Momento de Retroalimentación Estadística 2

Ricardo Arriaga A01570553

Introduccion y Resumen

Este análisis busca identificar cuáles son los principales factores que influyen en el nivel contaminación de mercurio en peces de lagos de florida. Para esto se buscó si existía una correlación entre el nivel de alcalinidad en el agua.

Para obtener este resultado se buscó la correlación con todas las variables disponibles del dataset y se encontró que la alcalinidad tenía la mayor correlación además de variables relacionadas con el mercurio pero estas se descartaron por su alta dependencia.

Análisis de los resultados

Descripción de datos

Para comenzar, lo primero que debemos hacer es conocer nuestros datos y entender cómo es que se comportan entre sí y que es lo que aportan a nuestro sistema. Para realizar este análisis tomaremos las variables numéricas de nuestro dataset y buscaremos en ellas su promedio, mediana, moda, desviación estándar, varianza y máximo y mínimo valor. Por último, para la variable "Lago" buscaremos solo su moda, aunque sabemos que será 1 ya que el dataset hace observaciones de 53 lagos diferentes.

```
## ------------------------ Alcalinidad ------------------------
## Promedio:  37.53019   Mediana:  19.6   Moda:  17.3 25.4
## Desviacion estandar:  38.20353   Varianza:  1459.509
## Minimo:  1.2   Maximo:  128
##
## ---------------------------- PH ----------------------------
## Promedio:  6.590566   Mediana:  6.8   Moda:  5.8 6.9
## Desviacion estandar:  1.288449   Varianza:  1.660102
## Minimo:  3.6   Maximo:  9.1
##
## -------------------------- Calcio --------------------------
## Promedio:  22.20189   Mediana:  12.6   Moda:  3 3.3 5.2 6.3 20.5
## Desviacion estandar:  24.93257   Varianza:  621.6333
## Minimo:  1.1   Maximo:  90.7
##
## ------------------------ Clorofila ------------------------
```

```
## Promedio:  23.11698   Mediana:  12.8   Moda:  1.6 3.2 9.6
## Desviacion estandar:  30.81632   Varianza:  949.6457
## Minimo:  0.7   Maximo:  152.4
##
## ---------------------- con_med_mercurio ---------------------
## Promedio:  0.5271698   Mediana:  0.48   Moda:  0.34
## Desviacion estandar:  0.3410356   Varianza:  0.1163053
## Minimo:  0.04   Maximo:  1.33
##
## ------------------------ num_peces ------------------------
## Promedio:  13.0566   Mediana:  12   Moda:  12
## Desviacion estandar:  8.560677   Varianza:  73.2852
## Minimo:  4   Maximo:  44
##
## ---------------------- min_con_mercurio ---------------------
## Promedio:  0.2798113   Mediana:  0.25   Moda:  0.04
## Desviacion estandar:  0.2264058   Varianza:  0.05125958
## Minimo:  0.04   Maximo:  0.92
##
## ---------------------- max_con_mercurio ---------------------
## Promedio:  0.8745283   Mediana:  0.84   Moda:  0.06 0.26 0.4 0.48 0.69
0.84 1.4 1.5 1.9
## Desviacion estandar:  0.5220469   Varianza:  0.2725329
## Minimo:  0.06   Maximo:  2.04
##
## --------------------- prom_mercurio_pez ---------------------
## Promedio:  0.5132075   Mediana:  0.45   Moda:  0.16
## Desviacion estandar:  0.3387294   Varianza:  0.1147376
## Minimo:  0.04   Maximo:  1.53
##
## --------------------------- edad ---------------------------
## Promedio:  0.8113208   Mediana:  1   Moda:  1
## Desviacion estandar:  0.3949977   Varianza:  0.1560232
## Minimo:  0   Maximo:  1

## --------------------------- Lago ---------------------------
##
##         Alligator             Annie            Apopka      Blue
Cypress
##                 1                 1                 1
1
##             Brick            Bryant            Cherry
Crescent
##                 1                 1                 1
1
##        Deer Point              Dias              Dorr
Down
##                 1                 1                 1
1
## East Tohopekaliga             Eaton           Farm-13
```

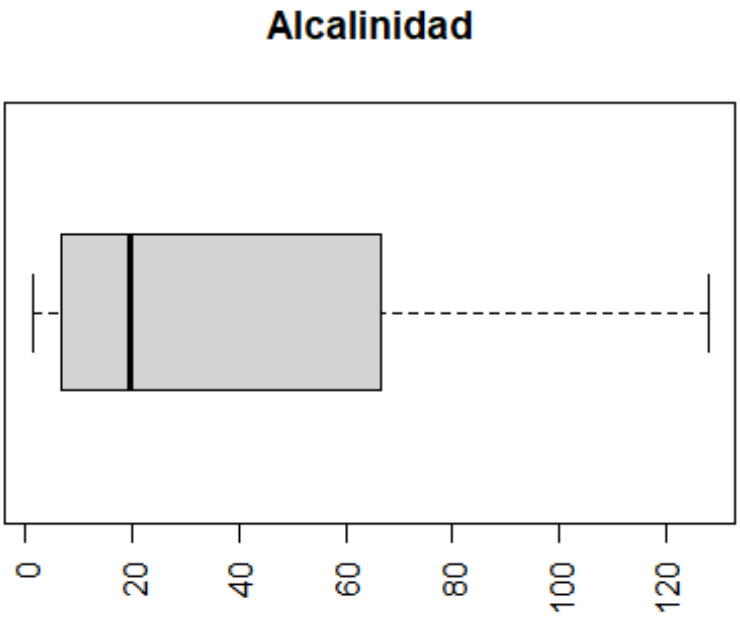
```
George
##                 1                 1                 1
1
##           Griffin            Harney              Hart
Hatchineha
##                 1                 1                 1
1
##           Iamonia         Istokpoga           Jackson
Josephine
##                 1                 1                 1
1
##          Kingsley         Kissimmee         Lochloosa
Louisa
##                 1                 1                 1
1
##        Miccasukee          Minneola            Monroe
Newmans
##                 1                 1                 1
1
##        Ocean Pond      Ocheese Pond        Okeechobee
Orange
##                 1                 1                 1
1
##       Panasoffkee            Parker            Placid
Puzzle
##                 1                 1                 1
1
##            Rodman          Rousseau           Sampson
Shipp
##                 1                 1                 1
1
##           Talquin            Tarpon      Tohopekaliga
Trafford
##                 1                 1                 1
1
##             Trout      Tsala Apopka              Weir
Wildcat
##                 1                 1                 1
1
##              Yale
##                 1
```

Viendo los resultados podemos comenzar a ver valores que podría servir y valores que podemos limpiar de nuestro dataset más adelante.

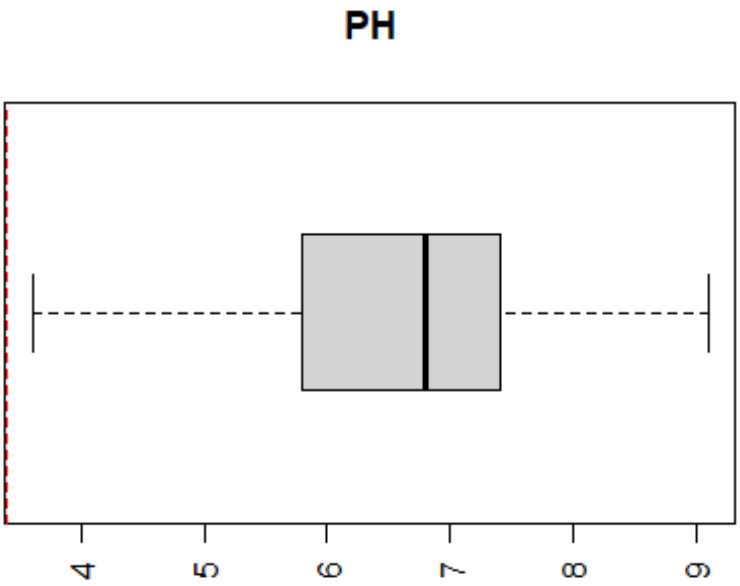
Cuartiles

Podemos hacer una exploración más profunda buscando los cuartiles de nuestras variables.

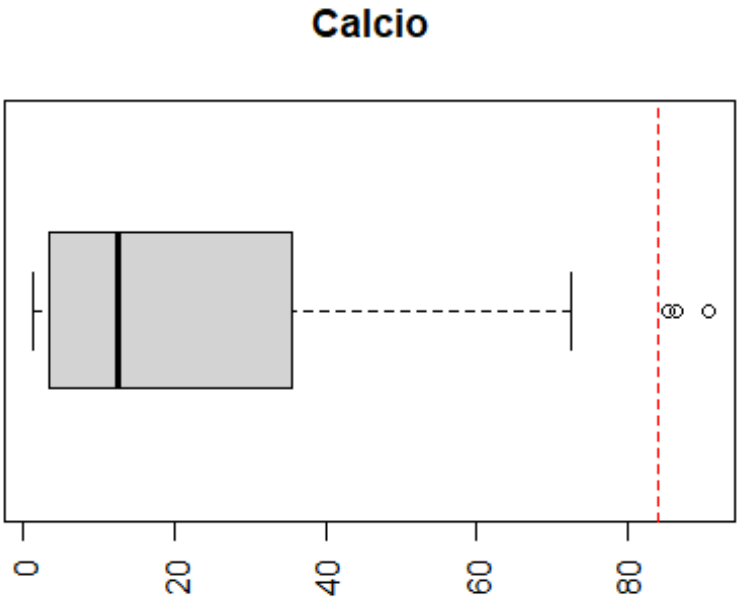
```
## ------------------------ Alcalinidad ------------------------
## Quartil 1:  6.6   Quartil 3:  66.5
```



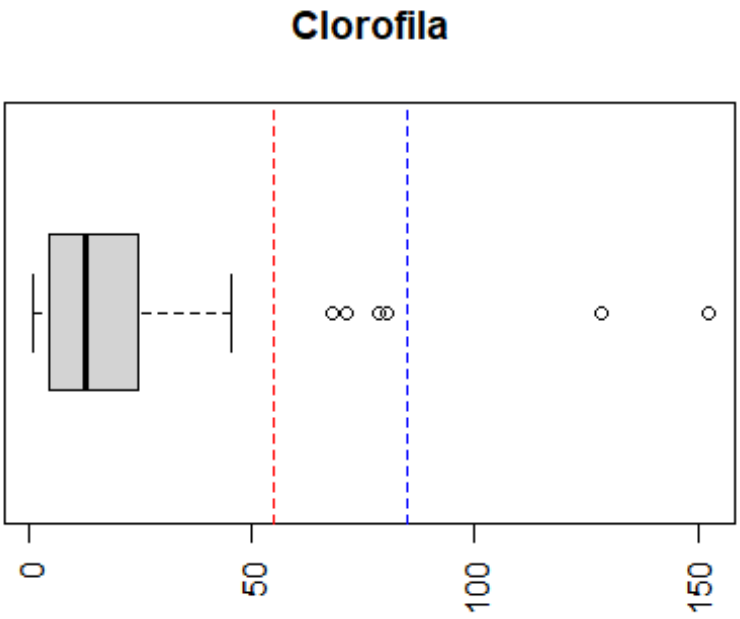

```
## ---------------------------- PH ----------------------------
## Quartil 1:  5.8   Quartil 3:  7.4
```



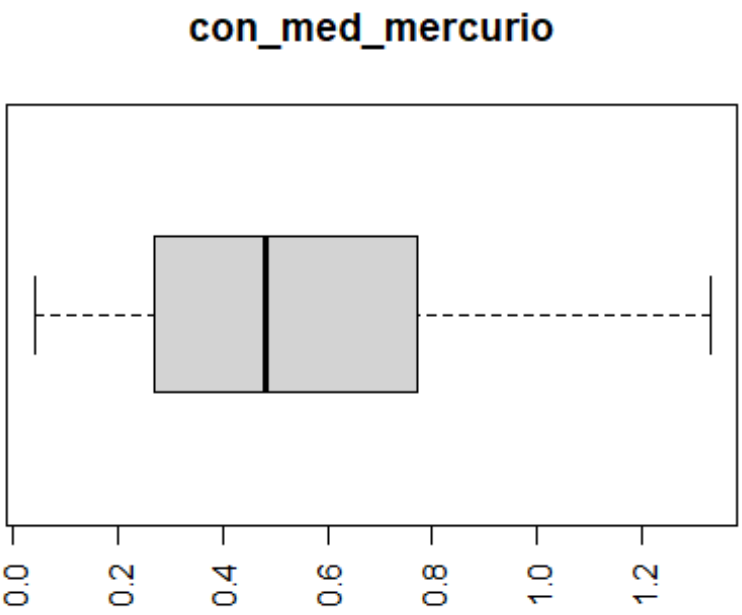

```
## --------------------------- Calcio --------------------------
## Quartil 1:  3.3   Quartil 3:  35.6
```



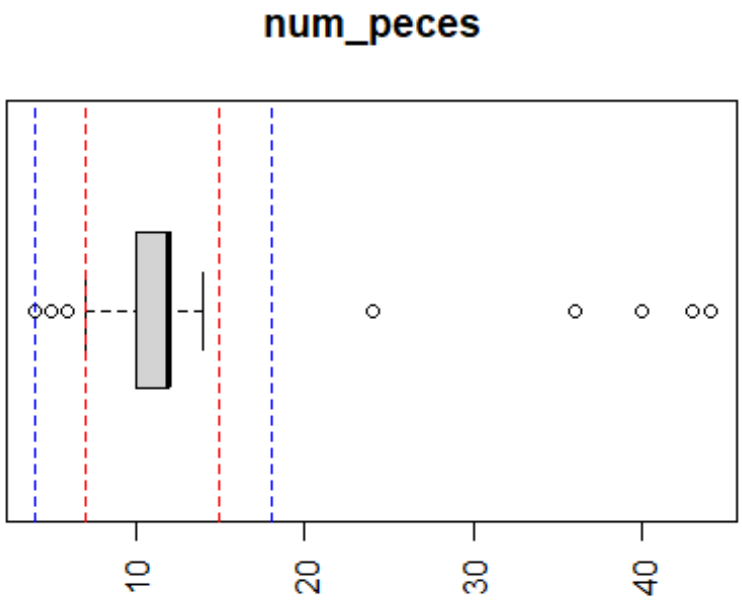
```
## ------------------------- Clorofila -------------------------
## Quartil 1:  4.6   Quartil 3:  24.7
```



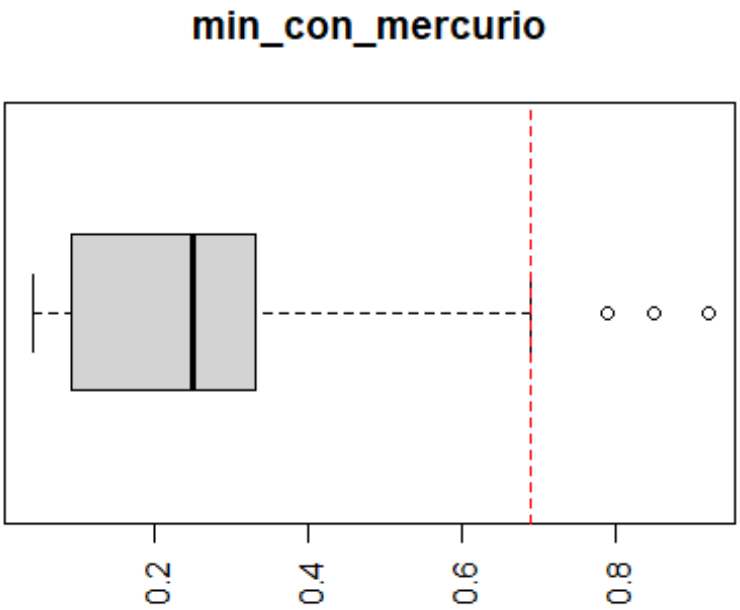

```
## ---------------------- con_med_mercurio ----------------------
## Quartil 1:  0.27   Quartil 3:  0.77
```



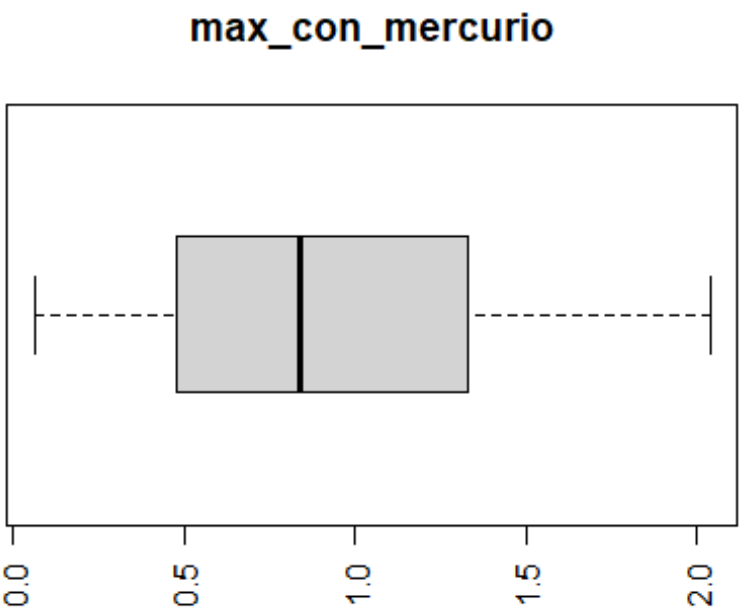

```
## ------------------------ num_peces ------------------------
## Quartil 1:  10   Quartil 3:  12
```



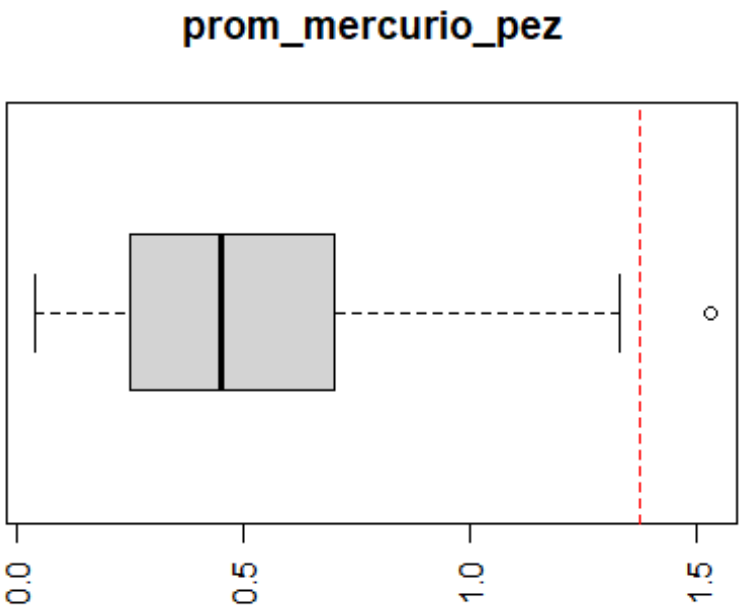
```
## ---------------------- min_con_mercurio ----------------------
## Quartil 1:  0.09   Quartil 3:  0.33
```



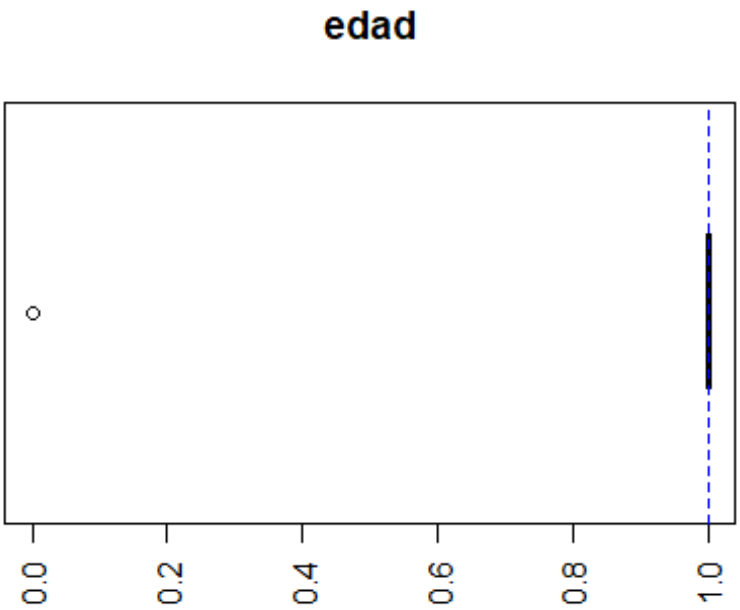

```
## ---------------------- max_con_mercurio ----------------------
## Quartil 1:  0.48   Quartil 3:  1.33
```




```
## ---------------------- prom_mercurio_pez ----------------------
## Quartil 1:  0.25   Quartil 3:  0.7
```



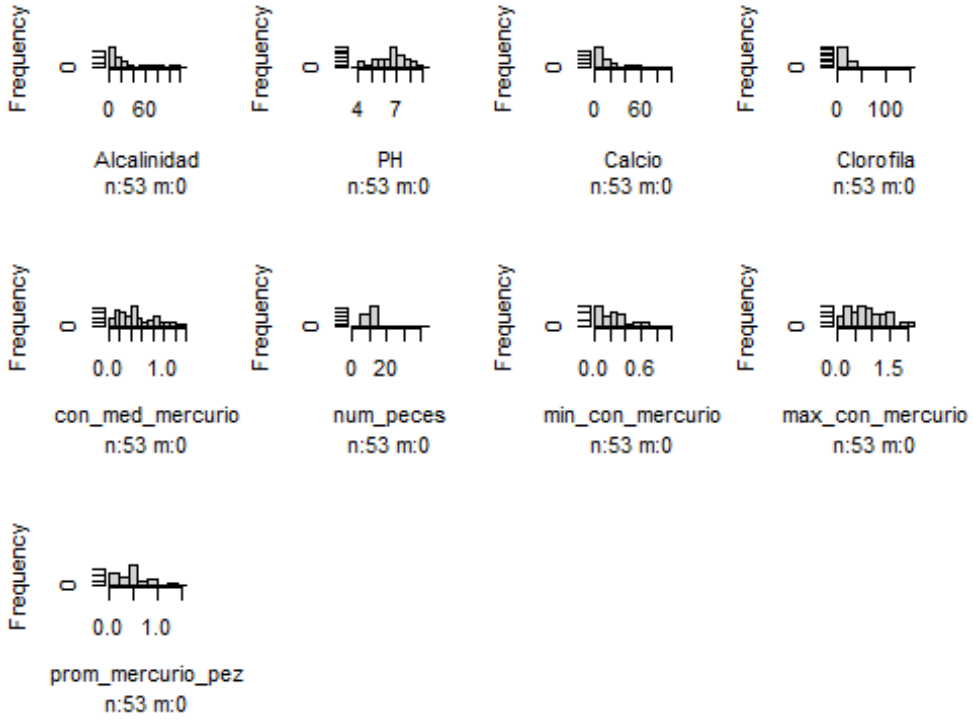
```
## ---------------------------- edad ----------------------------
## Quartil 1:  1   Quartil 3:  1
```

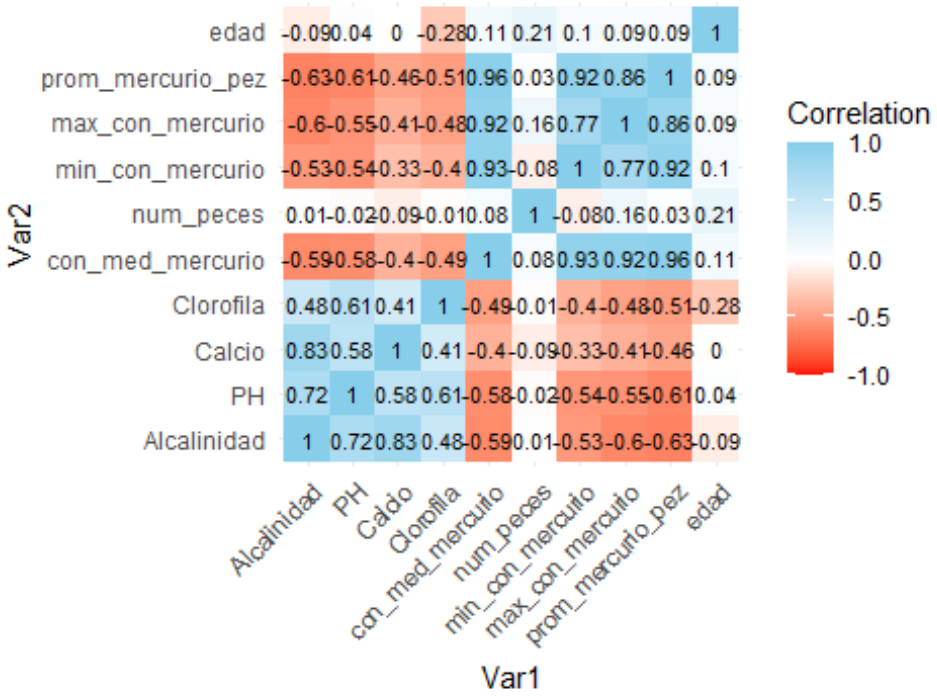


Histogramas y correlaciones

Para continuar con el análisis graficaremos en histogramas la frecuencia de las variables numéricas para ver si distribución y crearemos una matriz de correlación de todas contra todas nuestras variables numéricas para encontrar las variables que

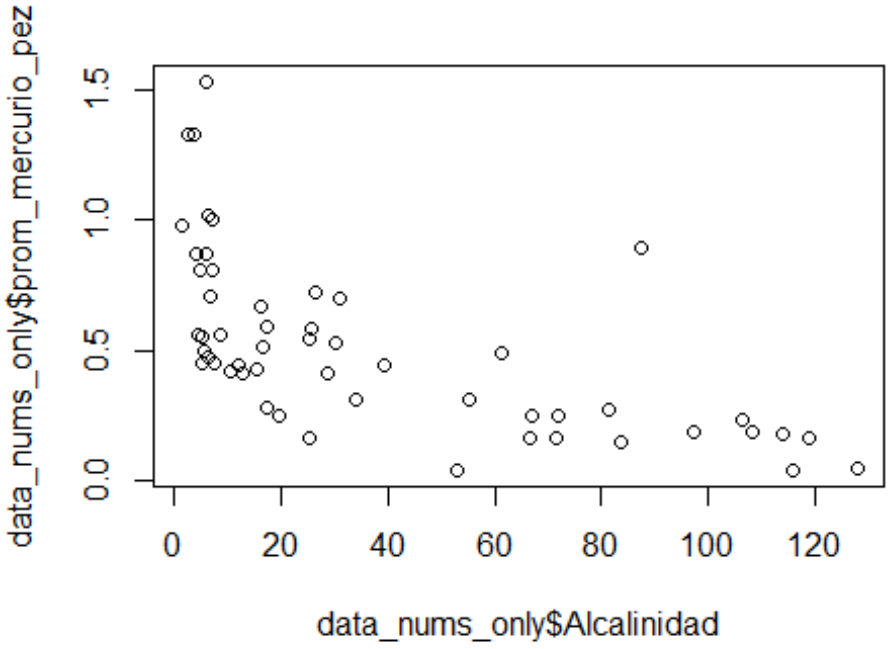
tengan mayor relevancia para realizar el análisis estadístico.





Con esta matriz podemos observar la correlación que tienen nuestras variables y podemos observar comportamientos interesantes que servirán para nuestro análisis,

específicamente variables como la alcalinidad o el calcio nos pueden decir información necesaria para entender la concentración de mercurio. También podemos ver como todas las variables relacionadas con la concentración de mercurio en general tienen una correlación alta, esto puede parecer bueno a primera vista, pero al tratarse de información muy similar existe dependencia entre estas variables. Para evitar crear un modelo erróneo habrá que limpiar nuestro dataset para dejar solamente las variables que tengan correlación entre ellas y que sean independientes.



Ya que la concentración media de mercurio (con_med_mercurio) y el promedio de mercurio por pez (prom_mercurio_pez) tienen la correlación más alta podemos elegir estas variables para incluirlas en un dataset limpio que creara el modelo de regresión.

Excluimos variables con correlación alta entre ellas

Regresión lineal con todas las variables

Realizamos una regresion lineal con nuestro dataset limpio.

```
##
## Call:
## lm(formula = prom_mercurio_pez ~ ., data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42440 -0.17183 -0.03828  0.11235  0.84598
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6905181  0.1048999   6.583 2.94e-08 ***
## Alcalinidad -0.0055498  0.0009894  -5.609 9.30e-07 ***
## num_peces    0.0010755  0.0044936   0.239    0.812
## edad         0.0208675  0.0978255   0.213    0.832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2712 on 49 degrees of freedom
## Multiple R-squared:  0.3959, Adjusted R-squared:  0.359
## F-statistic: 10.71 on 3 and 49 DF,  p-value: 1.59e-05
```

Búsqueda del mejor modelo

Ahora buscamos el mejor modelo para conseguir los parametros de nuestra ecuacion y realizar una segunda regresion.

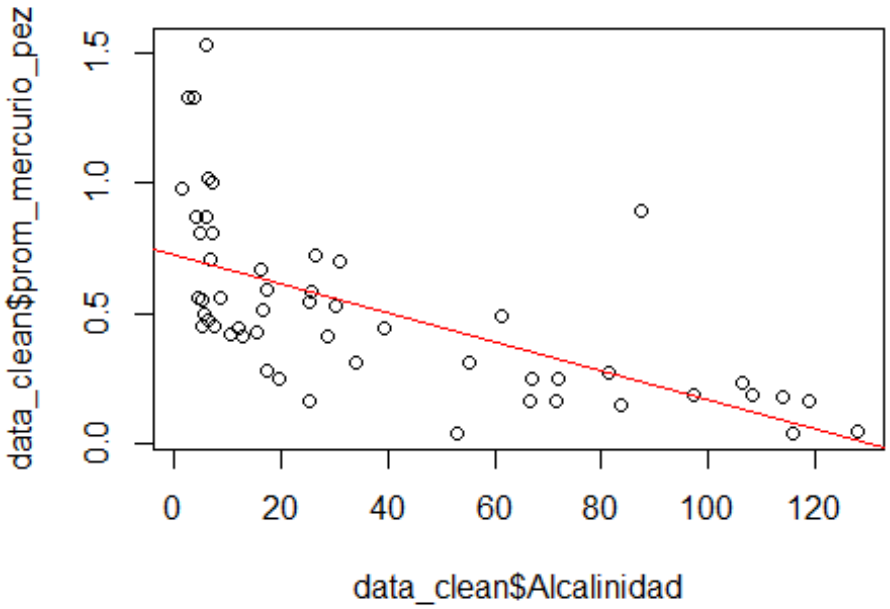
```
##
## Call:
## lm(formula = prom_mercurio_pez ~ Alcalinidad, data = data_clean)
##
## Coefficients:
## (Intercept)  Alcalinidad
##    0.722167    -0.005568
```

Regresion lineal con el mejor modelo

```
##
## Call:
## lm(formula = prom_mercurio_pez ~ Alcalinidad, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42075 -0.19154 -0.03631  0.10914  0.84068
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.7221665  0.0514965  14.024  < 2e-16 ***
## Alcalinidad -0.0055678  0.0009662  -5.762 4.82e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2662 on 51 degrees of freedom
## Multiple R-squared:  0.3943, Adjusted R-squared:  0.3825
## F-statistic:  33.2 on 1 and 51 DF,  p-value: 4.822e-07
```

Ecuacion de la regresion lineal

```
## con_med_mercurio =  0.7222  + -0.0056 * prom_mercurio_pez
```



Validación del modelo

Pruebas de hipótesis

Realizaremos las siguientes pruebas de hipótesis para validar que nuestro modelo sea correcto:

h0: beta1 = 0 h1: beta1 != 0

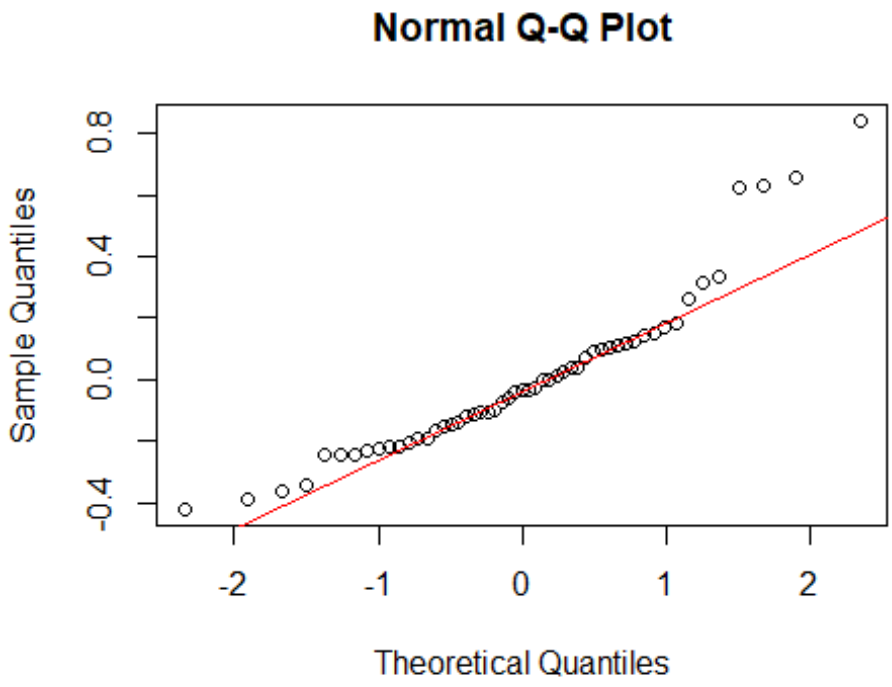
Reglas de decision: * Si p-value < alpha, se rechaza H0 y se acepta H1 * Si p-value > alpha, se rechaza H1 y se acepta H0 * Si t* > t, se rechaza H0 y se acepta H1 * Si t* < t, se rechaza H1 y se acepta H0

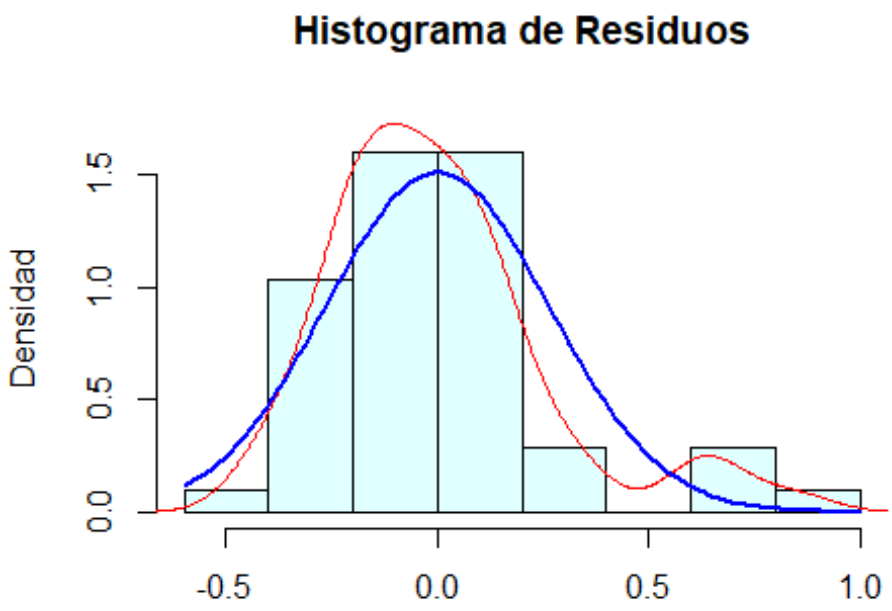
```
## La variable Alcalinidad es significativa. (t* > t0 & p < alpha)
##  t* = -5.7623 , t0 = 2.0076
##  p-value = 4.822478e-07 , alpha = 0.05
```

Verificación de supuestos

Normalidad de los residuos

En este grafico podemos observar que los residuos del modelo intentan seguir una distribución normal pero en las colas la normalidad se pierde.





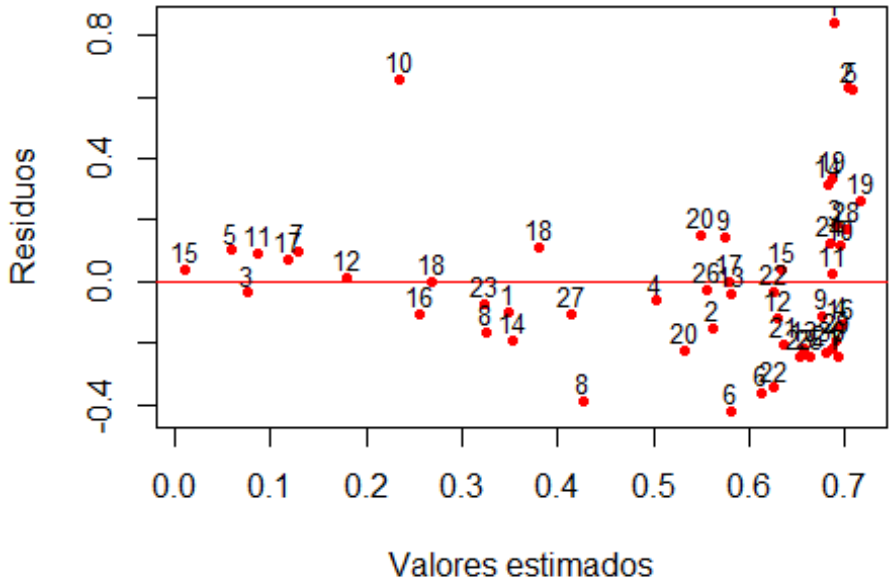

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  E
## W = 0.90775, p-value = 0.0005999
```

Homocedasticidad y modelo apropiado

Gráfica Valores estimados vs Residuos

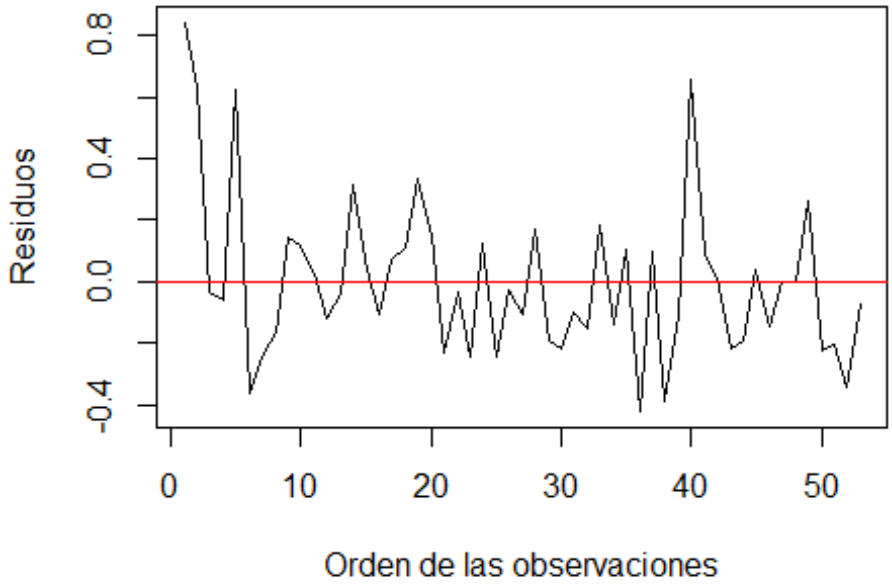
Con esta grafica podemos comprobar la homocedasticidad del modelo. Podemos observar un cambio drástico en la varianza de los valores graficados cambia a tener una tendencia heterosemántica ya que podemos ver desde la X = 0.0 hasta 0.6 los valores se agrupados se comienzan a expandir.



Independencia

Errores vs Orden de observación

En esta grafica podemos ver como nuestros ejes no siguen ningún tipo de patrón y demostrando independencia entre las variables elegidas para el modelo.

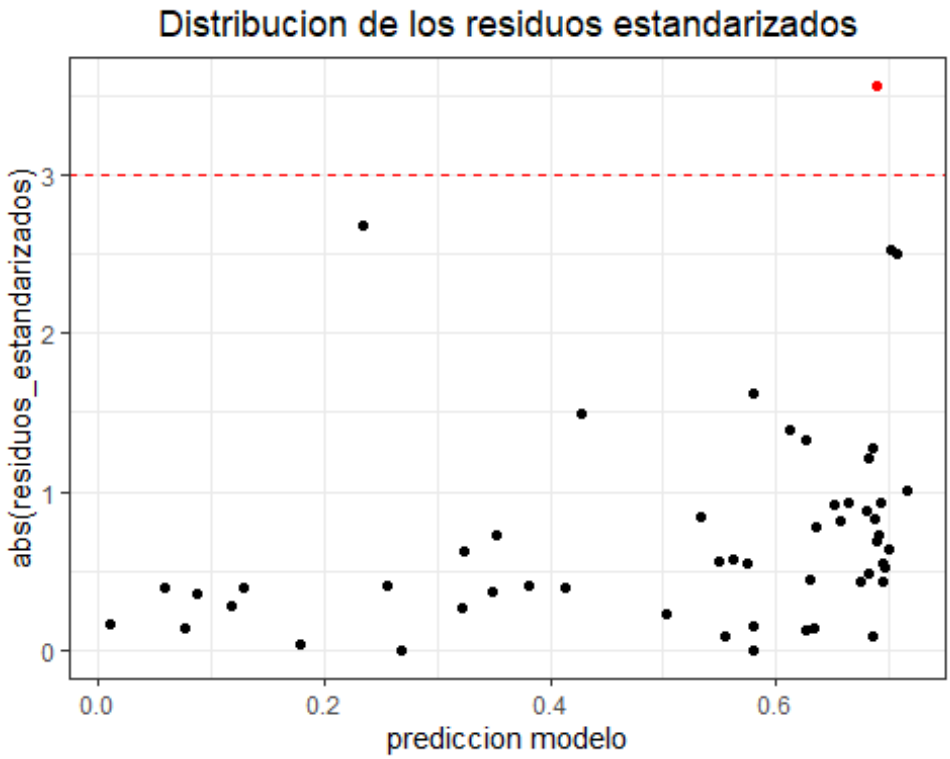



```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.09941852      1.604134   0.166
##  Alternative hypothesis: rho != 0
```

Datos atípicos o influyentes

Datos atípicos

Se estandarizan los residuos y se observa si hay distancias mayores a 3.



Conclusión

Con este análisis podemos concluir que la alcalinidad puede no ser un factor para la contaminación de mercurio ya que a pesar de que el modelo cumple con algunas pruebas la variable no se normaliza cuando se mueve hacia las colas. Una corrección que se podría hacer seria intentar usar alguna variable relacionada con el mercurio ya que Florida es un estado en el que se reportar niveles peligrosos de mercurio en el aire y ese es un factor por el cual los ríos del estado están contaminados.