Procesamiento de Datos Multivariados

Code

Gilberto Ramos Salinas A01734128

Introduccion

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra las muestras de 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.

Code

id	lake	alkalinity	…	calcium	chlorophyll	mercury_con	fish_number	min_r
<int>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	
1	Alligator	5.9	6.1	3.0	0.7	1.23	5	
2	Annie	3.5	5.1	1.9	3.2	1.33	7	
3	Apopka	116.0	9.1	44.1	128.3	0.04	6	
4	Blue Cypress	39.4	6.9	16.4	3.5	0.44	12	
5	Brick	2.5	4.6	2.9	1.8	1.20	12	
6	Bryant	19.6	7.3	4.5	44.1	0.27	14	
7	Cherry	5.2	5.4	2.8	3.4	0.48	10	
8	Crescent	71.4	8.1	55.2	33.7	0.19	12	
9	Deer Point	26.4	5.8	9.2	1.6	0.83	24	
10	Dias	4.8	6.4	4.6	22.5	0.81	12	

1-10 of 53 rows | 1-9 of 12 columns Previous **1** 2 3 4 5 6 Next

Code

1. Realice un análisis denormalidad de las variables continuas para identificar variables normales. Tome en cuenta los puntos que se sugieren a continuación (no son exhaustivos):

A. Realice la prueba de normalidad de Mardia y la prueba de Anderson Darling para identificar las variables que son normales y detectar posible normalidad multivariada de grupos de variables.

Code

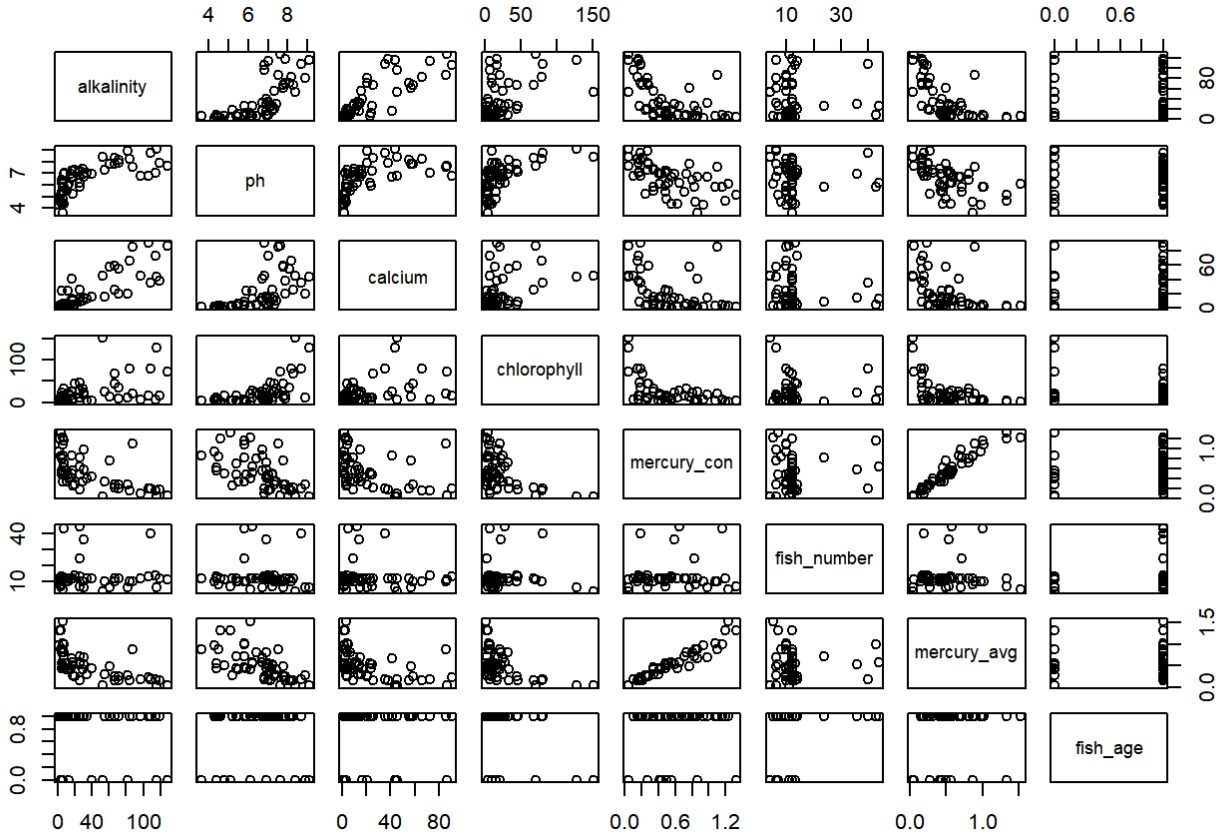
```
$multivariateNormality
           Test       HZ p value MVN
1 Henze-Zirkler 2.188291       0  NO

$univariateNormality
              Test    Variable Statistic   p value Normality
1 Anderson-Darling alkalinity     3.6725  <0.001      NO
2 Anderson-Darling     ph         0.3496  0.4611      YES
3 Anderson-Darling   calcium      4.0510  <0.001      NO
4 Anderson-Darling chlorophyll    5.4286  <0.001      NO
5 Anderson-Darling mercury_con    0.9253  0.0174      NO
6 Anderson-Darling fish_number    8.6943  <0.001      NO

$Descriptives
             n       Mean    Std.Dev Median  Min    Max  25th  75th       Skew
alkalinity  53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50  0.9679170
ph          53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40 -0.2458771
calcium     53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60  1.3045868
chlorophyll 53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70  2.4130571
mercury_con 53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77  0.5986343
fish_number 53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00  2.5808773
              Kurtosis
alkalinity  -0.4705349
ph          -0.6239638
calcium      0.6130359
chlorophyll  6.1042185
mercury_con -0.6312607
fish_number  6.0089455
```

Code

Test <chr>	Statistic <fct>	p value <fct>	Result <chr>
Mardia Skewness	219.53259918407	3.20436177199441e-21	NO
Mardia Kurtosis	4.4345204376395	9.22774660794268e-06	NO
MVN	*NA*	*NA*	NO

3 rows

Code


alkalinity
ph
calcium
chlorophyll
mercury_con
fish_number
mercury_avg
fish_age

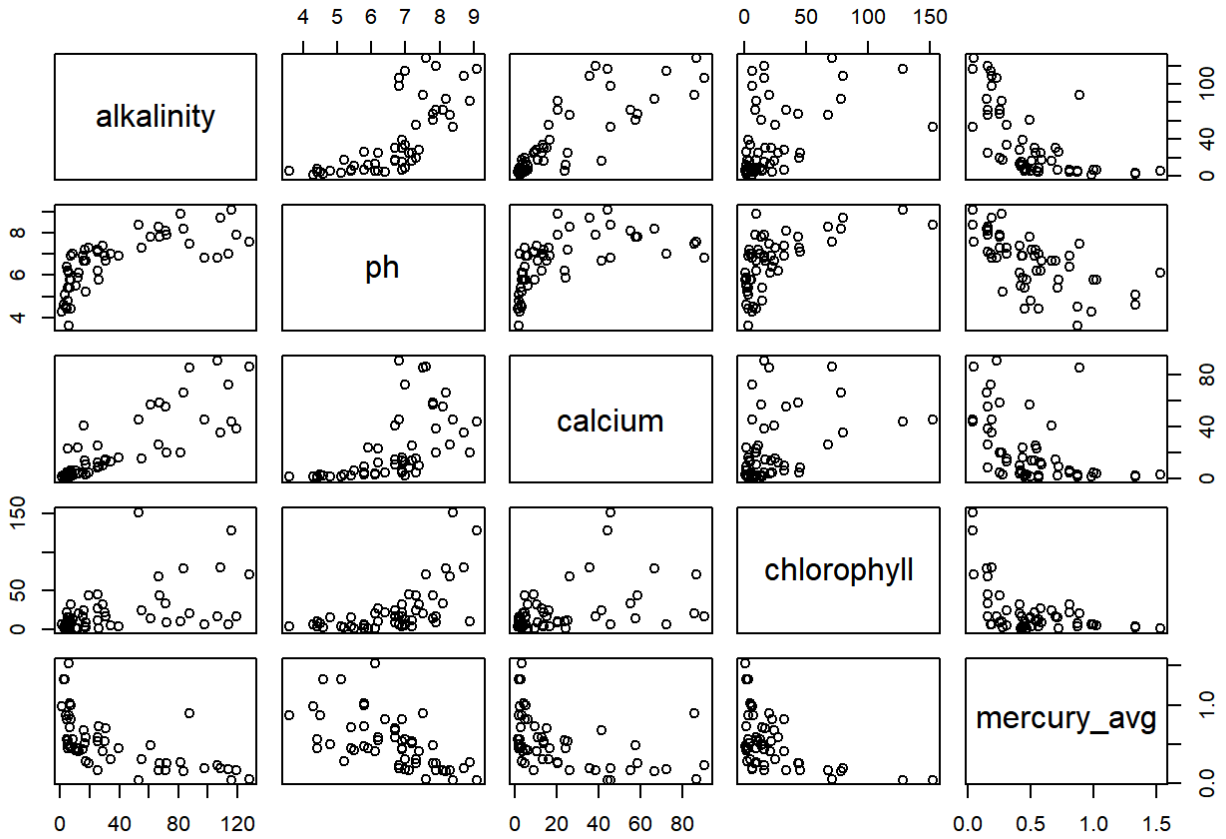
Code

```
$multivariateNormality
             Test        Statistic              p value Result
1 Mardia Skewness 339.026287053957 8.83292442948867e-23     NO
2 Mardia Kurtosis 4.35093055335355 1.35560986049832e-05     NO
3             MVN             <NA>                 <NA>     NO

$univariateNormality
              Test    Variable Statistic   p value Normality
1 Anderson-Darling alkalinity     3.6725  <0.001      NO
2 Anderson-Darling     ph         0.3496  0.4611      YES
3 Anderson-Darling   calcium      4.0510  <0.001      NO
4 Anderson-Darling chlorophyll    5.4286  <0.001      NO
5 Anderson-Darling mercury_con    0.9253  0.0174      NO
6 Anderson-Darling fish_number    8.6943  <0.001      NO
7 Anderson-Darling mercury_avg    1.0469  0.0086      NO
8 Anderson-Darling  fish_age     14.3350  <0.001      NO

$Descriptives
             n       Mean    Std.Dev Median  Min    Max  25th  75th       Skew
alkalinity  53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50  0.9679170
ph          53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40 -0.2458771
calcium     53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60  1.3045868
chlorophyll 53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70  2.4130571
mercury_con 53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77  0.5986343
fish_number 53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00  2.5808773
mercury_avg 53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70  0.9449951
fish_age    53  0.8113208  0.3949977   1.00 0.00   1.00  1.00  1.00 -1.5465748
              Kurtosis
alkalinity  -0.4705349
ph          -0.6239638
calcium      0.6130359
chlorophyll  6.1042185
mercury_con -0.6312607
fish_number  6.0089455
mercury_avg  0.5733500
fish_age     0.4005116
```

B. Realiza la prueba de Mardia y Anderson Darling de las variables que sí tuvieron normalidad en los incisos anteriores. Interpreta los resultados obtenidos con base en ambas pruebas y en la interpretación del sesgo y la curtosis de cada una de ellas.

Code

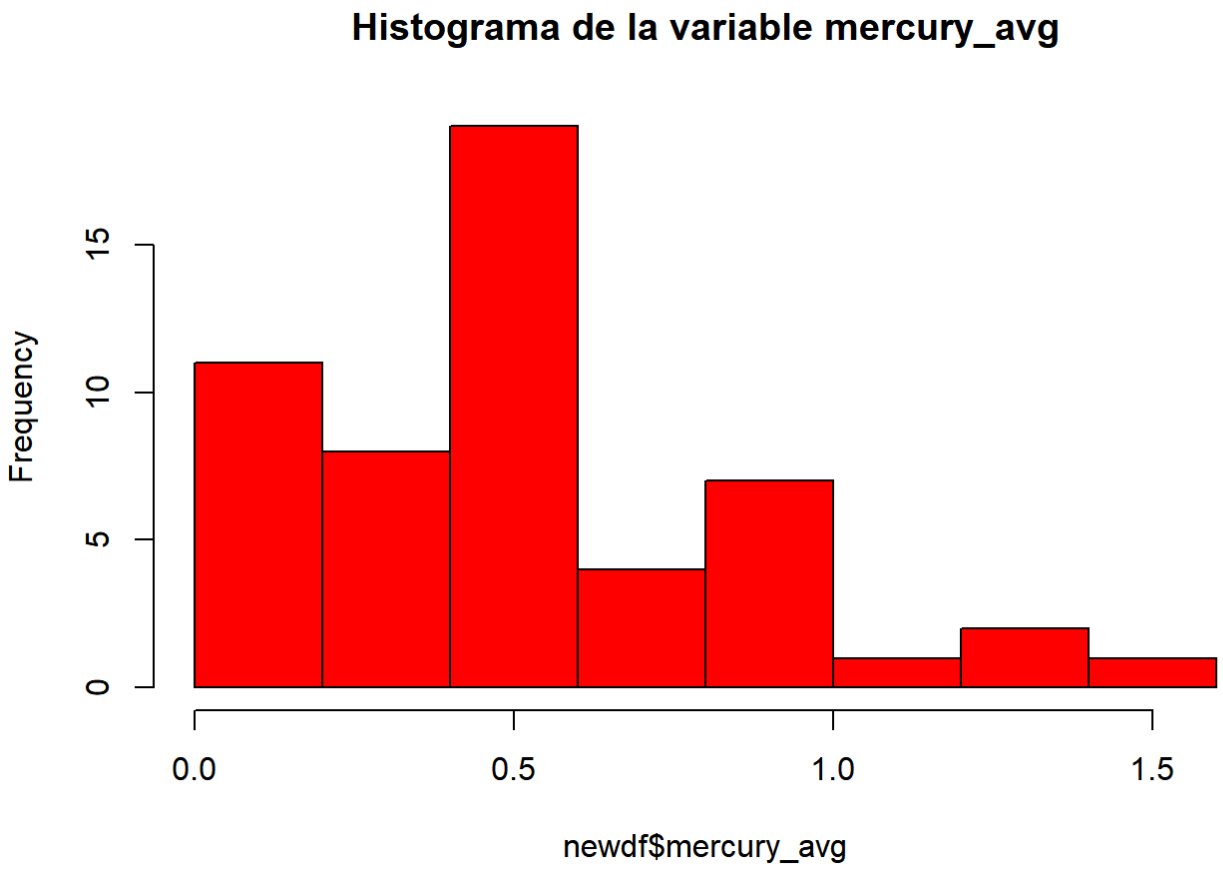
Code


alkalinity
ph
calcium
chlorophyll
mercury_avg
4 5 6 7 8 9
0 50 100 150
0 40 100
4 6 8
0 40 80
0 50 150
0.0 1.0
0 40 80 120
0 20 40 60 80
0.0 0.5 1.0 1.5

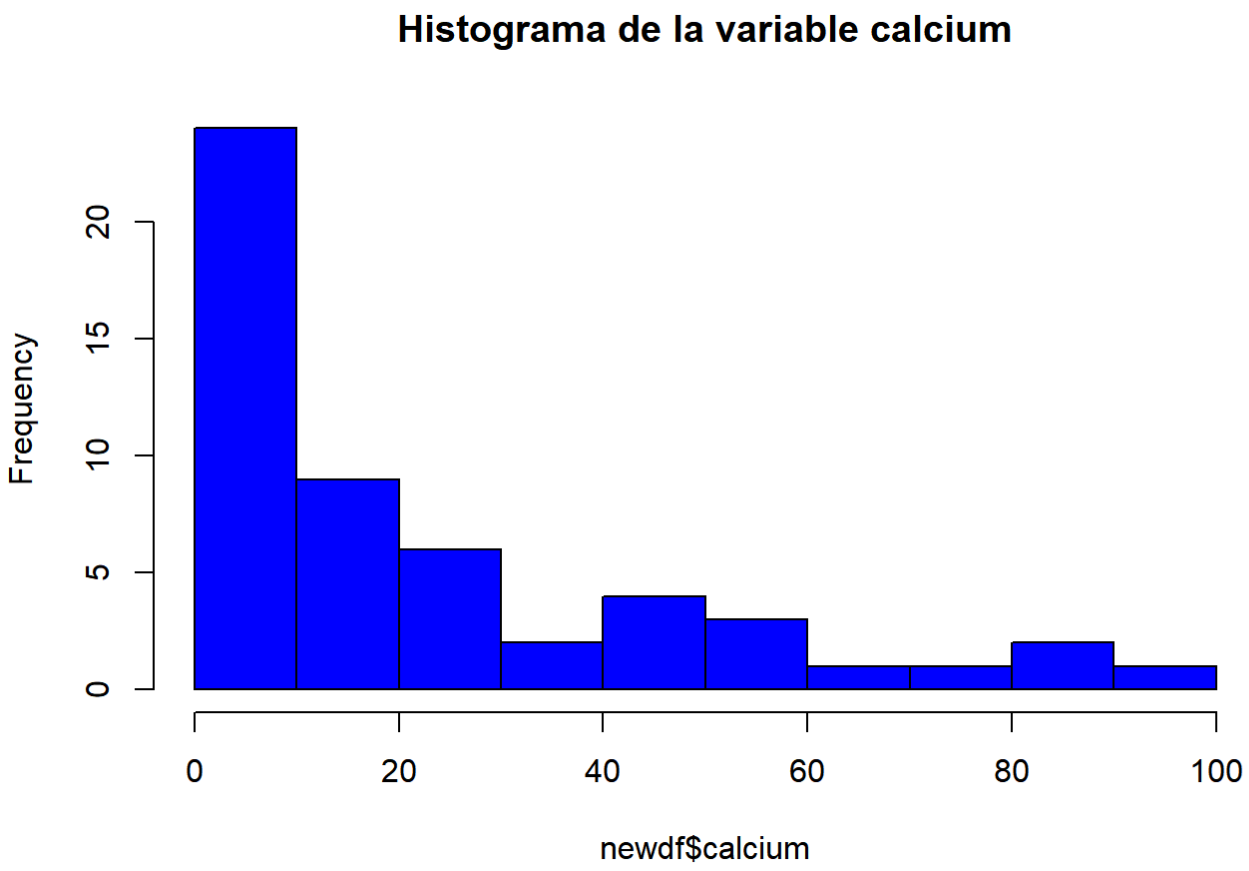
Code

```
$multivariateNormality
             Test        Statistic              p value Result
1 Mardia Skewness 139.639027903543 1.92942465275976e-14     NO
2 Mardia Kurtosis 3.98326989284336 6.79734854871494e-05     NO
3             MVN             <NA>                 <NA>     NO

$univariateNormality
              Test    Variable Statistic   p value Normality
1 Anderson-Darling alkalinity     3.6725  <0.001      NO
2 Anderson-Darling     ph         0.3496  0.4611      YES
3 Anderson-Darling   calcium      4.0510  <0.001      NO
4 Anderson-Darling chlorophyll    5.4286  <0.001      NO
5 Anderson-Darling mercury_avg    1.0469  0.0086      NO

$Descriptives
             n       Mean    Std.Dev Median  Min    Max 25th 75th       Skew
alkalinity  53 37.5301887 38.2035267  19.60 1.20 128.00 6.60 66.5  0.9679170
ph          53  6.5905660  1.2884493   6.80 3.60   9.10 5.80  7.4 -0.2458771
calcium     53 22.2018868 24.9325744  12.60 1.10  90.70 3.30 35.6  1.3045868
chlorophyll 53 23.1169811 30.8163214  12.80 0.70 152.40 4.60 24.7  2.4130571
mercury_avg 53  0.5132075  0.3387294   0.45 0.04   1.53 0.25  0.7  0.9449951
              Kurtosis
alkalinity  -0.4705349
ph          -0.6239638
calcium      0.6130359
chlorophyll  6.1042185
mercury_avg  0.5733500
```

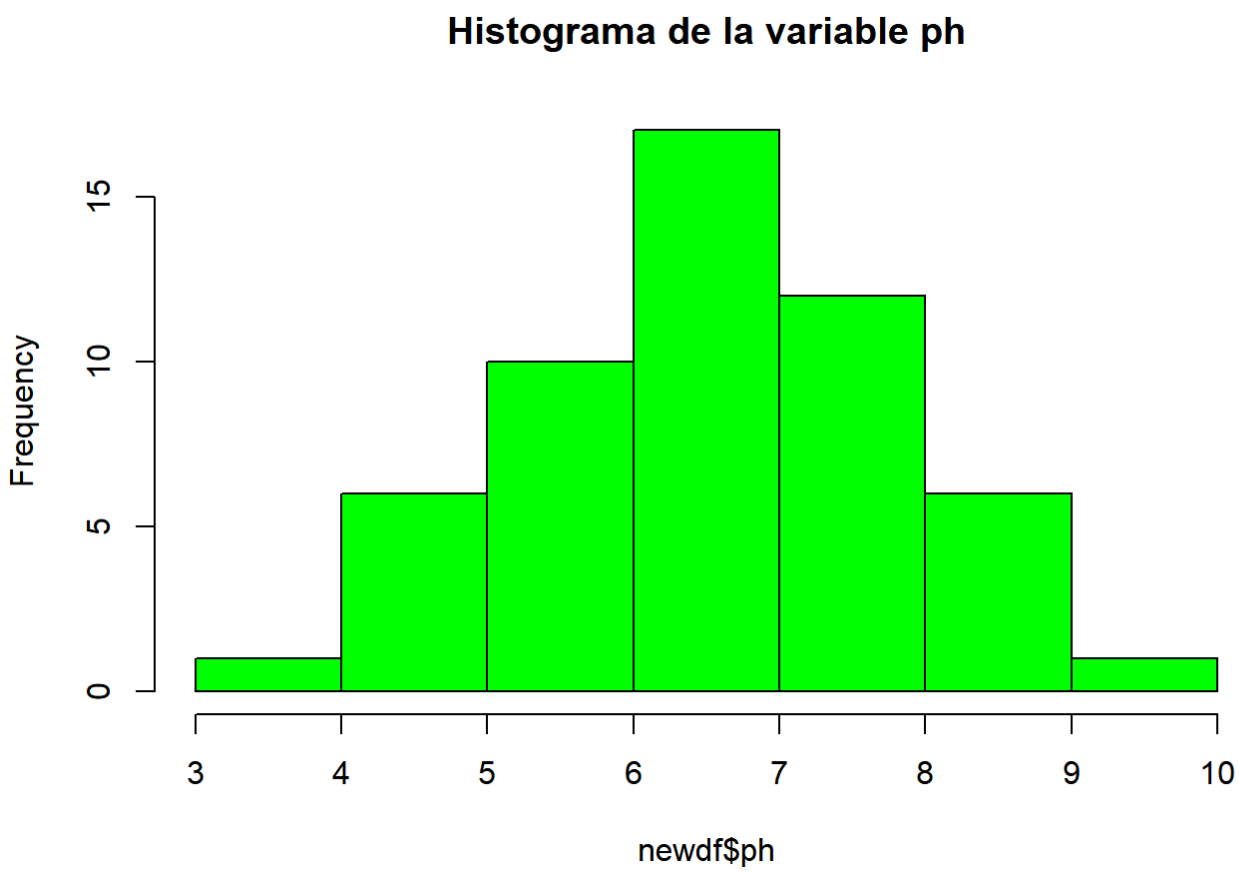
Code


Histograma de la variable mercury_avg
Frequency
0
5
10
15
0.0
0.5
1.0
1.5
newdf$mercury_avg

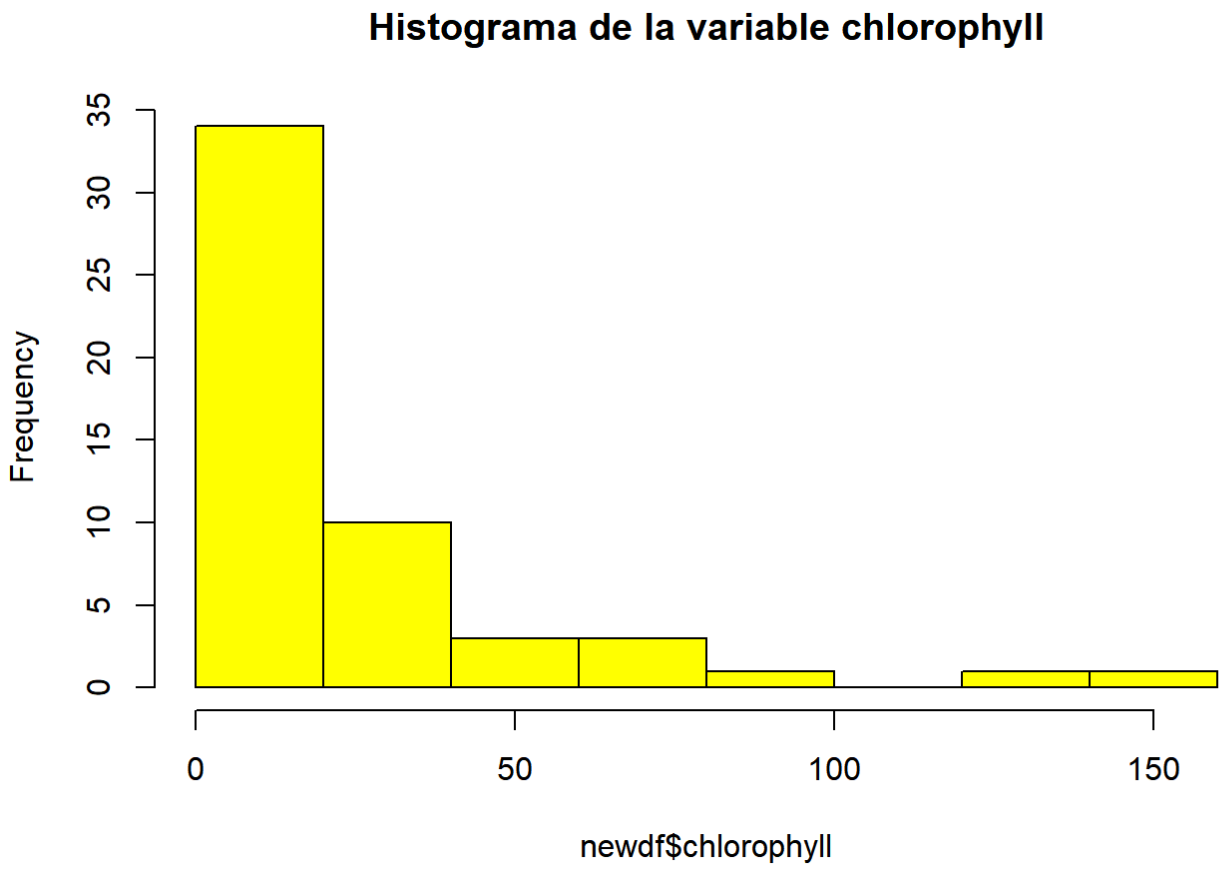
Code


Histograma de la variable calcium
Frequency
0
5
10
15
20
0
20
40
60
80
100
newdf$calcium

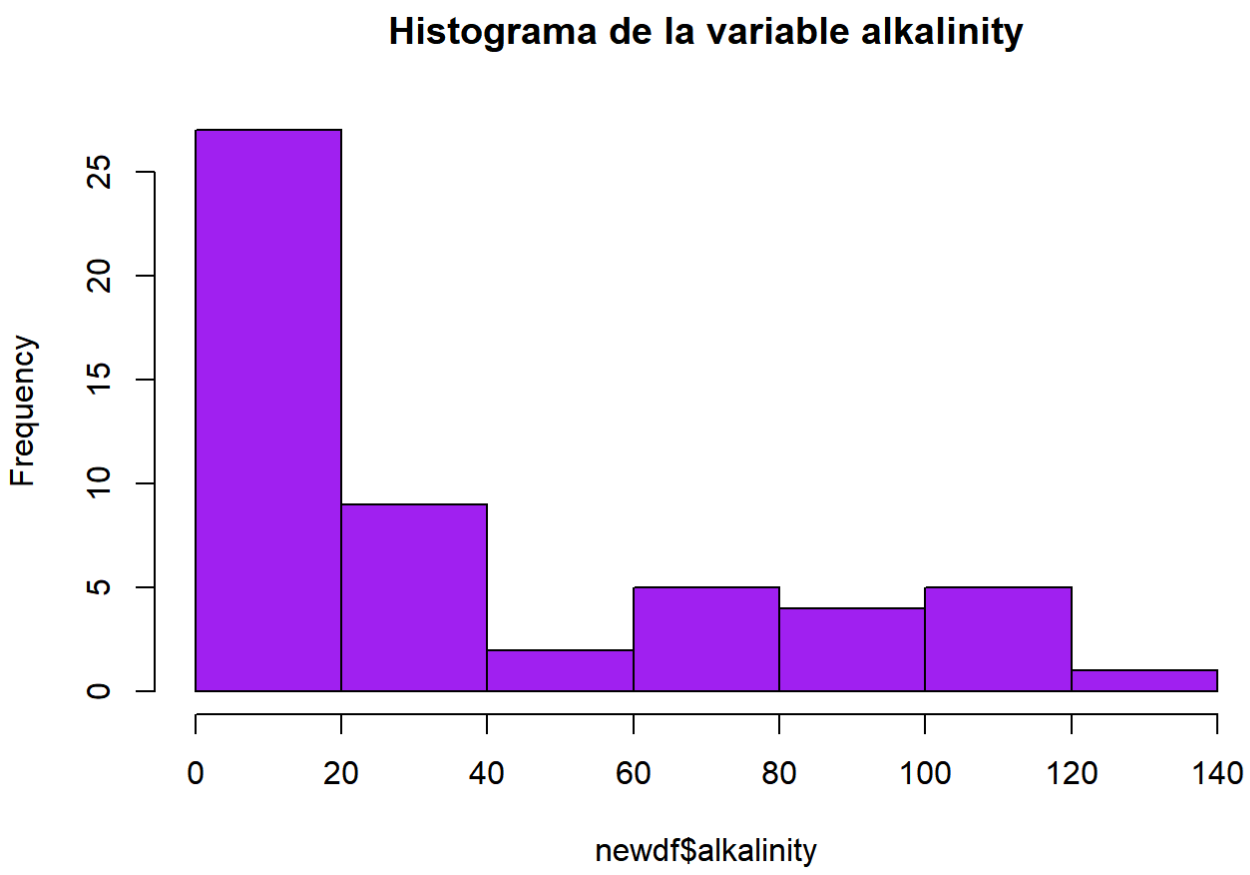
Code


Histograma de la variable ph
Frequency
0
5
10
15
3
4
5
6
7
8
9
10
newdf$ph

Code


Histograma de la variable chlorophyll
Frequency
0
5
10
15
20
25
30
35
0
50
100
150
newdf$chlorophyll

Code

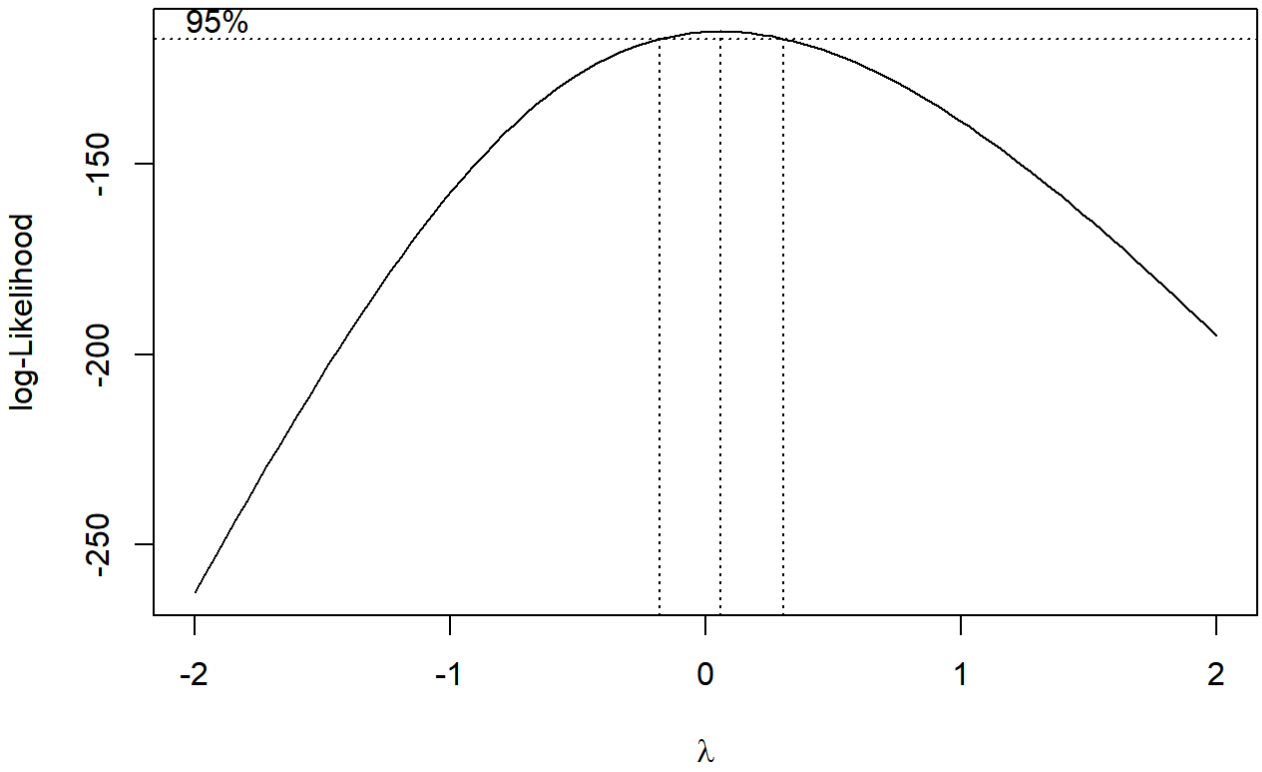


Code

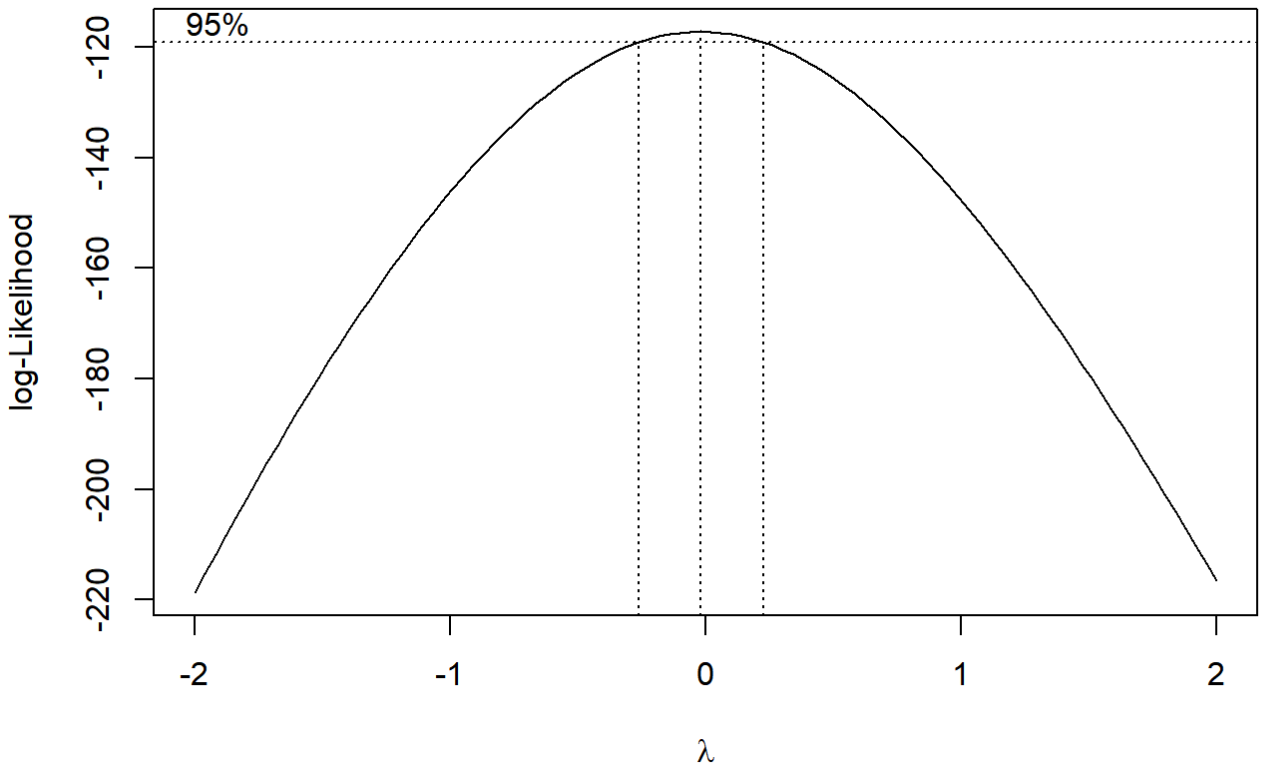
```
Attaching package: 'MASS'
```

```
The following object is masked from 'package:dplyr':

    select
```

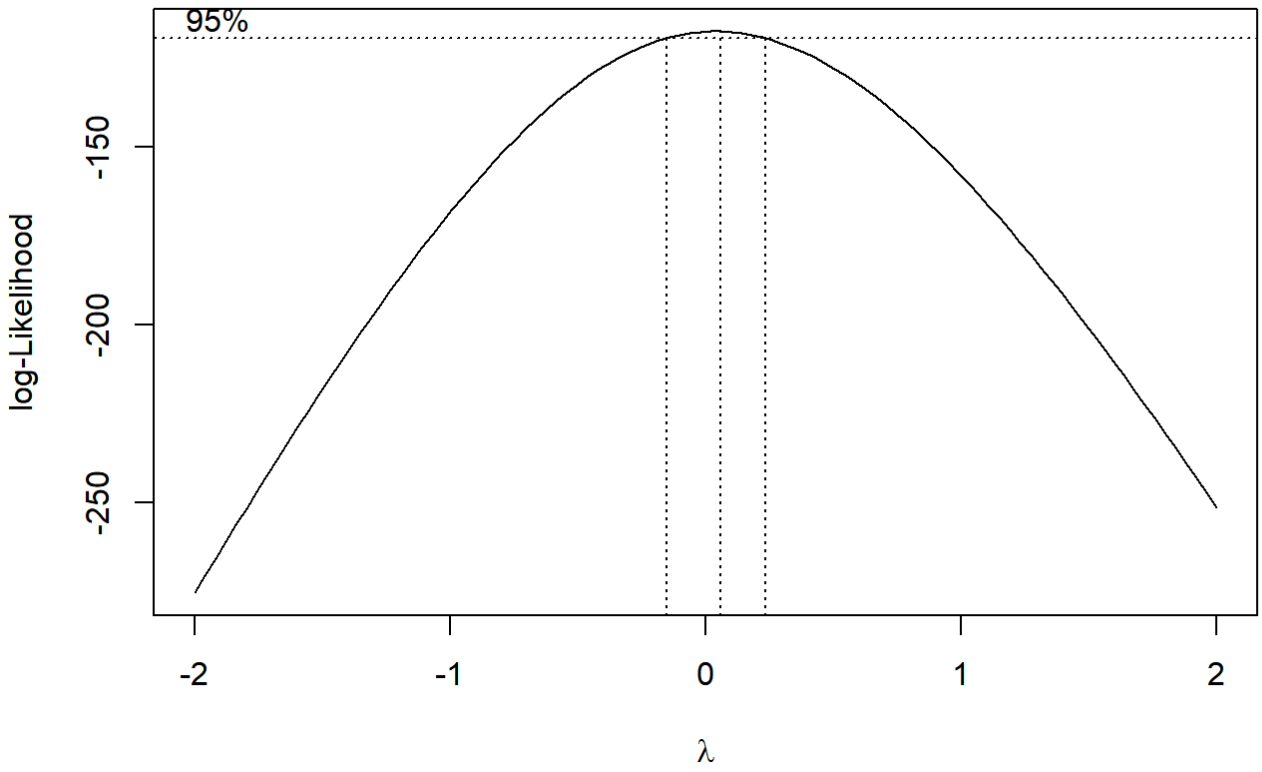
Code



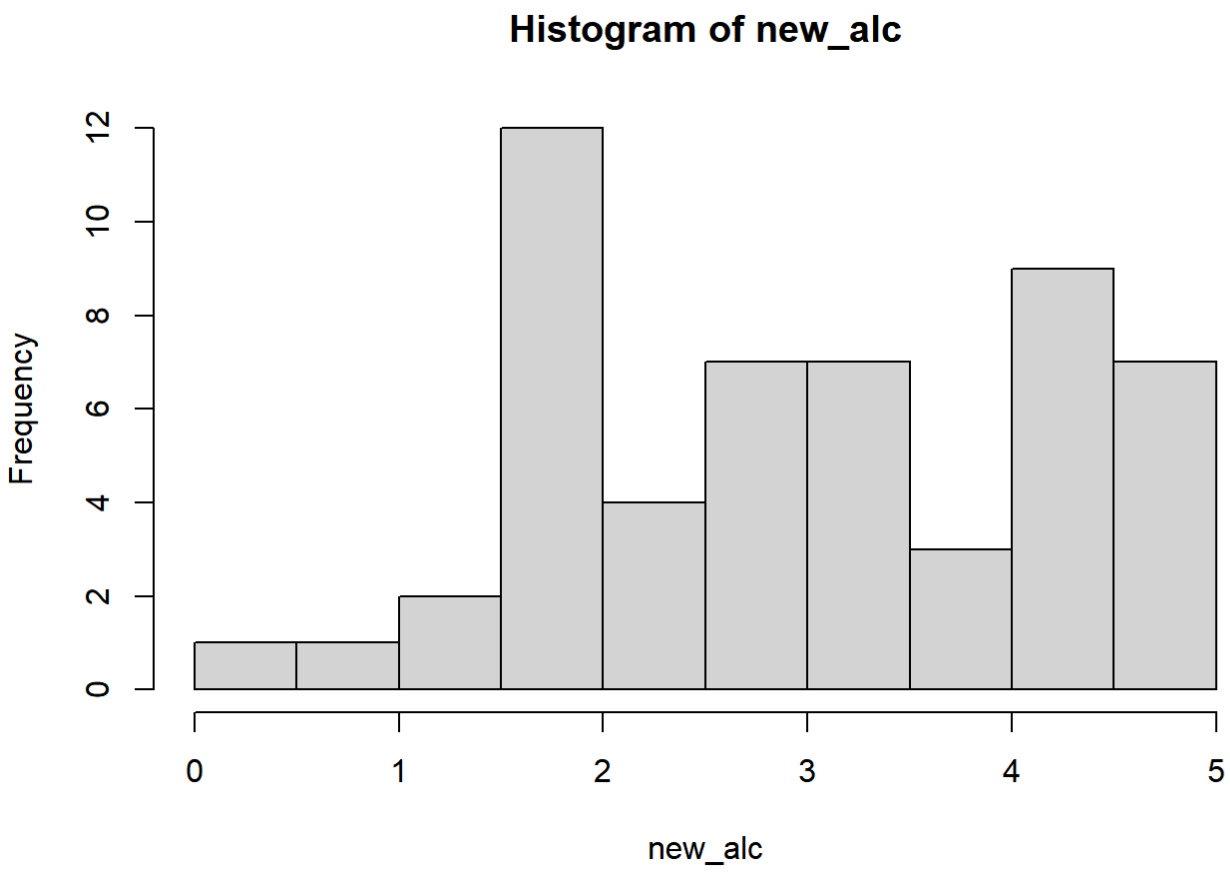
Code



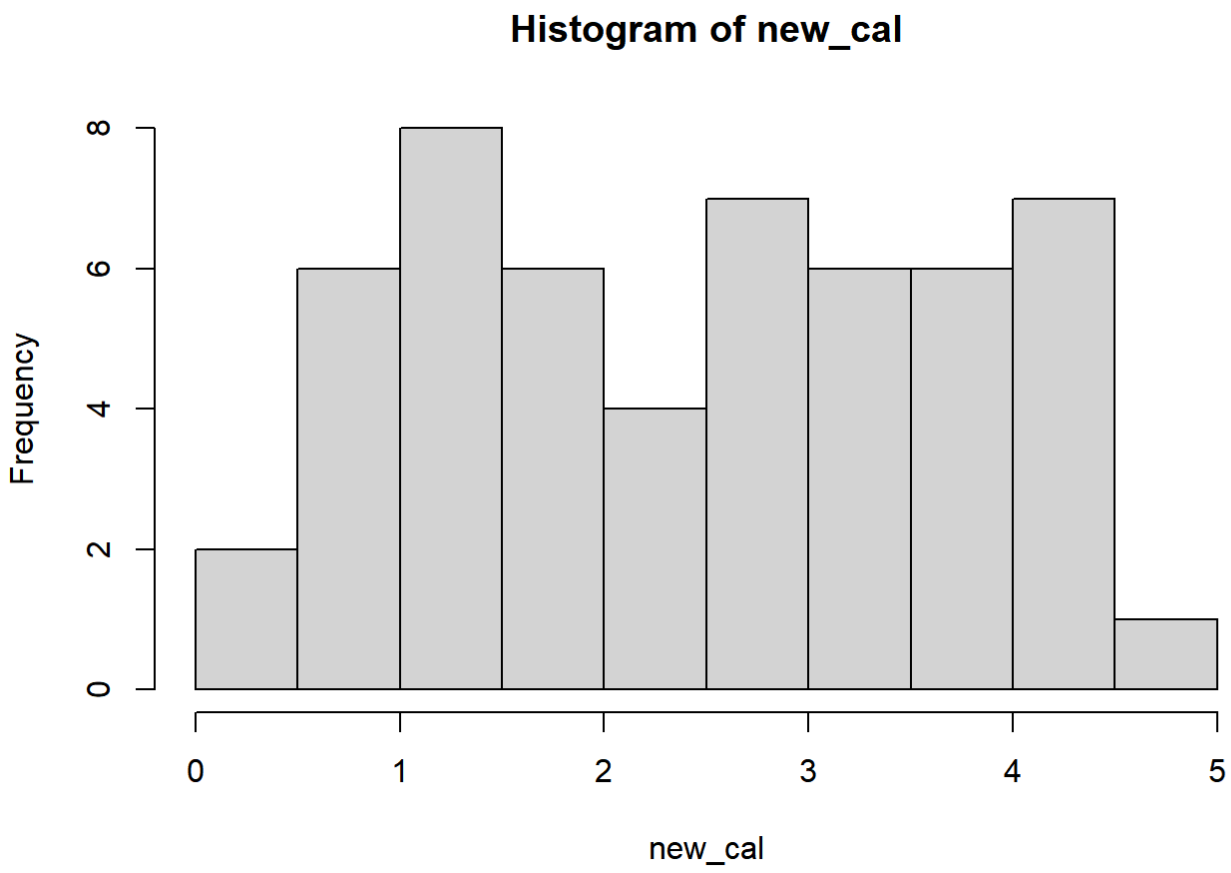
Code



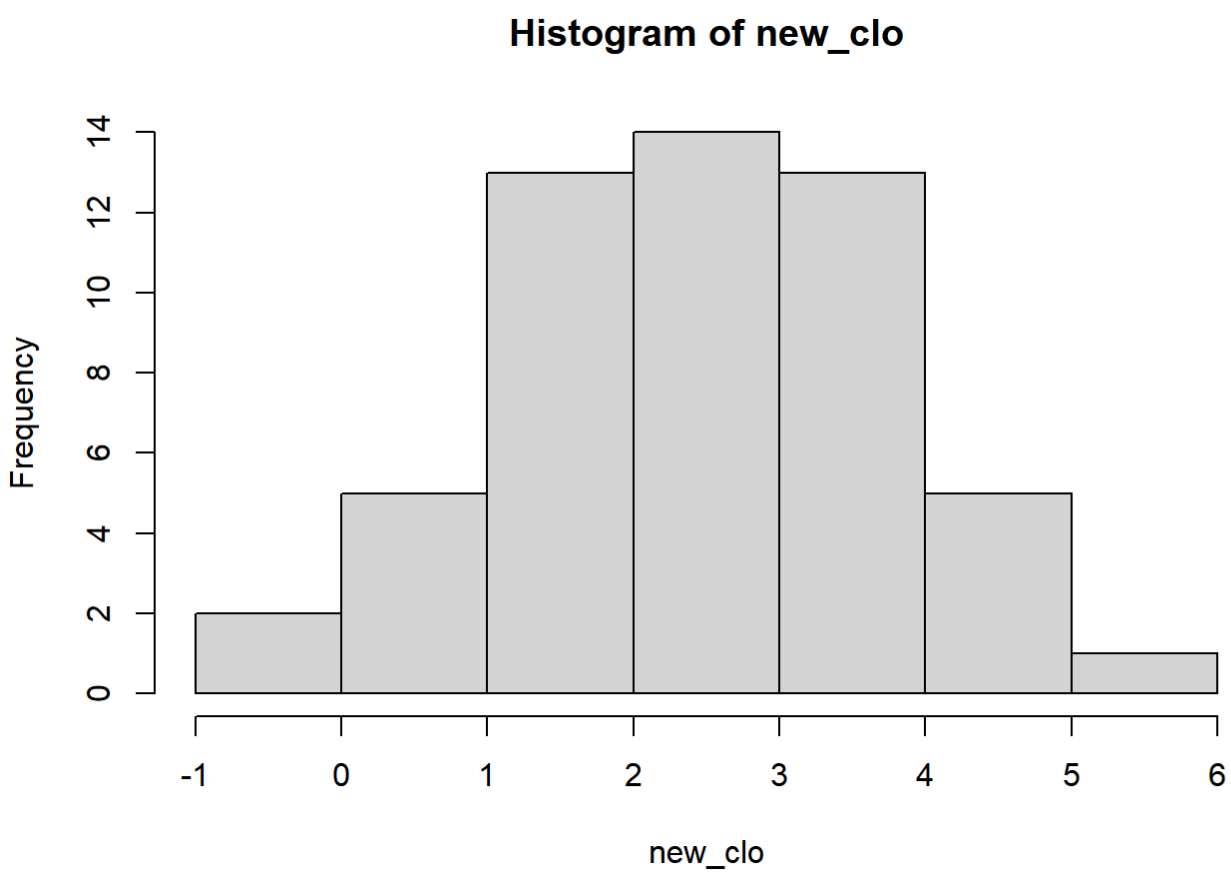
Code



Code



Code



Code

alkalinity <dbl>	ph <dbl>	calcium <dbl>	chlorophyll <dbl>
1.7749524	6.1	1.09861229	-0.3566749
1.2527630	5.1	0.64185389	1.1631508
4.7535902	9.1	3.78645978	4.8543713
3.6737658	6.9	2.79728133	1.2527630
0.9162907	4.6	1.06471074	0.5877867
2.9755296	7.3	1.50407740	3.7864598
1.6486586	5.4	1.02961942	1.2237754
4.2682979	8.1	4.01096295	3.5174978
3.2733640	5.8	2.21920348	0.4700036
1.5686159	6.4	1.52605630	3.1135153

1-10 of 53 rows Previous **1** 2 3 4 5 6 Next

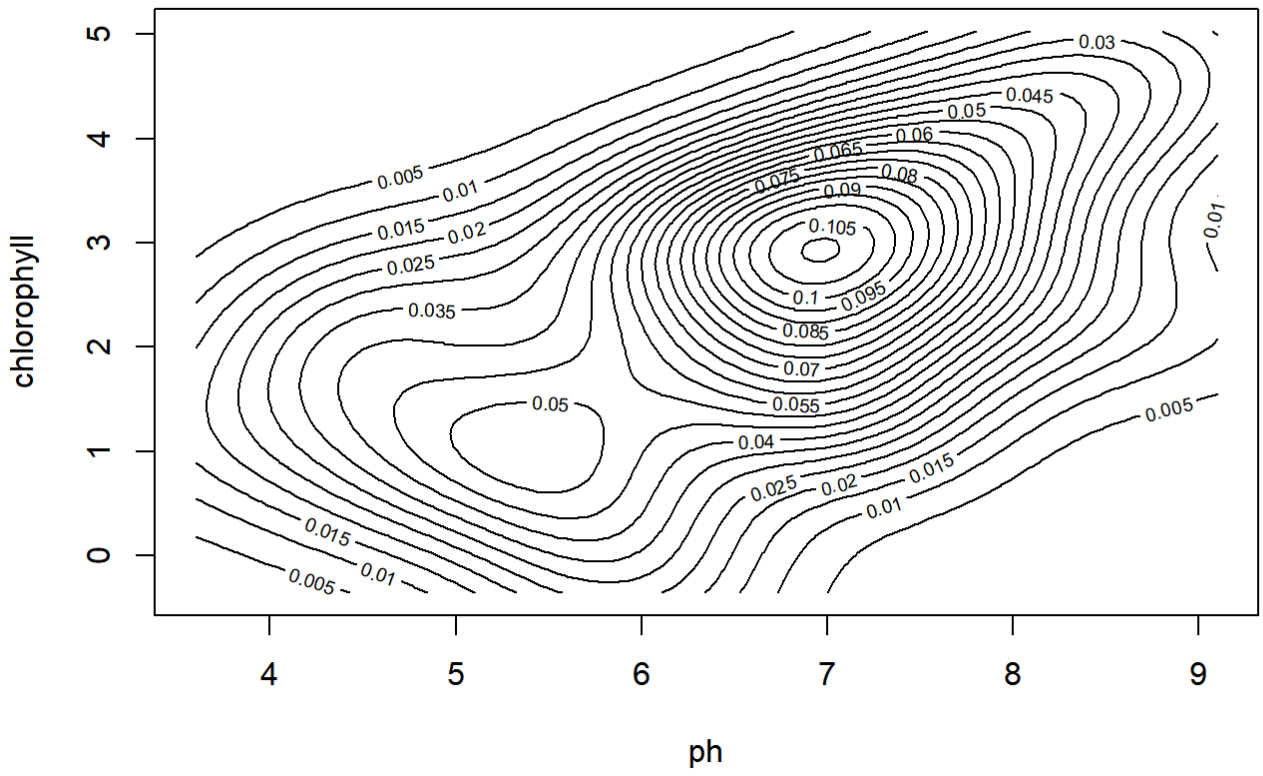
Code

```
$multivariateNormality
             Test          Statistic           p value Result
1 Mardia Skewness   28.4037564402811 0.100180836177609    YES
2 Mardia Kurtosis -0.919973180011851  0.35758677479266    YES
3             MVN               <NA>              <NA>    YES

$univariateNormality
              Test    Variable Statistic   p value Normality
1 Anderson-Darling alkalinity     0.8704    0.0239    NO
2 Anderson-Darling     ph         0.3496    0.4611    YES
3 Anderson-Darling   calcium      0.7818    0.0398    NO
4 Anderson-Darling chlorophyll    0.1744    0.9213    YES

$Descriptives
             n     Mean  Std.Dev   Median         Min      Max     25th
alkalinity  53 3.001047 1.220579 2.975530  0.18232156 4.852030 1.887070
ph          53 6.590566 1.288449 6.800000  3.60000000 9.100000 5.800000
calcium     53 2.404765 1.266219 2.533697  0.09531018 4.507557 1.193922
chlorophyll 53 2.419753 1.277739 2.549445 -0.35667494 5.026509 1.526056
                75th        Skew   Kurtosis
alkalinity  4.197202 -0.11596564 -1.1313325
ph          7.400000 -0.24587711 -0.6239638
calcium     3.572346  0.03301418 -1.3017952
chlorophyll 3.206803 -0.12172629 -0.5803851
```

A traves del analisis de normalidad y la transformacion de las variables con boxcox, se puede observar que las variables ph y chlorophyll cumplen con la normalidad del sesgo y la kurtosis, al ser validadas con el test de Mardia y Anderson-Darling.

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

Code

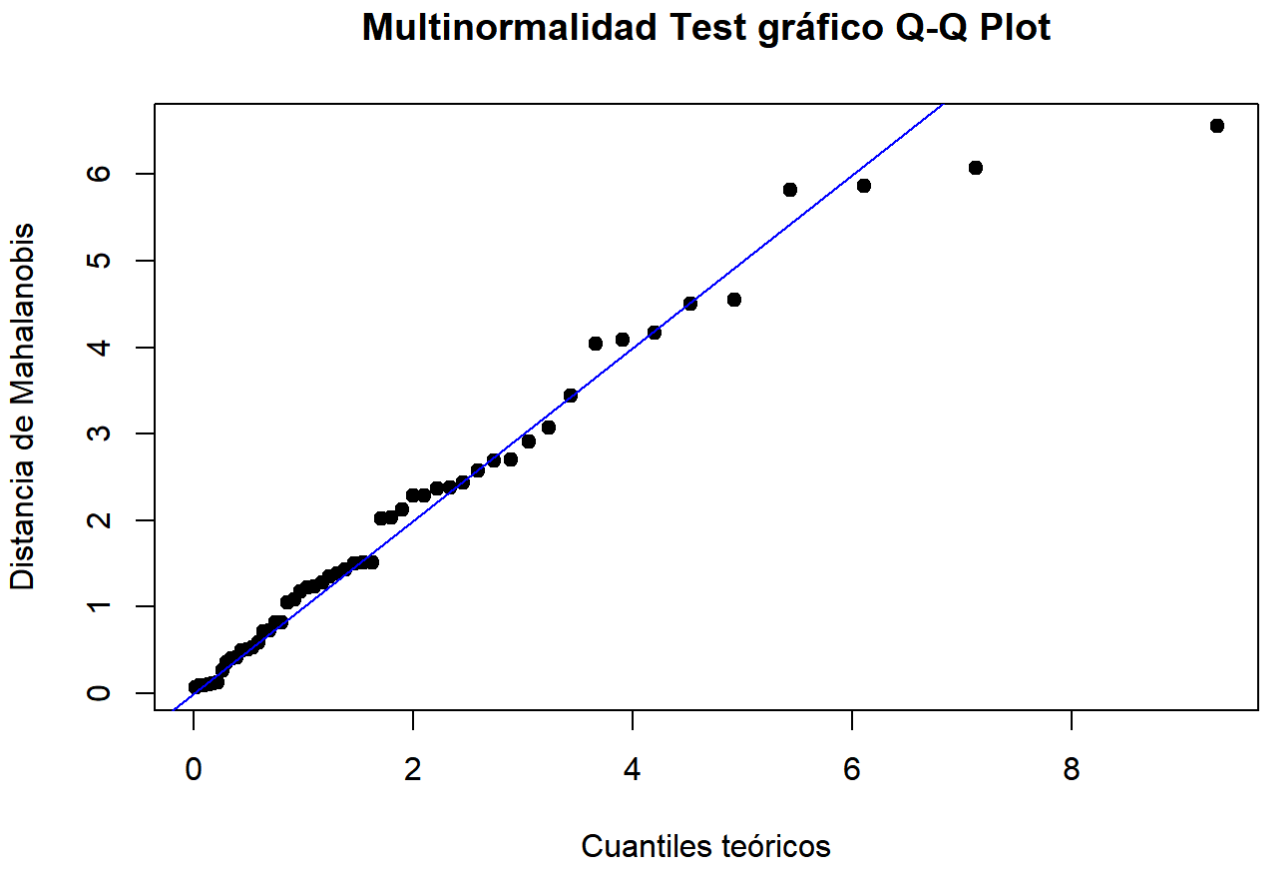


D. Detecta datos atípicos o influyentes en la normal multivariada encontrada en el inciso B (auxíliate de la distancia de Mahalanobis y del gráfico QQplot multivariado)

Code

```
[1] 6.554116 1.431772 4.503727 2.033376 2.707709 1.179796
```

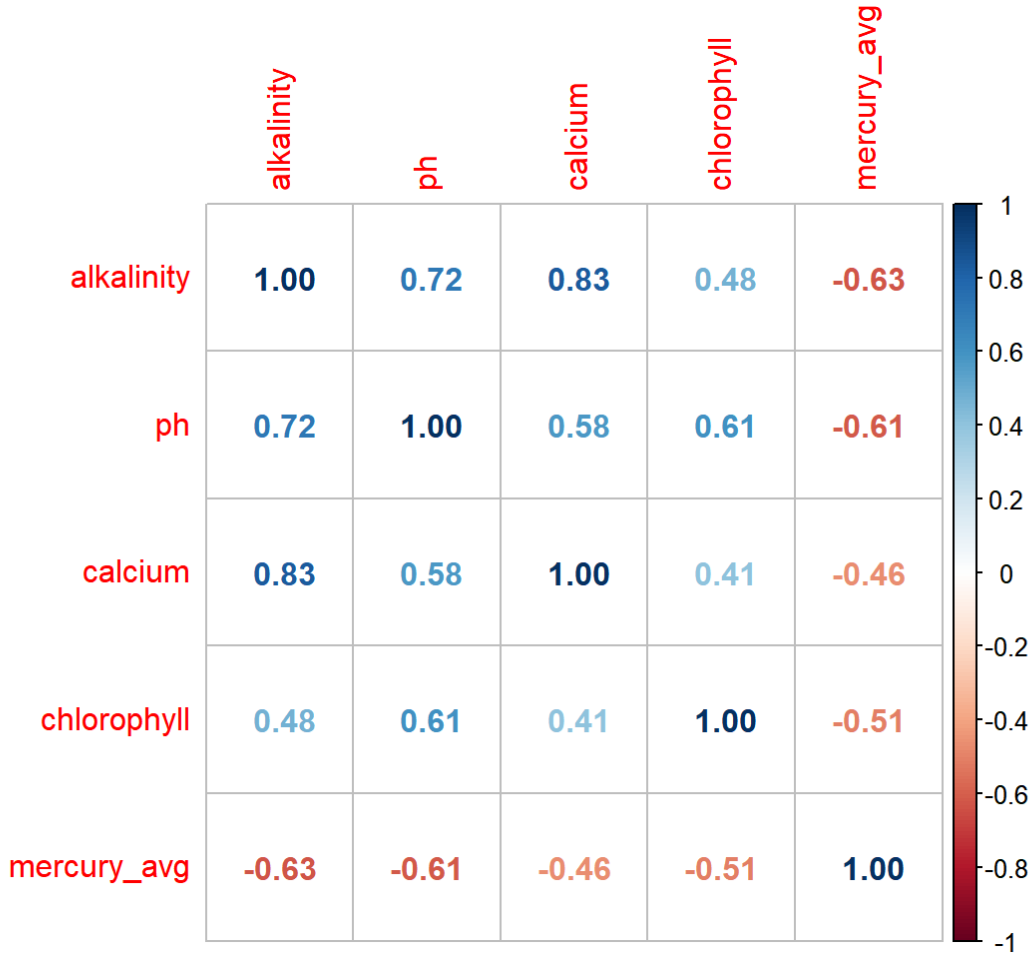
Code



2. Realice un análisis de componentes principales con la base de datos completa para identificar los factores principales que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce. Tome en cuenta los puntos que se sugieren a continuación (no son exhaustivos):

A. Justifique por qué es adecuado el uso de componentes principales para analizar la base (haz uso de la matriz de correlaciones)
El uso de componentes principales es conveniente gracias a que las variables que se encuentran en la base de datos son cuantitativas. Por lo tanto, es posible realizar un analisis de componentes principales para encontrar la variabilidad de los datos y asi deducir cuales son las variables que mas influyen en el caso de trabajo.

Code



B. Realiza el análisis de componentes principales y justifica el número de componentes principales apropiados para reducir la dimensión de la base

Code

```
[1] "Media de las variables"
```

Code

```
 alkalinity          ph     calcium chlorophyll mercury_avg
 37.5301887   6.5905660  22.2018868  23.1169811   0.5132075
```

Code

```
[1] "Varianza de las variables"
```

Code

```
  alkalinity           ph      calcium  chlorophyll  mercury_avg
1459.5094557    1.6601016  621.6332656  949.6456676    0.1147376
```

Code

```
[1] "sdev"     "rotation" "center"   "scale"    "x"
```

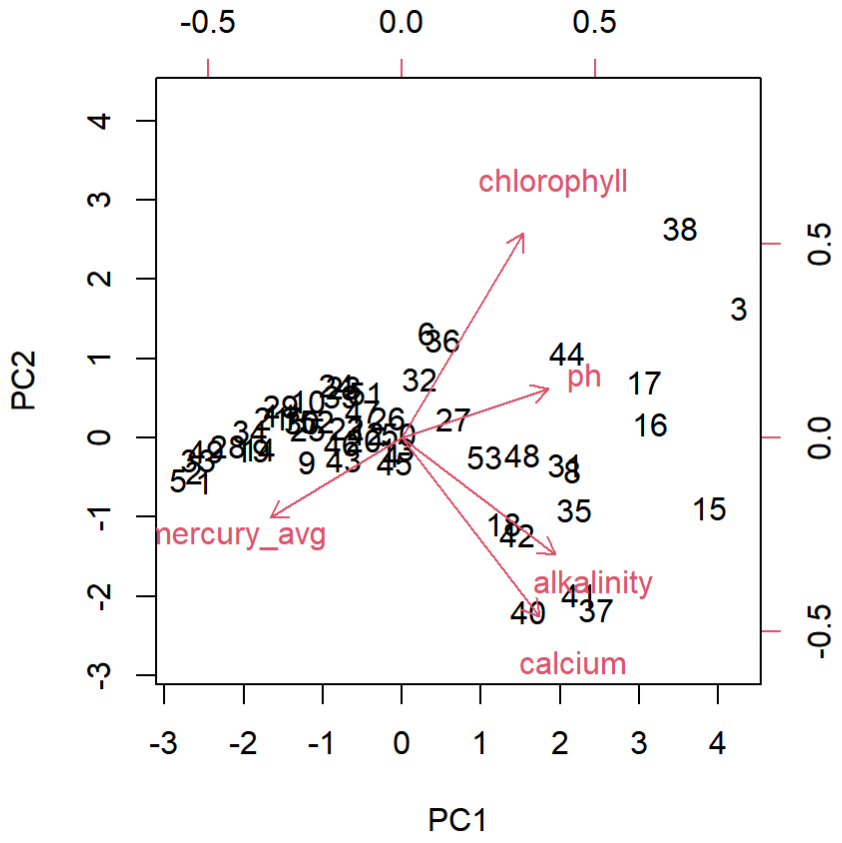
```
sdev

rotation

center

scale

x
```

Code

```
                  PC1        PC2         PC3          PC4         PC5
alkalinity    0.4946582 -0.3766753  0.01262775 -0.002757119  0.78311041
ph            0.4723842  0.1576048 -0.05172452 -0.835975061 -0.22468682
calcium       0.4433694 -0.5771095 -0.26705356  0.306689143 -0.55226107
chlorophyll   0.3916530  0.6589432 -0.52795953  0.356885456  0.07933007
mercury_avg -0.4268465 -0.2569335 -0.80442915 -0.282334264  0.15801386
```

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

Code



D. Interprete los resultados. Explique brevemente a qué conclusiones llega con su análisis y qué significado tienen los componentes seleccionados en el contexto del problema

3. Emite una conclusión general: Une las conclusiones aquí obtenidas con las ya obtenidas en el análisis que ya habías realizada anteriormente, ¿de qué forma te ayuda este nuevo análisis a contestar la pregunta principal del estudio:

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

   Por medio de este estudio se extrapola que las variables que mas influyen en el nivel de contaminacion por mercurio en los peces de los lagos de Florida son: ph, chlorophyll, alkalinity y calcium.

2. ¿En qué puede facilitar el estudio la normalidad encontrada en un grupo de variables detectadas?

   La normalidad encontrada en un grupo de variables detectadas puede facilitar el estudio ya que se puede realizar un analisis de componentes principales para encontrar la variabilidad de los datos y asi deducir cuales son las variables que mas influyen en el caso de trabajo.

3. ¿Cómo te ayudan los componentes principales a abordar este problema?

   Para poder encontrar la relacion entre las variables que influyen en el nivel de contaminacion por mercurio en los peces de los lagos de Florida, se puede realizar un analisis de componentes principales para encontrar la variabilidad de los datos entre si.