Reporte Final: Peces y Mercurio

Russel Rosique

2022-12-04

RESUMEN 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.

INTRODUCCIÓN

Anteriormente se trabajó con este conjunto de datos, en el cual se analizaron las variables, así como el peso que ejercía cada una sobre el nivel de mercurio en los peces. Utilizando algunas de las herramientas vistas en clase, así como la normal multivariada y los componentes principales, se responderá la siguiente pregunta:

¿Cuáles son las variables determinantes en la contaminación por mercurio de los peces?

DESCRIPCIÓN DE LOS DATOS
|——————————|——————————————————————————————
—————— | X1 | Número de identificación | | X2 | Nombre del lago | | X3 | Alcalinidad | | X4 | PH | | X5 | Calcio (mg/l) | | X6 | Clorofila (mg/l) | | X7 | concentración media de mercurio (parte por millón) en el tejido muscualar del grupo de peces estudiados en cada lago
| | X8 | Número de peces estudiados en el lago | | X9 | mínimo de la concentración de mercurio en cada grupo de peces | | X10 | máximo de la concentración de mercurio en cada grupo de peces | | X11 | estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible) | | X12 | indicador de la edad de los peces (0: jóvenes; 1: maduros) |

ANÁLISIS DE LA BASE DE DATOS

El primer paso en esta análisis del conjunto de datos es la limpieza de los datos, para poder realizar las pruebas pertinentes de normalidad. De esta forma se podrán identificar a las variables con una distribución normal, así como una posible normal multivariada. Para realizar estas pruebas primero se tienen que separar a las variables que no son númericas continuas. Posteriormente obtener la matriz de covarianza y correlación con las variables numéricas restantes.

```
##              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
```

```
##              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
```

Posterior a esto, se realizará el test de anderson-darling a todas las variables numéricas de la base de datos para encontrar aquellas que están distribuidas de forma normal.

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
##
##  Anderson-Darling normality test
##
## data:  alcalinidad
## A = 3.6725, p-value = 2.706e-09
```

```
##
##  Anderson-Darling normality test
##
## data:  PH
## A = 0.34956, p-value = 0.4611
```

```
##
##  Anderson-Darling normality test
##
## data:  calcio
## A = 4.051, p-value = 3.193e-10
```

```
##
##  Anderson-Darling normality test
##
## data:  clorofila
## A = 5.4286, p-value = 1.4e-13
```

```
##
##  Anderson-Darling normality test
##
## data:  mediaMercurio
## A = 0.92528, p-value = 0.0174
```

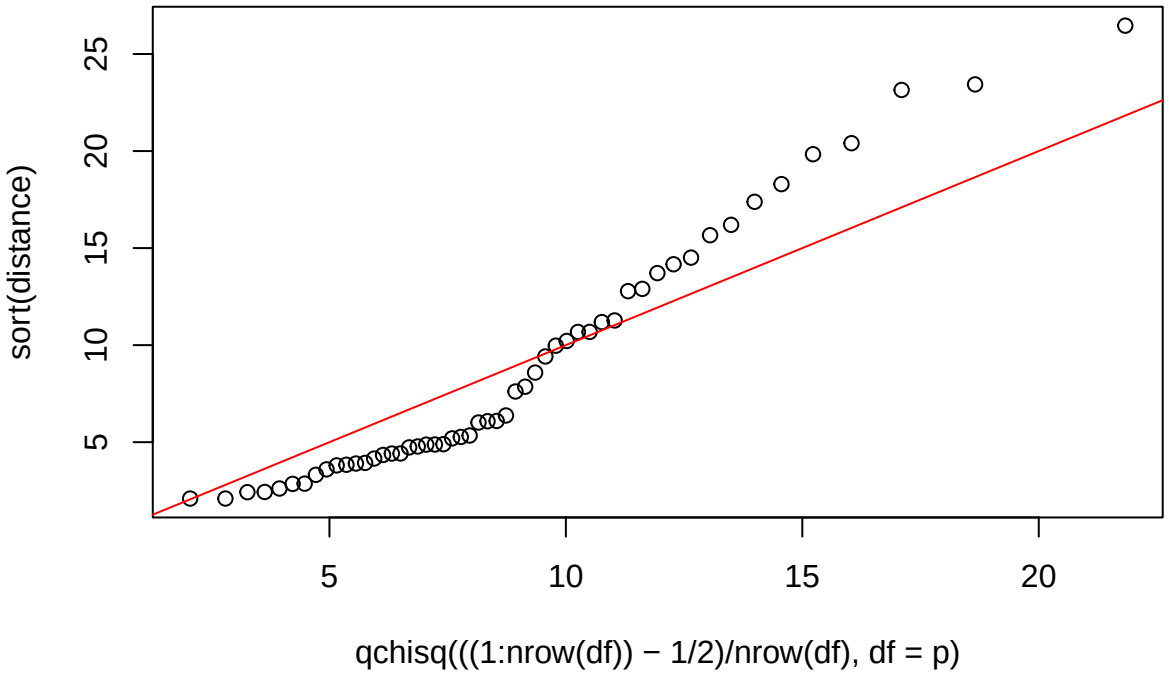
```
##
##  Anderson-Darling normality test
##
## data:  minMercurio
## A = 1.977, p-value = 4.161e-05
```

```
##
##  Anderson-Darling normality test
##
## data:  maxMercurio
## A = 0.65847, p-value = 0.08099
```

Tras realizar el test de anderson-darling en las variables numéricas continuas del conjunto, se concluye que existen 2 variables con normalidad: PH y máximo en la concentración de mercurio en cada grupo de peces.

Una vez hecho lo anterior, se obtiene la Distancia de Mahalanobis, al igual que el gráfico Q-Q Plot multivariado para observar datos atipicos:

```
##  [1] 23.148248 10.214173 17.385270  2.613110 12.781132  4.880536  3.317725
##  [8]  3.598373  8.585982  2.850207  4.342272 11.186573  4.414540 15.664149
## [15]  9.971855  6.093529 18.296407 13.711814  6.078138 16.197452 10.681426
## [22]  2.863256  2.096404 26.461606  4.730028  7.858117  2.423587  6.374972
## [29]  3.899854  4.418718  3.802860  2.434265 20.404938  4.873017 11.272724
## [36]  3.838086 14.166717 23.435759  4.779110 19.836807  9.415256  7.611127
## [43]  5.195259  3.932293 10.678490  4.897695 14.509979 12.898648  5.343472
## [50]  4.156698  2.095601  5.269222  6.012521
```



Se logra observar en la gráfica que existe una gran cantidad de datos atípicos.

Por esta razón, para el MVN test que se realizará a continuación, se dejará el argumento de 'showOutliers'

como verdadero, y se utilizarán únicamente las variables las cuales su normalidad ya fue comprobada, es decir, PH y máximo de mercurio.

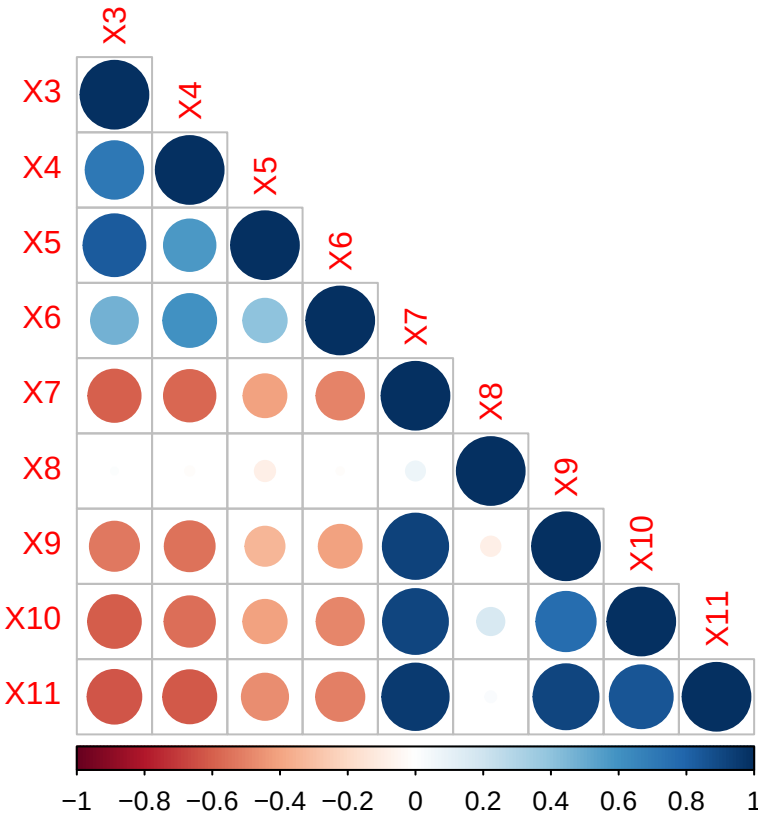
```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  6.17538668676458 0.186427564928852    YES
## 2 Mardia Kurtosis -1.12820795824432  0.25923210375991    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##               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
##
## $multivariateOutliers
## NULL
```

Estos resultados indican que el nivel de PH, y el máximo de concentración de mercurio tienen una distribución normal multivariada.

COMPONENTES PRINCIPALES

Se realizarán una serie de pruebas para determinar los componentes principales de este conjunto de datos, y de esta forma obtener información con mayor pertinencia.

```
## corrplot 0.92 loaded
```



En el plot anterior se puede observar la correlación de las distintas variables con un mapeo de calor. Como era de esperarse, la X9, X10, y X11 (Min. de mercurio, Max. de mercurio, y Estimación de mercurio respectivamente) tienen una alta correlación.

A continuación se obtienen los componentes principales del conjunto de datos:

```
## C.P. Covarianza:  0.7264164 0.9300767 0.9775266 0.9996963 0.9999067 0.9999882 0.9999979 0.9999994 1
```

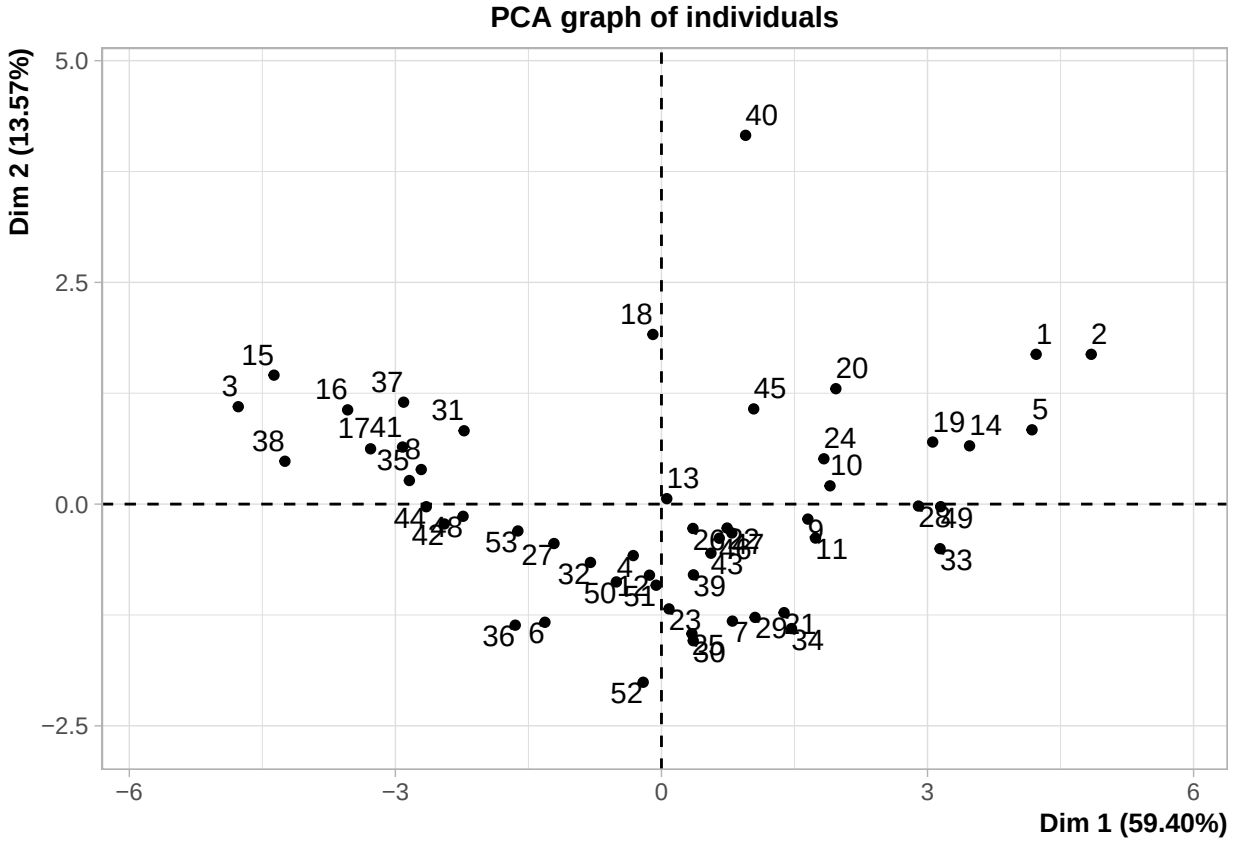
Se obtienen un total de 5 componentes que explican arriba del 99% de la varianza. Enseguida, utilizando las librerías de 'FactoMineR' y 'factoextra', se puede ver de forma gráfica la influencia de cada componente y la relación entre ellos.

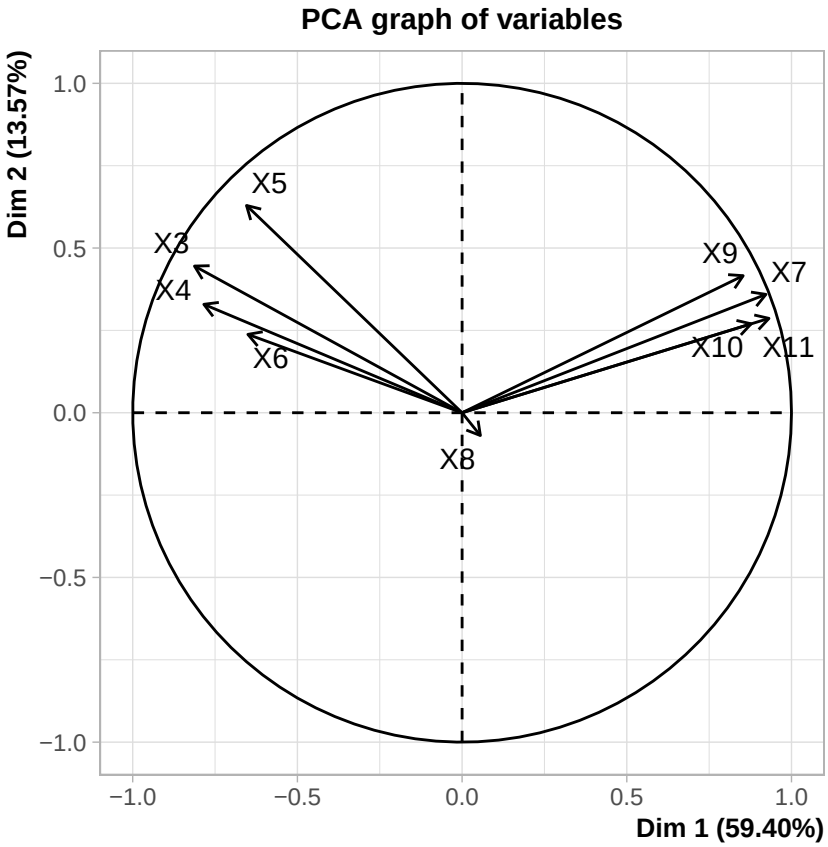
```
## Loading required package: ggplot2
```

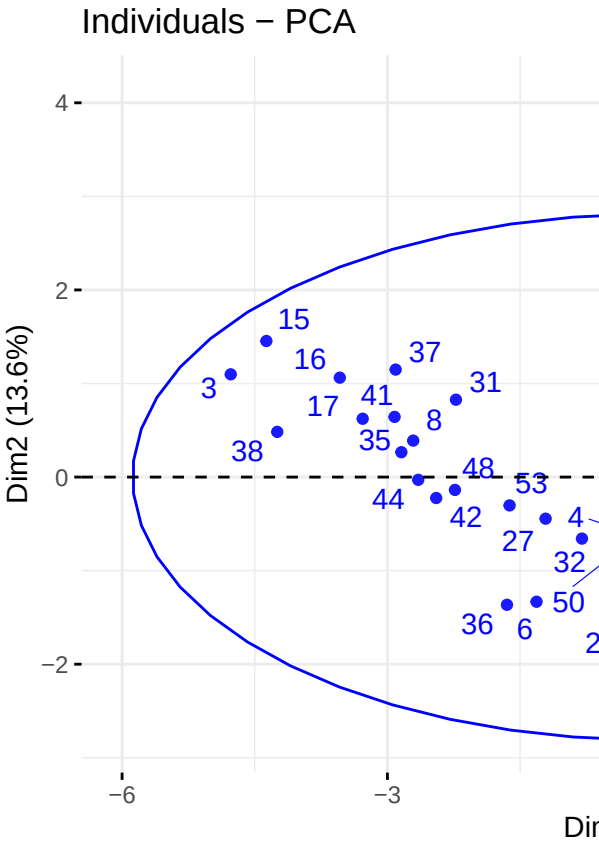
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

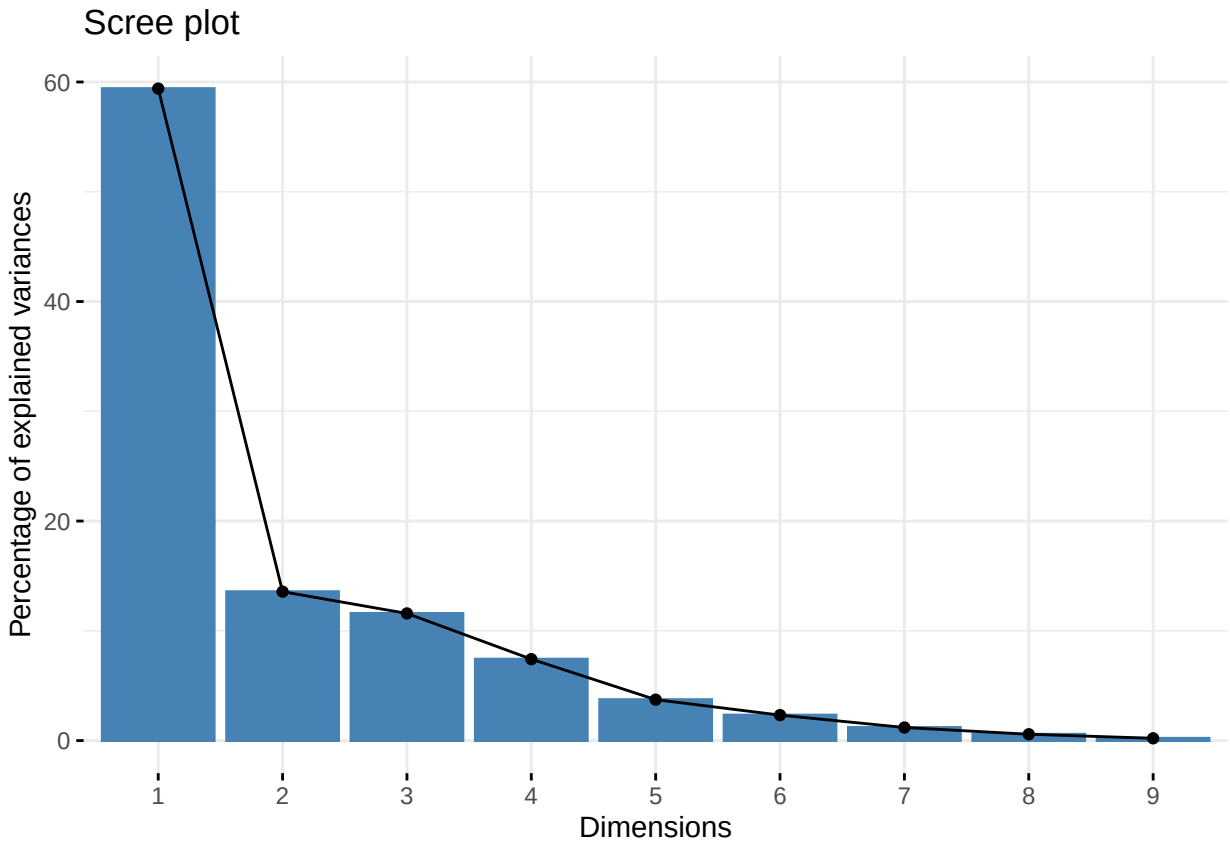
```
##
## Loadings:
##     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## X3   0.770  0.360  0.512  0.121
## X4                              -0.989  0.140
## X5   0.459  0.261 -0.825 -0.203
## X6   0.442 -0.896
## X7                                      0.472 -0.282  0.307  0.773
## X8                 0.237 -0.971
## X9                                      0.295 -0.466  0.587 -0.589
## X10                                     0.694  0.693        -0.182
## X11                                     0.435 -0.471 -0.749 -0.147
##
##                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
```

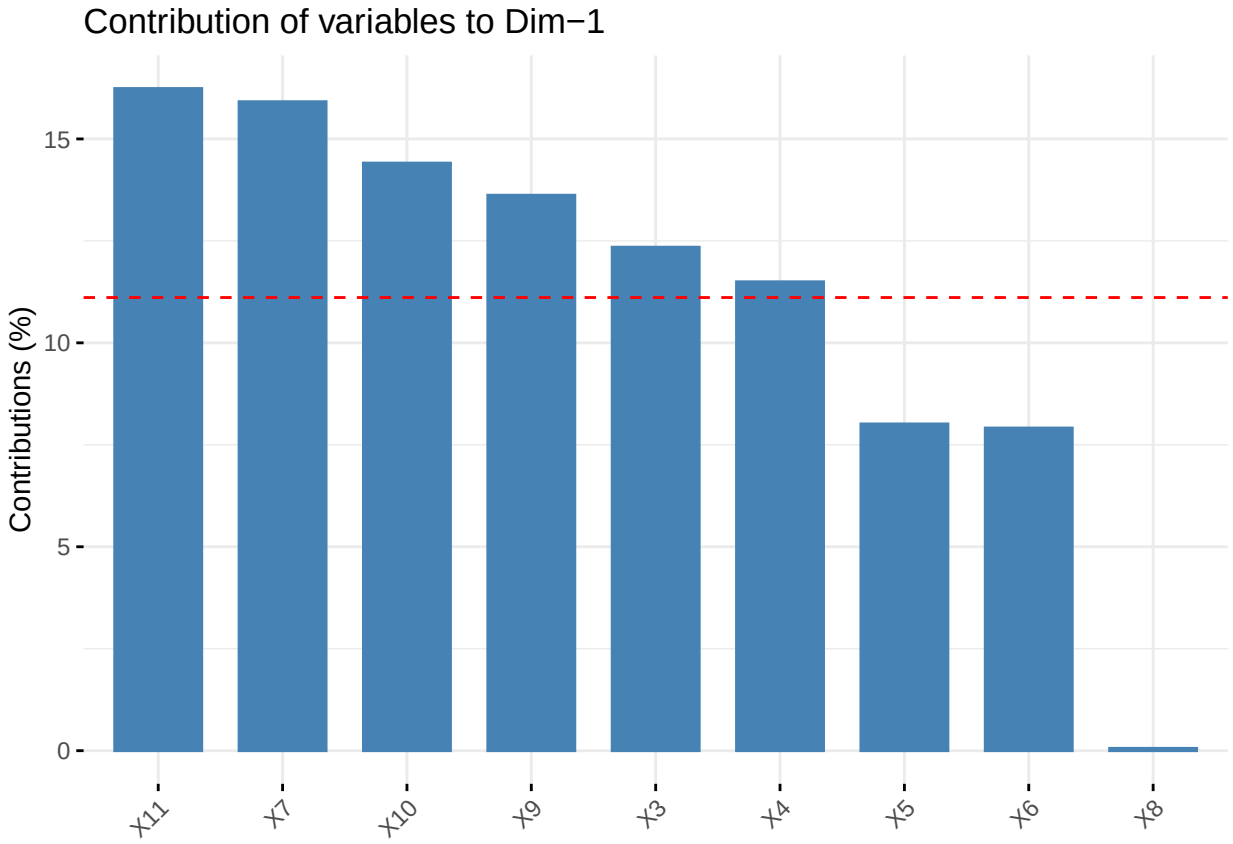
```
## Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
## Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889  1.000
```




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


Individuals − PCA
Dim2 (13.6%)
15
16
37
3
31
17
41
8
38
35
48
53
44
42
4
27
32
50
36
6
4
2
0
−2
−6
−3


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



En estas gráficas se puede ver la influencia de cada variable en los compontentes principales, así como la varianza explicada por cada compontente. En este caso, el primer componente es el que mayormente explica dicha varianza.

CONCLUSIÓN

En este proyecto, se tuvo la oportunidad de retomar un trabajo anterior, y expandirlo utilizando nuevas herramientas. De esta forma se puede ampliar el entendimiento del conjunto de datos dado. No obstante, se analiza que es un conjunto con muy pocas observaciones, por lo que aunque se obtuvieron resultados muy acertados, hace falta una gran cantidad de observaciones para determinar con certeza la causa o causas principales del problema de contamincación por mercurio en los peces.

No obstante, de todas formas se llegaron a varias conclusiones interesantes. Después de emplear herramientas para consultar normalidad multivariada y compontentes principales. Gracias a estas herramientas se llegó a la conclusión de que los factores más determinantes en la contaminación por mercurio en los peces son los niveles de alcalinidad y PH (variables que pasan la línea de contribución del primer componente), debido a sus valores dentro del análisis de componentes principales.

ANEXOS

https://github.com/russelr01/PortafolioIA