Preprocesamiento de datos multivariados o de una serie temporal

Samuel Méndez Villegas - A01652277

En este reporte, se presenta la continuación del trabajo “Análisis para determinar lo principales factores que influyen en el nivel de contaminación en los peces de los lagos de Florida”. La principal diferencia con el proyecto anterior, es que aquí se amplía el análisis al hacer uso de herramientas estadísticas más avanzadas como lo son los análisis multivariados y componentes principales.

El objetivo del trabajo es de igual forma determinar cuáles son los principales factores que influyen en el nivel de contaminación en los peces de los lagos de Florida. Adicionalmente, se buscará resolver algunas interrogantes que se derivan de la problemática principal, cómo lo son: ¿Hay evidencia para suponer que la concentración promedio de mercurio en los lagos es danñino para la salud humana? ¿Habrá diferencia significativa entre la concentración de mercurio por la edad de los peces? ¿Habrá influencia del número de peces encontrados en la concentración de mercurio en los peces? ¿Las concentraciones de alcalinidad, clorofila, calcio en el agua del lago influyen en la concentración de mercurio de los peces?

Carga de la base de datos

Primeramente se cargará la base de datos con la que se estará trabajando. Observemos que es la misma que la utilizada en trabajos anteriores.

```
X = read.csv("mercurio.csv")
head(X)
```

```
##   X1           X2    X3  X4   X5    X6   X7 X8   X9  X10  X11 X12
## 1  1    Alligator   5.9 6.1  3.0   0.7 1.23  5 0.85 1.43 1.53   1
## 2  2        Annie   3.5 5.1  1.9   3.2 1.33  7 0.92 1.90 1.33   0
## 3  3       Apopka 116.0 9.1 44.1 128.3 0.04  6 0.04 0.06 0.04   0
## 4  4 Blue Cypress  39.4 6.9 16.4   3.5 0.44 12 0.13 0.84 0.44   0
## 5  5        Brick   2.5 4.6  2.9   1.8 1.20 12 0.69 1.50 1.33   1
## 6  6       Bryant  19.6 7.3  4.5  44.1 0.27 14 0.04 0.48 0.25   1
```

```
# Número de variables
len = ncol(X)
# Total de registro de la base de datos
rows = nrow(X)

cat('El número de variables dentro de la base de datos es:', len, '\n')
```

```
## El número de variables dentro de la base de datos es: 12
```

```
cat('El número de registros dentro de la base de datos es:', rows)
```

```
## El número de registros dentro de la base de datos es: 53
```

Como se observa, se tienen en total 12 variables distintas. A continuación se presenta el significado de cada una:

- 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)

Exploración de las variables

El data set cuenta con 12 variables, las cuales se clasifican de la siguiente forma:

- **Variables cualitativas:**
  - Número de identificación (ordinal)
  - Nombre del lago (nominal)
  - Indicador de la edad de los peces (0: jóvenes; 1: maduros) (nominal)
- **Variables cuantitativas:**
  - Alcalinidad (mg/l de carbonato de calcio) (escala de razón)
  - PH (escala de razón)
  - Calcio (mg/l)
  - Clorofila (mg/l)
  - Concentración media de mercurio (parte por millón) en el tejido muscualar del grupo de peces estudiados en cada lago (escala de razón)
  - Número de peces estudiados en el lago (escala de razón)
  - Mínimo de la concentración de mercurio en cada grupo de peces (escala de razón)
  - Máximo de la concentración de mercurio en cada grupo de peces (escala de razón)
  - Estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible) (escala de razón)

El análisis descriptivo de estas variables se encuentra en el trabajo anterior, por lo que dicha parte se omitirá en este trabajo y se pasará a realizar análisis utilizando estadística multivariada.

Análisis de normalidad

Matriz de covarianzas y varianzas, y matriz de correlación

Para realizar las pruebas de normalidad, primeramente se obtendrá tanto la matriz de varianzas y covarianzas como la matriz de correlaciones. De igual forma, se obtendrá los vectores de media de cada una de las variables cuantitativas.

```
M = X[c("X3", "X4", "X5", "X6", "X7", "X8", "X9", "X10", "X11", "X12")]

# Vector de medias. Se calcula la media de cada columna
mius = colMeans(M)
cat('El vector de medias es:\n', mius)
```

```
## El vector de medias es:
##  37.53019 6.590566 22.20189 23.11698 0.5271698 13.0566 0.2798113 0.8745283 0.5132075
0.8113208
```

```
# Se obtiene la matriz de varianza y covarianza entre las variables.
cat('\n\nLa matriz de varianzas y covarianzas es:\n')
```

```
##
##
## La matriz de varianzas y covarianzas es:
```

```
S = cov(M)
S
```

```
##              X3          X4         X5         X6          X7           X8
## X3  1459.509456 35.39971335 793.065711 562.193324 -7.73773984   3.36556604
## X4    35.399713  1.66010160  18.540018  24.159971 -0.25283491  -0.20522496
## X5   793.065711 18.54001814 621.633266 314.949198 -3.40693687 -19.07703193
## X6   562.193324 24.15997097 314.949198 949.645668 -5.16408563  -3.11828737
## X7    -7.737740 -0.25283491  -3.406937  -5.164086  0.11630530   0.23074020
## X8     3.365566 -0.20522496 -19.077032  -3.118287  0.23074020  73.28519594
## X9    -4.544071 -0.15809797  -1.876788  -2.793997  0.07159176  -0.15825835
## X10  -12.062062 -0.37116800  -5.309432  -7.802021  0.16305729   0.71993106
## X11   -8.126195 -0.26746916  -3.922122  -5.286440  0.11080733   0.07481495
## X12   -1.432656  0.01933962  -0.020791  -3.444811  0.01464804   0.70319303
##               X9          X10         X11          X12
## X3  -4.544071118 -12.06206241 -8.12619485 -1.432656023
## X4  -0.158097968  -0.37116800 -0.26746916  0.019339623
## X5  -1.876788099  -5.30943179 -3.92212155 -0.020791001
## X6  -2.793996734  -7.80202068 -5.28644013 -3.444811321
## X7   0.071591763   0.16305729  0.11080733  0.014648041
## X8  -0.158258345   0.71993106  0.07481495  0.703193033
## X9   0.051259579   0.09046049  0.07048523  0.009002177
## X10  0.090460486   0.27253295  0.15203327  0.019332366
## X11  0.070485232   0.15203327  0.11473759  0.011962990
## X12  0.009002177   0.01933237  0.01196299  0.156023222
```

```
cat('\n\nLa matriz de correlación es:\n')
```

```
##
##
## La matriz de correlación es:
```

```
P = cor(M)
P
```

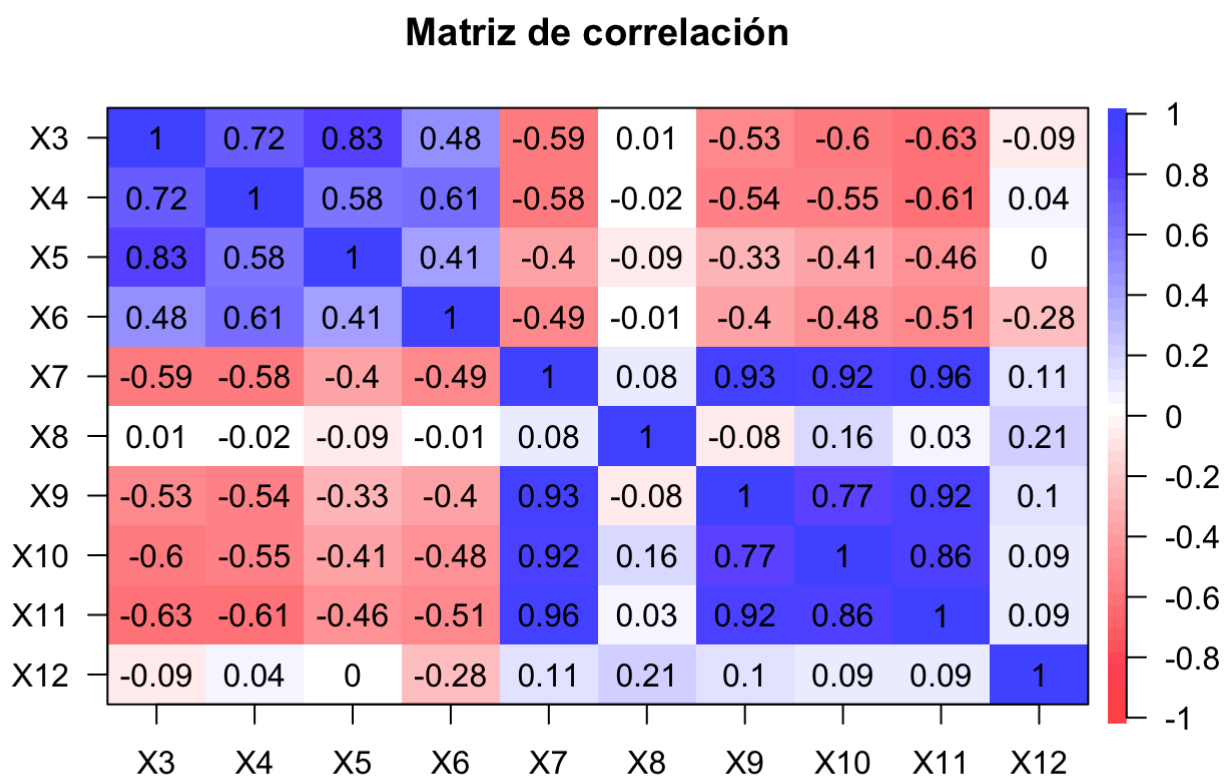
```
##              X3          X4           X5          X6          X7          X8
## X3   1.00000000  0.71916568  0.832604192  0.47753085 -0.59389671  0.01029074
## X4   0.71916568  1.00000000  0.577132721  0.60848276 -0.57540012 -0.01860607
## X5   0.83260419  0.57713272  1.000000000  0.40991385 -0.40067958 -0.08937901
## X6   0.47753085  0.60848276  0.409913846  1.00000000 -0.49137481 -0.01182027
## X7  -0.59389671 -0.57540012 -0.400679584 -0.49137481  1.00000000  0.07903426
## X8   0.01029074 -0.01860607 -0.089379013 -0.01182027  0.07903426  1.00000000
## X9  -0.52535654 -0.54196524 -0.332476229 -0.40045856  0.92720506 -0.08165278
## X10 -0.60479558 -0.55181523 -0.407916635 -0.48497215  0.91586397  0.16109174
## X11 -0.62795845 -0.61284905 -0.464409465 -0.50644193  0.95921481  0.02580046
## X12 -0.09493882  0.03800021 -0.002111124 -0.28300234  0.10873896  0.20795617
##              X9         X10          X11          X12
## X3  -0.52535654 -0.60479558 -0.62795845 -0.094938825
## X4  -0.54196524 -0.55181523 -0.61284905  0.038000214
## X5  -0.33247623 -0.40791663 -0.46440947 -0.002111124
## X6  -0.40045856 -0.48497215 -0.50644193 -0.283002338
## X7   0.92720506  0.91586397  0.95921481  0.108738958
## X8  -0.08165278  0.16109174  0.02580046  0.207956171
## X9   1.00000000  0.76535319  0.91908939  0.100661967
## X10  0.76535319  1.00000000  0.85975810  0.093752072
## X11  0.91908939  0.85975810  1.00000000  0.089411267
## X12  0.10066197  0.09375207  0.08941127  1.000000000
```

También, la matriz de correlación se puede visualizar con colores, en donde éste indica qué tan correlacionadas están las variables entre ellas. Mientras más cercano a 1 o -1 sea el valor del coeficiente de correlación, más linealmente relacionadas estarán las variables.

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.2
```

```
corPlot(P, cex = 1, main = "Matriz de correlación")
```



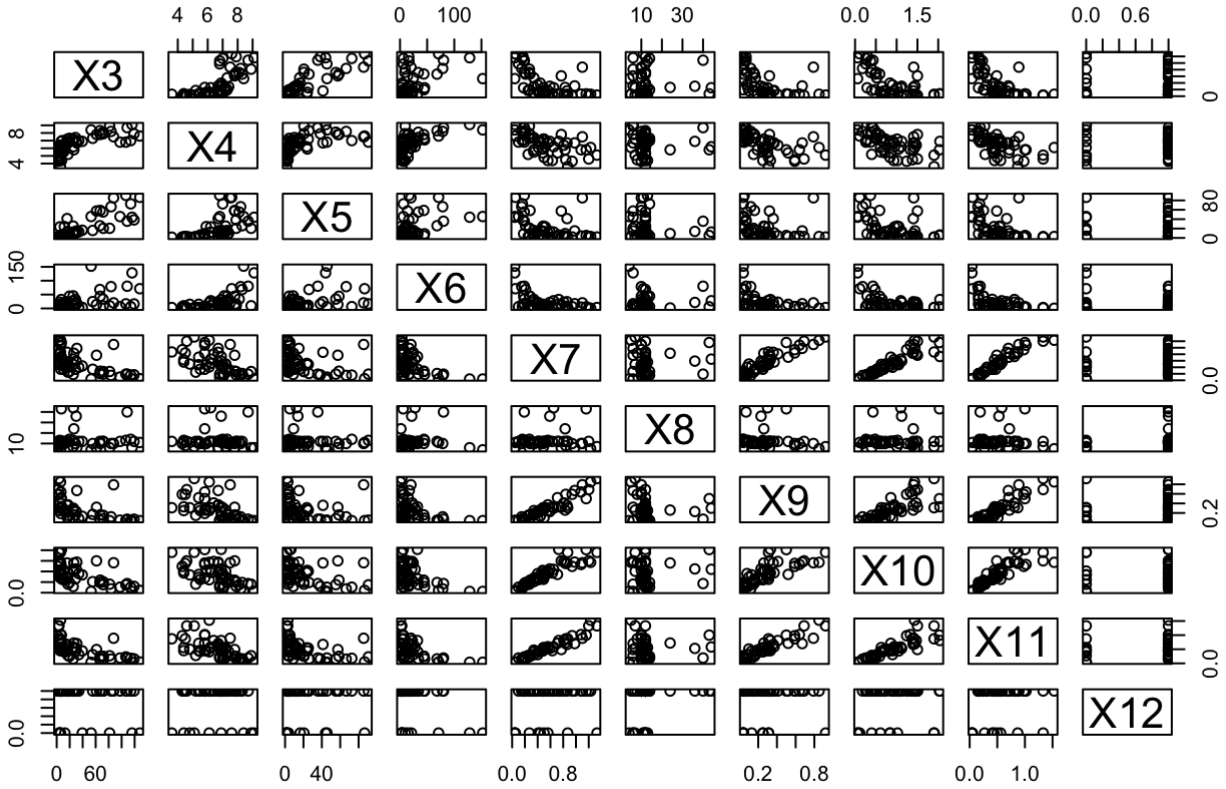
La matriz de correlaciones se describió a detalle en el trabajo anterior, por lo que a continuación se analizará la matriz de varianzas y covarianzas. Esta matriz como su nombre lo indica, contiene las covarianzas y varianzas entre cada una de las variables. Las varianzas se ubican en la diagonal principal e indican la variabilidad del conjunto de datos respecto a su media. En este contexto, la varianza se podría definir como la covarianza entre la misma variable.

Por otro lado, las covarianzas entre variables se expresan en el resto de la matriz (es una matriz simétrica), en donde se expresa qué tanto varían las variables en forma conjunta respecto a sus medias. Una covarianza positiva indica que mientras el valor de una variable crece, la otra variable de igual forma crece. Por el otro lado, una covarianza negativa indica que mientras una variable aumenta, la otra disminuye. La covarianza entre las variables nos resulta muy útil para determinar si dos variables son independientes entre sí, pues si la covarianzas es distinta de 0, se indica que existe una relación entre la variables y por lo tanto son dependientes. Cabe mencionar que el signo de la covarianza es el que determina de igual forma el signo de la correlación entre variables.

Ya obtenidas las matrices, se pueden graficar todas las variables entre sí con el propósito de ver cómo se comporta cada variable con respecto a las demás. De igual forma si en alguna gráfica se obtiene una nube de puntos similar a una elipse, podríamos estar hablando de una posible distribución normal bivariada.

```
## Se grafica todos las vriables contra todas las variables
pairs(M, main  = "Diagramas de dispersión entre variables")
```

Diagramas de dispersión entre variables



En el gráfico anterior, se muestran los diagramas de dispersión entre todas variables numéricas de la base de datos. Al ser tantas las variables, no se alcanza apreciar del todo bien el comportamiento de las nubes de puntos, sin embargo, se puede observar cierta linealidad entre $X7$, $X9$, $X10$ y $X11$. Esto tiene sentido, pues son las variables que están relacionadas de una forma con la concentración media de mercurio.

Las gráficas anteriores se pueden visualizar de mejor forma graficando por pedazos, u omitiendo algunas variables que se sabe que no es posible formar un comportamiento lineal, como lo es el caso de $X12$, pues esta solo puede tener 2 valores. Por lo tanto, se podría representar de la siguiente forma:

```
# install.packages("GGally")
library(GGally)
```

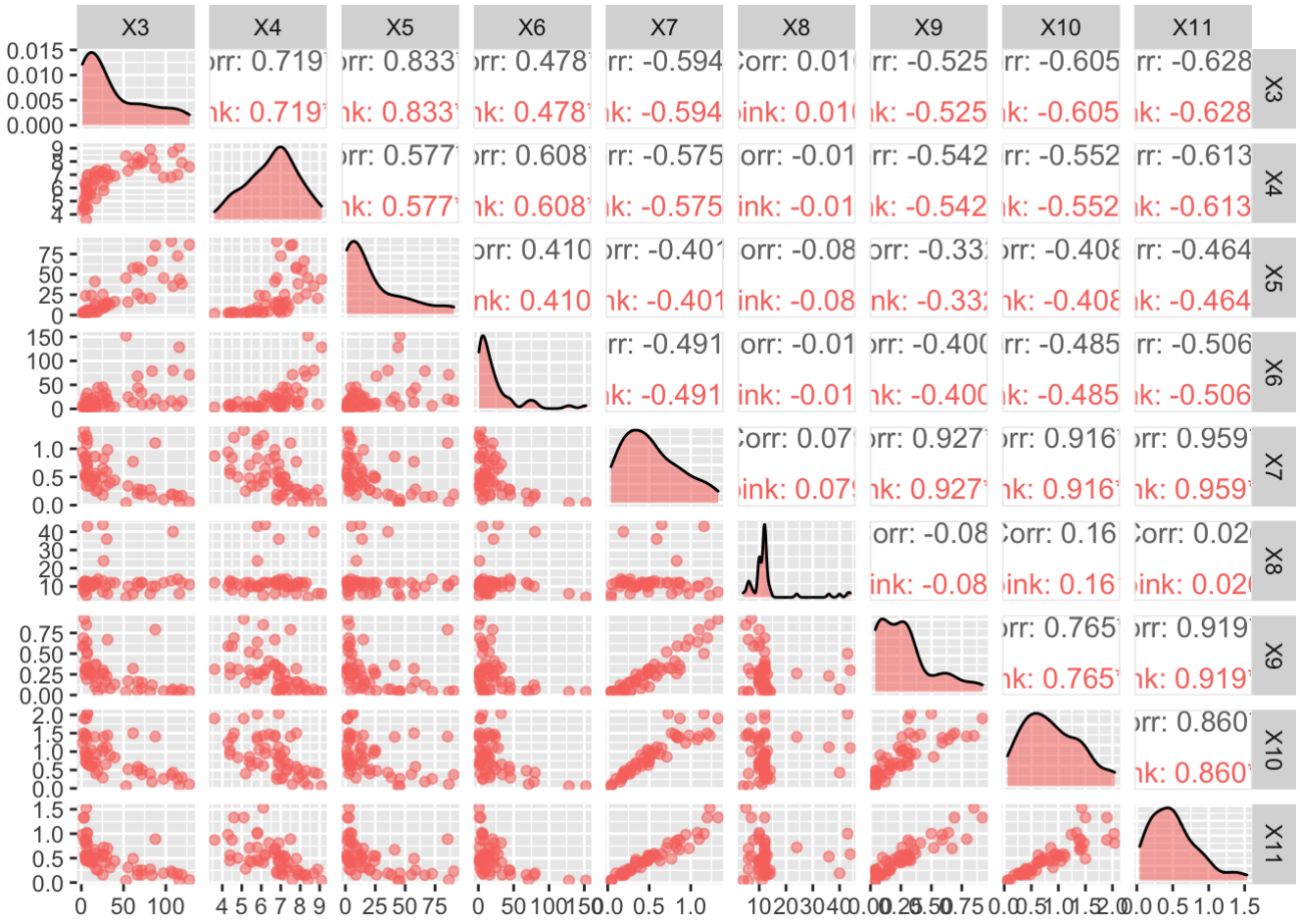
```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(X, columns = 3:11, aes(color = 'pink',
                               alpha = 0.5))
```



Pruebas de normalidad

Una vez realizados los gráficos, se pasará a realizar las pruebas de normalidad de Mardia y de Anderson Darling con el objetivo de identificar las variables que son normales y detectar posible normalidad multivarida de grupos de variables.

```
## Se utiliza la siguiente librería para aplicar pruebas de normalidad multivariada
library(MVN)
## Test de Multinomalidad: Método Sesgo y kurtosis de Mardia
mvn(M, subset = NULL, mvn = "mardia", covariance = FALSE, showOutliers = FALSE)
```

```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 474.747945136975 8.64265750182786e-21     NO
## 2 Mardia Kurtosis 3.59794900484948 0.000320736483631068     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##                Test  Variable Statistic   p value Normality
## 1  Anderson-Darling    X3        3.6725  <0.001      NO
## 2  Anderson-Darling    X4        0.3496  0.4611      YES
## 3  Anderson-Darling    X5        4.0510  <0.001      NO
## 4  Anderson-Darling    X6        5.4286  <0.001      NO
## 5  Anderson-Darling    X7        0.9253  0.0174      NO
## 6  Anderson-Darling    X8        8.6943  <0.001      NO
## 7  Anderson-Darling    X9        1.9770  <0.001      NO
## 8  Anderson-Darling    X10       0.6585   0.081      YES
## 9  Anderson-Darling    X11       1.0469  0.0086      NO
## 10 Anderson-Darling    X12      14.3350  <0.001      NO
##
## $Descriptives
##      n       Mean    Std.Dev Median  Min    Max  25th  75th       Skew
## X3  53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50  0.9679170
## X4  53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40 -0.2458771
## X5  53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60  1.3045868
## X6  53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70  2.4130571
## X7  53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77  0.5986343
## X8  53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00  2.5808773
## X9  53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33  1.0729099
## X10 53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33  0.4645925
## X11 53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70  0.9449951
## X12 53  0.8113208  0.3949977   1.00 0.00   1.00  1.00  1.00 -1.5465748
##       Kurtosis
## X3  -0.4705349
## X4  -0.6239638
## X5   0.6130359
## X6   6.1042185
## X7  -0.6312607
## X8   6.0089455
## X9   0.4060828
## X10 -0.6692490
## X11  0.5733500
## X12  0.4005116
```

La celda de código anterior expresa los resultados obtenidos de la pruebas de normalidad de Mardia y de Anderson Darling. En la primera salida se pueden observar los resultados de Mardia, en los cuales se observa que se toman como estadísticos de prueba el sesgo y la kurtosis. Se observa que el $p - value$ del sesgo es alto, al igual que su estadístico correspondiente, indicando asi que se tiene una gran variación en los datos. Por lo tanto, no se tiene una distribución simétrica. En el caso de la kurtosis, que simboliza que tan achatada está la campana formada, se tiene un valor pequeño en el $p - value$ el cual debería de ser mayor 0.05, suponiendo que

estamos trabajando con una confianza del 95%. Por lo tanto, al tener una kurtosis relativamente pequeña, se dice que está achatada la campana. De esta forma, se concluye que de esta prueba, se dice que no hay normalidad multivariada entre las 10 variables numéricas analizadas.

El siguiente análisis que se realiza es el de Anderson Darling, el cual analiza variable por variable con el objetivo de determinar cuáles variables siguen una distribución normal. En este caso únicamente se obtuvo que $X4$ y $X10$ se comportan de forma normal. Dichas variables corresponden al PH del lago, y al máximo de la concentración de mercurio en cada grupo de peces respectivamente.

Finalmente, la última pestaña que se arroja como resultado muestra un resumen de alguna medidas de tendencia central, dispersión y posición de los datos, así como la kurtosis y el sesgo de cada una.

Sabiendo que no se tiene una normalidad multivariada entre las 10 variables, se volverá a realizar el test, pero ahora utilizando únicamente aquellas variables que se comportan normalmente según la prueba de Anderson-Darling, es decir $X4$ y $X10$.

```
## Test de Multinomalidad: Método Sesgo y kurtosis de Mardia
M_subconjunto = X[c("X4", "X10")]

mvn(M_subconjunto, subset = NULL, mvn = "mardia", covariance = FALSE, showOutliers = FAL
SE)
```

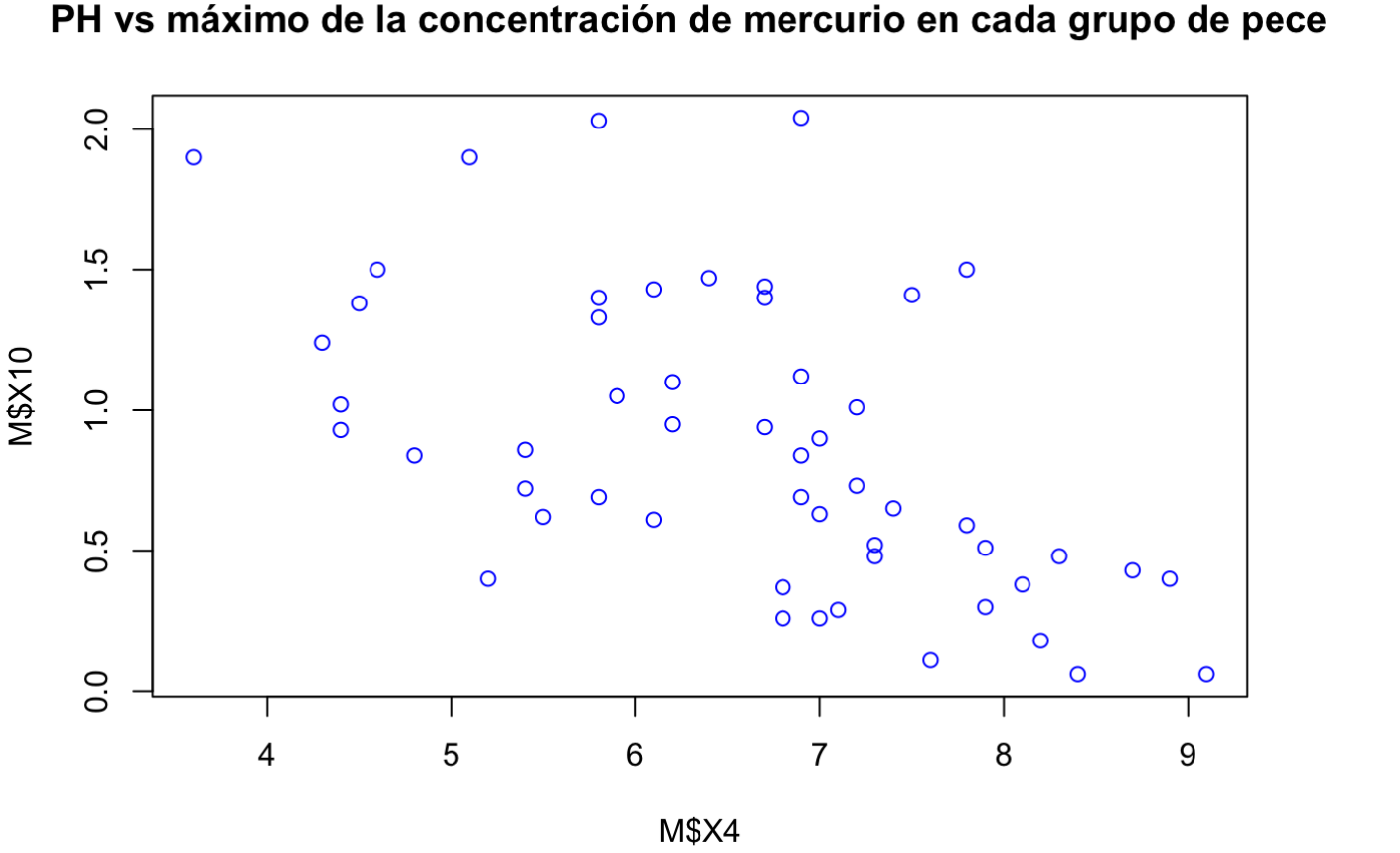
```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  6.17538668676458 0.186427564928852    YES
## 2 Mardia Kurtosis -1.12820795824432  0.25923210375991    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    X4        0.3496    0.4611    YES
## 2 Anderson-Darling    X10       0.6585    0.0810    YES
##
## $Descriptives
##      n      Mean   Std.Dev Median  Min  Max 25th 75th       Skew   Kurtosis
## X4  53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -0.2458771 -0.6239638
## X10 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925 -0.6692490
```

En este caso, en el test de Mardia se obtuvo que efectivamente las variables tienen un comportamiento binormal, ya que los $p - values$ de la kurtosis y del sesgo son mayores a 0.05. Tenemos que la curtosis obtenida es negativa, lo que indica que los datos presentan valores atípicos menos extremados a una distribución normal. En cuanto al sesgo, este se nota que es mucho menor que el obtenido con todas las variables, indicando así que no se tiene una gran variación entre los datos. Los resultados de los demás componentes siguen manteniéndose iguales. Por lo tanto, se puede decir que la combinación de estas variables forman una distribución normal bivariada, la cual se pasará a graficar a continuación.

Gráficos de normalidad

Primeramente, se grafica el diagrama de dispersión entre las variables:

```
plot(M$X4, M$X10,
     main = "PH vs máximo de la concentración de mercurio en cada grupo de peces", col =
'blue')
```



Como se observa, los datos forman más o menos una elipse diagonal, lo cual es sinónimo de una distribución normal bivariada. Dicho gráfico también se puede visualizar en 3D para que se observe de mejor forma la elipse que se forma, así como la kurtosis y el sesgo de la distribución bivariada.

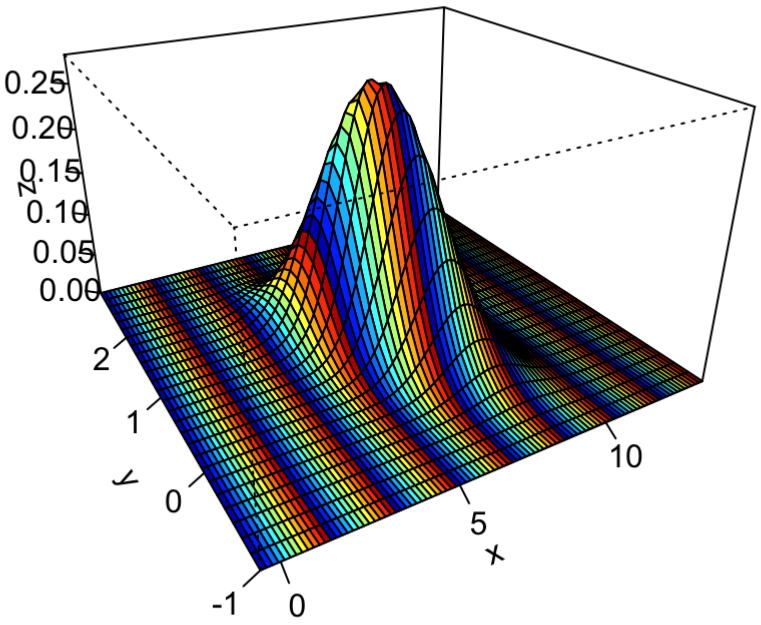
```
library(colorRamps)
```

```
## Warning: package 'colorRamps' was built under R version 4.1.2
```

```
library(mnormt)
```

```
## Warning: package 'mnormt' was built under R version 4.1.2
```

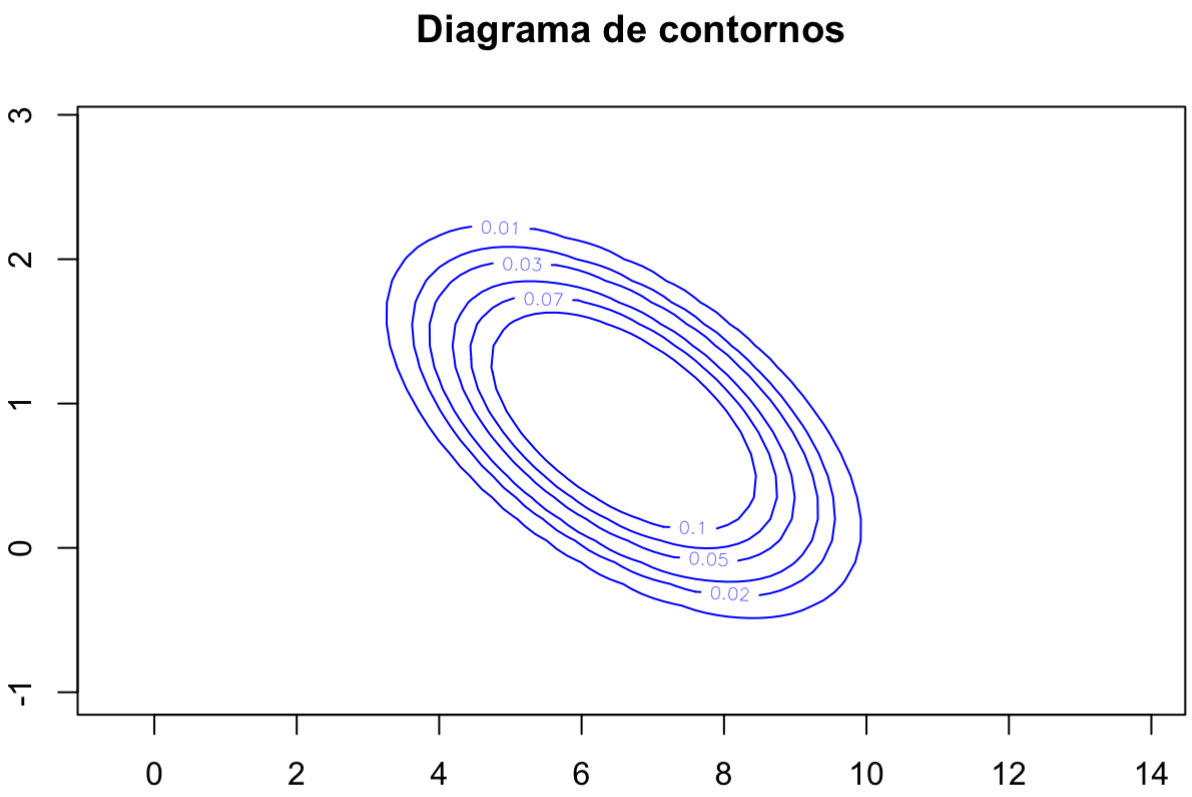
```
X = colMeans(M_subconjunto)
S = cov(M_subconjunto)

## Vectores de limites de graficación de las distribuciones
x = seq(-0.5, 14, 0.15)
y = seq(-1, 3, 0.15)

f = function(x, y) dmnorm(cbind(x, y), X, S)
z = outer(x, y, f)

## Se crea la gráfcia correspondiente
persp(x, y, z, theta=-30, phi=25, expand=0.6, ticktype='detailed', col = matlab.like(12
))
```



En el gráfico anterior se puede observar la campana que se forma por la distribución normal bivariada. Como se observa, no se tiene tanto sesgo, pues se muestra como una campana simétrica. Sobre la kurtosis, se nota que se tiene un valor muy pequeño, dado a que la campana no está tan achatada. De igual forma, se puede ver que la campana está colocada diagonalmente, al igual que la nube de puntos que se pudo observar al graficar $X4$ contra $X10$.

Ahora, se pasará a realizar la gráfica de contorno de la normal multivariada con el objetivo de visualizar más claramente la elipse que se forma.

```
#create contour plot
contour(x, y, z, col = 'blue', main = 'Diagrama de contornos',levels = c(0.01, 0.02, 0.0
3, 0.05, 0.07, 0.1))
```



La gráfica anterior representan los contornos de la ditribución bivariada. Se observa que los contornos tienen forma de elipse, lo que se esperaría de una distribución bivariada. La dirección en la que se expanden las elipses coinciden con los datos del diagrama de dispersión.

Valores atípicos

Finalmente, se pasará a detectar aquellos valores atípicos que se presentan en la distribución. Si observamos de vuelta el diagrama de dispersión entre las variables, se nota que se tienen algunos puntos que no están del todo cercanos a los demás, pudiéndose trata de valores atípicos. Por lo tanto, para determinar estos datos, se hará uso de la distancia de Mahalanobis y del gráfico QQplot multivariado.

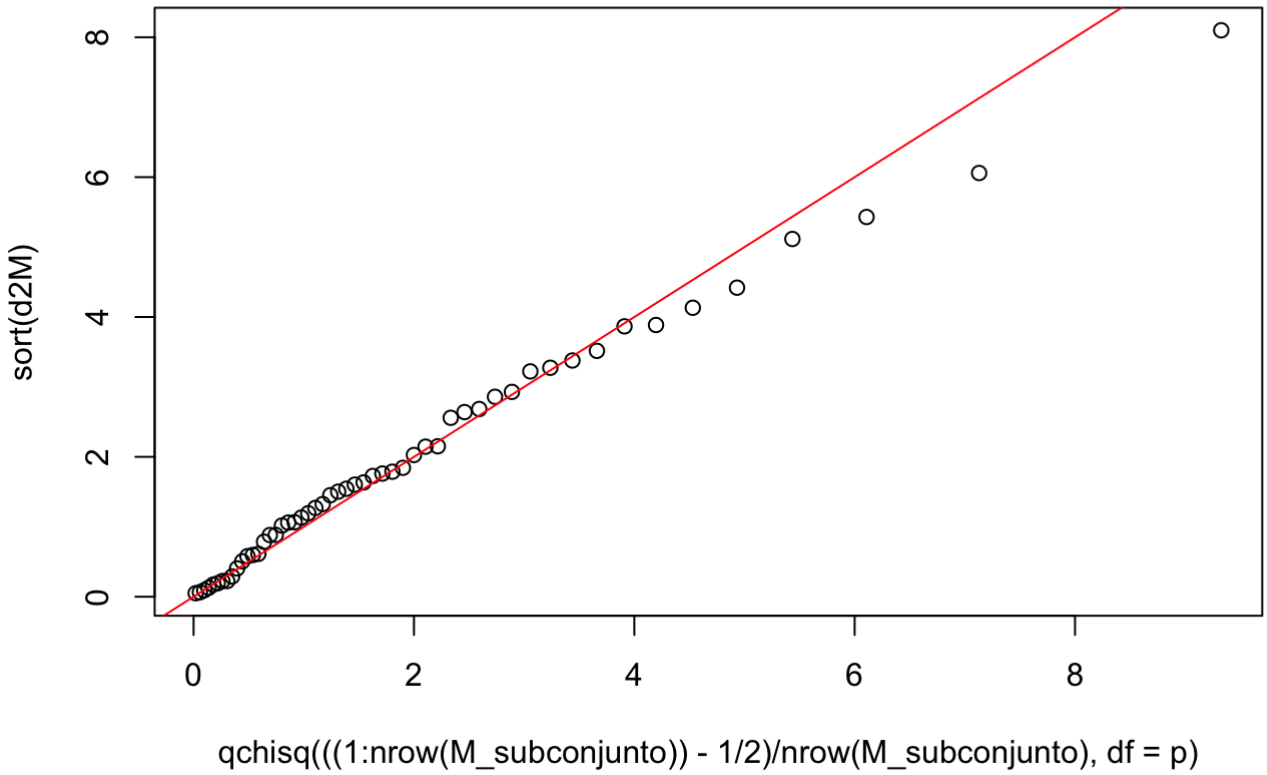
Primeramente se calcula la distancia de Mahalanobis de cada registro.

```
## Se calcula la distancia estadística de Mahalanobis de cada uno de los registros
d2M =  mahalanobis(M_subconjunto,X,S)
d2M
```

```
##  [1] 1.19340732 3.86623312 4.13223576 0.06401297 2.55854258 0.59679099
##  [7] 1.78765285 1.50256004 1.01802464 1.63445643 1.26957081 0.22408796
## [13] 0.61327299 5.43003539 2.14558317 2.15128332 2.68426029 5.11541618
## [19] 0.78631278 1.84332204 3.88561874 0.04988963 0.88374242 8.09988683
## [25] 2.02671417 0.58001802 0.50566324 2.64029211 2.92897033 1.06510339
## [31] 0.88216107 0.40467236 6.05908470 3.51592769 1.45168260 1.32485980
## [37] 1.13164718 2.85893264 0.22445752 3.37787450 1.54397421 1.72672287
## [43] 0.28960341 1.76103444 1.60277372 0.09262808 0.19257450 3.22197239
## [49] 3.27393627 0.17321589 0.12786732 4.41942776 1.06000856
```

Con dichas distancias, es posible obtener el QQ plot multivariado

```
p = 2 # indica el número de variables

## Multinormalidad Test gráfico Q-Q Plot
plot(qchisq(((1:nrow(M_subconjunto)) - 1/2)/nrow(M_subconjunto),df=p),sort( d2M ) )
abline(a=0, b=1,col="red",
       main = "QQ plot multivariado")
```



El diagrama QQ que se observa indica que realmente no se tienen datos atípicos, pues todos los puntos se encuentran realmente cercanos a la recta de color rojo. Los únicos dos posibles datos que podrían tratarse de datos atípicos, son los dos últimos, pues estos se desvían bastante de la recta y son lo que tienen una mayor distancia de Mahalanobis.

Cabe resaltar que los puntos tienen un comportamiento tal que se podría inferir que se cuenta con un leve sesgo hacia la izquierda, es decir, se tiene asimetría negativa.

Análisis de componentes principales

Uso adecuado de componentes principales

Ahora, se pasará a realizar un análisis de componentes principales con los datos. Esto es con el objetivo de ver si es posible encontrar nuevas variables (componentes principales) las cuales expliquen una mayor parte de la varianza sin perder tanta información, con tal de reducir la dimensión de la base de datos.

Para este análisis, retomemos un poco la matriz de correlaciones.

```
## Se guarda en la variable "S" la matriz de varianzas y covarianzas
P = cor(M)
P
```

```
##              X3          X4           X5          X6          X7          X8
## X3   1.00000000  0.71916568  0.832604192  0.47753085 -0.59389671  0.01029074
## X4   0.71916568  1.00000000  0.577132721  0.60848276 -0.57540012 -0.01860607
## X5   0.83260419  0.57713272  1.000000000  0.40991385 -0.40067958 -0.08937901
## X6   0.47753085  0.60848276  0.409913846  1.00000000 -0.49137481 -0.01182027
## X7  -0.59389671 -0.57540012 -0.400679584 -0.49137481  1.00000000  0.07903426
## X8   0.01029074 -0.01860607 -0.089379013 -0.01182027  0.07903426  1.00000000
## X9  -0.52535654 -0.54196524 -0.332476229 -0.40045856  0.92720506 -0.08165278
## X10 -0.60479558 -0.55181523 -0.407916635 -0.48497215  0.91586397  0.16109174
## X11 -0.62795845 -0.61284905 -0.464409465 -0.50644193  0.95921481  0.02580046
## X12 -0.09493882  0.03800021 -0.002111124 -0.28300234  0.10873896  0.20795617
##              X9         X10         X11          X12
## X3  -0.52535654 -0.60479558 -0.62795845 -0.094938825
## X4  -0.54196524 -0.55181523 -0.61284905  0.038000214
## X5  -0.33247623 -0.40791663 -0.46440947 -0.002111124
## X6  -0.40045856 -0.48497215 -0.50644193 -0.283002338
## X7   0.92720506  0.91586397  0.95921481  0.108738958
## X8  -0.08165278  0.16109174  0.02580046  0.207956171
## X9   1.00000000  0.76535319  0.91908939  0.100661967
## X10  0.76535319  1.00000000  0.85975810  0.093752072
## X11  0.91908939  0.85975810  1.00000000  0.089411267
## X12  0.10066197  0.09375207  0.08941127  1.000000000
```

Como ya se explicó con anterioridad, la matriz de correlaciones indica qué tan linealmente están relacionadas las variables entre si. Antes de realizar el análisis de componenetes principales es importante verificar la correlación, ya que uno de los supuestos es que las variables estén correlacionadas para formar combinaciones lineales. Si las variables tienen una correlación distinta de 0, quiere decir que se relacionan de una forma, y por lo tanto es válido aplicar componentes principales.

Obteniendo los componentes principales

Ahora se pasará a obtener los componentes principales de los datos a través de la matriz de correlaciones. Para esto, se obtienen los valores y vectores propios de la matriz respectiva.

```
## Se obtienen los vectores y valores propios de la matriz S
S = cov(M)
P = cor(M)
eigen(P)
```

```
## eigen() decomposition
## $values
##  [1] 5.36122641 1.25426109 1.21668138 0.90943267 0.59141736 0.30314741
##  [7] 0.20673634 0.08682133 0.05163902 0.01863699
##
## $vectors
##              [,1]        [,2]       [,3]         [,4]        [,5]        [,6]
##  [1,] -0.35065869 -0.21691594 -0.3472906  0.009131194  0.34050534  0.07547497
##  [2,] -0.33700381 -0.21940887 -0.2360975 -0.017242162 -0.39396038  0.73121012
##  [3,] -0.28168286 -0.26250672 -0.5113780  0.146950070  0.36205937 -0.31342329
##  [4,] -0.28334182  0.10195058 -0.2639612 -0.432676049 -0.63093376 -0.44112169
##  [5,]  0.39830786 -0.12104244 -0.2996635 -0.080630070 -0.03046869  0.07436922
##  [6,]  0.02667579 -0.57556151  0.3050633 -0.692854505  0.19646415 -0.05926732
##  [7,]  0.36839224 -0.04432459 -0.3876861  0.044658983 -0.13236038 -0.19602465
##  [8,]  0.37893835 -0.14237181 -0.2024901 -0.167921215  0.02678086  0.26671839
##  [9,]  0.40206100 -0.05279514 -0.2562319 -0.042242268 -0.05607416  0.03863899
## [10,]  0.05931430 -0.67421026  0.2294446  0.521815581 -0.37253140 -0.21612970
##              [,7]        [,8]        [,9]       [,10]
##  [1,] -0.33823501  0.68622998  0.04284021 -0.02239801
##  [2,] -0.08629646 -0.28769221  0.01363551  0.04445261
##  [3,]  0.34312185 -0.45568753 -0.11508339  0.02634676
##  [4,]  0.13435159  0.19006976 -0.06333133 -0.03982419
##  [5,] -0.01377825 -0.01674789  0.06243320 -0.84827636
##  [6,] -0.14693148 -0.16809481  0.02532023  0.04805976
##  [7,] -0.45674057 -0.18260535  0.53803577  0.35020485
##  [8,]  0.67376588  0.33602914  0.18844932  0.30445219
##  [9,] -0.23387764  0.02613406 -0.80648296  0.24018040
## [10,]  0.05759514  0.16451240 -0.02782678 -0.01839703
```

Los vectores propios anteriores se componen de los coeficientes de cada una de las variables para formar una nueva variable. En cada componente, se nota que se tiene un mayor peso en algunas variables, dando así a entender que esas juegan un papel más importante para dicho componente. Por ejemplo, en el primer componente que proviene del primer vector propio, se tiene que las variables $X11, X7$ y $X9$ tiene mayor relevancia. Cada componente le da una importancia distinta a cada variable y lo que busca es relacionarlas con el fin de darles nostros mismo un significado al componente.

Una vez realizado lo anterior, es posible calcular el procentaje de varianza explicada por cada una de las componentes. A continuación se realiza el análisis

```
## Se calcula la proporción de varianza explicada por cada componente

## Valores lambda de la matriz de covarianzas:
valores_lambda = eigen(P)$values
valores_lambda
```

```
##  [1] 5.36122641 1.25426109 1.21668138 0.90943267 0.59141736 0.30314741
##  [7] 0.20673634 0.08682133 0.05163902 0.01863699
```

```
## Varianza total
varianza_total = sum(diag(P))
varianza_total
```

```
## [1] 10
```

```
## La varianza total también es la suma de los valores propios de la matriz de covarianz
as. Esto se debe a que las combinaciones lineales buscan reproducir la varianza de X.
sum(valores_lambda)
```

```
## [1] 10
```

```
## Se calcula la proporción de varianzas explicada
cat('\nLa variazas explicada de cada componente es:\n')
```

```
##
## La variazas explicada de cada componente es:
```

```
## La variazas explicada de cada componente es:
proporciones = valores_lambda/varianza_total
proporciones
```

```
##  [1] 0.536122641 0.125426109 0.121668138 0.090943267 0.059141736 0.030314741
##  [7] 0.020673634 0.008682133 0.005163902 0.001863699
```

La varianza explicada indica que tanta información de los datos se está explicando en cada componentes. Por ejemplo, el primer componente principal está explicando alrededor del 53% de la varianza total, Lo que se busca es expresar con el menor número de componentes, el mayor porcentaje de varianza, para así poder realizar la reducción de dimensionalidad sin pérdida de información.

Para poder decidir con qué y cuántos componentes principales nos debemos de quedar, nos apoyamos en algunos gráficos.

Gráficos

Primero realizamos dos gráficos que comparan el comportamiento de los dos primeros componentes principales. En el primer gráfico, se pueden observar los regitros ya transformados, es decir después de haberles aplicado su combinación lineal correspondiente. Como se puede observar, ambos componentes abarcan gran parte de los datos, indicando así que explican gran parte de la varianza, lo cual es bueno cuando se quiere reducir dimensiones. En los ejes de igual forma se ve el porcentaje de varianza explicado de cada componente.

En el segundo diagrama se muestran los vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes. Como se observa, la primer componente jala más a $X3$, $X4$, $X7$ y $X11$, mientras que el segundo componente jala más a $X12$. Sin embargo, de igual forma de ve que la mayoría de vectores tienen una longitud bastante similar, indicando así que estos dos primeros componentes principales están describiendo gran parte d ela variación.

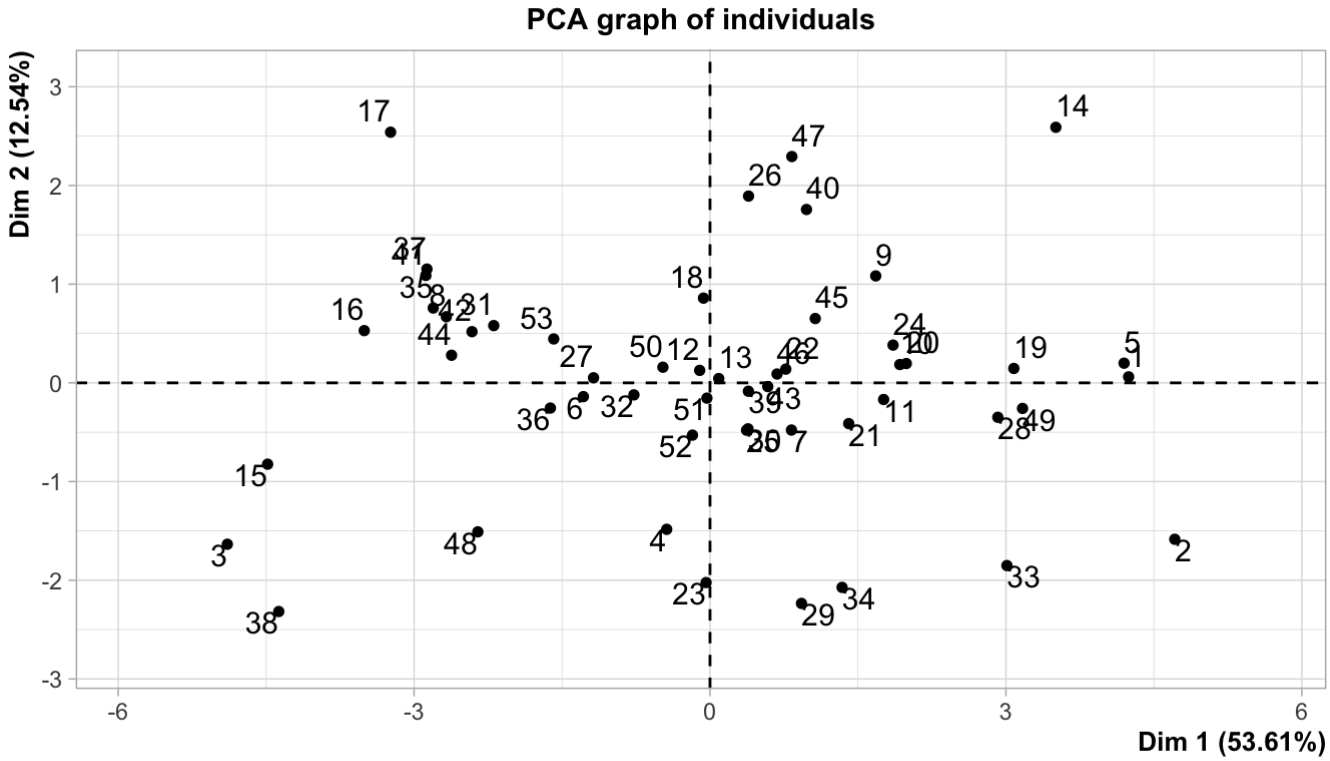
```
library(FactoMineR)
```

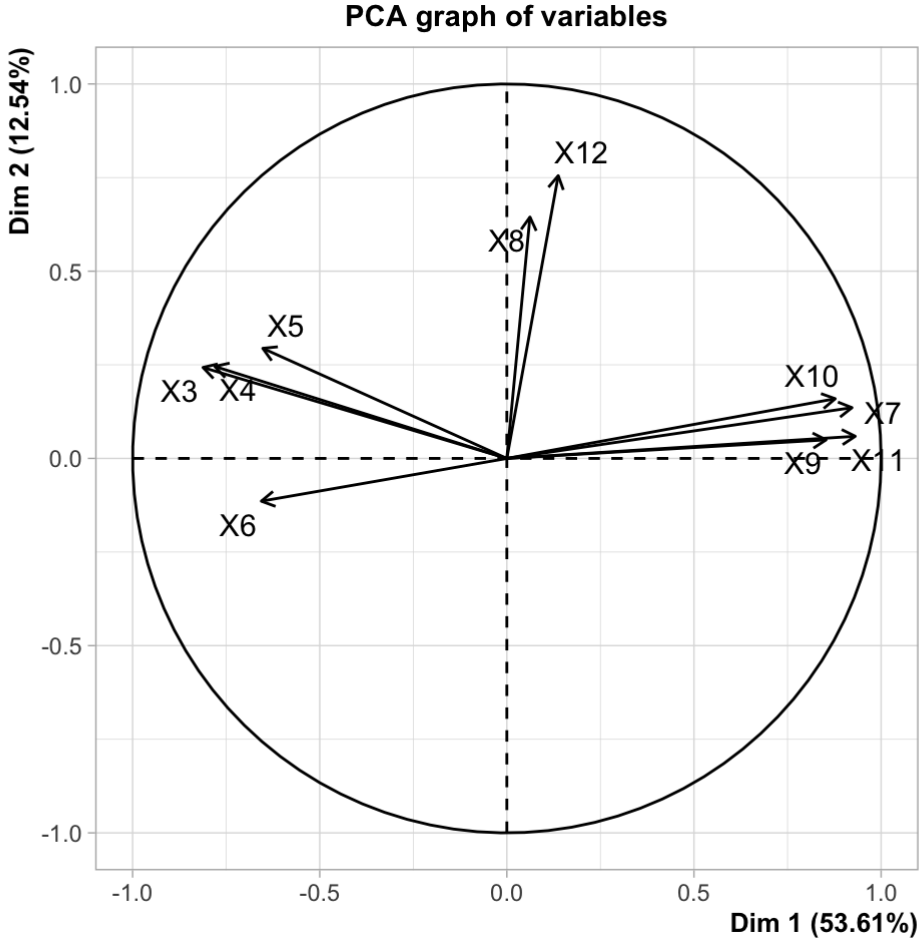
```
## Warning: package 'FactoMineR' was built under R version 4.1.2
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WB
a
```

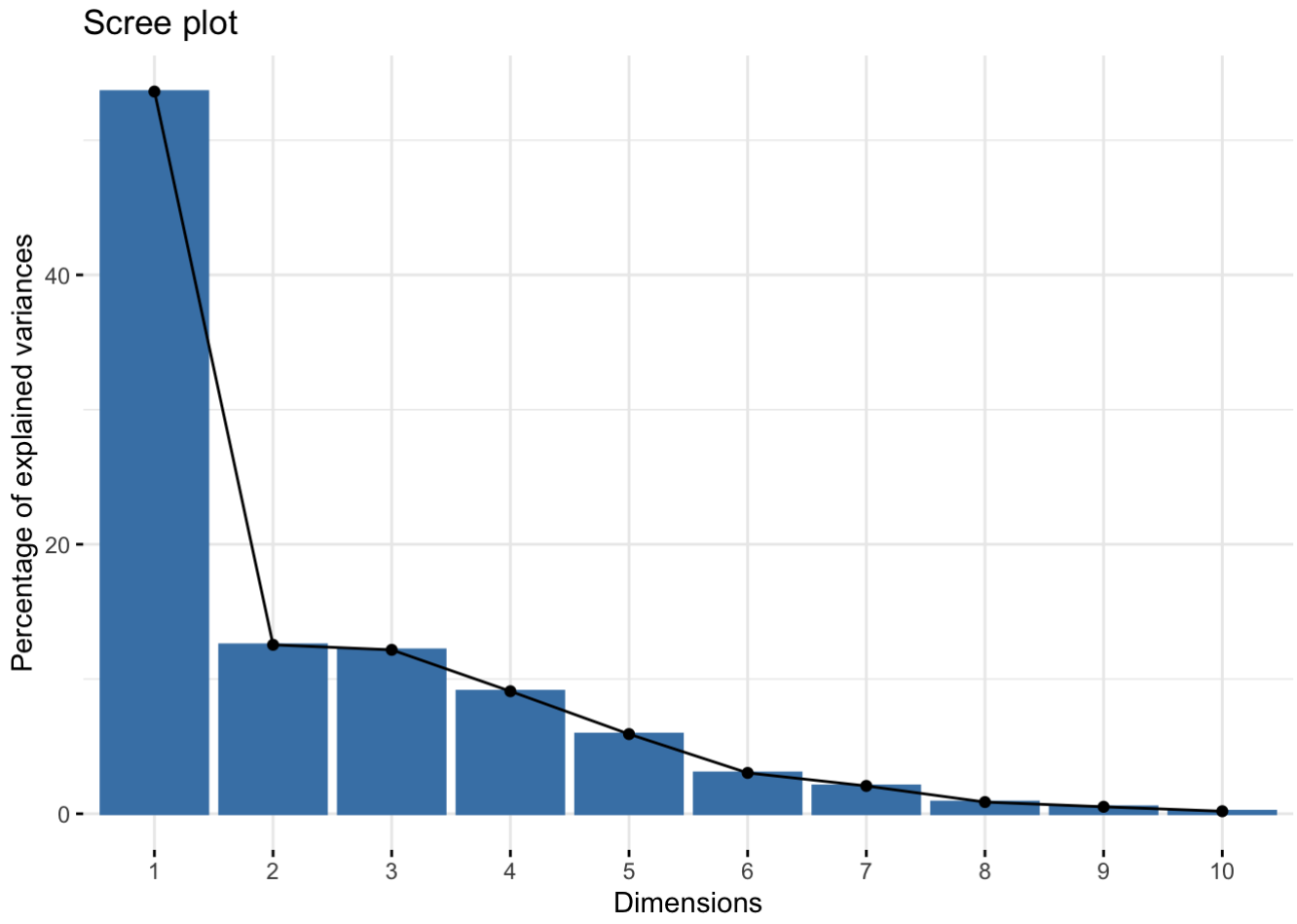
```
library(ggplot2)

datos= M
cp3 = PCA(datos)
```


PCA graph of individuals
Dim 2 (12.54%)
Dim 1 (53.61%)
3
2
1
0
-1
-2
-3
-6
-3
0
3
6
17
14
47
26
40
37
41
9
18
35
8
45
16
42
21
53
24
44
27
50
12
13
19
5
1
36
6
32
51
39
43
11
28
49
52
21
15
3
48
4
2
33
23
34
29
38


PCA graph of variables
Dim 2 (12.54%)
Dim 1 (53.61%)
1.0
0.5
0.0
-0.5
-1.0
X12
X8
X5
X3
X4
X10
X7
X9
X11
X6

Otro gráfico que es de mucho interés, es el de codo con base en la varianza explicada de cada variable. Aqui se muestra qué el primer componente representa el 53% de la información, mientras que el segundo y tercer componente tiene más o menos el mismo porcentaje. Si nos basamos en el codo de la gráfica, lo mejor sería tomar los tres pimeros componentes principales, ya que haciendolo de esa forma, se reduciería la dimensión original de 10 a 3 y se explicaría el 77% de la información, lo cual no está nada mal. Los demás componentes no tienen tanta relevancia.

```
fviz_screeplot(cp3)
```

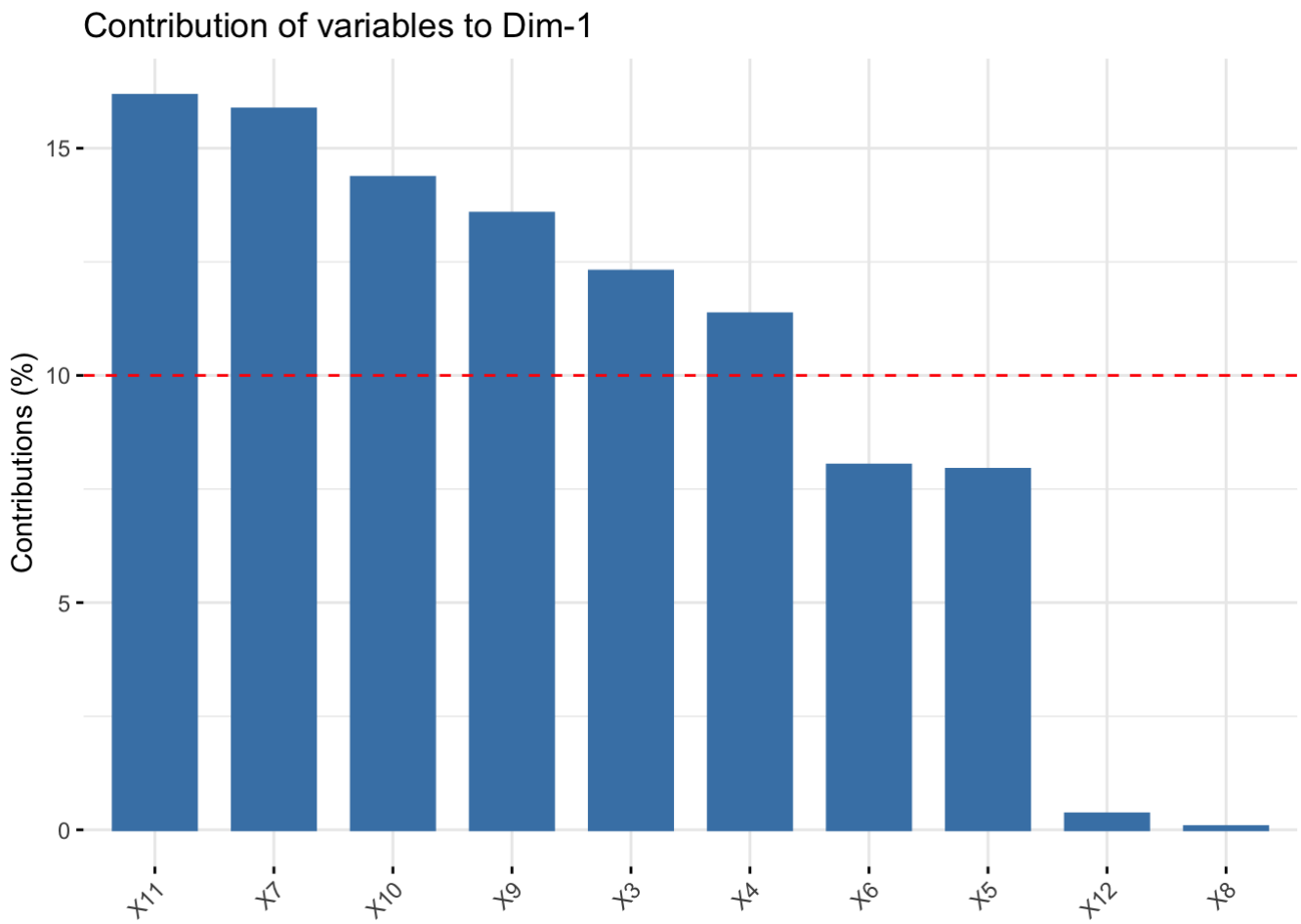


Finalmente, como se estará trabajando con los tres primeros componentes, se le debe de dar un significado a cada uno. Por ejemplo, el componente 1 relaciona varias variables, como se muestra en el siguiente gráfico, siendo las más representativas $X11$, $X7$, $X10$, $X9$, $X3$ y $X4$. Viendo las cuatro primeras variables, vemos que están relaconadas con la concentración de mercurio en los lagos, mientras mientras que las otras dos con el ph y la alcalinidad. Esto nos podría decir que tanto la alcalinidad del lago como su PH son relevantes para la concentración de mercurio, por lo que al componente principal se le podría dar el nombre de “Grado de acidez” o similar.

El segundo componente lo describe en su mayoría $X12$ y $X8$, en donde se puede vr que se relaciona la edad de los peces con el número de éstos. Por lo tanto el componente principal se podría interpretar como “características de la pobalción”, o similar.

Un análisis similar se tendría que hacer para cada uno de los componentes, es decir tratarles de dar un significado dentro del contexto.

```
fviz_contrib(cp3, choice = c("var"))
```



Conclusiones

PENDIENTES.