DSA Data Stories in the Classroom

Lesson Title: Learning with Data - PFAS Water Contamination
Grade(s): 8th - 12th
Subject(s): Science, Social Studies, Math

Lesson Overview:
This lesson is designed to walk students through the data-wrangling process. Students learn how to clean up raw datasets to get them ready for analysis. It is designed as a case study approach to encourage students to decide how they may want to analyze the data. The case study focuses on PFAS levels at various water collection sites along the Cape Fear River. It is a simple dataset with only 3 variables and introduces students to the CODAP platform for low-entry analysis. The Teacher Guide offers suggestions for group activities and ideas about other possible datasets that students can use to look for correlations and to build a data story.

How did it go?:
This lesson was presented to 8th grade - high school science teachers during a professional development workshop. The teachers appreciated the community-based relevance of the data and the ease of use of the CODAP platform. They found the data cleaning process easy to follow but did struggle during some parts of the analysis, particularly with making decisions about how to analyze the data.

Standards Alignment

8th Grade Science:

- Hydrosphere: potability, water treatment, Clean Water Act, human impact, stewardship
- Human health impacts - PFAS in human blood

Social Studies:

- Acts vs Laws - and the process of trying to regulate chemicals
- How to fix these issues in government
- How could you solve it?

Math:

- Statistical differences in raw vs filtered water
- Modeling - How could new regulations affect data?

TEACHER GUIDE: PFAS Water Contamination

Total Time: 1 hr 45 min - 2 hr 40 min

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

TASK 1: Data Collection (Why) (8 min)

1. Allow students 8 minutes to discuss and use their resources to answer the questions in the lesson.

TASK 2: Data Collection (What) (12 min)

2. Allow students 12 minutes to discuss and use the resources provided in the lesson to answer the questions in the lesson.

TASK 3: Data Collection (Where?) (3 min)

3. Allow students 3 minutes to brainstorm different locations that could be monitored for PFAS contamination.
    - Activity Option 1: Students could write ideas on sticky notes, and call them out as they stick them to their table.
    - Activity Option 2: Students could write their ideas on extra-large chart paper or a whiteboard at their table.
4. Allow groups to share out their ideas.
    - Some possible options may include: upstream and downstream from a chemical plant where products containing PFAS are made (or where the actual chemicals are produced), near water treatment plants, near farms, near recreation areas.

TASK 4: Reviewing and Questioning the Data (8 min)

Students should first look at both of the original datasets (**Nafion_byproduct_2** and **PFOA**). There is a lot of irrelevant data as well as missing data, and ERRORS.

5. Students should first take time to "sit with the data." Encourage them to look it over, discuss it within their groups, and try to determine the context as well as decipher the different parts.
6. Allow students 5 minutes to examine, discuss, and question the datasets.
    - Activity Option: What do you Notice/Wonder Burst - Students write an observation or a question on a sticky note and then call out the question and stick it to the table. (Use different colors for Notice vs Wonder)
7. AFTER the Notice/Wonder Burst, student groups should then answer the questions in the lesson.
8. Then share out what parts they can decipher as a class. **This is a good place to discuss with students that if they are not the ones collecting the information and no codebook is provided, then they should go back to the team who collected the data and ask questions to better understand it. (This lesson will better explain the data as students progress.)

TASK 5: Understanding Data in its Context (5 min in group + 5 min share out = 10 min)

It is important to understand the context of the data that you are working with, however, unless you were involved in the data collection process, there will likely be parts of the dataset that do not make sense.

9. Information about the types of PFAS in each dataset is provided in the lesson. Students should read over the information and then discuss why researchers may want to test for those chemicals individually versus PFAS in general (remember PFAS describes a *family* of chemicals.)
   - ➢ Thoughts that should come up in discussion:
     - ○ Each chemical in the PFAS family is produced by a specific company.
     - ○ Knowing which specific chemicals are present will help to track the source of the contamination, while only testing for PFAS in general doesn't help to locate potential contamination sources.
     - ○ The production of PFOA should have been discontinued in 2015. High amounts of PFOA immediately near the Dupont/Chemours factory might indicate that they are still producing the chemical. Finding PFOA in other areas might help to locate secondary contaminations and lead to considerations for cleanup.
     - ○ Nafion byproduct 2 is specifically produced by the Dupont/Chemours factory. Testing specifically for the type of PFAS may help to track the movement of the chemical through the environment, track its movement and potential contamination over time, and may even help with future investigations into disease clustering around high-concentration areas.
10. There are guiding questions in the lesson to encourage students to figure out how the different parts of the datasets are coded and the meaning of certain pieces of information. Students should first try to explain these parts on their own, but the answers are provided below:
    1. The **Locations** are coded by abbreviations of where the water was taken, which is described in the **Comment** column.
    2. "Standard" and "Instrument Response" are measurements used to help calibrate the water collection and testing devices. They are included in the datasets to show the accuracy of the tests.
    3. Only 2 locations have Raw or Finished attached to their data because both of those locations are from Water Treatment Plants (WTP).
       - ○ Raw = before processing
       - ○ Finished = after processing (what goes to homes and businesses)
    4. The samples used for calibrating the water testing devices do not have sample collection dates. Only the actual water test samples from the different bodies of water have collection dates.

TASK 6: Data Wrangling & Cleaning (15-20 min)

11. Both datasets contain raw data. They must be cleaned and organized (data wrangling) before they can be analyzed. The lesson will walk students through the data wrangling process.
12. First, allow student groups to look at the "messy" data and discuss some things that could be done to clean them.
13. Then look at the datasets together as a class and discuss the following:
    - ➢ What ideas did their group come up with for how to clean the data?

- ➢ Identify as a class how the data is *tidy*.
    - ○ Each row is a single **observation**
    - ○ Each column is a single **variable**
    - ○ Each **value** is in a single cell

14. Next, the lesson will walk students through how to clean the data to get ready for analysis. Make sure they are cleaning BOTH datasets.
    - ➢ First students will make a copy of each tab. **Make sure that students are not making changes to the original data**.
    - ➢ Next students will delete the **variables**/columns with “Standard Response / Instrument Response” because that information is irrelevant and does not fit into the specific problem they are trying to analyze.
    - ➢ Then students will delete **observations**/entries that say “Calibration standard”, “standard”, “Blank”, “Method blank”, “CC=Calibration check”, or “QC=Quality Control” because again, they are irrelevant.
    - ★ Students will analyze this data in CODAP. This platform does allow for hiding and deleting **variables** and **observations**, as well as making new spreadsheets.

TASK 7: Data Analysis (15 - 60 min)

Now it is time for students to work with and manipulate the data.

15. BEFORE students make any changes to the data or the spreadsheet, they should first **make a new tab** and then copy the Location column into the tab. It is advised to make a new tab each time you make any changes to the data, such as deleting columns or entries or pulling out certain entries to analyze separately.
16. It is important for students to have the time and freedom to brainstorm their own questions and ideas about how to analyze the data. Allow student groups discussion time before offering suggestions. Students may have ideas about what they want to investigate but are unsure about how. Offer scaffolded suggestions to help them get through these roadblocks as much on their own as possible and time allows.
17. Encourage students to use basic statistical skills, such as finding the mean, median, and range for data from each location.
18. Students should discuss in their groups what trends or possible correlations they want to look for.
    - ★ Discuss single variate, bivariate, and multivariate data.
        - ○ There are limited variables in this data (Location, Sample Date, Raw vs Finished), thus there will be limited correlations to potentially analyze.
    - ★ Some basic data analysis suggestions are listed in the lesson.
    - ★ Again, while this lesson instructs changes to be made in Sheets, students can hide **observations** in CODAP to visualize different trends.
    - ★ After some basic analysis, encourage students to find additional data to add to their analyses.
        - ○ Ex: longitude/latitude of each location, what towns are near the sample sites and their longitude/latitude, the demographics and household incomes of people in those nearby towns.

19. Data Analysis Examples:

- ★ **The levels of PFAS in our drinking water**
    a. Make a **new tab** and **copy variables**: location, concentration, and finished vs raw water.
    b. **Sort** the Finished vs Raw column and delete all entries that do not have data in this field.
    c. **Download** this spreadsheet as a **CSV** file and **Import** it into CODAP. (instructions on how to do this are in the lesson)
    d. Create a **Graph** and drag the concentration data to the x-axis.
    e. Drag the Finished vs Raw data to the graph (not to either axis, just into the middle of the graph).
    f. Now students can see concentration levels of Nafion BP2 in the drinking water supply both before and after the treatment process.
    g. Allow students to discuss what they notice and discuss the outliers.
    h. Encourage students to continue to ask questions as they analyze the data in different ways.
- ★ **Find the mean (average) concentrations at each location.**
    - ● Mean calculations can take place in both Sheets and CODAP.
    - ● Discuss why we should take multiple samples at each site.
        - ○ Because water is constantly moving and concentration levels can vary as the contaminant moves through the water.
        - ○ Discuss how numbers are more accurate/precise the more decimal places there are. BUT when doing calculations with data, you have to round to the decimal place where the original measurement was taken. For example, if your calculations give you numbers in the hundredths, but the original sample concentrations were only in the tenth place, then you must round to the tenth place. You CANNOT create false accuracy with mathematical calculations.
- ★ **Plot the data on a bar graph to compare concentrations at each location.**
    - ● Find the range of concentrations at each location.
        - ○ We would particularly be interested in the highest concentration levels.
    - ● Create **box plots** for the range of concentrations.
        - ○ Decide what ranges to look at:
            - ○ All concentrations at all locations
            - ○ Average concentrations for all locations
            - ○ Concentrations for Raw and Treated Drinking Water.
        - ○ This would be an especially good analysis of concentrations in drinking water.
        - ○ Math questions:
            - ➢ Calculate the inner quartile range.
            - ➢ How much of the data falls in the first two quartiles?
            - ➢ Are the distributions normal or skewed? Explain.
            - ➢ Math 3: Is this a normal curve? Explain.
            - ➢ Math 3: Calculate the standard deviation.
            - ➢ How do outliers affect the range or the mean?
            - ➢ If we removed an outlier, what would it affect more: Mean, Median, or Range? Will it impact the distribution?

- **Discuss how outliers affect the mean, median, range, and distribution.**
  - Look at what happens to each if outliers are removed
  - Discuss how keeping and removing outliers can impact how we look at the data and what decisions we make based on the data.
    - Some locations had a single sample that was really high or really low. Do we include those? Why or why not? What potential impacts would keeping or removing them have on the overall analysis for those locations?
- Graph on location's concentration levels over time on a scatter plot.
  - This requires **bivariate data** - concentration vs time
  - First, create a new column and turn "Date" into numbered day (Day 1, Day 3, Day 6, etc)
- **Look at the concentrations for both Nafion BP2 and PFOA downstream of the Chemours plant vs upstream (create a new column to add this categorical data)**
  - Since Chemours produces Nafion BP2, then it should be higher downstream of the plant. Is that the case? Does it decrease the further away you get from the plan? (indicates diluting)
  - What are the PFOA concentrations downstream vs upstream from Chemours? They should be evenly distributed since these chemicals should have been phased out since 2015.
    - If higher in certain areas, look to see if there are secondary contamination sites.
    - If higher specifically directly downstream from Chemours, are they still producing this chemical?
- **Try to map the data.**
  - Students will need to look up the longitude and latitude for each location and add them as separate columns.
  - Find the mean concentration for each location and map them to see if there are any trends in concentration depending on the location.
    - Ex: Are they higher/lower upstream vs downstream from Chemours?
  - Compare the amount of each contaminant moving through the environment over time.
    - Math 3 skill: **bivariate** data
      - Recursive equation - what you have now is what you had before plus some "change" (added or removed)
  - Students could look up the amount of water that flows through the Cape Fear River over time and compare it to the concentrations of each contaminant.
    - Math 3 skill: Equilibrium
      - The amount of water flowing through the environment over time.
  - Identify communities near the locations.
    - How could PFAS contamination affect these communities?
    - How do we decide how far out from the sample sites to look?
    - Discuss using a radius for distance and its limitations.
      - Radius does not account for topography.

- **Students could look up the weather data for each date and location.**
    - Are there any correlations between weather and concentration? (infer evaporation vs precipitation)
- **Have students visit the Environmental Working Group’s interactive map: PFAS Contamination in the U.S.**
    - Look for clustering:
        a. You can click the legend to turn the site selections on or off.
        b. Students could start by only looking at the “Other Known Sites”, then add the “Drinking Water: Above Proposed Limit”.
            - Do they notice any clustering trends?
        c. Turn on the “Military Sites”.
            - Do they notice any stronger clustering trends associated with military sites?
        d. Look for other clustering trends across the US.
            - Students could download a topographic map of the US. (See example)
                - Looking at the topography, do they notice any other trends?
                - (To help, students could paste both maps on top of each other in a program such as Google Slides, then make one map slightly transparent. After properly aligning the maps, students can see the topography of the land in relation to the drinking water contamination sites.
                - Discuss how water moves across the land: runoff, groundwater, high elevation to low, etc.
- **Have students look up the mean, median, and range of incomes (and other demographic information) for people living in the communities impacted by PFAS contamination or near contamination sites.**
    a. Create a new column(s) for this information.
    b. Create a scatter plot with the income and PFAS concentrations.
        - Add a best fit line.
    c. Are there any trends?
    d. Discuss the socioeconomic impacts:
        - Of the people in communities impacted by contamination.
        - Of the industries (manufacturing & military) that produce PFAS discharge that are housed in these communities. (trade-offs between economy and health)
        - Is the Median Income the best measure when considering impact? (Discuss the Vast Majority Income measure of inequality.)
- **More advanced students could try to compare NC cancer incidents with contamination sites.** (Instructions on how to import tables from PDF to Excel)
    - 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 for further research. **Discovering correlations between certain variables cannot be directly interpreted into causation**.

TASK 8: Communicating Your Findings (15 - 20 min)

Students should communicate what they've discovered in the data.

20. Have students report their findings. (Use a variety of reporting options: written, data visualizations, digital presentations, oral presentations, discussion boards, etc.)
21. Students should discuss what trends they found, what correlations they may have looked for but didn't find, what further information or data they might need, and what impacts these findings could have on the populations or communities where the data was collected.
    - ➢ It is important to again point out what data is missing or WHO is missing from the data.
22. Consider the Extensions listed in the Student Lesson:
    - Create a public information poster
    - Write a 2 minute speech for public comment to your state legislature
    - Design your own water quality testing experiment
    - ➢ You could set up a Gallery Walk or Expo where students can use their individual research to tell a story.

TASK 9: Data for Social Good (20 min)

23. In authentic data experiences, students should be given the time to discuss and ideate what to do with their findings, such as arguing for policies, public information campaigns, and human health monitoring (ideas are provided in the lesson, but students should explore their own ideas as well). This is an incredibly important part of the data process.
    - ➢ It is also incredibly important to highlight that **Data is Information**.
    - ➢ It comes with a certain degree of error, variability, exclusion/inclusion, and bias.
    - ➢ Data can help us recognize patterns and trends, which in turn can help us make decisions.
    - ➢ BUT data should not be treated as absolute truth, and in reality, we need multiple sets of data to reinforce if those patterns or trends are real. We also need different types of data to help us make educated decisions.
    - ➢ And we must **ALWAYS consider who is being impacted by the decisions we make based on the data we are using.**

Related Research Papers

Longitudinal assessment of point-of-use carbon filters for removal of per- and poly-fluoroalkyl substances from private well water

Measurement of Novel, Drinking Water-Associated PFAS in Blood from Adults and Children in Wilmington, North Carolina

Other Resources

Links to Nearpod version (not updated) of the Student Lesson: Tasks 1-4, Tasks 5-6

Guidance on PFAS Exposure, Testing, and Clinical Follow-Up

PFAS-Terminology-Breakdown

EPA: PFAS Explained

NC PFAS Testing Network