Procesamiento de Datos Multivariados

Módulo 5: Estadística Avanzada para ciencia de datos

Inteligencia Artificial Avanzada para la Ciencia de Datos.
Grupo 501

Autora
Emilia Victoria Jácome Iñiguez
A00828347

Docente:
Blanca Rosa Ruiz Hernandez

Tecnológico de Monterrey, Campus Monterrey
miércoles, 30 de Noviembre de 2022

Contents

Resumen

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 y el objetivo principal es hallar cuáles de estas variables son los factores más determinantes para las altas concentraciones de mercurio en los peces de los lagos de Florida, los cuales a su vez están provocando esta contaminación medioambiental. Tras los hallazgos y conclusiones obtenidas en el estudio previo, se determinó que el análisis de ANOVA no fue el modelo más apropiado para responder a la pregunta de investigación dado que las variables en cuestión poseen un alto grado de correlación entre sí. Por ende, en este estudio se procede a abordar la pregunta de investigación mediante la aplicación de un Análisis de Componentes Principales (PCA). Subsecuentemente, para poder abordar esta problemática, se hizo uso de varias herramientas estadísticas vistas durante el módulo 5 de la concentración Inteligencia Artificial Avanzada para la Ciencia de Datos. Principalmente, se aplicaron varias pruebas de normalidad multivariante a los datos en conjunto y un test de normalidad a cada variable por separado, la cuál sirvió para evaluar la tendencia de distribución de las variables. En segunda instancia, se procedió a aplicar un análisis de componentes principales (PCA) con la finalidad de simplificar el modelo reduciéndo la cantidad de variables en cuestión para no producir redundancia al utilizar variables fundamentando la selección de componentes en función de la varianza explicada acumulada.

Los principales resultados a los que se llegaron fueron:

- El conjunto de las 9 variables no posee normalidad multivariante.
- Las únicas variables que siguen una distribución normal corresponden a X4 y X9.
- Los componentes principales que explican el 85% de la varianza explicada del modelo corresponden a: PC1, PC2, y PC3.
- Las variables principales que producen mayor variación en cada componente, corresponden a:

  PC1:

  - X3: Alcalinidad
  - X4: PH
  - X7: Concentración media de mercurio
  - X9: Mínimo de concentración de mercurio en el lago
  - X10: Máximo de concentración de mercurio en el lago
  - X11: Estimación de concentración de mercurio.

  PC2:

  - X3: Alcalinidad
  - X4: PH
  - X5: Calcio
  - X7: Concentración media de mercurio
  - X9: Mínimo de concentración de mercurio en el lago

  PC3:

  - X8: Número de peces

Introducción

Al saber que las variables en cuestión están correlacionadas entre sí a través del estudio previo que se realizó donde se aplicó un modelo de ANOVA, se pretende indagar más a fondo sobre la interrelación de estas variables y averiguar si las variables en cuestión cumplen con la propiedad de normalidad multivariante. Adicionalmente, considerando que varias de ellas están fuertemente correlacionadas entre sí, a través de este estudio, se pretende construir un modelo donde se logre reducir el número de variables resumiéndolas en funciones lineales dependientes de varias variables. De modo que se utilizará la herramienta estadística de Análisis de Componentes principales para determinar las funciones estadísticas que explican mayormente la variación de los datos dentro del modelo, y, de esa manera, se procederá a dar una mejor aproximación a la pregunta de investigación que surge en este estudio:

Pregunta Base:

¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?

1. EXPLORACIÓN DE LA BASE DE DATOS

1.1 Descripción de cada variable:

- 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)

Identifica la cantidad de datos y variables presentes.

```
La cantidad de columnas es: 12


La cantidad de datos es de: 53
```

Clasificación de las variables de acuerdo a su tipo y escala de medición.

```
 X1 : int  1 2 3 4 5 6 7 8 9 10 ...
 X2 : chr  "Alligator" "Annie" "Apopka" "Blue Cypress" ...
 X3 : num  5.9 3.5 116 39.4 2.5 19.6 5.2 71.4 26.4 4.8 ...
 X4 : num  6.1 5.1 9.1 6.9 4.6 7.3 5.4 8.1 5.8 6.4 ...
 X5 : num  3 1.9 44.1 16.4 2.9 4.5 2.8 55.2 9.2 4.6 ...
 X6 : num  0.7 3.2 128.3 3.5 1.8 ...
 X7 : num  1.23 1.33 0.04 0.44 1.2 0.27 0.48 0.19 0.83 0.81 ...
 X8 : int  5 7 6 12 12 14 10 12 24 12 ...
 X9 : num  0.85 0.92 0.04 0.13 0.69 0.04 0.3 0.08 0.26 0.41 ...
 X10: num  1.43 1.9 0.06 0.84 1.5 0.48 0.72 0.38 1.4 1.47 ...
 X11: num  1.53 1.33 0.04 0.44 1.33 0.25 0.45 0.16 0.72 0.81 ...
 X12: int  1 0 0 0 1 1 1 1 1 1 ...
```

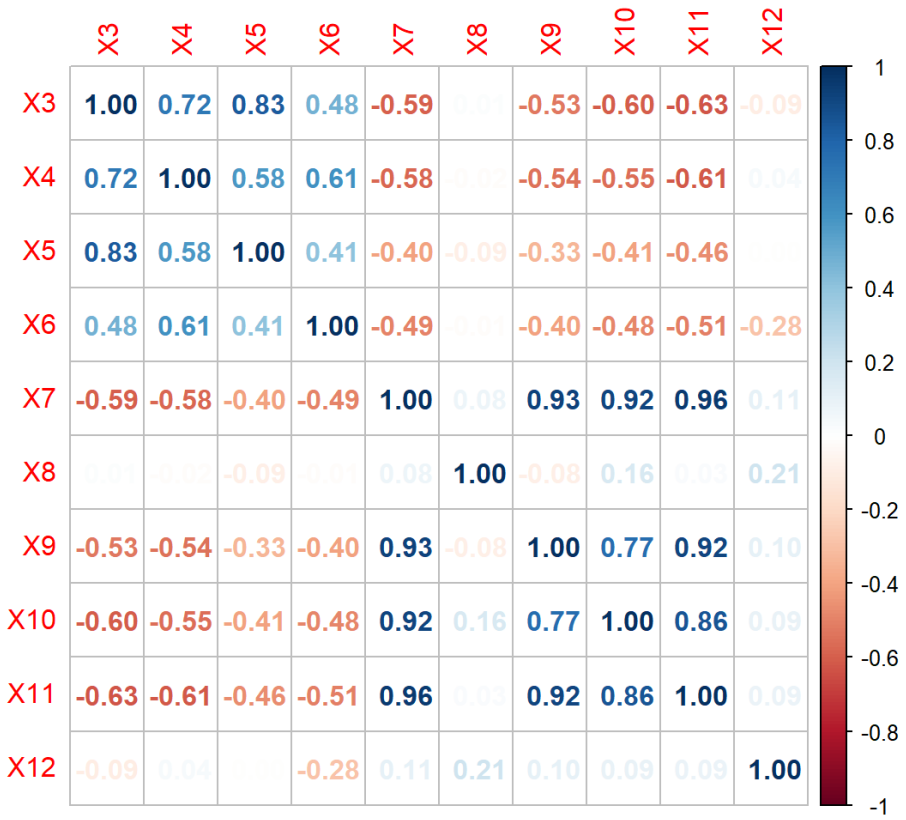
1.2 Exploración de la base de datos - Medidas estadísticas

Variables cuantitativas - Medidas de Tendencia Central

```
       X1             X2                  X3               X4              X5              X6        
 Min.   : 1   Length:53          Min.   :  1.20   Min.   :3.600   Min.   : 1.1   Min.   :  0.70  
 1st Qu.:14   Class :character   1st Qu.:  6.60   1st Qu.:5.800   1st Qu.: 3.3   1st Qu.:  4.60  
 Median :27   Mode  :character   Median : 19.60   Median :6.800   Median :12.6   Median : 12.80  
 Mean   :27                      Mean   : 37.53   Mean   :6.591   Mean   :22.2   Mean   : 23.12  
 3rd Qu.:40                      3rd Qu.: 66.50   3rd Qu.:7.400   3rd Qu.:35.6   3rd Qu.: 24.70  
 Max.   :53                      Max.   :128.00   Max.   :9.100   Max.   :90.7   Max.   :152.40  
       X7                X8              X9               X10              X11              X12        
 Min.   :0.0400   Min.   : 4.00   Min.   :0.0400   Min.   :0.0600   Min.   :0.0400   Min.   :0.0000  
 1st Qu.:0.2700   1st Qu.:10.00   1st Qu.:0.0900   1st Qu.:0.4800   1st Qu.:0.2500   1st Qu.:1.0000  
 Median :0.4800   Median :12.00   Median :0.2500   Median :0.8400   Median :0.4500   Median :1.0000  
 Mean   :0.5272   Mean   :13.06   Mean   :0.2798   Mean   :0.8745   Mean   :0.5132   Mean   :0.8113  
 3rd Qu.:0.7700   3rd Qu.:12.00   3rd Qu.:0.3300   3rd Qu.:1.3300   3rd Qu.:0.7000   3rd Qu.:1.0000  
 Max.   :1.3300   Max.   :44.00   Max.   :0.9200   Max.   :2.0400   Max.   :1.5300   Max.   :1.0000  
```

```
    X3          X4          X5          X6          X7          X8          X9          X10         X11
37.5301887  6.5905660 22.2018868 23.1169811  0.5271698 13.0566038  0.2798113   0.8745283   0.5132075

    X3              X4          X5          X6          X7          X8          X9
 1.459509e+03 1.660102e+00 6.216333e+02 9.496457e+02 1.163053e-01 7.328520e+01  5.125958e-02
     X10         X11
2.725329e-01 1.147376e-01
```

1.3 Exploración de la correlación entre las variables.



A partir de la matriz de correlación, se puede ver que las variables que tienen las correlaciones más fuertes corresponden a:

- X11 y X7: 0.96
- X9 y X7: 0.93
- X10 y X7: 0.92
- X11 y X10: 0.86
- X3 y X5: 0.83
- X10 y X9: 0.77
- X3 y X11: -0.63
- X4 y X11: -0.61
- X3 y X10: -0.60

2. ANÁLISIS DE RESULTADOS

2.1 Análisis de Normalidad

Prueba de normalidad multivariada de Mardia Hipótesis: H_0 (nulo): Las variables siguen una distribución normal multivariante. H_a (alternativa): las variables no siguen una distribución normal multivariante. a = 0.05

Criterios: Se rechazará la hipótesis nula si el valor p es menor a alfa.

```
           Beta-hat      kappa        p-val
 Skewness  46.43941 410.214791 0.000000e+00
 Kurtosis 116.76710   4.596126 4.304194e-06
```

Dado que los valores p son menores a alpha, se rechaza la hipótesis nula H_0.

Tras la prueba de normalidad multivariada, se procederá a rechazar la hipótesis nula H_0 debido a que los valores p son mucho menores que el valor de alpha establecido (0.05).

Prueba de normalidad multivariada de Anderson-Darling

```
          Beta-hat      kappa        p-val
 Skewness 2.190628 19.3505493 0.0006705956
 Kurtosis 8.706209  0.6426596 0.5204449820
```

```
            Anderson-Darling test for Multivariate Normality

   data : df_num

   AD              : 6.519289
   p-value         : 9.999e-05

   Result  : Data are not multivariate normal (sig.level = 0.05)
```

A partir del test de Anderson-Darling para la normalidad multivariada,se puede ver que el valor p es mucho menor al valor preestablecido de significancia representado por alpha (0.05) por lo que se determina que no existe un comportamiento normal entre todas las variables en cuestión.

```
            Anderson-Darling test for Multivariate Normality

   data : df_bi

   AD              : 0.547976
   p-value         : 0.4257574

   Result  : Data are multivariate normal (sig.level = 0.05)
```

A partir del test de Anderson-Darling para la normalidad multivariada entre X4 y X9, se puede ver que el valor p es mayor al valor preestablecido de significancia representado por alpha (0.05) por lo que se determina que si existe normalidad multivariada en el grupo de variables X4 y X9.

Prueba de normalidad de Shapiro-Wilk

```
 Normalidad Multivariada
            Test       HZ p value MVN
 1 Henze-Zirkler 1.403322       0  NO

Normalidad Univariada
           Test  Variable Statistic   p value Normality
 1 Shapiro-Wilk    X3        0.8203  <0.001      NO
 2 Shapiro-Wilk    X4        0.9810  0.5552      YES
 3 Shapiro-Wilk    X5        0.7913  <0.001      NO
 4 Shapiro-Wilk    X6        0.6817  <0.001      NO
 5 Shapiro-Wilk    X7        0.9421  0.0125      NO
 6 Shapiro-Wilk    X8        0.5830  <0.001      NO
 7 Shapiro-Wilk    X9        0.8770   1e-04      NO
 8 Shapiro-Wilk    X10       0.9555  0.0467      NO
 9 Shapiro-Wilk    X11       0.9258  0.0028      NO

Medidas Estadísticas Descriptivas
      n       Mean    Std.Dev Median  Min    Max  25th  75th       Skew   Kurtosis
 X3  53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50  0.9679170  -0.4705349
 X4  53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40 -0.2458771  -0.6239638
```

```
 X5  53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60  1.3045868   0.6130359
 X6  53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70  2.4130571   6.1042185
 X7  53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77  0.5986343  -0.6312607
 X8  53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00  2.5808773   6.0089455
 X9  53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33  1.0729099   0.4060828
 X10 53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33  0.4645925  -0.6692490
 X11 53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70  0.9449951   0.5733500
```

Prueba de normalidad de Anderson-Darling

```
Normalidad Multivariada
      Test        H      p value MVN
 1 Royston 89.88169 6.444298e-18  NO

Normalidad Univariada
              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

Medidas Estadísticas Descriptivas
      n       Mean    Std.Dev Median  Min    Max  25th  75th       Skew   Kurtosis
 X3  53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50  0.9679170  -0.4705349
 X4  53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40 -0.2458771  -0.6239638
 X5  53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60  1.3045868   0.6130359
 X6  53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70  2.4130571   6.1042185
 X7  53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77  0.5986343  -0.6312607
 X8  53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00  2.5808773   6.0089455
 X9  53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33  1.0729099   0.4060828
 X10 53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33  0.4645925  -0.6692490
 X11 53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70  0.9449951   0.5733500
```

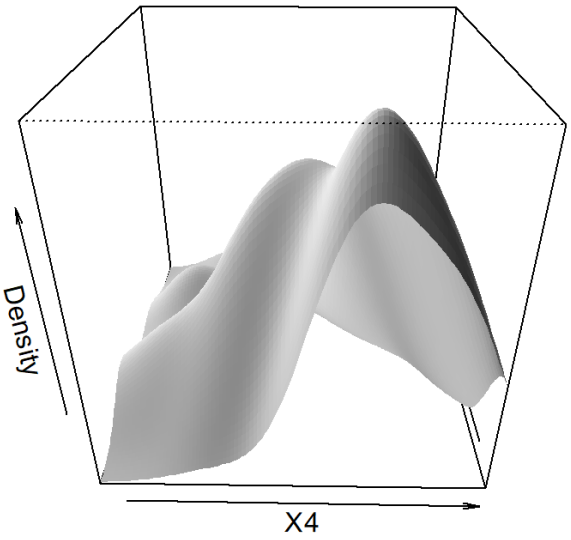
A partir de las pruebas de Anderson-Darling y el test de Shapiro-Wilk, se puede determinar que las variables que no tienen normalidad son:

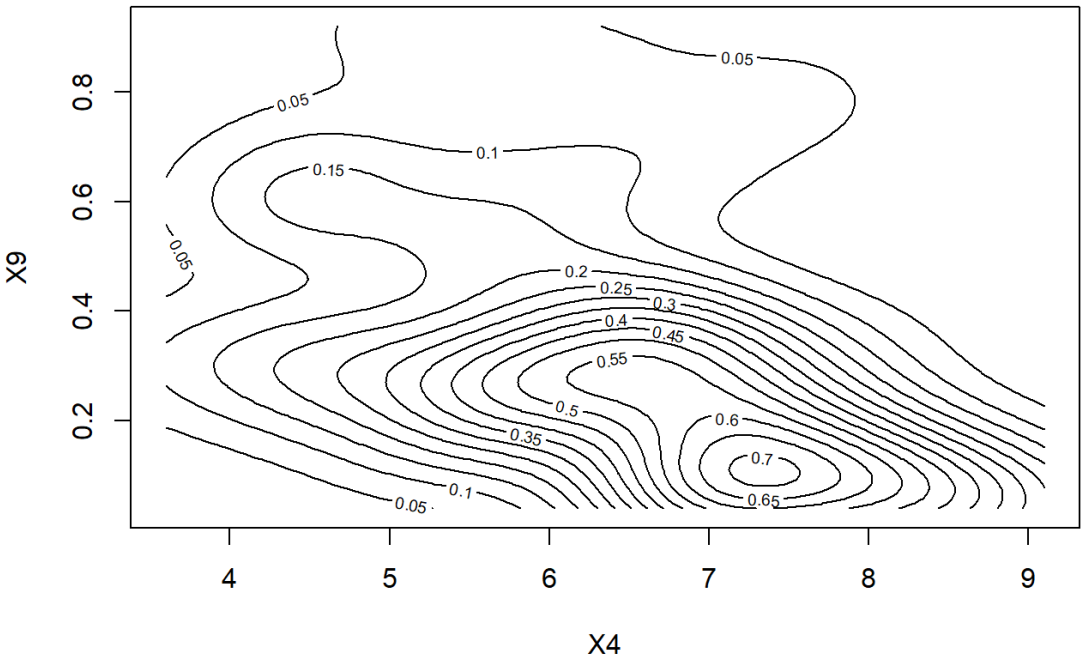
- X3 = alcalinidad (mg/l de carbonato de calcio)
- X5 = calcio (mg/l)
- X6 = clorofila (mg/l)
- X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago
- X8 = número de peces estudiados en el lago
- X9 = mínimo de la concentración de mercurio en cada grupo de peces
- 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)

Por otro lado, las variables que si tienen normalidad de acuerdo a este test corresponden a:

- x4: PH
- x10: máximo de la concentración de mercurio en cada grupo de peces

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






```
$multivariateNormality
           Test      HZ     p value MVN
1 Henze-Zirkler 1.46435 0.001324127  NO

$univariateNormality
              Test  Variable Statistic   p value Normality
```

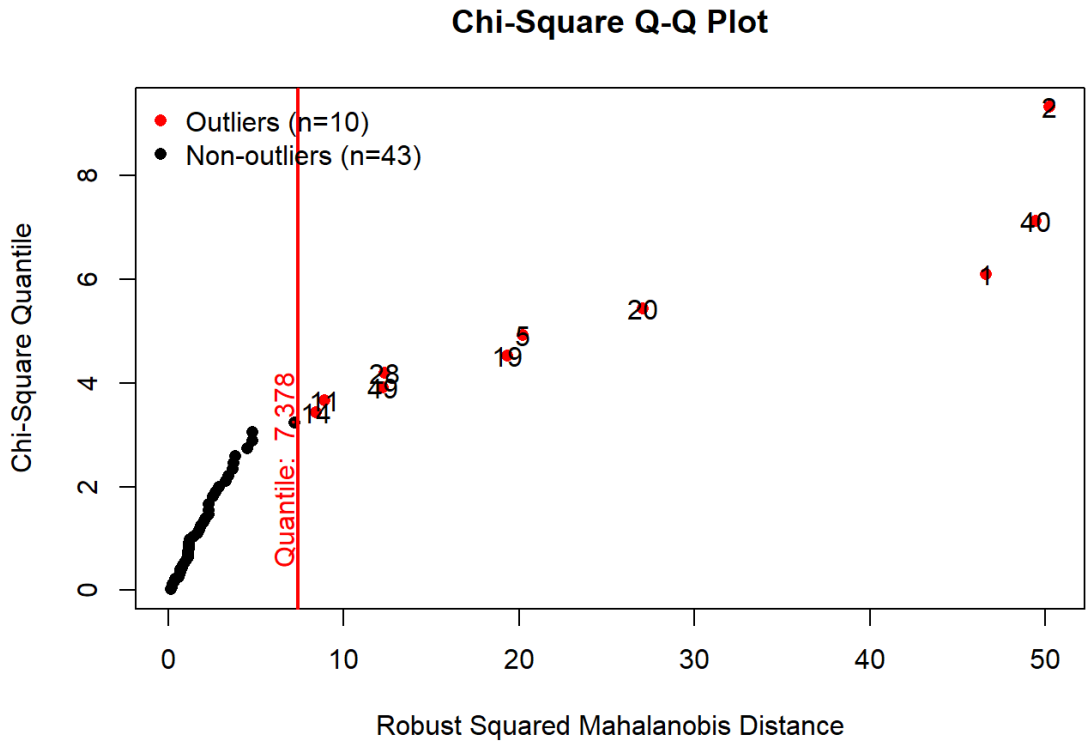
```
1 Anderson-Darling    X4        0.3496  0.4611      YES
2 Anderson-Darling    X9        1.9770  <0.001      NO

$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
X9 53 0.2798113 0.2264058   0.25 0.04 0.92 0.09 0.33  1.0729099  0.4060828
```

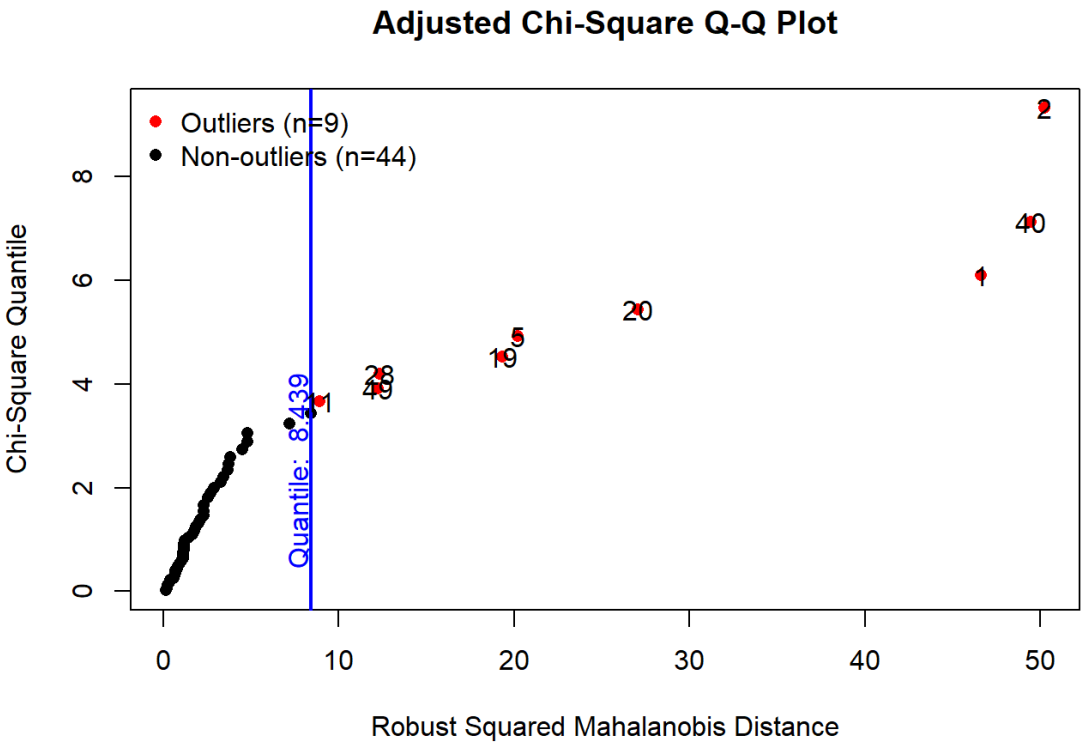
A través de las gráficas de contorno, se puede ver que ambas variables (X4, X10) tienen normalidad multivariada.

Detección de Datos Atípicos

La suposición de Normalidad Multivariada requiere la ausencia de valores atípicos multivariados. Por ende, es necesario verificar si los datos tienen valores atípicos multivariantes, antes de comenzar con el análisis de componentes principales. La MVN incluye dos métodos de detección de valores atípicos multivariados que se basan en distancias robustas de Mahalanobis (rMD (x)). La distancia de Mahalanobis es una métrica que calcula qué tan lejos está cada observación del centro de distribución o centroide en el espacio multivariable. Las distancias robustas se estiman a partir de estimadores determinantes de covarianza mínima.

Distancia Mahalanobis:



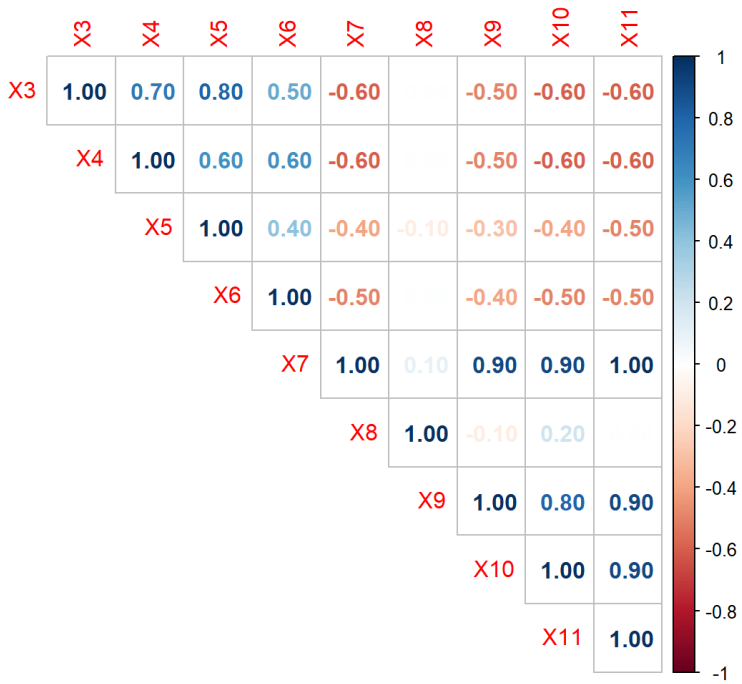


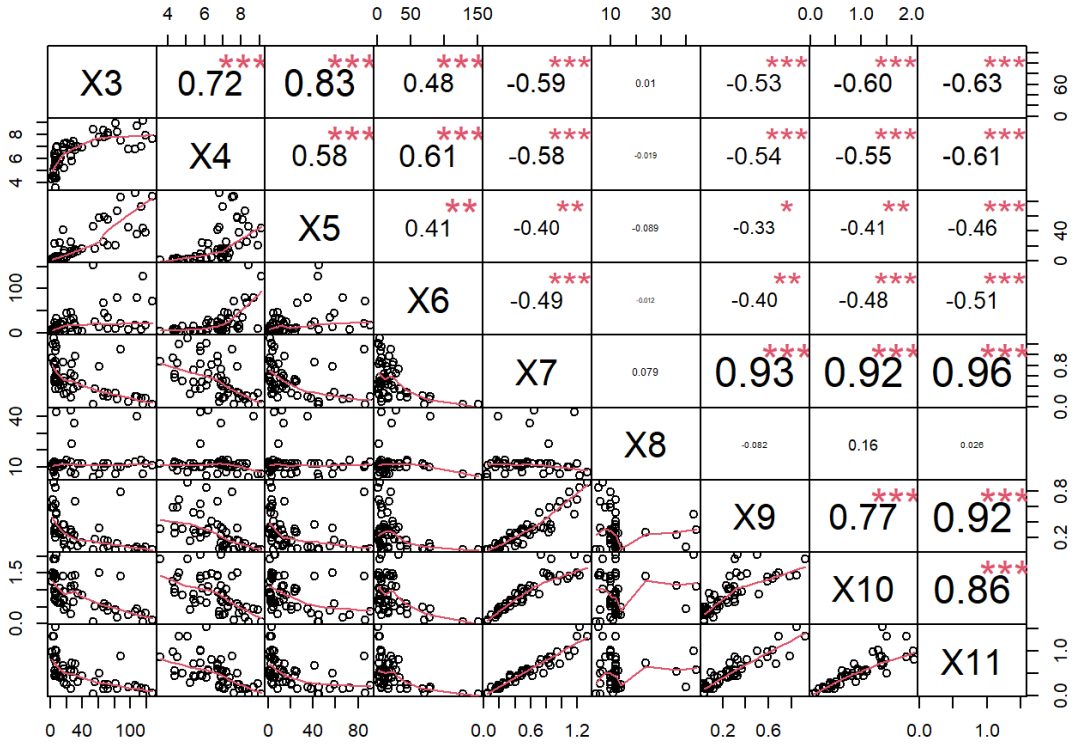
A partir de estas gráficas se puede observar que la distancia de Mahalanobis declara 5 observaciones como valores atípicos multivariados, mientras que la distancia de Mahalanobis ajustada no declara tan solo 4

2.2 Análisis de Componentes Principales

Justificación del uso de componentes principales

El análisis de componentes principales es adecuado para analizar la base de este problema ya que de acuerdo a lo que se determinó en el análisis previo de ANOVA, existen variables en este dataset que son codependientes entre sí y al momento de hallar cuáles son los factores determinantes en la contaminación de mercurio de los lagos de Florida, dichas variables solo traen redundancia al análisis ya que recaen dentro de la limitante de ANOVA dado que las variables no pueden estar correlacionadas entre sí y deben ser linealmente independientes unas de otras lo cuál no se cumpli al existir altos grados de correlación entre ellas tal y como se puede observar en la siguiente matriz de correlación.





De modo que el análisis de componentes principales (ACP) sirve justamente para resumir las variables que explican al modelo en funciones. Lo que busca es simplificar la cantidad de variables que se tienen al resumir a aquellas variables redundantes en funciones. Subsecuentemente, se generarán nuevas variables como resultado de la combinación lineal de las variables originales y de esa manera se agrupará la mayor variación posible. En efecto, su objetivo primordial es reducir lo más posible el número de variables en cuestión en función de la proporción de variabilidad que tienen sobre el modelo, descartando aquellas que realmente no apoyan mucho a la explicación de la variable que se pretende predecir. En otros términos, se elegirán las variables que introducen la mayor variabilidad al sistema. Estas nuevas variables que se crean, recibirán el nombre de componentes principales. El primer componente principal agrupa la mayor parte de variación, el segundo algo menos, y así sucesivamente.

Los supuestos que se van a tomar para aplicar este análisis corresponden a:

- Todas las variables deben ser numéricas.
- La cantidad de datos debe ser mayhor que el número de variables en cuestión (al menos 10 veces).
- No se requiere un supuesto de normalidad.
- Las variabls deben estar correlacionadas entre sí.

Análisis de componentes principales

Matriz de Covarianzas

	X3	X4	X5	X6	X7	X8
X3	1459.509456	35.3997134	793.065711	562.193324	-7.73773984	3.36556604
X4	35.399713	1.6601016	18.540018	24.159971	-0.25283491	-0.20522496
X5	793.065711	18.5400181	621.633266	314.949198	-3.40693687	-19.07703193
X6	562.193324	24.1599710	314.949198	949.645668	-5.16408563	-3.11828737
X7	-7.737740	-0.2528349	-3.406937	-5.164086	0.11630530	0.23074020
X8	3.365566	-0.2052250	-19.077032	-3.118287	0.23074020	73.28519594
X9	-4.544071	-0.1580980	-1.876788	-2.793997	0.07159176	-0.15825835
X10	-12.062062	-0.3711680	-5.309432	-7.802021	0.16305729	0.71993106
X11	-8.126195	-0.2674692	-3.922122	-5.286440	0.11080733	0.07481495

	X9	X10	X11
X3	-4.54407112	-12.06206241	-8.12619485
X4	-0.15809797	-0.37116800	-0.26746916
X5	-1.87678810	-5.30943179	-3.92212155
X6	-2.79399673	-7.80202068	-5.28644013
X7	0.07159176	0.16305729	0.11080733
X8	-0.15825835	0.71993106	0.07481495
X9	0.05125958	0.09046049	0.07048523
X10	0.09046049	0.27253295	0.15203327
X11	0.07048523	0.15203327	0.11473759

Matriz de Correlación

	X3	X4	X5	X6	X7	X8	X9
X3	1.00000000	0.71916568	0.83260419	0.47753085	-0.59389671	0.01029074	-0.52535654
X4	0.71916568	1.00000000	0.57713272	0.60848276	-0.57540012	-0.01860607	-0.54196524
X5	0.83260419	0.57713272	1.00000000	0.40991385	-0.40067958	-0.08937901	-0.33247623
X6	0.47753085	0.60848276	0.40991385	1.00000000	-0.49137481	-0.01182027	-0.40045856
X7	-0.59389671	-0.57540012	-0.40067958	-0.49137481	1.00000000	0.07903426	0.92720506
X8	0.01029074	-0.01860607	-0.08937901	-0.01182027	0.07903426	1.00000000	-0.08165278
X9	-0.52535654	-0.54196524	-0.33247623	-0.40045856	0.92720506	-0.08165278	1.00000000
X10	-0.60479558	-0.55181523	-0.40791663	-0.48497215	0.91586397	0.16109174	0.76535319
X11	-0.62795845	-0.61284905	-0.46440947	-0.50644193	0.95921481	0.02580046	0.91908939

	X10	X11
X3	-0.6047956	-0.62795845
X4	-0.5518152	-0.61284905
X5	-0.4079166	-0.46440947
X6	-0.4849721	-0.50644193
X7	0.9158640	0.95921481
X8	0.1610917	0.02580046
X9	0.7653532	0.91908939
X10	1.0000000	0.85975810
X11	0.8597581	1.00000000

Aplicar PCA Nuevas medias de las variables antes de la centralización

Alcalinidad	PH	Calcio	Clorofila	Concentración Mercurio	Cant. Peces
37.5301887	6.5905660	22.2018868	23.1169811	0.5271698	13.0566038

Mín Mercurio	Máx Mercurio	Estimación Merc
0.2798113	0.8745283	0.5132075

Nuevas varianzas de las variables antes de la centralización

```
Alcalinidad     PH         Calcio       Clorofila     Concentración Mercurio    Cant. Peces
37.8413998   1.2762362   24.6962413    30.5242170        0.3378030               8.4795316
            Mín Mercurio           Máx Mercurio        Estimación Merc
            0.2242597              0.5170985              0.3355186
```

Obtener eigenvalores y eigenvectores

```
Eigen valores:

5.34590819 1.22090789 1.04253153 0.66786333 0.33571266 0.20893778 0.10725403 0.05203127 0.01885332

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

Hallar Valor de los loadings para cada componente (eigenvector).

```
 Loadings:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
 Alcalinidad            0.351  0.403         0.304         0.284  0.726
 PH                     0.339  0.298        -0.232 -0.826        -0.223
 Calcio                 0.283  0.569         0.374  0.328 -0.298 -0.488 -0.141
 Clorofila              0.281  0.215        -0.831  0.395         0.111
 Concentración Mercurio -0.399  0.325
 Cant. Peces                           0.970                0.150 -0.140
 Mín Mercurio           -0.369  0.376 -0.117 -0.114  0.106  0.489 -0.224  0.528
 Máx Mercurio           -0.380  0.244  0.162        -0.165 -0.711  0.307  0.212
 Estimación Merc        -0.403  0.259                       0.223        -0.803
                       Comp.9
 Alcalinidad
 PH
 Calcio
 Clorofila
 Concentración Mercurio  0.850
 Cant. Peces
 Mín Mercurio           -0.340
 Máx Mercurio           -0.311
 Estimación Merc        -0.248

                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
 SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
 Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
 Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889  1.000
```

El primer componente es el resultado de la siguiente combinación lineal de las variables originales:

PC1 = 0.35 X3 + 0.34 X4 + 0.28 X5 + 0.28 X6 - 0.40 X7 - 0.02 X8 - 0.37 + X9 - 0.38 X10 - 0.40 X11

Los pesos asignados en el primer componente a las variables X11, X10, X9, X7 y X3 son aproximadamente iguales entre ellos y superiores al asignado a X8. Esto significa que el primer componente PC1 recoge mayoritariamente la información correspondiente a estimación de la concentración, Máxima concentración de mercurio, Mínimo de concentración, Concentración media de mercurio y alcalinidad.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
[1,]	-4.22502561	1.68907829	-1.065817636	-0.532543456	-0.51499921
[2,]	-4.84631607	1.68891743	-0.762426000	-0.475489713	0.05839408
[3,]	4.77162320	1.09818402	-0.517678570	-2.063000696	0.32062833
[4,]	0.31803491	-0.58045422	-0.076753567	0.504429476	-0.57413918
[5,]	-4.17839776	0.83761481	-0.252216036	-0.151362363	0.47414490
[6,]	1.31381567	-1.33242616	0.142291005	-0.869928396	-0.33158162
[7,]	-0.80010394	-1.31936879	-0.565715041	0.199612379	0.27425266
[8,]	2.70840701	0.38976584	-0.081736826	0.455025676	-0.21004457
[9,]	-1.65071136	-0.16910378	1.376322582	0.403120877	-0.09485627
[10,]	-1.90074578	0.20597518	-0.018860026	-0.686761339	-0.37906429
[11,]	-1.73659730	-0.38269057	-0.349673319	-0.299641926	0.42519921
[12,]	0.13511116	-0.80123647	-0.111193412	0.210575529	-0.81533471
[13,]	-0.06046233	0.06134507	-0.662331044	0.116393893	-0.62711693
[14,]	-3.47416785	0.65696008	3.644528982	-0.074812470	-0.09618200
[15,]	4.36875994	1.45422571	-0.158952530	0.529457366	1.15918018
[16,]	3.53832373	1.06247486	-0.293233041	-0.528739458	0.53715116
[17,]	3.27971478	0.62345211	3.367960835	-0.750058150	0.04547596
[18,]	0.09635673	1.91289444	-0.534193383	0.627063683	-0.71435220
[19,]	-3.05823124	0.69982100	-0.438693822	-0.292489699	-0.17612986
[20,]	-1.96676739	1.30130556	-0.773398222	-0.392740797	-0.43458086
[21,]	-1.38277926	-1.22378955	-0.302347874	0.182850369	0.93240034
[22,]	-0.74055314	-0.27007210	-0.191494273	-0.062864034	-0.42276279
[23,]	-0.08485435	-1.18128278	-0.265808689	-0.271498358	0.11847510
[24,]	-1.83181089	0.51096529	0.238912543	-0.973641361	-0.75846581
[25,]	-0.34375522	-1.46070399	-0.584930616	0.383698960	0.27655051
[26,]	-0.35524635	-0.27441691	2.747221853	-0.059877691	-0.21048436
[27,]	1.21142602	-0.44468811	-0.341572544	0.002405088	-0.41610519
[28,]	-2.89794608	-0.02190277	-0.737811265	-0.074079052	0.69002511
[29,]	-1.05532441	-1.27720433	-0.427038509	-0.034551976	0.76251528
[30,]	-0.35841804	-1.53996001	-0.521780268	0.290593965	-0.02976169
[31,]	2.22537931	0.82695403	-0.283851432	0.154707292	0.07669607
[32,]	0.80033208	-0.65776908	-0.322383685	-0.556281838	-0.53460678
[33,]	-3.14158395	-0.50275506	-0.020424792	0.317046302	0.94112877
[34,]	-1.46660647	-1.40318607	-0.149912546	0.348282285	0.77874288
[35,]	2.84223064	0.26579390	-0.036042751	1.110732754	-0.46650298
[36,]	1.64833449	-1.36504242	-0.073803729	-0.740777182	-0.02323613
[37,]	2.90714501	1.14903894	-0.039709626	1.986396431	0.90111978
[38,]	4.24560697	0.48314929	-0.867007919	-3.086157946	1.03109750
[39,]	-0.36115490	-0.79731460	-0.251578988	-0.338393933	-0.60249516
[40,]	-0.94871585	4.15827577	-0.291150414	0.806806567	0.13320882
[41,]	2.91923809	0.64338479	0.069940885	2.042702435	0.46006941
[42,]	2.44848393	-0.22376389	-0.186689073	1.530985926	0.20673766
[43,]	-0.55813765	-0.55383632	-0.433598602	0.534524368	0.09889868
[44,]	2.65161289	-0.02959875	0.075329029	-0.995017532	-0.33530782
[45,]	-1.04004850	1.07341851	-0.004153475	-0.089005463	-0.04972619
[46,]	-0.65215066	-0.38502956	-0.232081379	0.166589625	0.04466513
[47,]	-0.79389539	-0.32552707	3.592570523	-0.161482637	0.40822023
[48,]	2.23647757	-0.13738826	-0.669385467	0.468065872	-1.61083592
[49,]	-3.14852113	-0.02889182	-0.555467454	-0.079541389	0.82604286
[50,]	0.50837352	-0.87749489	-0.034318513	0.418539064	-0.69138833
[51,]	0.05972769	-0.91452278	-0.300042837	-0.256075045	-0.48731738
[52,]	0.20587593	-2.00871528	-0.382418621	0.533271927	0.52537867

```
 [53,]  1.61863757 -0.30285853 -0.085400422  0.572935790 -0.89902107

            Comp.6      Comp.7       Comp.8       Comp.9
 [1,]  1.069715328 -0.02667015 -0.685576169 -0.143781561
 [2,]  0.424266665  0.26015023  0.174094209 -0.114843858
 [3,]  0.241022355  0.74234179  0.022699515  0.045754194
 [4,] -0.186869637  0.14767405 -0.144096205  0.065909895
 [5,]  0.523077668  0.20721482 -0.547184278  0.134331174
 [6,] -0.101381085 -0.13451959 -0.110196401  0.154495770
 [7,]  0.164595140 -0.18123023  0.201043848  0.024798634
 [8,] -0.128276727 -0.41341996 -0.024072984 -0.027350989
 [9,] -0.320887448  0.32236779 -0.175860880  0.277167867
[10,] -0.391754524  0.08363606 -0.085014767 -0.047811507
[11,]  0.637834761 -0.20020933  0.180933460 -0.005105558
[12,] -0.183132539 -0.30534114 -0.456205540  0.267106959
[13,] -0.383380789 -0.23473052 -0.101453419 -0.166370362
[14,] -0.247357435 -0.04362125  0.031264316  0.068390992
[15,] -0.032459238  0.13938670 -0.093922054 -0.033365424
[16,]  0.005820923 -0.38414021 -0.099988573 -0.017097299
[17,]  0.686586511  0.33495949  0.062354182 -0.151154439
[18,] -1.074557999 -0.03218111  0.290188511  0.164749302
[19,]  0.462484313 -0.04598318 -0.025784229  0.220679317
[20,]  0.141561164  0.10801944  0.800434623  0.098320063
[21,] -0.057039740  0.14452418  0.392656575  0.253241775
[22,]  0.175301577 -0.24154432  0.099339302 -0.156662674
[23,]  0.231946171 -0.17753753  0.106000233  0.009866357
[24,] -1.365256212  0.45504175 -0.075059339 -0.475277789
[25,]  0.178803330 -0.19615819  0.085381342 -0.174923798
[26,]  0.033632373 -0.21714515  0.016193462 -0.042808156
[27,]  0.291645949  0.23584957  0.115689606  0.064852635
[28,]  0.072701194  0.21865581  0.200016880 -0.180996968
[29,]  0.025589190  0.05203280  0.153853980 -0.026932480
[30,]  0.005286579 -0.15847128 -0.139971766 -0.170184971
[31,] -0.336539879 -0.36956971  0.005246643 -0.091370217
[32,]  0.004794538 -0.03626014 -0.131895088 -0.068074852
[33,] -1.173139556  0.95582192 -0.209833251  0.008334444
[34,] -0.256306282  0.22475276 -0.023942487  0.052408709
[35,]  0.569674235  0.76608341  0.143779288 -0.004981972
[36,]  0.169730449 -0.24325785  0.117506210  0.042098236
[37,] -0.280114724 -0.24972990 -0.355813996 -0.033138233
[38,] -0.393922446 -0.25267975 -0.156157998  0.083290855
[39,]  0.484816571 -0.42438329 -0.087885184 -0.033596934
[40,]  0.241169028 -0.47425751  0.352015072  0.020124840
[41,]  0.149641149  0.09310211 -0.175405152  0.004209627
[42,]  0.337354774  0.36916616 -0.067758187 -0.138750517
[43,] -0.543130144 -0.34790246  0.180623049 -0.149217437
[44,] -0.118662437  0.25027007  0.054103269  0.055190329
[45,] -0.847295466 -0.49815150 -0.077605929  0.269523412
[46,] -0.266309741 -0.59103865 -0.045443875 -0.141073723
[47,]  0.322129810 -0.32867219  0.068250666 -0.037730625
[48,]  0.332481548  0.42133554 -0.120281088  0.153284382
[49,]  0.362871335  0.14079841 -0.098547569  0.068864450
[50,] -0.465543025  0.14464230  0.072024795  0.100735544
[51,]  0.126510340 -0.21780256  0.043726233  0.007058017
[52,]  0.279193244 -0.04976785  0.117769149 -0.011281100
[53,]  0.401078862  0.25854934  0.227767988 -0.070904333
```

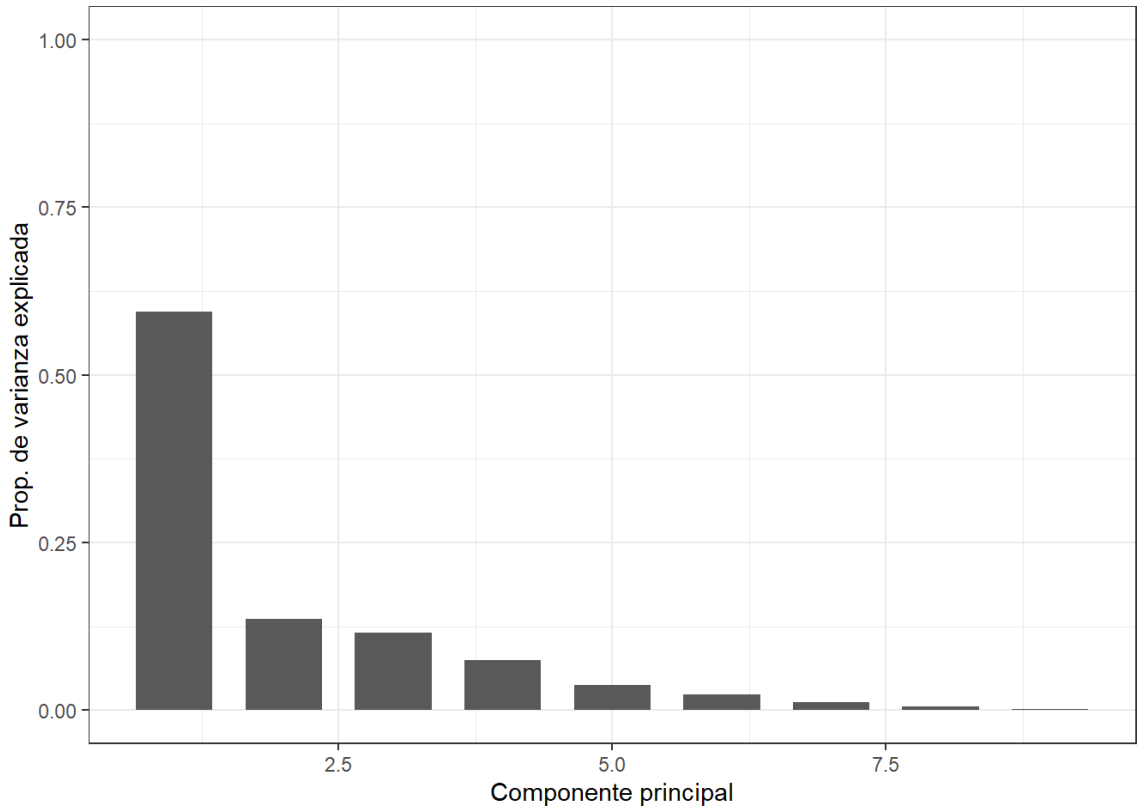
Conocer la varianza explicada por cada componente, la proporción respecto al total y la proporción de varianza acumulada.

Varianza explicada por cada componente:

```
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
5.34590819 1.22090789 1.04253153 0.66786333 0.33571266 0.20893778 0.10725403 0.05203127
                                       Comp.9
                                   0.01885332
```

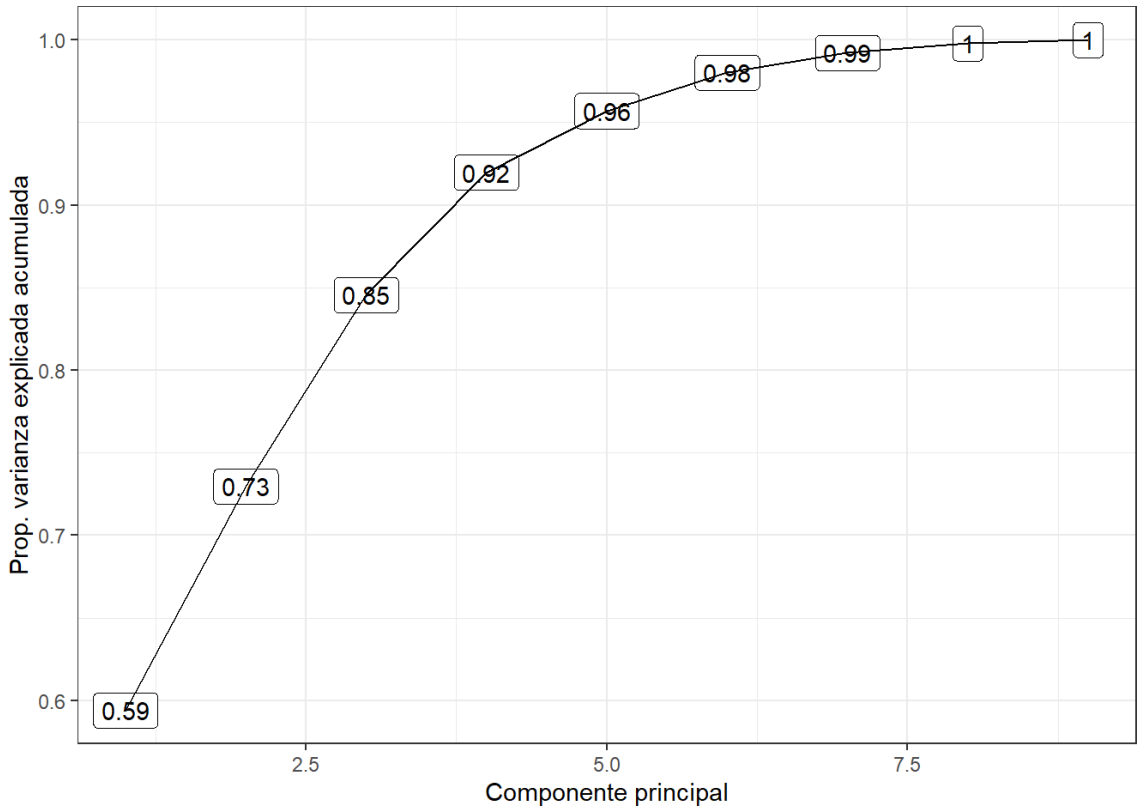
Proporción de Varianza explicada por cada componente:

```
     Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6      Comp.7     Comp.8
0.593989799 0.135656432 0.115836836 0.074207036 0.037301407 0.023215309 0.011917115 0.005781252
                                        Comp.9
                                    0.002094814
```



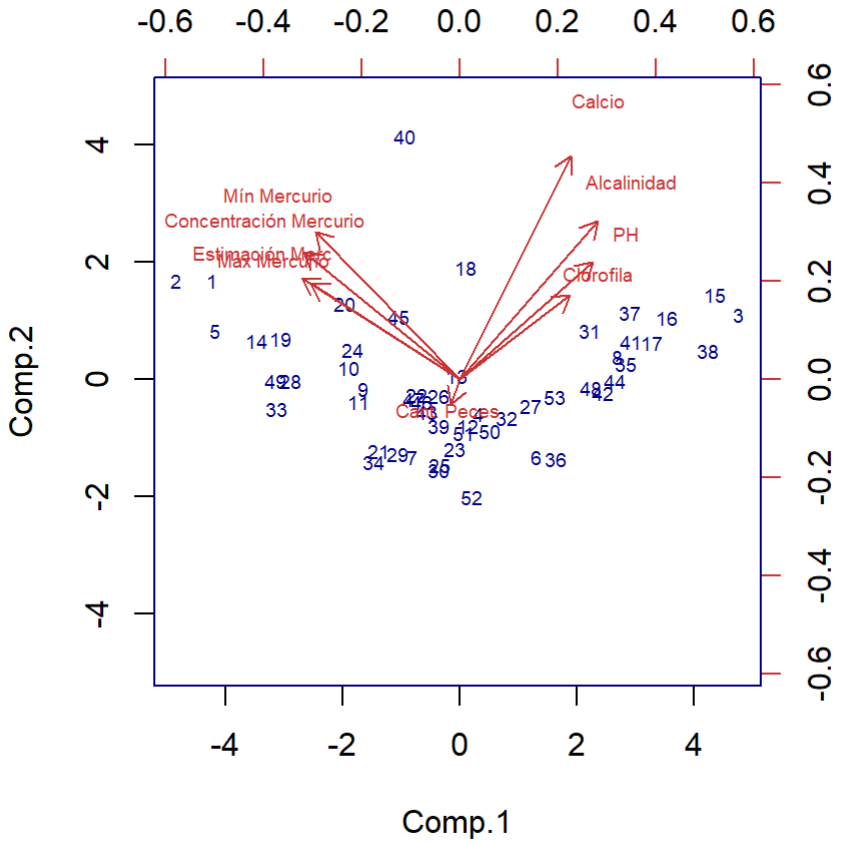
Se puede ver claramente que el primer componente explica la mayor cantidad de varianza en un 60%.

```
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9
0.5939898 0.7296462 0.8454831 0.9196901 0.9569915 0.9802068 0.9921239 0.9979052  1.0000000
```



Apreciándo la gráfica, se determina que el primer componente explica el 59% de la varianza observada en los datos, el segundo el 14% y el tercero el 12% .Los tres últimos componentes no superan por separado el 2% de varianza explicada. Si se aplicasen únicamente los tres primeros componentes se conseguiría explicar el 85% de la varianza observada.

Gráfico de vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes



Cuanto más largas sean las flechas rojas, más alto es el valor del coeficiente de esa variable en ese componente.Claramente se puede ver que el Componente 1 tiene una variabilidad más amplia que el componente 2 ya que los nuevos datos están más dispersos, por lo que se justifica graficamente la superioridad del Componente 1 en función de su varianza explicada. También se puede observar una grupación visual de las variables de acuerdo a su signo lo que se traduce en que X3, X4, X5 y X6 tienen una relación lineal positiva, mientras que X7, X9, X10 y X11 tienen una correlación lineal negativa con respecto al PC1. En lo que respecta a X8, se puede ver que esta es la variable que tiene menos influencia en ambos componentes y se comporta de manera isolada con respecto al resto de variables ya que tiene una tendencia negativa con respecto a ambos componentes.

CONCLUSIÓN

A partir del análisis de componentes principales, se obtuvo que los tres primeros componentes principales en su conjunto explican más del 85% de la varianza del modelo, por lo que se procederán a considerar estos componentes como los más significativos y exclusivos para determinar las variables que más variación causan dentro del contexto del problema. Dentro de cada componente, se procederá a escoger las variables más significativas (por encima de 0.30) dentro de cada componente.

Por ende, para el primer componente, las variables más importantes corresponden a:

- X3: Alcalinidad
- X4: PH
- X7: Concentración media de mercurio
- X9: Mínimo de concentración de mercurio en el lago
- X10: Máximo de concentración de mercurio en el lago
- X11: Estimación de concentración de mercurio.

Por ende, para el segundo componente, las variables más importantes corresponden a:

- X3: Alcalinidad
- X4: PH
- X5: Calcio
- X7: Concentración media de mercurio
- X9: Mínimo de concentración de mercurio en el lago

Por ende, para el tercer componente, las variables más importantes corresponden a:

- X8: Número de peces

En síntesis, se puede concluir que los factores que influyen mayormente en la contaminación de los lagos de la concentración media de merurio corresponden a:

- Alcalinidad
- PH
- Calcio
- Mínimo de Concentración de mercurio en el lago.
- Máximo de concentración de mercurio en el lago.
- Número de peces en el lago.

ANEXOS

Liga a Repositorio de Github:

https://github.com/emilyvic/ArtificialIntelligence/tree/main/Estadistical%20Models/Mercury%20in%20Fish%20v2