Lead Contamination Testing: Is water in Toronto safe for drinking?*

Water in Some Districts is more Contaminated than others

Amy Jin

September 26, 2024

The City of Toronto offers a free, non-regulated Residential Lead Testing Program, where residents collect and submit water samples. This paper examines how time and district affects the lead concentration with respect to safe lead concentration guidelines. Quality of drinking water is an important public health issue, since too much abosorbtion of lead might have detrimental effects especially for young children. It was found that the lead concentration does not have noticeble seasonality or autocorrelation, some areas in toronto tends to have higher lead concentrations than the rest.

1 Introduction

Lead contamination in drinking water poses significant public health risks, particularly for vulnerable populations such as children and pregnant women(Government Canada).

This paper takes in raw data from the City of Toronto(Open Data Toronto), where they implemented a non-regulated program called City's Residential Lead Testing Program. More information on this program can be explored at: www.toronto.ca/leadpipes. The results respond to the concerns about lead exposure in the city of Toronto.

While significant research has extensively documented the public health impacts of lead exposure, particularly focusing on its toxic effects on vulnerable populations like children and pregnant women (Levallois et al., 2018)(Jarvis & Fawell, 2021), a critical gap remains in understanding how lead concentrations in urban drinking water systems fluctuate both spatially and temporally.

*Code and data are available at: https://github.com/aj3616/Lead-Testing-Program-in-Toronto

This paper pre-processed the raw data and the cleaned data was used for visual representations and statistical analysis includes outliers analysis, time series analysis and linear regression analysis.

The results from the analysis show that there is no clear seasonality, which refers to recurring patterns or trends in data that repeat over specific time intervals, and no significant autocorrelation, which indicates a lack of correlation between current and past lead concentrations. In other words, future lead concentrations do not depend on previous lead levels, and there is insufficient evidence to suggest that certain months consistently exhibit higher trends of lead concentrations compared to others.

However, some district does have higher trend of lead concentration. There are clusters of outliers within some specific district and some district have higher mean lead concentrations than other. It is safe to say that geographic locations and community water maintenance have an effect on water safety.

This topic is important because even low levels of lead exposure can have serious health consequences, including developmental delays, cognitive impairments and even detrimental effects(Government Canada).

We use R Core Team (2023) for data analysis, Gelfand (2022) for collected data set, and Wickham et al. (2019) for directions and template.

The remainder of this paper is structured as follows. Section 2 Data, further explains the data set we are working with and the prepossessing done for further analysis. A discussion with the measurement used when collecting the data is also included. Graphs and tables were used to show important features of the data set. Section 3 Discussion, talks about, firstly, to what extent will Lead Concentration influence Health and how much concentration will be categorized as "dangerous". Secondly, the paper went into a detailed analysis of how districts affect lead concentration. Graphs and outliers analysis was used to present the result. Thirdly, the paper performs Times Series analysis to analyze the lead concentration over time. Lastly, there is a subsection discussing weakness of the paper and some further steps to further support the question.

2 Data

2.1 Tools Used

#TODO: change library citation R (R Core Team 2023) was used for the data analysis and the productions of all graphs and tables. The libraries tidyverse (Wickham et al. 2019), here (J. B. Kirill Müller 2020), dplyr (Hadley Wickham 2023), tibble (R. F. Kirill Müller Hadley Wickham 2023), janitor (Sam Firke 2023), ggplot2 (Wickham 2016), and knitr (Xie 2023) were also used for organizing data and performing analysis.

2.2 Data Set and Context

The data set "Non Regulated Lead Sample" by Open Data Toronto is to address the concerns Toronto residents have for their drinking water. The data set contains information on lead concentrations in water samples. Each row is from a water sample, with corresponding information on the date, partial postal code where the sample was provided, and the main variable of interest, lead concentration in parts per million (ppm).

Lead exposure is an important public health issue. High levels of lead in drinking water can pose severe health risks, particularly to vulnerable populations.(WHO) Chronic exposure to lead can result in developmental delays, neurological damage, and a range of other health problems.(WHO) Tracking lead concentrations in water is crucial for ensuring safe water quality and identifying areas requiring intervention to effectively manage health risks.

There are similar similar data sets that addresses the problem of hard water contamination. One example is in state of Michigan, USA, MIDHHS collected similar sample with months, years and lead concentrations(MIDHHS). There are also other data set that analyze a different metal in water. However, this data set is used because lead is a very common metal in hard water with severe damage to public health when overdose, and being a resident in city of Toronto, this should be a concern that needs attention.

2.3 Variables

There are five variables in total: _id, Sample Number, Sample Date, PartialPostalCode and Lead Amount. While the first two are keys of this data set for identification purpose, the rest are our variable of interest:

Date: Represents the date on which the water sample was collected, can be used to detect trends and seasonal variations.

Partial Postal Code: A truncated form of the postal code indicating the general area where the water sample was taken, can be used to identify potential geographical clusters.

Lead Concentration (ppm): The measured amount of lead in the water sample, expressed in parts per million. This is the primary variable of interest, as it directly relates to public health risks associated with lead exposure. We can also further analyze combinations of lead concentration with other variables.

Other than the given variables, there are three additional variables constructed during analysis: Months: Extracted from the Date variable to analyze seasonal patterns in lead concentration. Years: Extracted from the Date variable to evaluate trends over time and aggregate the data for yearly analysis. District (First Two Digits of Postal Code): Created from the Partial Postal Code to group data into broader geographic areas within the city, aiding in the spatial analysis of lead concentration.

2.4 Cleaning

For the purpose of data analysis the raw data set went through some pre-processing: NA Values: All missing (NA) values were removed to ensure accurate analysis. Lead Concentration: The Lead Amount (ppm) variable was cleaned to remove any anomalies and ensure that values were properly numeric. Date Variable Type: The Date variable was converted to Date type in R to facilitate temporal analysis and aggregation by months and years.

After cleaning the data looks like:

Date(Date): A Date type, ranges from 2014-01-01 to 2024-08-21, covering over a decade of water sampling data.

District(chr): A Character type, represents various regions within the city of Toronto based on the first two characters of the Partial Postal Code.

Lead Concentration(num): A num type, which represents the summary statistics of lead concentration typically include measures like the minimum, maximum, mean, median, and standard deviation of lead levels across the data set.

2.5 Summary

Table 1 is a summary table for num type variable lead amount. We can see that the minimum and maximum value differs by 168,800 times, which is very significant. Therefore, we can assume that there are water samples that are very clean in terms of lead concentration and there are samples that are contaminated. We can also see that the mean and median differs a lot. Which means that the distribution of this data set is highly right skewed. It also shows evidence that there are many outliers that will elevated mean lead concentration level in this data set.

Table 1: Distribution Statistics Measurements of Lead Amount

Summary of Lead Amount (ppm)

Statistic	Value
Min.	0.000050
1st Qu.	0.000125
Median	0.000473
Mean	0.006401
3rd Qu.	0.001480
Max.	8.440000

2.6 Relationships

These relationships are upon further discussion in Section 3.

Lead Concentration and Time (Months/Years): Analyzing lead levels over time can reveal temporal trends, such as reductions in lead levels following interventions, seasonal spikes, or gradual increases over certain periods.

Lead Concentration and Location (District): By examining lead concentrations across districts, one can identify areas with consistently high or low lead levels, indicating possible regional differences in water quality or pipe infrastructure.

Seasonality: The monthly variable can help detect seasonal patterns, such as increased lead concentration during warmer months due to factors like water temperature affecting pipe corrosion.

Temporal Trends: By grouping data by year, any long-term trends, such as a decrease in lead levels due to policy interventions, can be detected.

2.7 Measurement

In this dataset, we are interested in the concentration of lead in drinking water across different regions of a city and over time. The measurement process transforms a real-world phenomenon—lead contamination in water—into quantitative entries in a dataset that can be analyzed for trends.

Lead concentration: Measured in parts per million (ppm), indicating the amount of lead in the water relative to its volume. This choice of unit is small enough to capture low but potentially harmful concentrations and is standard in water quality testing. This transformation from the chemical presence of lead to a numeric value is a key aspect of measurement.

Sample Date: The date the sample was taken is recorded to allow for analysis over time, detecting patterns.

Partial Postal Code: Postal code represents where the sample was collected, allowing comparisons between different geographic areas. This made it easy to get a result using data analysis to public health control organizations.

Some challenges are that this is a non-regulated program, therefore the measurement process is not controlled, decreasing accuracy and precision. However, this is minimized by providing a detailed instructions with the sample kit used to collect the water samples and the deviation will only affect Lead concentration. Date and district are accurate.

Choices about measurement such as the instruments used, units of measurement, and protocols for sample collection—affect how well the data set represents the real-world phenomenon of lead contamination.

3 Discussion

3.1 Lead Concentration influencing Health

3.1.1 WHO and Canadian federal guidance on lead concentration in water

The main variable we are examining is the Lead Amount in ppm. The distribution of the data is highly right skewed with most of its data being very small, as shown in (Figure 1), most data are almost 0 with outliers which are mostly less than 2 ppm and 4 of them ranging from 3 to 8.44 which is the maximum value as shown in the summary Table 1. We can also verify our observation since the summary Table 1 shows that 75% of the data is below 0.001480. Government of Canada had specified that the maximum acceptable concentration (MAC) of drinking water is 0.010mg/L(Government Canada), which converts to 0.010011423 part/million(unitconverters.net), this is shown by the red line in the figure 1(Figure 1).

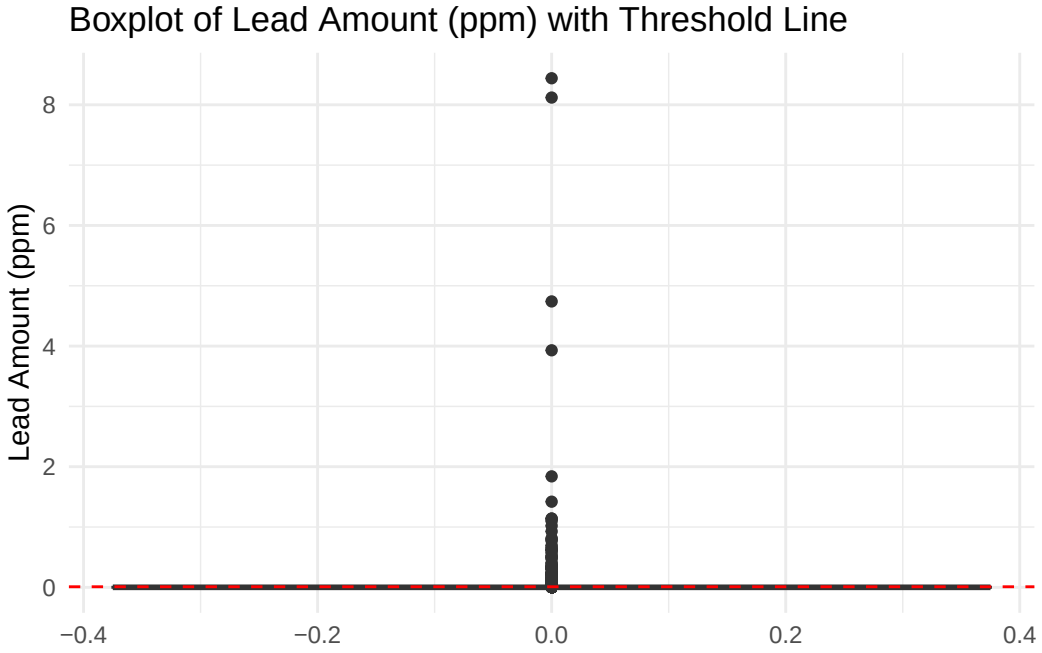

Figure 1: Boxplot showing the lead concentration above the saftey threshold

3.1.2 Percentage of data above this threshold

From calculations by R, there are 97.26% of the data collected indicating that the lead amount in the water is safe. As shown in (Table 2), there are only 280 counts of water sample collected that is above MAC, which takes up only 2.74% of the whole data set. Even though this seems

like a good result, those 280 measurements still effects people and their family's health since those lead concentrations will have negative effect on health(WHO)(Canadian Government).

Table 2: Number and Percentage of data above MAC

Summary of Lead Concentration Above Threshold

Metric	Value
Count of Samples Above Threshold	280.00
Percentage of Samples Above Threshold	2.74

3.1.3 plot of the data above this threshold

To visualize those data above MAC, data below MAC and above 2 were filtered out, because from (Figure 1), we can see that there are only 4 data points above 2, therefore, with that in mind it is safe to temporarily ignore to get a scale that will better show the most data points above MAC. A scatterplot was used with x axis being date and y axis being lead concentration give a graph that spreads out lead concentrations instead of on a 1-dimensional line. From (Figure 2) we can see that there are clusters of data points below 0.25 and for higher lead concentrations, there are less data points and more spread out than those below 0.25.

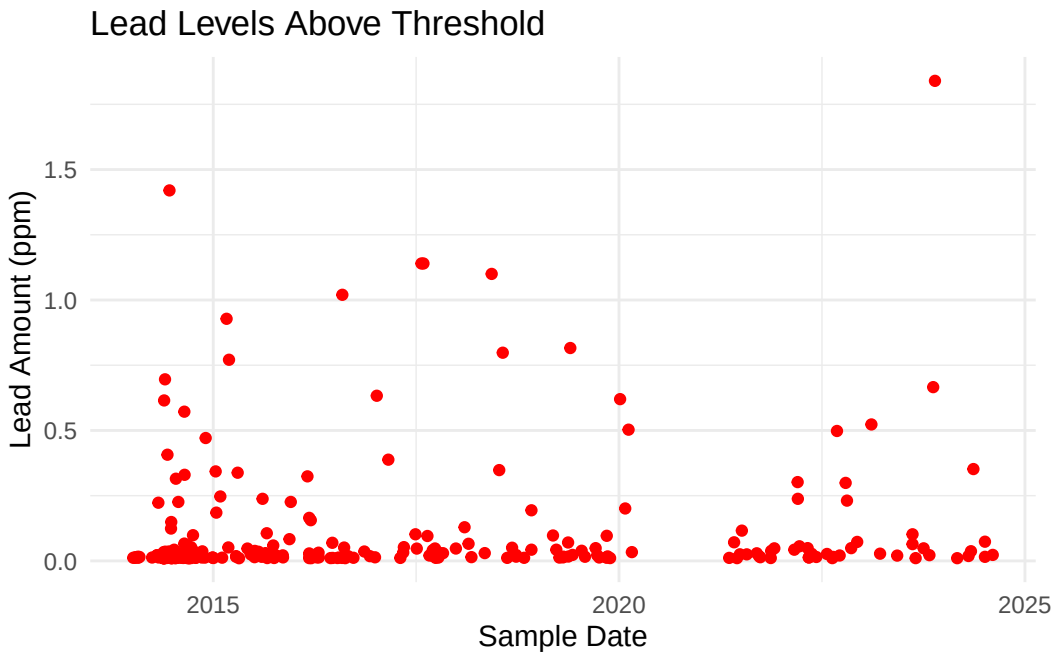

Figure 2: Data between 2 and MAC representing with Date

3.2 How districts affect Lead Concentration

3.2.1 graphs of postal code vs lead concentrations

Nine boxplot (Figure 3) each representing a district, which is the constructed variable from the first two letters of partial postal code. Again, most data points are around 0, and the outliers streched the scale so that this way we can easily analyze the outliers. The districts containing outliers were also shown in (Figure 3). We can see that the "M4" has two data points above 8 ppm, as well as a cluster between 0 and 1 ppm. "M6" also have many data between 0 and 2ppm. It is safe to assume that these two districts have water sources with high lead concentrations.

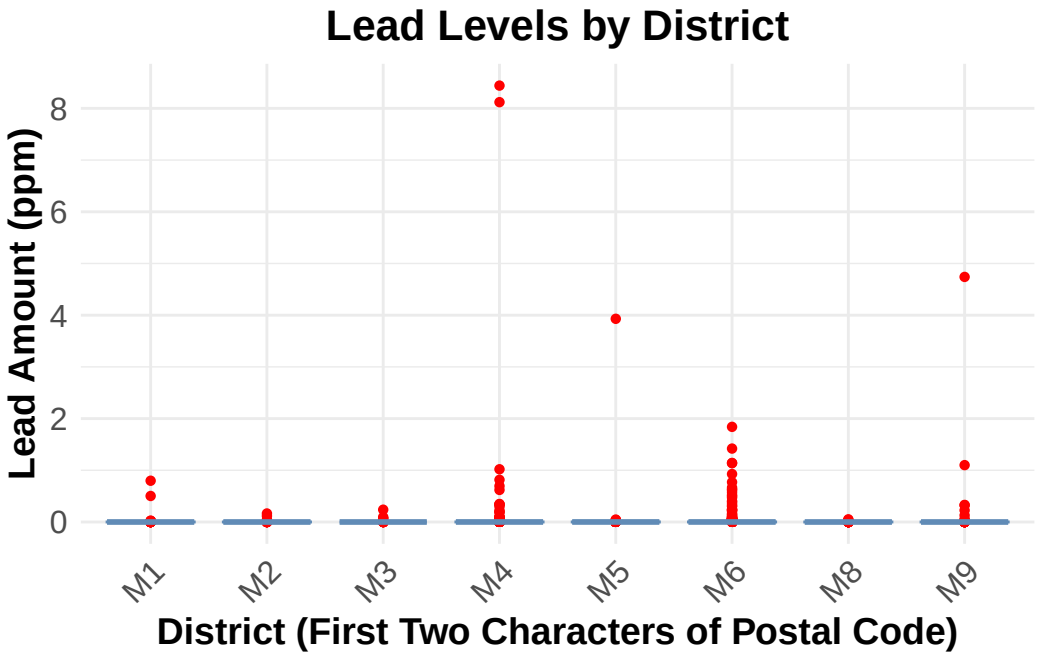

Figure 3: Boxplot Showing Outliers categorized by district

3.2.2 extract outliers above the threshold

We will verify which district have a higher overall lead concentration by calculating the mean lead concentration when still categorize the partial postal codes by only the first two numbers. Since outliers might influence the mean, especially when in our dataset we are dealing with median 0.000473 and outliers as high as 8ppm, the outliers above 2 ppm were omitted before caculating the mean lead concentration.

3.2.3 mean value of lead concentration for different districts

The bargraph (Figure 4) compares the mean values of lead concentration without outliers above 2 ppm and categorized by only the first two characters of the partial postal code. We can see that the top two district of the highest lead concentration in water is M9 and M6 respectively. This shows that without the data above 2 ppm, M4 does not have a overall high lead concentration, but only some very high outliers. However, M9 and M6 will be the two most dangerous district in Toronto in terms of lead concentration in water. Even though all the mean values shown in (Table 3) are all below MAC by Canadian Government, the M9 and M6 district might need more attention in the topic of water saftey given that the rest districts perform better especially M8.

Table 3: Specific Numbers of the mean for each district

Mean Lead Concentration by District

District	Mean Lead Amount (ppm)
M1	0.00343
M2	0.00200
M3	0.00314
M4	0.00304
M5	0.00153
M6	0.00584
M8	0.00075
M9	0.00745

3.3 How months and year affect Lead Concentration

3.3.1 Times Series analysis for autocorrelation

In this dataset, we have variables time and lead amount, therefore, we can perform times series analysis. As shown in (Figure 5), we can see that except for some sharp increase in 2014, 2019, 2021, 2023, the overall trend is stays around 0 and remains positive. There are no increasing or decreasing trend overtime and no obvious seasonality present.

From the ACF and PACF figures, we can conclude that the previous values of lead concentrations does not impact future values. For ACF, when lag is 0, ACF is always 1 but for all other lags in ACF(Figure 6) and PACF(Figure 7) they are all within the dotted blue lines which is the evidence that there are no significant correlation at any lags present in the data.

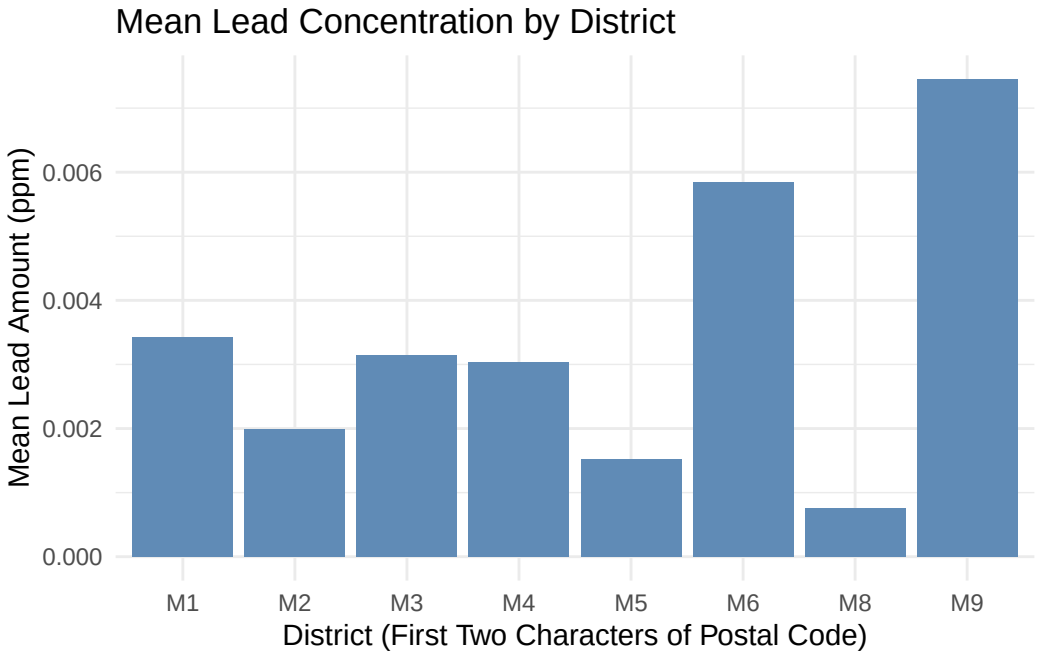

Figure 4: Mean lead concentration categorized by first two characters in partial postal code

This observation is reasonable because the samples are collected in different district in Toronto, and it is safe to assume that one incident of a very high lead concentration might again only be outliers instead of due some factors that will effect the overall water quality in Toronto.

Lead Concentration in water can vary overtime(WHO), therefore the high lead concentration outliers might not be caused by correlation instead, "a probability-based adaptive sampling plan should be used to access exposure(WHO)".

3.3.2 graphs of months vs mean lead concentration

With all data points, we do not see obvious trends over time. Here we examine closely for the constructed variable month from Date. (Figure 8) is the boxplot showing the outliers, but there are not enough evidence to show that there are specific month that will have a higher lead concentration. Therefore, we perform a similar filtering on data that is larger than 2 ppm.

Bar graph (Figure 9) shows the mean lead concentration for each month in the year. April and September has higher lead concentrations that is about 500% of the other months. However, this graph does not provide enough evidence to prove seasonality. We need we examine whether April and September for every year will repeatedly have higher measured lead concentration. This will be further discussed in Section 3.4.

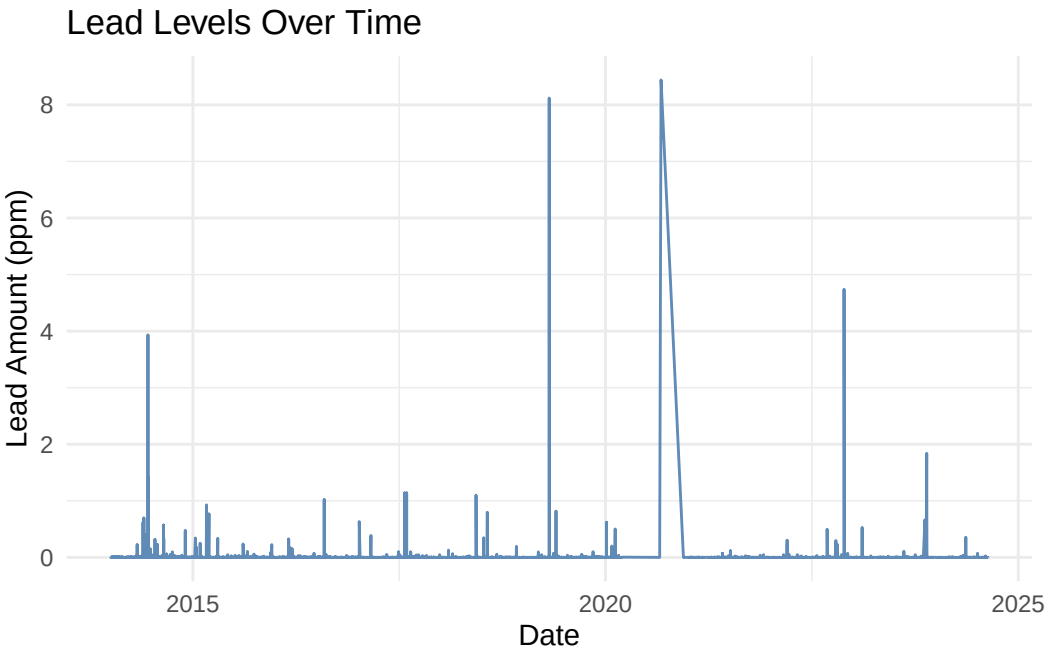

Figure 5: Lead Levels changed over time for times seriese analysis

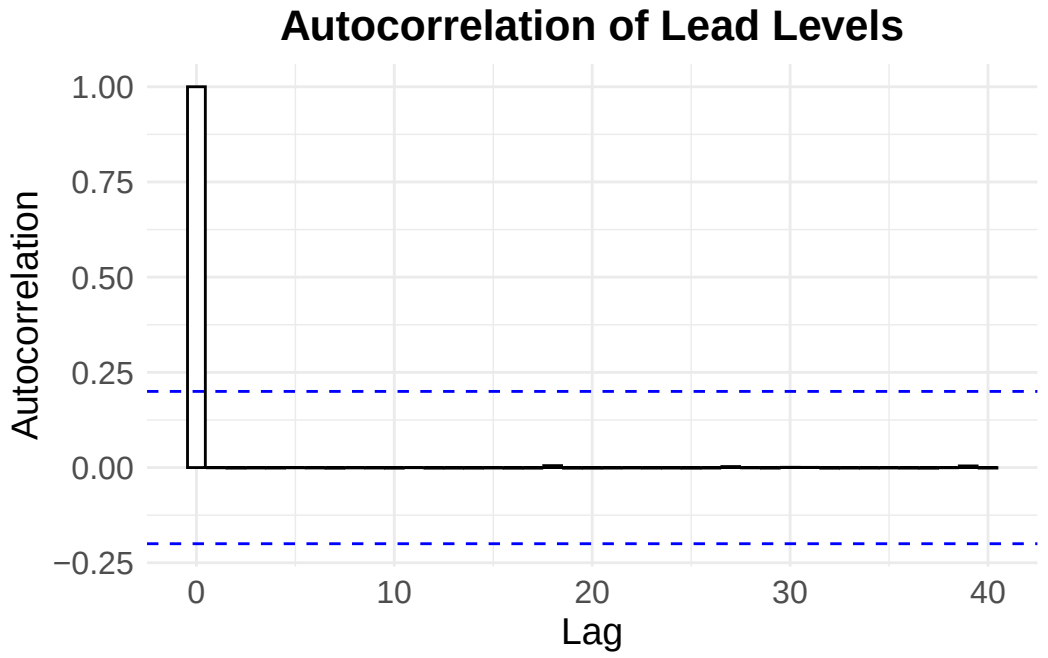

Figure 6: Autocorrelation function for lead concentrations

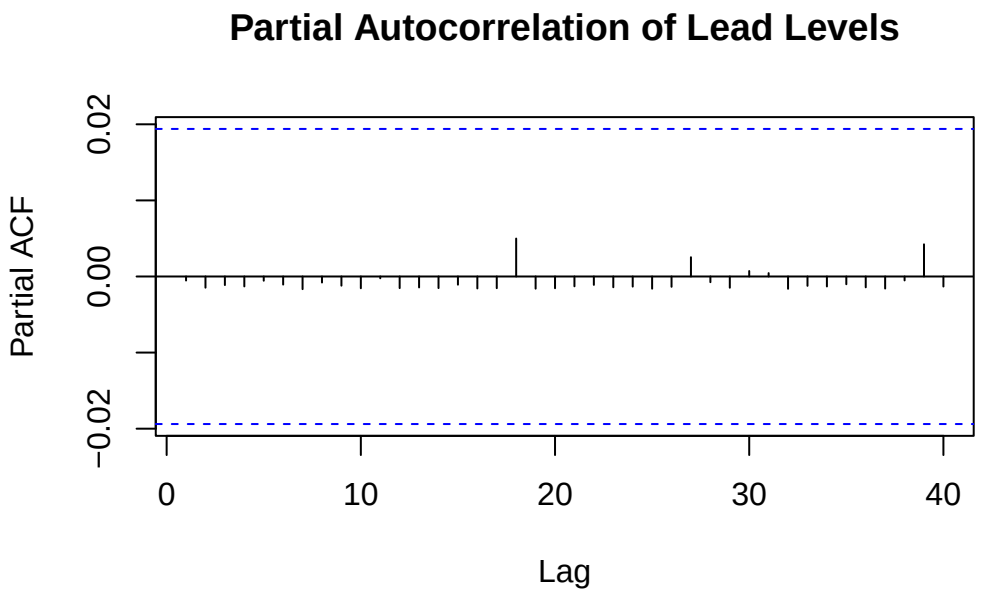

Figure 7: Partial Autocorrelation function for lead concentrations

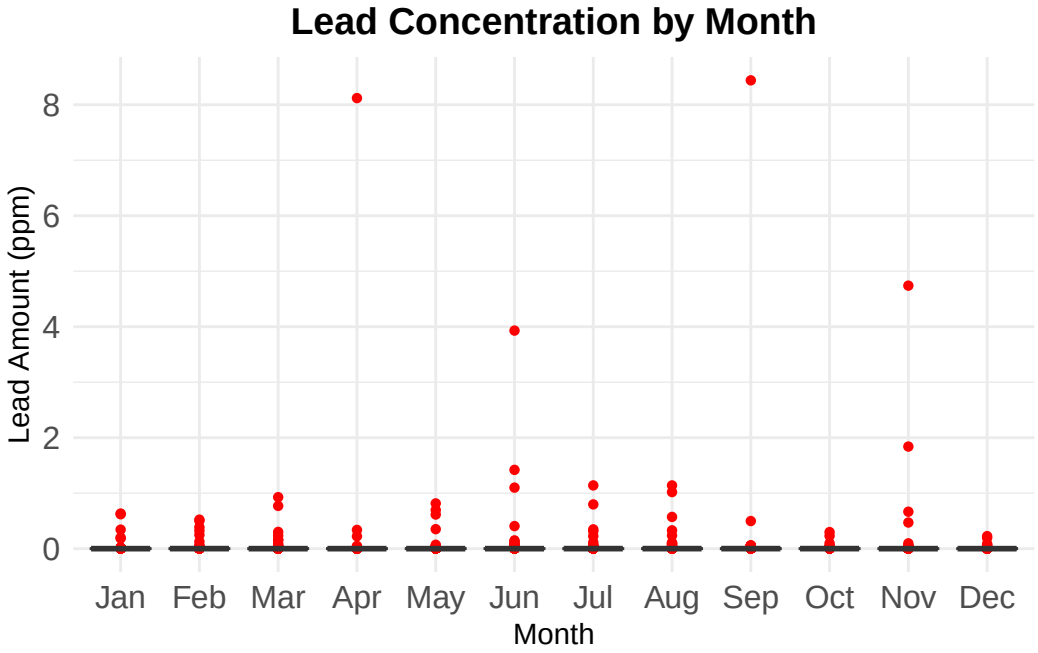

Figure 8: Boxplots that shows the outliers categorized by Months

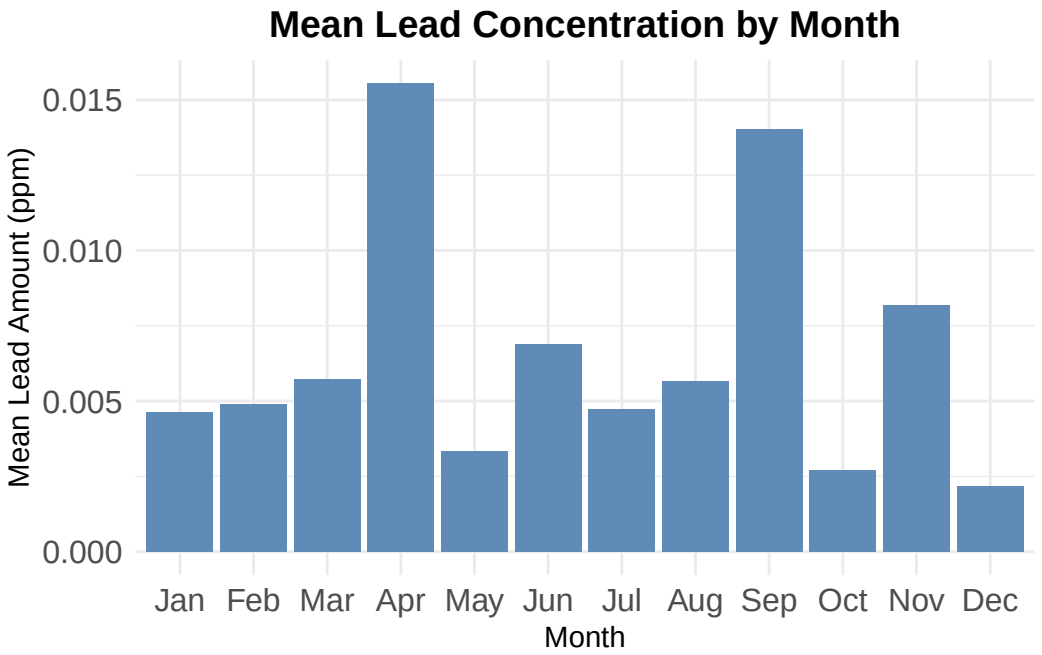

Figure 9: Mean Lead Concentration Without Data Greater than 2 ppm categorized by Months

3.3.3 graphs of years vs mean lead concentration

Similar to months, we also constructed a variable call year. However, both the boxplots and the mean values without data below 2 ppm does not show trend from year to year. This aligns with our conclusion in Section 3.3.1.

3.4 Weaknesses and next steps

A key weakness of the current dataset is that it originates from a variety of water sources and is not necessarily limited to direct drinking water. Since the water sample collection process was not controlled, it is uncertain whether all samples reflect tap water quality. Despite this, the guidelines used for analysis are intended for drinking water safety, which introduces a potential mismatch between the data and its intended use. The program is non-regulated, leaving the sampling process up to residents and without any standardization; as a result, the reliability and comparability of the data across different samples may be compromised.

Considering these weaknesses, it is important to recognize the role of pre-treatment, such as water filtration systems, in affecting lead concentration. This necessitates the documentation of whether samples are taken before or after filtration, which could significantly influence the interpretation of lead levels in relation to health guidelines. To improve the reliability of future data collection, a probability-based adaptive sampling plan should be considered. This

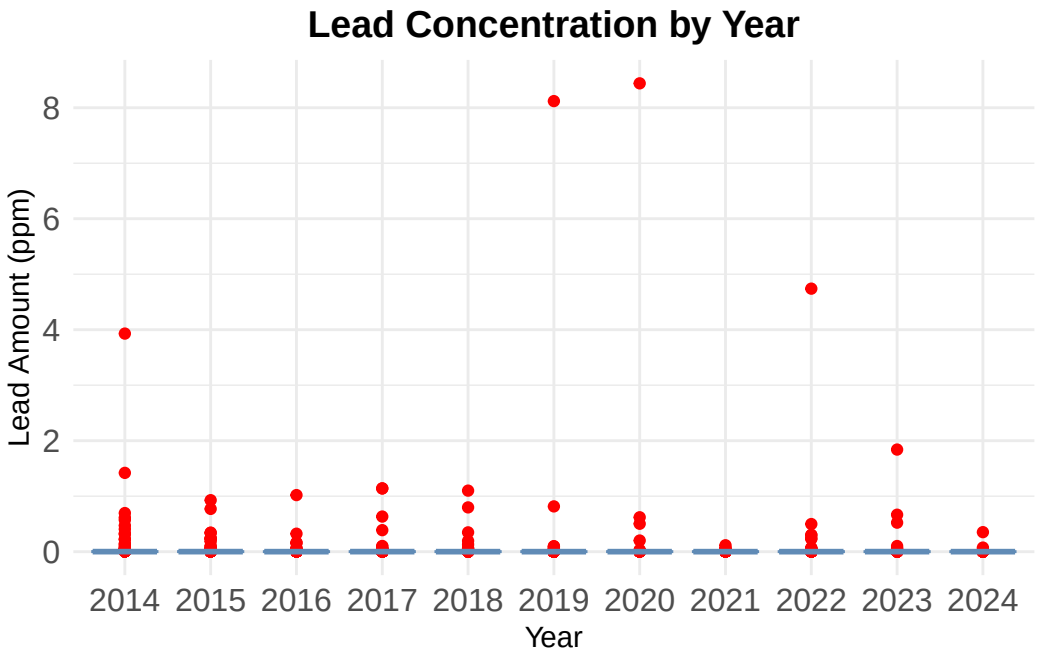

Figure 10: Boxplots that shows the outliers categorized by Year

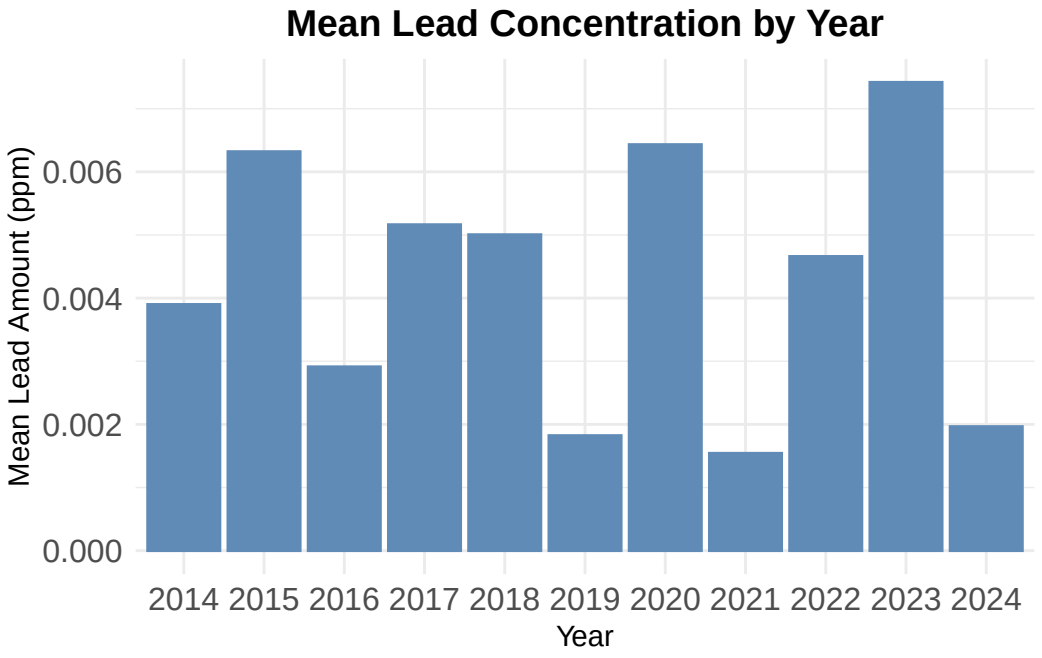

Figure 11: Mean Lead Concentration Without Data Greater than 2 ppm categorized by Year

approach would ensure more representative sampling of tap water from various districts, providing a better understanding of the exposure to lead in drinking water. Additionally, further analysis on the seasonality of lead levels could reveal patterns influenced by environmental factors such as temperature, water flow rates, and seasonal water treatment processes. Identifying these trends is crucial for developing targeted interventions to reduce lead exposure throughout the year.

Appendix

The color used in graphs : blueishgrey <- rgb(96, 139, 182, maxColorValue = 255) is the color of lead in reality.

A Additional data details

References

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

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

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.