Public Data in the Public Interest: A Spreadsheet-Based Project for High School Computing

Emily Nutwell
enutwell@umass.edu
University of Massachusetts Amherst
Amherst, Massachusetts, USA

Kelsey Badger
badger.60@osu.edu
The Ohio State University
Columbus, Ohio, USA

Course CS0
Programming Language None
Knowledge Unit Programming Concepts
CS Topics Tabular Data, Mathematical Reasoning, Functions
Resource Type Project

Synopsis

Data for Healthy Communities (DHC) is a 15-hour high school project that uses spreadsheets and public data to provide an accessible introduction to data science in the broader context of decision making for complex societal problems. Students work with real-world government data in the context of public health and will learn how to use data as evidence to support an argument for investment in their local communities. The no-code interface of spreadsheet software allows students to explore basic computing concepts such as variables and functions while engaging with authentic public health challenges like air quality, health inequity, and environmental burden. The intention is to lower the barrier for students' first introduction to computing and to present options for embedding data science education in a wider variety of curricular areas.

The project is scaffolded by skill-building modules that use a mix of lectures, class demonstrations, and labs to guide students through exploring, analyzing, and visualizing data to make an argument. Students then apply these skills in a team-based activity that challenges them to use real neighborhood-level community indicators to create a structured presentation to a city council to advocate for community investment. After the student presentations, it is recommended to hold a class discussion to highlight the complexity of real-world problems where there is no one right answer. Rather, complex problems often require teams to consider several factors to build an argument for a proposed solution.

Keywords

Data Science, Public Interest Technology, Computational Thinking, K-12 Instruction, Interdisciplinary Programs (CS + X)

ACM Reference Format:

Emily Nutwell and Kelsey Badger. March 2025. Public Data in the Public Interest: A Spreadsheet-Based Project for High School Computing. In ACM *EngageCSEdu*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3724411

ACM EngageCSEdu, March 2025.

ACM ISBN 979-8-4007-1570-9/2025/03.
https://doi.org/10.1145/3724411

1 Engagement Highlights

DHC uses multiple engagement strategies to increase participation in computing among high school students. Underpinning all of these practices is the decision to develop lessons that are entirely spreadsheet-based. This lowers the technological barrier to engagement for both students and teachers, allowing the activities to be integrated across the high school curriculum.

In addition to the use of spreadsheets, all four of the engagement practices under the NCWIT framework principle "Make It Matter" are incorporated in the design, as described below.

1.1 Make Interdisciplinary Connections to CS

DHC centers on public health challenges that are relevant to environmental science, civics, and healthcare. This breadth creates engagement opportunities for students with a variety of personal and professional interests while also creating multiple avenues for incorporation into existing courses.

1.2 Address Misconceptions about the Field of CS

DHC contextualizes data science as a problem-solving tool that can be used to create public interest technology (PIT) solutions. PIT is an emerging field focused on the thoughtful application of technology in the service of public good [8]. This context enables students to see how data science concepts are relevant to all professions, including public service roles that are not historically associated with computing.

1.3 Use Meaningful and Relevant Content

DHC activities use granular sources of public data that empower students to learn about their communities at the neighborhood-level. Students are then challenged to use this data to make evidence-based arguments for local community investment.

1.4 Incorporate Student Choice

DHC teaches students to subset large datasets to examine the community indicators they consider most important. This leads to classroom discussion about the challenge of analyzing data in real-world contexts, where there may not be a single "correct" answer to solving complex problems.

2 Recommendations

2.1 Prerequisites

The materials include instructions for Microsoft Excel but can be adapted to use other spreadsheet software such as Google Sheets if desired. Teachers do not need expertise in Excel and all skills used in the lessons are detailed in the teacher notes provided. Microsoft Excel provides excellent documentation, and there are several free courses available to provide further support [9]. Students are not required to have prior experience with spreadsheets, but should have completed Algebra I.

2.2 Scaffolding

Students work in teams of three to four students to complete the project. The teams select two to four features from over fifty provided to generate an analysis and provide recommendations based on their work. The materials provided illustrate one specific method to accomplish a given task, such as sorting or filtering. Although there are often several techniques that can be used to accomplish the same task using spreadsheets, students can learn the tool by focusing on one method to avoid confusion. Additionally, students who are novice users of spreadsheets may benefit from first observing the demonstrations. They may become overwhelmed "learning and doing" by trying to follow along with demonstrations before gaining a baseline comfort level with spreadsheets.

2.3 Modifications

The materials can be offered as a standalone unit or incorporated into a larger curriculum and are appropriate for a variety of content areas. The project can also be implemented as a standalone activity. This may be appropriate for more experienced students who do not require the additional scaffolding provided in the complete set of materials.

The project dataset was curated to include data from the local community to provide meaningful context for the students. It is recommended that the dataset be adapted for the location it is being taught and to keep the data current. The data curation process is included in the next section to provide guidance for this modification. The data provided with this resource was downloaded in fall 2023 and primarily focuses on Franklin County, Ohio.

3 Curating Public Data

Public data is not always analysis-ready and is often larger and more difficult to derive meaning from than demonstration datasets. To mitigate these challenges, the public data used in DHC was carefully selected after extensive testing, downloaded in advance of the lessons, and curated to promote ease of use by students. The process used for curating each dataset is described below. While most of these steps were specific to the given dataset, a few activities were applied uniformly to all data sources:

- Data Dictionary: a new sheet named 'dictionary' was added to the data workbook with variable definitions adapted from the documentation of the original source
- Source Information: a new sheet named 'about' was added to the data workbook to provide information on who collected the data and where it can be accessed

3.1 Air Quality Data

The air quality data is used across multiple skill-building modules and instructor demonstrations. It primarily comes from the U.S. Environmental Protection Agency's Air Quality System, which has a query tool for downloading daily data by pollutant and location [6]. The remainder of the air quality data comes from Air Quality Ontario and must be manually compiled from tables of hourly readings for three-day periods [7]. The preparatory work for the Canadian data is significantly more extensive than the U.S. locations, but because the data is used to analyze the impact of Canadian wildfires, it is a worthwhile investment of time.

3.1.1 EPA Air Quality System

- *Select locations and stations:* Many urban counties will have more than one air monitoring station, whereas rural counties may not have one at all. For counties with more than one station, download data from all stations and inspect them for missing or sparse data before selecting the location.
- *Append locations*: Create a single file by appending the exports from each location.
- *Reduce data*: Simplify the file by removing columns that will not be used in the activity. Some of this information can be retained for context in the data dictionary, such as the units of measurement.

3.1.2 Air Quality Ontario

- *Compile daily readings*: Data is only available in rolling three-day periods and needs to be compiled to examine changes over longer time periods. To reduce transcription errors, data should be either copied and pasted or retrieved programmatically.
- *Average hourly readings:* Use an average function to convert from hourly to daily readings.
- *Calculate Air Quality Index (AQI):* Use the AirNow AQI calculator to compute the daily AQI [1]. While it is possible to compile these readings using the same process as above, the Candian index for air quality is different than the one used in the U.S. and would be not be comparable to the EPA data.
- *Append to U.S. data file*: Be sure to check that all data fields match the structure of the U.S. data file first.

3.2 Environmental Justice Index

The DHC project uses the Environmental Justice Index (EJI), a Centers for Disease Control and Prevention project that uses data from multiple federal agencies [4].

- *Filter by county:* Data is downloaded at the state-level and should be reduced to a single county so students can examine indicators in their immediate geographic area.
- *Join with place names:* Each row in the EJI is a Census tract, a neighborhood-level geographic unit represented by an identifying number. A spatial join with place name data, such as the Census Designated Places from the Boundary and Annexation Survey [2], allows students to refer to more familiar neighborhood names instead of tract numbers. A low-tech alternative to this is to have students use a Census reference map to identify where a tract is located [3].
- *Reduce data:* The EJI was intentionally selected for its breadth of indicators, but some reduction is still useful

to make the data less overwhelming for students. One easy reduction step is to remove the columns that begin with E_, F_, or SPL_. These are included for transparency in the original dataset, since they are used to calculate the percentile ranks. Removing them leaves only one column per indicator or index and reduces the total number of columns by almost 40.

- *Add visual formatting:* While it is not generally a good practice to convey meaning only with visual styling, color-coding the student file to match the design of the EJI indicator table [5] can help students make connections between the data and the concepts represented. It is also recommended to paste an image of the indicator table into the data dictionary worksheet to further reinforce this connection.

4 Assessment

Since the goal of DHC is to connect foundational computing skills to data-driven decision making for societal problems, formative and summative assessments focus on students' abilities to support an action based on the analysis of the data rather than assessing the spreadsheet and analytical skills directly. The intention is to emphasize that data analytics is a powerful tool used to support decisions, including public health policies that impact local communities.

4.1 Formative Assessment

Formative assessments are included in several modules to assess students' understanding of the concepts presented. The strategies used in formative assessments include brief presentations or written explanations and structured sentences to reinforce abstract concepts. For example, since the EJI measures health burdens, higher percentile values indicate higher burden. This may be counterintuitive for students who are most familiar with the use of percentiles for test scores, where a higher percentile is more desirable. Using a templated sentence connecting percentile value to meaning, students can practice interpreting the EJI values appropriately.

4.2 Summative Assessment

The project concludes with group presentations of their analysis and subsequent recommendation for community investment in a specific neighborhood. Students are provided strategies to create visualizations based on the claim they are making. For example, if a student team wants to illustrate that a community exhibits high burden in two aspects of the EJI, a scatter plot with that specific community highlighted can be created to illustrate this point. Students are also provided a presentation template that scaffolds the idea of using data as evidence, since this is the primary skill assessed in the presentation.

5 Acknowledgements

We gratefully acknowledge our colleagues Chris Orban and Wonjung Hwang who collaborated on course design. This work is based upon work supported by the Public Interest Technology University Network.

6 Materials

- Overview Document dhc_intro.docx: Provides teacher with an outline of materials, information about the datasets, learning objectives, using Excel, linked resources, and frequently used terms.
- Lesson 1 PIT: Introduction to PIT, exploring PIT careers
  - Lecture slides dhc-lesson1_pit.pptx
  - Teacher notes Lesson 1 Intro to Public Interest Technology.docx
- Lesson 2 Public Health: Introduction to public health concepts, health inequities, and data to be used to support decision making in public health
  - Lecture slides dhc-lesson2_public-health.pptx
  - Teacher notes Lesson 2 Intro to Public Health.docx
- Lesson 3 Data Collection: Data types, where data comes from, Tidy Data principles [10].
  - Lecture slides dhc-lesson3_data-collection.pptx
  - Teacher notes Lesson 3 Data Collection.docx
  - Student-facing worksheet Lesson 3 Data Collection Student Worksheet.docx
- Lesson 4 Data Viz: Fundamental data visualization concepts, choosing graphs, and deception with data visualization
  - Lecture slides dhc-lesson4_data-viz.pptx
  - Teacher notes Lesson 4 Data Viz.docx
- Lesson 5 Analysis Demos: data types and types of analysis. Demonstrations of spreadsheet features such as sorting and filtering.
  - Lecture slides dhc-lesson5_analysis-demos.pptx
  - Teacher notes Lesson 5 Analysis Demos.docx
  - dhc_air-quality_all_Clean.xlsx: Excel file of air quality data which has not been manipulated
  - dhc_air-quality_all_Explore.xlsx: Excel file of air quality data which has been manipulated as described in all demonstrations outlined in the teacher notes
- Lesson 6 Analysis Practice: data types and types of analysis. Demonstration of sub-setting the project dataset and applying the ranking techniques to the project dataset.
  - Lecture slides dhc-lesson6_analysis-practice.pptx
  - Teacher notes Lesson 6 Analysis Practice.docx
  - Student-facing worksheet Lesson 6 Subsetting and Ranking Student Worksheet.docx
  - dhc_data-eji_student-version.xlsx: Excel file of project dataset which has not been manipulated
  - dhc_data-eji_teacher-version.xlsx: Excel file of project dataset which has as described in all demonstrations outlined in the teacher notes
  - dhc-community-garden_handout.pdf: Handout describing benefits of a community garden with respect to the EJI dataset

- Lesson 7 Relationships: Introduction of relationships between two quantitative variables using linear regression and the regression coefficient R^2
    - Lecture slides dhc-lesson7_relationships.pptx
    - Teacher notes Lesson 7 Relationships.docx
    - Student-facing worksheet Lesson 7 Relationships Student Worksheet.docx
    - dhc_data-eji_student-version.xlsx: Excel file of project dataset which has not been manipulated
    - dhc_data-eji_teacher-version.xlsx: Excel file of project dataset which has as described in all demonstrations outlined in the teacher notes
- Lesson 8 Communicating with Data: Communicating with data by building an argument and using data as evidence to support the claim
    - Lecture slides dhc-lesson8_communication.pptx
    - Teacher notes Lesson 8 Communication.docx
    - Student-facing worksheet Lesson 8 Data Communication Student Worksheet.docx
    - dhc_data-eji_student-version.xlsx: Excel file of project dataset which has not been manipulated
    - dhc_data-eji_teacher-version.xlsx: Excel file of project dataset which has as described in all demonstrations outlined in the teacher notes
- Lesson 9 Presentations: Student presentations
    - Lecture slides dhc-lesson9_presentations.pptx
    - Teacher notes Lesson 9 Presentations.docx
- Project: Files that can be used to teach the DHC project as a standalone activity, for students that already have some familiarity with spreadsheet tools
    - dhc_data-eji_student-version.xlsx: Excel file of project dataset which has not been manipulated
    - dhc_data-eji_teacher-version.xlsx: Excel file of project dataset which has as described in all demonstrations outlined in the teacher notes
    - dhc-community-garden_handout.pdf: Handout describing benefits of a community garden with respect to the EJI dataset
    - Lecture slides dhc_slides_project.pptx
    - Teacher notes Project Step1 Health Inequity and the EJI.docx
    - Teacher notes Project Step2 Exploring the EJI - Web map.docx
    - Teacher notes Project Step3 Exploring the EJI – Excel.docx
    - Teacher notes Project Step4 Presentation prep.docx
    - Teacher notes Project Step5 Presentations.docx
    - Student Worksheet - Subsetting and Ranking.docx for use in Step 3
    - Student Worksheet - Relationships.docx for use in Step 3
    - Student Worksheet – Data Communication.docx for use in Step 4

References

[1] AirNow. AQI Calculator. Retrieved from https://www.airnow.gov/aqi/aqi-calculator-concentration

[2] Census Bureau. Census Designated Places. *Census.gov*. Retrieved August 23, 2024 from https://www.census.gov/programs-surveys/bas/information/cdp.html

[3] Census Bureau. Reference Maps. *Census.gov*. Retrieved August 23, 2024 from https://www.census.gov/geographies/reference-maps.html

[4] Centers for Disease Control and Prevention and Agency for Toxic Substances Disease Registry. 2022. Environmental Justice Index. Retrieved January 12, 2025 from https://eji.cdc.gov/

[5] Centers for Disease Control and Prevention and Agency for Toxic Substances Disease Registry. 2023. Environmental Justice Index Indicators. Retrieved January 12, 2025 from https://www.atsdr.cdc.gov/place-health/php/eji/eji-indicators.html

[6] Environmental Protection Agency. Download Daily Data. *Outdoor Air Quality Data*. Retrieved from https://www.epa.gov/outdoor-air-quality-data/download-daily-data

[7] Ministry of the Environment Government of Ontario. Ottawa Downtown: Hourly Fine Particulate Matter Readings. Retrieved from https://www.airqualityontario.com/history/pollutant.php?stationid=51001&pol_code=124&showType=table

[8] Tara Dawson McGuinness and Hana Schank. 2021. *Power to the Public: The Promise of Public Interest Technology*. Princeton University Press, Princeton. Retrieved from https://muse.jhu.edu/pub/267/monograph/book/81957

[9] Microsoft. Excel help & learning. *Microsoft Support*. Retrieved August 23, 2024 from https://support.microsoft.com/en-us/excel

[10] Hadley Wickham. 2014. Tidy Data. *J. Stat. Softw.* 59, (September 2014), 1–23. doi:10.18637/jss.v059.i10