Reflections on Water: A Data Story of Quality and Ecology in Prince Albert's Lakes

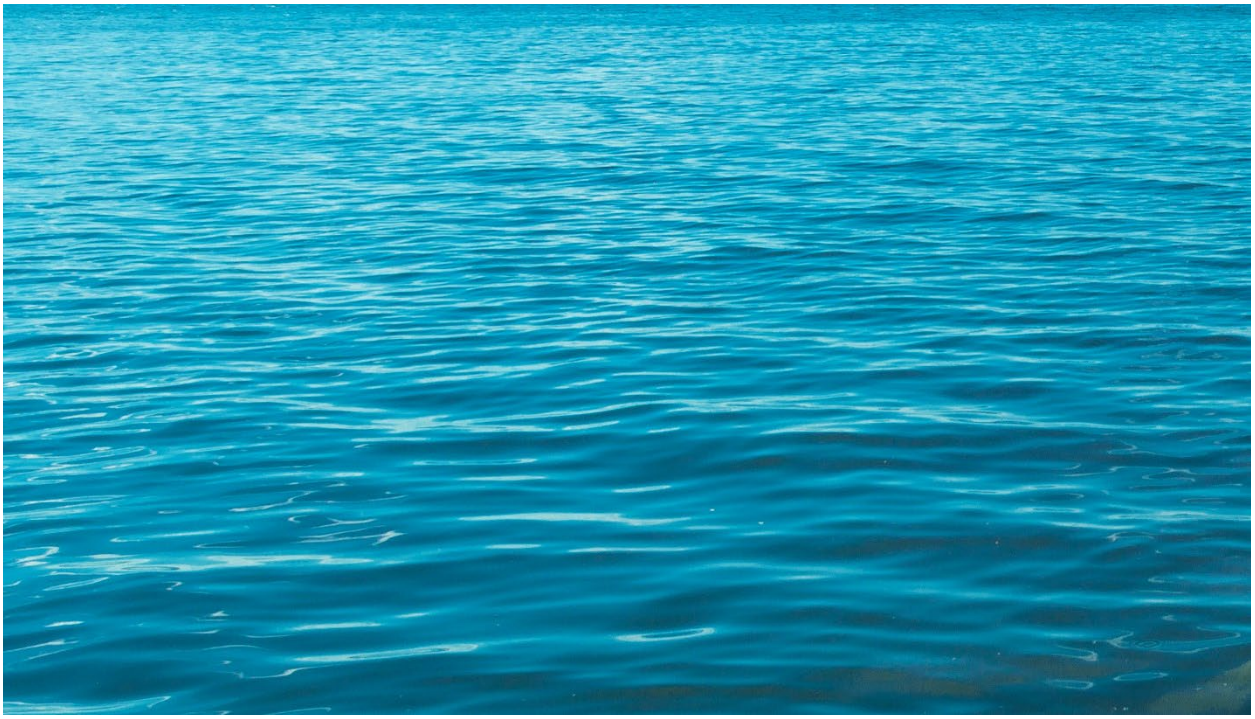
Zhiqi Pang

Introduction

Water, often referred to as the lifeblood of our planet, carries within its currents, intricate tales of ecological, chemical, and biological interactions. Our data-driven narrative, "Reflections on Water: A Data Story of Quality and Ecology in Prince Albert's lakes," aims to elucidate these silent stories, focusing on the water quality and ecological health of Kingsmere and Waskesiu lakes. These aquatic bodies, nestled within the verdant expanses of Prince Albert, Saskatchewan, Canada, play an instrumental role in our exploration of water's diverse narratives.

Kingsmere, a tranquil haven in the northwest of Prince Albert National Park, is known for its pristine waters surrounded by vast forests, attracting outdoor enthusiasts and nature lovers. In contrast, Waskesiu lake, located at the heart of the park, is a bustling hub of recreation and tourism, characterized by its rich biodiversity, expansive sandy beaches, and an array of flora and fauna. Both lakes are vibrant ecosystems, providing habitats for numerous species, each with its own unique tale to tell.

Through the lens of data analysis, our story embarks on a journey to uncover the delicate balance of these aquatic ecosystems. By studying the water quality metrics, we hope to unveil the intricate relationships and patterns that underpin their overall environmental health and sustainability.



Guiding Questions

In this data narrative, I aim to unravel the available water quality data to address the following five questions:

1. How do interactions between chemical attributes in water samples influence the overall water quality in aquatic ecosystems?

2. How does the Secchi depth change with seasons, and what does it tell us about variations in lake water transparency?

3. What patterns of change have been observed in the temperature, pH, and dissolved oxygen levels in the lake over time?

4. How do the total coliform concentrations contrast between Kingsmere lake and Waskesiu lake?

5. How do contamination levels in the lake vary across different months, and what environmental factors might be contributing to these changes?

Tools

In this data exploration, I predominantly employed Python and R for data manipulation, cleaning, and analysis. For visualization, I turned to tools such as Tableau, seaborn, matplotlib, and ggplot2. These tools help convert complex data into easily comprehensible insights and visually compelling narratives.

The Data

The data utilized in this context is sourced from the "Water Quality - Prince Albert" dataset, published by the Government of Canada in 2023. This dataset, a product of extensive collaboration between Parks Canada, the Saskatchewan Research Council, and Environment Canada, provides a comprehensive view of the water quality in two of the largest lakes in Prince Albert National Park - Kingsmere and Waskesiu.

This dataset includes six individual CSV files, each focusing on a specific aspect of water quality. The data has been carefully collected through data collection from May to September annually, augmented by additional water chemistry sampling in March. Water chemistry and Secchi disk depths are sampled by boat at three open water locations, and E.coli levels are tested from three high-use beach locations. The dataset encompasses diverse variables from water chemistry to discharge measurements and constitutes a robust, multi-dimensional dataset providing a comprehensive overview of the water quality in the lakes of Prince Albert National Park.

The dataset includes the following variables:

- Site: The location where the data was collected.
- Date: The exact date of data collection.
- Depth: Depth at which data was collected, measured in m.
- Chlorophyll: The amount of chlorophyll-a in the water, measured in mg/L.
- pH: The pH level of the water, indicating acidity or alkalinity.
- Dissolved Organic Carbon (DOC): The amount of dissolved organic carbon in the water, measured in mg/L.
- Discharge: The volume of water flowing through a given cross-section, measured in cubic meters per second (m^3/s).
- Dissolved Oxygen: The amount of dissolved oxygen in the water, presented in terms of saturation percentage and mg/L.
- Secchi Disk Depth: The depth to which a Secchi disk can be seen in the water, indicating water clarity, measured in m.
- Ammonia: The amount of ammonia in the water, measured in mg/L.
- Nitrite and Nitrate: The combined amount of nitrite and nitrate in the water, measured in mg/L.
- Nitrogen Oxides (Nox): A group of seven gases and compounds composed of nitrogen and oxygen, usually measured in ppb or ppm, depending on the specific compound.
- Temperature: The temperature of the water at the time of data collection, measured in °C.
- Total Dissolved Solids (TDS): The total amount of dissolved solids in the water, measured in mg/L.

- Total Phosphorus: The total amount of phosphorus in the water, measured in mg/L. This includes both organic and inorganic forms of phosphorus.
- E.Coli: The amount of E.Coli present in the water, measured in CFU/100 mL.
- Total Coliforms: The total amount of coliform bacteria in the water, measured in CFU/100 mL.

The dataset is licensed under the Open Government Licence - Canada. This licensing permits users to copy, modify, publish, translate, adapt, and distribute.

Data wrangling

The dataset was comprehensive, with each sub-dataset serving a specific purpose. The data wrangling phase sought to standardize, clean, and structure the data for in-depth analysis. I standardized the column names for clarity and removed null values to ensure accuracy and relevance. Some CSV files initially presented all information in a single column separated by semicolons. I split this into distinct columns, ensuring each attribute and its value were correctly associated. This wrangling process has made the data more navigable and primed for extensive exploration and analysis.

With our dataset prepared and our questions poised, we embark on a journey to weave together the narratives hidden within the depths of Kingsmere and Waskesiu.

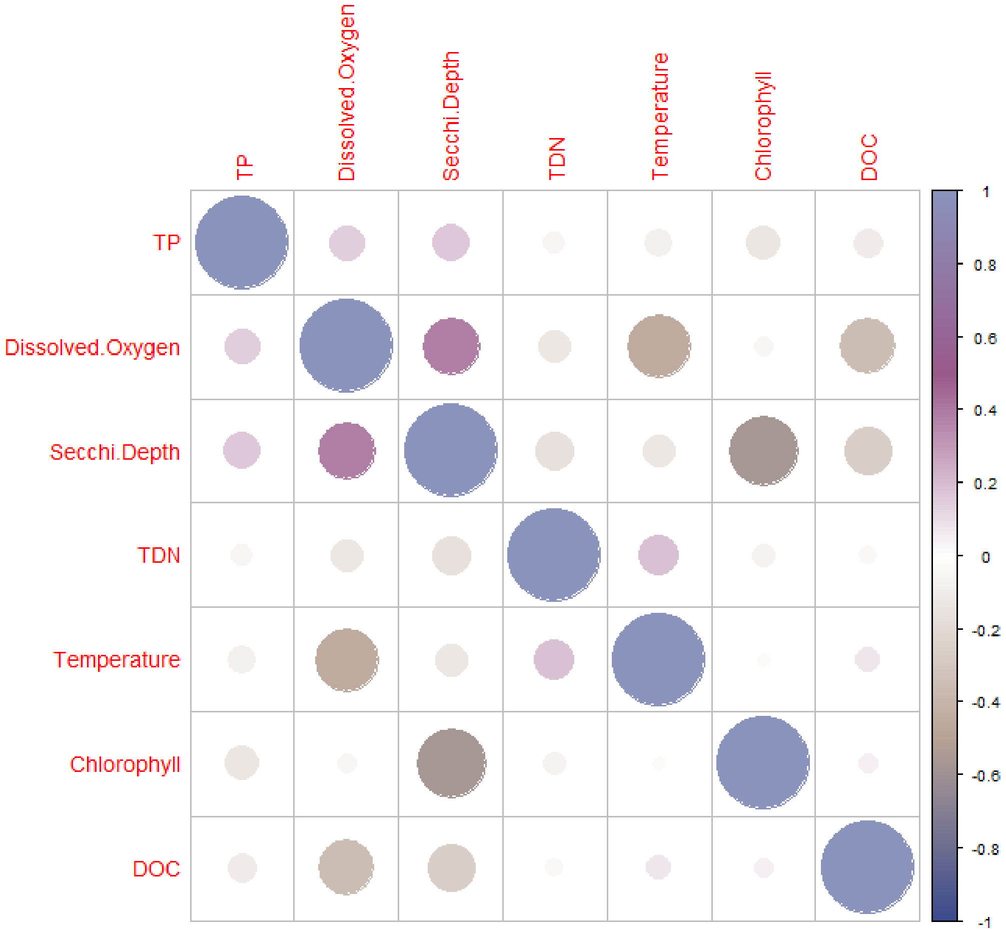
Guiding Question 1: How do interactions between chemical attributes in water samples influence the overall water quality in aquatic ecosystems?

Investigating the quality of lake water is a multifaceted endeavor. The clarity of the narrative emerges from understanding the intricate relationships among diverse chemical attributes that collaboratively shape the water quality. To this end, we applied correlation analysis, a statistical method that reveals the degree and direction of association between variables.

Correlation analysis illuminates the linear relationships between variables, quantified by the correlation coefficient, which ranges between -1 and 1. A coefficient close to -1 or 1 implies a strong relationship between the variables, with positive values indicating a direct relationship and negative values signifying an inverse relationship. Conversely, a value close to 0 indicates a weak or negligible relationship.

By examining the "Water Quality - Prince Albert" dataset, we derived Pearson correlation coefficients for different chemical attributes, revealing the following insights:

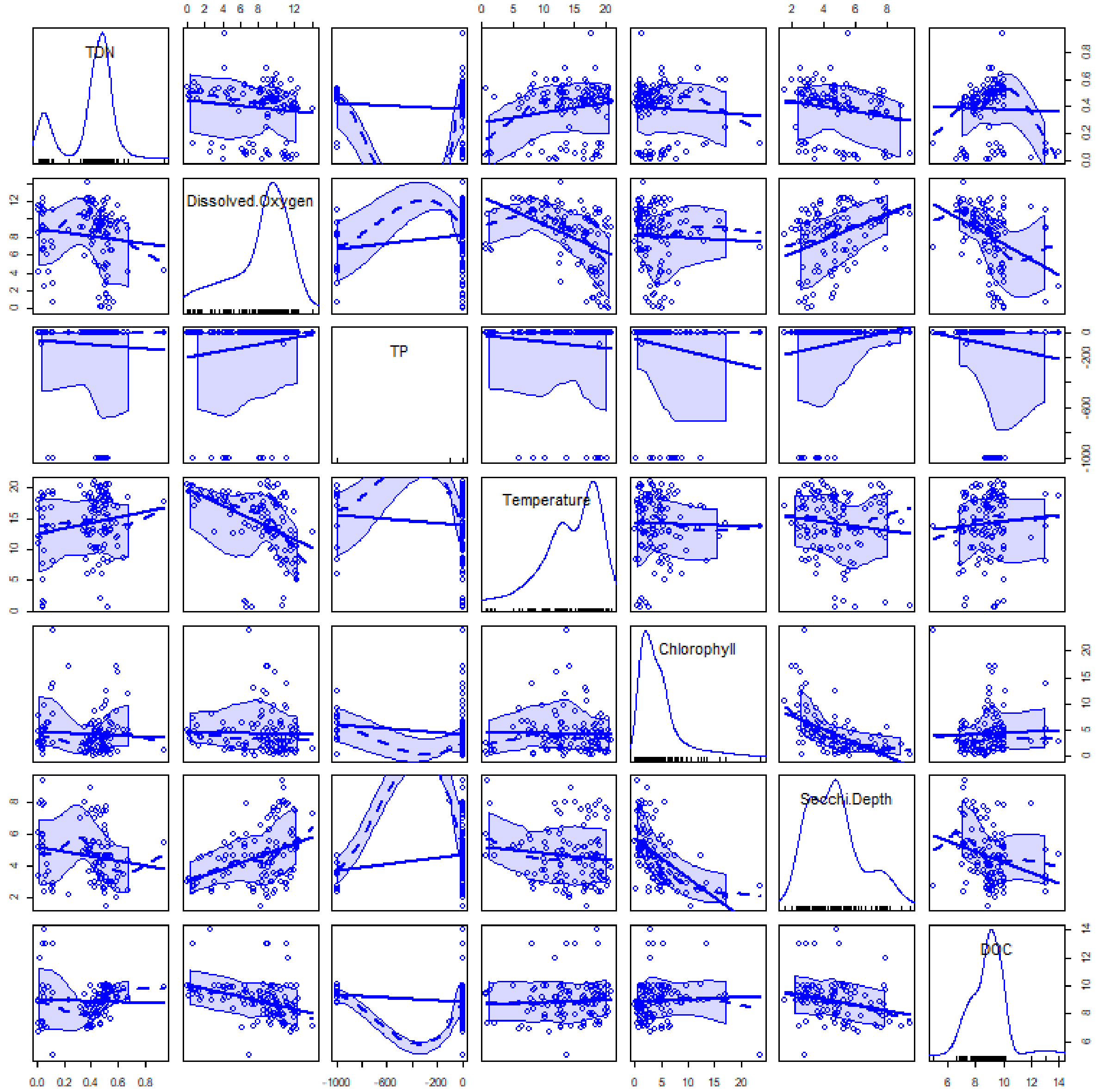
- Ammonia and Total Dissolved Nitrogen (TDN) show a minor negative correlation of -0.0688, indicating that an increase in one might correspond to a slight decrease in the other.
- A noticeable negative correlation of -0.4587 exists between Dissolved Oxygen and Temperature, suggesting that as the water's temperature increases, the level of dissolved oxygen might decrease.
- A pronounced negative correlation of -0.5613 is observed between Secchi Depth and Chlorophyll, suggesting that an increase in the concentration of chlorophyll in the lake water may result in reduced water clarity.
- Dissolved Organic Carbon (DOC) and Dissolved Oxygen show a moderate negative correlation of -0.3598, indicating that an increase in dissolved organic carbon may accompany a decrease in the amount of dissolved oxygen.



To visualize these relationships, I constructed a scatterplot matrix showcasing pairwise relationships between the selected chemical attributes of lake water. This comprehensive visualization helps identify correlations, distributions, and potential patterns or outliers within the data:

Diagonal Plots: These plots show the distribution of individual variables through histograms (or density plots), providing insights into the underlying distribution of the data. Most variables follow a near-normal distribution, suggesting a balanced dataset. Off-diagonal Plots: These scatter plots present relationships between pairs of variables. The direction of correlation is indicated by the positive or negative slope of the data points. For instance, the scatter plots show a potential negative correlation between Dissolved Oxygen and Temperature. Symmetry: The scatterplot matrix is symmetric, offering two perspectives on each relationship. This symmetry is useful for corroborating observations.

In conclusion, the scatterplot matrix serves as a foundation for our analysis, enabling us to identify variables that likely have significant relationships. It illustrates the complex and intertwined relationships between these variables, underscoring the multi-dimensional nature of water quality in aquatic ecosystems.



Guiding Question 2: How does the Secchi depth change with seasons, and what does it tell us about variations in lake water transparency?

In the realm of environmental science and water treatment, Secchi depth serves as a key indicator of water transparency, playing a crucial role in assessing the health and quality of aquatic environments. When Secchi depth is low, it often suggests an elevated level of turbidity, potentially indicative of a high concentration of suspended materials such as sediments, algae, or various pollutants. These particulates can significantly limit light penetration, disrupting aquatic life and potentially compromising the aesthetic and recreational value of the water body.

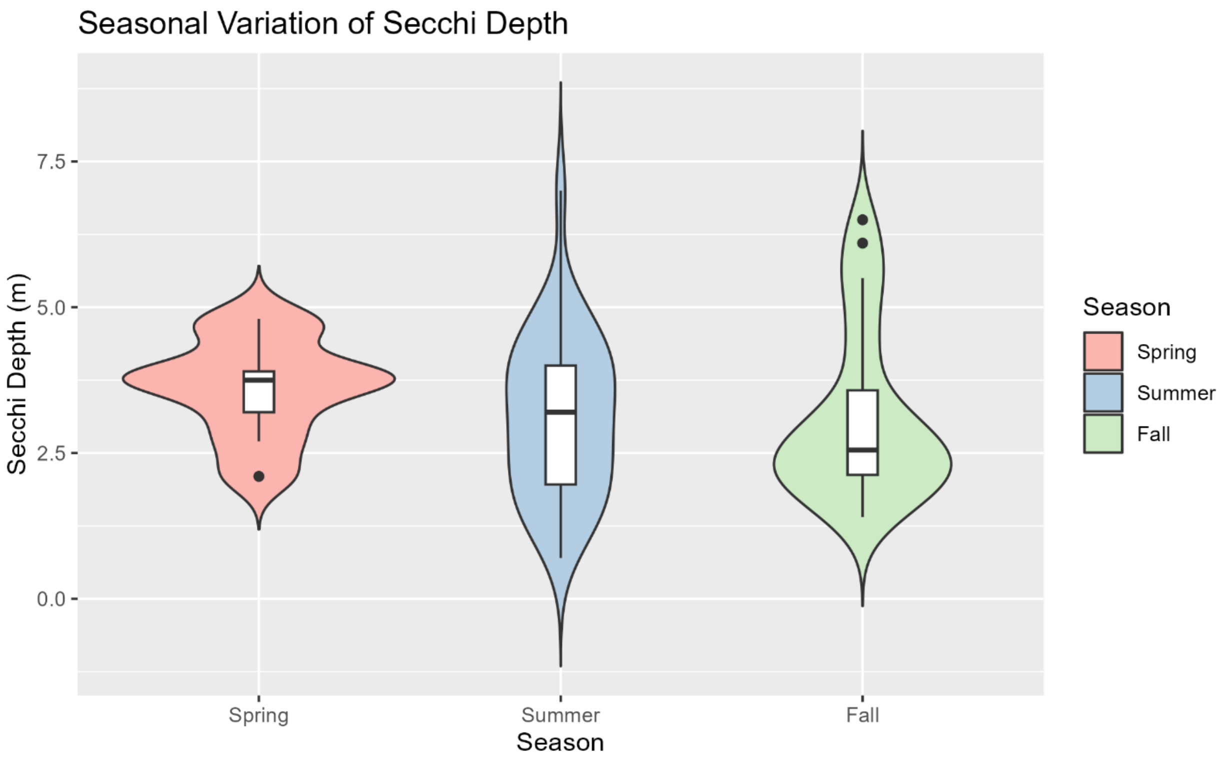
On the other hand, a higher Secchi depth signifies clearer water, typically synonymous with improved water quality. Such clarity implies a reduced concentration of suspended particles, allowing for greater light penetration, which in turn enhances photosynthetic activity, fostering a healthier and more diverse aquatic ecosystem. Additionally, clearer water is more visually appealing and promotes recreational activities, such as swimming and boating.

The analysis of the data reveals distinct seasonal variations in Secchi Depth values, as depicted in the violin plot. Spring exhibits the highest Secchi Depth values, with a median of 3.75 meters, indicating exceptional water clarity during this season. As we transition into summer, the median Secchi Depth slightly decreases to 3.20 meters, indicating a marginal reduction in water clarity compared to spring. Autumn experiences the lowest Secchi Depth values, with a median of 2.55 meters, indicating reduced water clarity during this season. However, it is important to note that data for the winter season is unavailable, limiting our understanding of Secchi Depth during that time.

These seasonal fluctuations in Secchi Depth can be attributed to various factors. During spring, the enhanced clarity can be attributed to lower nutrient levels and reduced algal blooms due to colder temperatures and decreased sunlight. Additionally, the melting snow and increased freshwater input during this season contribute to improved water quality.

In autumn, the lower Secchi Depth values can be linked to heightened biological activity, warmer water temperatures, and increased nutrient inputs, facilitating algal growth and diminishing water clarity. Moreover, recreational activities and increased human visitation during the summer months may lead to sediment disturbance and the presence of suspended particles in autumn, further influencing Secchi Depth values.

However, it's important to underscore that extremely clear water may also suggest a deficiency in the nutrients necessary to sustain a balanced aquatic ecosystem. Therefore, while Secchi depth serves as a valuable tool in evaluating water quality, it must be complemented by other parameters for a holistic assessment of water body health.

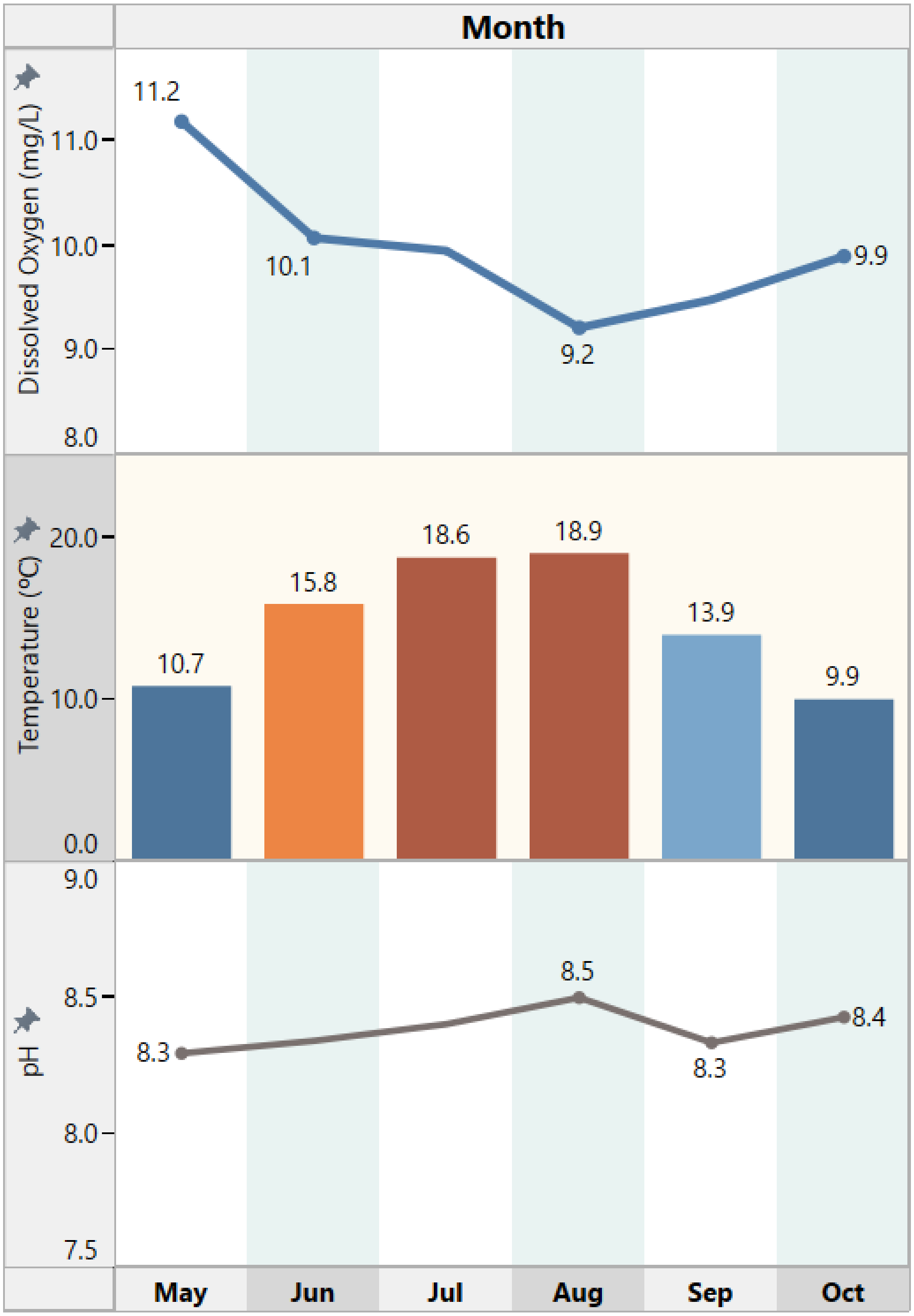


Guiding Question 3: What patterns of change have been observed in the temperature, pH, and dissolved oxygen levels in the lake over time?

Temperature plays a vital role in the solubility of oxygen in water; as water temperatures increase, the level of dissolved oxygen typically decreases. This is because warmer water holds less dissolved oxygen than cooler water. Therefore, during the hotter months such as August, when water temperatures peak, we tend to see lower levels of dissolved oxygen. In contrast, during cooler months such as May and October, water temperatures are relatively lower, and consequently, levels of dissolved oxygen are higher. This pattern can be seen clearly in our data, confirming the theoretical relationship between temperature and dissolved oxygen.

pH, on the other hand, is a measure of how acidic or basic the water is. The pH of a water body can influence the types of organisms that live in it and can also affect chemical reactions that take place in the water. In Kingsmere and Waskesiu lakes, the pH levels remained relatively stable, hovering around 8.5, indicating that the water bodies are mildly alkaline. The steadiness in pH may be attributed to the lakes' buffering capacity, which enables them to resist significant pH fluctuations.

However, it's important to remember that these relationships are not exclusive and can be influenced by many other factors. For instance, an increase in temperature can also increase the metabolic rates of aquatic organisms which may further decrease dissolved oxygen due to increased respiration.


Month
Dissolved Oxygen (mg/L)
11.0
10.0
9.0
8.0
11.2
10.1
9.2
9.9
Temperature (ºC)
20.0
10.0
0.0
10.7
15.8
18.6
18.9
13.9
9.9
pH
9.0
8.5
8.0
7.5
8.3
8.5
8.3
8.4
May
Jun
Jul
Aug
Sep
Oct

Guiding Question 4: How do the total coliform concentrations contrast between Kingsmere and Waskesiu lakes?

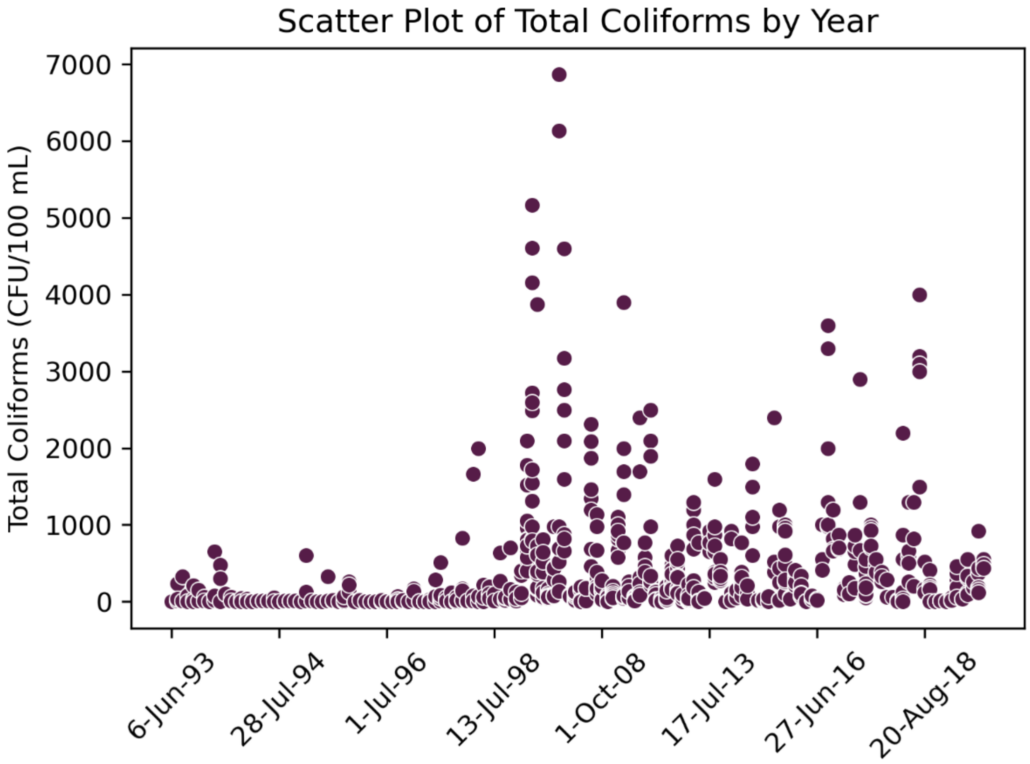
Analyzing the data through scatter plots, we discovered significant outliers in the Total Coliforms data. This observation is further underscored by the summary statistics of the total coliform data, which show considerable variance. The extreme range between the lowest recorded value of 0 and the highest value of 6870 CFU/100 mL signifies the presence of extreme outliers or a highly skewed data set.

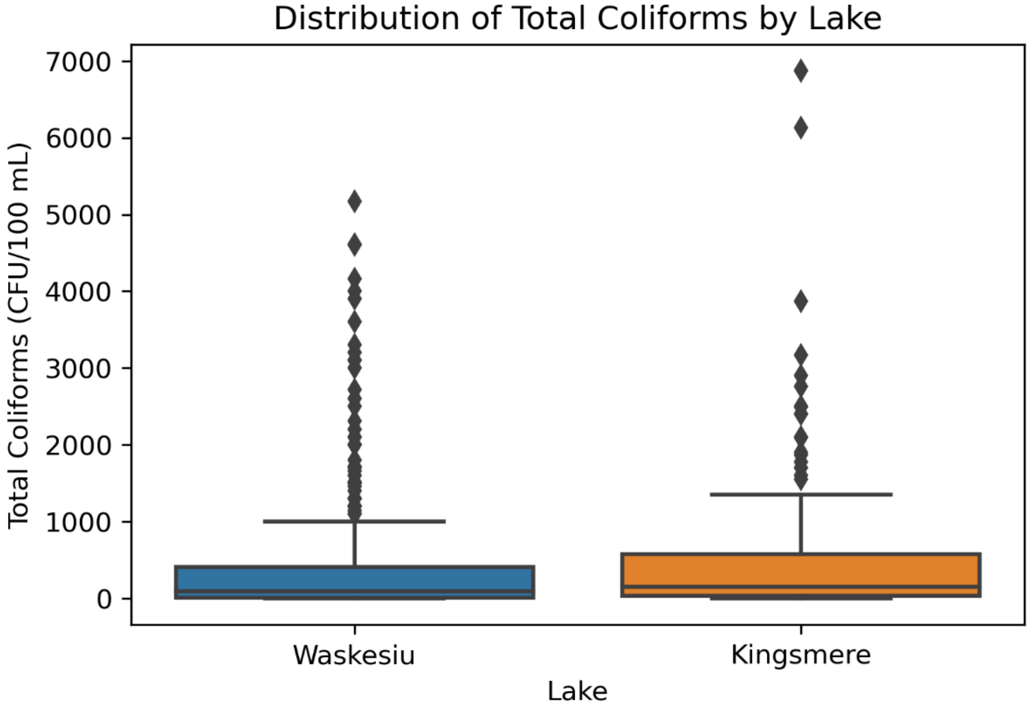
This extensive variability is echoed in the boxplot, which demonstrates a widespread of values, pointing to considerable differences in the levels of coliforms across various data points.

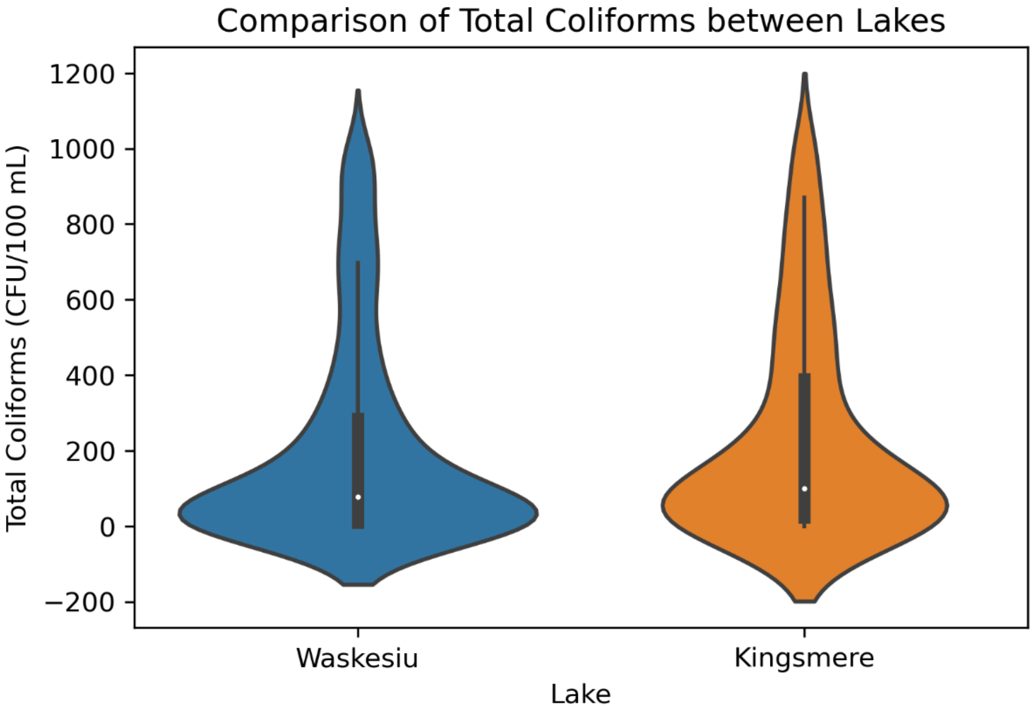
These disparities in Total Coliforms levels can be attributed to a host of factors, such as different sampling locations, water sources, seasonal changes, or human activities. To enable a more accurate comparison of coliform distributions between Kingsmere lake and Waskesiu lake, extreme values above 1000 CFU/100 mL were first removed. Following this, a violin boxplot was employed to visualize the distributions.

After excluding the outliers, the density distribution of Total Coliforms in both Waskesiu and Kingsmere lakes displays noticeable similarity. The 25th percentile and median values show considerable alignment between the two lakes. Nonetheless, Kingsmere lake presents a slightly elevated value for the 75th percentile.

This data suggests that these two bodies of water may share similar levels of contamination from coliform bacteria, as evidenced by the comparable lower quartile and median values. However, the marginally higher 75th percentile value for Kingsmere lake could hint at a slightly elevated level of contamination in the upper range, which may be attributable to specific influencing factors such as local pollution sources or fluctuations in environmental conditions.

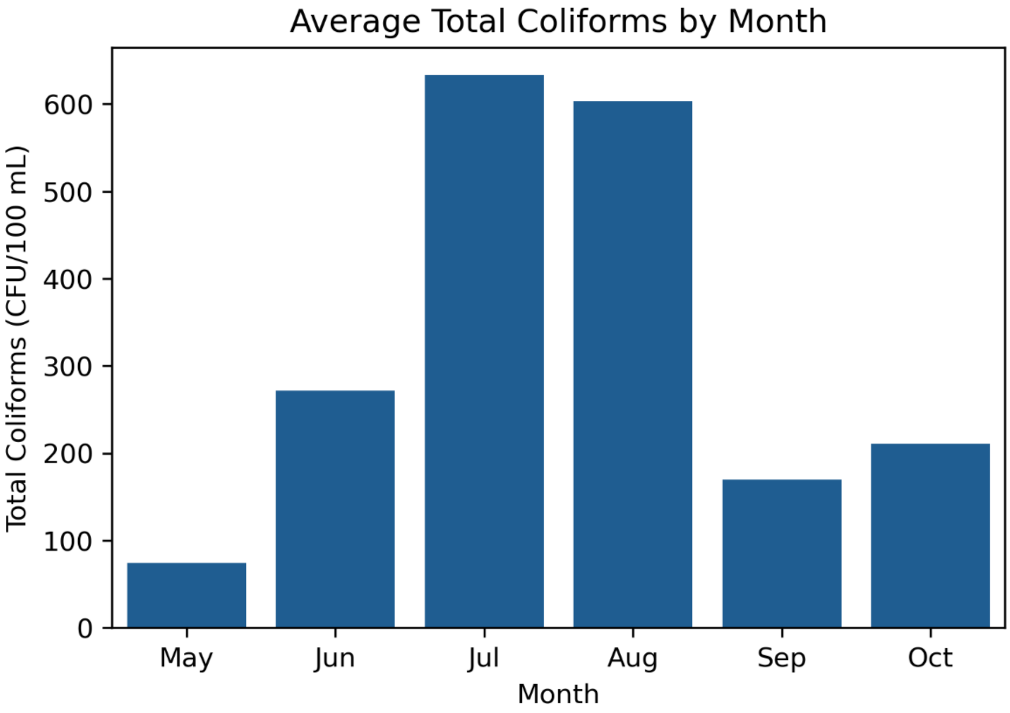

Scatter Plot of Total Coliforms by Year
Total Coliforms (CFU/100 mL)
7000
6000
5000
4000
3000
2000
1000
0
6-Jun-93
28-Jul-94
1-Jul-96
13-Jul-98
1-Oct-08
17-Jul-13
27-Jun-16
20-Aug-18


Distribution of Total Coliforms by Lake
Total Coliforms (CFU/100 mL)
7000
6000
5000
4000
3000
2000
1000
0
Waskesiu
Kingsmere
Lake

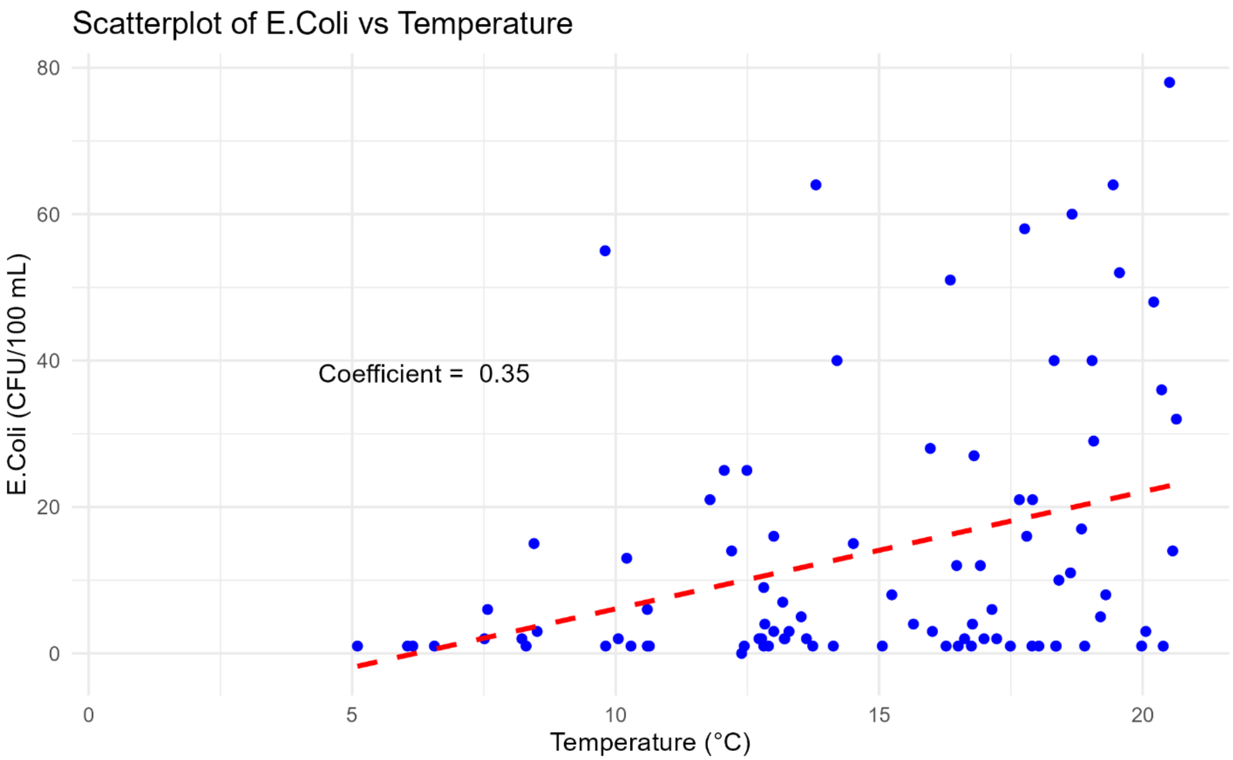


Guiding Question 5: How do contamination levels in the lake vary across different months, and what environmental factors might be contributing to these changes?

The bar plot of average total coliforms by month provides a visual comparison of the average levels of total coliforms for each month, allowing us to identify which months require closer attention to microbial levels in the water samples. It can be observed that the average total coliform levels are highest in July and August, while they are lowest in May. These variations could potentially be attributed to specific environmental conditions, temperature, or other factors. It is important to note that these observations are based on the provided dataset, and other factors such as sample size, data collection methods, and sample sources could also influence the results.



Following this, I explored the relationship between Escherichia coli (a key constituent of the total coliform group) and temperature to expose any hidden patterns or trends. Examining this relationship yields essential insights into how temperature variations impact the prevalence and proliferation of E. coli in the water samples. E. coli serves as an immediate indicator of fecal contamination in water, with high counts suggesting recent fecal contamination that could pose health risks upon human exposure.



Upon conducting a Pearson correlation test, the correlation coefficient between E. coli and Temperature stands at 0.35. This figure signifies a moderate positive correlation between the two, suggesting that as the temperature ascends, there's a corresponding increase in E. coli counts, albeit to a certain degree.

Like many bacteria, E. coli may find warmer conditions more conducive for multiplication, leading to a rise in its count. This phenomenon elucidates the positive correlation observed between E. coli and temperature in the data.

Conclusion

Through careful analysis of the "Water Quality - Prince Albert" dataset, we unveiled the relationships between chemical attributes in the lake water, exploring how these interactions affect overall water quality. This exploration led us to crucial insights about the correlations among chemical attributes such as pH value, ammonia, dissolved oxygen, temperature, etc.

The use of correlation analysis and visualizations, such as scatterplot matrices, allowed us to decipher complex patterns, providing a comprehensive view of water quality trends. By transforming abstract data into meaningful information, we deepened our understanding of these vibrant aquatic ecosystems.

Yet, as much as we learned, our study also reminded us that water quality is a dynamic, multifaceted phenomenon, influenced by a multitude of factors that are not static but in constant flux. As such, monitoring water quality is an ongoing process, and our data-driven narrative, though enlightening, represents only a snapshot of these ever-changing stories.

In conclusion, this exploration underscores the critical importance of continual data collection and analysis in our collective mission to understand, preserve, and cherish our natural water bodies. This data story is but one chapter in the never-ending book of the intricate life of water, each drop holding tales of ecological, chemical, and biological interactions that are yet to be discovered.

Reference

GOVERNMENT OF CANADA, 2023. WATER QUALITY - PRINCE ALBERT. AVAILABLE AT: HTTPS://OPEN.CANADA.CA/DATA/EN/DATASET/A0096EA6-6CE0-4007-B43E-9D535CD1A32C. CONTAINS INFORMATION LICENSED UNDER THE OPEN GOVERNMENT LICENCE – CANADA.