R Final Project - Water quality

Dongni Li & Ziliang Song

December 12, 2024

1.Introduction

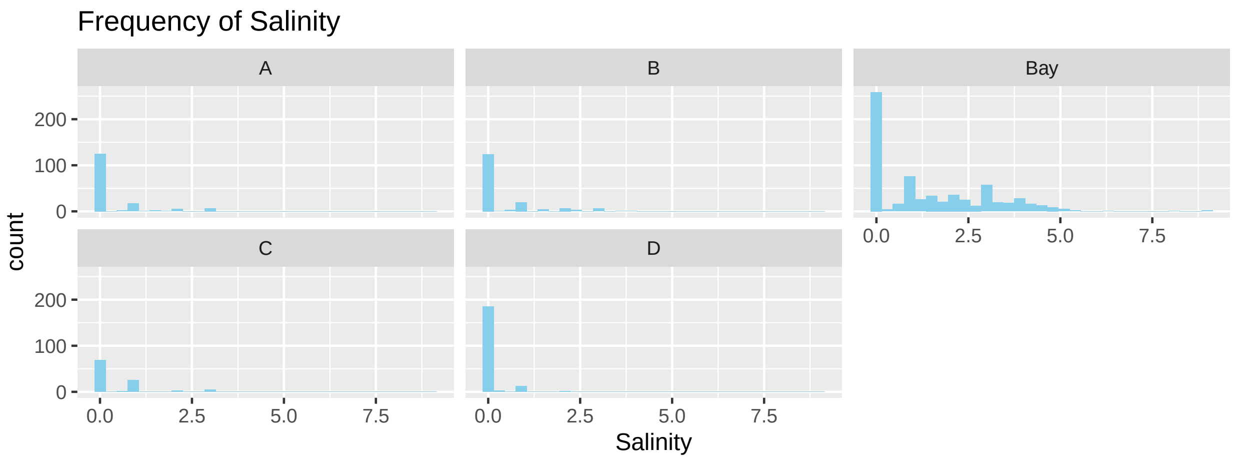
This project analyzes water quality data, including seven quantitative variables (Salinity, Dissolved Oxygen, pH, Water Temperature, Water Depth, Water Visibility Depth, and Air Temperature) and one categorical variable (Location). The study focuses on three key questions: whether the salinity-pH relationship varies by location, which environmental factors most affect water visibility (Secchi Depth), and whether increased water temperature decreases dissolved oxygen.There is a large body of literature exploring the relationship between water quality and environmental factors, some of which are relevant to the hypotheses of this project include whether different locations have an effect on the relationship between salinity and pH, for example, Rugebregt & Nurhati (2020) examined the waters of the Au Ocilir and found that there was a weak correlation between salinity and pH (correlation coefficient of -0.054), suggesting that changes in salinity have a limited direct effect on pH. However, a subsequent study by Rugebregt et al. (2023) reported a moderate correlation (0.314), suggesting that salinity does affect pH to some extent. These studies highlight the complexity of the salinity-pH relationship, suggesting that it may vary by site, which was the focus of our investigation. According to Man'Kovsky & V. I. (2001), water visibility (Secchi depth), an important indicator of water quality, is often used as an indirect assessment of water clarity. The relationship between water visibility and water quality may be affected by a combination of factors, and although the study concluded that secchi depth was not significantly related to water temperature and dissolved oxygen, this provides the research background for our hypothesis 2. Jane et.al (2021) found that there is an inverse relationship between temperature and dissolved oxygen concentration, specifically, as the water temperature increases, the solubility of oxygen decreases, which further leads to a decrease in dissolved oxygen concentration. This supports our hypothesis 3, "Does an increase in water temperature lead to a decrease in dissolved oxygen concentration".

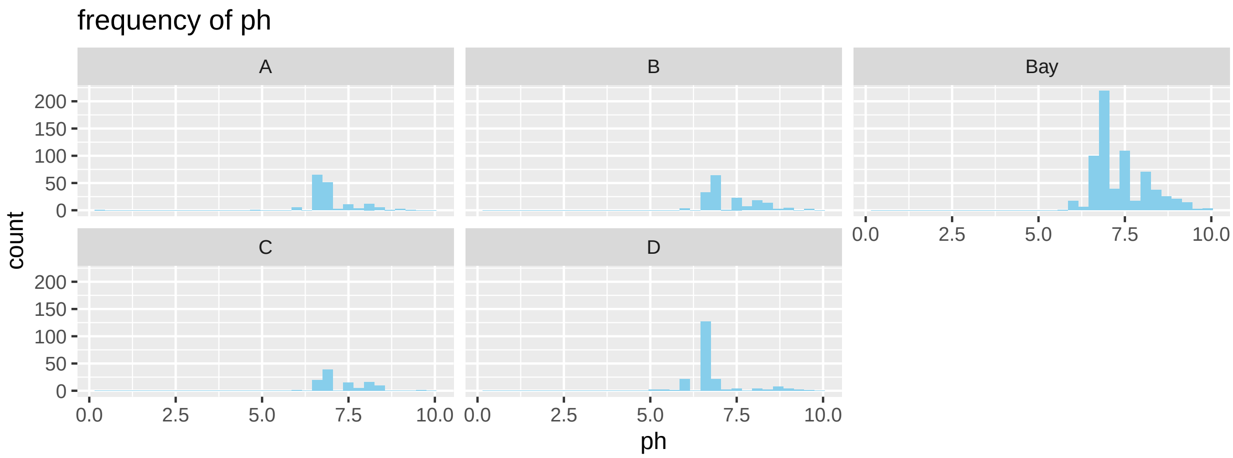
2.Initial Hypotheses

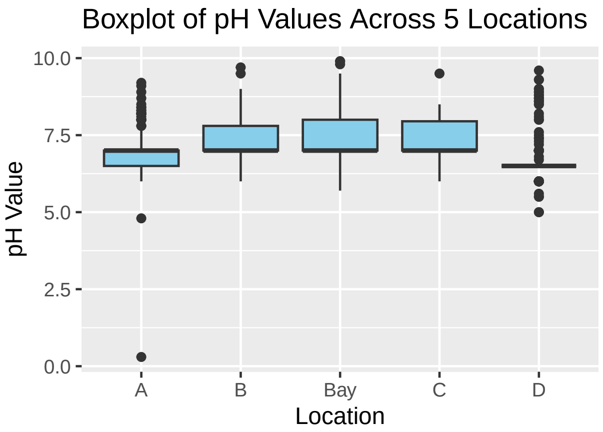
- Hypothesis 1: Does the relationship between salinity and pH vary across different locations?
- Hypothesis 2: What environmental factors (e.g., depth, salinity, dissolved oxygen, pH, temperature, etc.) most strongly affect water clarity (Secchi depth)?
- Hypothesis 3: Does an increase in water temperature result in a decrease in dissolved oxygen concentration?

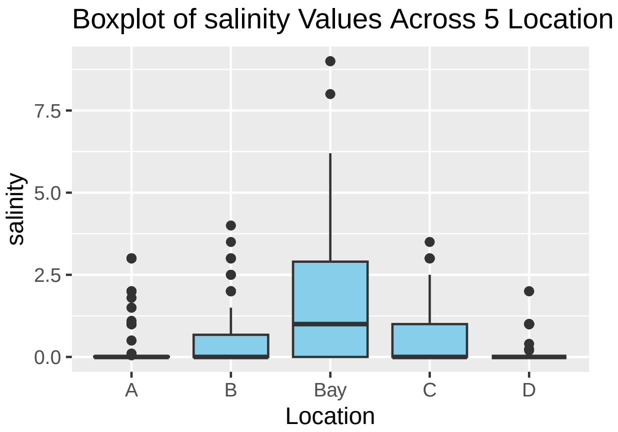
3.Exploratory Data Analysis

- Hypothesis 1:









answer:

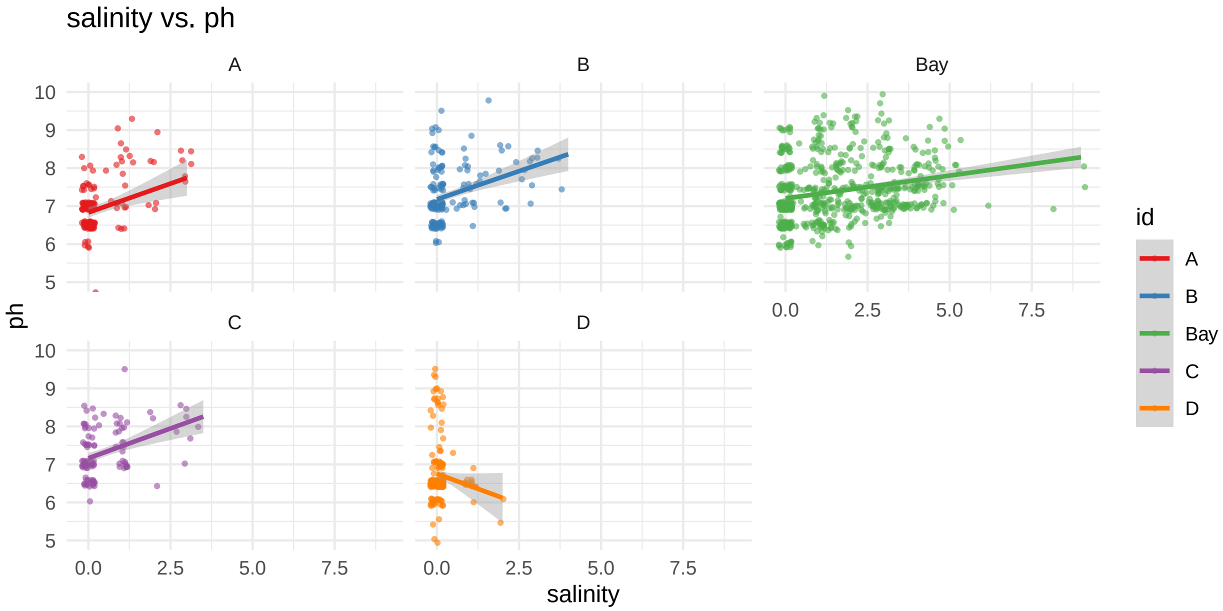
The distribution of salinity at the five locations is similar, all showing a strong lright-skewed trend.

The distribution of pH at the five locations is similar, showing a certain right-skewed trend.

The histograms of the two variables at the five locations are not normally distributed.

The location “Bay” has a larger sample size, so its frequency is generally higher.

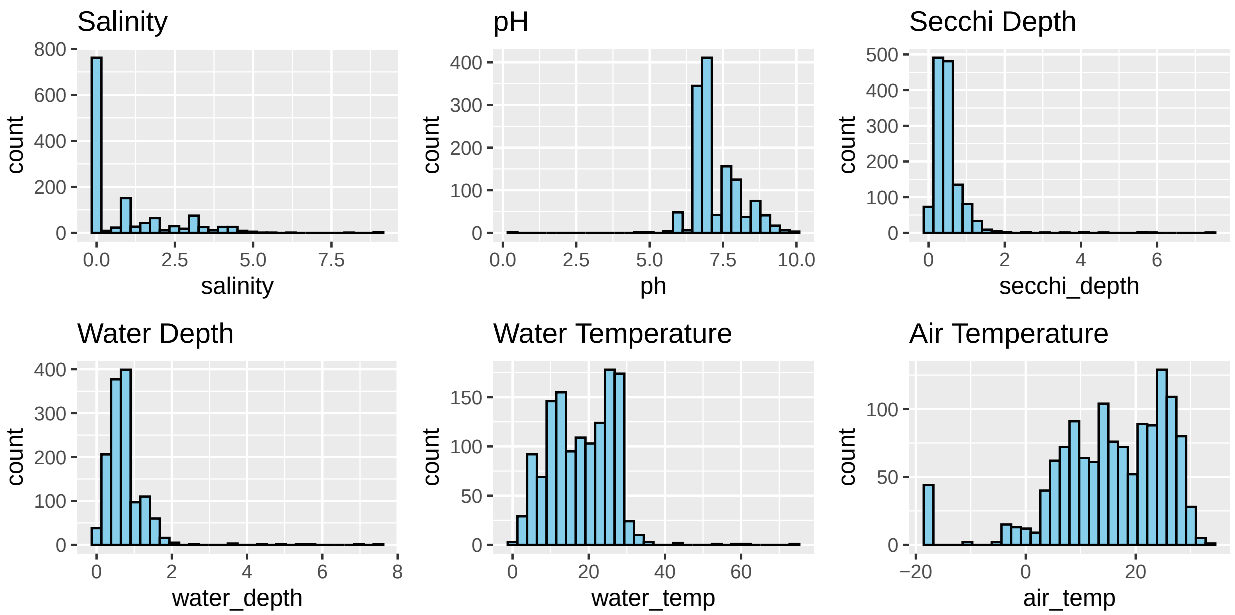
From the boxplot, we can see that locations A and D have smaller sample sizes but contain more outliers.



answer:

At locations A, B, Bay, and C, salinity and pH show a positive linear relationship, but at location D, they exhibit a negative correlation.

- Hypothesis 2:

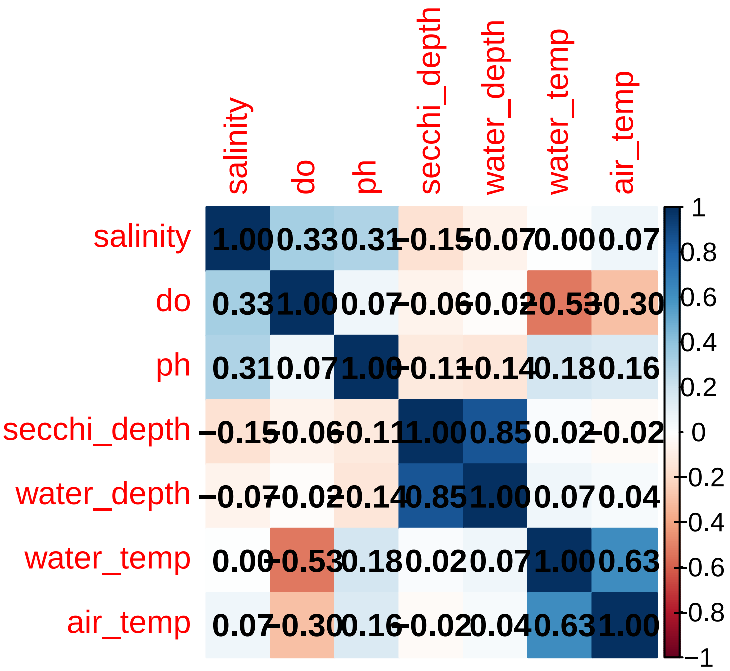


answer:

Based on the plots above, the distributions of the variables salinity, secchi depth, and water depth all show a clear right-skewed pattern. And the pH distribution is close to a normal distribution,

However, based on the histogram of the water temperature variable, we can clearly see two peaks and some relatively large values that occur less frequently, which could potentially be considered outliers.

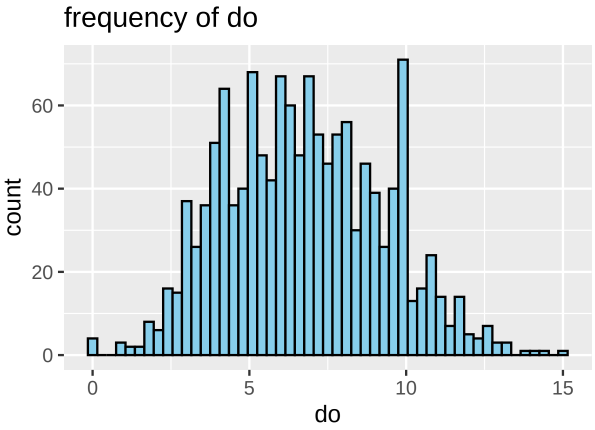
In the histogram of air temperature, we can observe three peaks, with the smallest peak around -20 being far from the main body of the distribution, and there is a gap between them.

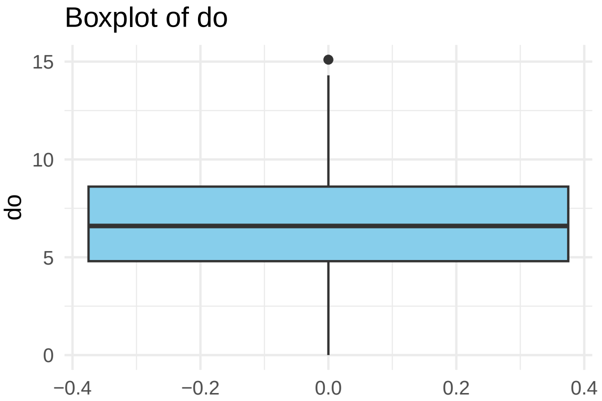


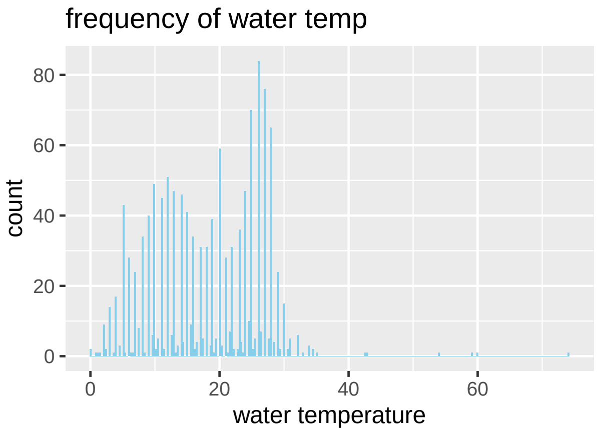
answer:

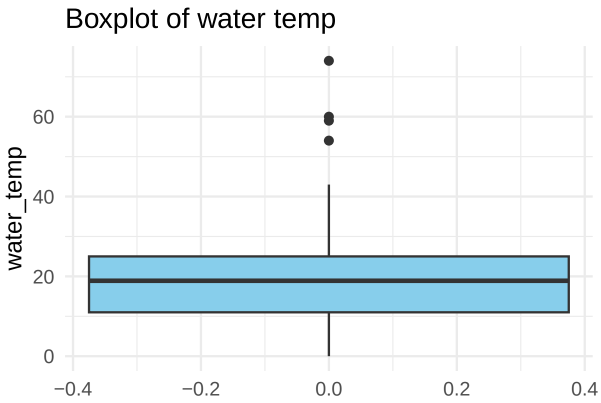
Look at the row secchi depth, the value 0.85 that water depth is strongly and positively associated with water clarity. Greater depth likely corresponds to clearer water. However, the relationship between secchi depth and other variables is not very significant.

- Hypothesis 3:





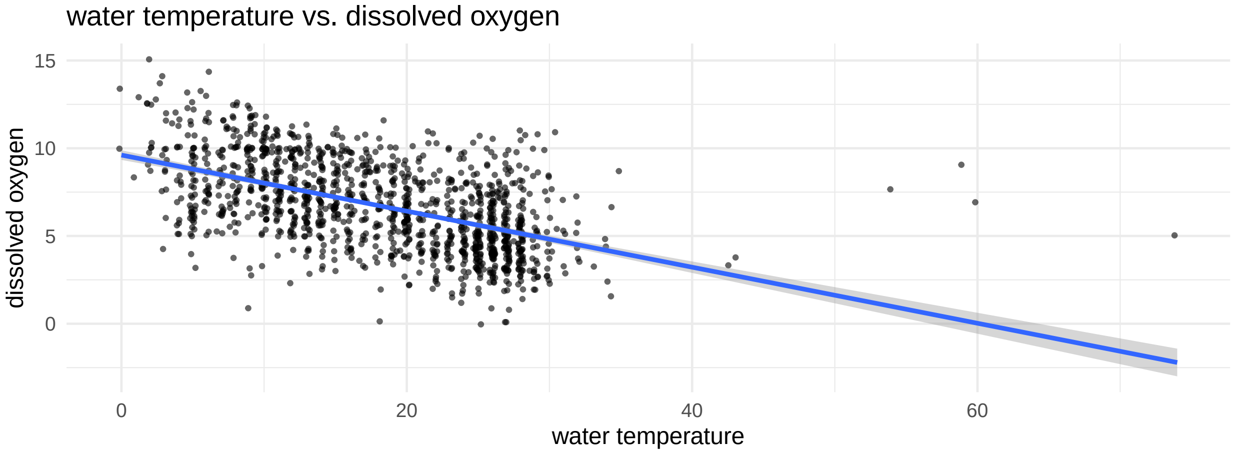




answer:

The distribution of dissolved oxygen is closed to normality. Most of them concentrate around 5 to 10. There is one considered as a outlier.

The distribution of water temperature is right-skewed, with a majority of observations clustered in the lower range. There are also some higher water temperature values greater than 40°C, but these are very rare. According to the box plot, we can see that there are four outliers here.



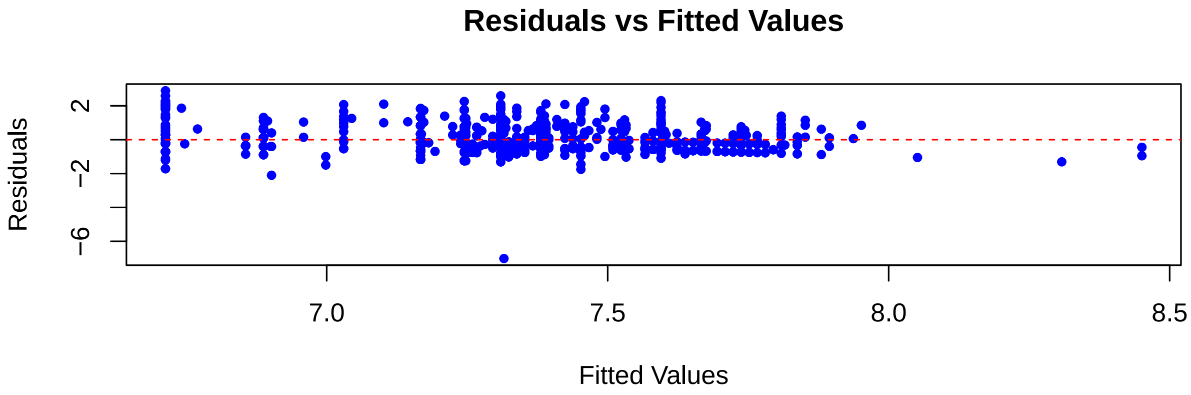
answer:The plot clearly shows a negative trend, indicating that as water temperature increases, dissolved oxygen levels tend to decrease.

4.Statistical Methods, Modeling, and Results

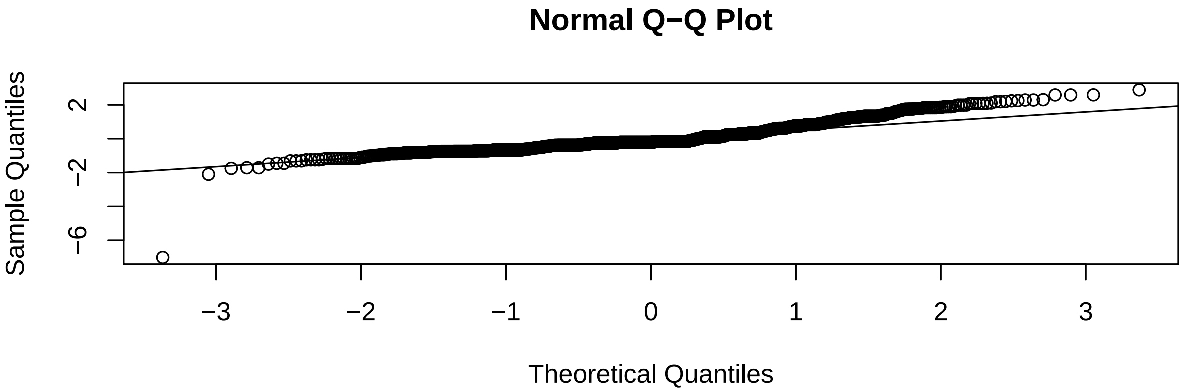
- Hypothesis 1: Model1 equation:$pH = \alpha + \beta_1 \cdot \text{Salinity} + \beta_2 \cdot \text{Location}$

Answer:

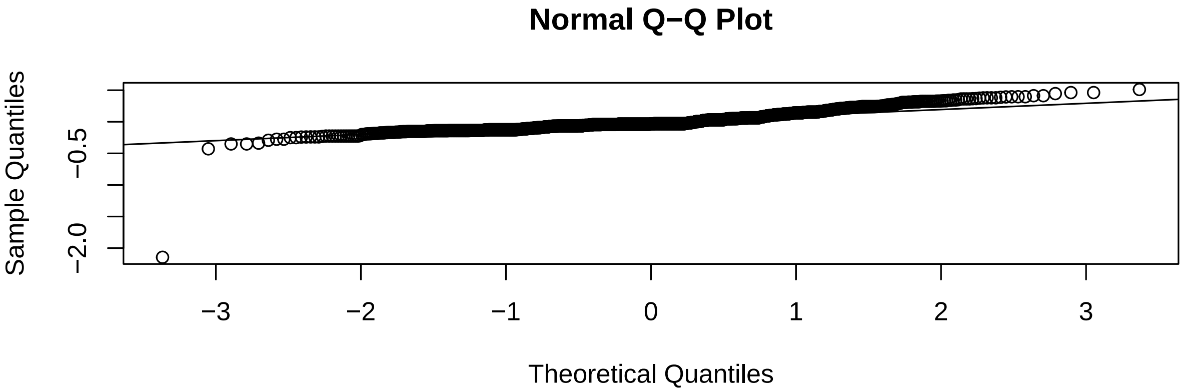
Pool B's response value is 0.3573 units higher than the reference pool (adjusted for other predictors). Bay's response value is 0.2795 units higher than the reference pool (adjusted for other predictors). Pool C's response value is 0.36 units higher than the reference pool (adjusted for other predictors). Pool D's response value is 0.1745 units lower than the reference pool (adjusted for other predictors).

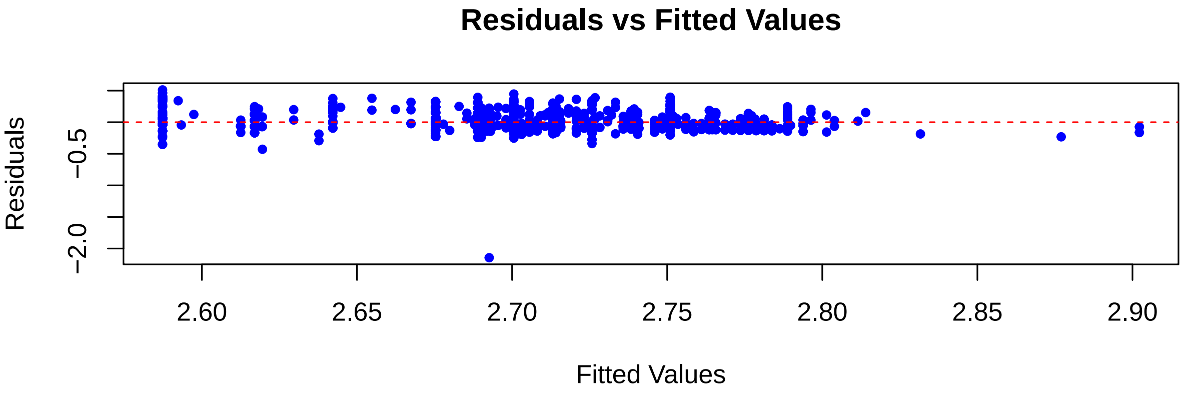


Answer:From the residuals vs. fitted values plot, it can be observed that the variance of residuals shows a slight decreasing trend from large to small; overall, the distribution appears relatively random. There is something special about this graph: many points on the far left are concentrated around 7.1.



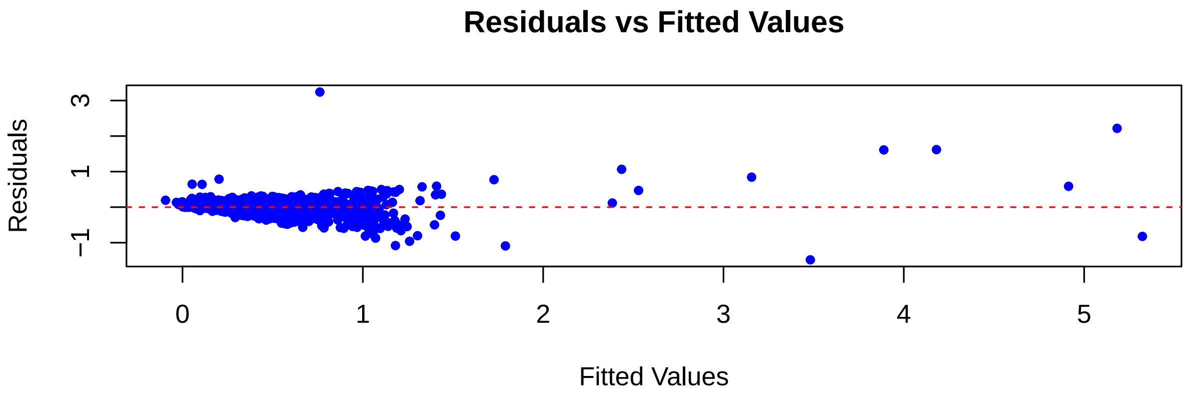
Answer:As can be seen from the figure above, most of the points in the front follow the normality line, and only the back shows obvious deviation.



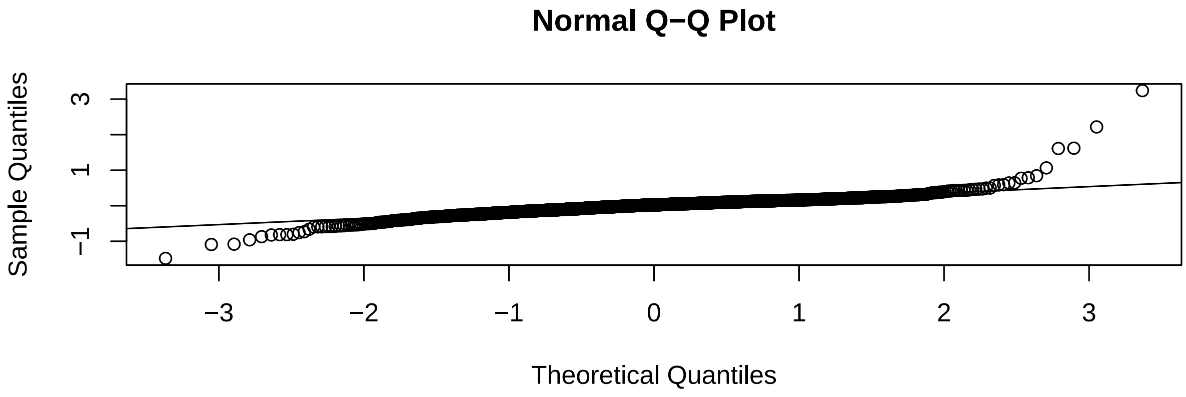


Answer:From the enhanced model, we can see that although the residual plot hasn't changed much, we have improved the normality of the residuals. Plots around the tail get closer to the line than before.

- Hypothesis 2: $SecchiDepth = \alpha + \beta1 * Salinity + \beta2 * DissolvedOxygen + \beta3 * pH + \beta4 * WaterDepth + \beta5 * WaterTemp + \beta6 * AirTemp$



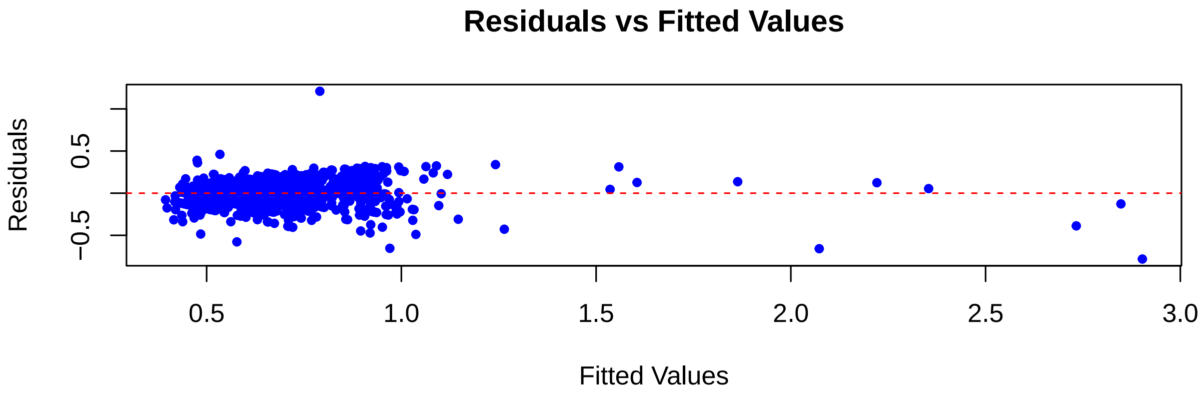
Answer:We can see that the variance of the residuals increases over time and is not randomly distributed.

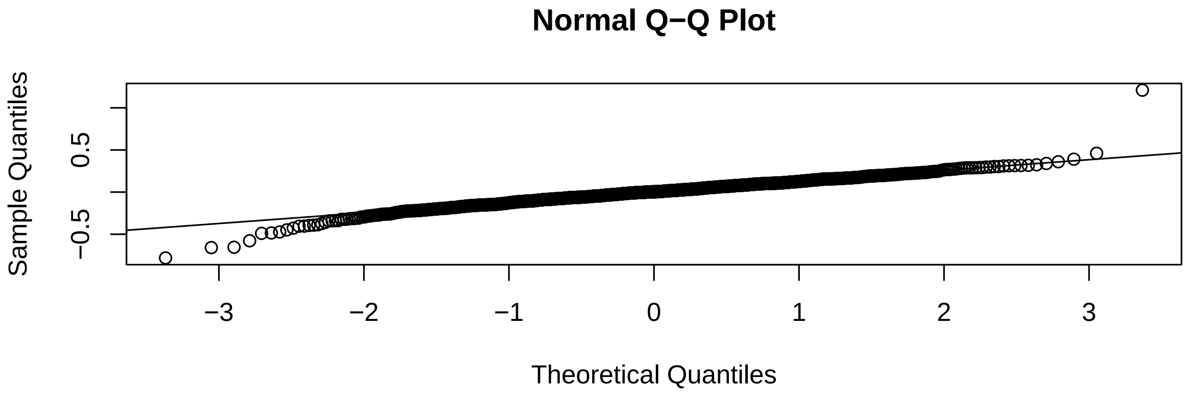


Answer:The data points generally follow the theoretical normal line, but notable deviations occur in the tails.

$$sqrtY = \alpha + \beta 1 * Salinity + \beta 2 * DO + \beta 3 * pH$$

$$+ \beta 4 * WaterDepth + \beta 5 * WaterTemp + \beta 6 * AirTemp$$

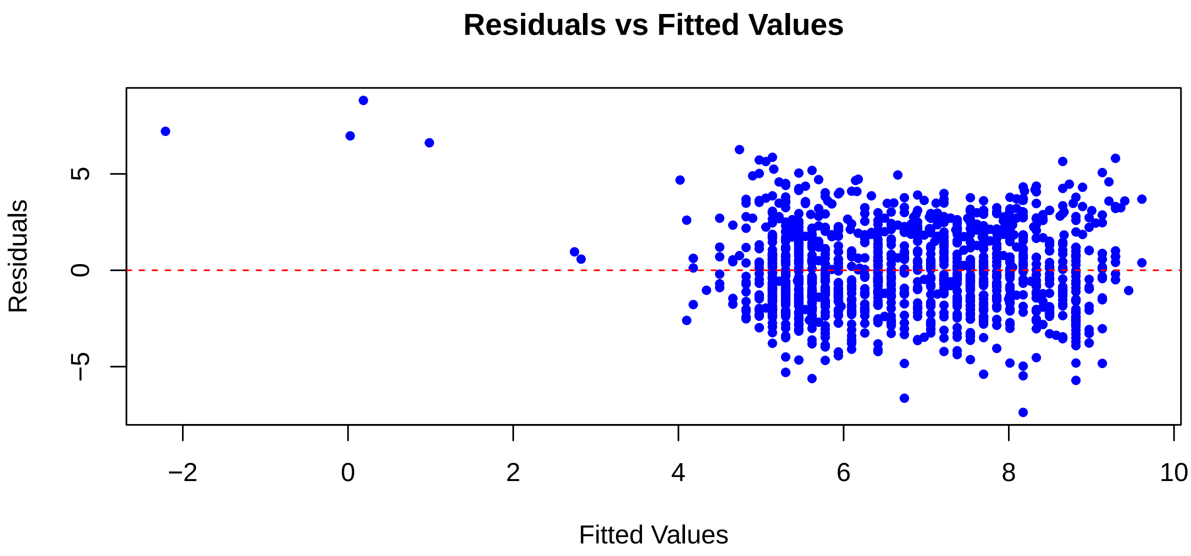




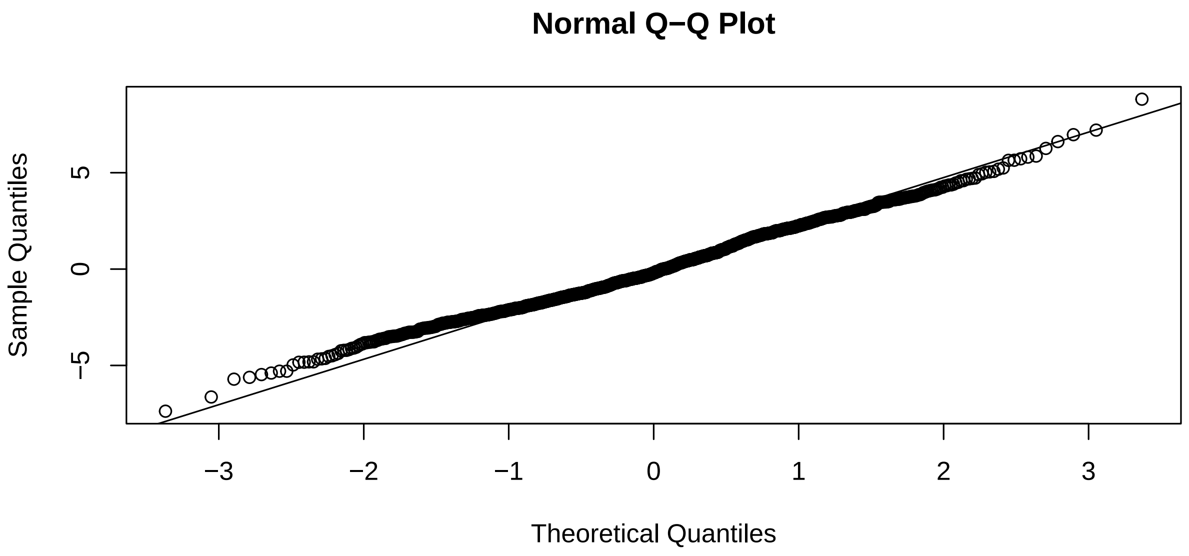
Answer:In the transformed model for the Y values, compared to the previous model, the residuals in the enhanced model exhibit a wider distribution without any noticeable trends. Additionally, in the normal plot, the points near both tails are closer to the normal line.

Answer:We also performed VIF analysis, The VIF analysis shows that the values of these predictors range between 1 and 5, indicating that there is no issue of multicollinearity among them.

- Hypothesis 3: $DissolvedOxygen = \alpha + \beta_1 \cdot \text{Water Temperature}$



Answer:Overall, the residuals show no noticeable trend and appear to be randomly distributed, although some points are farther away from the main cluster.



Answer:Most of the points in the plot closely follow the normal line.

5. Data-driven Hypotheses

During our data analysis, we came up with a new assumption of normality as part of the model diagnostic process. During the data cleaning process, we discovered a bimodal distribution in the histogram of the water temperature variable, characterized by continuous fluctuations. We suspect this may be related to the timing of data collection. We also found out that all the histograms of all variables show a left-skewed or right-skewed trend, except for the dissolved oxygen variable.

6. Discussion:

- Hypothesis 1: The results of regression analysis showed that the relationship between salinity and pH varied from site to site, a finding that partially supports existing studies. For example, Rugebregt & Nurhati (2020) in their study of Ohoililir waters found that the correlation between salinity and pH was weak (correlation coefficient -0.054), suggesting that changes in salinity have a limited direct effect on pH. However, a study by Rugebregt et al. (2023) yielded a moderate correlation (correlation coefficient 0.314) between salinity and pH. In contrast to these studies, our analysis further explored the effect of salinity on pH at different sites and revealed subtle differences in this relationship by grouping salinities together. This approach provides a deeper understanding of interregional heterogeneity without directly comparing the relative impacts of salinity to other environmental variables as in the previous studies.
- Hypothesis 2: Using the stepwise method to build the model, we identified water

depth as the primary factor influencing Secchi depth. After refining the model and reapplying the stepwise method, water depth remained the most significant factor affecting Secchi depth. The study by Man'Kovsky & V. I. (2001) indicated that the relationships between water temperature, dissolved oxygen (DO), and Secchi depth were not significant, suggesting that these factors have a minimal impact on water transparency. Although the paper did not specifically mention water depth, it analyzed the influence of other factors on Secchi depth.

- Hypothesis 3: Analysis through the development of a regression model indicated that an increase in water temperature was associated with a decrease in dissolved oxygen (DO) levels. This is in line with the findings of Jane et al (2021), which emphasized an inverse relationship between temperature and oxygen solubility.

7. References:

- (1) Rugebregt, M. J., Opier, R. D. A., Abdul, M. S., Triyulianti, I., Kesaulya, I., Widiaratih, R., Sunuddin, A., & Kalambo, Y. (2023). Changes in pH associated with temperature and salinity in the Banda Sea. IOP Conference Series. Earth and Environmental Science, 1163(1), 12001-. https://doi.org/10.1088/1755-1315/1163/1/012001
- (2) Rugebregt, M. J., & Nurhati, I. S. (2020, December). Preliminary study of ocean acidification: relationship of pH, temperature, and salinity in Ohoililir, Southeast Maluku. In IOP Conference Series: Earth and Environmental Science (Vol. 618, No. 1, p. 012004). IOP Publishing. DOI 10.1088/1755-1315/618/1/012004

- (3) Man'Kovsky, V. I. (2001). Relationship between the depth of visibility of Secchi disk and biooptical characteristics of the Black Sea waters. Physical Oceanography, 11(5), 491–494. https://doi.org/10.1007/BF02509714
- (4) Jane, S. F., Hansen, G. J. A., Kraemer, B. M., Leavitt, P. R., Mincer, J. L., North, R. L., Pilla, R. M., Stetler, J. T., Williamson, C. E., Woolway, R. I., Arvola, L., Chandra, S., DeGasperi, C. L., Diemer, L., Dunalska, J., Erina, O., Flaim, G., Grossart, H.-P., Hambright, K. D., … Rose, K. C. (2021). Widespread deoxygenation of temperate lakes. Nature, 594(7861), 66–70. https://doi.org/10.1038/s41586-021-03550-y

8.Appendix:

We found water quality data on the U.S. Government's Open Data. Here is the link to the data:This is the data source.,and there are 17 columns and 2371 rows of data