DSA Data Stories in the Classroom

Lesson Title: Learning with Data - NC Survey on Human Health and the Environment
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 the results from the 2021 and 2022 NC Surveys on Human Health and the Environment. These datasets are more complex using Likert Scale data and a large number of variables (100+). The lesson 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.

Survey Data Standards Alignment

Social Studies:

- People’s trust in politicians (political ideology)
- Civic engagement
- Media sources and bias

Earth & Environmental Science:

- Most pressing environmental issues in the state
- PFAS vs COVID perceptions

Biology:

- PFAS vs COVID perceptions
- Understanding of heart and lung health

Math:

- Play with regressions
- T-tests
- Difference of means

TEACHER GUIDE

Case Study: NC Survey on Human Health and the Environment

Total Time: 1 hr 23 min - 2 hr 13 min

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

TASK 1: Data Collection (How) (5 min)

1. Allow 3-5 minutes for students to brainstorm how they would collect information from the public about their knowledge, perceptions, and opinions. Students should record their brainstorming ideas somewhere that can then be shared with the class, such as on Padlet or chart paper.
   - ➢ A **Question Burst** activity can also be great here.
     - ○ Students write a question on a sticky note and then call out the question and stick it to the table.
2. Next, have each team share out their top 1-2 ideas. Collect the ideas on the board (allowing duplicates). Guide students to settle on the idea of surveying residents and then discuss how they might do that. All groups don't need to agree on HOW they would survey the public.
   - ➢ For context, this survey was originally going to be sent out by mail, but that was expensive. So instead it was sent out by a panel survey company. You pay them to implement the survey to X number of people and to stay open until X number of people from different population subgroups respond so that it's relatively representative of the NC population

TASK 2: Data Collection (What) (8 min)

3. Allow 5-8 minutes for students to brainstorm survey questions. Offer encouragement and guidance to each group to think about **social and environmental** factors that may be impacting human health in NC. If students conduct a web search of ***current environmental health threats to NC residents***, they should find things like climate change, water and air pollution, chemical contaminants such as PFAS and pesticides, and respiratory diseases such as COVID-19.

   Also, help students consider the various attributes that make up one's social demographics, as well as all of the ways people attain and interact with information (various types of media, voting, etc).

TASK 3: Reviewing and Questioning the Data (10 min)

4. **3a:** Students should first look at both datasets. This data is simply a collection of numbers and has no context. Allow students 2-3 minutes to record questions they may have about these datasets.
5. **3b:** The Codebooks for each dataset show what questions were asked and how the data was cleaned (what they did with the blank cells). Each dataset has a second tab with the cleaned data, which will be addressed in Task 4. Allow students 2-3 minutes to record additional questions they may now have about these datasets.
6. **3c:** First allow students 2-3 minutes to discuss possible answers to the contextual questions, then provide them with the actual answers.
   1. How was the data collected?
      a. An online survey company called Dynata was used. They wait until they get enough responses from people from different race and age groups to be generally representative of the entire state.

2. Who collected the data?
    a. The survey company collected it from people who signed up to do online surveys.
3. What was their intent (purpose)?
    a. Get a representative sample to understand how people in NC think about environmental issues.
    b. Get information about the environment and health, and see if there are patterns.
4. Who is represented in the data? Who is missing?
    a. People who sign up to take online surveys are represented, so people who don't are inherently missed.
    b. Ex: Older people? Lower socioeconomic status people?
    c. The good news is that they do try to be representative by not just taking the first X number of people who respond, but waiting until they get the right percentages of different population subgroups.
5. What is represented in the data? What is missing?
    a. [This is objective. Have students look at the questions asked and identify other relevant information that could have been collected.]
6. Are there any potential biases that you can recognize?
    a. Question wording can lead to bias. Online delivery methods can as well.

TASK 4: Data Wrangling & Cleaning (10 min)

7. **4a:** Both datasets have already been wrangled and cleaned. Data wrangling is the most important and most time-consuming part of data science and it is important for students to talk through and understand this process.
    - ➢ As a class, look at each dataset:
    - ➢ In the first tab, identify as a class how the data is tidy.
        - ○ Each row is a single **observation**
        - ○ Each column is a single **variable**
        - ○ Each **value** is a single cell
    - ➢ In the second tab, identify as a class how the data is clean and what further cleaning they could do in order to analyze the data to answer their own questions.
        - ○ Hide/Remove duplicate or irrelevant observations
            - ■ Irrelevant observations are those that do not fit into the specific problem you are trying to analyze.
        - ○ Fix structural errors
            - ■ Strange naming conventions, typos, or incorrect capitalization
        - ○ 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
        - ○ Handle missing data
            - ■ Many programs will not accept missing values
            - ■ Options:
                1. Drop observations that have missing values, but then you will lose information
                2. Input missing values based on other observations, but you will lose the integrity of the data because you may be operating from assumptions and not actual observations

3. Alter the way the data is used to effectively navigate null values

- **This step has already been completed and is described in the Codebooks:
  - “No Answer” is now -99 (This indicates that respondents did not provide an answer to each question.
  - “Not Applicable” is now -999 (This indicates that respondents did not receive the question. Some of the questions were given to all respondents, but some of the questions were randomly selected among the respondents.)

TASK 5: Data Analysis (15 - 60 min)

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

8. BEFORE students make any changes to the data or the spreadsheet, they should first **make a new tab** and then copy the CASEID 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 **variables** to analyze separately.
9. It’s 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.
10. Encourage students to use basic statistical skills, such as finding the median and mode.
11. Students should discuss in their groups what trends or possible correlations they want to look for.
    - ➢ Discuss single variate, bivariate, and multivariate data.
    - ➢ Discuss numerical vs categorical data
    - ➢ Some basic data analysis suggestions are listed in the lesson.
    - ➢ While this lesson instructs changes to be made in Sheets, students can hide **observations** and **variables** in CODAP to visualize different trends.
12. Survey data is typically collected on a Likert scale. This type of data is ordinal and the mean can be calculated in some instances, it is better measured by the median and mode.
13. Data Analysis Examples:
    - ➢ Calculate the mode for each question and the frequency of each response.
      - Calculate the percentage of respondents who answered each response option. (Ex: 80% of respondents strongly agreed that…)
    - ➢ Plot the answers to one specific question on a histogram.
      - Make 2 different histograms and stack them.
      - In CODAP, Click on THE ruler and select “show %”.
    - ➢ Create composite scores for related questions:

      a. Group related questions together in a new tab. (Ex. on 2021 Codebook Q32-53 are all about heart & lungs.)

      b. For each person, add up all of their answers to each question in that group.

      c. Find the mean answer.

        - Make a bar graph from just the composite scores of one group.
        - Compare the composite score to some categorical variable (i.e. sex)
    - ➢ Create box plots on bar graphs.
      - 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?
    - Create a two-way / cross-tabulation table
    - For open-ended questions, find themes in the answers and create categories.

TASK 6: Communicating Your Findings (15 - 20 min)

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

14. Have students report their findings. (Use a variety of reporting options: written, data visualizations, digital presentations, oral presentations, discussion boards, etc.)
15. 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.

TASK 7: Data for Social Good (20 min)

16. 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.**