```
Mounted at /content/drive
Mounted at /content/gdrive
/content
/content/gdrive/MyDrive/Colab Notebooks/SEMESTRE_7/Modulo1/Mercurio
EvidenciaEstadistica_2.ipynb  mercurio.csv
```

Reporte final de "Los peces y el mercurio"

Inteligencia Artificial Avanzada para la ciencia de datos (grupo 101)

Diego Solis Higuera

A00827847

18 de septiembre de 2022

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

X_1 = número de indentificación
X_2 = nombre del lago
X_3 = alcalinidad (mg/l de carbonato de calcio)
X_4 = PH
X_5 = calcio (mg/l)
X_6 = clorofila (mg/l)
X_7 = concentración media de mercurio (parte por millón) en el tejido muscualar del grupo de peces estudiados en cada lago
X_8 = número de peces estudiados en el lago
X_9 = mínimo de la concentración de mercurio en cada grupo de peces
X_{10} = máximo de la concentración de mercurio en cada grupo de peces
X_{11} = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
X_{12} = indicador de la edad de los peces (0: jóvenes; 1: maduros)

```
(53, 12)
```

Cálculo de medias estadísticas

Variables cuantitativas

	X3	X4	X5	X6	X7	X8	X9	X10	
count	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	53.000000	53
mean	37.530189	6.590566	22.201887	23.116981	0.527170	13.056604	0.279811	0.874528	
std	38.203527	1.288449	24.932574	30.816321	0.341036	8.560677	0.226406	0.522047	
min	1.200000	3.600000	1.100000	0.700000	0.040000	4.000000	0.040000	0.060000	
25%	6.600000	5.800000	3.300000	4.600000	0.270000	10.000000	0.090000	0.480000	
50%	19.600000	6.800000	12.600000	12.800000	0.480000	12.000000	0.250000	0.840000	
75%	66.500000	7.400000	35.600000	24.700000	0.770000	12.000000	0.330000	1.330000	
max	128.000000	9.100000	90.700000	152.400000	1.330000	44.000000	0.920000	2.040000	

Varianza de variables cuantitativas

```
Alcalinidad                1459.5094557329464
PH                         1.6601015965166905
Calcio                     621.6332656023222
Clorofila                  949.6456676342525
Concentracion Mid Hg       0.11630529753265603
Num Peces Lago             73.28519593613936
Min Concentracion Hg       0.05125957910014513
Max Concentracion Hg       0.27253294629898406
Concentracion Hg 3 Years   0.1147375907111756
```

Moda de variables cuantitativas

```
Moda
Repeticiones   Valor
2              17.3
4              5.8
2              3
3              1.6
4              0.34
20             12
6              0.04
2              0.06
4              0.16
43             1
```

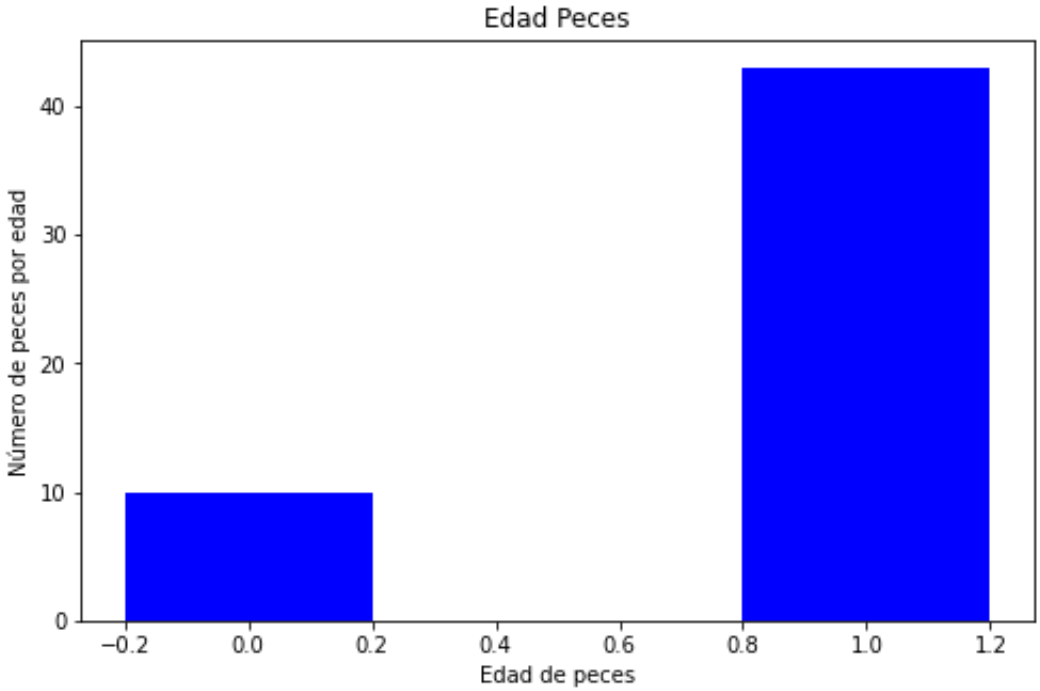
Variables cualitativas

Moda de variables cualitativas

	Nombre Lago
0	Alligator
1	Annie
2	Apopka
3	Blue Cypress
4	Brick
5	Bryant
6	Cherry
7	Crescent
8	Deer Point
9	Dias
10	Dorr
11	Down
12	East Tohopekaliga
13	Eaton
14	Farm-13
15	George
16	Griffin
17	Harney
18	Hart
19	Hatchineha
20	Iamonia
21	Istokpoga
22	Jackson
23	Josephine
24	Kingsley
25	Kissimmee
26	Lochloosa
27	Louisa
28	Miccasukee
29	Minneola
30	Monroe
31	Newmans
32	Ocean Pond
33	Ocheese Pond
34	Okeechobee
35	Orange
36	Panasoffkee
37	Parker
38	Placid
39	Puzzle
40	Rodman
41	Rousseau
42	Sampson
43	Shipp
44	Talquin
45	Tarpon
46	Tohopekaliga
47	Trafford
48	Trout
49	Tsala Apopka
50	Weir
51	Wildcat

```
52                  Yale
dtype: object
Edad Peces

0    1
dtype: int64
```

Tablas de distribución de frecuencias




```
~~~ Nombre Lago ~~~
Alligator            1
Louisa               1
Minneola             1
Monroe               1
Newmans              1
Ocean Pond           1
Ocheese Pond         1
Okeechobee           1
Orange               1
Panasoffkee          1
Parker               1
Placid               1
Puzzle               1
Rodman               1
Rousseau             1
Sampson              1
Shipp                1
Talquin              1
Tarpon               1
Tohopekaliga         1
Trafford             1
Trout                1
Tsala Apopka         1
Weir                 1
```

```
Wildcat              1
Miccasukee           1
Lochloosa            1
Annie                1
Kissimmee            1
Apopka               1
Blue Cypress         1
Brick                1
Bryant               1
Cherry               1
Crescent             1
Deer Point           1
Dias                 1
Dorr                 1
Down                 1
Eaton                1
East Tohopekaliga    1
Farm-13              1
George               1
Griffin              1
Harney               1
Hart                 1
Hatchineha           1
Iamonia              1
Istokpoga            1
Jackson              1
Josephine            1
Kingsley             1
Yale                 1
Name: X2, dtype: int64


~~~ Edad Peces ~~~
1    43
0    10
Name: X12, dtype: int64
```

Valores nulos

Al dentificar cantidad de valores nulos en dataset observamos que no hay valores nulos.

```
X1      0
X2      0
X3      0
X4      0
X5      0
X6      0
X7      0
X8      0
X9      0
X10     0
X11     0
X12     0
dtype: int64
```

Identificar los nombres de los lagos (Valores únicos)

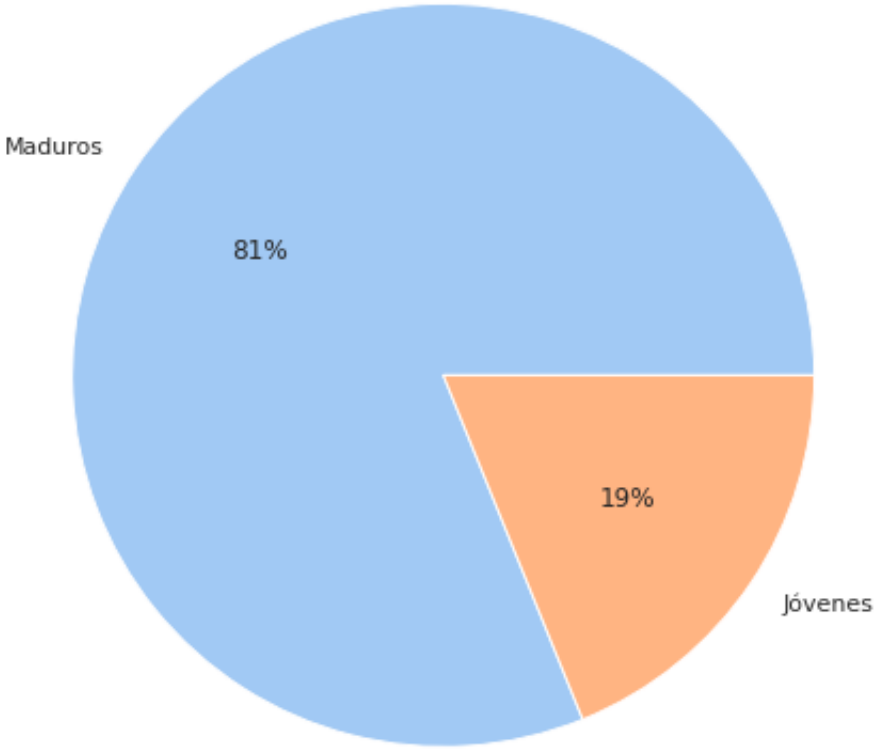
Podemos observar que cada fila corresponde a 1 lago, no hay datos repetidos para cada lago

```
53
```

Explora los datos usando herramientas de visualización

Variables categóricas

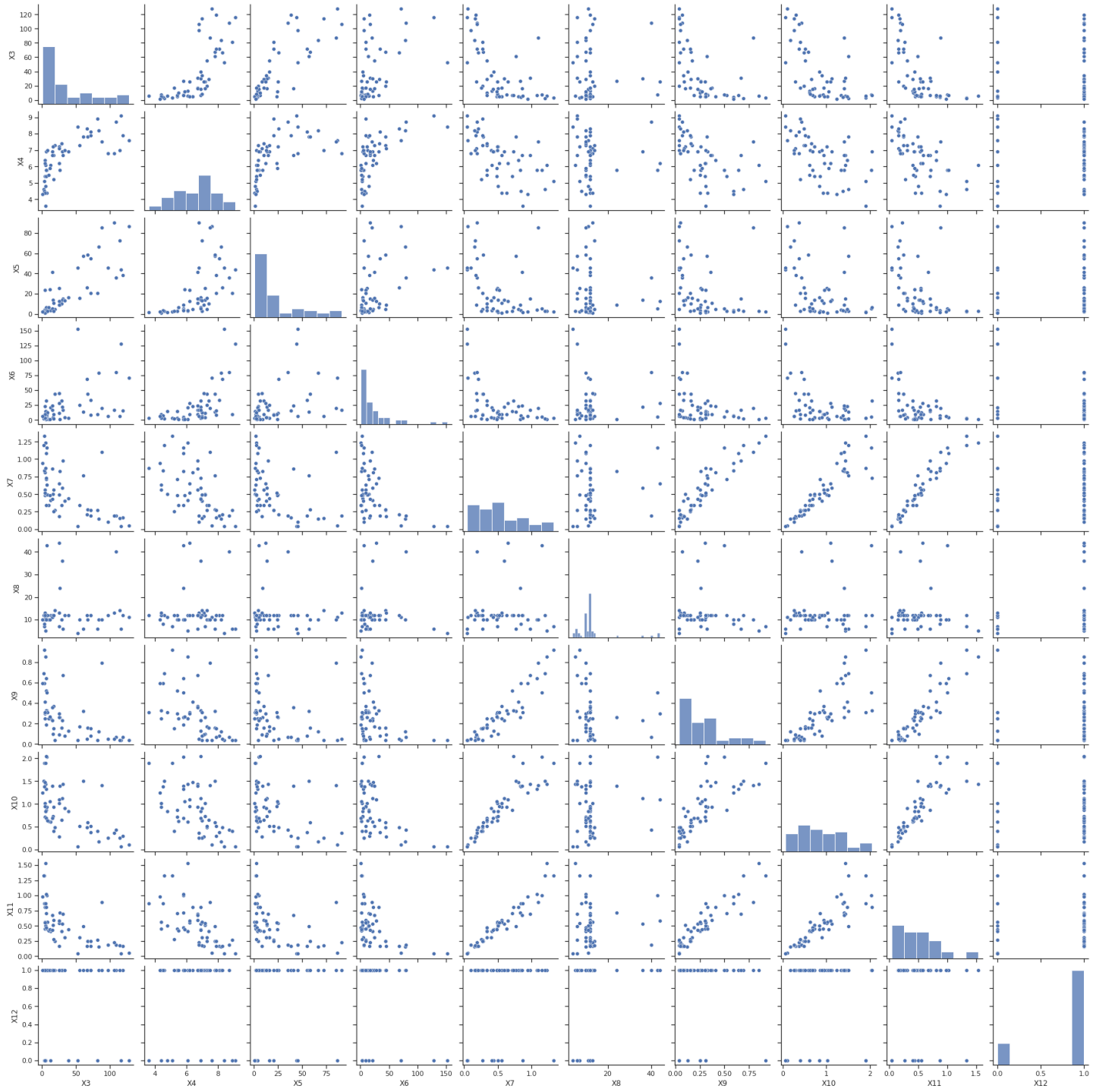
```
([<matplotlib.patches.Wedge at 0x7fa831dda2d0>,
  <matplotlib.patches.Wedge at 0x7fa831ddaa90>],
 [Text(-0.9123463061021291, 0.614511365022487, 'Maduros'),
  Text(0.9123463636368311, -0.6145112796024589, 'Jóvenes')],
 [Text(-0.49764343969207037, 0.33518801728499287, '81%'),
  Text(0.497643471074635, -0.33518797069225026, '19%')])
```



Variables Cuantitativas

Podemos observar un scatter matrix, el cual nos permite ver de manera visual cómo se comportan los datos y cómo se relacionan entre sí. Por ejemplo, vemos cómo X7 tiene una línea de tendencia con X9, X10 y X11.

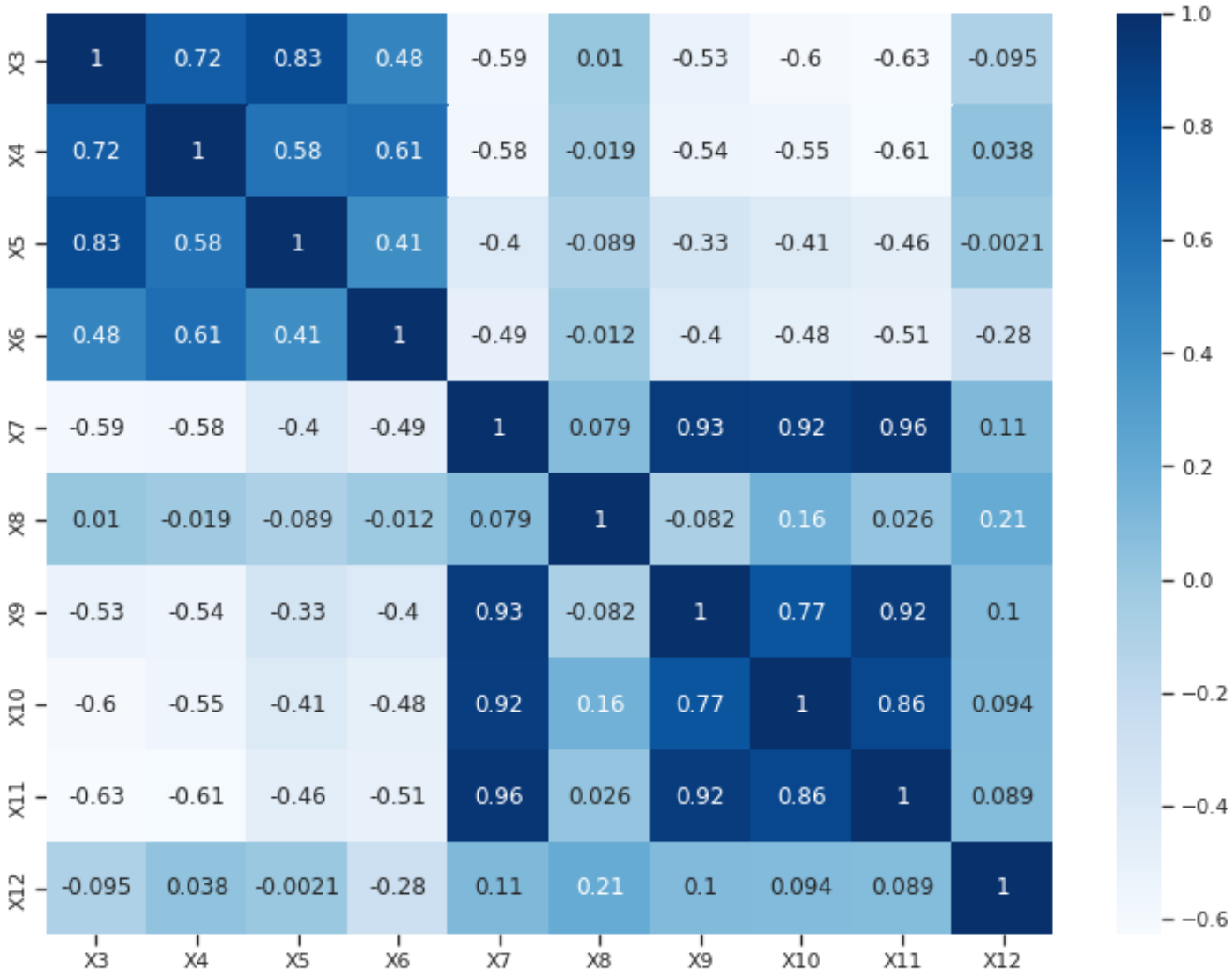
```
<Figure size 432x432 with 0 Axes>
```


X3
X4
X5
X6
X7
X8
X9
X10
X11
X12

Correlación entre variables

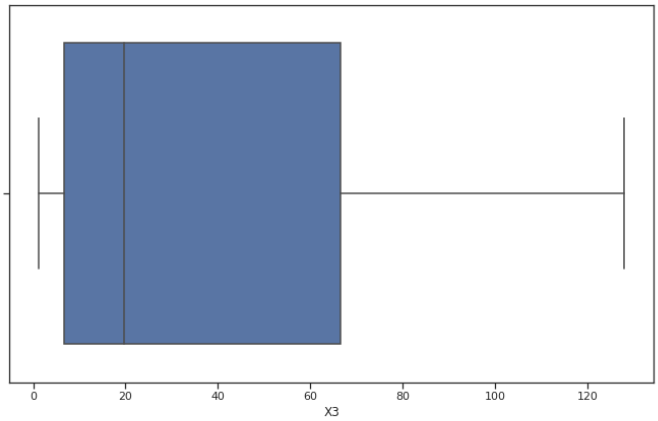
A continuación, podemos observar una matriz de correlación, que nos permite anazlizar la correlación entre variables. Esta nos permitirá encontrar las variables que mejor se ajusten para generar un modelo.

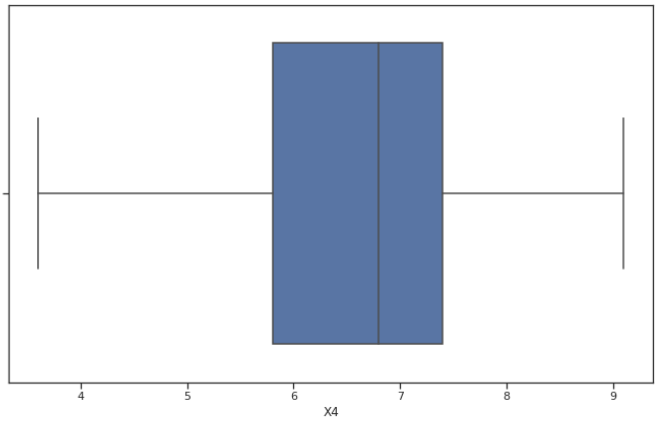
> En este caso, como mencioné anteriormente, X9, X10 y X11 son las variables que mayor correlación tienen con X7, por lo que para elegir una de las 3 y evitar colinealidad, analizamos la correlación que tiene cada una con las otras variables, y se determinó que X9 es la mejor opción, pues tiene baja correlación con las demás variables, lo que puede mejorar nuestro modelo.
>
> Por lo tanto, se eligieron X9, X12, X8, X6 y X5 como variables independendientes.

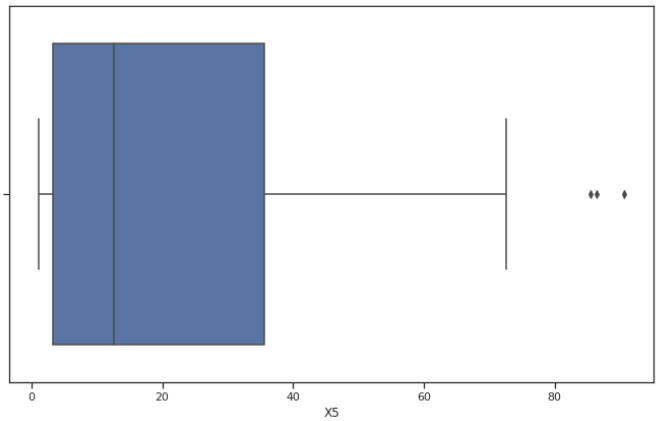


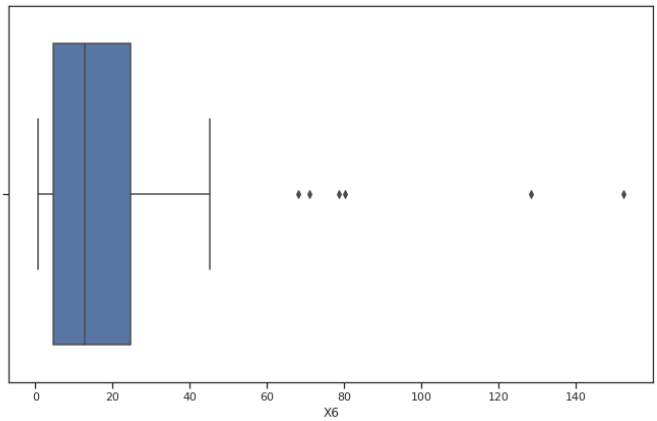
Medidas de posición

A continuación, podemos observar los boxplots de cada una de nuestras variables, donde se pueden apreciar los cuartiles, el rango intercuartílico, así como nuestros datos atípicos.









	X6	X8	X9
0	0.7	5	0.85
1	3.2	7	0.92
2	128.3	6	0.04
3	3.5	12	0.13
4	1.8	12	0.69

Creación del modelo

Modelo 1

En primera instancia, creamos un modelo con las variables mencionadas anteriormente. Analizamos los resultados, y vemos que no existe evidencia para determinar que X12 y X5 producen un efecto significativo en nuestro modelo; por lo tanto las eliminamos.

OLS Regression Results

Dep. Variable:	X7	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.893
Method:	Least Squares	F-statistic:	87.60
Date:	Mon, 19 Sep 2022	Prob (F-statistic):	1.21e-22
Time:	02:45:26	Log-Likelihood:	44.168
No. Observations:	53	AIC:	-76.34
Df Residuals:	47	BIC:	-64.51
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1585	0.054	2.917	0.005	0.049	0.268
X5	-0.0005	0.001	-0.756	0.453	-0.002	0.001
X6	-0.0015	0.001	-2.490	0.016	-0.003	-0.000
X8	0.0062	0.002	3.308	0.002	0.002	0.010
X9	1.3216	0.077	17.212	0.000	1.167	1.476
X12	-0.0438	0.042	-1.034	0.306	-0.129	0.041

Omnibus:	6.384	Durbin-Watson:	2.000
Prob(Omnibus):	0.041	Jarque-Bera (JB):	6.166
Skew:	0.835	Prob(JB):	0.0458
Kurtosis:	2.968	Cond. No.	261.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

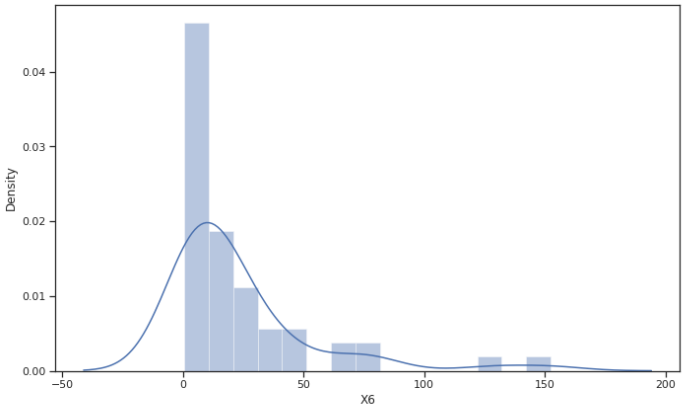
Modelo 2

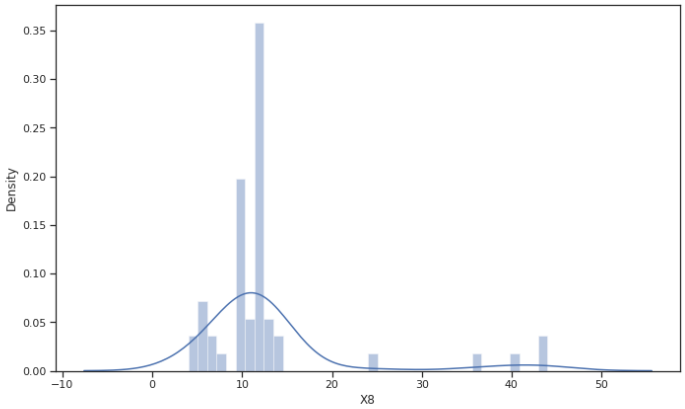
Repetimos el mismo paso, esta vez utilzando solamente 3 variables; X6, X8 y X9.

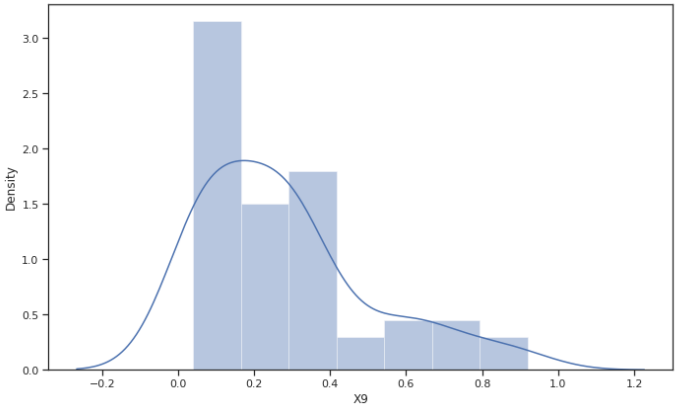
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     X7   R-squared:                       0.899
Model:                            OLS   Adj. R-squared:                  0.893
Method:                 Least Squares   F-statistic:                     145.5
Date:                Mon, 19 Sep 2022   Prob (F-statistic):           2.12e-24
Time:                        02:45:26   Log-Likelihood:                 43.097
No. Observations:                  53   AIC:                            -78.19
Df Residuals:                      49   BIC:                            -70.31
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.1107      0.042      2.657      0.011       0.027       0.194
X6            -0.0015      0.001     -2.724      0.009      -0.003      -0.000
X8             0.0060      0.002      3.285      0.002       0.002       0.010
X9             1.3336      0.075     17.798      0.000       1.183       1.484
==============================================================================
Omnibus:                        6.105   Durbin-Watson:                   1.943
Prob(Omnibus):                  0.047   Jarque-Bera (JB):                6.020
Skew:                           0.822   Prob(JB):                       0.0493
Kurtosis:                       2.844   Cond. No.                         207.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
'\nX2 = sm.add_constant(X)\nest2 = sm.OLS(y, X2).fit()\nprint(est2.summary())\n'
```

Distribución de variables

Como podemos observar, nuestras variables no tienen una distribución normal, por lo que no podemos explicar los efectos de mercurio en la saludo humana con el modelo generado. Las 3 tienen un sesgo a la derecha.







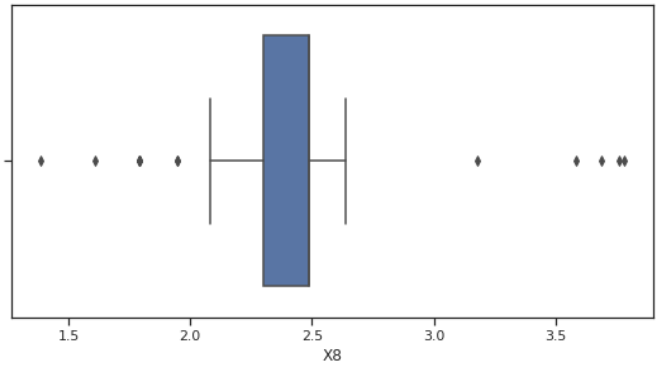
Ajuste de modelo

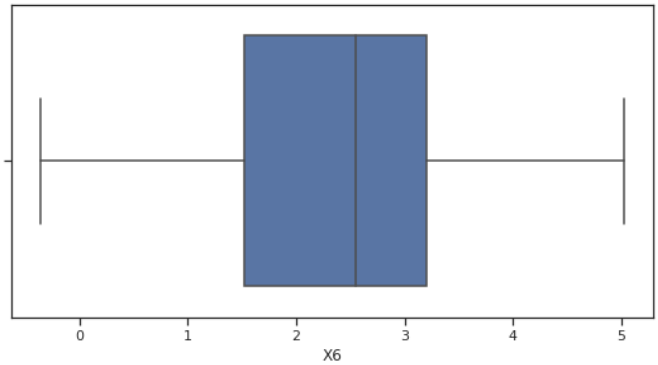
Transformación de datos

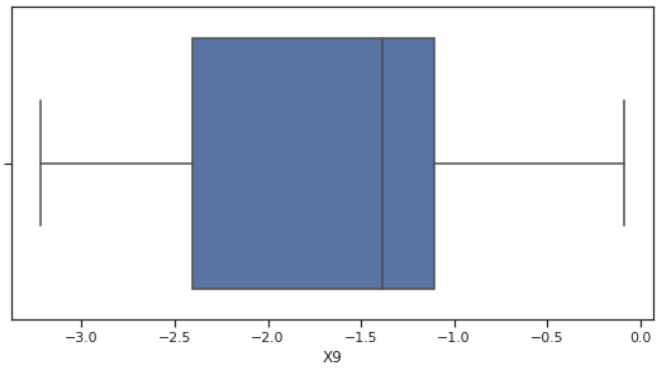
Al observar que prácticamente todas nuestras distribuciones tienen un sesgo positivo, se aplicará una transformación logarítmica para normalizar la distribución de datos. Esto nos permitirá generar un mejor modelo.

Boxplots

A continuación, podemos ver los boxplots de nuestras variables independientes, y vemos la gran cantidad de datos atípicos que contiene X8, lo que también puede afectar los resultados del modelo.

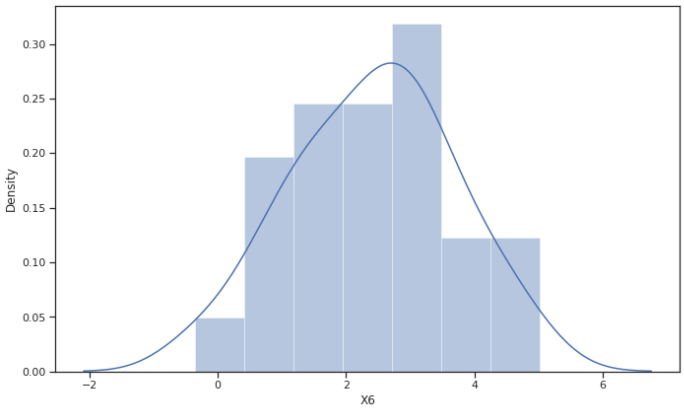


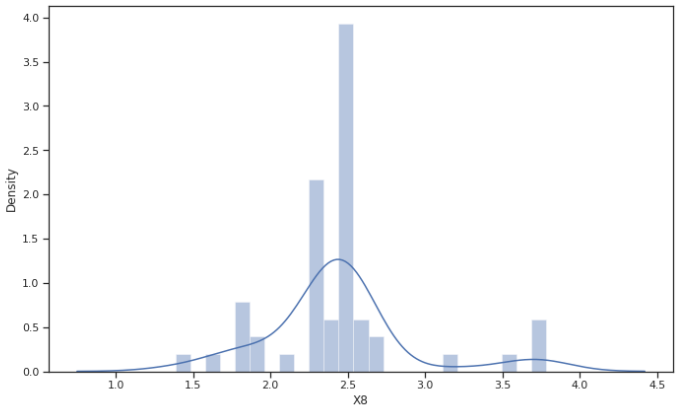


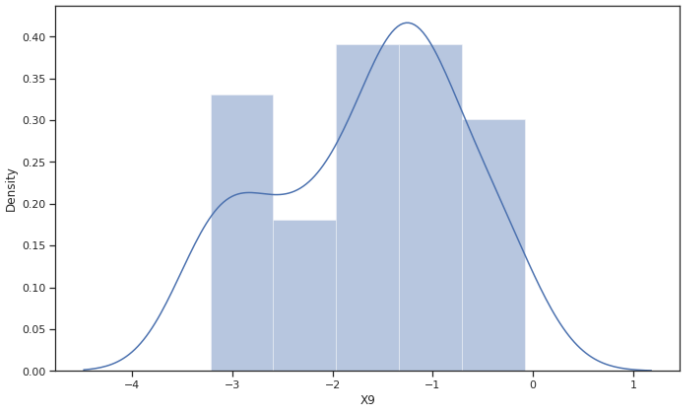


Normalización

Una vez se aplicó la función logarítmica, podemos ver cómo prácticamente desapareció el sesgo en las tres.







Eliminación de datos atípicos

A continuación, limpiamos nuestros, datos; calculamos el rango intercuartílico y quitamos todos los datos atípicos.

```
X6
X8
X9
```

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
0	1	Alligator	5.9	6.1	3.0	0.7	1.23	5	0.85	1.43	1.53	1
1	2	Annie	3.5	5.1	1.9	3.2	1.33	7	0.92	1.90	1.33	0
2	3	Apopka	116.0	9.1	44.1	128.3	0.04	6	0.04	0.06	0.04	0
3	4	Blue Cypress	39.4	6.9	16.4	3.5	0.44	12	0.13	0.84	0.44	0
4	5	Brick	2.5	4.6	2.9	1.8	1.20	12	0.69	1.50	1.33	1
5	6	Bryant	19.6	7.3	4.5	44.1	0.27	14	0.04	0.48	0.25	1
6	7	Cherry	5.2	5.4	2.8	3.4	0.48	10	0.30	0.72	0.45	1
7	8	Crescent	71.4	8.1	55.2	33.7	0.19	12	0.08	0.38	0.16	1
8	9	Deer Point	26.4	5.8	9.2	1.6	0.83	24	0.26	1.40	0.72	1
9	10	Dias	4.8	6.4	4.6	22.5	0.81	12	0.41	1.47	0.81	1
10	11	Dorr	6.6	5.4	2.7	14.9	0.71	12	0.52	0.86	0.71	1
11	12	Down	16.5	7.2	13.8	4.0	0.50	12	0.10	0.73	0.51	1
12	13	Eaton	25.4	7.2	25.2	11.6	0.49	7	0.26	1.01	0.54	1
13	14	East Tohopekaliga	7.1	5.8	5.2	5.8	1.16	43	0.50	2.03	1.00	1
14	15	Farm-13	128.0	7.6	86.5	71.1	0.05	11	0.04	0.11	0.05	0
15	16	George	83.7	8.2	66.5	78.6	0.15	10	0.12	0.18	0.15	1
16	17	Griffin	108.5	8.7	35.6	80.1	0.19	40	0.07	0.43	0.19	1
17	18	Harney	61.3	7.8	57.4	13.9	0.77	6	0.32	1.50	0.49	1
18	19	Hart	6.4	5.8	4.0	4.6	1.08	10	0.64	1.33	1.02	1
19	20	Hatchineha	31.0	6.7	15.0	17.0	0.98	6	0.67	1.44	0.70	1
20	21	Iamonia	7.5	4.4	2.0	9.6	0.63	12	0.33	0.93	0.45	1
21	22	Istokpoga	17.3	6.7	10.7	9.5	0.56	12	0.37	0.94	0.59	1
22	23	Jackson	12.6	6.1	3.7	21.0	0.41	12	0.25	0.61	0.41	0
23	24	Josephine	7.0	6.9	6.3	32.1	0.73	12	0.33	2.04	0.81	1
24	25	Kingsley	10.5	5.5	6.3	1.6	0.34	10	0.25	0.62	0.42	1
25	26	Kissimmee	30.0	6.9	13.9	21.5	0.59	36	0.23	1.12	0.53	1
26	27	Lochloosa	55.4	7.3	15.9	24.7	0.34	10	0.17	0.52	0.31	1

27	28	Louisa	3.9	4.5	3.3	7.0	0.84	8	0.59	1.38	0.87	1
28	29	Miccasukee	5.5	4.8	1.7	14.8	0.50	11	0.31	0.84	0.50	0
29	30	Minneola	6.3	5.8	3.3	0.7	0.34	10	0.19	0.69	0.47	1
30	31	Monroe	67.0	7.8	58.6	43.8	0.28	10	0.16	0.59	0.25	1
31	32	Newmans	28.8	7.4	10.2	32.7	0.34	10	0.16	0.65	0.41	1
32	33	Ocean Pond	5.8	3.6	1.6	3.2	0.87	12	0.31	1.90	0.87	0
33	34	Ocheese Pond	4.5	4.4	1.1	3.2	0.56	13	0.25	1.02	0.56	0
34	35	Okeechobee	119.1	7.9	38.4	16.1	0.17	12	0.07	0.30	0.16	1
35	36	Orange	25.4	7.1	8.8	45.2	0.18	13	0.09	0.29	0.16	1
36	37	Panasoffkee	106.5	6.8	90.7	16.5	0.19	13	0.05	0.37	0.23	1
37	38	Parker	53.0	8.4	45.6	152.4	0.04	4	0.04	0.06	0.04	0
38	39	Placid	8.5	7.0	2.5	12.8	0.49	12	0.31	0.63	0.56	1
39	40	Puzzle	87.6	7.5	85.5	20.1	1.10	10	0.79	1.41	0.89	1
40	41	Rodman	114.0	7.0	72.6	6.4	0.16	14	0.04	0.26	0.18	1
41	42	Rousseau	97.5	6.8	45.5	6.2	0.10	12	0.05	0.26	0.19	1
42	43	Sampson	11.8	5.9	24.2	1.6	0.48	10	0.27	1.05	0.44	1
43	44	Shipp	66.5	8.3	26.0	68.2	0.21	12	0.05	0.48	0.16	1
44	45	Talquin	16.0	6.7	41.2	24.1	0.86	12	0.36	1.40	0.67	1
45	46	Tarpon	5.0	6.2	23.6	9.6	0.52	12	0.31	0.95	0.55	1
46	51	Tohopekaliga	25.6	6.2	12.6	27.7	0.65	44	0.30	1.10	0.58	1
47	47	Trafford	81.5	8.9	20.5	9.6	0.27	6	0.04	0.40	0.27	0
48	48	Trout	1.2	4.3	2.1	6.4	0.94	10	0.59	1.24	0.98	1
49	49	Tsala Apopka	34.0	7.0	13.1	4.6	0.40	12	0.08	0.90	0.31	1
50	50	Weir	15.5	6.9	5.2	16.5	0.43	11	0.23	0.69	0.43	1
51	52	Wildcat	17.3	5.2	3.0	2.6	0.25	12	0.15	0.40	0.28	1
52	53	Yale	71.8	7.9	20.5	8.8	0.27	12	0.15	0.51	0.25	1

Prueba de modelo 2

Una vez hayamos realizado las modificaciones necesarias a nuestros datos para limpiarlos, volvemos a correr nuestro modelo, para ver el impacto que esto tuvo en el, y si es necesario realizar cambios.

> Como podemos observar, nuestro modelo cambió bastante, podemos ver como las variables X6 y X8 no generan un impacto significativo en nuestro modelo. Además, nuestra R cuadrada ajustada disminuyó. Esta variable explica la manera en la que nuestro modelo se ajusta a los datos.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     X7   R-squared:                       0.790
Model:                            OLS   Adj. R-squared:                  0.777
Method:                 Least Squares   F-statistic:                     61.40
Date:                Mon, 19 Sep 2022   Prob (F-statistic):           1.27e-16
Time:                        02:45:28   Log-Likelihood:                 23.660
No. Observations:                  53   AIC:                            -39.32
Df Residuals:                      49   BIC:                            -31.44
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.0314      0.126      8.161      0.000       0.777       1.285
X6            -0.0210      0.020     -1.073      0.289      -0.060       0.018
X8             0.0234      0.048      0.491      0.626      -0.072       0.119
X9             0.3110      0.027     11.588      0.000       0.257       0.365
==============================================================================
Omnibus:                        6.674   Durbin-Watson:                   1.661
Prob(Omnibus):                  0.036   Jarque-Bera (JB):                2.439
Skew:                           0.100   Prob(JB):                        0.295
Kurtosis:                       1.968   Cond. No.                         24.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
```

Modelo 3

Una vez realizado el análisis anterior, determinamos que la variable X9 es la que mejor describe nuestra variable dependiente, por lo que nuestro modelo de regresión lineal múltiple se convierte en un modelo de regresión lineal simple. Podemos ver cómo incrementó nuestro R-cuadrado ajustado con respecto al modelo anterior.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     X7   R-squared:                       0.784
Model:                            OLS   Adj. R-squared:                  0.780
Method:                 Least Squares   F-statistic:                     185.2
Date:                Mon, 19 Sep 2022   Prob (F-statistic):           1.32e-18
Time:                        02:45:28   Log-Likelihood:                 22.935
No. Observations:                  53   AIC:                            -41.87
Df Residuals:                      51   BIC:                            -37.93
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.0587      0.045     23.621      0.000       0.969       1.149
X9             0.3238      0.024     13.608      0.000       0.276       0.372
==============================================================================
Omnibus:                        7.162   Durbin-Watson:                   1.695
Prob(Omnibus):                  0.028   Jarque-Bera (JB):                2.928
Skew:                           0.260   Prob(JB):                        0.231
Kurtosis:                       1.973   Cond. No.                         4.71
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
```

Prueba de modelo 3

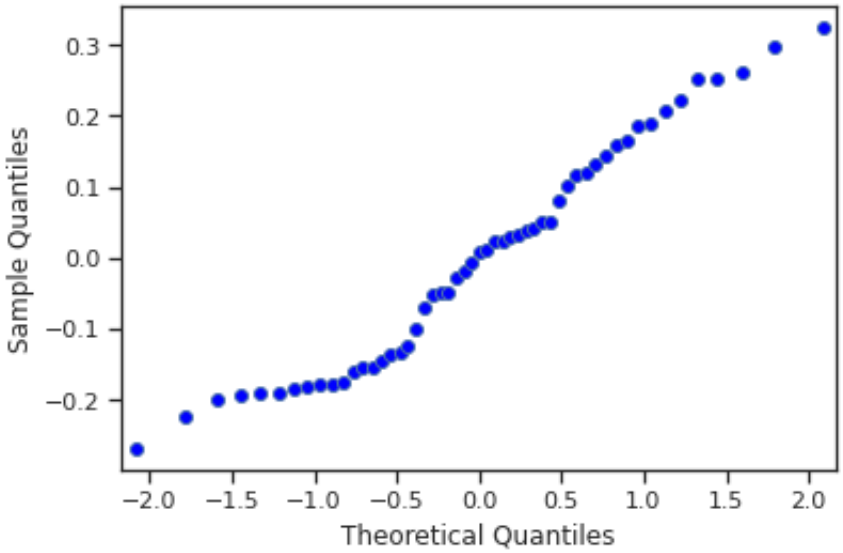
A continuación, una vez generado nuestro modelo, podemos generar las predicciones de X7, nuestra variable dependiente. Posteriormente, obtenemos los residuos.

Residuos

Interpretación

Como podemos observar, generamos una QQPlot, la cual nos permite determinar la normalidad de nuestros residuos; como podemos observar, nuestros residuos se acercan a una recta diagonal, lo que indica un supuesto de que tenemos un buen modelo que puede explicar la concentración media de mercurio. Por otro lado, realizamos la prueba de Shapiro-Wilk, la cual sugiere que nuestros residuos se comportan de manera normal, pues el valor-p obtenido es significativamente menor que nuestro alfa de 0.05.

```
<Figure size 1080x576 with 0 Axes>
```




```
ShapiroResult(statistic=0.9524779319763184, pvalue=0.03442264720797539)
```

Anova

Tenemos nuestra variable categórica X12, que es un indicador de la edad de los peces. Para responder a la siguiente pregunta:

> ¿Habrá diferencia significativa entre la concentración de mercurio por la edad de los peces?

Podemos concluir que no existe una diferencia significativa entre ambos grupos de edad de peces, pues tenemos un valor-p que supera 0.05.

	sum_sq	df	F	PR(>F)
X12	0.071511	1.0	0.610248	0.438306
Residual	5.976364	51.0	NaN	NaN

Conclusiones

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

Una vez finalizado el análisis anterior, podemos concluir que el principal factor que influye en el nivel de contaminación es el mínimo de la concentración de mercurio en cada grupo de peces. Esto también nos lleva a responder una pregunta paralela, donde podemos concluir que las concentraciones de alcalinidad, clorofila y calcio en el agua no influyen de manera significativa en los niveles de contaminación por mercurio. Además, pudimos observar cómo la interacción con las distintas variables no tenía un efecto significativo en el efecto final que buscábamos. A pesar de que algunas variables parecieran redundantes, pudimos descartarlas de nuestros modelos por su baja correlación o si baja significancia.