Módulo 1: Estadística para ciencia de datos

Reporte "Los peces y el mercurio"

Inteligencia artificial avanzada para la ciencia de datos I (Grupo 102)

Daniel Salvador Cázares García A01197517
Septiembre de 2022

Resumen

En este trabajo, por medio del uso de datos de un estudio realizado en 53 lagos de Florida, se busco realizar un análisis de las variables que influyen en la concentración de mercurio en los peces. Para dicha tarea, se emplearon metodos para calcular medidas estadísticas y visualizar datos. Así mismo, se busco construir modelos estadisticos y validar su utilidad por medio del uso de supuestos e hipotesis. Esto, con el objetivo encontrar modelos que permitan entender el compartamiento actual de diferentes variables sobre la concentración de mercurio en peces, así como hacer predicciones a futuro del comportamiento de este fenomeno.

Introducción

La contaminación por mercurio en el los cuerpos de agua y los peces comestibles, es un riesgo para la salud que cada vez se hace más presente. ¿Qué factores influyen en esta concentración? ¿Esta concentración es dañina para la salud humana? En este trabajo se buscará examinar los factores que influyen en el nivel de contaminación por mercurio para comprender las relaciones existentes y utilizar tecnicas estadisticas que permitan modelar esta situación.

1. EXPLORACIÓN DE LA BASE DE DATOS

Base de datos

Base de datos: Mediciones del estudio de Mercurio en los lagos de Florida

id	lago	alcalinidad	PH	calcio	clorofila	avg-mercurio	num-peces	min-mercurio	max-mercurio	est-mercurio	edad
1	Alligator	5.9	6.1	3.0	0.7	1.23	5	0.85	1.43	1.53	1
2	Annie	3.5	5.1	1.9	3.2	1.33	7	0.92	1.90	1.33	0
3	Apopka	116.0	9.1	44.1	128.3	0.04	6	0.04	0.06	0.04	0
4	Blue Cypress	39.4	6.9	16.4	3.5	0.44	12	0.13	0.84	0.44	0
5	Brick	2.5	4.6	2.9	1.8	1.20	12	0.69	1.50	1.33	1

	alcalinidad	PH	calcio	clorofila	avg-mercurio	num-peces	min-mercurio	max-mercurio	est-mercurio	
count	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	5
mean	37.530189	6.590566	22.201887	23.116981	0.527170	13.056604	0.279811	0.874528	0.513208	
std	38.203527	1.288449	24.932574	30.816321	0.341036	8.560677	0.226406	0.522047	0.338729	
min	1.200000	3.600000	1.100000	0.700000	0.040000	4.000000	0.040000	0.060000	0.040000	
25%	6.600000	5.800000	3.300000	4.600000	0.270000	10.000000	0.090000	0.480000	0.250000	
50%	19.600000	6.800000	12.600000	12.800000	0.480000	12.000000	0.250000	0.840000	0.450000	
75%	66.500000	7.400000	35.600000	24.700000	0.770000	12.000000	0.330000	1.330000	0.700000	
max	128.000000	9.100000	90.700000	152.400000	1.330000	44.000000	0.920000	2.040000	1.530000	

Variables y significado

Cantidad de datos y variables presentes

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 53 entries, 1 to 53
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   lago          53 non-null     object 
 1   alcalinidad   53 non-null     float64
 2   PH            53 non-null     float64
 3   calcio        53 non-null     float64
 4   clorofila     53 non-null     float64
 5   avg-mercurio  53 non-null     float64
 6   num-peces     53 non-null     int64  
 7   min-mercurio  53 non-null     float64
 8   max-mercurio  53 non-null     float64
 9   est-mercurio  53 non-null     float64
 10  edad          53 non-null     int64  
dtypes: float64(8), int64(2), object(1)
memory usage: 5.0+ KB
```

Se tiene un total de 53 registros (1 registro por lago) y 12 columnas

Variables presentes

1. id = número de indentificación
2. lago = nombre del lago
3. alcalinidad = alcalinidad (mg/l de carbonato de calcio)
4. PH = PH
5. calcio = calcio (mg/l)
6. clorofila = clorofila (mg/l)

7. avg-mercurio = concentración media de mercurio (parte por millón) en el tejido muscualar del grupo de peces estudiados en cada lago
8. num-peces = número de peces estudiados en el lago
9. min-mercurio = mínimo de la concentración de mercurio en cada grupo de peces
10. max-mercurio = máximo de la concentración de mercurio en cada grupo de peces
11. est-mercurio = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
12. edad = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Clasificación de variables por tipo y medida

1. id: cuantitativa (númerica) | discreta
2. lago: cualitativa (categórica) | nominal
3. alcalinidad: cuantitativa (númerica) | continua
4. PH: cuantitativa (númerica) | continua
5. calcio: cuantitativa (númerica) | continua
6. clorofila: cuantitativa (númerica) | continua
7. avg-mercurio: cuantitativa (númerica) | continua
8. num-peces: cuantitativa (númerica) | discreta
9. min-mercurio: cuantitativa (númerica) | continua
10. max-mercurio: cuantitativa (númerica) | continua
11. est-mercurio: cuantitativa (númerica) | continua
12. edad: cualitativa (categórica) | nominal/ordinal

Exploración de la base de datos

Medidas estadísticas

Variables cuantitativas

Medidas de tendencia central

Promedio, mediana y moda

	promedio	mediana
alcalinidad	37.530189	19.60
PH	6.590566	6.80
calcio	22.201887	12.60
clorofila	23.116981	12.80
avg-mercurio	0.527170	0.48
num-peces	13.056604	12.00
min-mercurio	0.279811	0.25
max-mercurio	0.874528	0.84
est-mercurio	0.513208	0.45

	alcalinidad	PH	calcio	clorofila	avg-mercurio	num-peces	min-mercurio	max-mercurio	est-mercurio
0	17.3	5.8	3.0	1.6	0.34	12.0	0.04	0.06	0.16
1	25.4	6.9	3.3	3.2	NaN	NaN	NaN	0.26	NaN
2	NaN	NaN	5.2	9.6	NaN	NaN	NaN	0.40	NaN
3	NaN	NaN	6.3	NaN	NaN	NaN	NaN	0.48	NaN
4	NaN	NaN	20.5	NaN	NaN	NaN	NaN	0.69	NaN
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.84	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.40	NaN
7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.50	NaN
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.90	NaN

Medidas de dispersión

Rango máximo - mínimo, varianza, desviación estándar

	min	max	varianza	std
alcalinidad	1.20	128.00	1459.509456	38.203527
PH	3.60	9.10	1.660102	1.288449
calcio	1.10	90.70	621.633266	24.932574
clorofila	0.70	152.40	949.645668	30.816321
avg-mercurio	0.04	1.33	0.116305	0.341036
num-peces	4.00	44.00	73.285196	8.560677
min-mercurio	0.04	0.92	0.051260	0.226406
max-mercurio	0.06	2.04	0.272533	0.522047
est-mercurio	0.04	1.53	0.114738	0.338729

Variables cualitativas

Tabla de distribución de frecuencia

col_0	count
edad	
0	10
1	43

Se puede observar que la mayoría de los peces son maduros

Moda

```
0    1
Name: edad, dtype: int64
```

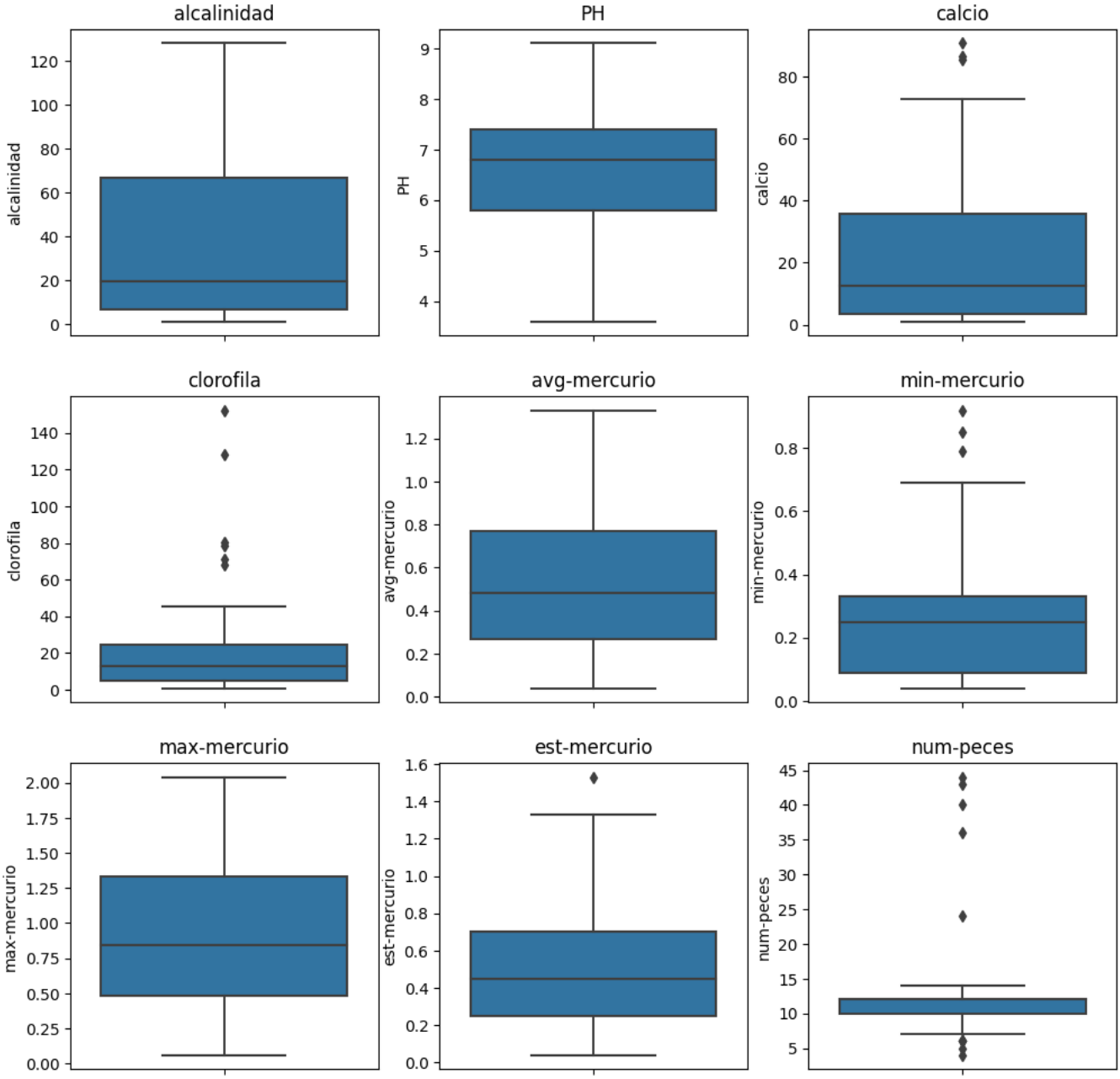
Visualización de datos

Variables cuantitativas

Medidas de posición

Boxplot, cuartiles, outliers

```
[Text(0.5, 1.0, 'num-peces')]
```



Observaciones:

- Se observan outliers en las variables: calcio, clorofila, min-mercurio, est-mercurio y num-peces
- Se puede ver que cada variable tiene escalas distintas, por lo que será necesario realizar un escalamiento de los mismos

Outliers

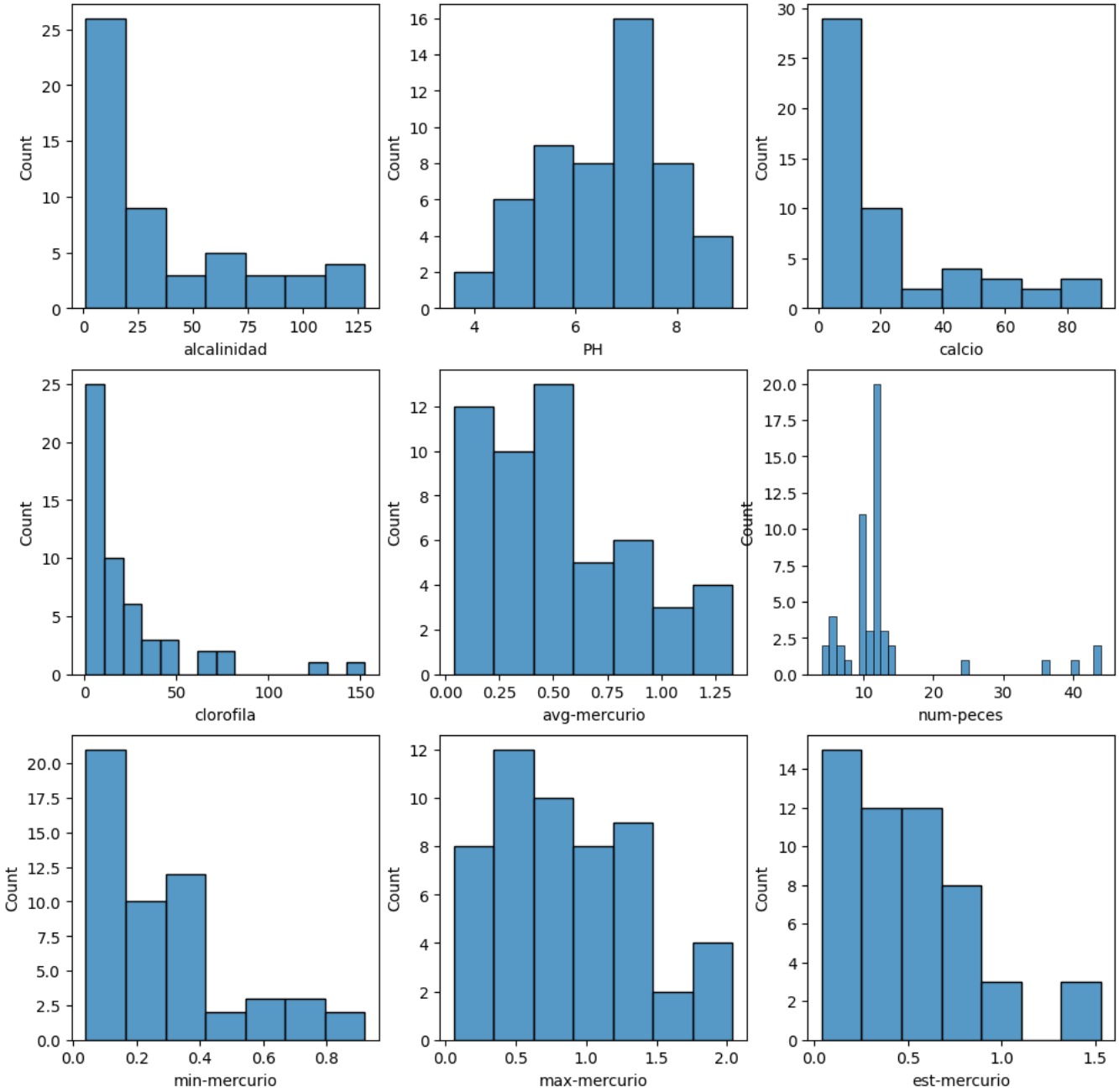
Se utilizo el rango intercuartil (IQR) para encontrar los outliers de cada variable.

```
alcalinidad:  0
PH:  0
calcio:  3
clorofila:  6
avg-mercurio:  0
num-peces:  11
min-mercurio:  3
max-mercurio:  0
est-mercurio:  1
```

Histogramas

Análisis de distribución de los datos y forma

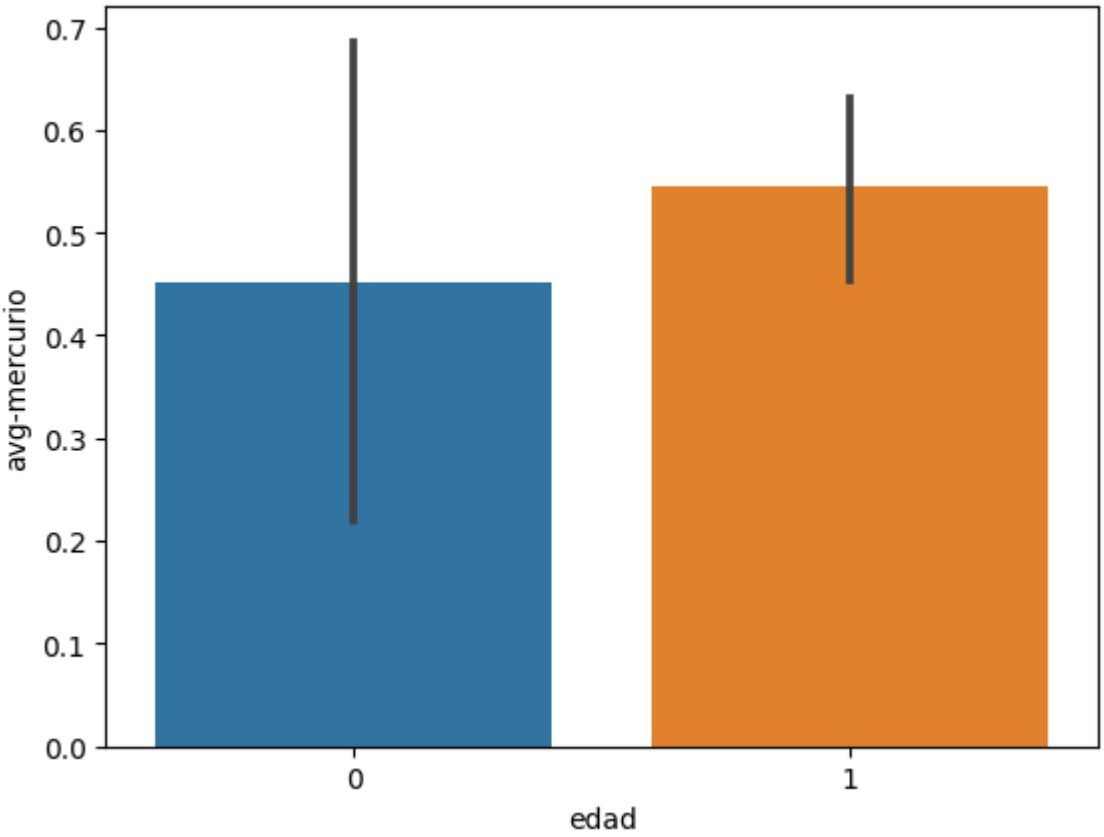
```
<AxesSubplot:xlabel='est-mercurio', ylabel='Count'>
```



La mayoría de las bariables no se distribuyen de forma normal, con excepción de max-mercurio y PH.

Mercurio y edad

```
<AxesSubplot:xlabel='edad', ylabel='avg-mercurio'>
```

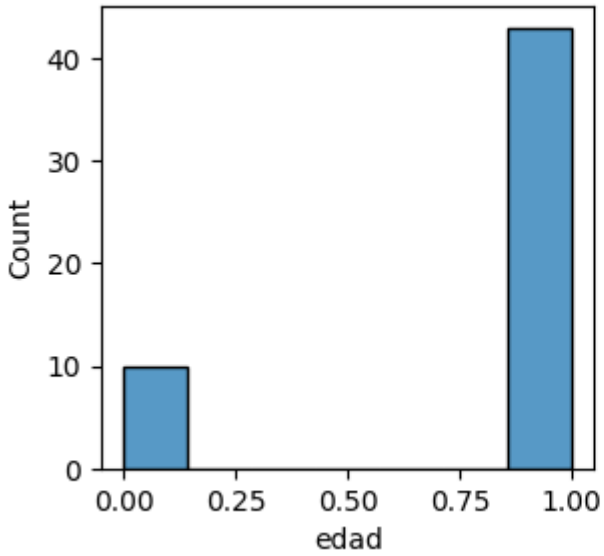


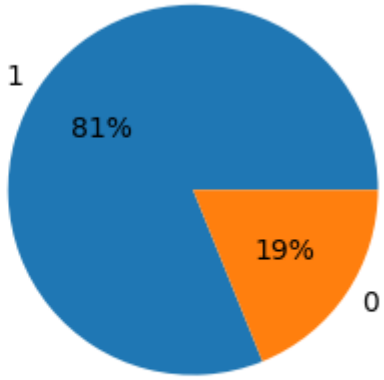
No parece que la concentración de mercurio varíe con relación a la edad de los peces. Sin embargo, aquellos más jovenes presentan una concentración ligeramente menor.

Variables categóricas

Distribución de los datos (diagramas de barras, diagramas de pastel)

```
([<matplotlib.patches.Wedge at 0x259c515a6b0>,
  <matplotlib.patches.Wedge at 0x259c515ac50>],
 [Text(-0.9123463061021291, 0.614511365022487, '1'),
  Text(0.9123463636368311, -0.6145112796024589, '0')],
 [Text(-0.49764343969207037, 0.33518801728499287, '81%'),
  Text(0.497643471074635, -0.33518797069225026, '19%')])
```

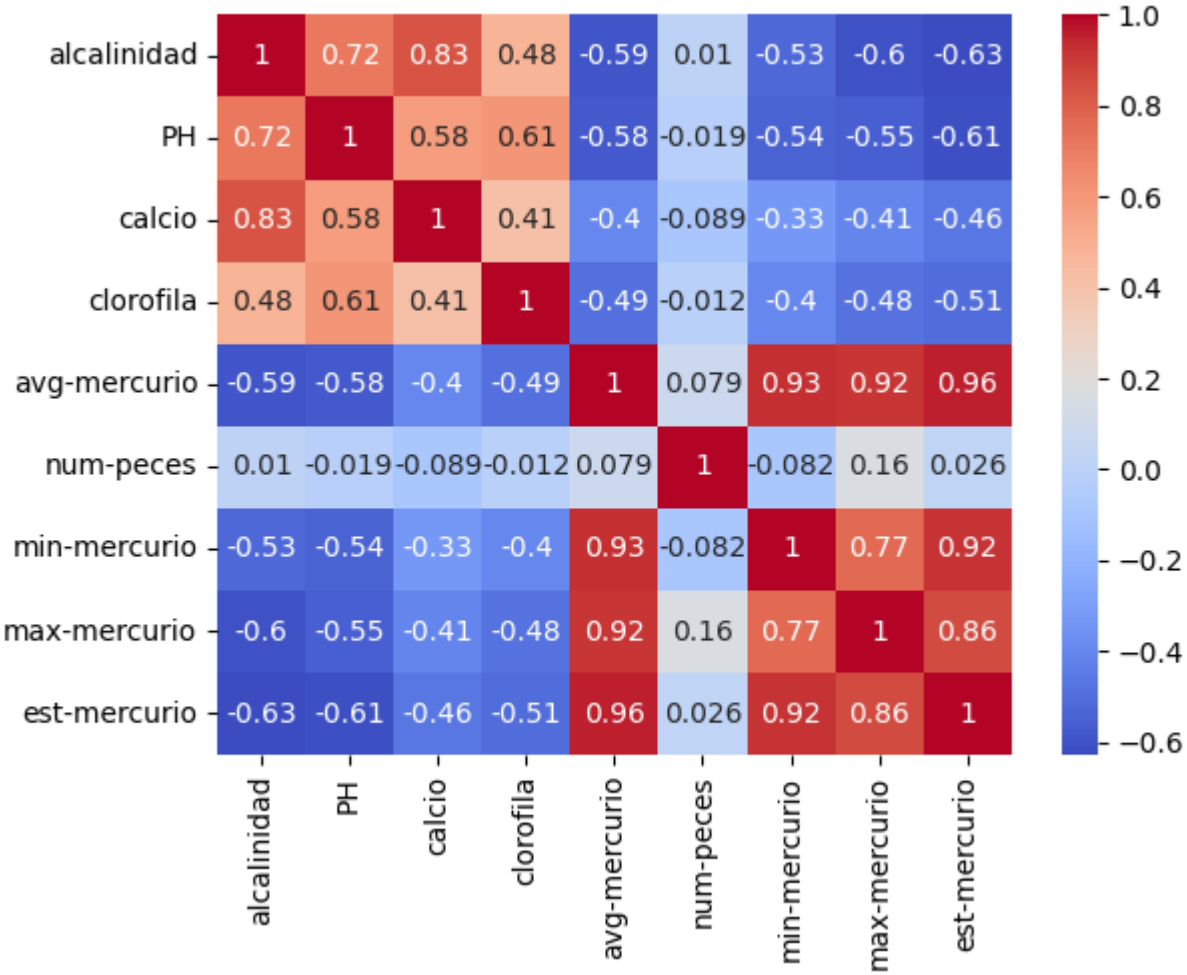




81% de los peces estudiados son maduros

Correlación entre variables

```
<AxesSubplot:>
```



Resultados: Correlación alta:

- alcalinidad: PH, calcio
- PH: alcalinidad
- calcio: alcalinidad

Correlación moderada:

- alcalinidad: mercurio
- PH: calcio, clorofila, mercurio
- calcio: PH
- clorofila: PH
- mercurio (avg): alcalinidad, PH

Observaciones:

- Calcio y alcalinidad tienen una correlación fuerte de 0.83, este se puede explicar debido a que esta variable se mide en mg / listros de carbonato de calcio.
- La alcalinidad también tiene una correlación fuerte de 0.72 con el PH, lo cual hace sentido pues el PH también es una medida de alcalinidad, pero para el agua de los lagos.

En este caso, las variables que mayormente nos interesan, son aquellas que se correlacionan con la cantidad de mercurio en los peces.

2. ANÁLISIS DE DATOS Y PREGUNTAS BASE

Modelos estadísticos y selección de variables

A continuación se realizaran modelos de regresión para las variables utilizadas.

Elección de variables

Como se pudo observar en la correlación, las variables que podrían tener una relación significativa con la cantidad promedio de mercurio son:

- Alcalinidad
- Ph
- Calcio
- Colorofila
- Mínimo, Máximo y Estimación de mercurio

Las realacionadas con mercurio no tiene sentido utilizarlas, pues ya se está utilizando el promedio. De este modo, las variables seleccionados son las siguientes:

- **Predictores**: `alcalinidad` , `PH` , `calcio` , `clorofila`
- **Objetivo**: `avg-mercurio`

Preprocesamiento de los datos

Antes de buscar un modelo que permita observar la relación de las diferentes variables con la cantidad de mercurio, es necesario procesar los datos para un correcto entrenamiento y ajuste. Este preprocesamiento incluye escalar y normalizar los datos.

Creación del modelo

Se realizara un modelo de regresión multiple con las variables seleccionadas

Evaluación del modelo

```
MSE:  0.048553141020570935
R^2:  0.4927067444223685
```

Se puede observar que el modelo tiene un error pequeño. En cuanto al coeficiente de determinación, hay cierto grado de relación, pero no es el más óptimo.

Verificación del modelo

En esta sección se utilizará el lenguaje R para una mayor facilidad en el manejo estadistico.

```
Call:
lm(formula = y ~ ., data = cbind(X, y))

Residuals:
     Min
      1Q
  Median
      3Q
     Max
-0.42260
-0.19155
-0.08438
 0.14334
 0.62234
Coefficients:

  Estimate
 Std. Error
 t value
 Pr(>|t|)

(Intercept)
  1.004440
   0.257561
   3.900
 0.000299
 ***
alcalinidad
 -0.005503
   0.002028
```

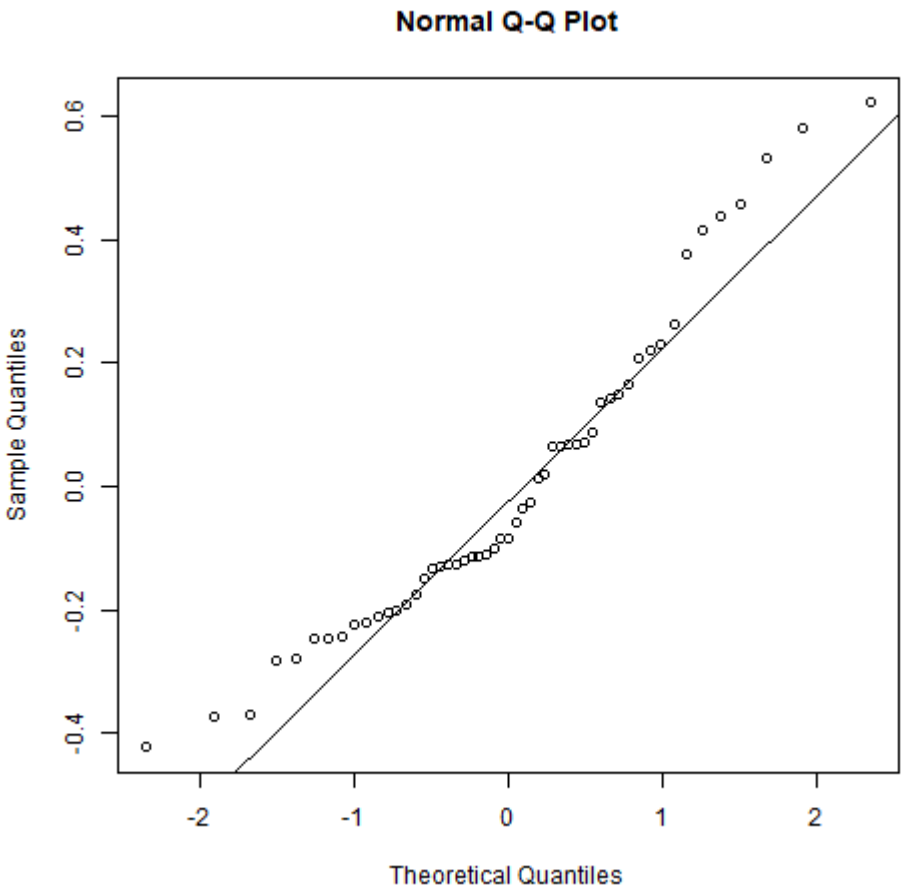
```
  -2.713
 0.009224
 **
PH
 -0.046709
   0.045329
  -1.030
 0.307968

calcio
  0.004129
   0.002648
   1.559
 0.125484

clorofila
 -0.002361
   0.001497
  -1.577
 0.121257

---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error:

0.2629

on

48

degrees of freedom
Multiple R-squared:

0.4515
,       Adjusted R-squared:

0.4058

F-statistic:

9.879

on

4

and

48

DF,  p-value:

6.499e-06
```

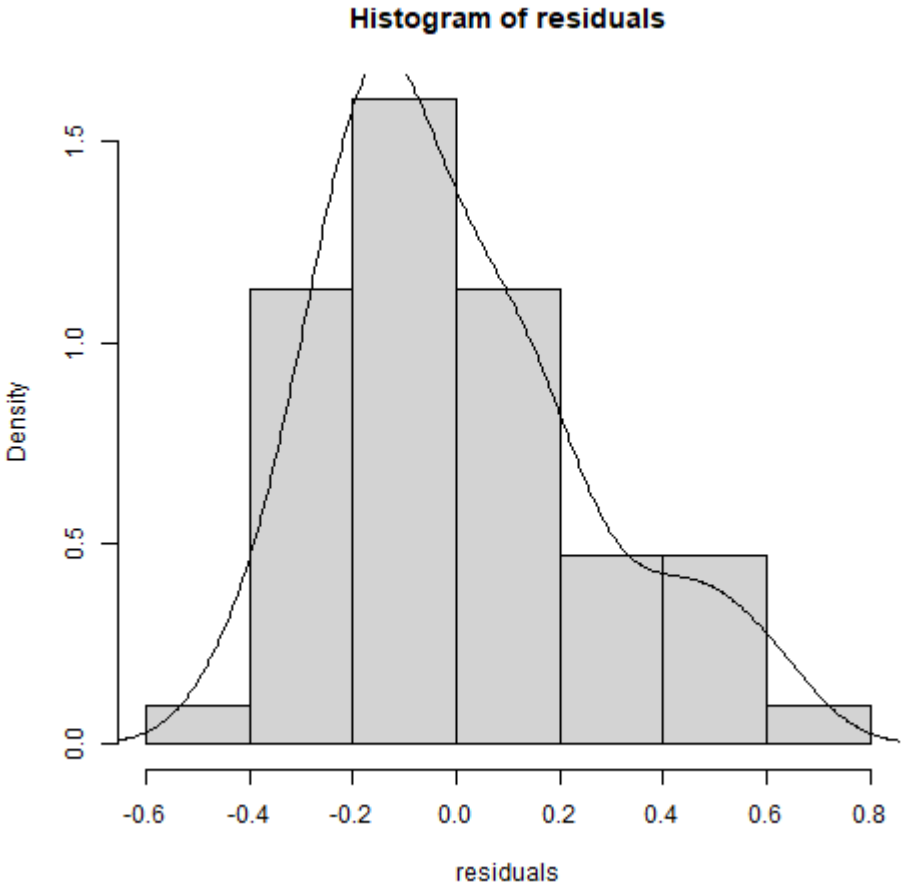
Normalidad de los residuos

Hipótesis:

- H0: población normal
- H1: población NO normal

Regla de decisión: $p < 0.05$






```
        Shapiro-Wilk normality test
data:
residuals
W = 0.94148, p-value = 0.01176
```

Observaciones:

- En la qqplot se observa que la mayoría de los residuos se acercan a la línea normal. Sin embargo, posiblemente es mejorable.
- El histograma se acerca a una distribución normal.
- Valor p = 0.01176 y p es menor a 0.05

Conclusión: Los residuos son normales

Verificación de media cero

Hipótesis:

- H0: media = 0
- H1: media != 0

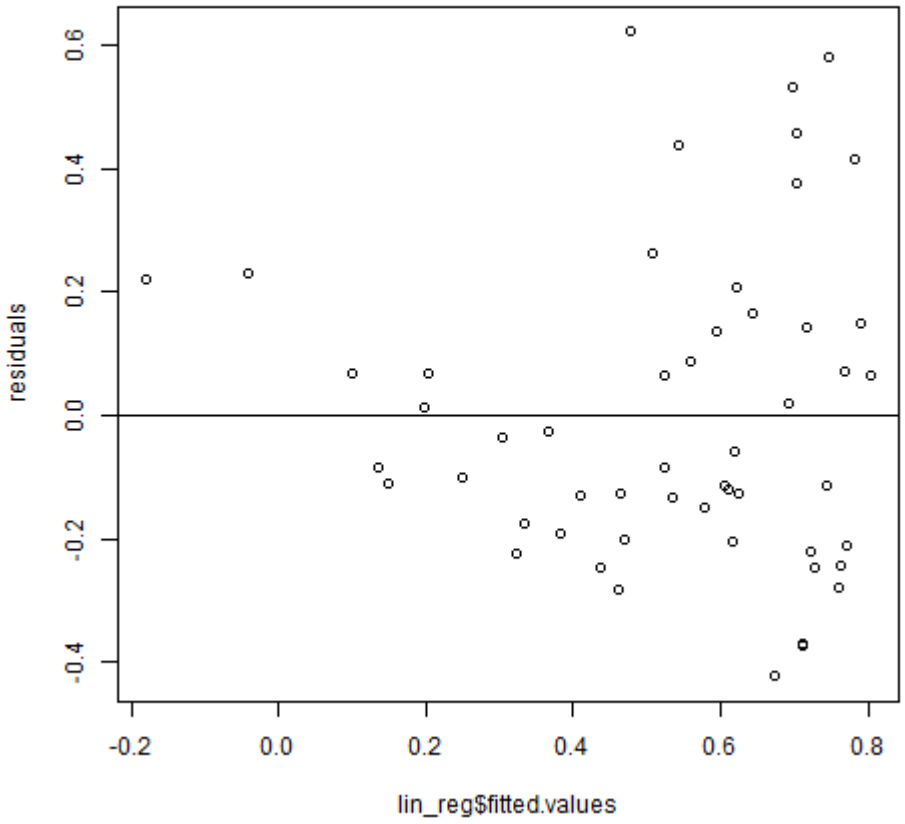
Regla de decisión: p < 0.05

```
        One Sample t-test
```

```
data:
residuals
t = -2.862e-16, df = 52, p-value = 1
alternative hypothesis:
true
mean
 is
not equal to

0
95
 percent confidence interval:

-0.06961592  0.06961592
sample estimates:
   mean of x
-9.92908e-18
```

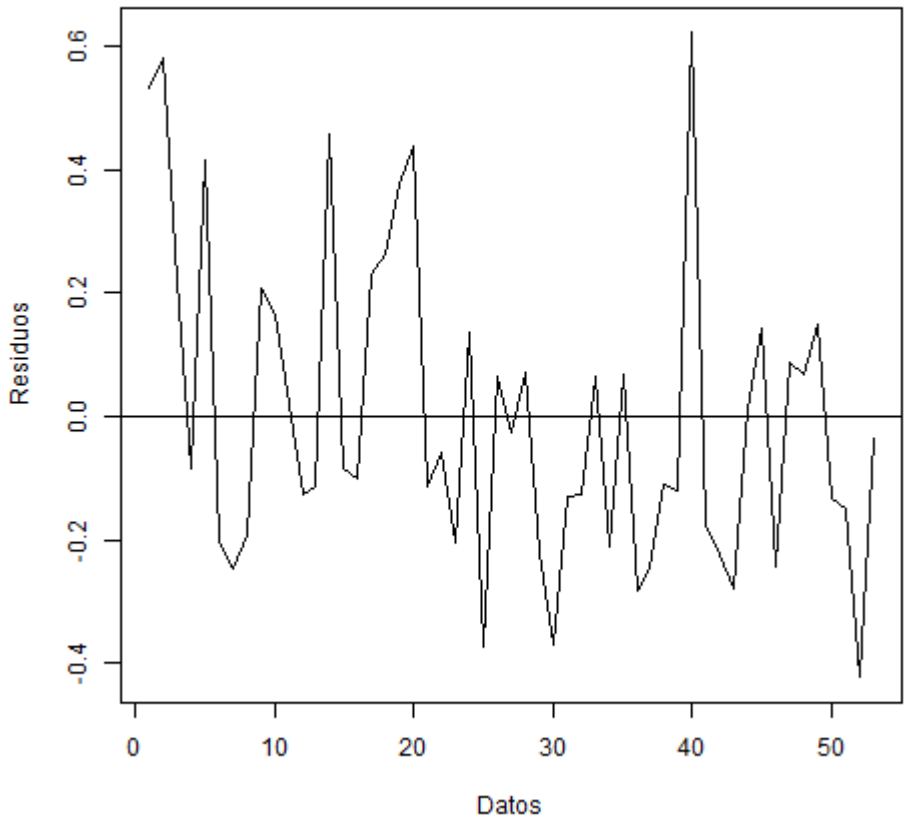
Conclusión: p > 0.05, por lo que media de los residuos es 0. (No se rechaza hipotesis nula)

Homocedasticidad



Los puntos parecen tener un ligero patrón de inclinación hacia la derecho, por lo que puede que no tengan homocedasticidad.

Independencia



Los datos no parecen seguir algún patrón aparente y ser aleatorios, por lo que se podría decir que son independientes.

Conclusiones

Los son normales e independencia. Sin embargo, pareciera que no hay homocedasticidad. Por lo tanto, es posible que el modelo no sea el mejor para predecir las concentraciones de mercurio y sería necesario hacer más pruebas.