E. coli Testing: Safeguarding Public Health*

Tianrui Fu

September 26, 2024

Swimming in beaches with high E. coli levels can lead to gastrointestinal issues, rashes, and infections, particularly affecting young children, the elderly, and those with weakened immune systems. In 1998, the implementation of the Ministry of Health Beach Management Protocol project led the Toronto department to start collecting beach water quality data in 2007, aiming to reduce the incidence of waterborne diseases in the population. This paper prove that the mean E.coli is decrease by time and site location. Higher than 50% time and site location satisfy the E.coli standard.

Table of contents

*Code and data are available at: https://open.toronto.ca/dataset/toronto-beaches-water-quality/.

1 Introduction

The increase in E. coli concentration is a serious problem for humans. Elevated concentrations make beach water unsafe, hindering safe human use. Although some strains of E.coli are harmless, pathogenic microorganisms may still be present. In cases of increased concentration, healthy individuals among those infected can recover on their own; however, vulnerable populations such as children, the elderly, and those with compromised immune systems may suffer more severe illnesses, including general gastrointestinal discomfort, rashes, ear and eye infections, and in severe cases, bloody diarrhea, kidney failure, and death. The rise in E.coli concentration is caused by various factors, such as the runoff of animal feces carrying pathogens from land or sewage systems, aging sewage pipelines, and wastewater infrastructure issues. Considering the rapid growth of E. coli and the associated public health risks, real-time monitoring of beach water quality is essential.

According to the requirements of the Ministry of Health Beach Management Protocol (January 1, 1998), the Beach Water Sampling Program for the City of Toronto was implemented to reduce the incidence of waterborne diseases. Toronto's beaches have been certified under the Blue Flag program. From June to September each year, the relevant departments in Toronto collect water samples daily from all regulated beaches in the city to test for E.coli bacteria concentrations.

The number of E.coli in freshwater is determined by counting the number of yellow and yellow-brown colonies that grow after placing a 0.45-micron filter on m-TEC medium and incubating it at 35.0ºC for 22-24 hours. The water quality standard for E.coli in Ontario and federally is 200 E.coli per 100 milliliters of water, while Toronto's beach water quality standard is 100 E.coli per 100 milliliters of water. This article utilizes 22,000 data points collected by government departments from June to September from 2007 to 2024 to assess whether the E.coli is much more than the Toronto water quality standard. Meanwhile, it try to prove if the E.coli is influenced by the change of time and site location.

2 Data

2.1 Raw Data

The data used in this paper is from Open Data Toronto and download by the Gelfand (2020). The Toronto Public Health (2024) is used to analyze whether the E.coli value in the beach during June and September is satisfy for people to swim. All the data analysis is through the R Core Team (2023) and the completed by following packages Wickham et al. (2019), Wickham et al. (2023), Robinson (2023), Wickham (2016), Allaire (2023), Peters (2023) and Xie (2023). The dataset is published by Toronto Public Health and is part of The Beach Water Sampling Program, which collects E. coli values from two different beaches. The data is updated daily, and the data used for this analysis spans from June 3, 2007, to September 8, 2024. The raw dataset contains 21,882 water quality samples testing for E.coli concentrations. The dataset also includes beach itendifiers, names, sample site names, collection dates for each sample, and geographic coordinates (latitude and longitude).

2.2 Cleaned Data

In the raw data provided by Toronto Public Health, there were missing values (NA). During the data cleaning process, rows containing these NA values were completely removed to ensure that the NA values would not affect the analysis output and to simplify the analysis. Since the raw data also contained some particularly large outliers, those were also removed. The cleaned data only includes the necessary columns for analysis, such as collection date, E.coli, site name, and beach name. Figure 1 shows the cleaned data samples, listing the results of tests conducted at different locations on the same day.

```
Attaching package: 'kableExtra'


The following object is masked from 'package:dplyr':

    group_rows
```

Beach_Name	**Date**	**E.coli**	**Site_Name**
Marie Curtis Park East Beach	2024-09-01	230	29W
Marie Curtis Park East Beach	2024-09-01	220	33W
Marie Curtis Park East Beach	2024-09-01	240	32W
Marie Curtis Park East Beach	2024-09-01	200	31W
Marie Curtis Park East Beach	2024-09-01	200	30W
Sunnyside Beach	2024-09-01	910	18W
Sunnyside Beach	2024-09-01	210	17W
Sunnyside Beach	2024-09-01	170	20W
Sunnyside Beach	2024-09-01	80	21W
Sunnyside Beach	2024-09-01	400	19W
Sunnyside Beach	2024-09-01	360	22W

Figure 1: Bills of penguins

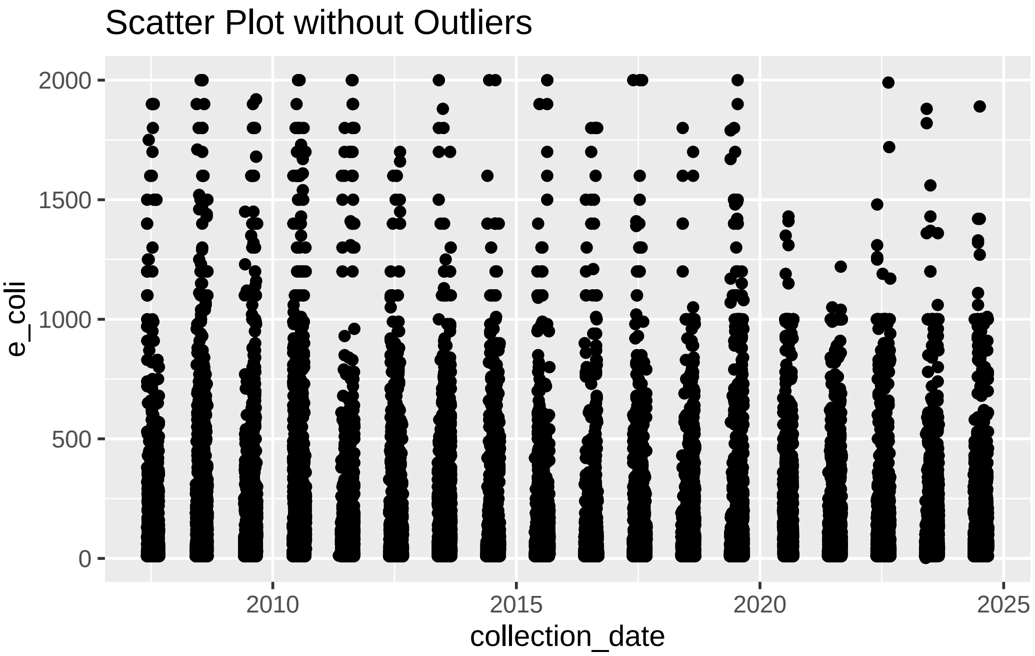

Figure 2: Relationship between wing length and width

2.3 Basic Summary of cleaned Dataset

Figure 1 shows the attributes of the data, with most E.coli values being below 500, meaning there are 500 E.coli per 100 ml of water. However, the scatter plot also reveals that the data from 2020 stands out compared to other years, but a more precise graphical output is needed for a better understanding of the dataset. Therefore, in Table 2, a bar chart was used to show the total amount of valid data collected each year. It is clear that due to the COVID-19 pandemic, the data collected by Toronto Public Health in 2020 is significantly less compared to other years. Figure 2 also provides a calculation of the distribution of collected E.coli data. Based on Canada's water quality standard of 200 E.coli per 100 ml, about 78% of the data falls within this standard, meaning that approximately 78% of the days from 2007 to the present were safe for beach activities. However, based on Toronto's stricter standard of 100 E.coli per 100 ml, only around 60% of the data meets this standard, meaning that about 60% of the days were safe for beachgoers. However, considering the additional variables of location and date, it is not yet possible to draw a reasonable conclusion about the overall water quality. A more in-depth analysis will be conducted in Section 3.

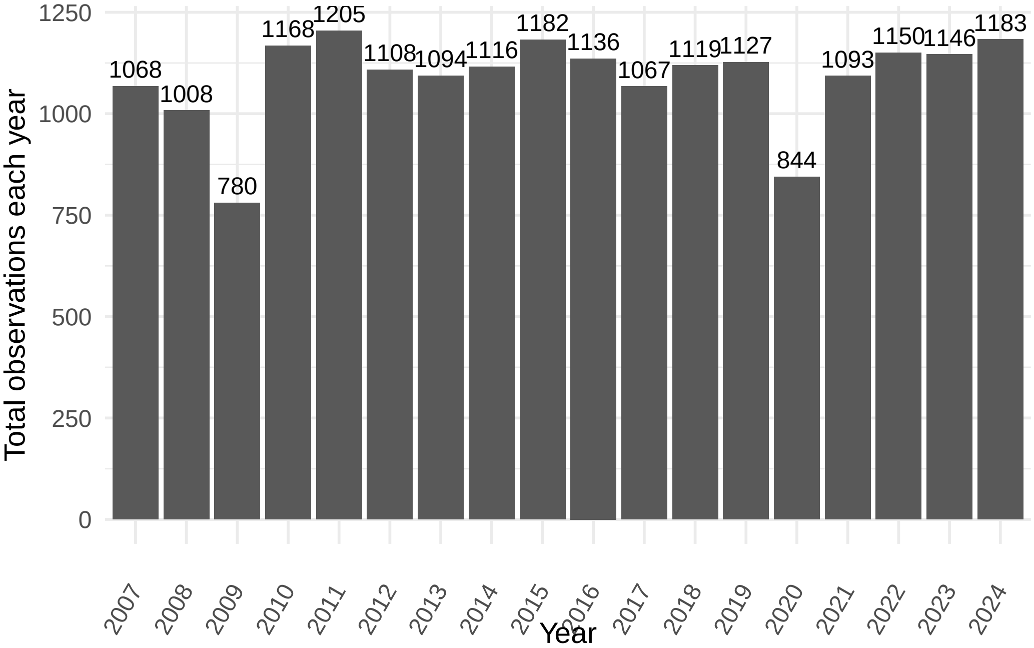

Figure 3: Number of observations by year

2.4 Dataset justification

The reason for choosing this dataset is that, during this summer, there were news reports about several incidents of people defecating on Toronto beaches, with even photos of the feces

Table 2: Portion of E.Coli levels in different ranges

E.Coli Levels	Sample Count	Portion
0-50	8588	43.82974
50-100	3609	18.41890
100-200	3009	15.35674
>200	4388	22.39461

Figure 4: Bills of penguins

circulating. This caused some panic among people in Toronto who were planning to visit the beach. To verify whether the beach closures were indeed caused by these incidents and to reduce potential bias, as well as out of personal interest, this paper uses data published by Toronto Public Health. The dataset spans 18 years, from 2007 to the present, containing E.coli test results for beach water quality. It effectively reveals the trends in water quality over the years and helps to project future trends, reducing doubts about the validity of the analysis due to a small data sample.

3 Result

3.1 Studying the relationship between time and E. coli concentration

Though section 2.3 finds that about 60% of the beaches meet the Toronto Beach water quality standard, we need to further explore the changes in the average E.coli concentration over the years and whether the average values truly reflect that the water quality meets the standard more than half of the time. Figure 3 shows the yearly average E.coli values. Through Figure 4, we can observe that none of the years had an average E.coli concentration that meets the stricter Toronto Beach standard, only four years exceeded the Canadian water quality standard. The average E.coli concentration was at its lowest in 2016, approaching 100 E.coli/100ml (Toronto Beach standard), and peaked in 2008, nearly reaching 300 E.coli/100ml. Over time, while there was a slight increase in the average E.coli concentration from 2016 to 2020, the overall trend shows a decline, stabilizing around 150 E.coli/100ml. This trend suggests that while water quality has improved over the years, it still fluctuates and does not consistently meet the Toronto Beach water quality standard.

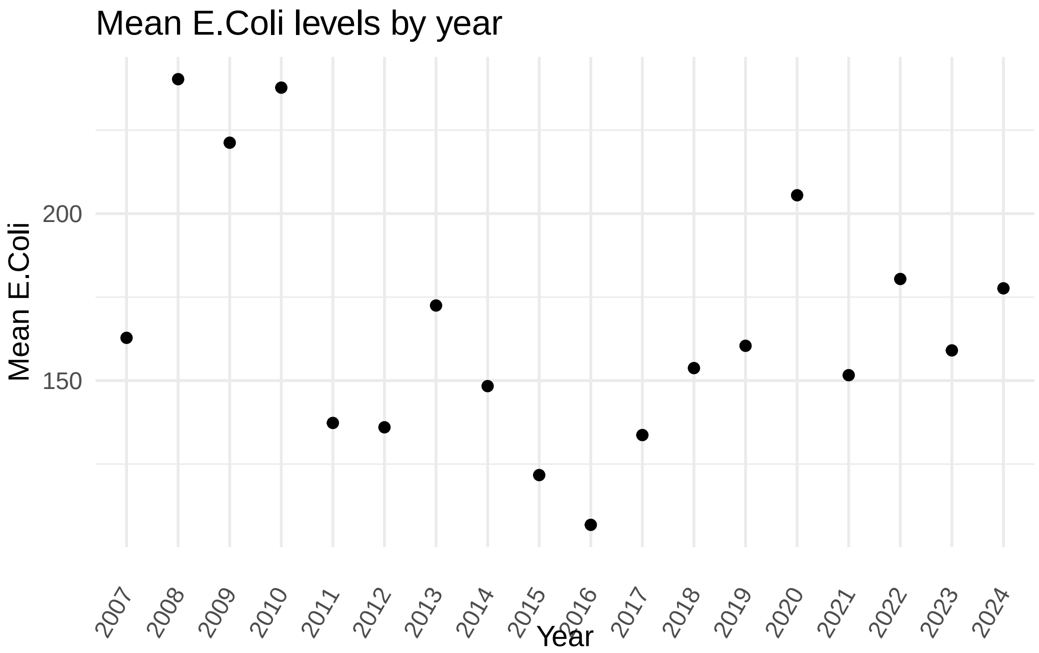

Figure 5: Bills of penguins

3.2 Examing the portion of E.coli higher than the standard level by year

To further understand the data, we calculated the proportion of E.coli levels exceeding the Toronto Beach standard each year, providing a clearer view of water quality trends at Toronto beaches. From Figure 4, we can see that the highest and lowest average E.coli concentrations, observed in 2008 and 2016 respectively, correspond to the highest and lowest proportions of exceedance. In 2008, over 50% of the samples exceeded the Toronto standard, while in 2016, this proportion dropped to as low as 20%. Additionally, the overall proportion of exceedances has been declining over time and has stabilized in recent years.

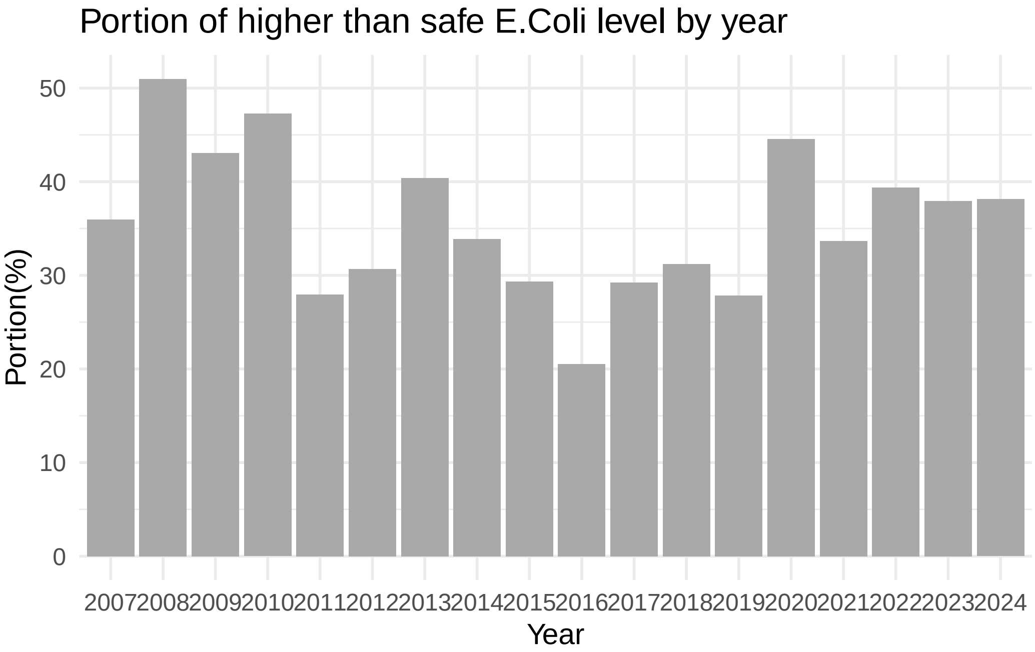

Figure 6: Bills of penguins

3.3 Investigating the relationship with E.coli and site location

By observing the trend of average E.coli levels across testing sites, as shown in Figure 5, we can gain further insights into the data. The E.coli exceedance percentages for each site are presented in Table 2. Sites 17W-23W correspond to Sunnyside Beach, while sites 29W-33W are located at Marie Curtis Park East Beach. Although all of these sites exceed Toronto's beach standards, the E.coli levels at sites 17W-23W are generally lower than those at sites 29W-33W. The lowest E.coli levels were recorded at the 20W testing site. Table 2 also shows that the data from site 23W is noticeably lower than the other sites, making it less significant for reference. The exceedance percentage at site 20W is the lowest, at 35.09%, and the exceedance rates for most sites range between 30% and 40%.

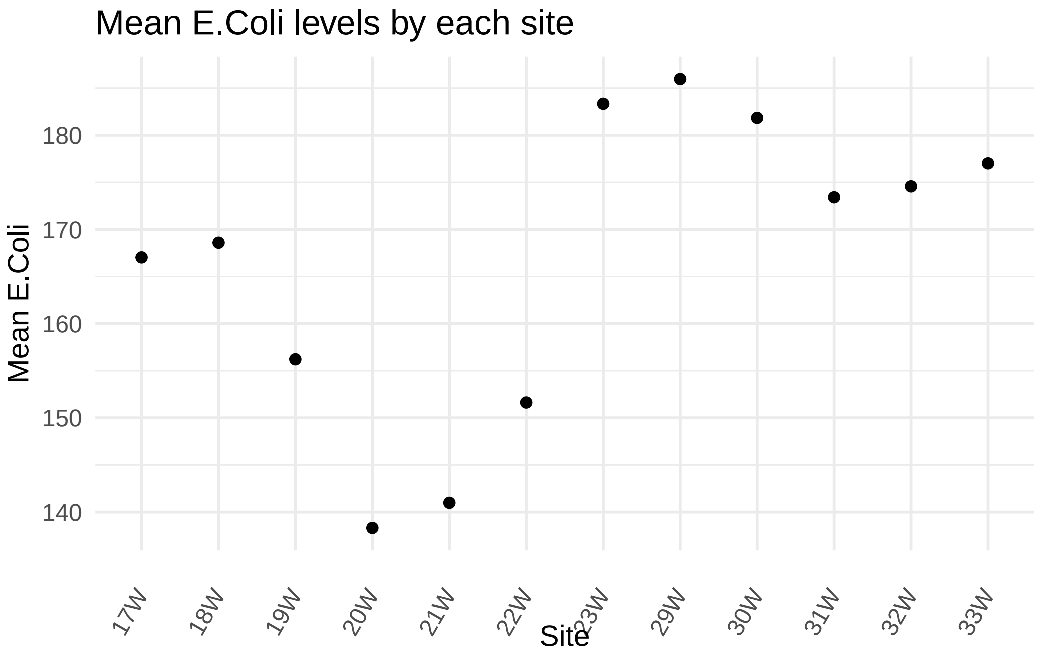

Figure 7: Bills of penguins

Table 3: Portion of E.Coli levels above safe level in different site

Site name	**Total Samples**	**Above 100 count**	**Portion(%)**
17W	1778	690	38.80765
18W	1774	681	38.38782
19W	1778	628	35.32058
20W	1782	554	31.08866
21W	1783	564	31.63208
22W	1770	625	35.31073
23W	460	197	42.82609
29W	1691	610	36.07333
30W	1693	609	35.97165
31W	1697	577	34.00118
32W	1696	590	34.78774
33W	1692	596	35.22459

Figure 8: Bills of penguins

4 Discussion

4.1 The time influence E.coli

From a time perspective, based on the information provided in the previous section's graphs and tables, there have been significant changes in Toronto beach water quality between 2007 and 2024. Particularly in 2008, E.coli levels peaked, with an average concentration approaching 300 E.coli/100ml. After that, the E.coli levels showed a downward trend and became more stable. This result may be related to several factors mentioned in the introduction, such as animal feces, human activities, and the aging of the stormwater drainage system. Based on the evidence provided by the trend in Figure 3, there is a clear negative correlation between the average E.coli levels and time. However, since the average value alone cannot be used as a final indicator to judge the trend, we can refer to the percentage of exceedance bars in Figure 4. The exceedance rate of E.coli in 2008 exceeded 50%, indicating that water quality from June to September of that year was poor, leading to reduced activities and potential impacts on health and hygiene. In contrast, 2016 had the lowest E.coli levels, being closest to Toronto's beach water quality standard of 100 E.coli/100ml, with an exceedance rate of only about 20%. This may be related to the government's efforts to improve the beach environment and favorable weather conditions from June to September. Data from 2020 was limited due to collection restrictions caused by the COVID-19 pandemic, but there was no significant increase compared to other years, indicating that the overall water quality remained stable. Over time, both graphs show that after two peaks, E.coli levels have generally declined and stabilized.

4.2 Whether the Site location influence E.coli

In addition to time factors, evaluating whether the test locations affect E.coli levels can further reveal water quality differences across various measurement sites. Based on Figure 5, we can clearly see that the average E.coli levels vary significantly depending on the test location. E.coli levels at Sunnyside Beach are generally lower, indicating relatively better water quality in this area, especially at the 20W test site, where the exceedance rate is only 35.08866%, according to Table 3. On the other hand, at the Marie Curtis Park East Beach, E.coli levels are relatively higher, particularly at the 29W and 33W test sites, where exceedance rates are close to 40%. However, although differences exist between test sites, the average levels at all sites still exceed Toronto's water quality standards, though they are below Canadian water quality standards. Since average concentration alone cannot fully determine whether the water quality of an area is non-compliant, the calculation of exceedance rates in Table 2 helps reveal that only a small portion of water quality is non-compliant. The high exceedance values contribute to the elevated averages. By calculating exceedance rates by test location, the analysis becomes more insightful and evidence-based, providing more convincing results.

4.3 Conclusion

Overall, through the analysis of E.coli data from 2007 to 2024 in Toronto beaches, we have uncovered trends in concentration changes over time and across different locations. While there has been an improvement in overall water quality, some years and measurement sites still exhibit high E.coli concentration exceedances. Additionally, the influence of time on E.coli levels cannot be fully confirmed due to the variability of factors other than data and the uncertain governmental interventions.

4.4 Limitations and Next Step

Although the visualizations from the graphs and tables provide an overview of E.coli concentrations in Toronto beaches over the past 18 years, there remain limitations and the need for more detailed data to address these issues. The article outlines the overall concentration trend, but Toronto Public Health did not mention whether interventions were conducted in the middle years, which may explain the overall downward trend in concentration. To avoid the impact of outliers on average value plots, this study excluded large outliers at the start of the analysis. However, TPH did not specify whether all data were accurate, so the analysis after removing some data may lack reference value.

Additionally, while there were many measurement points, the beach locations were limited to only two, narrowing the dataset's scope. As a result, the data may not fully represent the trends in Toronto's beach water quality, leading to potential inaccuracies in observations and analysis. Future analyses could increase the number of beaches to provide more comprehensive data and improve credibility. Moreover, by controlling for other variables and confirming outliers, tracking E.coli levels from the same source over time or by location would provide deeper insights into beach water quality, helping the government take effective intervention measures.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

Checking the convergence of the MCMC algorithm

References

Allaire, J. J. 2023. *Here: A Simpler Way to Find Your Files*. https://cran.r-project.org/package=here.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Peters, Dominik. 2023. *kableExtra: Construct Complex Table with 'Kable' and 'Knitr' Packages*. https://cran.r-project.org/package=kableExtra.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://cran.r-project.org/package=janitor.

Toronto Public Health. 2024. *Toronto Beaches Water Quality*. https://open.toronto.ca/dataset/toronto-beaches-water-quality/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley et al. 2023. *Dplyr: A Grammar of Data Manipulation*. https://cran.r-project.org/package=dplyr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://cran.r-project.org/package=knitr.