

Learning with Data

PFAS Water Contamination

NC STATE



https://go.ncsu.edu/dsa.tsh_pfasstudentlesson

Taryn Shelton
The Science House
Data Science Academy

Katlyn May
Lacey Brown
Center for Human Health and the Environment
Center for Environmental and Health Effects of PFAS

Water Contamination

Case Study:

You are a data scientist working for an environmental consulting company. Your team has been tasked with a water quality monitoring project in NC.

TASK 1: Data Collection (Why?)

- 1) What is NC's largest river basin and to how many people in NC does it supply water?
- 2) In what ways do people and businesses use water from the river?
- 3) Name several water quality standards that can be measured and discuss as a team why it is important to monitor the quality of our waterways.

TASK 2: Data Collection (What?)

Use the following resources to answer the questions on the next page.

- ❖ [Guide to Understanding and Addressing PFAS in our Communities](#)
- ❖ [What are PFAS and how do we classify them?](#)
- ❖ [Environmental Working Group \(EWG\): Short-chain PFAS Chemicals](#)
- ❖ ['Forever chemicals' found in drinking water in dozens of cities](#)
- ❖ [Utah Department of Environmental Quality](#)

TASK 2: Data Collection (What?)

Answer the following questions to learn more about NC's largest river basin and the threats to our public water supply.

- 1) What are PFAS and what are they used for?
- 2) What are the risks of PFAS to human health?
- 3) What are the different ways that PFAS get into groundwater vs surface water?
- 4) Why are PFAS called “forever chemicals”?
- 5) What is the difference between long-chain and short-chain PFAS?

TASK 3: Data Collection (Where?)

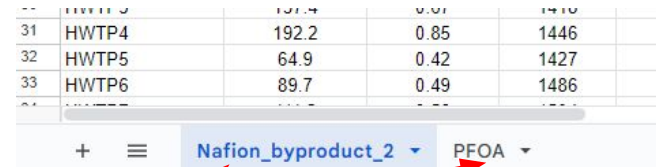
Brainstorm with your team what types of areas or locations you think should be tested for PFAS and why.

Then share out as a class.

TASK 4: Reviewing and Questioning the Data

The linked dataset below contains PFAS/PFOA concentration data from several waterways along the Cape Fear River.

[Nafionbp2 PFOA Water Testing Data](#)



31	HWTP4	192.2	0.85	1446
32	HWTP5	64.9	0.42	1427
33	HWTP6	89.7	0.49	1486

Open the file and review the data in **each tab**.

1) What do you notice? What do you wonder?

*Investigate the data and try to make sense of it.

TASK 4: Reviewing and Questioning the Data

In reviewing the data, consider the following questions:

- 1) What questions do you have about this data?
- 2) What parts of this dataset can you understand?
- 3) What parts are you uncertain about?
- 4) Try to explain the parts you are uncertain about.
(It's ok if you are wrong!)

TASK 5: Understanding Data in its Context

When given a dataset to analyze, you should ask questions about it to better understand the data in its **context**.

Look at the data in both the “**Nafion_byproduct_2**” and “**PFOA**” tabs.

* The next two slides gives more information about these two types of chemicals that were tested for in the water samples.

- 1) As a team, think about why researchers may want to test for each of these chemicals individually instead of PFAS in general?

TASK 5: Understanding Data in its Context

PFOA is a legacy, longer-chain PFAS. This type of PFAS was phased out by 2015 due to its [health risks](#), and should no longer be in use per the EPA's [PFOA Stewardship Program](#). However, PFOA can linger in the environment.

Because legacy long-chain PFAS don't degrade well over time, they can continue to show up in the environment from secondary sources, such as prior disposal into landfills.

TASK 5: Understanding Data in its Context

[Nafion byproduct 2](#) is a newer, [short-chain](#) PFAS molecule produced by the Dupont/Chemours company. Nafion byproduct 2 replaces the legacy longer-chain PFAS, such as PFOA and is not yet regulated by the EPA. The [health effects](#) of this specific chemical are unknown, however, it is in the PFAS family. Nafion-BP2 and other PFAS can contaminate the environment from air emissions, runoff from contamination sites like airports or landfills, and wastewater discharges from the factories that produce them.

The [Chemours](#) factory is located just south of Fayetteville, NC.

TASK 5: Understanding Data in its Context

Look again at the data in the “**Nafion_byproduct_2**” and “**PFOA**” tabs.

- 1) How are the sample **Locations** (column A) coded? [Explain what the letters mean. Refer to the **Comments** (column E).]
- 2) What do you think “Standard” and “Instrument Response” mean and why are they included?
- 3) In column G, why are only some of the samples labeled as “raw” or “finished”? [What do those designations mean and why don't all of the samples have them?]
- 4) Why do you think not all of the samples have a collection date?

Good Data Practices!

If you were not a part of the team collecting the data and no codebook explaining the variables was provided, then you should **go back to the team** who collected it **and ask questions** to better understand each of the variables and the context of the data.

(This lesson will better explain the data as you progress.)

(Explanations of each variable are also in the Teacher Guide.)

TASK 6: Data Wrangling & Cleaning

Both of the datasets are **Raw** datasets, also known as “messy data”. This is data in its original form after it is collected.

- Raw data needs to be cleaned up before it can be analyzed.
- This process is known as “Data Wrangling.”

1) Review the following slides on **Data Wrangling** and then, as a team, decide at least 2 things you could do to clean up both datasets.

Practicing data wrangling and cleaning

Key Idea: Understand how to clean up data and make important decisions during that process and how those decisions can affect analysis

Data Wrangling - formatting and structuring raw data into a more useable form

Tidy Data - organizing data so that it can be easily analyzed

Data Cleaning - fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset

Good Data Practices!

A good data science practitioner ALWAYS keeps a copy of the original dataset AND tracks every step made during the data wrangling and cleaning process.

- You never want to lose your original data AND you want to record each change made because...
 - a. Mistakes happen
 - b. You may need to reference the original data
 - c. It is important to be transparent about how you manage and analyze your data

Data Wrangling - Technical Terms

variable - a characteristic, number, or quantity that can be measured.

observation - all of the measurements for a given entity.

value - a single measurement of a single variable for a given entity.

region	year	population
Toronto	2016	2235145
Vancouver	2016	1027613
Montreal	2016	1823281
Calgary	2016	544870
Ottawa	2016	571146
Winnipeg	2016	321484
Hamilton	2016	306034
Edmonton	2016	537634
Halifax	2016	187478
London	2016	220452
Victoria	2016	172559
St. John's	2016	92353
Saskatoon	2016	124766

Data Wrangling - Tidy Data

- Each row is a single observation,
- Each column is a single variable, and
- Each value is a single cell

rows = observations

region	year	population
Toronto	2016	2235145
Vancouver	2016	1027613
Montreal	2016	1823281
Calgary	2016	544870
Ottawa	2016	571146
Winnipeg	2016	321484

columns = variables

region	year	population
Toronto	2016	2235145
Vancouver	2016	1027613
Montreal	2016	1823281
Calgary	2016	544870
Ottawa	2016	571146
Winnipeg	2016	321484

cells = values

region	year	population
Toronto	2016	2235145
Vancouver	2016	1027613
Montreal	2016	1823281
Calgary	2016	544870
Ottawa	2016	571146
Winnipeg	2016	321484

Data Wrangling - Cleaning Data

- 1) Hide/Remove duplicate or irrelevant observations
 - Irrelevant observations are those that do not fit into the specific problem you are trying to analyze.
- 2) Fix structural errors
 - Strange naming conventions, typos, or incorrect capitalization
- 3) Filter unwanted outliers
 - Just because an outlier exists, doesn't mean it is incorrect!
 - Determine the validity and relevance of the outlier before deciding to delete or hide

Data Wrangling - Cleaning Data

4) Handle missing data

- Many programs will not accept missing values
- Options:
 - a) Drop observations that have missing values, but then you will lose information
 - b) Input missing values based on other observations, but you will lose integrity of the data because you may be operating from assumptions and not actual observations
 - c) Alter the way the data is used to effectively navigate null values

Good Data Practices!

Before making any changes to the data, such as deleting observations that have missing values, first **consult with those who collected the data** ...

- a. To gain a better understanding of why values are missing
- b. To better understand the importance of those observations to the overall data analysis

TASK 6: Data Wrangling & Cleaning

Let's clean this data together!

Before making ANY changes to a dataset, **ALWAYS** make a copy so that the original data remains intact.

1. Make a copy of both tabs.

Copy of Nafionbp2_PFOA Water Testing Data

File Edit View Insert Format Data Tools Extensions Help

100% 123 Arial 10

A2 Method blank

	A	B	C	D	E	F	G	H	I	J
1				Nafion byproduct 2 internal standard instrument response						
2	Method blank	10.4	0.23	1408	DI water that has been processed like a sample	(has the addition of internal standard)				
3	0.5 ng/L standard	97.5	0.54	1381	Calibration standard					
4	1 ng/L standard	236.9	1.02	1407	Calibration standard					
5	2 ng/L standard	473.8	1.75	1391	Calibration standard					
6	5 ng/L standard	1184.5	4.78	1367	Calibration standard					
7	10 ng/L standard	2369.0	6.99	1385	Calibration standard					
8	25 ng/L standard	4738.0	23.97	1381	Calibration standard					
9	50 ng/L standard	9476.0	47.94	1388	Calibration standard					
10	100 ng/L standard	18952.0	97.35	1324	Calibration standard					
11	250 ng/L standard	47380.0	254.70	1172	Calibration standard					
12	500 ng/L standard	94760.0	501.06	1059	Calibration standard					
13	1000 ng/L standard	189520.0	883.34	906	Calibration standard					
14	Blank		414.40	1						
15	QC 25 ng/L		20.45	1377	QC=Quality control					
16	QC 250 ng/L		220.91	1303						
17	Blank		2.49							
18	Blank		17.45							
19	CFPUA1		37.33							
20	CFPUA1_dup		38.72							
21	CFPUA2		52.18							
22	CFPUA3		51.36	1381						
23	CFPUA4		40.25	1409						
24	CFPUA5		45.9	1406						

Click the dropdown arrow

5/16/2017 Finished water
5/16/2017 Finished water
5/18/2017 Raw water
5/18/2017 Finished water
5/17/2017 Finished water
5/17/2017 Raw water

Nafion_byproduct_2 PFOA

TASK 6: Data Wrangling & Cleaning

Look at columns **B** and **D**.
The **variables** “Standard Response / Instrument Response” are used to calibrate the devices and are not relevant for analyzing this data.

2. You can delete **variables/columns B & D** from both copied tabs.

Click the dropdown arrow

Copy of Nafionbp2_PFOA Water Testing Data

File Edit View Insert Format Data Tools Extensions Help

100% | \$ % .0_ .00 123 | Arial

	A	B	C	D	E	F	G
1		Nafion byproduct 2 instrument response	Nafion byproduct 2 concentration (ng/L)	Nafion byproduct 2 internal standard instrument response		Sample collection date	Finish or water: WTP
2	Method blank	10.4	0.23	1408			
3	0.5 ng/L standard	97.5	0.54	1381			
4	1 ng/L standard	236.9	1.02	1407			
5	2 ng/L standard	437.0	1.75	1391			
6	5 ng/L standard	1270.6	4.78	1367			
7	10 ng/L standard	1907.3	6.99	1385			
8	25 ng/L standard	6522.9	23.27	1381			
9	50 ng/L standard	13736.2	47.80	1388			
10	100 ng/L standard	27531.8	97.35	1324			
11	250 ng/L standard	69799.1	254.70	1172			
12	500 ng/L standard	140571.5	501.06	1059			
13	1000 ng/L standard	250870.7	883.34	906			
14	Blank	65.8	414.40	1			
15	QC 25 ng/L	5698.9	20.45	1377			
16	QC 250 ng/L	66065.4	220.91	1303			
17	Blank	3.5	2.49	8			
18	Blank	14.9	17.45	4			
19	CFPUA1	10861.3	37.33	1417	CFPUA=Cape Fear Public Utility Authority	5/16/2017	Finishe
20	CFPUA1_dup	11408.3	38.72	1433	duplicate sample preparation	5/16/2017	Finishe
21	CFPUA2	15738.9	52.18	1453		5/18/2017	Raw w
22	CFPUA3	14717.4	51.36	1381		5/18/2017	Finishe
23	CFPUA4	11680.0	40.25	1409		5/17/2017	Finishe
24	CFPUA5	13144.3	45.28	1408		5/17/2017	Raw w

Dropdown menu options:

- Delete selected columns
- Clear selected columns
- Hide columns
- Resize selected columns
- Create a filter
- Conditional formatting
- Data validation
- Dropdown
- Smart chips
- View more column actions

Nafion_byproduct_2 | Copy of Nafion_byproduct_2 | PFOA | Copy of PFOA

TASK 6: Data Wrangling & Cleaning

In the datasets, you see **observations** that say “Calibration standard”, “standard”, “Blank”, “Method blank”, “CC=Calibration check”, or “QC=Quality Control”.

You should notice that these **observations** also do not have sample collection dates.

These are also not relevant for analyzing this data.

3. You can delete these **observations**. But **WAIT**. If this were a very large dataset, it would be very difficult to delete each entry individually.

** Follow the directions on Slide 26 to delete these more efficiently.

Good Data Practices!

When preparing data for analysis, you can either delete or simply **HIDE** certain **variables** or **observations** from the COPIED version of the dataset.

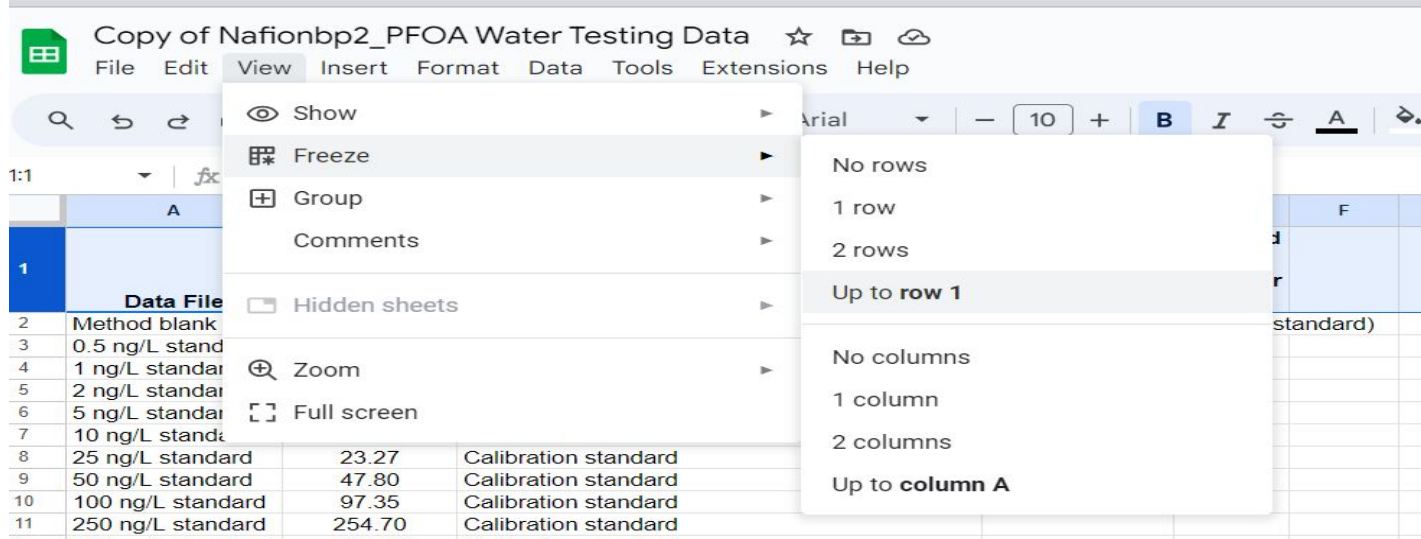
- If you have confirmed that the information is not at all relevant, you can **delete** it.
- If the information is not relevant at the moment but may be useful under different circumstances, the you can **hide** it.

★ **Keep track of all changes that you make!**

TASK 6: Data Wrangling & Cleaning

3a. Highlight the top row, containing the **variables**/column headings.

3b. Then go to **View** and freeze that top row. (This will keep the row at the top when you sort the entries.)



The screenshot shows a Google Sheets interface with the following data:

	A	B	C	D	E	F
1	Data File					
2	Method blank					
3	0.5 ng/L stand					
4	1 ng/L standar					
5	2 ng/L standar					
6	5 ng/L standar					
7	10 ng/L stand					
8	25 ng/L standard	23.27	Calibration standard			
9	50 ng/L standard	47.80	Calibration standard			
10	100 ng/L standard	97.35	Calibration standard			
11	250 ng/L standard	254.70	Calibration standard			

The View menu is open, showing the Freeze option with a sub-menu that includes 'Up to row 1' (highlighted).

TASK 6: Data Wrangling & Cleaning

3c. Now select the **variable: Location** in column A and sort it A-Z by clicking the dropdown arrow.

The screenshot shows a Google Sheets interface with a spreadsheet titled "Copy of Nafionbp2_PFOA Water Testing Data". The spreadsheet has columns D, E, and F. Column D is labeled "Sample collection date", column E is "Finished or raw water for WTPs", and column F is "Authority". The data in these columns includes dates like 5/16/2017 and 5/18/2017, and the text "Finished water" and "Raw water".

Column A is selected, and a context menu is open. The menu options are:

- Cut (Ctrl+X)
- Copy (Ctrl+C)
- Paste (Ctrl+V)
- Paste special
- Insert 1 column left
- Insert 1 column right
- Delete column
- Clear column
- Hide column
- Resize column
- Create a filter
- Sort sheet A to Z** (highlighted with a red oval)
- Sort sheet Z to A

TASK 6: Data Wrangling & Cleaning

3c. After sorting, select the **observations** that list:

- “Calibration standard”
- “Standard”
- “Blank”
- “Method blank”
- “CC=Calibration check”
- “QC=Quality Control”

Right-click and Delete those **observations**. Do this for both datasets.

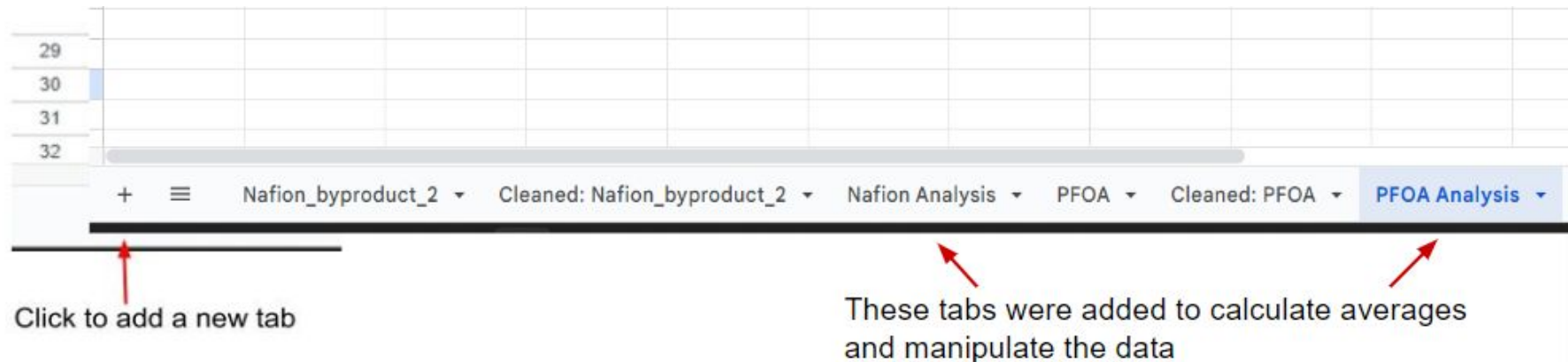
	A	B	C	D	E	F	G
		Nafion byproduct 2 concentration (ng/L)		Finished			
1	Data File		Comment				
5	1 ng/L standard	1.01					
6	10 ng/L standard	6.99	Calibration standard				
7	10 ng/L standard	7.01					
8	100 ng/L standard	97.35	Calibration standard				
9	100 ng/L standard	102.92					
10	1000 ng/L standard	883.34	Calibration standard				
11	1000 ng/L standard	875.35					
12	2 ng/L standard	1.75	Calibration standard				
13	25 ng/L standard	23.27	Calibration standard				
14	25 ng/L standard	23.90					
15	250 ng/L standard	254.70	Calibration standard				
16	5 ng/L standard	4.78	Calibration standard				
17	5 ng/L standard	4.72					
18	50 ng/L standard	47.80	Calibration standard				
19	500 ng/L standard	501.06	Calibration standard				
20	500 ng/L standard	497.09					
21	Blank	414.40					
22	Blank	2.49					
23	Blank	17.45					
24	Blank	37.85					
25	Blank	12.56					
26	Blank	11.85					

- ✂ Cut Ctrl+X
- 📄 Copy Ctrl+C
- 📄 Paste Ctrl+V
- 📄 Paste special ▶
- + Insert 35 rows above
- + Insert 35 rows below
- 🗑 Delete rows 2 - 36**
- ✖ Clear rows 2 - 36
- 👁 Hide rows 2 - 36
- 📏 Resize rows 2 - 36
- 🔍 Create a filter
- 🎨 Conditional formatting

TASK 7: Data Analysis

When analyzing data, each time you want to make changes or focus on certain variables, you should work in a new tab so that you don't lose the original data.

1. At the bottom, click the + and add a new tab, then rename it.



Good Data Practices!

You should always track your processing steps.

Option 1: Create a new tab and record each step you take in processing the data (such as when you hide or delete **variables/observations.**)

Option 2: Make a new tab each time you make changes to analyze the data in a different way.

- a. Decide on a naming convention so that the steps are clear when referred back to.
- b. For each new tab add a note beneath the new data that describes what was done from the previous step.

TASK 7: Data Analysis

Now you are ready to analyze!

Think about what types of trends you want to investigate.

From your cleaned tab, select the data that you want to work with and paste it into the spreadsheet on the new tab. Label that tab.

**** Remember to include the **Location** column in each new spreadsheet**

2. Decide how you want to manipulate the data.
 - a. Ex: Find the mean for each testing site
 - b. Ex: Categorize by Water Treatment Plant (WTP) vs surface water (anything not “WTP”)
 - c. Add any new information (up or downstream from the Chemours plant)
 - d. Compare Raw vs Finished water at the Water Treatment Plants

TASK 7: Data Analysis

- When manipulating your data, such as counting entries or calculating averages, you should work in a spreadsheet, such as Excel or Sheets.
 - Just remember to make any changes in a **new tab**
- You can then visualize the data either in Excel/Sheets or you can upload the data that you want to visualize into [CODAP](#).
 - First download the data tab you want to work with as a CSV file. Then **IMPORT** that file into CODAP

Nafionbp2_PFOA Water Testing Data

File Edit View Insert Format Data Tools Extensions Help

- New
- Open
- Import
- Make a copy
- Share
- Email
- Download
- Approvals
- Labels
- Rename
- Move
- Add shortcut to Drive
- Move to trash

Microsoft Excel (.xlsx)

OpenDocument (.ods)

PDF (.pdf)

Web Page (.html)

Comma Separated Values (.csv)

Tab Separated Values (.tsv)

CODAP

About Forums Contact Help [Launch CODAP](#)

Common Online Data Analysis Platform (CODAP)

Open-source software for dynamic data exploration

For Educators For Developers

Untitled Document

Tables Graph Map Slider Calc Text Plugins

WHAT WOULD YOU LIKE TO DO?

OPEN DOCUMENT OR BROWSE EXAMPLES

CREATE NEW DOCUMENT

Untitled Document

New

Open...

Close

Import...

Revert...

Save...

Create a copy

Share...

Rename

TASK 7: Data Analysis - Visualizations

- What do you see? What do you wonder?
 - Explore the data, look for trends or patterns, and ask questions.
 - Exploring the data through a variety of visualizations may reveal different patterns, interpretations, or insights.
- ★ *Note: Data might be viewed in different ways, but the data point never changes or disappears.*
- ★ Click [here](#) to learn more about the different types of graphs.

TASK 8: Communicating Your Findings

- 1) What interesting trends or patterns did you find?
- 2) What impacts could those trends have on certain communities or populations?
 - Your communication piece should include both written explanations and graphic visualizations.
 - Visualizations should be interesting enough to grab the reader's attention, but clear enough to understand the trends you are trying to convey.

Good Data Practices!

It is important to understand that correlation does NOT equal causation.

Data scientists may investigate possible correlations to help formulate data stories or identify potential avenues of further research.

Discovering correlations between certain variables cannot be directly interpreted into causation.

TASK 8: Communication - What do we do with the insights we have learned from the data?

- 3) What decisions can be made based on your findings?
 - a) What additional information would you need first?
 - b) Who is not included in the data?
 - c) How can this information be attained in the future?
- 4) What questions are there that the given dataset can't answer or can only partially answer?
- ★ It is good to acknowledge that the data is not perfect but can still be used for modeling

TASK 8: Communication - What do we do with the insights we have learned from the data?

- You can discuss and argue certain ideas, while realizing that datasets cannot provide pure answers

★ Remember:

Data is information - not truth - with error, variability, and degrees of inclusion/exclusion

TASK 9: Data for Social Good

- 1) Compare the data to the [EPA's proposed maximum contamination level \(MCL\) for drinking water](#).
 - a) Note: There are no regulations as of yet for Nafion_BP2
- 2) How can this data be used to create & advocate for policies?
- 3) How can this data be used to create environmental plans?
 - Filtration; Cleanup; Further monitoring; etc
- 4) How can this data be used to plan for human health monitoring?
 - a) What data is missing that you may first need to make this plan?
- 5) How can we use the data to make decisions that can affect the health and well-being of communities?

Extensions

- Create a public information poster
- Write a 2 minute speech for public comment to your state legislature
- Design your own water quality testing experiment

Additional Resources

- [Per- and Polyfluoroalkyl Substance \(PFAS\): Overview and Prevalence](#)
- [Nafion byproduct 2 Found in Blood of Well Users Near Fayetteville, N.C.](#)
- [National Academies: Guidance on PFAS Exposure, Testing, and Clinical Follow-Up](#)
- [PFAS Contamination Site Database](#)
- [EPA: The Fifth Unregulated Contaminant Monitoring Rule \(UCMR 5\) Fact Sheet](#)
- [EWG Interactive Map: PFAS Contamination in the U.S.](#)
- [EWG: PFAS Chemicals](#)
- [Topographic Map of the US](#)
- [North Carolina Income - Table](#)
- [U.S. Census: Quick Facts](#)
- [NC Cancer Incidents](#)
- [Instructions on how to import tables from PDF to Excel](#)

Additional Videos

- [Drinking water in North Carolina being tested for toxic substance](#)
- [The PFAS Puzzle: Lessons On Science - Interview with NCSU's Dr. Knappe](#)
- [How an unregulated chemical entered a North Carolina community's drinking water](#)
- [WRAL Documentary: 'Forever Chemicals: North Carolina's Toxic Tap Water'](#)
- [Researchers discover new forever chemicals in Cape Fear River](#)
- [Team Of Scientists Formed To Study PFAS In Cape Fear River](#)