HARVARD DATA SCIENCE REVIEW
Just Accepted
DOI: 10.1162/99608f92.4f3ac3da
ISSN: 2644-2353

Understanding the Data in K-12 Data Science

Rotem Israel-Fishelson*, Peter F Moon, Rachel Tabak, David Weintrop
College of Education, University of Maryland, College Park, Maryland

Contribution of the authors [CRediT]:
Rotem Israel-Fishelson: Formal Analysis, Data Curation, Writing – Original Draft, Writing – Reviewing and Editing, Visualization; **Peter F Moon:** Formal Analysis, Data Curation, Writing – Reviewing and Editing; **Rachel Tabak:** Formal Analysis, Data Curation, Writing – Original Draft; **David Weintrop:** Conceptualization, Methodology, Validation, Writing – Reviewing and Editing, Supervision

Abstract

Our increasingly data-driven world is amplifying the need for everyone to develop foundational data literacy skills. In response, a growing number of K-12 data science curricula are being designed to introduce all students to data. These curricula define what data science is at the high school level and directly shape how students are introduced to and understand the discipline. Ensuring these curricula are effective, engaging, and, most critically, equitable is of paramount importance. This paper presents a qualitative analysis of four curricula, focusing on the data used to introduce learners to the field of data science. The analysis uses a series of analytical lenses to evaluate the 296 distinct datasets used across the curricula and identifies trends and best practices in dataset selection. The analysis includes using data collected from high school students about their interests and experiences with data to understand if and how contemporary data science curricula are tapping into students' lived experiences to situate data science learning experiences. The findings show that the curricula use relatively recent and small datasets covering a range of topics and that there is limited learner involvement in dataset selection. Further, the analysis reveals gaps between the datasets used and students' self-reported interests. This work highlights the importance of dataset selection, especially as it relates to supporting learners from historically excluded populations in technology fields. Finally, this paper provides practical implications to assess existing curricula and advances our understanding of how to situate the field of data science in the interests, ideas, and values of today's students.

Keywords: data science education, curricula analysis, datasets, high school

Media Summary

The datasets used in K-12 data science curricula directly shape the way students are introduced to the discipline and shape their impressions of the importance and relevance of the field. These curricula impart fundamental skills for collecting, analyzing, and interpreting the data surrounding students that impact their daily lives. Choosing the right datasets is crucial in creating engaging and inclusive introductory data science learning experiences. Selected compelling and relevant datasets help learners connect with the content and critically communicate whose ideas, values, and cultures are important in the field. Unfortunately, this aspect is often overlooked, making it critical to prioritize dataset selection to ensure that all learners feel welcome and engaged. This study presents a qualitative analysis of the datasets used in four major high school data science curricula: Bootstrap:Data Science, CodeHS, Introduction to Data Science, and YouCubed Explorations in Data Science. Our findings show that the curricula use relatively recent and small datasets covering a range of topics and that there is limited learner involvement in selecting the datasets used. Using data collected from high school students, the analysis reveals gaps between the datasets currently used in introductory contexts and the interests and lived experiences of today's high school students. The findings from this work shed light on the current state of high school data science education with a particular focus on if, how, and when introductory materials are tapping into the data-rich lived experiences of today's youth. This paper provides direction for future revisions and innovations in data science instruction that better situate instruction in the data-rich lives of today's students.

1. Introduction

Data is increasingly surrounding us and affecting our lives in both visible and invisible ways. The rapid growth in the amount of data being created, collected, and analyzed shows no signs of slowing down. Further, the expanding set of uses for these data means the number of ways they are impacting our lives, both directly and indirectly, is rapidly growing. In this increasingly data-driven landscape, understanding the role of data and the technologies and algorithms built around it is becoming an essential literacy (Bargagliotti et al., 2020). This is especially critical for individuals from populations historically excluded from computational and technological fields (e.g., BIPOC individuals, women, neurodiverse individuals, non-native English speakers) as they are disproportionately likely to be negatively impacted by biases and predatory uses of data (Benjamin, 2019; O'Neil, 2016). It is thus imperative to provide opportunities for all youth to learn foundational data science concepts and practices as part of K-12 education (LaMar & Boaler, 2021; Weiland & Engledowl, 2022).

Only recently has data science emerged as an independent discipline in K-12 classrooms. The last few years have seen the publication of several year-long curricula designed for high school contexts (e.g., Gould et al., 2022; Schanzer et al., 2022). These curricula play an essential role in defining data science at the high school level and how students are being introduced to the discipline. As part of these curricula, students not only learn about algorithms, manipulating datasets, and

creating and interpreting data visualizations (Biehler et al., 2022) but also learn that a core aspect of data science is considering how data is collected, produced, and used and the implications and impacts of data use on individuals, groups, and society (Lee et al., 2021).

As the name suggests, data is central to data science. Given the increasingly data-rich world youth are growing up in, there is tremendous potential for situating data science instruction to draw on learners' interests, experiences, and cultures (Lee & Delaney, 2022). Selecting engaging datasets related to learners can foster agency and ownership (Lee et al., 2021) and is especially important for learners from populations historically excluded from computing (Calabrese Barton et al., 2013; Ladson-Billings, 2014). Therefore, it is crucial to consider the provenance and cultural resonance of the datasets used in modern data science curricula, as these are critical for creating inviting, engaging, and equitable data science learning opportunities for all students.

Given the critical role dataset selection plays in engaging learners and signaling whose ideas, values, and cultures are valued by the field, selecting what data students will engage with should be carried out with great thought and care. Selecting the right datasets to situate data science instruction can be the difference between generative, engaging, and equitable instruction that welcomes students to the field or instruction that alienates and further perpetuates existing racial, gender, and socioeconomic gaps. Although the academic literature often highlights the importance of reflecting on youth experiences in the datasets they study (e.g., Stornaiuolo, 2020), little work has been done to understand these datasets more broadly.

In this work, we investigate the datasets used in four of the most widely used high school data science curricula in the United States: Bootstrap:Data Science (Bootstrap, 2022), CodeHS (CodeHS, 2022), Introduction to Data Science (Gould et al., 2022), and YouCubed Explorations in Data Science (YouCubed, 2022). In conducting this analysis, we recognize how these curricula shape the emerging landscape of K-12 data science and seek to highlight the importance of choosing datasets that draw on the cultural knowledge and lived experiences of the youth who will be learning with them. Stated more explicitly, in this work, we pursue the following research questions:

RQ1. What datasets are being used in high school data science curricula?

RQ2. How do the characteristics of these datasets differ across curricula?

RQ3. How do the datasets in high school data science curricula align with the interests and lived experiences of high school students?

To answer these questions, we reviewed four high school data science curricula and identified 296 distinct datasets used across them. We then qualitatively analyzed all 296 datasets using a series of analytical lenses that considered various dimensions of the datasets. These dimensions are Topic, Recency, Dataset Type, Student Choice, Size, Proximity, and Alignment with Students' Interests. The result of this analysis is a comprehensive understanding of the datasets being used to introduce the current generation of learners to data science. The findings from applying these analytic lenses provide direction for future revisions and innovations in data science instruction that better situate instruction in the lived experiences of today's students. Researchers can use these analytical dimensions to assess other datasets, while educators can use our findings as a scaffold for

selecting curriculum materials that will resonate with their students and reflect their interests. This work contributes to the larger goal of preparing today's students to succeed in the data-rich world that awaits them.

2. Literature Review

To situate this research, we draw on prior research on data science education, prior work investigating the role and use of data across K-12 education, and research on social and cultural aspects of data and data science.

2.1. Data Science Education

The substantial increase in the amount of data in the world and its widespread use in various sectors, along with the improvement of computational capabilities, greatly motivated the advancement of data science. Academic discourse and attempts to define data science have been ongoing for over 50 years, with researchers and educators from various disciplines seeking to frame data science as an extension of their fields. Statisticians and statistics educators have sought to frame data science as an extension of statistics, manifested in acquiring "new" skills for investigating, analyzing, and modeling large datasets (Ridgway, 2016). This has led to calls to reform and expand the field of statistics to emphasize the preparation, analysis, and presentation of data (Cleveland, 2001; Donoho, 2017). Others have argued that data science is more closely aligned with computer science (Hazzan et al., 2020) and mathematics (LaMar & Boaler, 2021).

In 2010, Conway proposed a Venn diagram positioning data science at the intersection of three areas of knowledge: hacking skills (i.e., the knowledge and skills to accomplish tasks with computational tools), math and statistics knowledge, and substantive disciplinary expertise (Conway, 2010). Three years later, Finzer (2013) released a revised Venn diagram, in which data science sits between computing and data skills, math and statistics, and disciplinary expertise. Since then, different versions of this Venn diagram have been released, further highlighting the interdisciplinary of data science.

Today, the prevailing view is that data science is an independent discipline that occupies the intersection of statistics, computer science, and application fields (Biehler et al., 2022). Data science includes computational practices for collecting, storing, extracting, and analyzing data, which are used to draw conclusions, predict, and classify knowledge structures from different sources. It can be integrated into existing content areas, such as biology, but does not have to be related to any scientific discipline (Lee & Delaney, 2022). At its core, data science is multifaceted, and its pedagogy must include a variety of topics, concerns, and practicalities associated with dealing with the data (Donoghue et al., 2021). For this study, we are embracing the prevailing multifaceted view of data science as an independent discipline, as suggested by Biehler et al. (2022) and Donoghue et al. (2021).

2.2. High School Data Science

The need to foster data literacy is expanding as data science becomes more significant in academia, industry, and society (Gould, 2021). Numerous higher-education institutions have started to offer specialized bachelor's and master's degrees in data science in response to market and research

needs (Berman et al., 2018; Irizarry, 2020). Following these trends, educational initiatives began to emerge to promote data science in primary and secondary education (Rosenberg et al., 2020). An example is the International Data Science in Schools Project (IDSSP). This cross-disciplinary and multi-national project offers training programs and frameworks for designing introductory high school data science courses (IDSSP Curriculum Team, 2019). In addition, academic standards, such as GAISE II (Bargagliotti et al., 2020), and operative frameworks for performing data science in practice have been suggested. However, the discipline is still in its infancy as different stakeholders formulate what data science is and what the curriculum must include (Schanzer et al., 2022).

Data science curricula at the high school level seek to incorporate core technical skills in statistics and mathematics while allowing students to investigate phenomena and pursue questions arising from datasets in domains ranging from health and politics to ecology and sports. In addition, the curricula may also include computing elements as students work to make data-driven insights accessible by encoding data processing and analysis systems in the form of programs. Finally, and critically, high school data science curricula can promote civic responsibility to help students understand their role as producers and consumers of data and the dangers of using data without considering its social effects (Schanzer et al., 2022). Incorporating ethical, social justice, and equity considerations into introductory data science courses is crucial in addressing the historical inequities associated with data and algorithms (Benjamin, 2019; O'Neil, 2016). To achieve this, it is essential to include in the curricula datasets that will cover a range of topics and content areas and different types of data. Additionally, students should be given the opportunity to research and synthesize these datasets (Donoghue et al., 2021). Varied curricula allow students to learn data science concepts and apply them to real-world issues of their interest (Drozda, 2021). This indicates the need to examine the state of existing curricula and assess whether the embedded datasets help achieve the goal of placing data science in the lives of today's students.

As an initial step towards understanding the current state of high school data science, Lee and Delaney (2022) developed a framework for data science curricula and applied it to two popular data science curricula: Introduction to Data Science and Bootstrap: Data Science. The framework attends to various aspects of the curricula, including topics covered, pedagogical strategies employed, how the data relates to students, and the size of the datasets. The work presented below builds on the work by Lee and Delaney, attending specifically to the role of data within high school data science curricula and extending the analysis to two additional curricula.

2.3. Culturally Oriented Data Science

The long-term impacts data scientists have and will have on society and our lived experiences underscore the importance of data science curricula reflecting the full breadth of cultures, ideas, and identities. Further, ongoing and recurring consideration of the impact of data and data-adjacent activities as they relate to issues of equity and real-world societal impact must be a central instructional goal of K-12 data science instruction (Donoghue et al., 2021).

Various factors influence students' experiences in data science, including their history and the cultural knowledge they bring to the classroom. Designing instruction to build an authentic connection between the students' identities and social and cultural backgrounds through carefully selected datasets and activities deepens learners' affiliation with data science and promotes interest

and self-efficacy (Wilkerson & Polman, 2020). Moreover, it can motivate students who lose interest in the field due to continuous preoccupation with the technical or computational aspects of data analysis (Donoghue et al., 2021). By situating the learning activities in familiar contexts, students develop a sense of agency and ownership over the datasets provided and the field more broadly (Lee et al., 2021). In addition, connecting the narratives, educational activities, and the studied data to students' interests, views, and previous experiences can increase the likelihood of knowledge acquisition (Brooks et al., 2021; Van Wart et al., 2020). The curricula should give students access to meaningful and appealing data, photos, and text, enabling them to perform first- or second-hand investigations using scientific methods (Penuel & Bell, 2016).

Data science curricula can integrate meaningful and multifaceted datasets that capture details about our environment, society, and behavioral habits from a variety of sources, such as social networks and media sources, civic and community-based data, climate data, and other socially relevant and culturally responsive topics (Wilkerson & Polman, 2020). A cross-disciplinary framework proposed by Lee et al. (2021) emphasizes the importance of attending to the personal, cultural, and sociopolitical dimensions of data and data practices. As part of the framework, they developed the notion of proximity to measure the personal relevance of the datasets to the students. Our work adopts these lenses, specifically the proximity measure, as a key dimension for understanding the data used in high school data science.

3. Method

This section presents our approach to answering the stated research questions. We begin by discussing how we selected the four focal data science curricula, briefly describing each. We then present our analytic approach, discussing each dimension used to characterize the datasets identified across the focal curricula.

3.1. Focal High School Data Science Curricula

This paper presents the analysis of four widely used data science curricula designed for high school students. To identify the focal curricula, we reviewed the curricular resources on the DataScience4Everyone website in early 2023. DataScience4Everyone is a coalition dedicated to advancing data science education in partnership with policymakers, industry leaders, schools, and scholars. Its website compiles educational materials related to data science and hosts a catalog of the curricula developed by leading universities and organizations (Data Science for Everyone, 2022). At the time of the analysis, the coalition listed 12 curricula for grades 9-12: Bootstrap:Data Science (BS:DS), Code.org, CodeHS, CourseKata, Data8, DataCamp, Education Development Center, Key2Stats, STEMcoding, Stats Medic, Introduction to Data Science (IDS), and YouCubed. It should be noted that this list and the curricular materials linked to it have been updated since we performed the analysis. In deciding which of these curricula to include, we defined four criteria a curriculum must meet: (1) it must be focused on data science (rather than a curriculum with some elements of data science); (2) it must be high-school focused; (3) it must be a fully realized curriculum (i.e., not a collection of activities/lessons to be curated by an educator); (4) it must be school/classroom-ready (i.e., must include student assignments, teacher-focused resources, supporting materials and datasets, etc.).

Each of the 12 curricula listed above was examined by two researchers using the four criteria. After the review, the researchers discussed each curriculum to be included in the research until full agreement was reached. The review of the curricula by the exclusion criteria narrowed the list to four curricula: Bootstrap: Data Science, CodeHS, Introduction to Data Science, and YouCubed. These curricula have been distributed in recent years across the US and have been collectively taught to over 200,000 students (IDC, 2023; Schanzer et al., 2022; YouCubed, 2022).

Bootstrap:Data Science (BS:DS)

BS:DS is a curriculum aimed at grades 7-10, focused on coding and data analysis. BS:DS is designed to be implemented as a standalone course or integrated into existing courses across disciplines. The BS:DS curriculum consists of 29 lessons, including presentation slides, a student workbook, programming exercises, pre-made worksheets, a teacher forum, and the online Pyret development environment (Bootstrap, 2022). During the academic year 2022-23, the BS:DS curriculum was introduced in 49 states around the US, reaching nearly 30,000 students.

CodeHS

CodeHS is a semester-long data science curriculum introducing data collection, cleaning, transformation, analysis, and visualization skills. The curriculum includes 58 lessons consisting of video tutorials, example programs to explore, written programming exercises in Python, and offline handouts. It is accompanied by lesson plans and tools for grading and tracking student progress (CodeHS, 2022).

Introduction to Data Science (IDS)

IDS is a year-long high school curriculum focusing on practical data analysis applications to develop students' computational and statistical thinking skills. The curriculum was developed by researchers from the University of California-Los Angeles in partnership with the Los Angeles Unified School District. We analyzed the fifth version, which includes four units containing 81 lessons, lab activities, practicums, and summative projects (Gould et al., 2022). The IDC curriculum was introduced in 74 districts, reaching over 40,000 students.

YouCubed Explorations in Data Science (YouCubed)

YouCubed is a project-based data science curriculum developed at Stanford's Graduate School of Education. The curriculum includes eight units, each broken into a series of sections, introducing the main ideas in data science using tools such as Google Sheets, Python, Data Commons, and Tableau. It provides detailed lesson plans and various resources for teachers, students, and parents (YouCubed, 2022). The YouCubed curriculum was previously approved in California as an alternative to Algebra 2 (LaMar & Boaler, 2021), although the State Board of Education recently reversed this decision. It has been introduced to more than 160,000 students across the US.

3.2. Analytic Approach

Our analysis aims to characterize the data used to introduce high school students to the field of data science. In looking across four widely used curricula, we seek to understand the state of data

in data science in aggregate and identify similarities and differences across different teaching approaches. To do so, we first identified all the datasets in each curriculum. For our purposes, we defined all datasets presented either in the form of data visualization (graphs, tables, maps, etc.), raw data that the learners needed to analyze, or datasets that the learners needed to produce themselves.

Once the datasets had been identified, two researchers independently analyzed each dataset, evaluating it along eight distinct dimensions: Topic, Recency, Dataset Type, Student Choice, Size, Proximity, and Students' Interests. After independently coding the datasets, the researchers compared the coding scheme results. We ran an inter-rater reliability assessment using Cohen's Kappa (Cohen, 1960), which yielded a satisfactory coefficient of 0.8. The researchers discussed and resolved any discrepancies between their analyses. Below is a discussion of each dimension in the coding scheme used to analyze each curriculum.

Topic

This category is meant to capture what real-world topic the dataset represents. We adopted the topic list for classifying datasets from the BD:DS curriculum (Schanzer et al., 2022), as it provided a concise and useful classification of topics covered by datasets in a data science context. This list includes the following topics: Sports, Politics, Entertainment, Environment and Health, Education, and Nutrition. We broadened Entertainment to Entertainment & Media and Nutrition to Food & Nutrition. Classification by topic reveals the breadth of areas covered by the datasets. Datasets that did not fit into any of these predefined categories were labeled under the "other" category, as detailed in the Results section.

Recency

This category captures the time period the data represents. This may be a single year (e.g., top 100 songs of 2022) or a time span (e.g., top 100 songs of the 2010s). Note: This category is meant to describe when the data is from, not necessarily when it was collected. If a dataset generated in 2020 contains data on crop yields in the 1800s, its recency will be 1800-1899 rather than 2020. When datasets cover a timespan, we use the most recent date to determine the category, so the 1800-1899 dataset will use 1899 as the year to determine where it fits in the coding scheme. The coding scheme developed to capture Recency includes five categories (Table 1).

Table 1. The Recency coding scheme

Recency Level	Description
Fresh	Data just-collected or just-created (e.g., students create a survey for classmates to complete and use that as their dataset; students query a public data repository for live weather conditions)
Recent	Data from the last 3 years
The Last Decade	Data from the last 10 years (but not the most recent 3 years)
Over 10 Years Old	Data from more than 10 years ago

Not Relevant	Data that does not have a time period associated with it (e.g., top speed of land animals) or fictitious data (e.g., lemonade stand sales)

Proximity

Proximity captures how the dataset relates to the learners themselves. This measure is derived from Lee and Delaney's (2022) work, which proposed a 5-point scale ranging from 0-4, with 0 describing content-agnostic data and 4 capturing data that students collected about themselves and their peers. Levels 0 and 1 capture fictional datasets, while levels 2-4 capture data representing something from the real world. In this work, we slightly expand the categories to reflect the broader set of datasets encountered in this work. Table 2 details the five levels of the proximity coding scheme.

Table 2. The proximity categorization, adapted from Lee & Delaney (2022).

Authenticity	Level	Description
Fictional Data	0	Data have no context (i.e., lists of numbers).
	1	Data have context but it is either fictional or so generalized that it is indistinguishable from fiction (e.g., list of ages of employees with no sense of where the data is coming from).
Real Data	2	Data is about a topic that may be familiar to some but not all students. This often means datasets from very niche topics (e.g., speed of various types of birds) or from primarily adult or working professional contexts (e.g., salaries, home prices).
	3	Data is on a topic one could reasonably expect learners to be familiar with but not about the learner OR data is learner-created or -generated but not about the learner. This includes learners collecting data that is not on themselves (e.g. skin tones represented in magazines), or when learners select a subset of data that is of interest to them (e.g. choosing certain states' data to look at out of a larger dataset).
	4	Data is learner-created or -generated and is about the learners themselves.

Dataset Type

This category is meant to capture the dataset's form when presented to the learner. It has three codes. Static - Datasets that have already been analyzed and are presented in a finalized form that students can interpret but not interrogate (e.g., an infographic or chart); Provided - Datasets given to the students to be analyzed (e.g., a raw CSV file containing information about the students in a fictitious school); and Student-Generated - Datasets created by the students using a survey or other direct collection of data (e.g., rolling a die and recording the results, conducting a survey).

Student Choice

This binary category captures whether students have agency in selecting which dataset they will use. In other words, are students assigned a specific dataset to use, or do they have a choice of what dataset to use?

Size

This category represents the number of observations or entries in the dataset (i.e., the sample size or the number of rows). We classified the datasets into five sizes: very small (< 25), small (25 – 100), medium (101 – 1,000), large (1,001 – 10,000), and very large (> 10,000).

Alignment with Students' Interests

The final category relates to students' interests and is used to answer the third research question about dataset alignment with student interests. For this analytic category, we conducted a pair of participatory design sessions and a 3-day extended co-design session with 28 high school students from an urban school district in the Mid-Atlantic region of the United States to gain insight into their experiences with and perceptions of data science. Of the participants, 79% were Black or African American; the remainder were American Indian, Hispanic, and White. The sessions included discussions and activities designed to shed light on topics and questions they found interesting and datasets that aligned with them.

As part of the design activities, students were asked to create an Empathy Map for a typical school student. This activity was based on the User-Centered Design concept of a persona (Miaskiewicz & Kozar, 2011) and had participants put themselves in other students' shoes and express interests and feelings while incorporating and reflecting on their identities. Students were asked to sketch their typical student and describe, among other things, what are they interested in. Using these empathy maps, students were asked to review the interests expressed by their peers and write as many questions as possible for each topic. Students were then asked to vote on the topics/questions they found most exciting and the topics they thought were most important. We then analyzed the votes from the students, which identified the topics of greatest interest to the students to be Social Media, Sports, Video Games, Animals, Going Out, and Cooking. The rationale behind this activity was to engage the participants and to understand better what interests them and what is important to them.

Alongside the design activities, at the end of each participatory design session, all students completed a short online questionnaire that included questions about their perceptions of, and interests related to data and data science. The students were specifically asked about their preferences for topics to study within data science courses. The students rated different topics on a 5-point Likert scale. Calculating the weighted average of each topic revealed that the topics they were most interested in were Music (3.9), Video Games (3.86), TV Shows and Movies (3.86), Sports (3.43), and Art and Design (3.43). Combining results from the design activities and the questionnaire, our final list of student-expressed interests includes Social Media, Sports, Video Games, TV Shows and Movies, Music, Art and Design, Animals, Going Out, and Cooking. We use this list of interests to serve as a demonstration of if and how the existing datasets in the four curricula align with the interests of the students who participated in the co-design session. In practice, we coded each of the datasets according to the areas of interest they covered, according to the mentioned list of topics, and then compared the coding between the researchers and discussed

disputes until complete agreement. This analysis is meant to serve as a demonstration of one potential way to gain insight into the interests of a set of students and then how those interests align (or do not align) with four widely used data science curricula.

4. Results

The primary goal of this study is to examine the datasets used in leading high school data science curricula to introduce learners to the field of data science. To answer the first research question focused on what datasets are used, we analyzed the characteristics of the datasets, aggregating the variables across the four curricula to provide an overarching view of the state of introductory high school data science. To answer the second research question focused on differences across curricula, we analyzed the results of our analysis comparatively. To answer the third research question, we analyzed the topics covered by the data sets and examined whether they coincided with topics that students indicated as particularly interesting to them. The results reflect the analysis of 296 datasets across the four curricula: 80 from BS:DS, 68 from CodeHS, 92 from IDS, and 56 from YouCubed.

4.1. Topics

Our analysis reveals that the most frequent topic, comprising 62 out of 296 datasets, was Environment & Health. These datasets focus on climate and weather, animals and ecosystems, and sickness and life expectancy. The next most frequent topic is Entertainment & Media (43 datasets) concerning music and movie popularity, followed by Politics (36 datasets), which included topics such as demographics, economics, and social data. The next most common topic was Education (30 datasets) relating to student grades and college acceptance rates. Additionally, we categorize 91 datasets as "Other", including datasets created to demonstrate a concept (e.g., distribution of dice rolls), fictional datasets (e.g., lemonade stand sales), and datasets meant to demonstrate the phenomenon of spurious correlations (e.g., Arcade games compared to CS grads). Datasets chosen by students and art-related datasets (for example, Visualizing RGB Space) were also included in this category as they did not fit into other categories. Figure 1 shows the frequency of the various topics by curriculum. Looking across the four curricula, the focus varies. While in BD:DS the common topic is Education, in IDS and CodeHS, it is Environment & Health (after "Other"), and in YouCubed, it is Entertainment & Media. Interestingly, we found no overlap between the datasets across the curricula. However, we did find similar datasets. For example, we found that in three curricula (IDS, CodeHS, and BS:DS) they used data on NFL players. We also found that datasets related to movies were used in three curricula (YouCubed, IDC, and BS:DS). Similarly, we found that data related to climate and animals were used in three of the curricula. In addition, we found that in YouCubed and IDS, students were asked to collect and use data related to water consumption and that in IDS and BS:DS, students analyzed data on nutritional values.

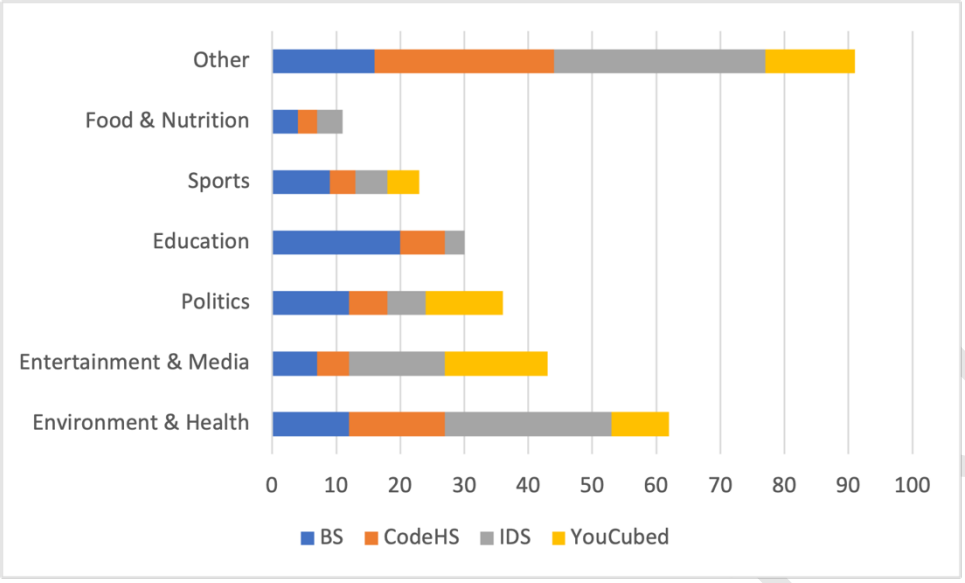

Figure 1. Frequency of the topics covered in the datasets by curriculum.

4.2. Recency

Our recency analysis shows that 41% of the datasets across the curricula (122 out of 296) are from the last decade, while only a tenth of the datasets (31 out of 296) are from more than a decade ago. In addition, about half of the datasets (143 out of 296) did not have any associated date, including fictitious datasets and data that was not time-dependent (i.e., dataset representing the lifespan of mammals).

When examining the recency of the datasets across the four curricula, we discover similarities and differences (Figure 2). CodeHS and YouCubed tend to be more current than the other two curricula as they have more datasets coded as fresh, meaning the data was/will be collected by the students, or recent, meaning the data is from the last three years. Additionally, IDS mainly has two distinct recency types – Fresh or over 10 years old data. For BS:DS, we see an opposite trend in recency types, where most datasets are either recent or from the past decade. Moreover, BS:DS has very little fresh data, meaning the curriculum rarely asks students to generate data. One thing to remember when thinking about recency is how datasets may age. By this, we mean that a dataset that is currently coded as Recent will eventually move to the Past decade, whereas a Fresh dataset will always be Fresh as it was created during the course.

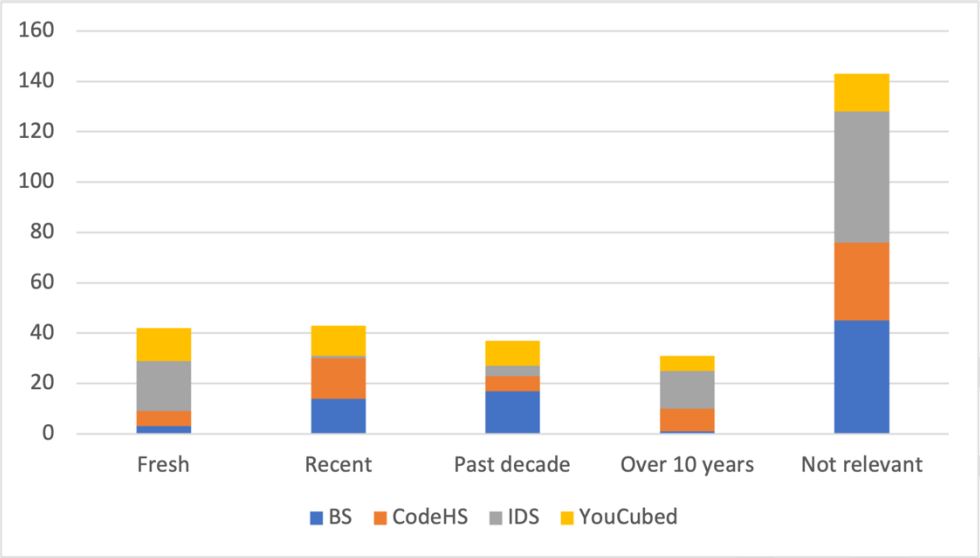

Figure 2. Frequency of datasets by recency

4.3. Proximity

Looking at Proximity, that is, how the data relates to the students, we observe a bell-shaped distribution of proximity levels across each of the four curricula. Most datasets (90 out of 296) are rated as Level 2, representing real data that is not closely related to the students' lives (e.g., economic data). Level 3 is the next most common level, representing 87 datasets that include topics more relevant to students (e.g., data from youth-oriented pop culture). This is followed by Level 1, which represents 68 of the total datasets and includes fictional data with context (e.g., lemonade stand sales figures). In the tails of the distribution, we observed fictional data without context (Level 0) and data closely related to the students themselves (Level 4).

When comparing the curricula in terms of proximity (Figure 3), our findings reveal that BS:DS had the highest proximity scores based on the datasets included in the curricula, with almost half of the datasets rated as Level 3 or Level 4. In contrast, IDS and CodeHS have the highest number of fictional data (i.e., Levels 0 or 1), while YouCubed has the least fictional data. Considering how datasets relate to students is critical, especially if one of the goals of the curriculum is to help learners understand the role data and data science play in their daily lives.

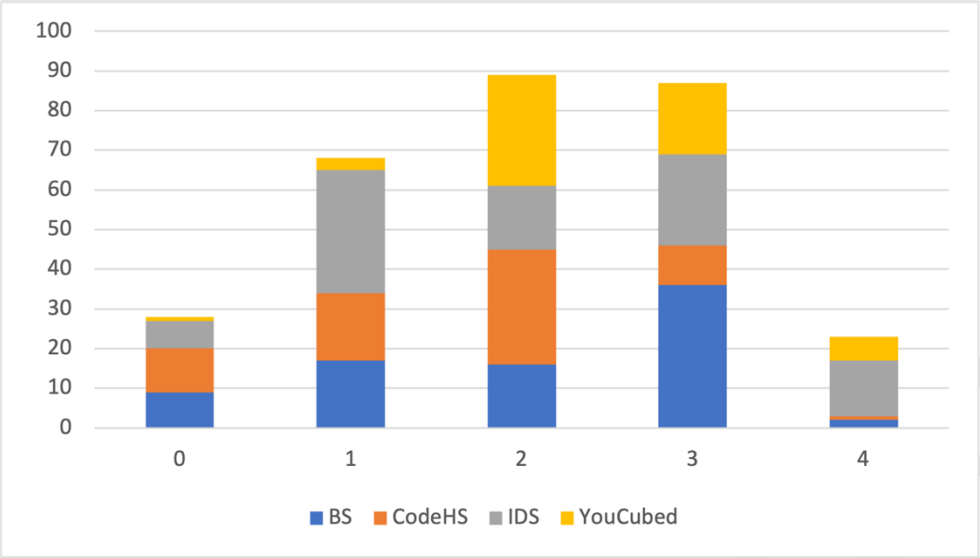

Figure 3. Frequency of datasets by proximity level

4.4. Dataset Types

Our analysis of the types of datasets learners use and how they interact with them reveals that almost half of the total datasets analyzed (140 out of 296) are provided, meaning that the data were collected in advance and given to the students for analysis. 64 of these 140 datasets included raw data in CSV or Google Sheet formats, such as raw data from the Centers for Disease Control or the American Community Survey. Six additional datasets were based on data from sites such as Wikipedia, Yelp, and ESPN. Other provided datasets included data pre-entered into RStudio, Pyret, or Colab. It could not be determined if these datasets had undergone any processing. Our analysis revealed many static datasets (110 out of 296), including graphs and infographics. These datasets included only processed data. The remaining datasets, 46 out of 296, were Learner-Generated datasets that included raw data the students collected by themselves.

We see some differences emerge when we look at the dataset types used in each curriculum (Figure 4). CodeHS has the highest number of provided datasets in its curriculum (50 out of 68), followed by BS:DS, which has 48 out of 80. In contrast, they have relatively few Learner-Generated datasets (3 datasets in CodeHS and 7 datasets in BS:DS). YouCubed, on the other hand, has the fewest provided datasets (9 out of 56) but the highest percentage of static datasets (57%, 32 out of 56) compared to the other curricula. A relatively high percentage of static datasets is also present in IDS (41%, 38 out of 92). In these curricula, students spend more time looking at data analyzed by others rather than analyzing it themselves. The low number of static datasets in CodeHS suggests a de-emphasis on students' learning from analysis performed by others. It is worth noting that there is value in including both interactive and static datasets as part of a data science curriculum, given the way students will encounter data outside of the classroom.

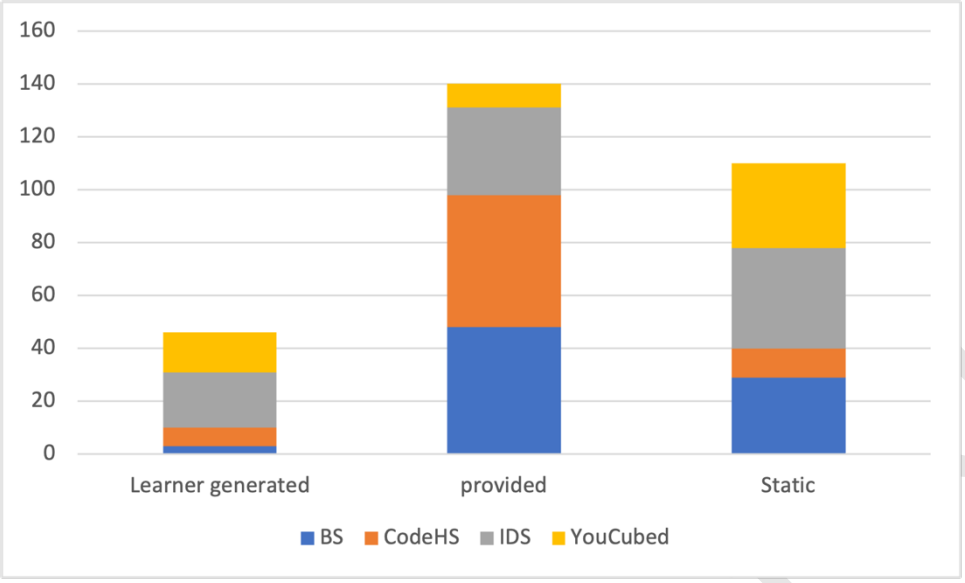

Figure 4. Frequencies of dataset types per curricula

4.5. Students' Choice

Our analysis of Student Choice indicates that in all curricula we analyzed, students rarely have agency in deciding what datasets to analyze. Across the four curricula, 235 out of 296 datasets were prescribed by the curriculum, meaning students have no say in the datasets they analyze. It should be noted that BS:DS has the largest share of datasets for students to choose from (44%, 35 out of 80 datasets). These datasets cover a wide range of topics, but most relate to Politics (10 out of 35) or Environment & Health (7 out of 35). There are a few datasets on Entertainment & Media, Sports, Education, and Food & Nutrition, exemplifying that the data selection can be completely open to the students or depend on a certain context/topic. Considering the goal of helping students see the importance and relevance of data science in their lives, revisiting how and when students have agency in selecting datasets may be an important step.

4.6. Alignment with Student's Interests

Our analysis reveals that 35% of the datasets (102 out of 296) align with one of the nine topics that students expressed interest in (Figure 5). Sports and Music were the only topics that appeared in all four curricula. Sports was the most frequent topic that appeared in 23 datasets. Music and Animals were the next most frequent topics (each appeared in 14 datasets), with most music-related datasets found in the YouCubed curriculum and most animal-related datasets found in CodeHS. Cooking was the next most common topic (12 datasets), followed by TV Shows & Movies, which appeared in a total of 10 datasets and mainly in the IDS curriculum. Other topics of interest to students that the datasets covered were Art & Design (8 datasets), Social Media (8 datasets), Going Out (7 datasets), and Video Games (6 datasets). Additionally, we found that 11 out of 296 datasets dealt with topics chosen by the students. These datasets were based on data collected by the students or datasets they found by themselves around the internet. These datasets were found in all the curricula reviewed except for BS:DS.

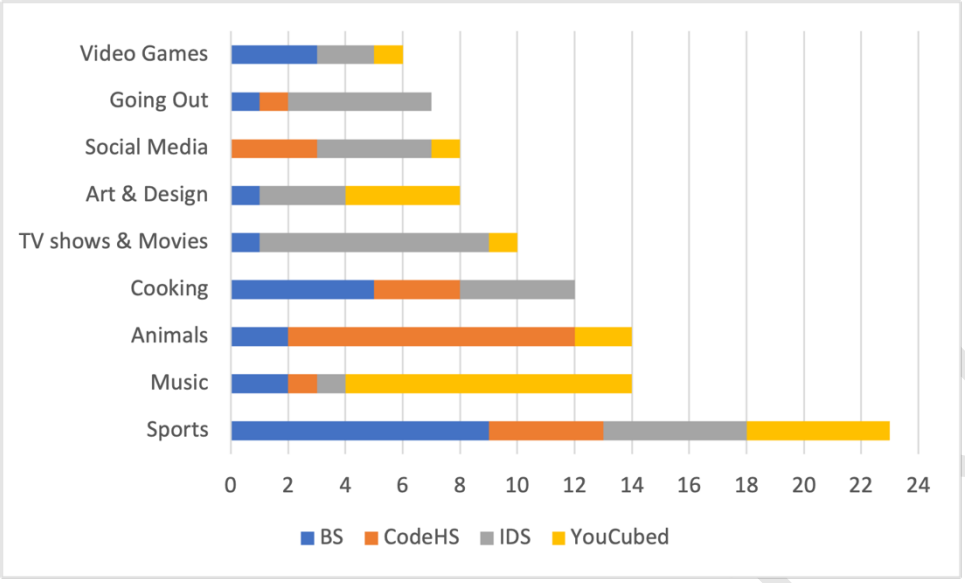

Figure 5. Frequency of datasets by topics of interest to students

Surprisingly, we found that over 60% (183 out of 296) of the datasets were from topics that did not align with students' self-identified interests. The most common topics for data sets not aligned with students' interests were related to Environment & Health (45 datasets), Education (30 datasets), and Politics (29 datasets). It should be noted that some of these topics were partially represented in topics of interest to the students. For example, the Environment & Health category included datasets that dealt with animals that students perceived as interesting but also included datasets that dealt with health, disease, climate, and water consumption that did not emerge in our work to draw out students' self-reported interests. None of the 30 datasets related to Education were seen as interesting for the students. Further, most of the datasets included under the Politics category (29 out of 36) did not coincide with students' interests as they dealt with demographic and economic information.

4.7. Intersections between Analytic Dimensions Across Curricula

Having presented findings for each dimension in isolation, we now look across our analytic dimensions to identify trends and potential opportunities with how various categories and patterns interact, with a particular emphasis on the topics students identified as being of interest. In doing so, we pursue questions such as: Do topics that align with students' interests tend to be newer than data sets not aligned with their interests? And is there a difference in the size of the datasets that deal with topics of interest to students? In this section, we highlight particular interactions of note from our analyses.

In examining the topics that align with students' interests in relation to the size of the datasets across the curricula, we see that some of the topics tend to have smaller datasets than others. More concretely, the datasets dealing with Cooking, Art & Design were relatively small (two very small datasets, seven small datasets, and three medium datasets). At the same time, topics related to Politics, Education, and Environment & Health, which were overwhelmingly not found to be interesting topics for the students, were responsible for providing the largest datasets. This finding

is not surprising as these datasets may be sampled over a long period of time. We also see that datasets dealing with sports were found in all sizes, though most of them are small to medium. Additionally, datasets dealing with Social Media were mostly medium to very large, which reflects a key characteristic of social media sites, their ability to generate vast amounts of data. In looking at the intersection of topics that align with students' interests and recency, we see that the datasets related to topics of Cooking, Social Media, and Music are the most recent. While it is not surprising that Social Media and Music are more contemporary, it is surprising that Cooking was comprised of recent datasets. Among the datasets that dealt with topics that generally did not interest the students, we found many of the older datasets. Figure 6 summarizes this analysis by illustrating how many datasets we identified for each topic/size/recency tuple (the bubble size reflects the number of datasets that match that size/recency/topic combination). This figure shows the value of using recent data to create a data science curriculum that is relatable and of interest to students. Moreover, it highlights the importance of the size of the datasets, as working with different datasets can meet different pedagogical requirements.

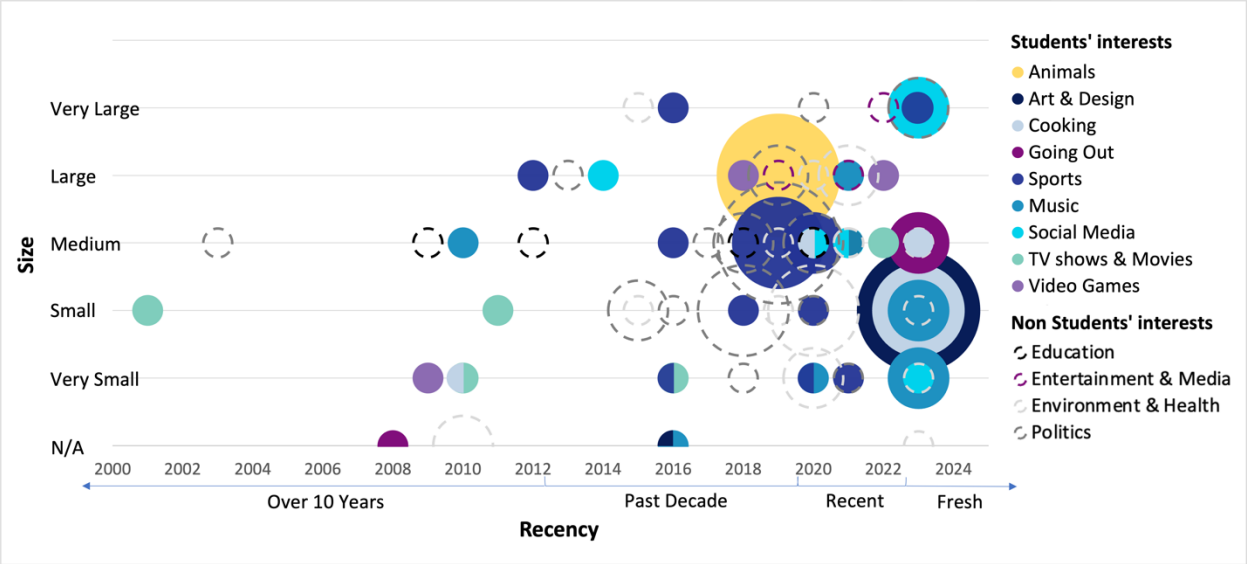

Figure 6. The size of datasets across the curricula, organized on a timeline. Colored bubbles represent topics found to be of interest to students, and bubbles with a dashed line show the topics that are not in alignment with students' interests. The bubble size represents the number of datasets with that size/recency, with four levels - the largest bubble represents four datasets, and the smallest represents one dataset. Halved bubbles represent the overlap between bubbles from different topics that have the same size.

5. Discussion

In analyzing the datasets present in four high school data science curricula, we sought to deepen our understanding of the data used to introduce students to the field of data science. Our analysis reveals the unique characteristics of each curriculum's chosen datasets alongside the similarities and shared trends. Examining the topics, recency, size, and proximity of the datasets teaches us about the data itself, while the type of the datasets and whether students have agency in choosing the datasets sheds light on pedagogical aspects of how the data is consumed. Examining the

datasets from the perspective of the students, i.e., focusing on the students' topics of interest as they came up in the co-design and participatory design sessions, shows whether and how the datasets draw on student interests or align with their lived experiences.

It is important to acknowledge that students' areas of interest may not be uniformly shared. Consequently, a dataset selection tailored to the preferences of a specific focus group, as illustrated in the case study, may not be universally representative. Therefore, curriculum designers should think of ways to allow students to choose the datasets themselves or alternatively to empower teachers to make adaptations in the data selection.

The study results show how the curricula integrate relatively recent datasets that cover a range of topics. In addition, they show that the curricula are limited in the involvement of the students in the selection of the datasets used. Moreover, they indicate significant gaps between the topics covered by the datasets and the interests of the high school student participants and emphasize the nuances between and within these topics. The findings show that some topics are more captivating than others, but it's also crucial to be responsive to the interests of the students in the classroom. It is possible that in some classrooms, students may have been more interested in the topics covered in these datasets. However, in this classroom and potentially many others, increased ability for students to choose datasets personally interesting to them could support better alignment.

Two important caveats must be considered in the selection process of the datasets. The initial concern involves students' limited exposure to datasets beyond the immediate scope of their community and environment. This restriction raises the issue of potentially hindering their exploration of new topics or areas of interest that they may not have encountered previously. Despite these considerations, educators and curriculum designers should prioritize the overarching goals of learning and the practical skills students need to acquire.

The emphasis on fostering students' interest in learning is a central objective in education due to its significant impact on motivation, curiosity, engagement, and, ultimately, academic success (Hidi & Harackiewicz, 2000; Rotgans & Schmidt, 2017). The interest development theory emphasizes that interests thrive when individuals can explore, interact with, and derive meaning from topics that captivate them (Renninger & Hidi, 2015). Aligning the narratives of educational activities and the studied datasets with students' interests, perspectives, and prior experiences can intensify engagement and heighten the probability of knowledge acquisition (Brooks et al., 2021). Despite the potential drawbacks of limited dataset exposure, the emphasis remains on nurturing a dynamic and meaningful learning experience that aligns with students' educational and personal needs.

Our analysis also reveals opportunities to improve the curricula by, for example, re-examining datasets' themes and replacing dated ones with ones that are more recent, proximate, and relevant to the students' daily lives. Given the recent nature of the datasets, it is plausible that older data pertaining to student interests may remain more relevant and preferable than more recent data that does not align with the student's areas of interest. For instance, a student passionate about art may lean towards working with data from the 17th century rather than more contemporary data on galaxies. However, the converse scenario is also conceivable, where a student may prioritize learning about a current topic outside their usual interests over engaging with outdated data. Therefore, it is crucial to consider and find a balance between the age of the subjects, their specific

interests, and the data to which they are exposed. These efforts can increase students' motivation and interest in the material studied and in data science in general (Donoghue et al., 2021).

We acknowledge that updating datasets is a difficult and time-consuming process to implement, especially in curricula where updating a dataset may have cascading effects on task prompts and other activities used within the curriculum. There are several ways in which this challenge might be alleviated in a way that supports student interests: first, curricula could lean more on live-updated API data sources; second, curricula could include flexible activities that point students towards a choice of several different datasets for analysis; third, some activities could center student-generated datasets. These suggestions would improve students' ability to choose data that is interesting to them and/or increase the recency of included datasets without dramatically increasing the workload on the curriculum designers' part after the initial change.

Moreover, the various indicators presented in our study can help researchers, curriculum developers, and educators examine the curricula and the datasets integrated into them and adapt them to better meet learning needs. For example, different datasets achieve different educational goals. If the goal is to train students to think like data scientists, then working with authentic datasets (e.g., large, messy) and performing the analyses is extremely important. To this end, it is essential to integrate more "provided" or "learner-generated" datasets and put an emphasis on practicing data analysis. Alternatively, if the goal is to help students become data-literate citizens and develop a basic understanding of data science principles and practices, then it may make more sense to spend time looking at analyses done by others and conducting analyses with more general-purpose data analysis tools like spreadsheets. While it was not the focus of the current study, it is crucial to match the datasets with the learning objectives to allow for the necessary concepts and practices to be taught effectively. In other words, it is critical not to sacrifice rigor or learning opportunities for the sake of relevancy or interest alignment. Our view of finding interesting and engaging datasets is not necessarily in tension with datasets that allow learners to effectively learn essential data science concepts and practices. We believe datasets spanning a range of topics can contribute significantly to achieving these educational goals.

The analysis presented above can also be valuable for educators using these curricula to identify opportunities to augment existing datasets, bring in different datasets that align with the interests of students, or offer opportunities for connecting with specific communities. If a teacher recognizes that a curriculum has relatively few datasets on a given topic or few opportunities for learners to select their own datasets, that teacher could make modifications for greater inclusivity. Alternatively, if there are unique opportunities within a specific school or community, say, datasets that highlight specific local ecologies or community events, incorporating those datasets may provide an opportunity to better situate data science as personally relevant to learners. Teachers and curriculum designers have at their disposal a variety of up-to-date and free-to-use databases, starting with K-12 data science tools that have built-in datasets (Israel-Fishelson et al., 2023) and repositories of government agencies (such as Data.gov) or Open Data portals by cities and states. As data science education is a relatively new initiative, research on its teaching remains in its infancy, and more research into how teachers may make decisions about dataset alignment and selecting new datasets to include and use is certainly warranted.

Similarly, the selection of the size of the datasets, the programming language, and the tool/environment for processing and analyzing the data are essential for designing the learning

experiences and achieving the educational goals. Our findings showed that the curricula mainly use small datasets and limited tools and programming languages. These can undoubtedly be beneficial in introducing the field of data science and imparting its principles to students. However, working with large datasets and various languages and analysis tools is important for preparing students for the jobs of data scientists whose work is often based on big data and advanced analytics (Coelho da Silveira et al., 2020; Zhang et al., 2017). Large datasets demonstrate how programming techniques, an essential part of data science, are necessary; manual techniques that might be workable for a small dataset are not practical in the work of the field. Including datasets of this size will help students understand more of the working principles of the field and why the emphasis on programming techniques is so important. It is important that students get this message early in learning about data science so that it becomes a fundamental understanding as they continue to learn about the field. Following that notion, due to the complexity required to analyze large datasets, it is necessary to ensure appropriate teacher training that will enable them to acquire and master the technical and practical skills needed for effective instruction (Lee & Delaney, 2022). It is important to ensure that schools possess the required tools and infrastructure to handle large-scale datasets.

Another important aspect of this analysis is thinking about ways these introductory data science curricula and the datasets they are using are welcoming, supporting, or valuing the perspectives of learners from historically excluded populations in computing. Do the datasets being included and analyzed in the curriculum reflect the interests, perspectives, or values of learners from historically excluded groups? This is one area where allowing for learner agency and having learners collect their own data may be useful, as it will provide opportunities for students to analyze data that has some personal or cultural relevance. Another important, related consideration is whether the curriculum and datasets provide opportunities to discuss topics like algorithmic bias or investigate ways that certain populations are being negatively impacted by data-driven algorithms. Given that a high school data science course may be a person's only opportunity to learn about the impact of data on their lives in a formal context, it is essential that the curricula attend to issues of equity, access, and social justice and that the datasets used to reflect the interests, ideas, and values of all students in the classroom.

6. Limitations

While the results and insights of this study contribute to a deeper understanding of how established curricula use datasets and what their characteristics are, we also want to highlight some limitations. The first is related to locating and extracting the data. We relied on the DataScience4Everyone database to locate established curricula but may have missed other programs that meet the inclusion criteria and are published elsewhere. We believe it is important to expand the canvas and that this study can be used to examine and evaluate additional curricula.

A second limitation is not attending to the depth of use of different datasets. For example, some datasets are the focus of weeks-long projects, while others are shown on a slide for only a few minutes. In our coding, we do not distinguish between these two datasets when considering the composition of datasets in the course. A deeper analysis could consider how long students will spend with a given dataset as a means of providing some form of weighting to the datasets.

A third limitation is related to the analysis of the datasets based on the interests of the research population, i.e., the 28 students who participated in the co-design session. Tailoring the analysis to a particular group's preferences may overlook key factors that could significantly affect a more diverse demographic. As a result, our findings may not accurately reflect the complexities present in the broader societal context, leading to biased conclusions and limited generalizability. We, therefore, intend to continue examining the preferences of additional populations and ways to integrate more diverse datasets (for example, using APIs).

A final limitation is recognizing that this analysis reflects the state of data use in high school curricula at a specific moment in time. Over time, new curricula will emerge, and new datasets and revisions will be introduced to the curricula analyzed. This happened throughout this project, where YouCubed released updated activities and datasets after completing the analysis. As such, the results presented in the paper represent a historical snapshot rather than a permanently up-to-date reflection of data use in high school data science curricula. As an aside, this is the exact challenge data science curricula designers and educators face in trying to keep their curricula up-to-date and relevant. While we see this as a limitation, we still think this work is useful as a means of taking stock of where we are early in the development of high school data science.

7. Conclusion

Given the growing prevalence and significance of data in society, all students must develop a basic understanding of what data science is and how it impacts their lives. A key component of data science instruction is the datasets that are included. To help understand the current state of data science and how learners are being introduced to it, this paper provided a systematic analysis of the 296 datasets used across four of the most widely used high school curricula. By analyzing each dataset along eight distinct dimensions, we can understand the breadth of topics, structures, and contexts in which data is being situated. This is important as developing curricula with relevant and engaging datasets is crucial to creating effective, engaging, and equitable data science educational experiences for all students. This work serves to help us understand a critical dimension of the nascent field of high school data science. Understanding what data is and is not, being included in high school learning experiences, can help us understand where we are in a field, what ideas, values, and practices are being prioritized, and serve to inform the next wave of tools, curricula, and innovations in data science education.

Disclosure Statement

This work is supported by the National Science Foundation (Award # 2141655)*. Any opinions, conclusions, and/or recommendations are those of the investigators 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).* American Statistical Association.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code* (1 edition). Polity.

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., & Stodden, V. (2018). Realizing the potential of data science. *Communications of the ACM*, *61*(4), Article 4.

Biehler, R., Veaux, R. D., Engel, J., Kazak, S., & Frischemeier, D. (2022). Research on data science education. *Statistics Education Research Journal*, *21*(2), Article 2. https://doi.org/10.52041/serj.v21i2.606

Bootstrap. (2022). *Bootstrap:Data Science*. https://bootstrapworld.org/materials/data-science/

Brooks, C., Quintana, R. M., Choi, H., Quintana, C., NeCamp, T., & Gardner, J. (2021). Towards culturally relevant personalization at scale: Experiments with data science learners. *International Journal of Artificial Intelligence in Education*, *31*(3), 516–537. https://doi.org/10.1007/s40593-021-00262-2

Calabrese Barton, A., Kang, H., Tan, E., O'Neill, T. B., Bautista-Guerra, J., & Brecklin, C. (2013). Crafting a future in science: Tracing middle school girls' identity work over time and space. *American Educational Research Journal*, *50*(1), 37–75. https://doi.org/10.3102/0002831212458142

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x

CODAP. (2022). *CODAP – Common Online Data Analysis Platform*. https://codap.concord.org/

CodeHS. (2022). *CodeHS – Data Science Course*. https://codehs.com/course/data_science/overview

Coelho da Silveira, C., Marcolin, C. B., Da Silva, M., & Domingos, J. C. (2020). What is a data scientist? Analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias*, *8*(1), 25–39. https://doi.org/10.5585/iptec.v8i1.17263

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Conway, D. (2010). *The data science Venn diagram*. Drew Conway. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science for Everyone. (2022). *Teaching data science*. K12data. https://www.datascience4everyone.org/teach-data-science

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching creative and practical data science at scale. *Journal of Statistics and Data Science Education, 29* (sup1), S27–S39. https://doi.org/10.1080/10691898.2020.1860725

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Drozda, Z. (2021). *Catalyzing a new field: Data science education in k-12.* National Center for Education Research. https://ies.ed.gov/ncer/whatsnew/techworkinggroup/pdf/DataScienceTWG.pdf

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2), Article 2. https://doi.org/10.5070/T572013891

Gould, R. (2021). Toward data-scientific thinking. *Teaching Statistics, 43*(S1). https://doi.org/10.1111/test.12267

Gould, R., Machado, S., Johnson, T. A., & Molyneux, J. (2022). *Introduction to data science curriculum.*

Hazzan, O., Ragonis, N., & Lapidot, T. (2020). Data science and computer science education. In O. Hazzan, N. Ragonis, & T. Lapidot (Eds.), *Guide to Teaching Computer Science: An Activity-Based Approach* (pp. 95–117). Springer International Publishing. https://doi.org/10.1007/978-3-030-39360-1_6

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, 70(2), 151–179. https://doi.org/10.3102/00346543070002151

IDC. (2023). *Introduction to Data Science.* https://www.idsucla.org/

IDSSP Curriculum Team. (2019). *Curriculum frameworks for introductory data science.* http://idssp.org/files/IDSSP_Frameworks_1.0.pdf

Israel-Fishelson, R., Moon, P. F., Tabak, R.& Weintrop, D. (2023). Preparing K-12 students to meet their data: Analyzing the tools and environments used in introductory data science contexts. *Proceedings of the 2023 Symposium on Learning, Design and Technology.* Evanston, Illinois, 29-42.

Irizarry, R. A. (2020). The role of academia in data science education. *Harvard Data Science Review.* https://doi.org/10.1162/99608f92.dd363929

Ladson-Billings, G. (2014). Culturally relevant pedagogy 2.0: A.K.A. the remix. *Harvard Educational Review, 84*(1), 74–84. https://doi.org/10.17763/haer.84.1.p2rj131485484751

LaMar, T., & Boaler, J. (2021). The importance and emergence of K-12 data science. *Phi Delta Kappan*, *103*(1), 49–53. https://doi.org/10.1177/00317217211043627

Lee, V. R., & Delaney, V. (2022). Identifying the content, lesson structure, and data use within pre-collegiate data science curricula. *Journal of Science Education and Technology*, *31*(1), 81–98. https://doi.org/10.1007/s10956-021-09932-1

Lee, V. R., Wilkerson, M. H., & Lanouette, K. (2021). A call for a humanistic stance toward K–12 data science education. *Educational Researcher*, *50*(9), 664–672. https://doi.org/10.3102/0013189X211048810

Miaskiewicz, T., & Kozar, K. A. (2011). Personas and user-centered design: How can personas benefit product design processes? *Design Studies*, *32*(5), 417–430. https://doi.org/10.1016/j.destud.2011.03.003

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Penuel, W. R., & Bell, P. (2016). *Qualities of a good anchor phenomenon for a coherent sequence of science lessons*. Researchandpractice.Org. http://researchandpractice.org/wp-content/uploads/2016/03/Anchor_Design_Problems_March2016.pdf

Renninger, K. A., & Hidi, S. (2015). *The power of interest for motivation and engagement*. Routledge.

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review*, *84*(3), 528–549. https://doi.org/10.1111/insr.12110

Rosenberg, J. M., Lawson, M., Anderson, D. J., Jones, R. S., & Rutherford, T. (2020). Making data science count in and for education. In *Research Methods in Learning Design and Technology* (pp. 94–110). Routledge.

Rotgans, J. I., & Schmidt, H. G. (2017). The role of interest in learning: Knowledge acquisition at the intersection of situational and individual interest. In P. A. O'Keefe & J. M. Harackiewicz (Eds.), *The Science of Interest* (pp. 69–93). Springer International Publishing. https://doi.org/10.1007/978-3-319-55509-6_4

Schanzer, E., Pfenning, N., Denny, F., Dooman, S., Politz, J. G., Lerner, B. S., Fisler, K., & Krishnamurthi, S. (2022). Integrated data science for secondary schools: Design and assessment of a curriculum. *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education*, 22–28. https://doi.org/10.1145/3478431.3499311

Stornaiuolo, A. (2020). Authoring data stories in a media makerspace: Adolescents developing critical data literacies. *Journal of the Learning Sciences*, *29*(1), 81–103. https://doi.org/10.1080/10508406.2019.1689365

Van Wart, S., Lanouette, K., & Parikh, T. S. (2020). Scripts and counterscripts in community-based data science: Participatory digital mapping and the pursuit of a third space. *Journal of the Learning Sciences*, *29*(1), Article 1. https://doi.org/10.1080/10508406.2019.1693378

Weiland, T., & Engledowl, C. (2022). Transforming curriculum and building capacity in K–12 data science education. *Harvard Data Science Review*, *4*(4). https://doi.org/10.1162/99608f92.7fea779a

Wilkerson, M. H., & Polman, J. L. (2020). Situating data science: Exploring how relationships to data shape learning. *Journal of the Learning Sciences*, *29*(1), 1–10. https://doi.org/10.1080/10508406.2019.1705664

YouCubed. (2022). *Explorations in data science*. Youcubed High School Data Science Course. https://hsdatascience.youcubed.org/

Zhang, Y., Zhang, T., Jia, Y., Sun, J., Xu, F., & Xu, W. (2017). DataLab: Introducing software engineering thinking into data science education at scale. *2017 IEEE/ACM 39th International Conference on Software Engineering: Software Engineering Education and Training Track (ICSE-SEET)*, 47–56. https://doi.org/10.1109/ICSE-SEET.2017.7