CURATING DATASETS TO SUPPORT MIDDLE SCHOOL STUDENT INQUIRY

Jan Mokros[2], Andee Rubin[1], Jacob Sagrans[2], and Traci Higgins[1]
[1]TERC, 2067 Massachusetts Ave., Cambridge, MA 02140, United States
[2]Tumblehome, Inc., 201 Newbury St. #201, Boston, MA 02116, United States
jan@tumblehomelearning.com

We examine how developers of data science curricula determine what makes a pedagogically effective dataset enabling 10–14 year-old students ("middle school" in the United States) to engage in the data investigation cycle by posing their own questions about relationships among variables. We describe strategies for curating existing datasets to address goals for learning about data, and for optimizing the use of these datasets once they are curated. We investigate how data science educators can transform existing datasets into ones appropriate for students with little data experience, drawing on our experience working with several publicly available datasets, which students explored in CODAP (the Common Online Data Analysis Platform) (Concord Consortium, n.d.).

INTRODUCTION AND APPROACH

Though there are millions of datasets available to the public (Noy, 2020), substantial work is often needed to make these datasets useful in educational settings. Educators may be able to examine existing datasets and find a reasonable match between their instructional goals and a dataset that has already been curated, but it is often difficult to find datasets that are a match or that enable students to conduct their own independent data investigations. Additional curation involving organizing, selecting and structuring data is often needed to make a dataset more accessible to students.

Our recommendations for curation of data for classroom use start with the premise that while students should be "awash in data" (Erickson, 2022), they should not drown in the data. Messy data are an inevitable part of data exploration; however, we agree with Erickson that students must start making "data moves" (such as filtering, grouping, and summarizing) relatively quickly in order to be engaged in data investigations. To do this, they must have datasets that are accessible. The theoretical approach to this work involves developing a "framework for action" (diSessa and Cobb, 2004, p. 83) that is based on many iterations of identifying pedagogical goals, curating datasets to match these goals, observing students using the data, and subsequently making changes or annotations to the dataset. As Nilsson and colleagues (2018) have pointed out, the primary goal of descriptive research like this is to "distinguish, narrate, and characterize teaching and learning processes." In our work, this translates to describing the ways in which middle school students use, explore, question, and are challenged by a wide variety of datasets with different parameters. In this paper, we make recommendations based on the parameters of datasets that we have found to be effective. The recommendations are not exhaustive but can be seen as starting places for dataset curation.

In our work, we have curated dozens of datasets, including datasets on COVID infections, shrimp viability in pools, puffin characteristics and behaviors, injuries and trips to emergency rooms, and ticks and Lyme disease. In our pedagogical work with these datasets, we are primarily addressing the first three key understandings in the GAISE II framework, namely: 1) Questioning in statistics; 2) Considering different data and variable types, and; 3) Multivariate thinking (Bargagliotti et al., 2020). Collectively, the data projects we have developed and researched with students reveal the following factors that we believe educators should consider as they curate datasets.

As the reader considers these recommendations, two things should be kept in mind: 1) While our recommendations are primarily tool-independent, our perspectives are shaped by the capacities of CODAP, which provides users with a rich selection of graph-construction and manipulation tools that do not require coding, and; 2) No matter how well curated a dataset is, it is important for educators to explore the data for themselves before using it with students, to determine the extent to which there are interesting patterns that can be discovered, as well as to determine potential challenges.

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada.

RECOMMENDATIONS FOR CURATING DATASETS

1. Number of cases

Students need to have enough cases to address their data questions and make meaningful comparisons. Unfortunately, most teachers select datasets that are too small to achieve these goals. For example, Rosenberg and colleagues found that 79% of teachers who used existing datasets chose ones with 20 cases or fewer (2022, p. 11). Only 18.8% used "large" datasets with 100 or more cases. The over reliance on small datasets is likely a relic of times when students would calculate statistics by hand; given that most teachers are now using software that can calculate these and other quantities easily, that restriction is no longer applicable. Often, instruction is limited when using such a small number of cases. When students employ digital tools like CODAP, it is technically easy to work with datasets consisting of hundreds of cases. In our experience, more cases do not necessarily make data investigation more difficult.

It is particularly important to begin with a larger dataset when there are categorical variables with many levels that students are likely to use to subset the data. Otherwise, cell sizes can become so small that the ability to draw conclusions is limited. In a dataset dealing with injuries, we included more than 800 cases and found that students could comfortably work with this number (Higgins et al., 2023). Because the dataset included categorical variables with many levels (e.g., various different body parts injured), a large number of cases was needed to ensure there were cases belonging to each level of these variables. Even with this larger sample size, students sometimes expressed the need for more data, especially when they found potentially interesting patterns but low cell frequencies. For example, there were relatively few burn injuries in the dataset, and these were predominantly in young children and in older adults. Some students wondered if it was true that burns occur more often in these age groups, or whether the number of cases was too small to be confident in this pattern.

2. Number of variables

Professionals almost always work with datasets with many variables, but most students do not. In the Rosenberg and colleagues study, the 79% of teachers who used existing datasets not only selected ones with few cases but also with only one or two variables. (p. 11). There is no need to restrict the number of variables, and achieving the instructional goal of multivariate thinking requires more than two variables. Introducing students to data investigations by looking at one or two variables is still an option with datasets with many variables: one can start by focusing on a subset of variables and then move on to consider more variables as students wish to explore additional relationships. Schanzer and colleagues (2022) have suggested guidelines for the number of variables, stipulating that a dataset for teaching should have at least three numerical variables and at least two categorical variables. Most of the datasets we have used with middle school students have at least five variables and have almost always included a mixture of categorical and numerical variables.

How many variables are too many? Scientists routinely work with dozens or hundreds of variables. But it can be overwhelming or confusing for students to consider many variables. We suggest students be given no more variables than can comfortably fit on a screen. In tables, variables generally appear as columns, and 10–12 columns can be examined at once on a typical laptop screen.

3. Types of variables

The types of variables in a dataset determine the kinds of analyses that can be conducted. Describing the distribution of a numerical variable requires different statistical concepts and terms than describing the distribution of a categorical variable; regression is an appropriate technique for exploring the relationship between numerical variables, while two-way tables are appropriate for relationships among categorical variables. But there are other possibilities that are worth keeping in mind. Even with a small number of variables, students can pursue multiple analyses by filtering the data to look at a subset of the cases. Sometimes a relationship between two variables only holds for some cases or changes from one subset to another. Categorical variables can be used to divide a dataset into several subsets that can be analyzed separately (for example, in working with a dataset on ticks and Lyme disease, students divided the data into subsets of US states where deer ticks were found vs. not found, and perhaps not unsurprisingly found that rates of Lyme disease were much higher in states where deer ticks were found). A numerical variable can be recoded into a categorical

variable that can then be used to filter the data (for example, dividing into groups or "bins" of a specified numerical range, such as age by decade). Considering these possibilities before introducing a dataset to students will help educators determine which tool capabilities are important to introduce to students.

4. Variability

Since variability is an important characteristic of many datasets, most of the variables in an educational dataset should exhibit substantial variability. Lack of variability in one or two out of many variables is not a problem as long as it is not a surprise to the teacher or curriculum developer and is handled appropriately in suggested analyses. For example, in a dataset of 256 Atlantic puffins in Maine, the vast majority of puffins come from one of two islands; in this case, having students analyze the relationship between location and other variables is unlikely to yield interesting results. On the other hand, it may lead to questions about why there are fewer puffins on the other island and could serve as a baseline for comparing geographical distributions of puffins in the future.

Lack of variability in numerical variables is also important to note. In a dataset of weekly stream measurements from the El Yunque rainforest in Puerto Rico, the pH of the streams is quite stable, around 7.1. Students interested in the impact of water conditions on shrimp populations in these streams were disappointed that pH had little effect (since it rarely changed). On the other hand, similar to the distribution of puffins on Maine islands, the lack of variability in pH levels provides a valuable baseline against which anomalies such as shifts in pH after hurricanes would stand out.

5. Complexity of variables

At the simplest level, variables involve counts, for example: "How many people have had Lyme disease in Vermont?" But many measurements need to be expressed as rates because they are measured across differently sized units. In our experience, students readily see that it is "unfair" to compare the numbers of COVID infections in the states of Maine and Mississippi because of the different-sized populations. To make comparisons fair, one needs a rate. Using rates means that it is important to be explicit about both the numerator and the denominator. In the case of COVID, some of the relevant questions about the numerator (COVID infections) include: Are all COVID infections being counted or are some going unreported?, and; Where are COVID cases counted that are reported by people who are visiting the state, but are not residents? Questions about the denominator (the overall population) might include: How is the state population determined?, and; How long ago was the census taken and has the state's population changed since then?

In data collected and made available by governments, variables are often even more complex than a simple rate. For example, Gross National Product (GNP) is a frequently used measure of a country's economy. The process of calculating GNP, however, is far from straightforward and is not easily understood by middle school students. In such cases, it is important to give students an opportunity to consider how to interpret such a measure, even if they do not understand all the details of how it was calculated. In such situations, the notion of benchmarks can be especially helpful. Comparing a measure to a quantity for which students already have an intuitive understanding can demystify complex variables. In the case of GNP, an easier measure for students to relate to could be GNP per capita, which is a number that they might be able to relate to their family income.

When curating a dataset, it is important to consider the overall complexity of the variables, since understanding each one can take significant effort. It is advisable to have several simpler variables as well to provide less challenging options.

6. Appropriate intervals for time-series data

Examining change over time is an important mathematical and scientific skill. Middle school students typically have not encountered calculus, but they are capable of identifying trends on a graph by noticing slopes that increase or decrease steeply or slowly and looking for periods of time where there is little change. But which time intervals should be selected to enable students to detect trends? This depends on the phenomenon being examined: For example, studying how water temperature increases as it is heated to a boil involves data collected in seconds, while looking at climate change involves data collected over years or decades. In our project on the COVID pandemic we looked at data reported daily to reflect the fact that national and international health organizations reported daily

infections during much of the pandemic. The pedagogical goal was for students to use their knowledge of the virus, as well as their knowledge of public health policies, to interpret trends in infections over time. We chose daily data because of its ubiquity, but as the pandemic wore on, data were reported less frequently, often on a weekly basis. The change in reporting led us to reconsider the unit of analysis with respect to time.

Working with ecological data also highlights issues surrounding measurement intervals. In data from the Puerto Rican rainforest, some measurements are made daily (e.g., air temperature and precipitation) and others are only made weekly (e.g., number of shrimp in a particular pool). When preparing a dataset for students, we had to decide whether to record data weekly (which required aggregating daily measurements) or daily (which would mean there was substantial missing data or, alternatively, repeated data). In curating other datasets, we found that agencies make changes in their reporting timelines and that we must either adjust our own dataset timelines (to focus on periods of consistent reporting) or annotate the dataset to enable students to interpret trends in light of irregular reporting. There is no one right answer for these situations; the choice of time interval depends on the instructional goals of the educator and the experience of the students.

7. Data structure

Even when the content of a dataset is decided, there are decisions to make about how the data are structured. There is no firm rule for these decisions, but it is important to think ahead about which analyses are supported by any given structure. For example, data science makes a distinction between "wide" and "long" formats for data. In general, wide formats have many columns and long formats have many rows; each format facilitates particular types of analyses.

Consider the following example of ecological data from Puerto Rico. Scientists are interested in how the populations of three different species of shrimp change over time, and they sample shrimp in each of three rainforest pools each week; therefore, there are nine weekly data points. We could organize the data by considering each row/case to be a week and have nine separate columns, each of which is a measure of a particular species in a particular pool, as shown in Table 1 below. This "wide" organization facilitates in-depth investigations of a particular species in a particular pool over time since all of those measurements are in a single column.

Week	Average shrimp per trap, species 1, pool 0	Average shrimp per trap, species 2, pool 0	Average shrimp per trap, species 3, pool 0
2	22.3	34.7	0.3
3	21.7	48.3	0

Table 1. Example of wide data format. Excerpt from the full table shows two weeks of observations for the three species in pool 0.

Another possible data structure would be to consider each measurement to be a row/case and the columns to indicate which species was measured and which pool the measurement was taken in, as shown in Table 2 below. This "long" format has fewer columns and more rows than the "wide" organization. This second data structure facilitates comparisons between species and/or pools but requires different data manipulations to isolate measurements of particular species or particular pools.

week	species	pool	Average shrimp per trap
1	Species 1	P0	68.3
1	Species 2	P0	36.5
1	Species 3	P0	0.6

Table 2. Example of long data format. Excerpt from the full table shows observations for the three species in pool 0 in week 1.

The contrast between wide and long format raises questions about not only what the variables are in a dataset but also what constitutes a case; in the example above, for instance, is a case one weekly observation for all of the species and all of the pools, or is it a weekly observation of just one specific species in one specific pool? Considering case structure more broadly, we have found that it is important

to clarify the definition of a case for students. Some students are prone to think of cases as individual people or animals, but sometimes the case definition is more complex. In the injury dataset we used, for example, the case is an injury, not a person. To underscore this point, we added this annotation to the dataset: "Each case is an injury that happened in the last six months. It is not a person. Some people have multiple injuries.' We also reminded educators to model observations that were appropriate to the case structure, e.g., " there were more injuries in July than other months," rather than "there were more people injured in July than other months.

8. Metadata and measurement issues

Most datasets are accompanied by metadata describing the variables and how the data were collected. Metadata can inform the construction of simple variable names that are understandable to students. The metadata provide descriptive information that should appear along with the data, such as how the data were collected, the sampling process, and how each variable is defined. As Rubin (2021) points out, information about how and when the data were collected, by whom, and for what purposes is often hidden. Revealing this information gives students a better sense of what is involved in conducting a data investigation as well as what can and cannot be concluded from a dataset.

As students examine a dataset, they are likely to raise questions about data collection and measurement, especially if the analysis yields puzzling results. For instance, when examining a dataset about social media platforms popular among teens, students questioned the veracity of the data, claiming that very few people used the social media platforms that were most common in the data (Rubin, 2019). As students examined the metadata, they saw that the data had been collected several years earlier, and they could examine the data with this contextual knowledge.

With the COVID data discussed above, measurement issues came up frequently. For example, daily infection data were often not released over the weekends, resulting in a backlog of cases reported early in the week. Thus, the infection rate appeared to spike at regular seven-day intervals. We could have eliminated this noise by transforming the data to weekly infection data, but instead decided to keep the daily data and ask students to think about possible reasons for the spikes. Some students interpreted the spikes as being indicative of weekend partying that resulted in infections a few days later, while other students thought people were less likely to see medical professionals over the weekend. These explanations are plausible but would not explain the lack of new cases during weekends. When we reminded students that public health officials probably were not working during weekends, they quickly came up with appropriate explanations for the spikes.

Interrogation of measurement methods is critical when there are anomalies in the data. In studying Lyme disease, students discovered a steep drop in cases in one state (Massachusetts) from 2015 to 2016. Further investigation revealed that the way Lyme infections were counted had changed, as there were new reporting requirements on the part of doctors. This caused the rate of Lyme to appear to drop (Palumbo, 2020). Students need to have experience probing anomalous data points, and this process can be informed by using the metadata.

In the situations described above, some students might reject these as "bad" datasets. However, the experience of working with such data can engage students in thinking deeply about the challenges of "datafying" complex real-world phenomena. Opportunities for interrogating the data, considering measurement issues, and using the metadata lead to more questioning and an understanding of the limits of data interpretation.

9. Issues of cleanliness

Real-world datasets are rarely clean. They often have missing data, anomalous values, and errors. Missing data can appear as empty cells or may be coded as an unlikely value (e.g., "999"). Educators often want to know how much these kinds of values should be modified or eliminated before students encounter the dataset. Our perspective is that these decisions depend both on the instructional goals of the class and on the software students are using. In our work, the goal has been to introduce students to the power and limitations of data, not to teach the technical aspects of data cleaning. Also, CODAP itself deals appropriately with missing values. Thus, we delete entries that indicate missing values and leave these cells blank. We check cells that have a value of 0 to determine if the data is missing (in which case we delete the 0) or that the value should actually be 0 (in which case we leave the value as 0). We do not eliminate cases with missing data because we want students to grapple with

why data might be missing and to draw them into questions about data collection methodology, as described more under the section above about metadata and measurement issues.

CONCLUSION

The recommendations in this paper focus on preparing datasets for classroom use, but they are only effective if classroom activities are carefully designed to encourage students to ask their own questions of the data, query the origins of the data, and debate their findings. Educators can prepare themselves for these classroom discussions by exploring the data themselves ahead of time and imagining how their students might interact with the data; however, any rich dataset is likely to raise unanticipated issues once students are invited to delve into the data. We consider such surprises a hallmark of student engagement and a sign of success.

ACKNOWLEDGEMENTS

The projects reported on in this paper have been funded by National Science Foundation grant numbers DRL-1742255, DRL-1917653, DRL-2241777, DRL-2313212, and DRL-2049061. Any opinions, findings, conclusions, or recommendations presented are only those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Concord Consortium. (n.d.). Common Online Data Analysis Platform (CODAP) [Computer software]. https://codap.concord.org/

diSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *The Journal of the Learning Sciences, 13*(1), 77–103. https://doi.org/10.1207/s15327809jls1301_4

Erickson, T. (2022). *Awash in data.* https://codap.xyz/awash/

Higgins, T., Mokros, J., Rubin, A., & Sagrans, J. (2023). Students' Approaches to Exploring Relationships between Categorical Variables. *Teaching Statistics*. https://doi.org/10.1111/test.12331

Nilsson, P., Schindler, M., & Bakker, A. (2018). The nature and use of theories in statistics education. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 359–386). Springer International Publishing. https://doi.org/10.1007/978-3-319-66195-7_11

Noy, N. (2020, January 23). Discovering Millions of Datasets on the Web. *The Keyword.* https://blog.google/products/search/discovering-millions-datasets-web/

Palumbo, A. (2020, June 5). State Health Leaders Dispute CDC's Claim of Drop in Lyme Disease Cases. *NECN*. https://www.necn.com/news/national-international/state-health-leaders-dispute-cdcs-claim-of-drop-in-lyme-cases/2013230/

Rosenberg, J. M., Schultheis, E. H., Kjelvik, M. K., Reedy, A., & Sultana, O. (2022). Big Data, Big Changes? The Technologies and Sources of Data Used in Science Classrooms. *British Journal of Educational Technology, 53*, 1179–1201. https://doi.org/10.1111/bjet.13245

Rubin, A. (2019). Facebook or Instagram? Teens Explore Data about Technology Use. *Hands On.* https://www.terc.edu/facebook-or-instagram-teens-explore-data-about-technology-use/

Rubin, A. (2021). What to Consider When We Consider Data. *Teaching Statistics, 43*(S1). https://doi.org/10.1111/test.12275

Schanzer, E., Pfenning, N., Denny, F., Dooman, S., Gibbs Politz, J., Lerner, B. S., Fisler, K., & Krishnamurthi, S. (2022). Integrated Data Science for Secondary Schools: Design and Assessment of a Curriculum. *ACM Technical Symposium on Computer Science Education*. https://cs.brown.edu/~sk/Publications/Papers/Published/spddplfk-integ-ds-desn-assm-curric/paper.pdf