I was working in project Indiana Clean Lakes monitoring program from 1988-2010. Below are my 3 research questions

- How does nutrient concentration (nitrate, ammonia, phosphorus) affect water clarity in Indiana lakes?
- What factors influence algal growth (as measured by chlorophyll-a) in Indiana lakes?
- What is the suitability of lake water in Indiana for treatment into potable drinking water, and how can it be efficiently supplied to households?

Research Question -1: How does nutrient concentration (nitrate, ammonia, phosphorus) affect water clarity in Indiana lakes?

1. Hypothesis –
   Null Hypothesis (H0): There is no significant relationship between nutrient concentrations (nitrate, ammonia, phosphorus) and water clarity (Secchi Depth).

   Alternative Hypothesis (H1): There is a significant relationship between nutrient concentrations (nitrate, ammonia, phosphorus) and water clarity.

   Statistical Test: Multiple linear regression

   $Secchi = \beta 0 + \beta 1 x NO3\_epi + \beta 2 x NH3\_epi + \beta 03 x Total\_Phos\_epi + \epsilon$

   If the p-values for any of the coefficients ($\beta 1$, $\beta 2$, or $\beta 3$) are < 0.05, the null hypothesis is rejected for that variable.

For calculating coefficients below is the code

```
data = pd.read_csv('/Users/abhijitghosh/Documents/DataScience/IN_chemistry.csv')
model1 = smf.ols('Secchi ~ NO3_epi + NH3_epi + Total_Phos_epi', data=data).fit()

# Summary of the model
print(model1.summary())
```

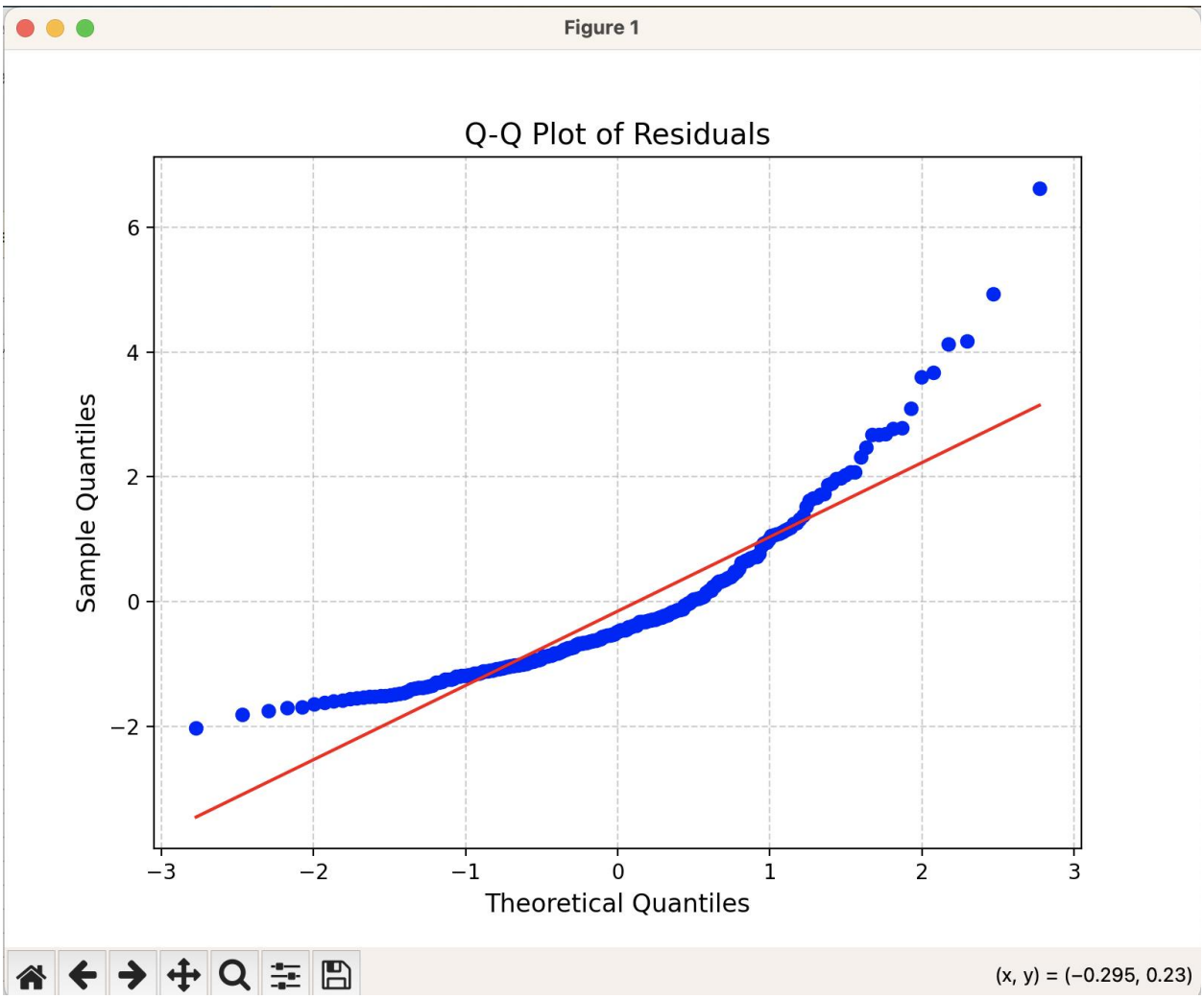
Output for coefficient/p value from summary table

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.3268	0.044	52.401	0.000	2.240	2.414
NO3_epi	−0.1530	0.027	−5.573	0.000	−0.207	−0.099
NH3_epi	−0.1633	0.198	−0.823	0.410	−0.552	0.226
Total_Phos_epi	−2.6125	0.331	−7.904	0.000	−3.261	−1.964

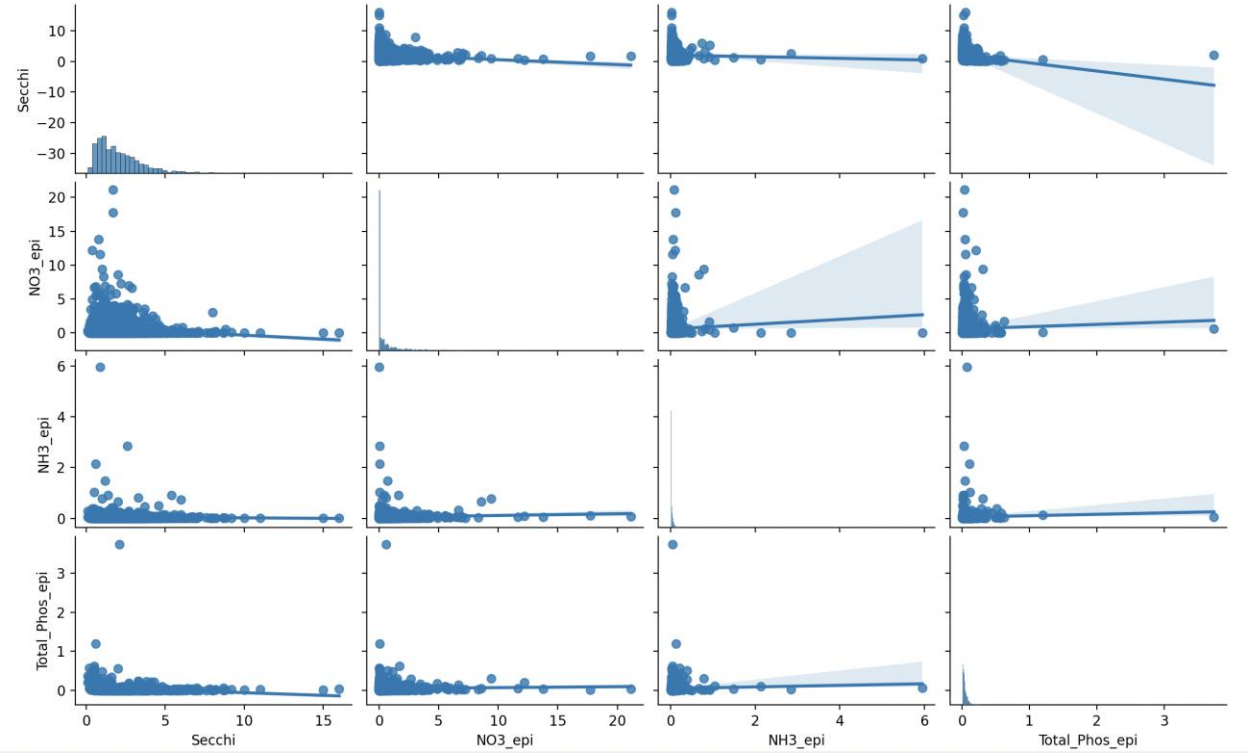
Decision: Reject the null hypothesis since all p-values are < 0.05. Nutrient concentrations significantly reduce water clarity.

Assumption:

- There is normality in residuals. From q-q plot we can verify it. Below is the plot.
- Plot residuals vs y^. Should not have fanning out or funneling in
- Plot residuals vs y^. Residuals shouldn't be uniformly above 0 or uniformly below 0 for any subsection.
- Plot residuals vs y^/ Shouldn't see any clear patterns.



Visualization: I'm trying to draw a pairplot.



Research Question -2 : What factors influence algal growth (as measured by chlorophyll-a) in Indiana lakes?

Null Hypothesis (H0): Nutrient levels (phosphorus, nitrogen) and water clarity (Secchi depth) have no significant impact on algal growth (Chlorophyll_a).
Alternative Hypothesis (H1): Nutrient levels and water clarity significantly influence algal growth.

Dependent Variable: Chlorophyll_a (indicator of algal growth).
Independent Variables: Total_Phos_epi, TKN_epi, Secchi.

Statistical Test: Multiple linear regression
Chlorophyl_a = $\beta 0 + \beta 1 x$Total_Phos_epi + $\beta 2 x$TKN_epi + $\beta 3 \cdot$Secchi + ϵ
If the p-values for any of the predictors ($\beta 1$, $\beta 2$, or $\beta 3$) are < 0.05, the null hypothesis is rejected for that variable.

For calculating coefficients below is the code

```
# Linear regression: Chlorophyll_a ~ Total_Phos_epi + TKN_epi + Secchi
model2 = smf.ols('Chlorophyll_a ~ Total_Phos_epi + TKN_epi + Secchi', data=data).fit()
```

```
# Summary of the model
print(model2.summary())
```

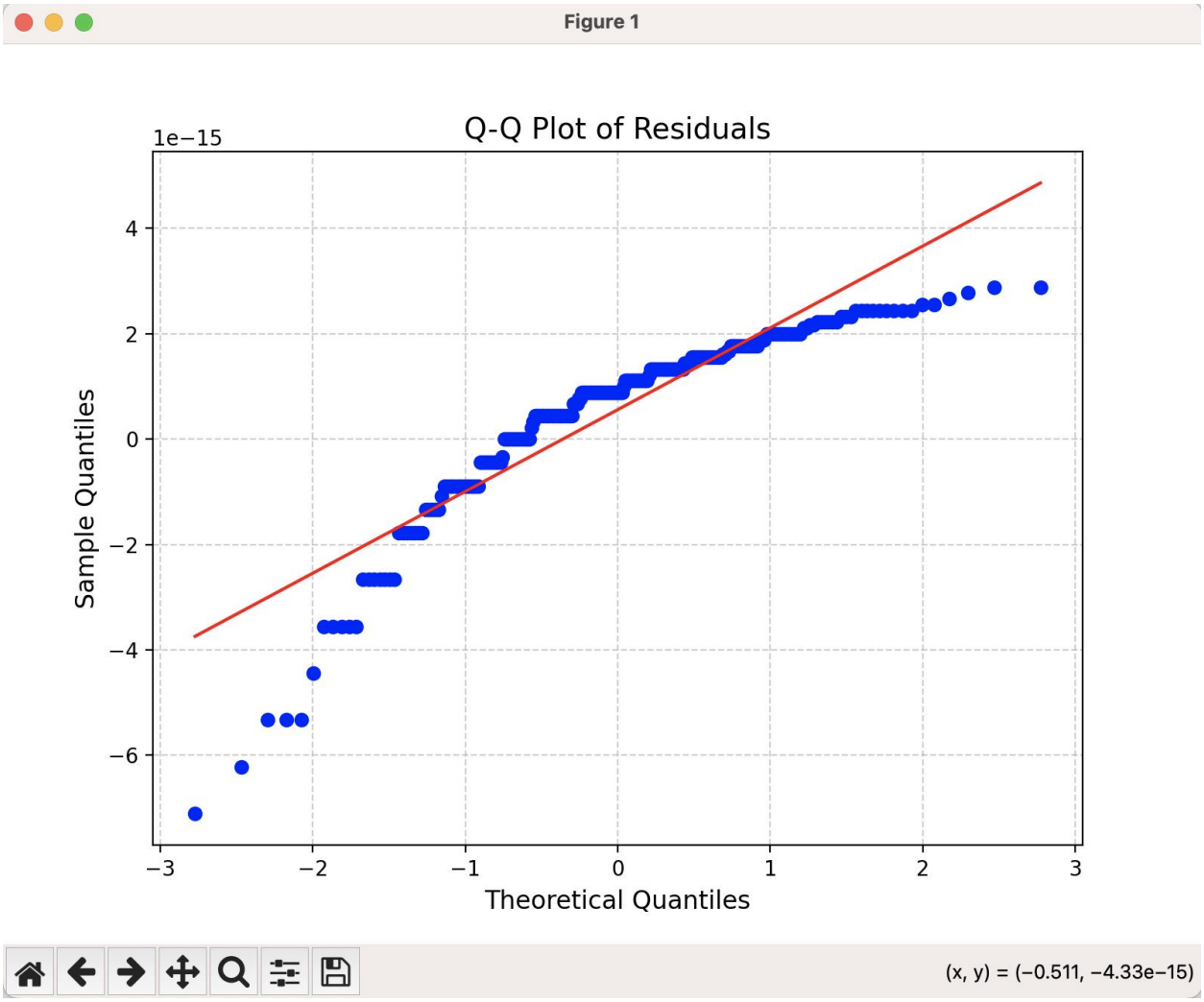
Summary output

```
                            OLS Regression Results                            
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.7937      1.355      5.013      0.000       4.135       9.452
Total_Phos_epi    89.7723      8.616     10.420      0.000      72.870     106.674
TKN_epi        8.9230      0.909      9.820      0.000       7.140      10.706
Secchi        -3.0093      0.356     -8.454      0.000      -3.708      -2.311
==============================================================================
Omnibus:                     1730.900   Durbin-Watson:                   1.823
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           601385.598
Skew:                           6.594   Prob(JB):                         0.00
Kurtosis:                     106.099   Cond. No.                         47.0
==============================================================================
```

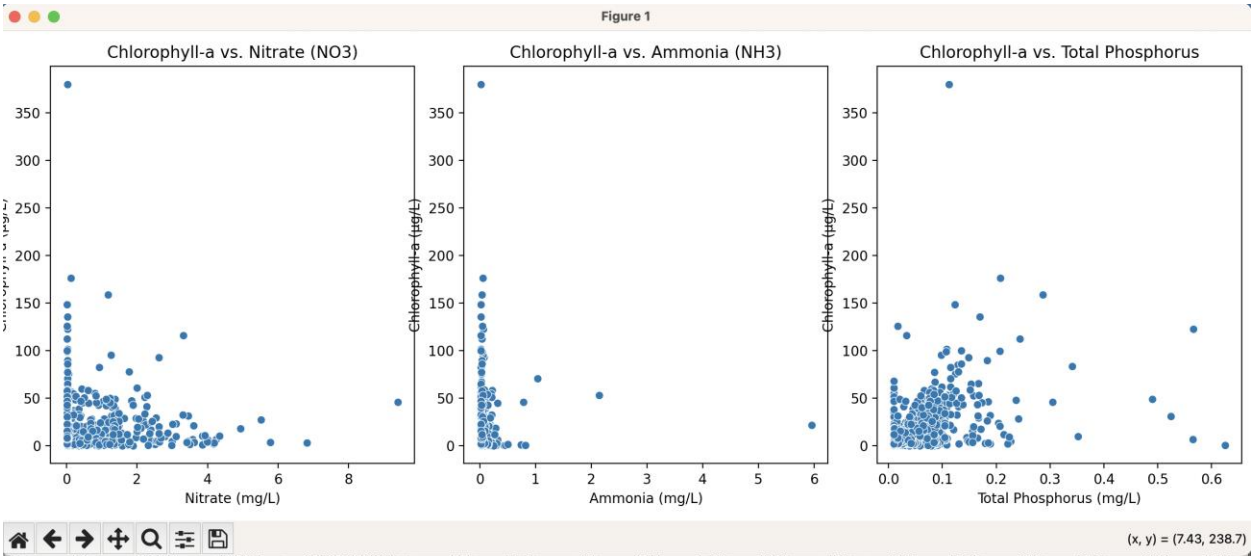
Decision: Reject the null hypothesis since all predictors have p-values < 0.05. Nutrient levels and water clarity significantly influence algal growth, with higher phosphorus and nitrogen increasing algae, and clearer water reducing it.

Assumption:

- There is normality in residuals. From q-q plot we can verify it (Shown below).
- Plot residuals vs y^. Should not have fanning out or funneling in
- Plot residuals vs y^. Residuals shouldn't be uniformly above 0 or uniformly below 0 for any subsection.
- Plot residuals vs y^/ Shouldn't see any clear patterns.



Visualization: Scatter plot



Research Question - 3: What is the suitability of lake water in Indiana for treatment into potable drinking water, and how can it be efficiently supplied to households?

Null Hypothesis (H0): The average concentrations of nitrate, ammonia, and water clarity (Secchi depth) in Indiana lakes meet the suitability thresholds:

- NO3_epi ≤ 10 mg/L
- NH3_epi ≤ 0.5 mg/L
- Secchi depth ≥ 1.5 m

Alternative Hypothesis (H1): The average concentrations of nitrate, ammonia, or water clarity do not meet the suitability thresholds.

Statistical Test:

One-sample t-tests to compare the mean of each parameter against its threshold (this threshold value I got from internet):

NO3_epi against 10 mg/L.
NH3_epi against 0.5 mg/L.
Secchi depth against 1.5 m.

Null hypothesis is rejected if p-value < 0.05 for any test.

We are running one sample t test because we are trying to compare sample mean and population mean. Below is the python code and output.

```
nitrate_levels = data['NO3_epi']  # Example nitrate data (mg/L)
ammonia_levels = data['NH3_epi']    # Example ammonia data (mg/L)
secchi_depths = data['Secchi']     # Example Secchi depth data (m)

# Thresholds for suitability
nitrate_threshold = 10.0
ammonia_threshold = 0.5
secchi_threshold = 1.5

# One-sample t-tests
# Nitrate
nitrate_t_stat, nitrate_p_value = ttest_1samp(nitrate_levels, nitrate_threshold)
print("Nitrate:")
print(f"T-Statistic: {nitrate_t_stat}, P-Value: {nitrate_p_value}")
if nitrate_p_value < 0.05:
    print("Reject the null hypothesis: Nitrate levels exceed the threshold.")
else:
    print("Fail to reject the null hypothesis: Nitrate levels meet the threshold.")

# Ammonia
```

```
ammonia_t_stat, ammonia_p_value = ttest_1samp(ammonia_levels, ammonia_threshold)
print("\nAmmonia:")
print(f"T-Statistic: {ammonia_t_stat}, P-Value: {ammonia_p_value}")
if ammonia_p_value < 0.05:
    print("Reject the null hypothesis: Ammonia levels exceed the threshold.")
else:
    print("Fail to reject the null hypothesis: Ammonia levels meet the threshold.")

# Secchi Depth
secchi_t_stat, secchi_p_value = ttest_1samp(secchi_depths, secchi_threshold)
print("\nSecchi Depth:")
print(f"T-Statistic: {secchi_t_stat}, P-Value: {secchi_p_value}")
if secchi_p_value < 0.05:
    print("Reject the null hypothesis: Secchi depth is below the threshold.")
else:
    print("Fail to reject the null hypothesis: Secchi depth meets the threshold.")
```

Nitrate:
T-Statistic: -411.07875042113716, P-Value: 0.0
Reject the null hypothesis: Nitrate levels exceed the threshold.

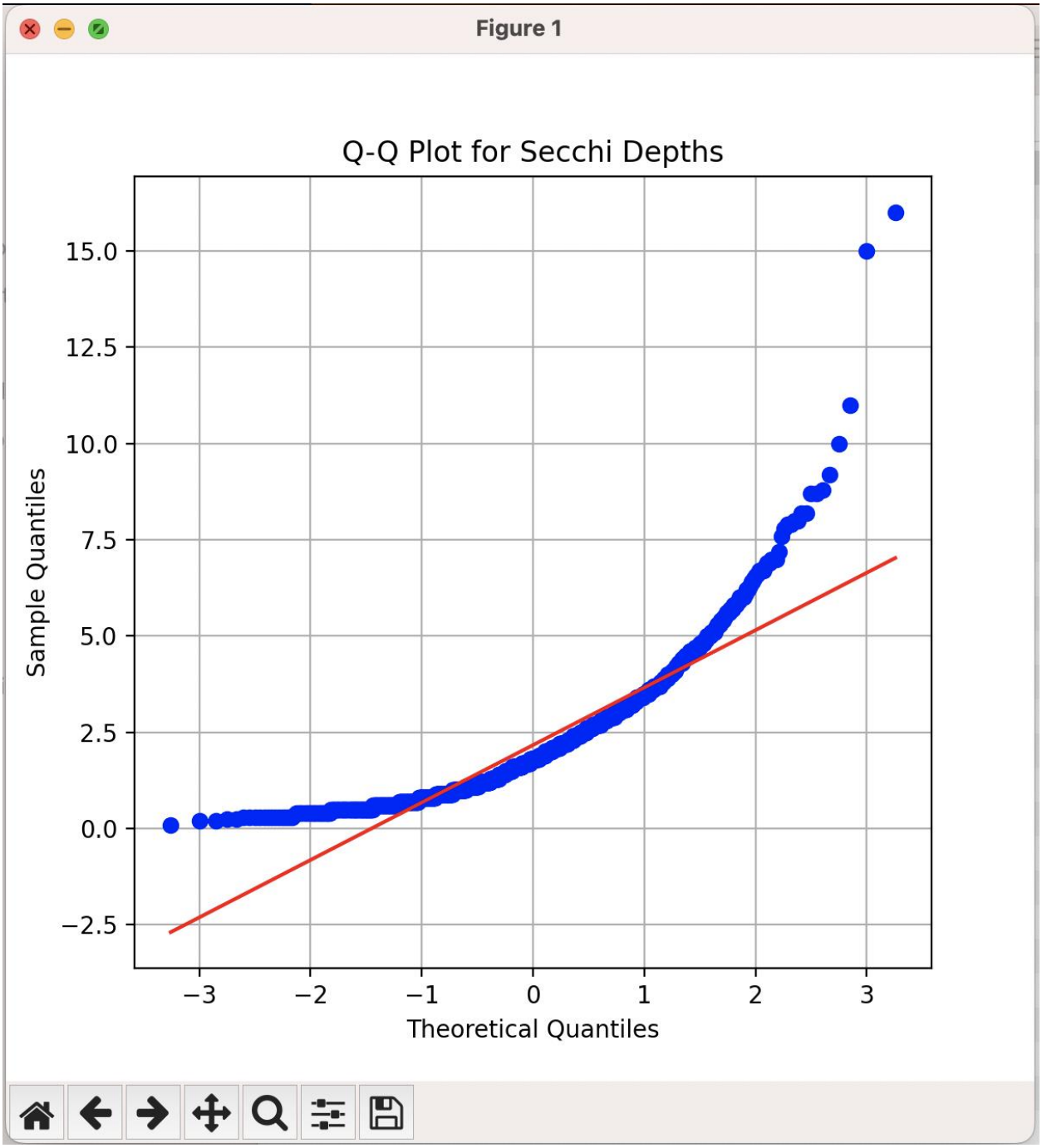
Ammonia:
T-Statistic: -83.33584956058084, P-Value: 0.0
Reject the null hypothesis: Ammonia levels exceed the threshold.

Secchi Depth:
T-Statistic: 14.384530495241984, P-Value: 1.7567647265826735e-43
Reject the null hypothesis: Secchi depth is below the threshold.

In all three cases we can reject the null hypothesis.

Assumption:

- Data are normally distributed for each variable. We can verify with q-q plot. (one example for Secchi depth is shown below)
- Observations are independent.



Final Visualization of Data:

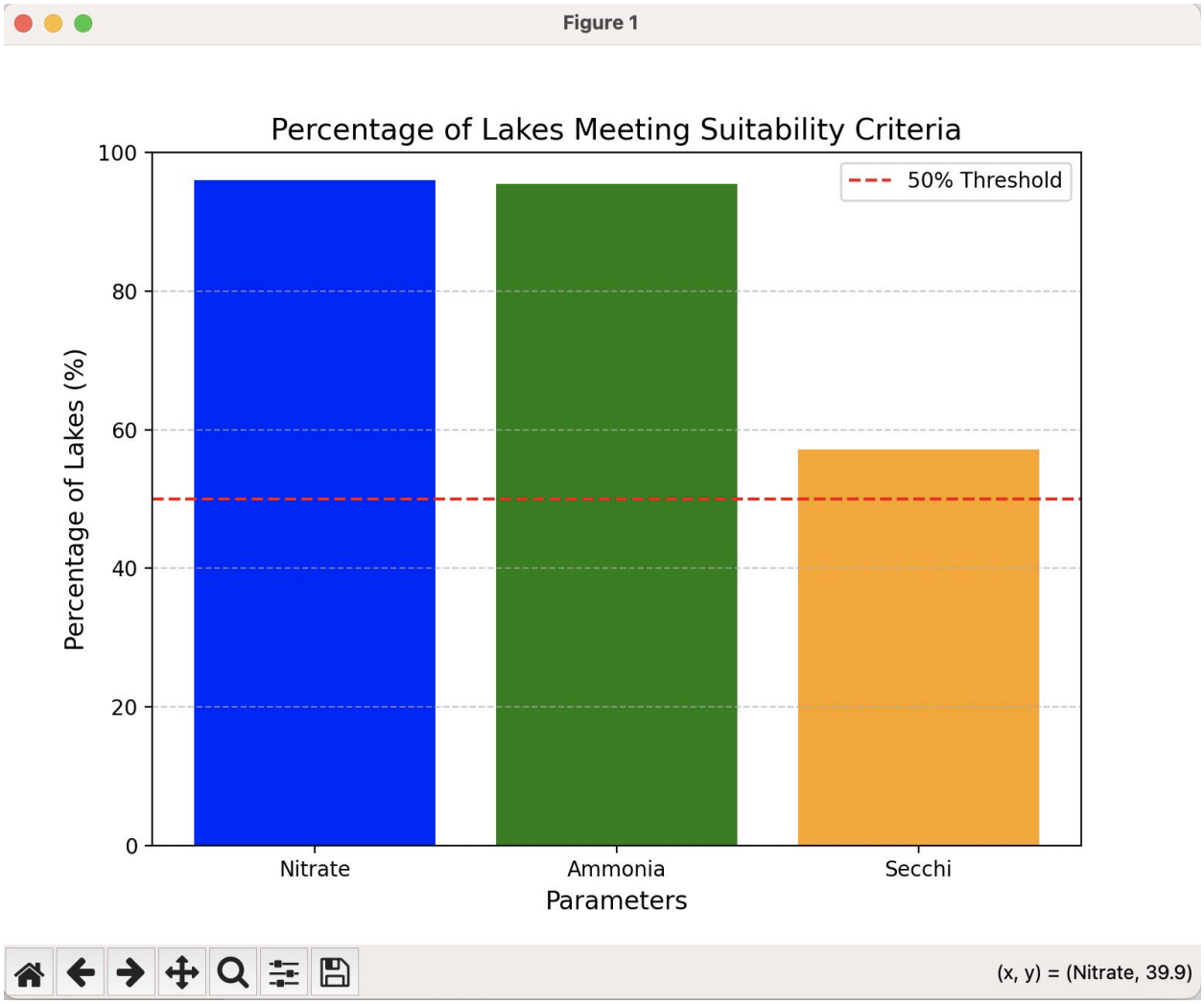


Figure 1
Percentage of Lakes Meeting Suitability Criteria
50% Threshold
Percentage of Lakes (%)
100
80
60
40
20
0
Nitrate
Ammonia
Secchi
Parameters
(x, y) = (Nitrate, 39.9)