         71.4 8.1   55.2      33.7                   0.19    12
## 9         26.4 5.8    9.2       1.6                   0.83    24
## 10         4.8 6.4    4.6      22.5                   0.81    12
## 11         6.6 5.4    2.7      14.9                   0.71    12
## 12        16.5 7.2   13.8       4.0                   0.50    12
## 13        25.4 7.2   25.2      11.6                   0.49     7
## 14         7.1 5.8    5.2       5.8                   1.16    43
## 15       128.0 7.6   86.5      71.1                   0.05    11
## 16        83.7 8.2   66.5      78.6                   0.15    10
## 17       108.5 8.7   35.6      80.1                   0.19    40
## 18        61.3 7.8   57.4      13.9                   0.77     6
## 19         6.4 5.8    4.0       4.6                   1.08    10
## 20        31.0 6.7   15.0      17.0                   0.98     6
## 21         7.5 4.4    2.0       9.6                   0.63    12
## 22        17.3 6.7   10.7       9.5                   0.56    12
## 23        12.6 6.1    3.7      21.0                   0.41    12
## 24         7.0 6.9    6.3      32.1                   0.73    12
## 25        10.5 5.5    6.3       1.6                   0.34    10
## 26        30.0 6.9   13.9      21.5                   0.59    36
## 27        55.4 7.3   15.9      24.7                   0.34    10
## 28         3.9 4.5    3.3       7.0                   0.84     8
## 29         5.5 4.8    1.7      14.8                   0.50    11
## 30         6.3 5.8    3.3       0.7                   0.34    10
## 31        67.0 7.8   58.6      43.8                   0.28    10
## 32        28.8 7.4   10.2      32.7                   0.34    10
## 33         5.8 3.6    1.6       3.2                   0.87    12
## 34         4.5 4.4    1.1       3.2                   0.56    13
## 35       119.1 7.9   38.4      16.1                   0.17    12
## 36        25.4 7.1    8.8      45.2                   0.18    13
## 37       106.5 6.8   90.7      16.5                   0.19    13
## 38        53.0 8.4   45.6     152.4                   0.04     4
## 39         8.5 7.0    2.5      12.8                   0.49    12
## 40        87.6 7.5   85.5      20.1                   1.10    10
## 41       114.0 7.0   72.6       6.4                   0.16    14
## 42        97.5 6.8   45.5       6.2                   0.10    12
## 43        11.8 5.9   24.2       1.6                   0.48    10
## 44        66.5 8.3   26.0      68.2                   0.21    12
## 45        16.0 6.7   41.2      24.1                   0.86    12
## 46         5.0 6.2   23.6       9.6                   0.52    12
## 47        25.6 6.2   12.6      27.7                   0.65    44
## 48        81.5 8.9   20.5       9.6                   0.27     6
## 49         1.2 4.3    2.1       6.4                   0.94    10
## 50        34.0 7.0   13.1       4.6                   0.40    12
```

```
## 51        15.5 6.9    5.2      16.5                  0.43    11
## 52        17.3 5.2    3.0       2.6                  0.25    12
## 53        71.8 7.9   20.5       8.8                  0.27    12
##    min_concentración max_concentración estimación_3_años
## 1               0.85              1.43              1.53
## 2               0.92              1.90              1.33
## 3               0.04              0.06              0.04
## 4               0.13              0.84              0.44
## 5               0.69              1.50              1.33
## 6               0.04              0.48              0.25
## 7               0.30              0.72              0.45
## 8               0.08              0.38              0.16
## 9               0.26              1.40              0.72
## 10              0.41              1.47              0.81
## 11              0.52              0.86              0.71
## 12              0.10              0.73              0.51
## 13              0.26              1.01              0.54
## 14              0.50              2.03              1.00
## 15              0.04              0.11              0.05
## 16              0.12              0.18              0.15
## 17              0.07              0.43              0.19
## 18              0.32              1.50              0.49
## 19              0.64              1.33              1.02
## 20              0.67              1.44              0.70
## 21              0.33              0.93              0.45
## 22              0.37              0.94              0.59
## 23              0.25              0.61              0.41
## 24              0.33              2.04              0.81
## 25              0.25              0.62              0.42
## 26              0.23              1.12              0.53
## 27              0.17              0.52              0.31
## 28              0.59              1.38              0.87
## 29              0.31              0.84              0.50
## 30              0.19              0.69              0.47
## 31              0.16              0.59              0.25
## 32              0.16              0.65              0.41
## 33              0.31              1.90              0.87
## 34              0.25              1.02              0.56
## 35              0.07              0.30              0.16
## 36              0.09              0.29              0.16
## 37              0.05              0.37              0.23
## 38              0.04              0.06              0.04
## 39              0.31              0.63              0.56
## 40              0.79              1.41              0.89
## 41              0.04              0.26              0.18
## 42              0.05              0.26              0.19
## 43              0.27              1.05              0.44
## 44              0.05              0.48              0.16
## 45              0.36              1.40              0.67
## 46              0.31              0.95              0.55
## 47              0.30              1.10              0.58
## 48              0.04              0.40              0.27
## 49              0.59              1.24              0.98
## 50              0.08              0.90              0.31
```

```
## 51              0.23              0.69              0.43
## 52              0.15              0.40              0.28
## 53              0.15              0.51              0.25
```

```
##              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
```

Según Mardia no hay normalidad en los datos, debido a que el p value es menor a nuestra $\alpha$ rechazamos la hipótesis nula.

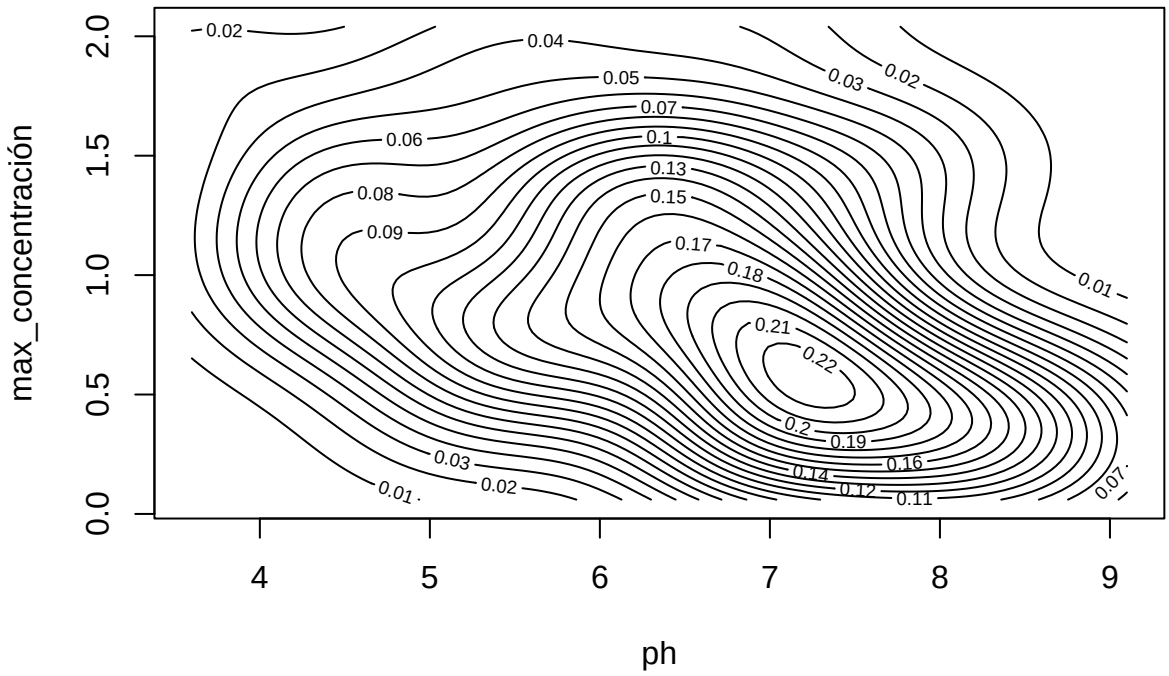
Debido a que no existe normalidad multivariada, procedemos a analizar la normalidad univariada.Para observar si hay variables que si tengan una distribución normal.

```
##               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            4.0510  <0.001      NO
## 4 Anderson-Darling       clorofila           5.4286  <0.001      NO
## 5 Anderson-Darling concentración_mercurio    0.9253  0.0174      NO
## 6 Anderson-Darling         peces             8.6943  <0.001      NO
## 7 Anderson-Darling   min_concentración       1.9770  <0.001      NO
## 8 Anderson-Darling   max_concentración       0.6585   0.081      YES
## 9 Anderson-Darling   estimación_3_años       1.0469  0.0086      NO
```

Según los resultados de Anderson Darling las variables que tienen un nivel de significancia mayor a 0.05 no tienen evidencia para rechazar la hipótesis nula, por lo que podemos decir que los datos siguen una distribución normal.

Las variables con distribución normal son: - ph - max_concentración

**B. Prueba de normalidad de Mardia y la prueba de Anderson Darling con variables con normalidad y C. Contorno de la normal multivariada**



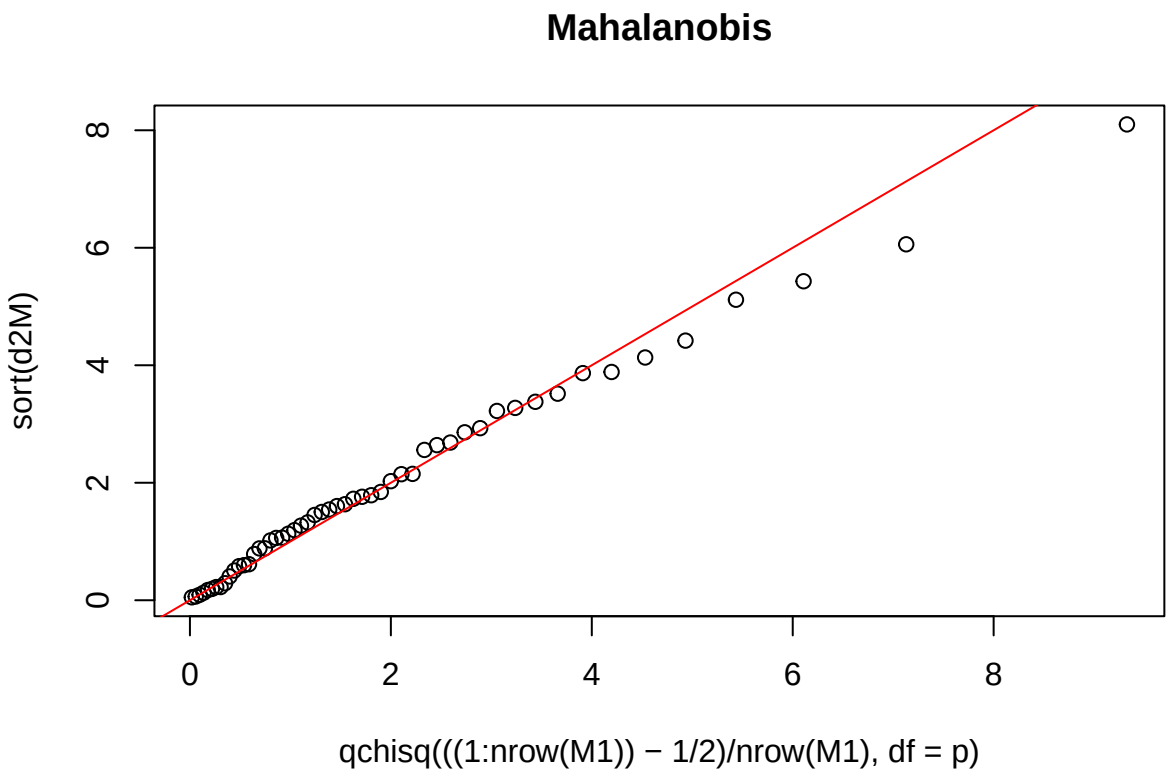

```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  6.17538668676458 0.186427564928852    YES
## 2 Mardia Kurtosis -1.12820795824432  0.25923210375991    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##               Test          Variable Statistic   p value Normality
## 1 Anderson-Darling        ph            0.3496    0.4611    YES
## 2 Anderson-Darling max_concentración    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_concentración 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925
##                      Kurtosis
## ph                 -0.6239638
## max_concentración -0.6692490
```

Como se puede observar tenemos los mismos valores de p en la prueba de Anderson que los obtenidos anteriormente.

Sin embargo, ahora ambos valores para el sesgo y la curtosis de Mardia son mayores a 0.05 por lo que no podemos rechazar la hipótesis nula y podemos decir que existe normalidad multivariada.

En la gráfica de los contornos podemos observar que las variables no tienen correlación por que no están centrados en 0, 0 ni tienen una forma circular

**D. Detecta datos atípicos o influyentes en la normal multivariada encontrada en el inciso B**




La distancia de Mahalanobis nos ayuda a medir la distancia entre un punto y una distribución.

Como podemos observar a medida que aumenta x también lo hace y por lo que no tenemos valores atípicos.

El único outlier que podríamos remover es el más alejado que se muestra en la esquina derecha superior.

## 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.

**A. Por qué es adecuado el uso de componentes principales para analizar la base**

Es importante usar los componentes principales ya que al existir muchas variables, nos ayuda a seleccionar las características más importantes. Con los componentes principales reducimos el conjunto de variables, quedandonos con las que ya no guardan correlación entre si y permitiéndonos reducir la dimensionalidad del problema.

Se usa la matriz de correlaciones debido a que con ella no tendremos problemas con las diferentes unidades de las variables.

```
##
##
## Matriz de correlaciones
```

**B. Análisis de componentes principales y justifica el número de componentes principales apropiados**

A continuación obtenemos los valores y los vectores de Eigen. Se usa la matiz de correlación ya que en esta todas las variables están en el mismo rango por lo que no habría diferencia en cuanto a unidades.

```
##
##
## Valores y vectores propios de la correlación
```

```
## eigen() decomposition
## $values
## [1] 5.34590819 1.22090789 1.04253153 0.66786333 0.33571266 0.20893778 0.10725403
## [8] 0.05203127 0.01885332
##
## $vectors
##              [,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
```

Para el componente 1 las variables que más influen son la 9 y 5 (seguidas de la 8 y 7), mientras que para el componente 2 las variables que más influyen son la 6 y la 10.

Calculamos la proporción de varianza explicada por cada componente

```
##
##
## Proporción de varianza explicada de la matriz de correlación
```

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

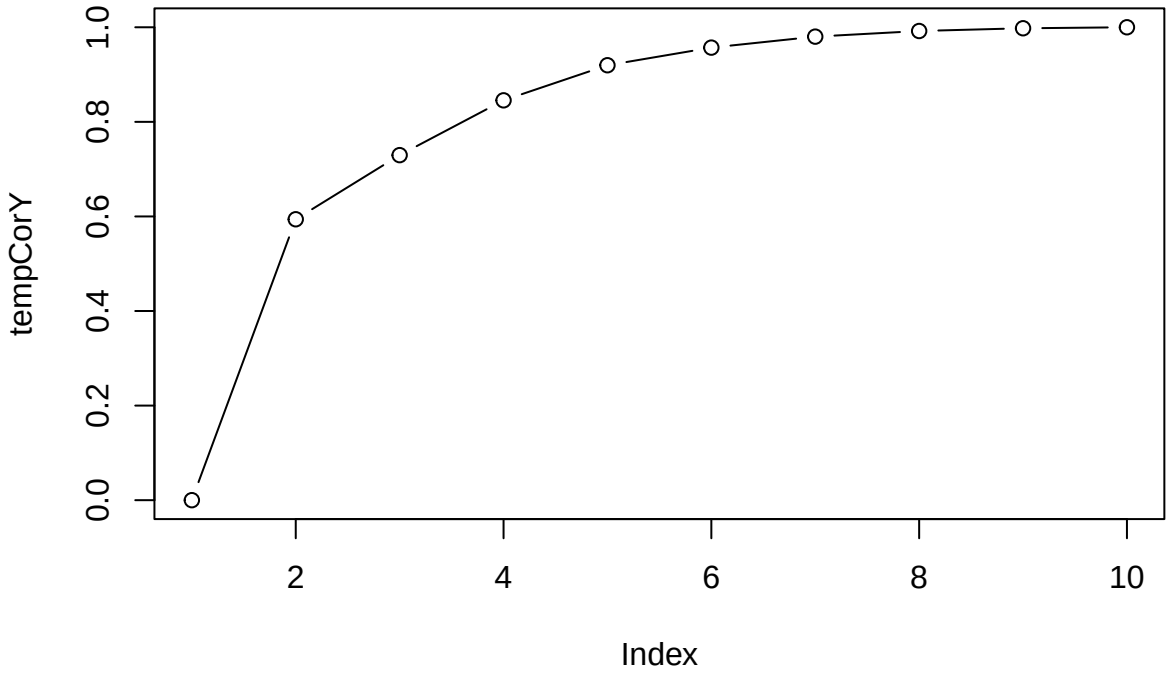
Acumulamos los resultados para visualizarlo de mejor manera

```
##
##
## Acumulativo de proporción de varianza para la matriz correlación
```

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

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

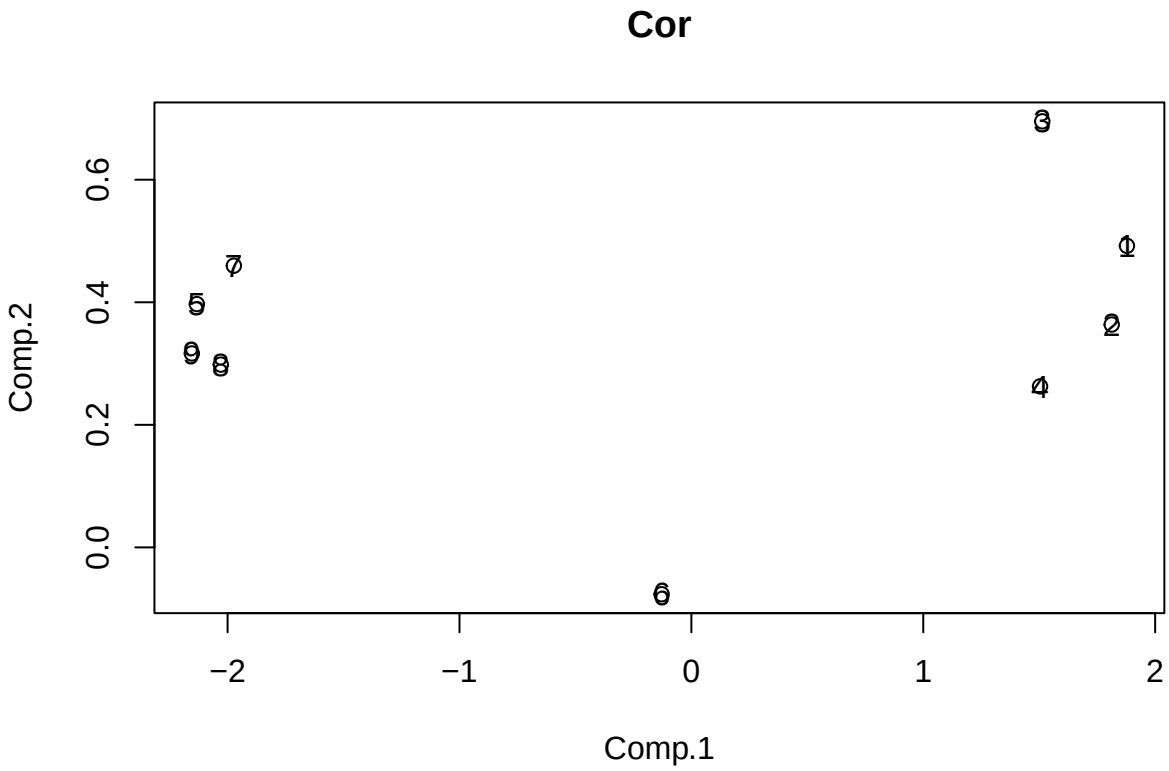
A continuación podemos ver las puntuaciones de los componentes principales y su nivel de explicabilidad.

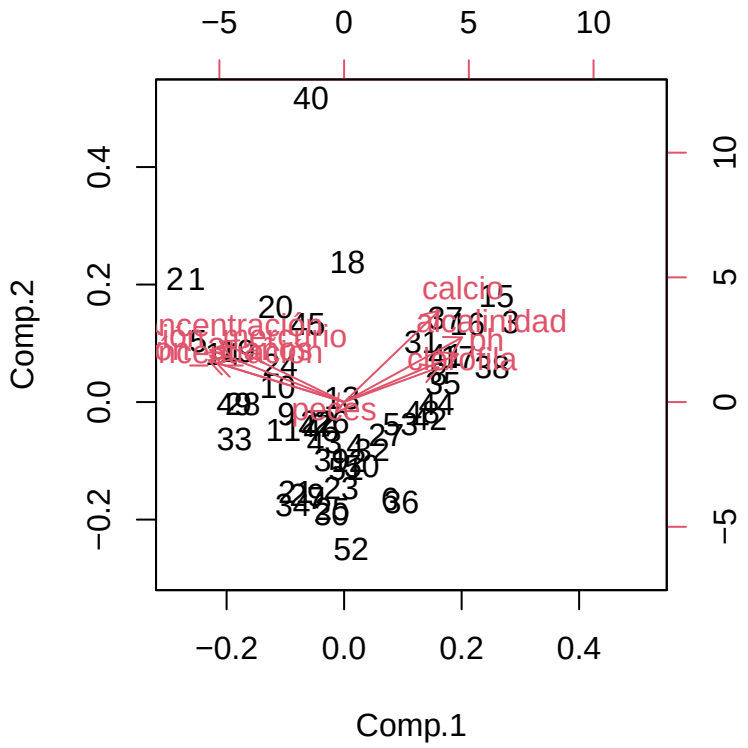



Como se puede observar llegamos a un nivel de explicación mayor al 90% hasta el 5° componente principal. Los dos primeros componentes explican el 72% de los datos, por lo que no es un muy buen resultado.

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

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


Cor
Comp.2
0.0
0.2
0.4
0.6
−2
−1
0
1
2
Comp.1

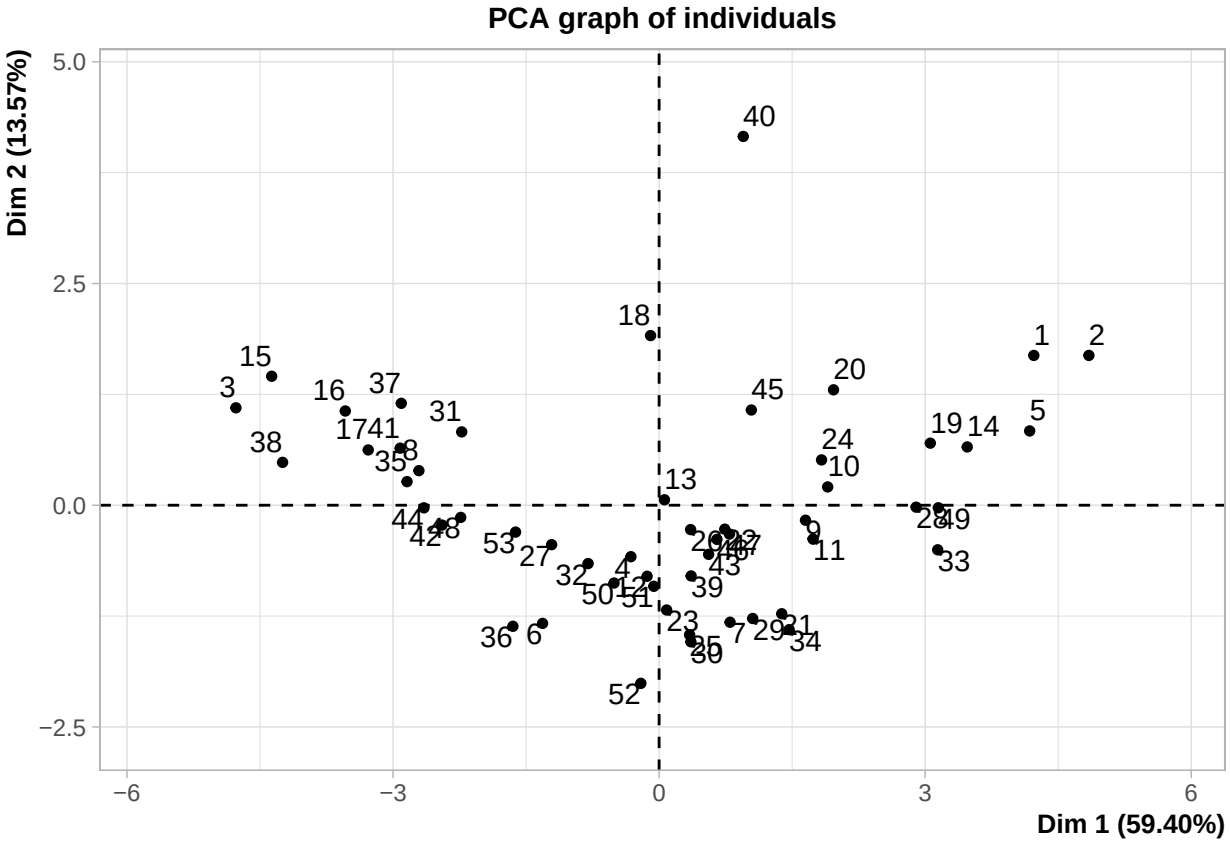


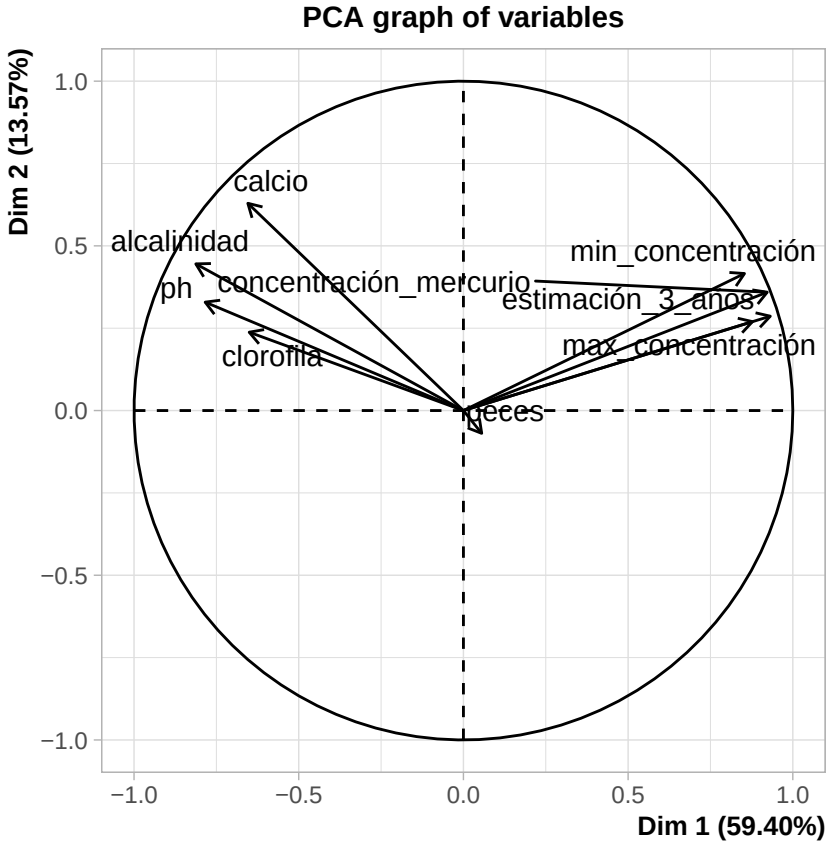

Como podemos observar en la primera gráfica, tenemos un poco de dispersión en los datos por lo que no se segmentan en grupos de manera muy definida, aunque si podemos ver un poco la segmentación.

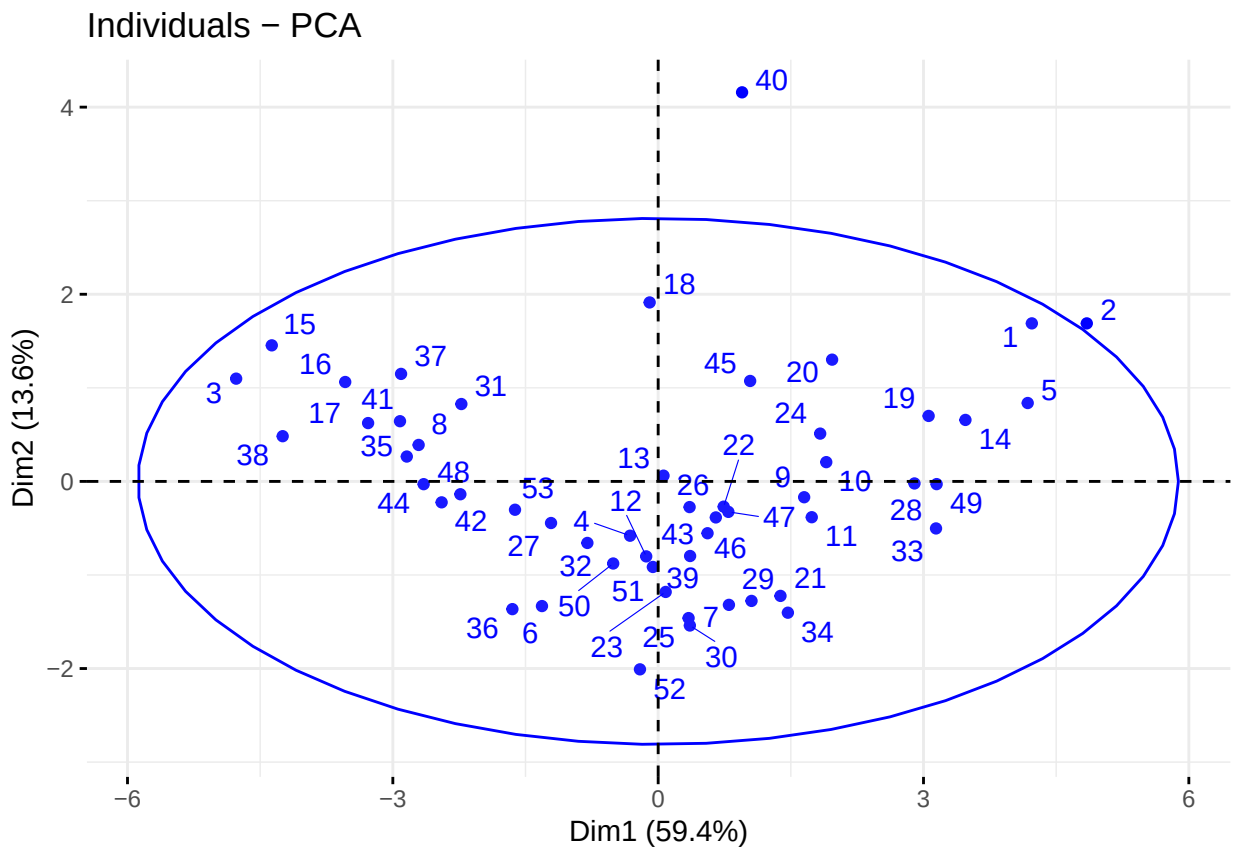
En la segunda gráfica podemos identificar cuales son las variables que tienen más efecto en los componentes principales. Es difícil saber cuales son las de mayor efecto pero parece que son alcalinidad, calcio y peces, aunque es difícil con este diagrama verificar si si son las correctas, esto lo haremos más adelante.

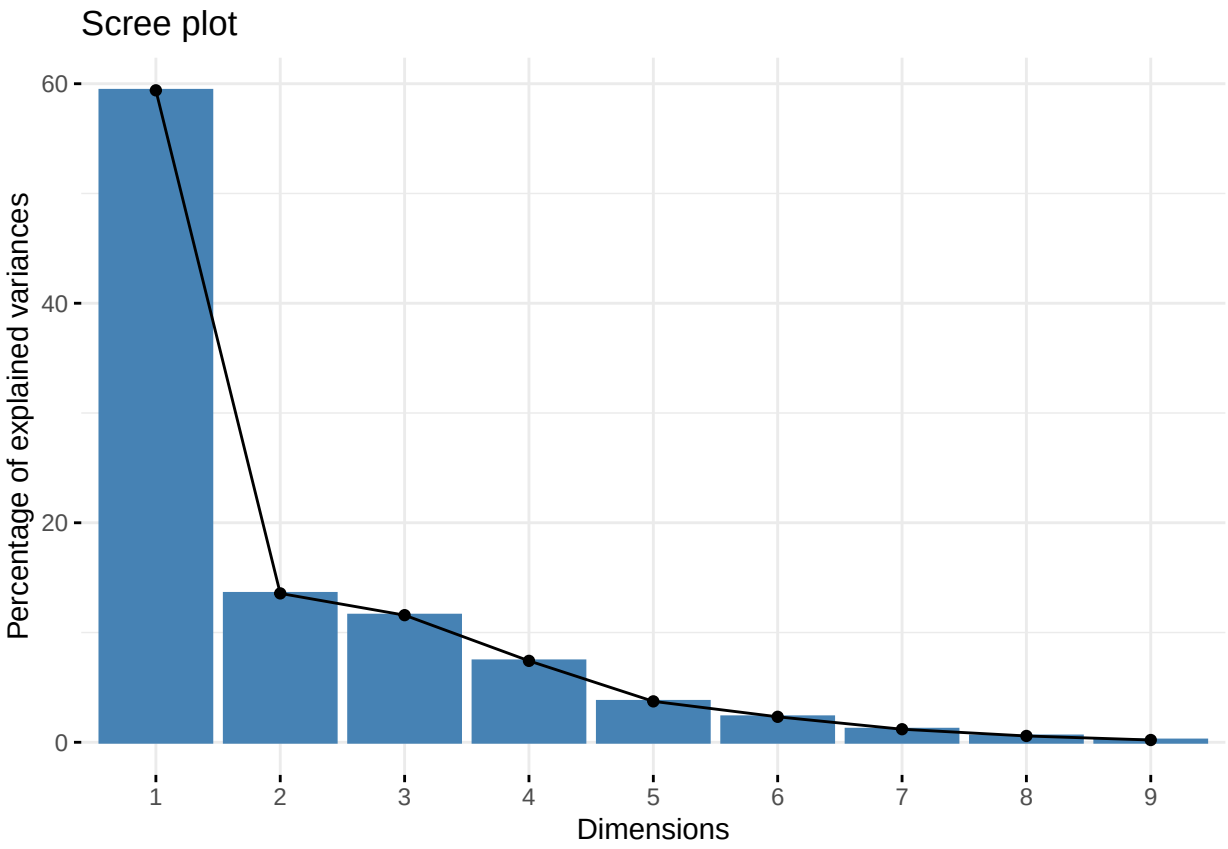
Si observamos la gráfica podemos ver que hay muchas variables en la izquierda que no se alcanza a visualizar muy bien si tienen influencia en los datos.

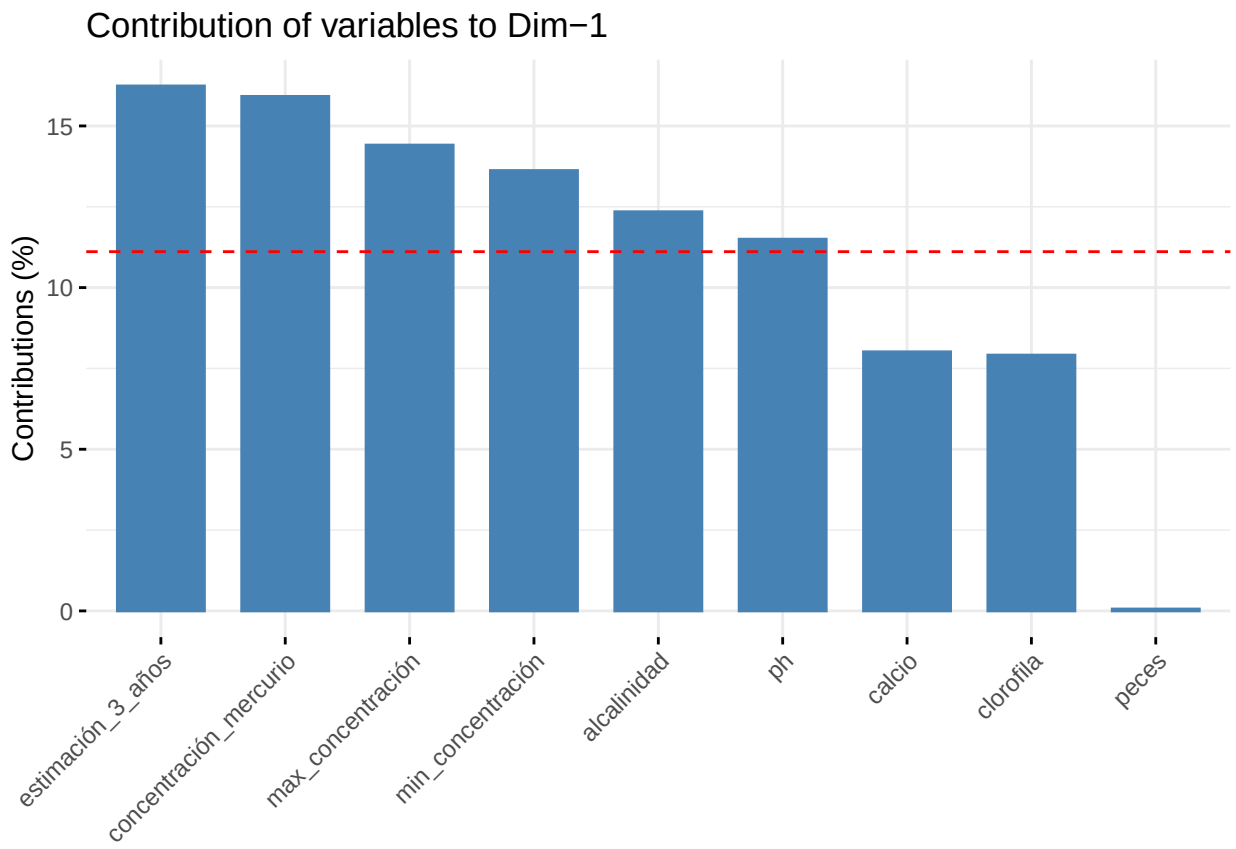
**D. Interprete los resultados.**




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
50
51
36
6
23
25
52
30
7
39
29
21
34
13
22
26
9
10
12
47
43
46
11
45
20
24
19
5
14
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




En esta parte estamos realizando PCA (Principal Component Analysis). Con esto obtenemos un análisis mucho más completo y realizado de manera mucho más sencilla.

En la **gráfica 1** podemos observar que los datos están un poco dispersos y no es fácil agruparles o segmentarles, por lo que la explicabilidad de los datos no es muy buena.

En la **gráfica 2** podemos observar que las variables que contribuyen de mayor manera a los componentes principales son: * Peces * Alcalinidad * Ph * Calcio

En la **gráfica 3** podemos observar una elipse en la gráfica con la dispersión de los datos, lo que nos indica la variablidad de los datos y podemos ver que hay muy pocos datos que salen fuera de la elipse.

En la **gráfica 4** podemos observar que el componente 1 y 2 explican poco más del 70% de los datos.

Y finalmente en la **gráfica 5** podemos observar la contribución de las variables en el componente principal 1 de una mejor manera. Como podemos ver las variables que más contribuyen son las relacionadas con los niveles de mercurio, pero estas serían las posibles para ser usadas como variable dependiente. Las variables que siguen en contribución al modelo son alcalinidad y ph como mayores, seguidos de calcio y clorofila. La edad de los peces y los peces finalmente tienen una contribución muy pequeña.

## Conclusión

Después del análisis podemos concluir que los factores que influyen en los niveles de mercurio en el agua son los niveles de alcalinidad y ph encontrados en ella.

En el análisis multivariado pudimos observar que las variables no eran normales; por eso realizamos el análisis univariado y sacamos las variables que resultaron normales. Después podemos aplicar PCA solo a las variables que son normales; sin embargo, como solo nos quedaron 2 variables multivariadas no nos sirve aplicar PCA en ellas, por lo que lo realizamos sin la normalidad e intentamos reducir la dimensionalidad.

De las gráficas podemos concluir que el componente principal 1 segmenta dos grupos claramente, hacia la derecha se encuentra la alcalinidad, clorofila, ph y calcio, y en la izquierda se encuentran las variables relacionadas con la concentración de mercurio. El componente principal 2 segmenta todas las anteriores y en otro grupo podemos observar la variable de número de peces..

Podemos concluir que con los dos primeros componentes ya tenemos forma de explicar todas las variables, por lo que si podríamos quedarnos con ellos y sería un buen análisis.

# Anexos

Código: https://drive.google.com/drive/folders/16z0d5gPp6IjctMk1Z4ugrgrs8-8U0hZ8?usp=sharing

# Bibliografía


- R: Mardia Test (Skewness and Kurtosis) for Multivariate... (2014). R-Project.org. https://search.r-project.org/CRAN/refmans/mvnormalTest/html/mardia.html
- Nistrup, P. (2019, January 29). Principal Component Analysis (PCA) 101, using R - Towards Data Science. Medium; Towards Data Science. https://towardsdatascience.com/principal-component-analysis-pca-101-using-r-361f4c53a9ff
- Normality Test in R: The Definitive Guide - Datanovia. (2019, November 30). Datanovia. https://www.datanovia.com/en/lessons/normality-test-in-r/