Indiana Clean Water Monitoring Program (1988-2010)

Context of the Study

The Indiana Clean Water Monitoring Program focuses on analyzing water quality across Indiana lakes. This dataset provides comprehensive water quality metrics to evaluate lake health and suitability for various uses, including recreation, ecological preservation, and potable water treatment. Indiana's lakes are critical resources that support local ecosystems, provide recreational opportunities, and serve as potential sources of drinking water.

Dataset Overview

The dataset, sourced from the Environmental Data Initiative, includes detailed measurements from various lakes across Indiana. Each data point corresponds to a specific sample collected on a given date at a specific location in the lake. The dataset consists of the following key columns:

- **General Lake Information**: Sample_ID, Lake_ID, Lake_Name, County, Date_Sampled, Sample_Location
- **Lake Characteristics**: Max_Depth (maximum depth in meters), Surface_Area (lake size in square meters)
- **Water Quality Metrics**:
    - **Water Clarity**: Secchi (Secchi depth in meters, a measure of water transparency)
    - **Nutrient Concentrations**:
        - NO3_epi, NO3_hypo (nitrate concentration in the surface and bottom layers)
        - NH3_epi, NH3_hypo (ammonia concentration in the surface and bottom layers)
        - TKN_epi, TKN_hypo (total amount of nitrogen in surface and bottom layers)
        - SRP_epi, SRP_hypo (soluble reactive phosphorus in surface and bottom layers)
        - Total_Phos_epi, Total_Phos_hypo (total phosphorus in surface and bottom layers)
    - **Chlorophyll-a**: Proxy for algal biomass

Research Objectives

The primary objectives of this project are to investigate the water quality of Indiana lakes by addressing the following research questions:

1. **Nutrient Concentration and Water Clarity**:
    - How do nitrate (NO3), ammonia (NH3), and phosphorus concentrations impact water clarity (measured by Secchi depth)?
2. **Factors Influencing Algal Growth**:

    - Which factors, such as nutrient concentrations, lake depth, and surface area, are most correlated with algal growth (measured by chlorophyll-a)?
3. **Suitability for Potable Water Treatment**:
    - What is the suitability of Indiana lake water for treatment into potable water?
    - Metrics like nitrate, ammonia, and Secchi depth are analyzed against recommended thresholds to assess suitability.

Three 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 relationship between one of the (at least one) nutrient concentrations (nitrate, ammonia, phosphorus) and water clarity.

   Statistical Test: F test in Multiple linear regression

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

In multiple linear regression, the **F-statistic** evaluates whether the regression model explains a significant portion of the variability in the dependent variable compared to a model with no predictors.

```
data = pd.read_csv('/Users/abhijitghosh/Documents/DataScience/IN_chemistry.csv')
columns = ["Secchi", "NO3_epi", "NH3_epi", "Total_Phos_epi"]
df = data[columns].dropna()

X = df[["NO3_epi", "NH3_epi", "Total_Phos_epi"]]
y = df["Secchi"]
# Summary of the model
X = sm.add_constant(X)
```

```
# Perform multiple linear regression
model = sm.OLS(y, X).fit()

f_statistic = model.fvalue
f_pvalue = model.f_pvalue
print(f"F-Statistic: {f_statistic}")
print(f"F-Test p-value: {f_pvalue}")
# Display regression results, including F-statistic and p-value
print(model.summary())
```

Output for coefficient/p value from summary table

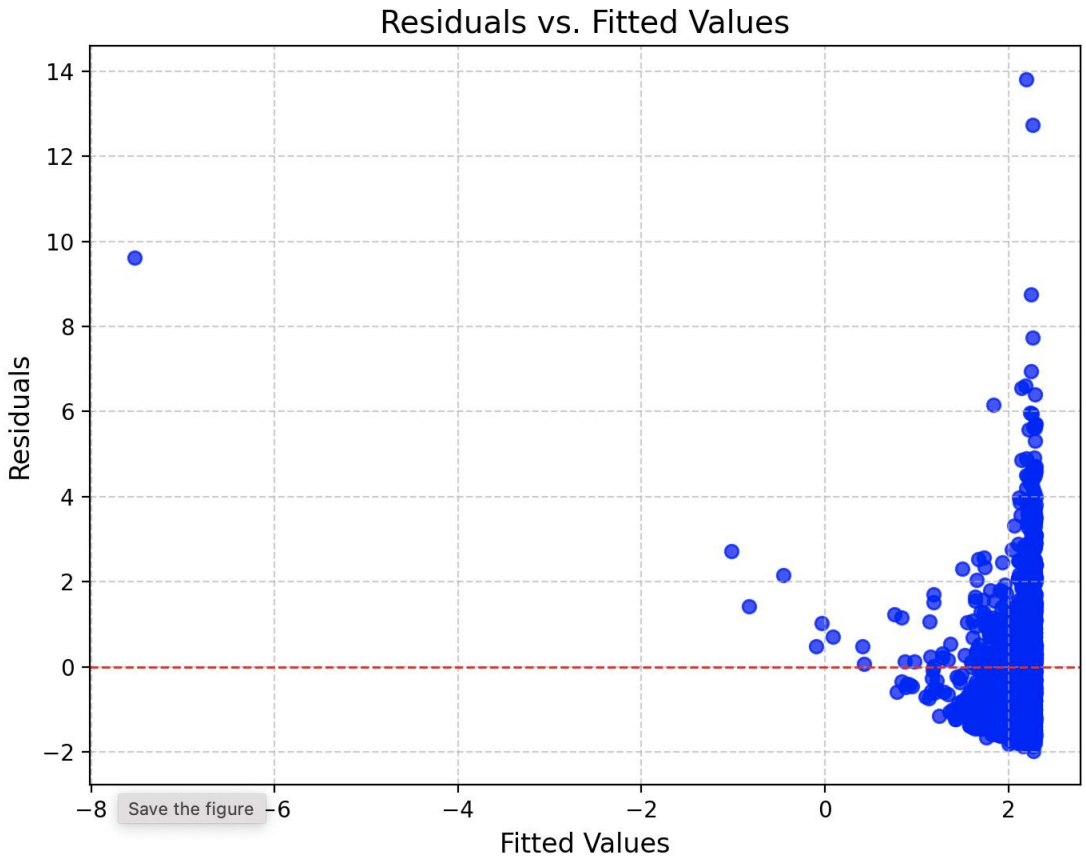
```
F-Statistic: 32.60093050465909
F-Test p-value: 1.7405924024121344e-20
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Secchi   R-squared:                       0.054
Model:                            OLS   Adj. R-squared:                  0.053
Method:                 Least Squares   F-statistic:                     32.60
Date:                Thu, 27 Feb 2025   Prob (F-statistic):           1.74e-20
Time:                        16:26:55   Log-Likelihood:                -3110.3
No. Observations:                1703   AIC:                             6229.
Df Residuals:                    1699   BIC:                             6250.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          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
==============================================================================
Omnibus:                      903.552   Durbin-Watson:                   1.875
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9190.220
Skew:                           2.278   Prob(JB):                         0.00
Kurtosis:                      13.429   Cond. No.                         14.0
==============================================================================
```

Decision: Here we can see that p value of N03_epi < .05 and Total_Phos_epi < .05 but NH3_epi > .05 Also p value for F test is less. So, we can reject the null hypothesis. That means at least one nutrient has relation with water clarity.

Plot for Residual Vs fitted value:

```
fitted_values = model.fittedvalues  # Predicted values (fitted by the model)
residuals = model.resid  # Residuals (actual - predicted)

# Plot residuals vs. fitted values
plt.figure(figsize=(8, 6))
plt.scatter(fitted_values, residuals, alpha=0.7, color='blue')
plt.axhline(y=0, color='red', linestyle='--', linewidth=1)
plt.title('Residuals vs. Fitted Values', fontsize=14)
plt.xlabel('Fitted Values', fontsize=12)
```

```
plt.ylabel('Residuals', fontsize=12)
plt.grid(True, linestyle='--', alpha=0.6)
plt.show()
```

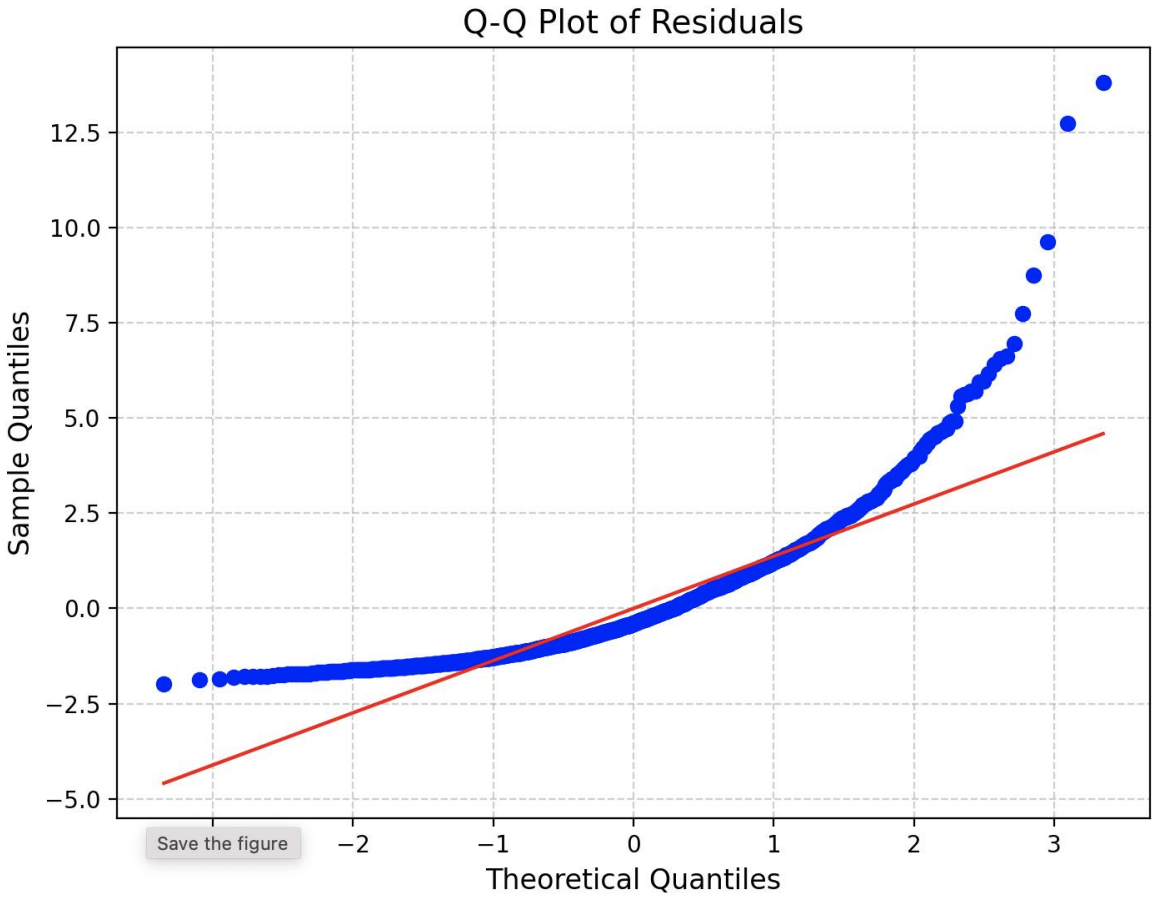


Observation

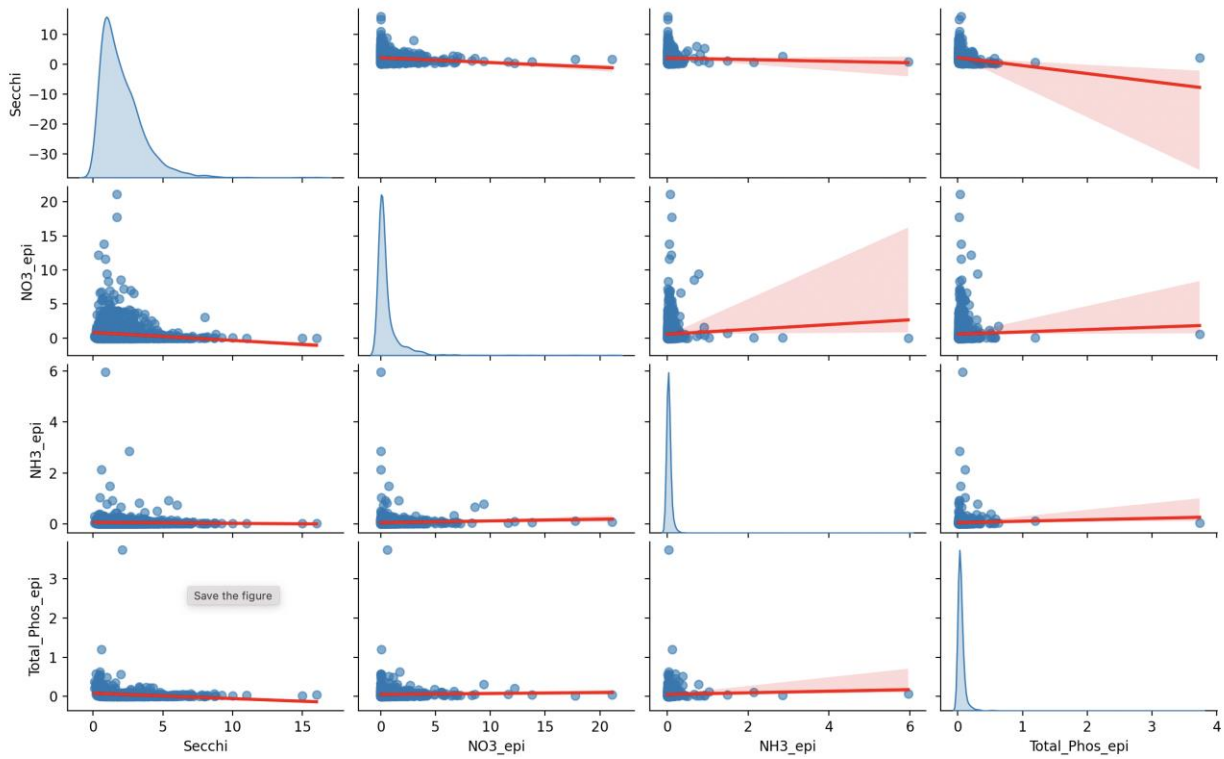
Most residuals are tightly clustered near zero for fitted values between 0 and 2. This indicates that the model predictions align reasonably well with the actual values in this range. There are a few residuals with very large values (e.g., residuals > 8 and fitted values < 0). These are likely outliers.

Assumption:

- There is normality in residuals. From q-q plot we can verify it. Below is the plot. This q-q plot looks left skewed.
- 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. The red line is the regression line with respect to data. This Plot explains relation with each field. Here we don't see any correlation between variables as per regression line.



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): At least one of the attributes (phosphrous, nitrogen, Secchi depth) has a relationship with algal growth.

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

Statistical Test: F test for 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

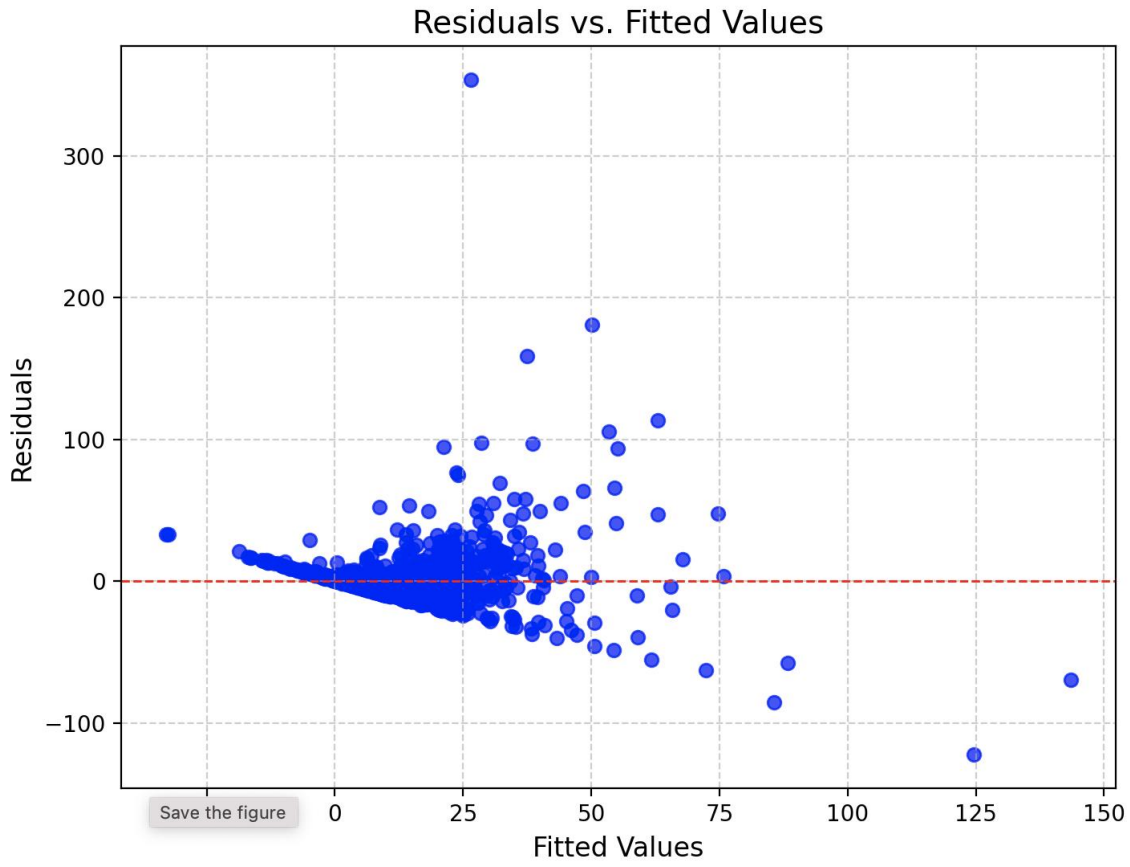
```python
columns2 = ["Chlorophyll_a", "Total_Phos_epi", "TKN_epi", "Secchi"]
df2 = data[columns2].dropna()

X2 = df2[["Total_Phos_epi", "TKN_epi", "Secchi"]]
y2 = df2["Chlorophyll_a"]
# Summary of the model
X2 = sm.add_constant(X2)

# Perform multiple linear regression
model2 = sm.OLS(y2, X2).fit()

f_statistic2 = model2.fvalue
f_pvalue2 = model2.f_pvalue
print(f"F-Statistic: {f_statistic2}")
print(f"F-Test p-value: {f_pvalue2}")
# Display regression results, including F-statistic and p-value
print(model2.summary())

fitted_values2 = model2.fittedvalues  # Predicted values (fitted by the model)
residuals2 = model2.resid  # Residuals (actual - predicted)
```

Summary output

```
F-Statistic: 212.99023710145536
F-Test p-value: 7.644236424565017e-113
                            OLS Regression Results
==============================================================================
Dep. Variable:          Chlorophyll_a   R-squared:                       0.324
Model:                            OLS   Adj. R-squared:                  0.323
Method:                 Least Squares   F-statistic:                     213.0
Date:                Thu, 27 Feb 2025   Prob (F-statistic):          7.64e-113
Time:                        17:29:34   Log-Likelihood:                -5832.6
No. Observations:                1336   AIC:                         1.167e+04
Df Residuals:                    1332   BIC:                         1.169e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          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. Also, p value for f test is very less than .05. Nutrient levels and water clarity significantly influence algal growth, with higher phosphorus and nitrogen increasing algae, and clearer water reducing it.

Plot for Residuals Vs Fitted Value

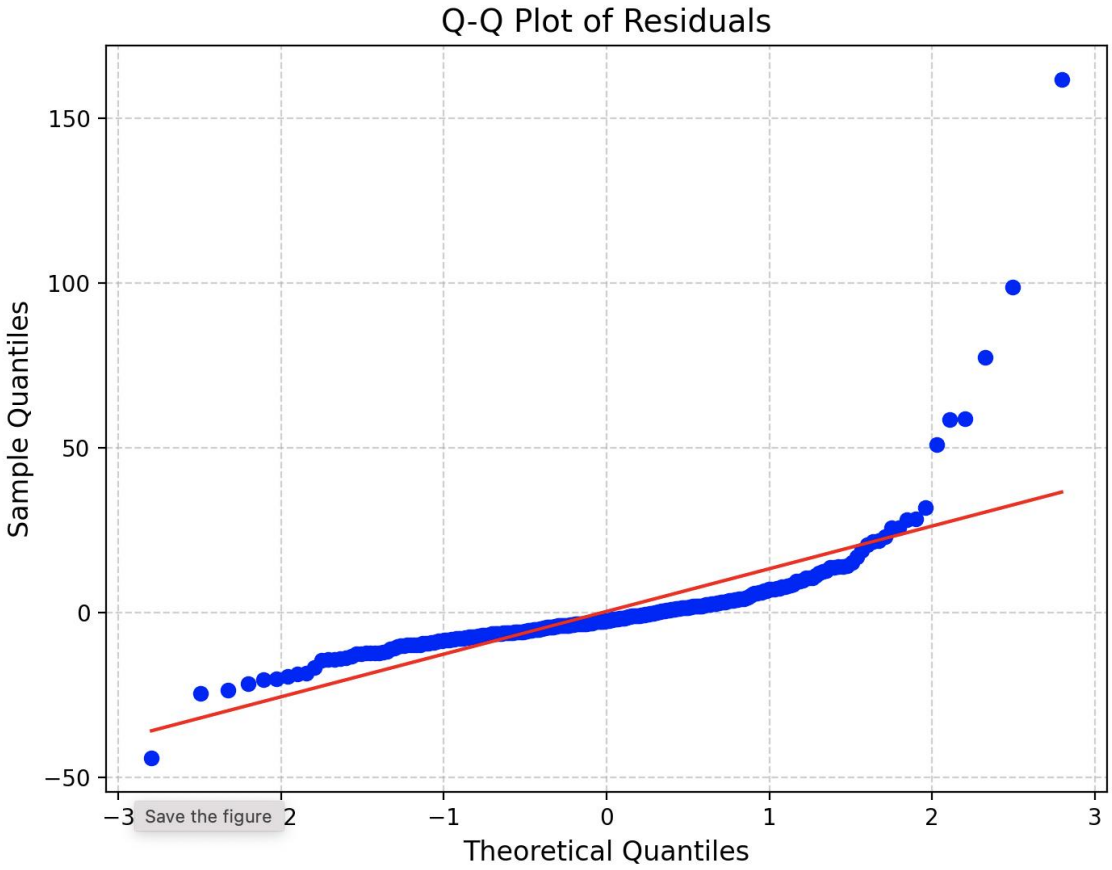


Observation

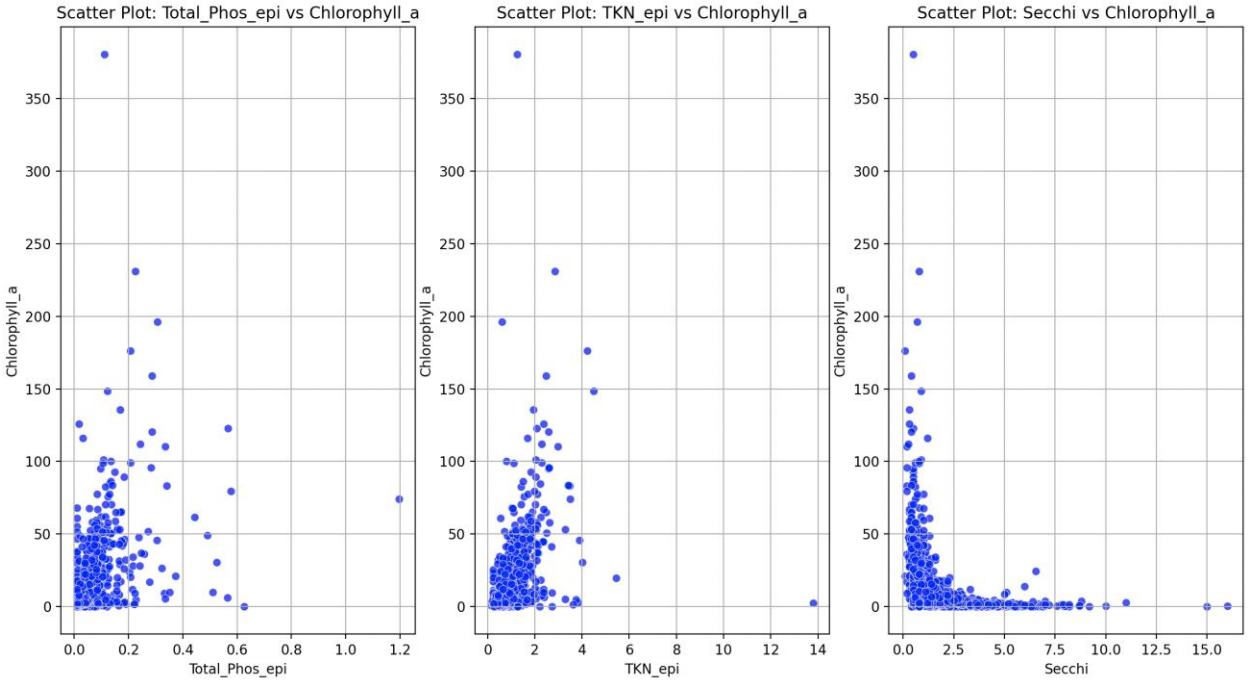
There appears to be a fan-shaped pattern in the residuals, where the spread increases as the fitted values increase. This indicates heteroscedasticity, meaning that the variance of the residuals is not constant across all levels of the fitted values. There are several points with very large residuals (both positive and negative), indicating outliers. There might be model improvement needed with transformation of the response variable by taking logarithmic values.

Assumption:

- There is normality in residuals. From q-q plot we can verify it (Shown below). The q-q plot is skewed.
- 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:

We are running one sided test for each variable. 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.

```
# Thresholds for suitability
thresholds = {
    "NO3_epi": 10.0,  # Nitrate concentration (mg/L)
    "NH3_epi": 0.5,   # Ammonia concentration (mg/L)
    "Secchi": 1.5     # Minimum Secchi depth (m)
}

# Define directions of comparison for null hypothesis
comparison = {
    "NO3_epi": "≤",  # Nitrate should be less than or equal to the threshold
    "NH3_epi": "≤",  # Ammonia should be less than or equal to the threshold
    "Secchi": "≥"    # Secchi depth should be greater than or equal to the threshold
}

# Perform one-sample t-tests (one-tailed) for the metrics
results = []
for col, threshold in thresholds.items():
    sample = df[col].dropna()
    sample_mean = sample.mean()  # Sample mean as an estimate of population mean
    t_stat, p_value = ttest_1samp(sample, threshold)

    # Adjust for one-tailed p-value based on direction of comparison
    if comparison[col] == "≤":
        one_tailed_p = p_value / 2 if t_stat < 0 else 1 - (p_value / 2)
    elif comparison[col] == "≥":
        one_tailed_p = p_value / 2 if t_stat > 0 else 1 - (p_value / 2)

    # Conclusion
    conclusion = "Fails to Reject H₀" if one_tailed_p > 0.05 else "Rejects H₀"

    results.append({
        "Metric": col,
        "Sample Mean (Estimate of Population Mean)": sample_mean,
        "Threshold": threshold,
        "T-Statistic": t_stat,
        "P-Value (One-Tailed)": one_tailed_p,
        "Conclusion": conclusion
    })

# Create a results DataFrame
results_df = pd.DataFrame(results)
print("Water Quality Analysis Results")
print(results_df)
```

Output
======

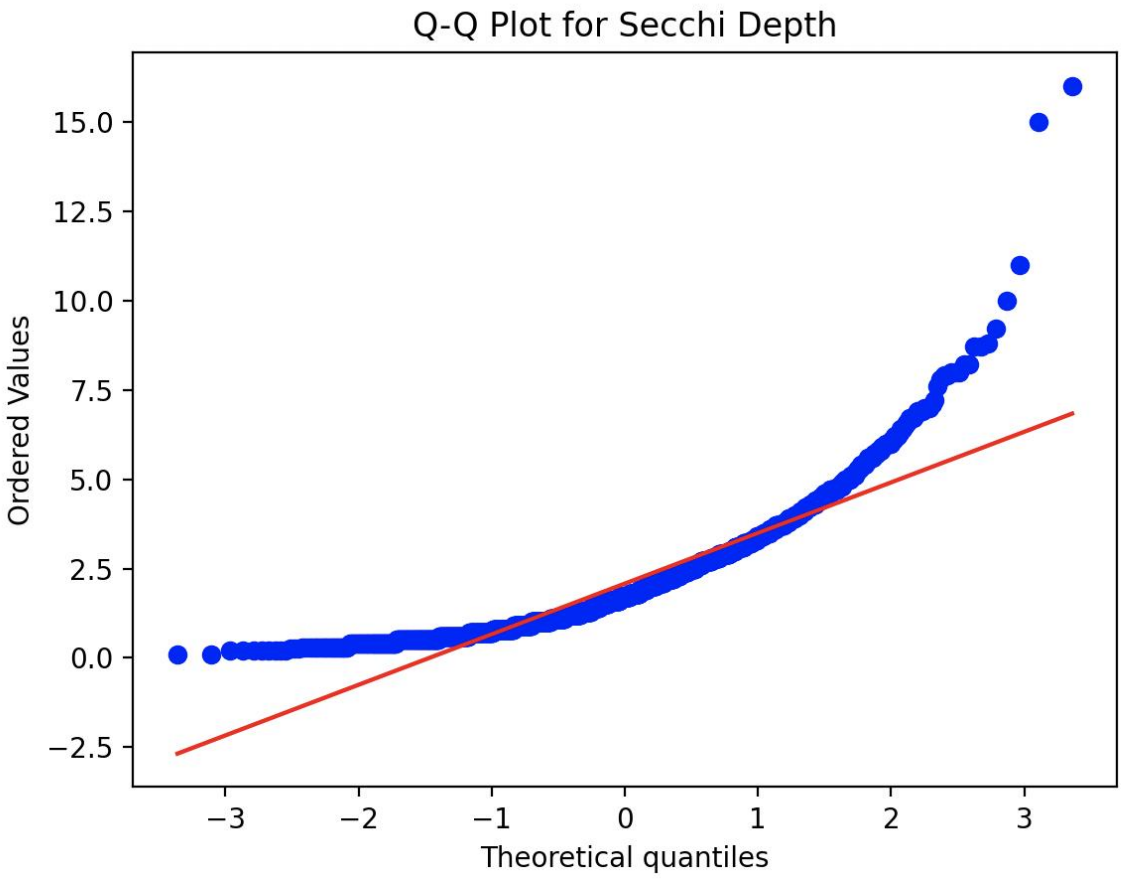
```
...
Water Quality Analysis Results
    Metric  Sample Mean (Estimate of Population Mean)  Threshold  T-Statistic  P-Value (One-Tailed)  Conclusion
0  NO3_epi                                   0.583238       10.0  -292.162100          0.000000e+00  Rejects H₀
1  NH3_epi                                   0.054575        0.5   -99.801253          0.000000e+00  Rejects H₀
2   Secchi                                   2.088890        1.5    15.718425          2.166596e-52  Rejects H₀
>>>
```

In all three cases we can reject the null hypothesis.

Assumption:

- Observations are independent.

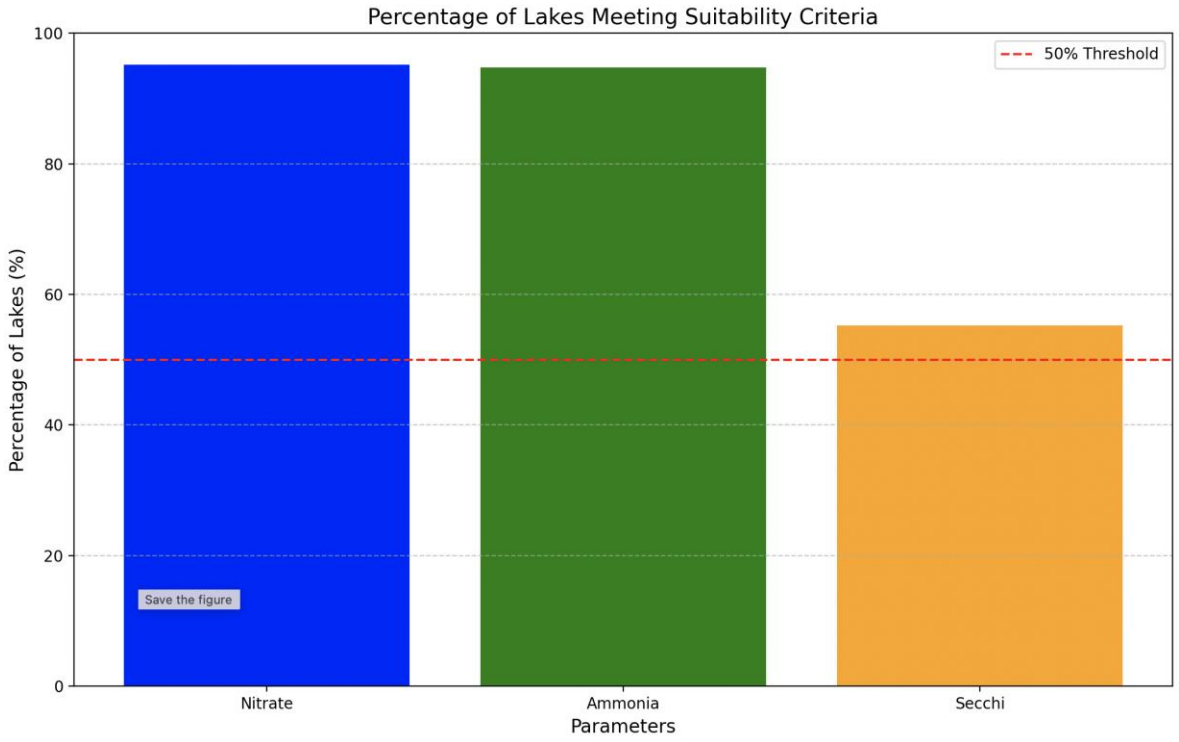
Q-Q Plot (left skewed)



Final Visualization of Data:

For final Visualization we are trying to evaluates the suitability of Indiana lakes for drinking water based on thresholds for nitrate concentration, ammonia concentration, and water clarity (Secchi depth). Then calculate and visualize the percentage of lakes meeting these criteria. We are summing up the number of lakes and checking how much is the percentage of lakes having threshold criteria. Finally, we draw a bar chart out of it.

```
nitrate_threshold = 10
ammonia_threshold = 0.5
secchi_threshold = 1.5

# Evaluate suitability
data['Nitrate_Suitable'] = df['NO3_epi'] <= nitrate_threshold
data['Ammonia_Suitable'] = df['NH3_epi'] <= ammonia_threshold
data['Secchi_Suitable'] = df['Secchi'] >= secchi_threshold

# Count the number of lakes meeting each criterion
suitability_counts = {
    'Nitrate': data['Nitrate_Suitable'].sum(),
    'Ammonia': data['Ammonia_Suitable'].sum(),
    'Secchi': data['Secchi_Suitable'].sum()
}

# Convert to percentages
total_lakes = len(data)

suitability_percentages = {key: (value / total_lakes) * 100 for key, value in suitability_counts.items()}

# Plot the bar chart
plt.figure(figsize=(8, 6))
plt.bar(suitability_percentages.keys(), suitability_percentages.values(), color=['blue', 'green', 'orange'])
plt.ylim(0, 100)
plt.title("Percentage of Lakes Meeting Suitability Criteria", fontsize=14)
plt.ylabel("Percentage of Lakes (%)", fontsize=12)
plt.xlabel("Parameters", fontsize=12)
plt.axhline(50, color='red', linestyle='--', label="50% Threshold")
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



Conclusion:

The red dotted line indicates the threshold line of the parameters. Here we can see most of the Indiana Lakes approx. 90% met the criteria of Nitrate and Ammonia which is great for aquatic life. Secchi also met the criteria. We can do it better by recycling or making frequent water treatment.

Appendix: Find all the relevant code in below git location

https://github.com/iamabhra/week8/blob/main/week8projectAssignment.py