Introduction

The Need

High-quality, engaging, relatable datasets are essential for data science education, and early research shows that a student's selection of dataset has a substantial impact on their engagement. There are many datasets freely available online or ready to be provided by industry partners. However, those datasets may not be appropriate for a classroom audience, nor are they guaranteed to include pedagogical outcomes required to teach introductory data science or statistical concepts in a relatable, engaging, and clear way. Additional work is required to select and clean datasets for use in data science classrooms or in data science curricula, which can be a barrier between teachers and curriculum writers creating engaging and accessible data science lessons.

The Action

This spec offers a pathway for individuals to find, clean, document, and upload datasets that can be used in data science tools (like Code.org's App Lab) or curricula (like Bootstrap's Data Science course), modeled after Bootstrap's pilot with Brown University students that eventually appear in Bootstrap's Data Science Curriculum. Its intended audience are industry or academia partners who would like to contribute work hours or datasets to data science education, but need guidance for how best to do so. The intended results are ready-to-use datasets that educators can incorporate into data science lessons.[1]

This spec also codifies a requirement that datasets include a datasheet, adapted from the requirements listed in the research paper Datasheets for Datasets. These datasheets provide necessary context when considering the source and use of data, information about any normalizing or cleaning that was done to make a dataset compatible with the spec, as well as pedagogical considerations for educators and curriculum developers to best inform how these datasets can be used with intentionality within a lesson or curriculum.

Initial Partners

- Shriram Krishnamurthi, Brown University and Bootstrap, shriram@brown.edu
- Emmanuel Schanzer, Bootstrap, schanzer@bootstrapworld.org
- Daniel Schneider, Code.org, dan@code.org
- Zarek Drozda, Data Science For Everyone, zarekd20@uchicago.edu

[1] This spec purposefully does not address datasets that can be instructional for cleaning or organizing "Messy Data". This term is too broad for a specification, as there are many ways that data can be messy. Instead, we defer to individual teachers on how to curate "messy" datasets to use for instruction

The Spec

This spec offers a pathway for individuals to find, process, document, and upload datasets. However, just because data are available doesn't mean it will be a good fit for a classroom! There are many reasons this could happen: eg, the data may be overly sensitive, it may not be age-appropriate, it may demand too much of a teacher, or processing may remove valuable information, etc. The requirements below serve as a high-level "filter" for dataset contributors.

Notes on terminology:

- When we refer to "classrooms", we either don't include a qualifier (ie: Datasets for Classrooms) or we qualify as 'secondary' (ie: Datasets for Secondary Classrooms). Students in these classrooms are presumed to be 10 years of age or older (e.g. - middle and high school in the US).
- "Accessible", "relevant", and "engaging" can be highly contextual to the students in the room and their prior lived experiences. When considering a dataset, at the very least, consider the student audience who will engage with the dataset and consider what choices are available for students to explore based on their interests.

Datasets for Classroom Data Science Spec

Requirement #1) **Dataset Relevancy.** The dataset should be about a topic that is:

- Accessible to students under the age of 18 (the time taken to explain should not be so involved as to detract from the learning goals of the class)
- A mix of relevant and engaging for students

Requirement #2) **Dataset Dimensions**

- Small enough that a 2yr-old Chromebook can sort or filter the dataset without more than a few seconds' lag. (Moving target - a 2yr old Chromebook will be faster than it is now!)
  - As of 2022: less than 10,000 rows
  - As of 2022: between 4 and 30 columns

Requirement #3) **Dataset Columns**

- Column names and descriptive metadata should be accessible to students
  - The *names* of columns should be simple enough to be accessible, while also accurately representing what is in those columns. If no accurate name can be found that is also accessible, the column should be removed from the dataset.
  - If it takes more than a minute to explain the *content* of a column, it should be removed from the dataset.
- The Dataset contains at least 1 categorical and 3 quantitative columns
- Categorical columns containing strictly "yes" or "no" values should be provided in boolean form (using standard `TRUE` and `FALSE` values)

Requirement #4) **Dataset Pre-Processing**

- No blank cells. This may require removing rows or columns with empty cells. Removing data to avoid empty cells must be done with care to avoid impacting the statistical properties of other data.
- Cell *values* should be normalized: dates, names, numbers, etc. should be represented uniformly.

Note: Dataset pre-processing is highly contextual and should be done with care to maintain the statistical and ethical integrity of the dataset. If processing a dataset results in changing the statistical properties of the dataset, then the dataset may not be a viable candidate for this setting.

Requirement #5) **No Sensitive / Confidential Data**

- Datasets for classroom use should not contain any personally identifiable information, passwords, financial information, etc. Consider how the columns in your dataset *could be aggregated to identify individuals*.
- If this isn't a dataset that everyone would want shared publicly, then it is not appropriate for this setting.

Requirement #6) **Free and Open use**

- Datasets should not contain data that "expire" after a certain amount of time (e.g. - data collected under IRB conditions that specify deleting the data after 2 years.)
- Datasets should not contain data that anyone could rightfully demand be removed after the fact. For example, if a survey allows people to retroactively revoke consent, then data from that survey should not be used in a classroom dataset.
- Datasets should be unencumbered by IP or copyright restrictions, licensing, export, or regulatory rules.

Bonus: **Educator's Guide**

Quantitative or Qualitative columns can align to a desired statistics or data science learning outcome. For example:

- A dataset could contain several columns that are correlated, or columns with left or right skew
- A dataset could have outliers that require additional research to explain, which then changes the interpretation of the dataset.
- A dataset for which linear regression will expose a meaningful result
- A dataset that includes desired visualizations, such as a location map or heat map or specific charts.

These items should be made clear in the Educator-Facing Datasheet to help answer the question "How could an educator use this?" and inform an educator or curriculum writer when a dataset is appropriate for a particular lesson.

Once a dataset has been curated to meet the requirements above, it should be saved in a .csv format and a datasheet should be prepared (see below)

The Datasheet

Why a Datasheet?

The Datasheet for Datasets paper outlines several key reasons why dataset authors should work to provide documentation and transparency in the form of a datasheet. Additionally, the information in these datasheets can inform important classroom discussions around the larger social and ethical context of data collection and usage. However, not all datasets have their own datasheet and for older datasets we may not know the answers to many of these questions.

Additionally, if work was done to clean or edit a dataset to meet the requirements above, a new "derivative" dataset has been created specifically for educational purposes. Any changes from the original dataset, no matter how small, should be documented to better understand how this derivative dataset has been adapted or changed from the original dataset. This is especially true when working with data that was curated without the involvement of the original dataset authors and without the original context from which the data was collected.

In an effort to document these derivative datasets and provide context to educators and curriculum writers, we ask all contributors to complete an Educator-Facing Datasheet. This datasheet includes a subset of questions from the Datasheets for Datasets paper, questions about how this dataset was adapted and derived to meet the requirements above, and a place to document the Educator's Guide that helps illuminate how a dataset is aligned to a statistical or data science learning objective.

The Datasheet: Educator-Facing Datasheet for Derivative Datasets

The most up-to-date version of the datasheet can be found at https://bit.ly/educator-datasheet, including any revisions based on feedback from educators and dataset curators. Additionally, this datasheet is maintained as a template in the following formats to help dataset curators easily create datasheets:

- A Google Doc with a table and blanks
- A README.md file on a public GitHub repository

Both of these resources are available to be copied or cloned by dataset curators and used as a template for their own datasheet.

Hosting and Availability

Rather than mandate a centralized repository or submittal process, we ask that contributors host the dataset and datasheet in a publicly-accessible location in a manner of their choosing. This could be through a dedicated webpage to this project, or in a GitHub repository, or even a public Google Drive folder. Accessing the datasets and datasheets should not require a login or access key, and the dataset itself should be in a .csv format.

This approach tries to balance open access and ensuring the most amount of classrooms can access data for their projects, while also providing flexibility for curators in how they would like to make these derivative datasets and datasheets available to the public. For example, Bootstrap hosts several datasets on their website and Code.org has dedicated pages to each dataset in their curriculum (example here).

Once the files are hosted on a publicly-accessible webpage, you can follow these steps to alert various curriculum providers of these datasets:

- **Data Science 4 Everyone**: Submit a resource here
- **Code.org:** follow the steps listed on this webpage
- **Bootstrap:** email contact@BootstrapWorld.org with information about the dataset

Feedback and Revisions

The authors acknowledge the value of feedback and iteration in making sure these guidelines are useful for educators and dataset curators. If you are an educator or dataset curator and would like to provide feedback, please email any of the authors. If you are a curriculum provider interested in adding your name to the list above, please email any of the authors.