We will use this document as our point of reference:

* Add your group project topic/title here: Surface Water Quality Analysis - Michigan
* Let's communicate (among the group including me) through this document
* A draft topic should be okay; we will adapt as we go
* Wait for my next instructions here ….

Understanding water quality is important for taking care of the environment. We need to check different things in the water, like how acidic or basic it is, how much oxygen is dissolved in it, and how warm or cold it is. These factors affect the health of aquatic life. We also look at how much oxygen is needed by tiny organisms to break down things in the water, and how many solid particles are floating around. Another thing we check is the amount of nitrate, which can be good for plants but bad for drinking water if there's too much of it.

This dataset has information from 200 different measurements of these factors. It's useful for research, learning, and making decisions about water quality. Scientists and analysts can use it to understand trends and patterns in water quality and figure out how to keep our water clean and safe for everyone.

The six things we measure are:

1. pH
2. Dissolved Oxygen (DO)
3. Temperature
4. Biochemical Oxygen Demand (BOD)
5. Total Suspended Solids (TSS)
6. Nitrate-Nitrogen (NO3-N)

pH: pH measures the acidity or basicity of a liquid on a scale from 0 to 14. Values less than 7 indicate acidic conditions, values greater than 7 suggest primary needs and a value of 7 indicates a neutral state.

Dissolved Oxygen (DO): DO measures the amount of oxygen dissolved in water and is essential for aquatic life. High DO levels are crucial for the survival of fish and other marine organisms.

Temperature: Temperature affects various physical, chemical, and biological processes in water bodies. It is an essential factor that influences the rate of many aquatic processes.

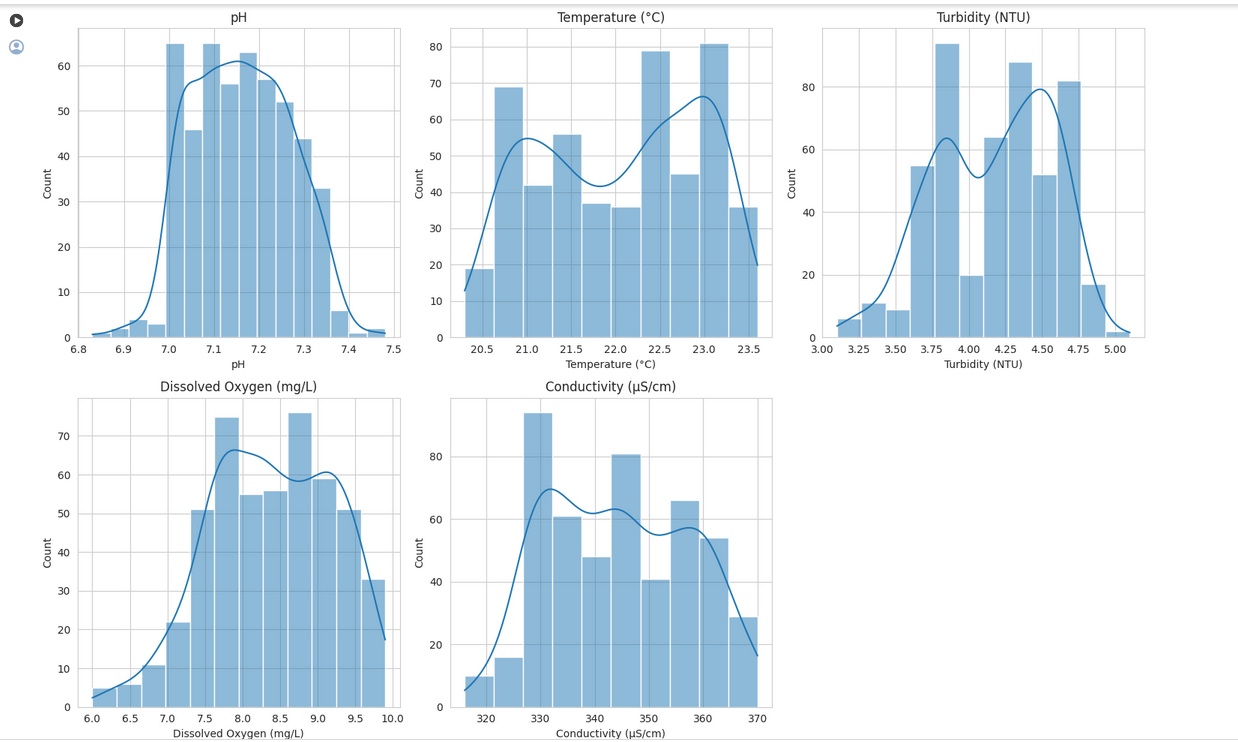
Biochemical Oxygen Demand (BOD): BOD measures the amount of oxygen microorganisms require to decompose organic matter in water. High BOD levels can indicate that the water is polluted with organic matter and may not be suitable for consumption or recreation.

Total Suspended Solids (TSS): TSS measures the amount of solids suspended in water, including organic matter, sediment, and other pollutants. High TSS levels can indicate poor water quality and affect aquatic life and other uses of water.

Nitrate-Nitrogen (NO3-N): NO3-N measures the amount of nitrate in water. Nitrate is an essential nutrient for plant growth, but high nitrate levels in drinking water can harm human health.

<https://colab.research.google.com/drive/1FSFPc4aIF_kbOtEXyS08FnBI3l_1zT-O?usp=sharing>

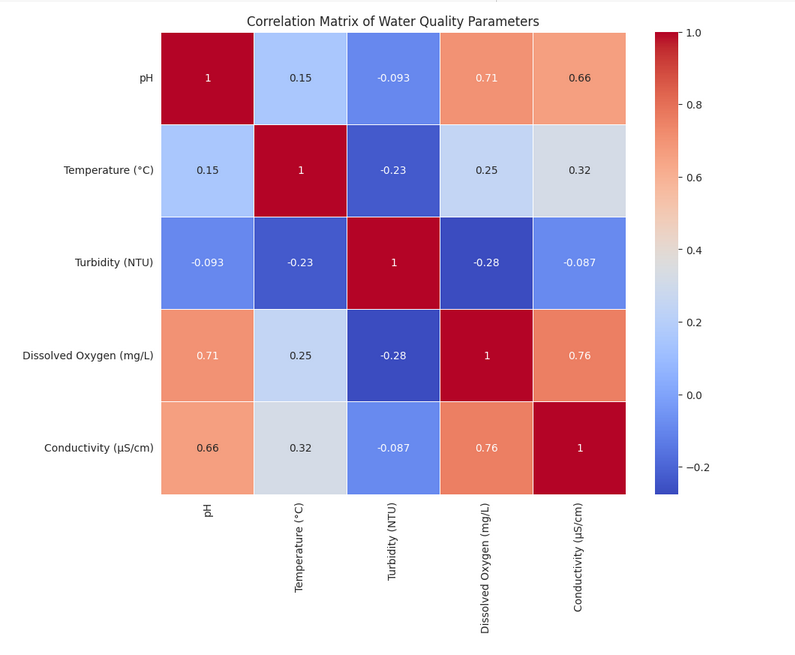
Above is the link to my dataset and here I have started to read the dataset and where I could eliminate the null values and could find out the values in the dataset and describe them thoroughly.



**pH**: The distribution of pH values is roughly normal, centered around the mean value, indicating a consistent water quality in terms of acidity or basicity across samples.

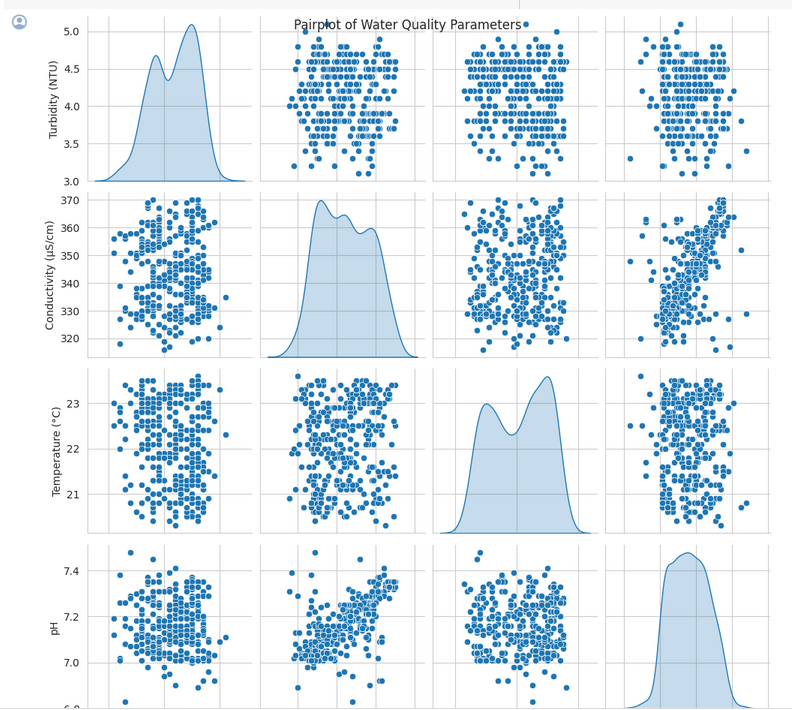
* **Temperature (°C)**: The temperature distribution is also approximately normal, with most samples clustering around the mean temperature.
* **Turbidity (NTU)**: Turbidity shows a slightly right-skewed distribution, suggesting that while most water samples have low turbidity, a few samples have higher turbidity levels.
* **Dissolved Oxygen (mg/L)**: The dissolved oxygen levels are fairly normally distributed, indicating a consistent availability of oxygen across the water samples.
* **Conductivity (µS/cm)**: The conductivity of the water samples is normally distributed, with a slight skew towards higher values, indicating variations in the ionic content of the water samples.

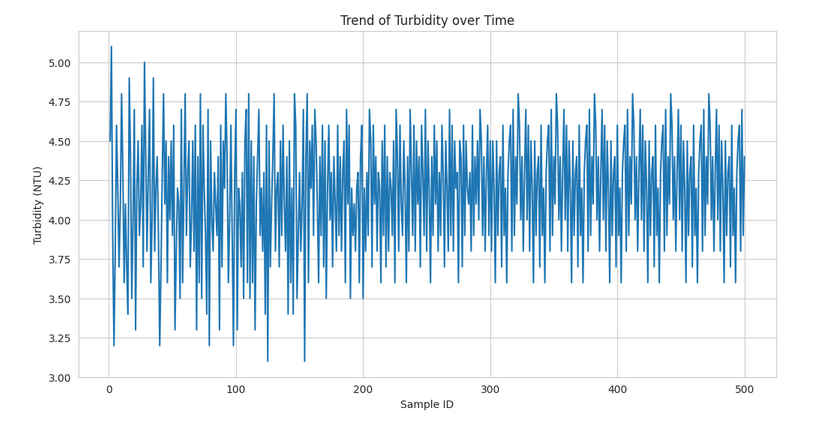
These distributions suggest that the water quality parameters are relatively stable across the samples, with a few exceptions in turbidity and conductivity that might indicate specific areas or conditions where water quality varies more significantly.



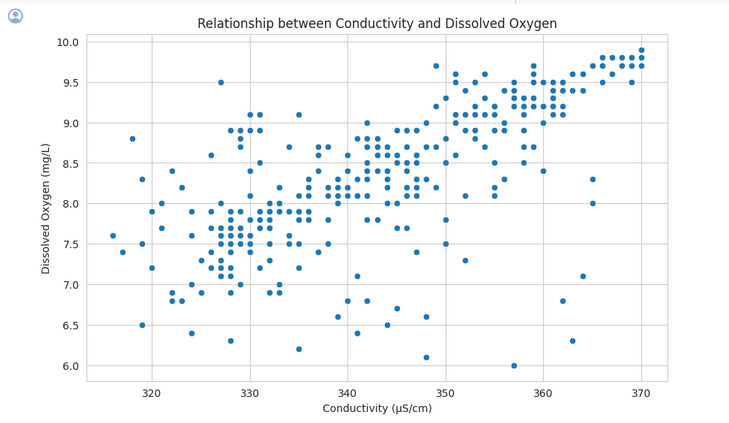
* There are no strong correlations between most of the parameters, indicating that they vary independently of each other to a large extent.
* The strongest correlation observed is a slight negative correlation between temperature and dissolved oxygen, which is expected as warmer water typically holds less dissolved oxygen than cooler water.
* Other correlations are relatively weak, suggesting that factors such as pH, turbidity, and conductivity do not significantly influence each other in this dataset.

These findings suggest that the water quality parameters are influenced by a variety of factors and that changes in one parameter do not necessarily predict changes in another. This independence is typical in natural water bodies where multiple factors can influence water quality.





* The detailed EDA revealed the distribution of Turbidity and Conductivity, showcasing the spread and central tendency of these parameters. The pairplot visualized the relationships between multiple water quality parameters, offering insights into potential correlations and patterns. Additionally, the line plot of Turbidity over time highlighted the trend of this parameter, indicating any significant changes or patterns over the sampling period



* The correlation coefficient between Conductivity and Dissolved Oxygen is approximately 0.76, indicating a strong positive correlation between these two parameters.
* The scatterplot shows a clear positive relationship between Conductivity (µS/cm) and Dissolved Oxygen (mg/L), where as Conductivity increases, Dissolved Oxygen also tends to increase. This visual representation confirms the strong positive correlation observed in the dataset.

[Michigan Water PFAS Concentration Levels.ipynb](https://colab.research.google.com/drive/1vm6Xiz2qgoyYA9RNh8y9j5lrHv9fv7p3#scrollTo=epKxPrjCNMFT)

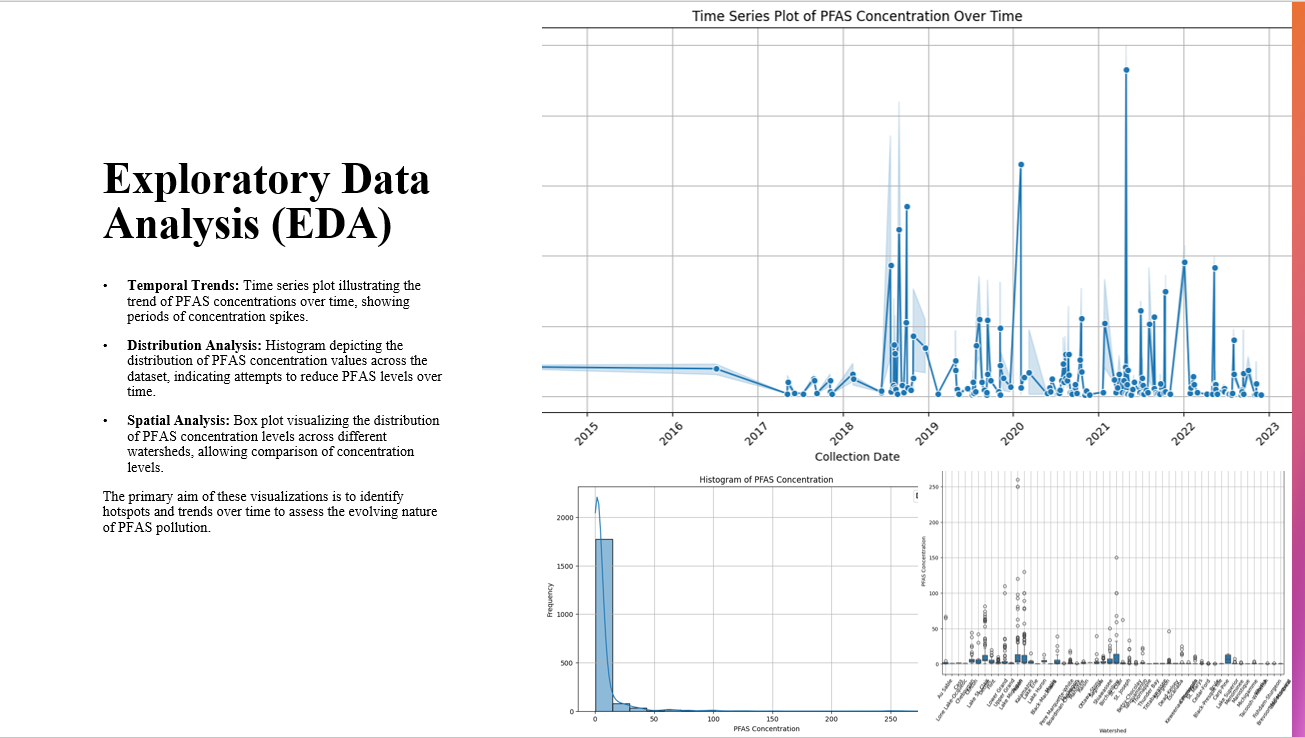
I attached some of the work that I have done on my dataset so far. It focuses on visualizing the PFAS Concentration levels in Michigan Water.

The data includes various fields such as sample IDs, sampling locations, collection dates, analysis methods, concentration units, and specific PFAS concentrations along with associated flags indicating detection limits. Notably, if a flag is listed as "Not Measured," the PFAS analyte wasn't included in the analysis for that sample.

This dataset serves as a critical component of Michigan's PFAS response efforts and is utilized in the MPART: PFAS Geographic Information System.

The data includes extensive information on PFAS concentrations across different waterbodies in Michigan, with corresponding geographic coordinates, sampling depths, and additional descriptions of sampling locations.

**EDA**



**Temporal Trends:** The time series plot illustrates the trend of PFAS concentrations over time, showing periods of concentration spikes. Temporal trend analysis tracks changes in PFAS levels over different time periods, uncovering patterns of persistence, attenuation, or emergence.

**Distribution Analysis:** The histogram depicts the distribution of PFAS concentration values across the dataset, indicating attempts to reduce PFAS levels over time.

**Spatial Analysis:** Box plot visualizing the distribution of PFAS concentration levels across different watersheds, allowing comparison of concentration levels. Spatial distribution analysis reveals areas with elevated PFAS concentrations, pinpointing potential sources of contamination.

**Key Points:**

The primary aim of these visualizations is to identify hotspots and trends over time to assess the evolving nature of PFAS pollution.

Comparison with historical data enables assessment of long-term trends and evaluation of the effectiveness of mitigation measures.

**Predictive Modeling**

We built a linear regression model to predict future PFAS concentration levels based on various variables such as location, sampling depth, and collection date. We assessed the model performance using metrics like Mean Squared Error (MSE) and R-squared score.

The following figure is the MSE for the PFAS dataset:

Mean Squared Error: *6.593209680763768e-05*

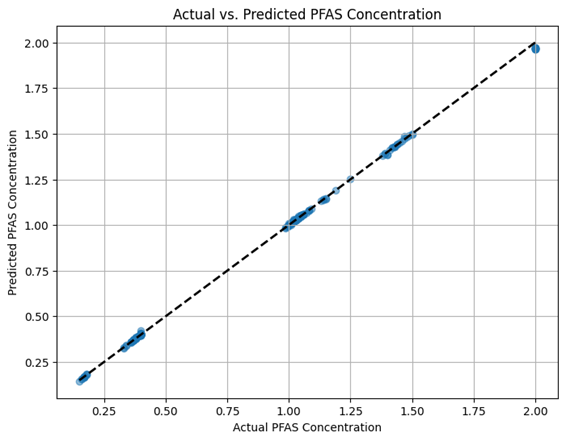
This figure measures the average squared difference between the predicted values and the actual values in the test set. Here's what this value implies:

*Magnitude of Error:* The smaller the MSE, the better the model performance. A lower MSE indicates that, on average, the model's predictions are closer to the actual values. In this case, 6.59 \* 10 \*\*-5 is a relatively small value, suggesting that the model's predictions are quite close to the actual values.

*Comparison:* To fully interpret the significance of this MSE, it's helpful to compare it to other models or to a baseline model (Baseline Mean Squared Error: 3.495244564175446e-05). If this MSE is smaller than that of other models or a baseline, it indicates that the current model is performing better in terms of prediction accuracy.

Context: The interpretation of MSE also depends on the context of the problem and the acceptable level of error. For this application, a MSE of 6.59 \* 10 \*-5 is considered acceptable, while for others, it might not be. Overall, a MSE of 6.59 \* 10 \*-5 suggests that the model is performing relatively well in predicting the target variable. However, for a more comprehensive assessment, it's essential to consider the context of the problem and compare the MSE to other models or benchmarks.

We also visualized our prediction using a scatter plot to display the relationship between actual and predicted PFAS concentrations, evaluating how well the model captures variation in the data.



By comparing observed PFAS concentrations with model-predicted values, the scatter plot offers insights into how effectively the model captures variation in the dataset across different sampling locations and time periods.

The visualization aids in identifying potential discrepancies or patterns in prediction accuracy, informing adjustments or refinements to the predictive model to enhance its performance and reliability.

**Unsupervised Learning**

After successfully applying clustering algorithms (KMeans and Agglomerative Clustering) to our dataset, these were the evaluation metrics found for the clustering performance:

KMeans Silhouette Score: 0.441

Agglomerative Silhouette Score: 0.538

KMeans Davies-Bouldin Score: 0.843

Agglomerative Davies-Bouldin Score: 0.608

Applying KMeans and Agglomerative Clustering helped us to identify patterns in our data. On the other hand, Silhouette score and Davies-Bouldin score was used to evaluate clustering performance. A higher silhouette score and a lower Davies-Bouldin score indicate better clustering performance. In this case, the Agglomerative Clustering algorithm appears to have slightly better performance based on silhouette score and Davies-Bouldin score.

[PFAS Concentration Levels - Michigan surface water](https://github.com/ArjunOn/Academic_Projects/blob/main/PFAS_Surface_Water_Sampling.ipynb)

Focused on the prediction of ADONA concentration levels in water.

Pending: Summarize the findings and make a story.

From the state of Michigan’s website:

*PFAS are a large group of man-made chemicals that include perfluorooctanoic acid (PFOA) and perfluorooctanesulfonic acid (PFOS). PFAS have been used globally during the past century in manufacturing, firefighting and thousands of common household and other consumer products. These chemicals are persistent in the environment and in the human body – meaning they don’t break down and they can accumulate over time. In recent years, experts have become increasingly concerned by the potential effects of high concentrations of PFAS on human health.*

PFAS contamination in West Michigan has unfortunately made [headlines](https://www.wzzm13.com/article/news/local/judge-approves-54m-settlement-kent-co-residents-wolverine-3m-pfas-case/69-091ea42c-125a-4341-8ea0-238fdac4d06e), reminiscent of other environmental advocacy efforts like that of renowned whistleblower [Erin Brockovich](https://en.wikipedia.org/wiki/Erin_Brockovich).

The original data dictionary for the full surface water data is available at <https://gis-egle.hub.arcgis.com/datasets/egle::pfas-surface-water-sampling/about>.

A non-comprehensive laboratory PFAS analyte list (showing for example that PFTeDA is Perfluorotetradecanoic acid) for analyzing data collected by Michigan’s Departments of Environment, Great Lakes, and Energy, Health and Human Services, Agriculture and Rural Development, and Natural Resources can be found [here](https://www.michigan.gov/pfasresponse/-/media/Project/Websites/PFAS-Response/Sampling-Guidance/Minimum-Laboratory-Analyte-List.pdf?rev=a35aba56ec5a4922b986f01e25c1a19d&hash=04E6F164AA5F5CD29B83B39983341345): <https://www.michigan.gov/pfasresponse/-/media/Project/Websites/PFAS-Response/Sampling-Guidance/Minimum-Laboratory-Analyte-List.pdf?rev=a35aba56ec5a4922b986f01e25c1a19d&hash=04E6F164AA5F5CD29B83B39983341345>.

Source

Data was obtained from the state of Michigan website <https://gis-egle.hub.arcgis.com/search?q=pfasgis>.

Data Dictionary

Surface water sampling data

| Variable | Description |
| --- | --- |
| geoid | geographic region ID with the first 2 digits being the state Federal Information Processing Standard (FIPS) code  and the last 3 digits the county FIPS code |
| longitude | The longitude coordinate for the sampling location |
| latitude | The latitude coordinate for the sampling location |
| lab\_sample\_id | The sample ID provided by the analytical laboratory |
| site\_code | The sampling location name |
| coc\_sample\_id | The sample ID listed on chain of custody provided to the analytical laboratory |
| sample\_type | An expanded description of what type of sample was collected |
| lab\_name | The name of the analytical laboratory used for analysis of the sample |
| lab\_job\_name | The analytical laboratories job/work order name |
| collection\_date | The date that the sample was collected |
| analysis\_method | The analysis method used by the analytical laboratory to analyze the sample |
| dilution\_factor | The dilution factor used by the analytical laboratory during sample preparation |
| analysis\_date | The date that the analysis was conducted by the analytical laboratory |
| duplicate | A description of what type of sample was collected |
| watershed | The watershed that the sample was collected from |
| waterbody | The waterbody the sample was collected from |
| location\_code | The sampling location code |
| huc10 | The 10 digit hydrological unit code for the sampling location |
| huc8 | The 8 digit hydrological unit code for the sampling location |
| project | The project that is associated with the sampling |
| description | A description of the sampling location |
| additional\_description | An additional description of the sampling location |
| visit\_id | The identification number of the sampling event |
| sample\_depth | An expanded description of the sampling depth |
| analyte | PFAS analyte |
| analyte\_value | The analyte concentration in the sample measured in parts per trillion (ppt or ng/L) |

Public water supply sampling data

From the [EGLE website](https://www.michigan.gov/egle/maps-data/mpart-pfas-gis):

*The public water supply results include information from both the State of Michigan-funded drinking water Per- and Polyfluoroalkyl Substances (PFAS) sampling efforts and ongoing PFAS compliance monitoring of public water supplies in Michigan. It includes, but is not limited to, sample location, sample date, and sample results…*

*The combined public water supply results data is representative of the PFAS sampling locations, with each location having multiple samples taken, dates for each sample, and the analytical results for each sample.*

| Variable | Description |
| --- | --- |
| geoid | geographic region ID with the first 2 digits being the state Federal Information Processing Standard (FIPS) code  and the last 3 digits the county FIPS code |
| longitude | The longitude coordinate for the sampling location |
| latitude | The latitude coordinate for the sampling location |
| system\_name | Name of water supply system |
| system\_type | Type of water supply system |
| sample\_date | The date that the sample was collected |
| lab\_name\_code | Code for processing lab |
| analyte | PFAS analyte |
| analyte\_value | The analyte concentration in the sample measured in parts per trillion (ppt or ng/L) |

More references for the Analysis:

1. [Analysis of F-53B, Gen-X, ADONA, and emerging fluoroalkylether substances in environmental and biomonitoring samples: A review - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S2214158819300145)
2. [Exposure to PFOS, PFHxS, or PFHxA, but not GenX, Nafion BP1, or ADONA, Elicits Developmental Neurotoxicity in Larval Zebrafish | Science Inventory | US EPA](https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=341170&Lab=NHEERL)
3. [PFAS – PFOA, PFOS, PFNA and More Lab Testing](https://www.suburbantestinglabs.com/pfas-pfoa-and-pfos-lab-testing.html)
4. [PFAS REPLACEMENT COMPOUNDS “The Next Generation”](https://esaa.org/wp-content/uploads/2021/02/ET2020-Obal1.pdf)