Los peces y el mercurio

Code

Yolanda Elizondo Chapa A01137848: 04 de diciembre del 20222

EL PROBLEMA

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio. Las variables que se midieron se encuentran en mercurio.csv y su descripción es la siguiente:

X1 = número de indentificación
X2 = nombre del lago
X3 = alcalinidad (mg/l de carbonato de calcio)
X4 = PH
X5 = calcio (mg/l)
X6 = clorofila (mg/l)
X7 = concentración media de mercurio (parte por millón) en el tejido 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 generals

Antes de iniciar a hacer las pruebas de normalidad y el análisis de componentes principales, se va a analizar de manera general el datset proporcionado, intentando identificar datos faltantes, atípicos, duplicados, etc. Lo primero que podemos observar es la dimensión del dataset, lo cual es conformada por 53 filas las cuales corresponden a los 53 lagos de Florida y 12 columnas las cuales corresponden a la descripción proporcionada anteriromente.

Code

```
[1] 53 12
```

A continuación se muestra los tipos de datos del dataframe, en donde se puede identificar una variable categórica (X2), la cual posiblemente se tenga que cambiar a una variable númerica o eliminar.

Code

```
'data.frame':   53 obs. of  12 variables:
 $ 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 ...
```

En esta parte se muestra las primeras 3 columnas del dataset, se puede observar que las variables X1 y X2 proporcionas la misma información un identificador unico, por lo tanto se va a cambiar el nombre del renglón por el valor de X2 y se eliminaran ambas variables X1 y X2.Terminando con unicamente 10 variables y sin ninguna variable categórica.

Code

	X1 <int>	X2 <chr>	X3 <dbl>	X4 <dbl>	X5 <dbl>	X6 <dbl>	X7 <dbl>	X8 <int>	X9 <dbl>
1	1	Alligator	5.9	6.1	3.0	0.7	1.23	5	0.85
2	2	Annie	3.5	5.1	1.9	3.2	1.33	7	0.92
3	3	Apopka	116.0	9.1	44.1	128.3	0.04	6	0.04

3 rows | 1-10 of 12 columns

Code

	X3 <dbl>	X4 <dbl>	X5 <dbl>	X6 <dbl>	X7 <dbl>	X8 <int>	X9 <dbl>	X10 <dbl>	X11 <dbl>
Alligator	5.9	6.1	3.0	0.7	1.23	5	0.85	1.43	1.53
Annie	3.5	5.1	1.9	3.2	1.33	7	0.92	1.90	1.33
Apopka	116.0	9.1	44.1	128.3	0.04	6	0.04	0.06	0.04

3 rows | 1-10 of 10 columns

A continuación se muestra el número de datos duplicados y datos nulos

Code

```
Datos duplicados  0
```

Code

```
Datos nulos  0
```

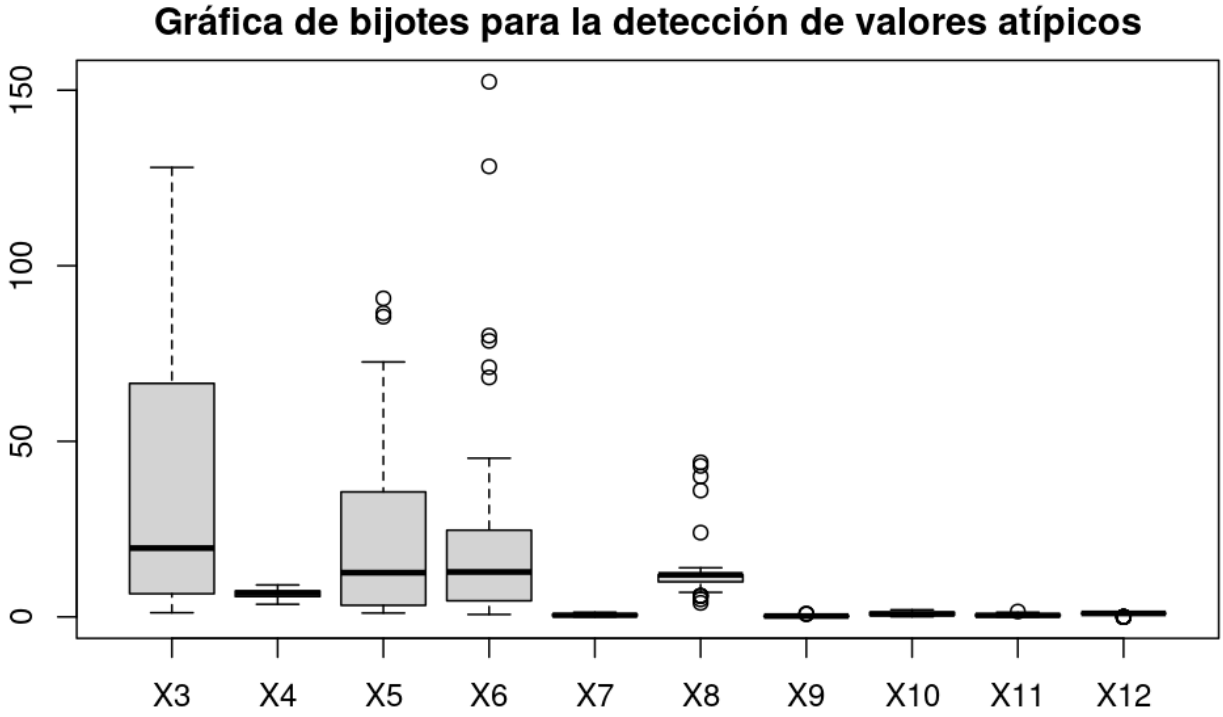
Con respecto al comportamiento de las variables, a continuación se muestra un resumen de estadística descriptiva (Min, Max, Median, Mean, 1stQu and 3rd Qu), se puede observar que el promedio de alcalinidad es de 37.5 y el del PH es de 6.5. También se observa que el máximo de calcio en el lago es de 90.7 y de clorofila es de 152.4. Y el promedio del mínimo de la concentración de mercurio es de .27, mientras que el promedio máximo es de .87 generando así un promedio de la estimación de la conecntración de un .5

Code

```
       X3               X4              X5             X6               X7              X8
 Min.   :  1.20   Min.   :3.600   Min.   : 1.1   Min.   :  0.70   Min.   :0.0400   Min.   :
4.00
 1st Qu.:  6.60   1st Qu.:5.800   1st Qu.: 3.3   1st Qu.:  4.60   1st Qu.:0.2700   1st Qu.:1
0.00
 Median : 19.60   Median :6.800   Median :12.6   Median : 12.80   Median :0.4800   Median :1
2.00
 Mean   : 37.53   Mean   :6.591   Mean   :22.2   Mean   : 23.12   Mean   :0.5272   Mean   :1
3.06
 3rd Qu.: 66.50   3rd Qu.:7.400   3rd Qu.:35.6   3rd Qu.: 24.70   3rd Qu.:0.7700   3rd Qu.:1
2.00
 Max.   :128.00   Max.   :9.100   Max.   :90.7   Max.   :152.40   Max.   :1.3300   Max.   :4
4.00
       X9              X10              X11              X12
 Min.   :0.0400   Min.   :0.0600   Min.   :0.0400   Min.   :0.0000
 1st Qu.:0.0900   1st Qu.:0.4800   1st Qu.:0.2500   1st Qu.:1.0000
 Median :0.2500   Median :0.8400   Median :0.4500   Median :1.0000
 Mean   :0.2798   Mean   :0.8745   Mean   :0.5132   Mean   :0.8113
 3rd Qu.:0.3300   3rd Qu.:1.3300   3rd Qu.:0.7000   3rd Qu.:1.0000
 Max.   :0.9200   Max.   :2.0400   Max.   :1.5300   Max.   :1.0000
```

Por úlitmo se muestra una gráfica de bigotes donde se puede observar que la variable X5, X6 y X8 tienen valores atípicos.

Code

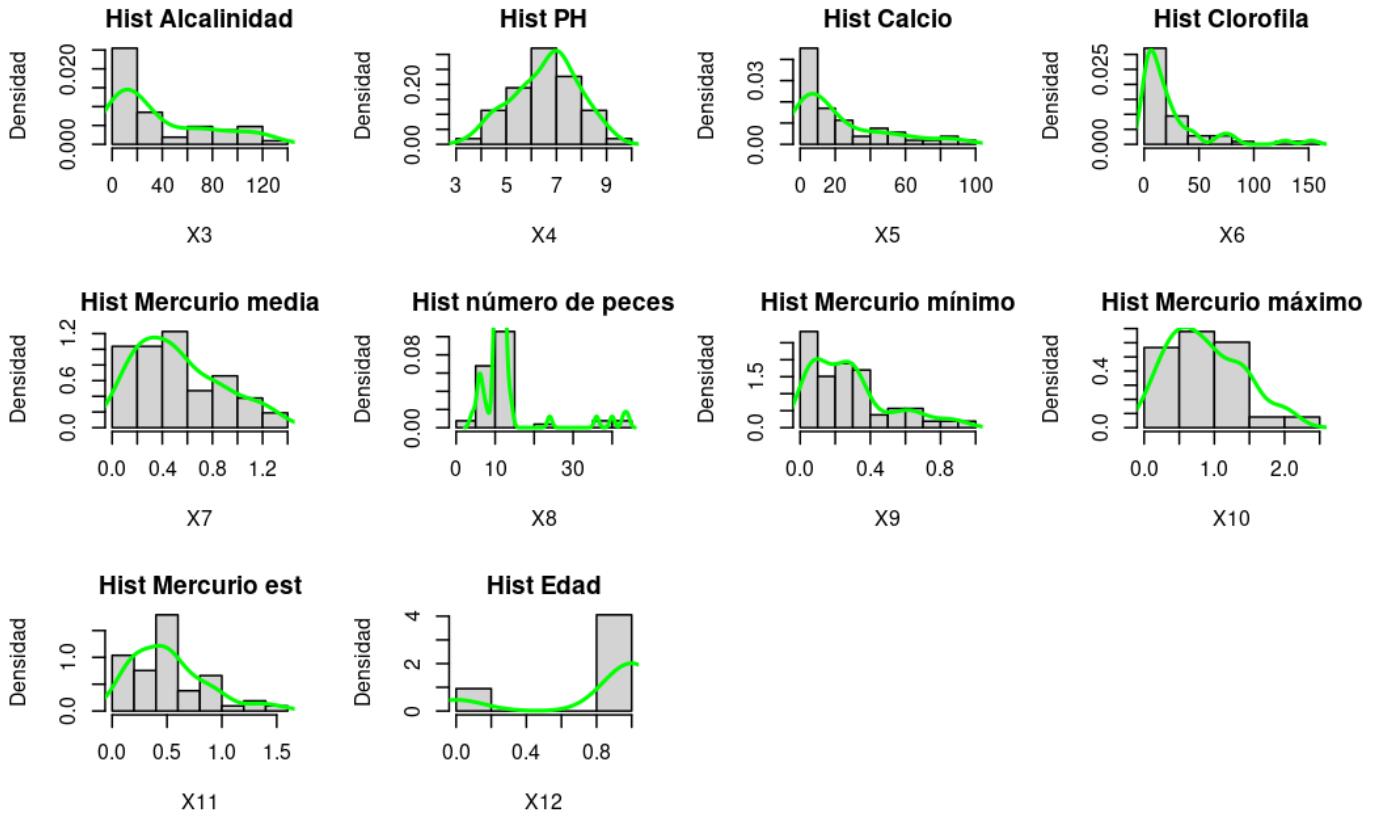


1. Análisis de normalidad de las variables continuas para identificar variables normales.

El análisis de normalidad es importate ya que varios algunas técnicas estadísticas clásicas asumen la normalidad univariante o multivariante de los datos. Las variables con comportamiento normal, se pueden identificar por medio de gráficas (Q-Q plot, distribución de frecuencia o densidad) y por medio de pruebas de normalidad como Shapiro-Wilk, Kolmogorov-Smirnov Anderson Darling, Mardia, etc. Estas pueden variar y usualmente se hacen más de una para comparar resultados, en caso de que no coincidan entonces no se puede concluir que las variables tengan una normalidad multivariada.

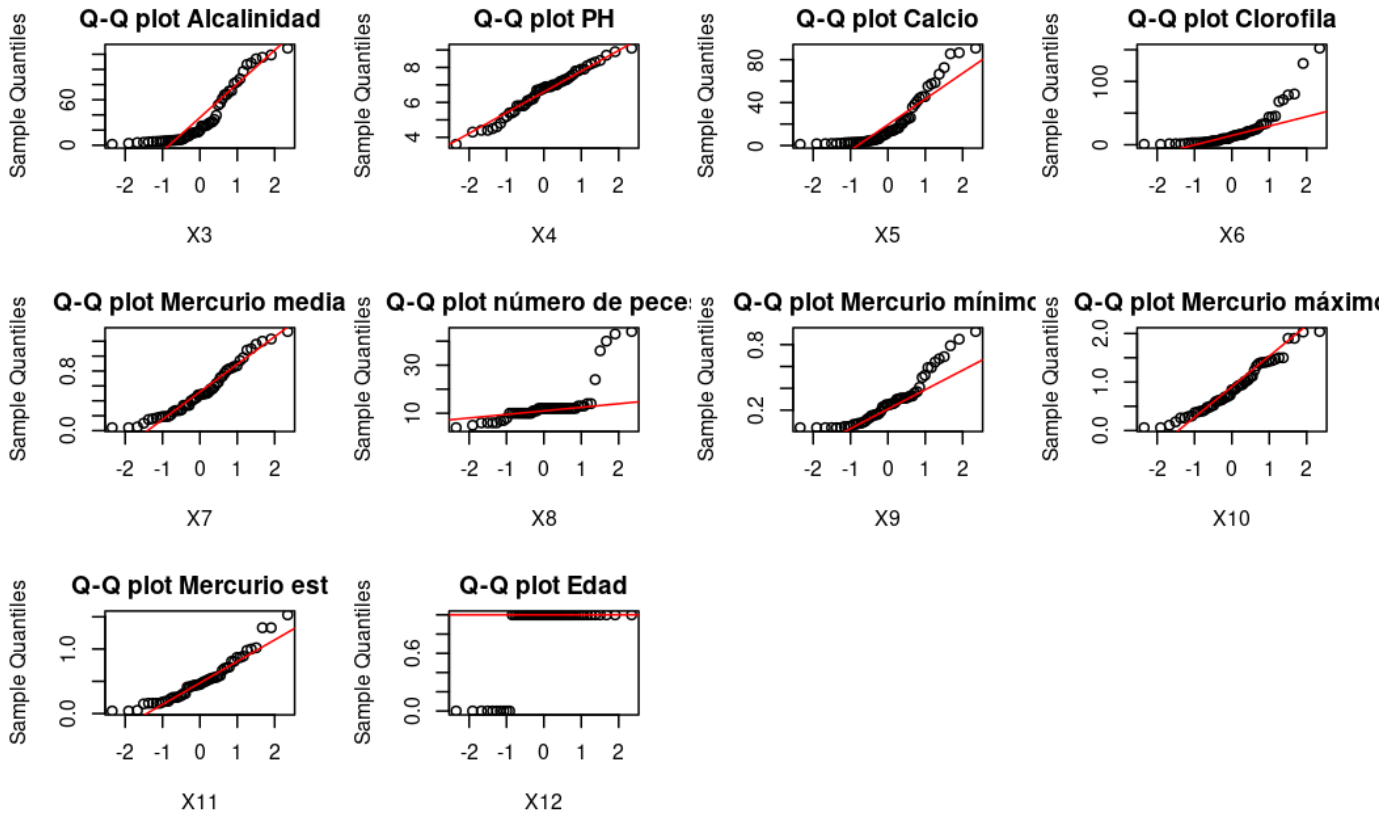
Primero vamos a observar los histogramas y tal vez solo observando X4, X7, X10 y X11 puedan tener una distribución normal debido a la densidad mostrada (linea verde)

Code
Code
Code
Code
Code
Code
Code
Code
Code
Code



Por otra parte, en una distibución normal iedal, los valores para los ejes x and y serán iguales, por lo que las observaciones estarán sobre la línea roja o con muy poca desviación. De acuerdo con esto las variables con posible normalidad serían igual X4, X7, X10 y X11 debido a que son los que se encuentran más cercanos a la linea roja y algunas otras variables como X6 tiene una cola al final salida lo cual hace que no se le considere una variable normal.

Code
Code
Code
Code
Code
Code
Code
Code
Code
Code



a) Prueba de normalidad de Mardia y de Anderson Darling

Ahora si a pesar de que ya sabemos que las variables X4, X7, X10 y X11 sean posiblemente normales hay que validarlo con algunas pruebas de normalidad en este caso vamos a usar Mardia para noralidad multivariante y Anderson Darling para normalidad univariable, en donde nos vamos a enfocar en el valor de P para aceptar o rechazar la hipótesis nula.

La hipótesis nula, la distribución de los datos es normal.
* $H_0: P > .05$
La hipóstesis alternativa, la distribución de los datos NO es normal.
* $H_1: P <= .05$

El valor de P es una medida de credibilidad de la hipótests nula, es decir si P > .05 es significativo por lo tanto tiene una distribución normal.De acuerdo con la prueba de Mardia se rechaza la hipótesis nula y se acepta la hipótesis alternativa, es decir los datos no son normales a pesar de presentar una Skewness (asimetría) con P-value mayor a .05, sin embargo presenta una kurtosis (mayor o menor concentración de datos alrededor de la media) menor a .05. Por lo tanto el resultado final es No tiene una normalidad multivariante.

Code

Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	474.747945136974	8.64265750182826e-21	NO
Mardia Kurtosis	3.59794900484947	0.000320736483631068	NO
MVN	*NA*	*NA*	NO

3 rows

Code

A continuación se observa el resultado de la prueba de normalidad univariante, en la cual si nos enfocamos solo en las variables antes visualizadas en las gráficas como posibles variable normales (X4, X7, X10 y X11), solo X4 y X10 son normales, X7 tiene un valor más cercano a .05 que otras variables pro aún asi se menor y X11 se aleja más que X7 con un valor de .0086 a comparación de X7 con .0174.

Code

	Test <S3: AsIs>	Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
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
10	Anderson-Darling	X12	14.3350	<0.001	NO

1-10 of 10 rows

b) Interpretación de las variables que sí tuvieron normalidad en los incisos anteriores.

A continuación se analizará el sesgo y la curtosis X4 y X10:

- X4 tiene una curtosis de -.62 y una asimetría de .96 indicando que se encuentra muy centrado ya que para estar perfectamente centrado debería de tener una asimetría de 1, el .96 es muy cercano.
- X10 tiene una curtosis de -.66 y una asimetría de .46 indicando que esta mas inclinada a la izquierda y presenta una cola a la derecha y estadísticamente el promedio es mayor que la mediana y la moda.

y debido a que ambas presentan una curtosis negativa (menor a 3) hay una menor concentración de datos en torno a la media y la gráfica es más achatada.

Code

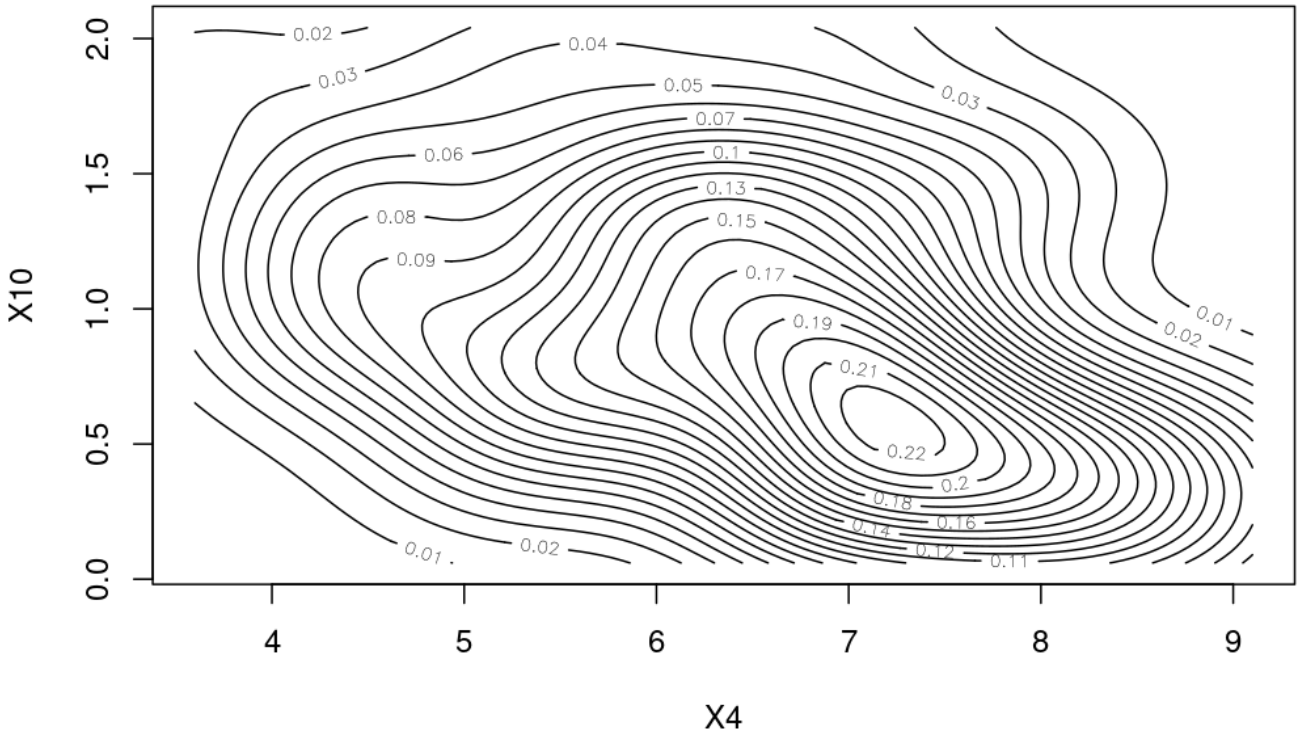
	n <int>	Mean <dbl>	Std.Dev <dbl>	Median <dbl>	Min <dbl>	Max <dbl>	25th <dbl>	75th <dbl>	Skew <dbl>
X3	53	37.5301887	38.2035267	19.60	1.20	128.00	6.60	66.50	0.9679170
X4	53	6.5905660	1.2884493	6.80	3.60	9.10	5.80	7.40	-0.2458771
X5	53	22.2018868	24.9325744	12.60	1.10	90.70	3.30	35.60	1.3045868
X6	53	23.1169811	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

	n <int>	Mean <dbl>	Std.Dev <dbl>	Median <dbl>	Min <dbl>	Max <dbl>	25th <dbl>	75th <dbl>	Skew <dbl>
X11	53	0.5132075	0.3387294	0.45	0.04	1.53	0.25	0.70	0.9449951
X12	53	0.8113208	0.3949977	1.00	0.00	1.00	1.00	1.00	-1.5465748

1-10 of 10 rows | 1-10 of 10 columns

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

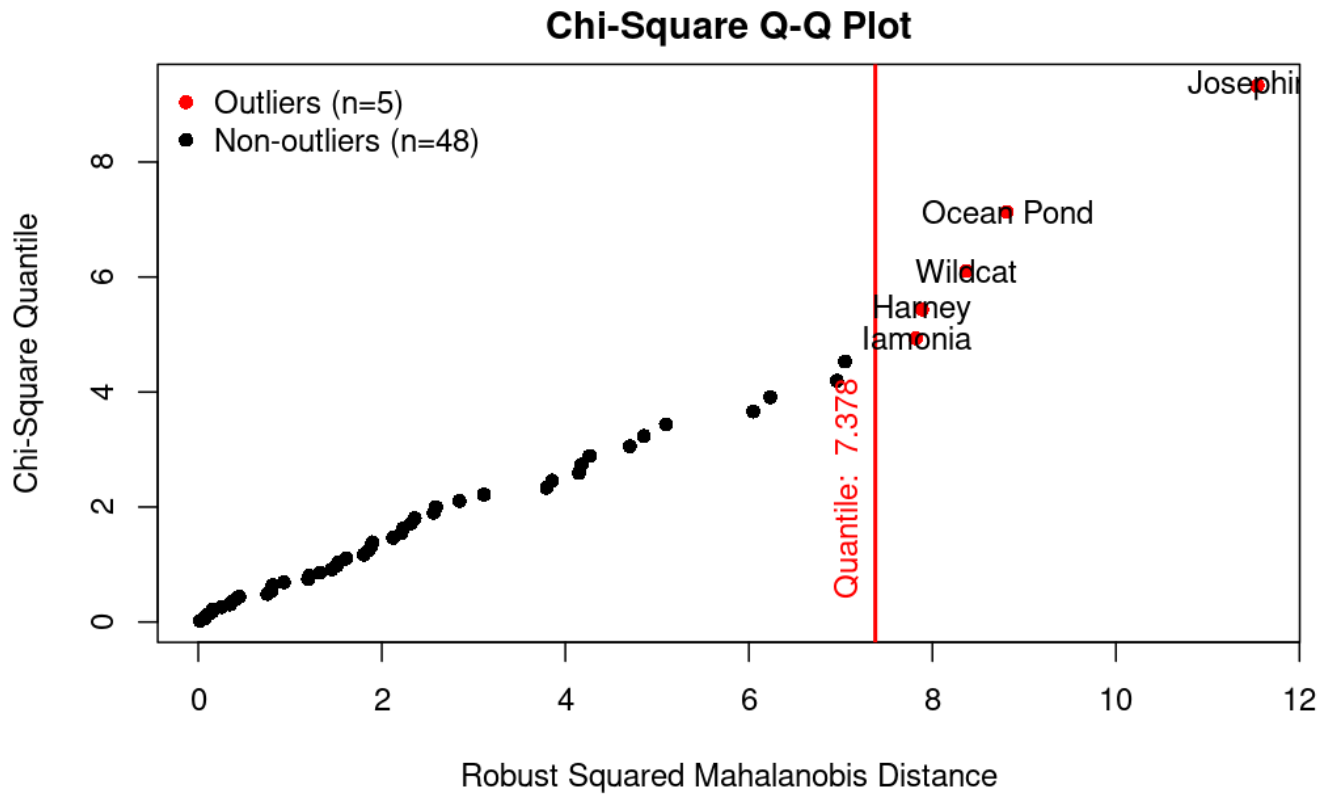
Code



d) Detecta datos atípicos o influyentes en la normal multivariada

De acuerdo con la Distancia de Mahalanobis hay 5 datos atípicos (Josephine, Ocean Pond, Wildcat, Harney y Iamonia) por lo que se eliminaran para así proceder al análisis de componentes.

Code



Code

Las dimensiones del dataset ahora son 48 renglones en vez de los 53 de un inicio

Code

```
[1] 48 10
```

2. Análisis de componentes principales

Un análisis de componentes principales tiene como objetivo reducir la dimensionalidad de la base de datos, mediante la reducción de variables pero manteniendo un alto porcentaje para pronosticar la variable objetivo.

a) Justificación del uso de componentes principales

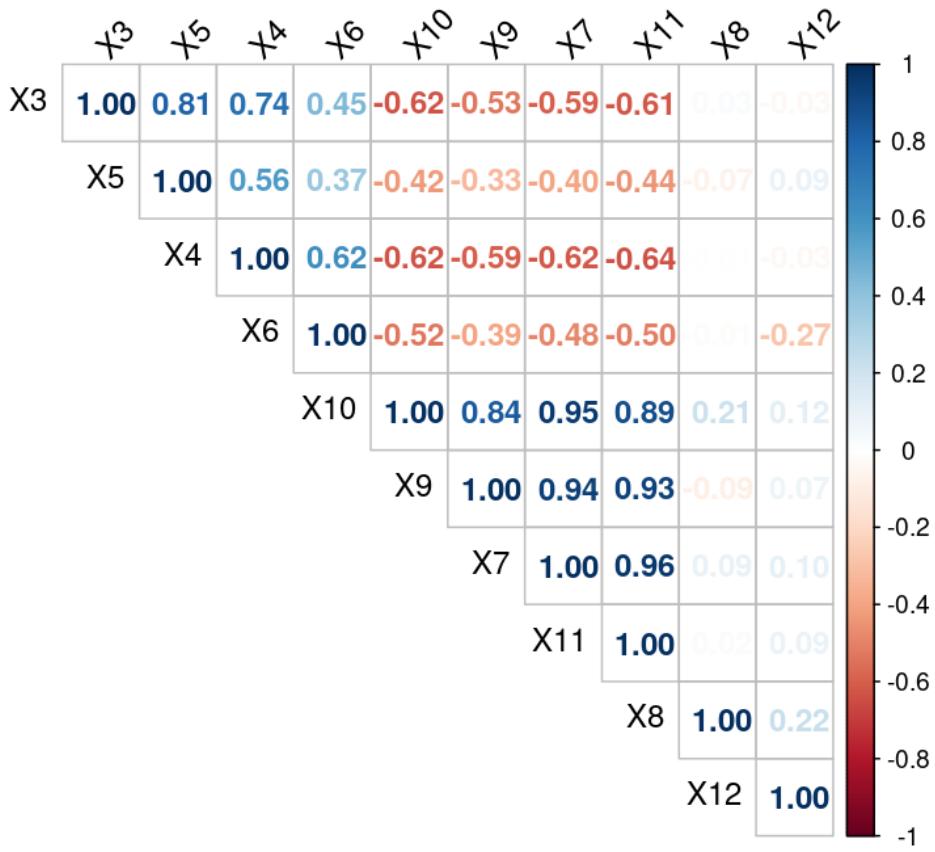
Matriz de varianza-covarianza

Code

```
             X3         X4          X5         X6          X7          X8           X9
X10
X3  1378.2030452 34.2205053 702.7471144 517.755465 -7.52179388   9.5695479 -4.585585106 -10.9
6703191
X4    34.2205053  1.5489184  16.4873316  24.009353 -0.26342642  -0.1029255 -0.173255319  -0.3
6549645
X5   702.7471144 16.4873316 550.3362722 275.498772 -3.19344193 -14.0222074 -1.796457447  -4.6
7742908
X6   517.7554654 24.0093528 275.4987722 980.740634 -5.19327704  -3.4014628 -2.891180851  -7.7
2930142
X7    -7.5217939 -0.2634264  -3.1934419  -5.193277  0.11756804   0.2770878  0.075367021   0.1
5407837
X8     9.5695479 -0.1029255 -14.0222074  -3.401463  0.27708777  79.7938830 -0.182021277   0.9
0414894
X9    -4.5855851 -0.1732553  -1.7964574  -2.891181  0.07536702  -0.1820213  0.054995745   0.0
9260851
X10  -10.9670319 -0.3654965  -4.6774291  -7.729301  0.15407837   0.9041489  0.092608511   0.2
2359716
X11   -7.7531090 -0.2715833  -3.5254832  -5.387600  0.11275488   0.0681117  0.074442553   0.1
4352837
X12   -0.4478723 -0.0141844   0.7514184  -3.240071  0.01322695   0.7553191  0.006382979   0.0
2134752
            X11          X12
X3  -7.75310904 -0.447872340
X4  -0.27158333 -0.014184397
X5  -3.52548316  0.751418440
X6  -5.38760018 -3.240070922
X7   0.11275488  0.013226950
X8   0.06811170  0.755319149
X9   0.07444255  0.006382979
X10  0.14352837  0.021347518
X11  0.11662535  0.011205674
X12  0.01120567  0.141843972
```

Se puede observar en la matriz de correlación que la variable X7 es la que tiene más correlación con las demás variables .95 con X10 y .94 con X9 mientras que las variables X8 y X12 son las que menos correlación con los demás tienen.

Code



La variable que más varia/cambia de un estado a otro, es la X3, la cual tiene 1378 de varianza y le sigue la variable X6 con 981

Code

```
         X3           X4           X5           X6           X7           X8           X9
X10
1.378203e+03 1.548918e+00 5.503363e+02 9.807406e+02 1.175680e-01 7.979388e+01 5.499574e-02 2.
235972e-01
         X11          X12
1.166254e-01 1.418440e-01
```

Code

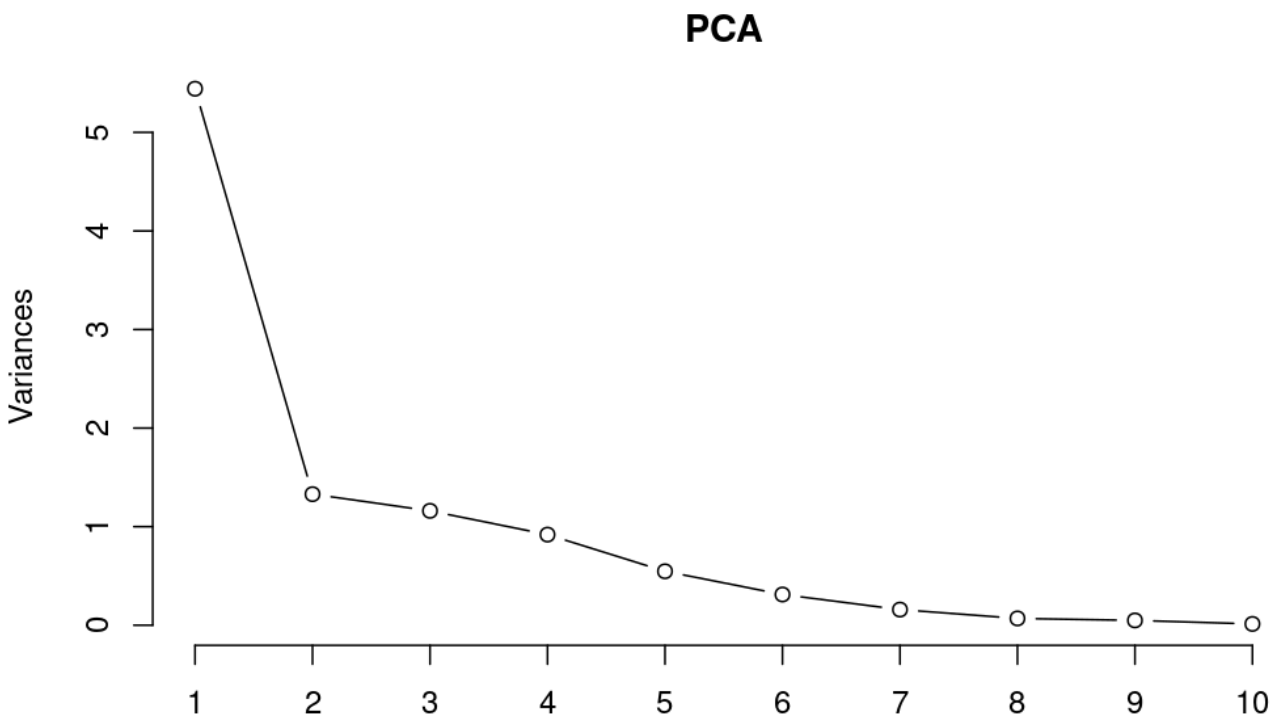
```
    X3     X4     X5     X6     X7     X8     X9    X10    X11    X12
"1378"    "2"  "550"  "981"    "0"   "80"    "0"    "0"    "0"    "0"
```

b) Justificación del número de componentes principales apropiados para reducir la dimensión de la base

La magnitud de las variables puede influir mucho en los resultados de la PCA. Por lo tanto hay que escalar (dividir por la desviación tipica) y centrar (restar la media) las funcionalidades es importante. Por lo que en la función prcomp ambos parametros son VERDADEROS.

A continuación se muestra una gráfica que representa el comportamiento de los componentes principales, en donde se puede observar que los primeros dos componentes aportan mucho más y que apartir del componente 6-7 ya no se logra observar mucho cambio.

Code



Después de observar la gráfica e imprimir el resumen de PCA, se decide que la mejor opción es escoger 6 componentes principales ya que con estos se logra un explicar hasta el 97% de los datos ya que con 7 componentes principales no varia, solo aumenta un 1%.

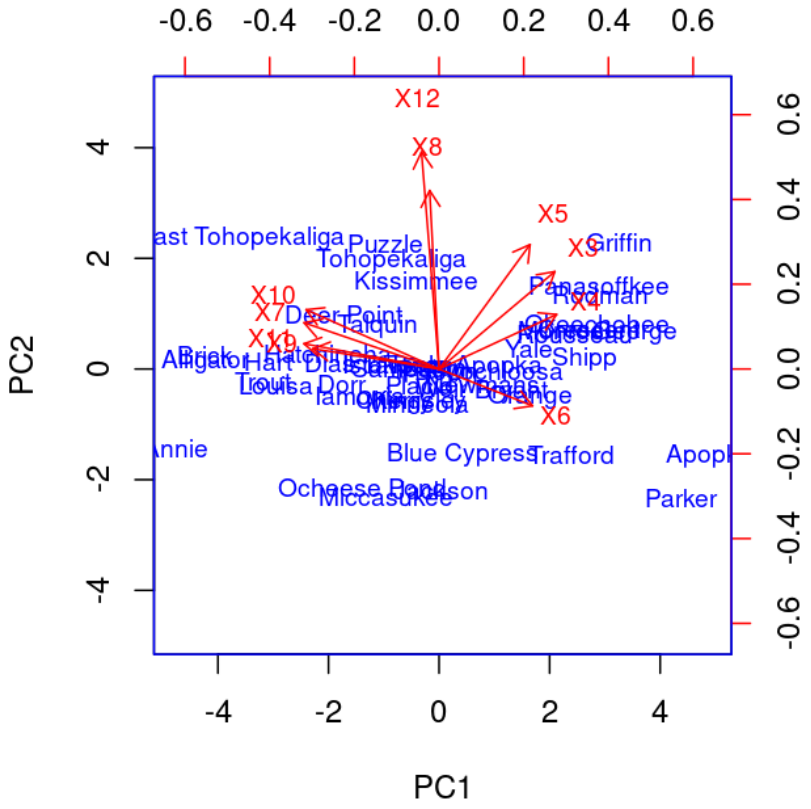
Code

```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7    PC8     PC9
PC10
Standard deviation     2.3330 1.1531 1.0772 0.95877 0.7396 0.55813 0.39796 0.2627 0.22104 0.1
1461
Proportion of Variance 0.5443 0.1330 0.1160 0.09192 0.0547 0.03115 0.01584 0.0069 0.00489 0.0
0131
Cumulative Proportion  0.5443 0.6773 0.7933 0.88521 0.9399 0.97106 0.98690 0.9938 0.99869 1.0
0000
```

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

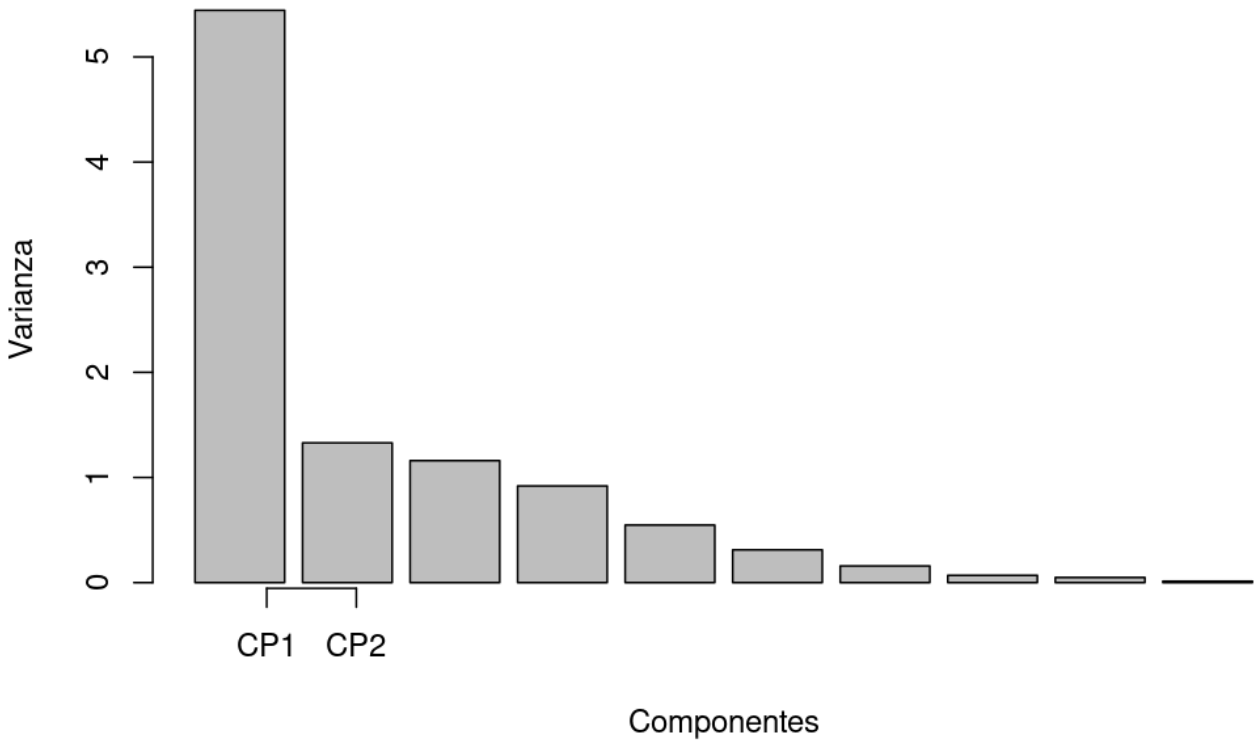
En esta gráfica se observar cómo la mayor parte de la variabilidad está dada por el componente 1 y el componente 2 lo hacer en un rango mucho más pequeño. Tambien se puede observar la correlación de algunas variables como X8 y X12.

Code



En esta otra gráfica, se muestra que la varianza del componente principal 1 es mucho mayor que las del resto de componente incluido el componente 2. Con esta observación se puede decir que con tan solo el primer componente es “suficiente” y el resto puede llegar a ser irrelevante para el análisis. Sin embargo el dejar solo un componente depende del tipo de problema al que este relacionado si es algo medico o con mucho riesgo, el 54% no es suficiente.Para este caso en particular vamos a dejar 6 componentes para reducir el dataset pero aún asi mantener arriba del 95% de precisión.

Code



3. Conclusión general

El análisis de componentes principales ayuda mucho a reducir el dataset, en algunos problemas el tener muchas variables puede afectar de manera negativa el resultado, dando paso a que se de overfittin o underfitting, por lo que el pre procesamiento es indispensable y el reducir estas variable puede ayudar a tener una mayor precisión sin sobrecargar un modelo con tantas variables. Aparte te puede ayudar a identificar las variables que más influyen y tomarlo en cuenta para cualquier toma de decisión.

Como ya se mencionó durante el reporte, se puede decir que el nivel de contaminación por mercurio en los peces es dada mas del 50% por un solo componente, por lo que se puede reducir drasticamente la contaminación reduciendo o quitando por completo este elemento.

Finalmente el estudio de normalidad es importante para algunos procedimientos estadisticos ya que muchas veces estos infieren que las variables se comportan de manera normal.