Exploring variability during data preparation: A way to connect data, chance, and context when working with complex public datasets

Note: This manuscript has been accepted for publication in *Mathematical Thinking & Learning* as of May 5, 2020

Abstract.

Data preparation (also called “wrangling” or “cleaning”)—the evaluation and manipulation of data prior to formal analysis—is often dismissed as a precursor to meaningful engagement with a dataset. Here, we re-envision data preparation in light of calls to prepare students for a data-rich world. Traditionally, curricular statistics explorations involve data that are derived from observations that students record themselves or that reflect familiar, relatively closed systems. In contrast, pre-constructed public datasets are much larger in scope and involve temporal, geographic, and other dimensions that complicate inference and blur boundaries between “signal” and “noise.” As a result, students have fewer opportunities to consider sources of variability in such datasets. Due to these constraints, we argue that data preparation becomes an important site for students to reason about variability with public data. Through analyses of repeated task-based interviews with five pairs of adolescent participants, we find that specific actions during data preparation, such as filtering data or calculating new measures, presented opportunities to engage leaners with variability as they prepared and analyzed several public socioscientific datasets. More broadly, our study highlights some changes to theory and curriculum in statistics education that are necessitated by a focus on “big data literacy”.

Keywords. data moves; data literacy; data analysis; variability; secondary school

Exploring variability during data preparation: A way to connect data, chance, and context when working with complex public datasets

The emerging field of data science is changing the nature of K-12 statistics education. Educators are increasingly expected to support learners in developing wide-ranging data literacies, including an understanding of how data may be used and misused; how politics and technology shape the ways data are constructed and analyzed; and how data often reflect, both implicitly and explicitly, the histories, lives, environmental contexts, and interests of students themselves (Engel, 2017, Philip, Schuler-Brown, & Way, 2013). These goals require more emphasis on students learning to work with *public* datasets that are often large in scale and complex in nature (Gould, 2017; Ridgway, 2015). Working with such datasets requires not only that students learn new technical competencies to organize and visualize data, but also that they learn how to judge the utility and validity of datasets constructed by others for their own investigative purposes.

A statistical modeling perspective behooves us to consider student investigation of pre-existing datasets as not passively exploratory, but actively goal-driven and constructive. In this view, public datasets are not static information sources. Instead, they are dynamic resources that a student may explicitly seek out, evaluate, transform, or otherwise repurpose (Wilkerson & Laina, 2018) to address specific needs. Within one dataset, what constitutes a meaningful pattern for one person may simply count as noise for another depending on investigative goals. Given this perspective, we argue *data preparation*—the process of evaluating and manipulating data in preparation for formal analysis—is an important way for students to consider variability and error in public datasets.

Our interest in data preparation complements current efforts in statistics education to engage learners with chance and variability. As we describe in more detail below, most such efforts focus on having learners construct their own data from observations, or model public data with simulations (e.g., Aridor & Ben-Zvi, 2018; Biehler, Frischemeier, & Podworny, 2018). Our focus on the preparation of large, public datasets adds a much-needed opportunity for learners to explore variability as it emerges across time, space, and in interaction across multiple variables in a dataset. Attending to data preparation highlights how even simple manipulations to datasets may expose, amplify, or hide variation in a dataset—with important implications for analysis and inference. We argue that such considerations are critical at a time when students are expected to become thoughtful and reflective consumers of public data and the inferences drawn from them.

This paper reports an initial attempt to explore data preparation as a site for student reasoning about variability in public datasets. It is part of a larger design-based research project to create tools, activities, and structures for engaging youth in manipulating and analyzing large, public datasets using the Common Online Data Analysis Platform (CODAP; Concord Consortium, 2014). As part of this work, we conducted semi structured, repeated task-based interviews with five pairs of youth as they explored researchers' and their own questions using a number of public datasets. Our analysis focuses on the question: What opportunities were there for participating youths to engage with variability when preparing public datasets for analysis?

Background Literature

In the educational research literature, statistical modeling is described as a response to an artificial separation in statistics education: Curricula often emphasized *relational* or *summary* patterns separately from the *probabilistic* patterns that emerge in observational data (Biehler, 1994). Using the language of Konold and Pollatsek (2002), such curricula highlighted either “signal” or “noise” in data without deep connections between the two. To address this separation, statistical modeling approaches engage learners in constructing data and then using statistics and simulations to model the patterns found within those data (Konold & Kazak, 2008). Statistical modeling highlights how expected patterns in data (“signal”) connect to context, because students observe or simulate the context first-hand. It also emphasizes how error and natural variation introduce variability (“noise”) into data, as students observe variability within the data context and work to reproduce it in their models (Pfannkuch, Ben-Zvi, & Budgett, 2018).

Our work is concerned with how learners use *existing* datasets constructed and made publicly available by others, often with no particular investigative purpose in mind (Casell & Topi, 2010). Learners are not involved in the construction of these datasets and thus have less access to the data context; at the same time such datasets are often about topics or events students are aware of or have been a part of (Wilkerson & Polman, 2020). This changes the types of opportunities students have make sense of patterns (“signal”) and variability (“noise”) within these datasets. In the following sections, we explore how the statistical modeling literature lends insight into learners’ opportunities to engage with variability in these larger, public datasets.

Possibilities and Challenges for Exploring Variability in Public Datasets

Early research on Exploratory Data Analysis (“EDA”) in educational settings, like EDA in the professional sphere (e.g., Tukey 1977), engaged learners in finding patterns in multivariate datasets using emerging visualization and analysis technologies. These approaches emphasize establishing familiarity with data before more formal analysis, for example by using visualizations such as scatterplots and boxplots to highlight variability in the “shape” of data (Cooper & Shore, 2010). Some statistics educators voiced concern, however, that students would struggle to draw proper inferences or conclusions from exploratory data analysis without a stronger focus on probability (Biehler, 1994; Konold & Kazak, 2008). These tensions, in fact, motivated much of the research on statistical modeling approaches, and remain an ongoing topic of conversation in the field (e.g., Biehler, Frischemeier & Podworny, 2017; Pfannkuch, Ben-Zvi, & Budgett, 2018).

Given recent interest in larger and more complex datasets, exploratory approaches are again emerging within the literature, along with the same concerns. In a recent paper examining high school students’ modeling “big data,” Gil and Gibbs (2017) reported students engaged in modeling *covariation* within data. However, there was not much evidence for students’ explicit engagement with *variability* in the data. When variability was mentioned, it was described with respect to “scatter” or other visual features of graphs, without interpretation of what the visual features mean for data or inference.

The statistical modeling literature offers tips for addressing this (re)emerging need to engage students with variability during exploratory work. Pfannkuch, Ben-Zvi, and Budgett (2018), highlighting work by Lehrer and colleagues (Lehrer, Kim, & Schauble, 2007; Lehrer, 2017), describe a statistical modeling approach that is grounded in exploratory data analysis but “pay[s] attention to *causes* for the way data are distributed” (p. 1114; emphasis added). Others (Biehler, 1994; Konold & Kazak, 2008; Konold & Pollatsek, 2002; Makar & Rubin, 2009) make similar calls to attend not only to variation in data, but also to the contexts, events, and processes that introduce variability to data.

For reasons we described above, learners are less likely to know much about the contexts and processes that generate *public* datasets. Additionally, many public datasets feature sparse observations made across temporal and/or geospatial dimensions. These characteristics make public datasets quite different from the repeated observations within controlled environments that are common in datasets used for statistical modeling activities. They complicate whether variation observed in a public dataset is due to error, natural variation, or related potential causes such as time, geographic distribution, and other factors. More generally, the multidimensional nature of such datasets means that any variation that has been recorded in observations is likely to reflect *multiple* sources of variability. Furthermore, since such datasets are often repurposed by investigators and educators to explore new questions, one person's "noise" within this variability may be another person's "signal," depending on one's investigative goals.

Modeling Variability in Public Datasets

How, then, might students reason about variability in the context of existing public datasets? In the Civic Statistics project, pre-service teachers create statistical models of publicly accessible civic datasets (Biehler, Frischemeier, & Podworny, 2018). Biehler and colleagues describe one activity in which teachers use Tinkerplots' sampler tool to simulate hospital admissions data. In creating and revising these simulations to better fit the actual available statistics, the participating teachers began to think deeply about why people of different ages and genders might be admitted to a hospital. For example, they began to investigate their own assumptions for how likely people of different ages and genders may be admitted to the hospital for an emergency, versus a planned hospitalization, or childbirth. Creating the simulation directly engaged learners in reasoning about *both* causes for variation observed in the pubic dataset, as well as how those causes may result in different distributions of hospitalized people.

This approach is powerful and establishes that public data can serve as a site for engaging deeply with variability. However, it relies on the use of a specific software tool, Tinkerplots. And, the type of inference it engages is also relatively guided and proximal—students are asked to reason about general hospitals using a specific dataset. We seek to extend statistical modeling approaches to consider how learners might explore variability in geographically and temporally rich public data, using functionalities that are available in a wide variety of professional and pedagogical data technologies.

Data Preparation **as a Site for Engaging with Variability in Public Datasets**

Given the challenges and opportunities presented by public datasets, we contend that *data preparation* warrants more attention as a component of data analysis—especially as a site to observe variation in data, and reason about and engage with potential sources of variability. Throughout this paper, we use the term data preparation to refer to the evaluation and manipulation of a dataset in preparation for analysis.

In prior work (Wilkerson & Laina, 2018), we found that when evaluating existing data about rodent populations in their city of residence, late elementary-aged children attended to variation and considered sources of variability in several public datasets. They critiqued the role of observational methods (reported sightings) in introducing error into one provided dataset, and questioned the validity of another dataset that included unexpected variability across different months of the year. These evaluative stances led some students to discard datasets, using other more reliable sources for analysis. Other students opted to manipulate the data—for example, by using the mean of reported sightings for a given season, rather than month-by-month reports, to smooth variation that they deemed inconsequential. Here, we extend these preliminary insights to a context where learners use computational tools to actively evaluate and manipulate data for analysis.

Figure 1 illustrates phases during modeling activity during which students have the opportunity to create, notice, or explore variation in data. During traditional statistical modeling activities (light grey shading), students have opportunities to observe, measure, and record data from a real-world context—often including making decisions about what measurement tools to use and what observations to make in the first place. They may also be invited to generate data through mechanistic simulations they build, based on observations or understandings of the real-world context they wish to describe. They then model the constructed data, evaluate the degree to which the model matches the context they which to describe, and revise their work accordingly.

In contrast, when modeling with public datasets, the space of interaction with data and variability is reduced significantly. The only place where students come into contact with the variation inherent in data is when they are approximating data by modeling it using a simulation (as in the case of the civic statistics project described above; medium grey box in figure), or *after* the data has been collected. If statistically modeling the data using software tools such as Tinkerplots is not an option, the only opportunity left for students to directly engage with variation and consider sources of variability in data is during preparation and analysis (focus of the current study, dark grey box in the figure).

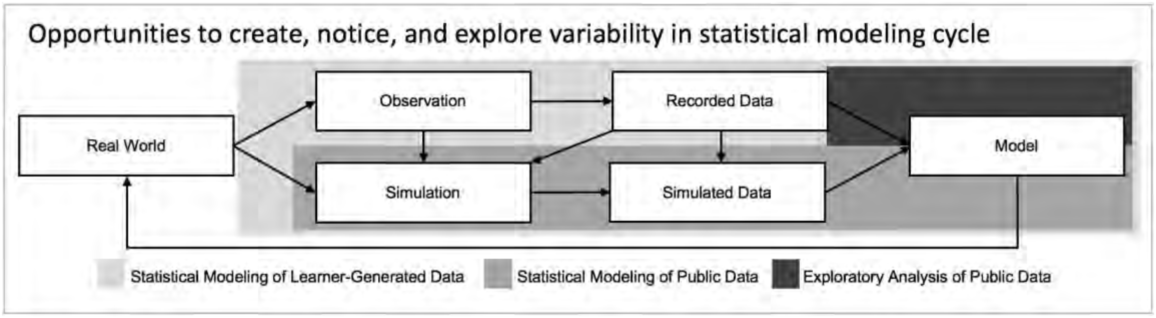

Opportunities to create, notice, and explore variability in statistical modeling cycle
Real World
Observation
Simulation
Recorded Data
Simulated Data
Model
Statistical Modeling of Learner-Generated Data
Statistical Modeling of Public Data
Exploratory Analysis of Public Data

Fig 1. Opportunities to explore variability in traditional modeling (light grey), statistical modeling of public data (med grey), and exploratory analysis of public data (dark grey).

These differences in opportunities to interact directly with data and variability may help explain why historically, data preparation—the step after observing and recording data, but before analyzing and modeling it—has not been a major area of focus in the statistics education literature. Certainly, the organization, transformation and representation of data, often referred to as "transnumeration" (Pfannkuch & Rubick, 2002; Chick, Pfannkuch, & Waston, 2005), has long been considered an important step in the modeling process. However, transnumeration is assumed to be a minor part of the statistical modeling cycle. For instance, "data cleaning" is listed alongside data collection and management in Wild & Pfannkuch's (1999) foundational work. Modeling with public datasets, in contrast, introduces a case where data preparation is the *only* available means for learners to (re)construct data in service of their investigative goal.

Theoretical Framework

To theorize how *data preparation* can serve a site for engaging with variability in public data, we consider (1) the conditions under which data preparation has become important, and (2) what constitutes data preparation actions *(data moves)* and *variability* in this context. It is well known that representational technologies have an important impact on student sensemaking in statistics. The proliferation of research exploring inference and variability using modeling with simulations, for example, has been explicitly linked to the development of the Sampler tool in Tinkerplots (Ben-Zvi, Bakker, & Makar, 2015; Pratt & Kazak, 2018; Watson & Donne, 2009). Our theoretical approach draws from the learning sciences literature to examine how representational technologies, as cultural artifacts, shape what *actions* learners can take during data preparation, and therefore what *types of variability* become evident to learners as they interact with public datasets in these ways.

Data Analysis Tools as Epistemic Forms

We draw from Collins and Ferguson's notion of *epistemic forms* and *epistemic games*[1] (1993) to guide our understanding of the origin and purpose of tools used to manipulate and analyze data. Epistemic forms are representational structures that can be populated by practitioners to organize, reflect upon, and expand their knowledge—such as lists, tables, or graphs. Epistemic games are the ways of thinking that allow them to effectively populate and make use of those forms—for example, reasoning about what might be reasonable axes on a graph, or recognizing and developing methods to fill in missing data in a table. Epistemic forms represent a particular type of convention shared by communities of practice, which has been developed over time to support patterns of activity characteristic of that community (Saxe, 1991). The Tinkerplots Sampler tool as an epistemic form, for example, has highlighted and supported the statistics education communities' interest in emphasizing causes for variability by supporting learners in a "data factory" (Konold, Harradine, & Kazak, 2007) epistemic game. In this game, random devices are configured to approximate real-world mechanisms so that they, in turn, are able to manufacture data that approximates real-world patterns.

[1] White, Collins, & Fredericksen (2011) call these "model types" and "modeling strategies". We use the original terms (a) to emphasize knowledge construction as result of data moves, and (b) to avoid confusion with other definitions of modeling used in the special issue.

We, in contrast, are centrally concerned with the processes of preparing public data for goal-driven analysis. The statistics and data science communities have developed a number of software packages (R, python pandas, Tableau, etc.) intended to support this goal. These tools, though diverse in technical implementation, all share a core purpose—to support the alignment of a complex dataset's structure and scope with an investigator's proposed questions. A corresponding collection of educational software, including the Common Online Data Analysis Platform (CODAP; Finzer & Damelin, 2014) used in this study, is now being developed. The emergence of these technologies represents a new *epistemic form*, complete with its own grammar (Wickham, Francois, Henry, & Müller, 2015). The epistemic game they are built to support is referred to in the professional sphere as "data wrangling" (Kandel, et al. 2011) and in education as "data moves" (Erickson, Wilkerson, Finzer & Reichsman, 2019). This game is guided by a given investigator's goals and values, and by the ways in which their understanding of the contexts in which data are collected impact the reliability and validity of those data (Wilkerson, et al., 2018).

Data Moves as Opportunities to Engage (or Ignore) Variability in Public Datasets

In this study we explore how the specific actions, or *data moves*, an investigator can execute during data preparation might expose or hide variation (and thus, potential sources of variability) in a dataset. Erickson and colleagues (2019) define a data move as:

> "an action that alters a dataset's contents, structure, or values. Altering a dataset's contents means changing the cases or attributes already present in the dataset: adding or removing rows or columns. Altering a dataset's structure means changing the way that cases, attributes, and values are related to one another. Altering a dataset's values is simply changing the values in the cells. Some data

> moves, such as merging, may alter both contents and structure." (Erickson, Wilkerson, Finzer & Reischmann, 2019, p.3)

Erickson et al (2019) describe in detail a number of specific data moves that they have found useful for novice data investigators; in this study, we observed the actions *sorting*, *filtering*, *grouping*, and *calculating* to be most often leveraged by students.

We are interested in how executing data moves provides students with opportunities to engage or ignore *variability* in public datasets. Reading and Shaugnessy (2004) describe variability as "something that is apt or liable to vary or change" (p.201), which is reflected within a given dataset as *variation* among recorded values in one or more dimensions. We consider students' *engagement with variability* as the ways in which students reason about the potential sources or causes for variability in a dataset. Given the complexity and multidimensionality of the types of public datasets that we are most interested in, we are especially concerned in how learners make sense of what Wild and Pfannkuch (1999) characterized as unexplained variation in data—through considering potential causes for variability including patterns and relationships among recorded variables, among recorded and unrecorded variables, and/or as a result of within-variable factors such as natural variation or measurement error.

As an example of how data moves might create or limit students' opportunities to reason about variability in a dataset, consider the following scenario. We illustrate the interaction between data moves and variability in this example graphically for clarity and brevity; however, the data we present in this study establish that data moves can exist somewhat independently from graphing in practice. The National Health and Nutrition Examination Survey (NHANES) is a publicly available collection of health and nutrition related data from a representative sample of participants in the United States. Imagine that a student investigator is exploring a dataset from NHANES that includes the heights and ages of 800 children, with a relatively uniform sampling across ages 5-19. Fig 2 displays a graph of all recorded heights for this dataset. The shape is unusual, with a "hump" between 150 and 175 cm. The clever analyst will note that childrens' growth slows as they approach adult height in their teen years, and may conclude that the variability observed within this graph can be largely explained by differences in age.

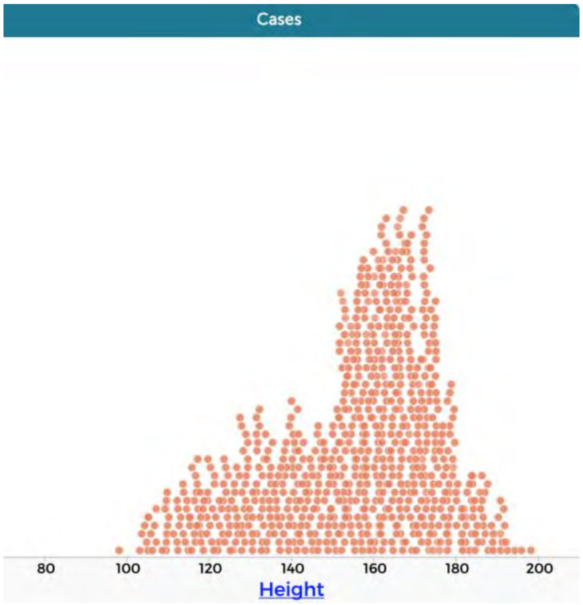
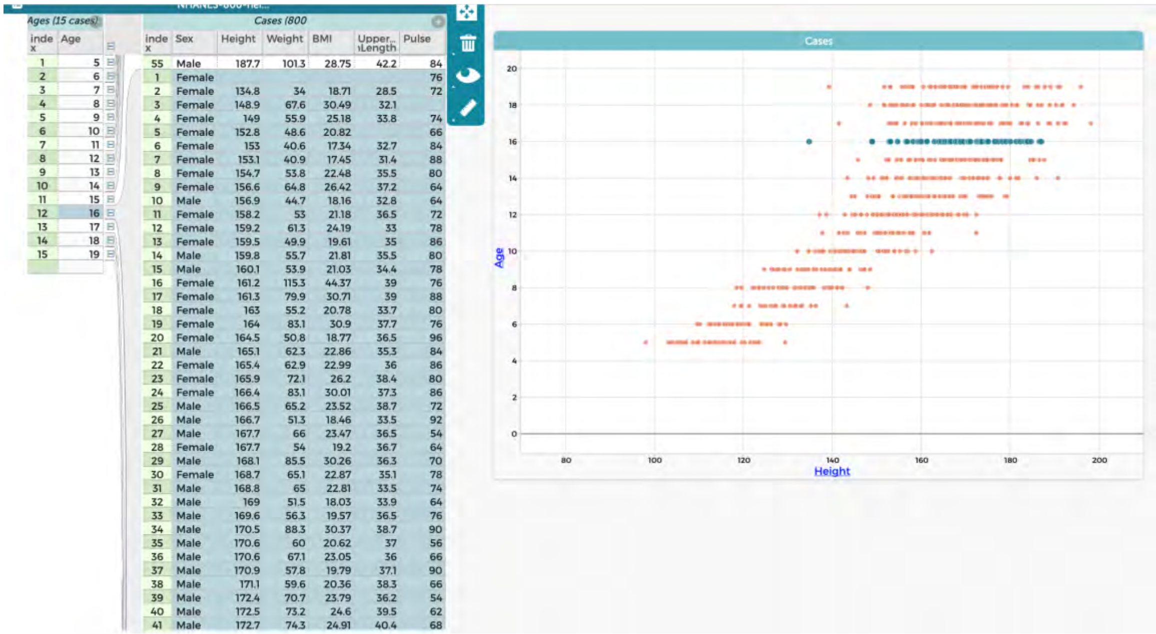


Fig 2. A plot of the distribution of heights of children age 5-19.

Grouping this dataset by age, however, can serve to foreground additional sources of variability. Fig 3 demonstrates that within age *groups*, there is still considerable variation in heights, including that some younger children have recorded heights that are taller than *all* of their older peers (see differences in ages 11 and 12, 16 and 17). This highlights that the dataset follows *different* participants at each age; therefore, it does not include information about how a participant's prior height impacts their current height and may not be appropriate for exploring patterns of individual growth. There are also differences in the extent of variation by age group, with younger children's heights varying less than those of older children. In this way, grouping the data reveals how natural variation emerges both across and within age in ways that, depending on an analyst's goals, might influence how these height data should be interpreted.



Ages (15 cases)

index	Age
1	5
2	6
3	7
4	8
5	9
6	10
7	11
8	12
9	13
10	14
11	15
12	16
13	17
14	18
15	19

Cases (800

index	Sex	Height	Weight	BMI	Upper... ıLength	Pulse
55	Male	187.7	101.3	28.75	42.2	84
1	Female					76
2	Female	134.8	34	18.71	28.5	72
3	Female	148.9	67.6	30.49	32.1	
4	Female	149	55.9	25.18	33.8	74
5	Female	152.8	48.6	20.82		66
6	Female	153	40.6	17.34	32.7	84
7	Female	153.1	40.9	17.45	31.4	88
8	Female	154.7	53.8	22.48	35.5	80
9	Female	156.6	64.8	26.42	37.2	64
10	Male	156.9	44.7	18.16	32.8	64
11	Female	158.2	53	21.18	36.5	72
12	Female	159.2	61.3	24.19	33	78
13	Female	159.5	49.9	19.61	35	86
14	Male	159.8	55.7	21.81	35.5	80
15	Male	160.1	53.9	21.03	34.4	78
16	Female	161.2	115.3	44.37	39	76
17	Female	161.3	79.9	30.71	39	88
18	Female	163	55.2	20.78	33.7	80
19	Female	164	83.1	30.9	37.7	76
20	Female	164.5	50.8	18.77	36.5	96
21	Male	165.1	62.3	22.86	35.3	84
22	Female	165.4	62.9	22.99	36	86
23	Female	165.9	72.1	26.2	38.4	80
24	Female	166.4	83.1	30.01	37.3	86
25	Male	166.5	65.2	23.52	38.7	72
26	Male	166.7	51.3	18.46	33.5	92
27	Male	167.7	66	23.47	36.5	54
28	Female	167.7	54	19.2	36.7	64
29	Male	168.1	85.5	30.26	36.3	70
30	Female	168.7	65.1	22.87	35.1	78
31	Male	168.8	65	22.81	33.5	74
32	Male	169	51.5	18.03	33.9	64
33	Male	169.6	56.3	19.57	36.5	76
34	Male	170.5	88.3	30.37	38.7	90
35	Male	170.6	60	20.62	37	56
36	Male	170.6	67.1	23.05	36	66
37	Male	170.9	57.8	19.79	37.1	90
38	Male	171.1	59.6	20.36	38.3	66
39	Male	172.4	70.7	23.79	36.2	54
40	Male	172.5	73.2	24.6	39.5	62
41	Male	172.7	74.3	24.91	40.4	68

Fig 3. A grouping data move reveals the distribution of children's heights by age.

Finally, consider two different data moves that an investigator might make to explore the relationship between children's height and their sex as assigned at birth. An analyst may calculate a new summary statistic, "Mean Height by Age and Sex," which computes a single mean statistic for each group of participants by Age and Sex identification. This mean highlights an interesting pattern whereby children identified as female are on average shorter than those identified as male, except for a few years in early adolescence. However, it obscures a different pattern made evident by filtering the data points so that those identified as female are highlighted, and comparing the distributions—in this case, the two groups share very similar distributions until age 10.

When preparing data for analysis in this way using data moves, students may have a number of opportunities to expose and explore, or to collapse, the variation in the data they are working with. These also serve as opportunities for the students to consider, more generally, the many potential sources of variability from which this variation arises. In large, publicly-available datasets such as NHANES, such explorations can also lead to larger reflections upon the data context (e.g., patterns of individual growth), the datasets' multidimensional nature, and the types of inferences that are possible.

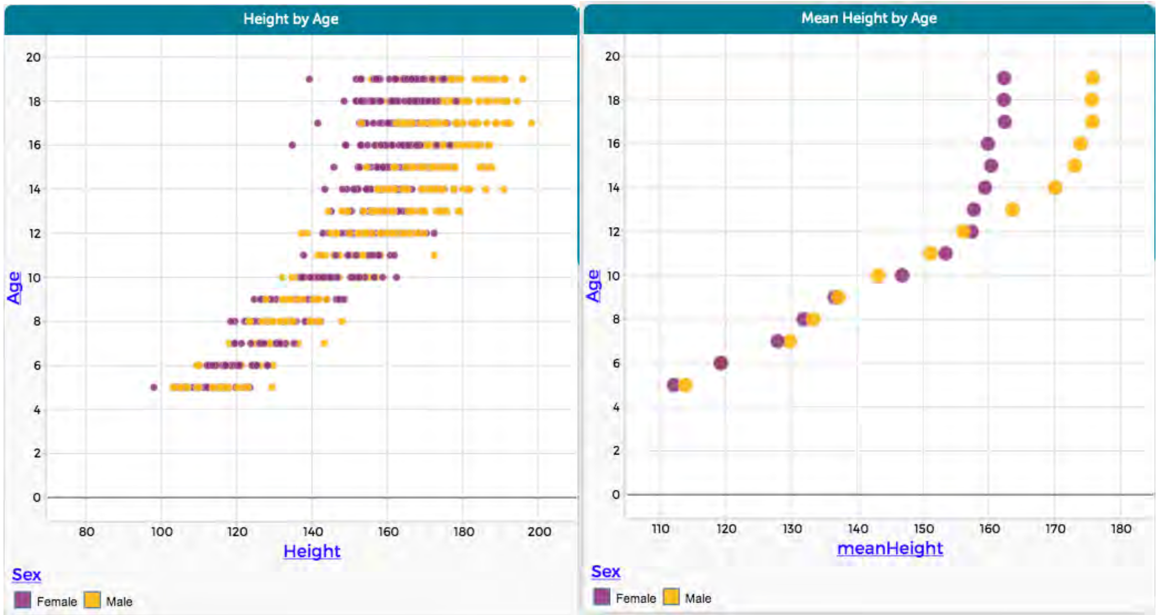


Fig 4. Two graphs, each produced by grouping data by sex assigned at birth. On the left, a filtering move highlights those children assigned female at birth. On the right, an additional calculation move is used to describe the mean of children's heights by age and sex.

Given the theoretical language elaborated in this section, we more precisely state our research question as: What opportunities were there for participating youths to *engage with variability*—that is, to reason about potential unexpected causes for variation—when preparing public datasets for analysis using *data moves*?

Methodology

We draw our data from a series of three interviews conducted 2-3 weeks apart with five pairs of young adults ages 15-18. Participants were recruited through emails and fliers distributed at public high schools in the San Francisco Bay Area of the U. S. State of California. The recruitment materials specifically invited youth to bring a partner, so all participant pairs were very comfortable working together and had pre-existing relationships as friends or siblings. Some participants reported limited experience with statistics and/or computer science through coursework; none reported instruction in data science or computational data analysis. None of our participants had prior experience with the CODAP tool used in this study (described below), or similar data analysis tools. To aid with study retention, participants received an increasing amount of compensation: US$15 for the first interview, US$20 for the second, and US$25 for the third interview.

As we describe in more detail below, our original interest was to explore how secondary-level students with limited experience manipulated and analyzed data within a tool-mediated instructional context. We chose to focus on this population because there are increasing calls for data-rich exploratory investigations to be introduced to high school and university students across the curriculum, regardless of students' statistical backgrounds or futures (Finzer, 2013; Gould, et al, 2017).

Design of the Interview

The study we present here is a re-analysis of interview data originally collected to study when and how students decide to enact data moves, rather than to more specifically explore how they engage with variability. Therefore, we did not directly ask students to consider variation in the data they were working with, or what potential sources of such variation might be. Instead, we presented them with a number of tasks that created an emergent need for particular data moves. Upon re-analysis of the data, we found that each data move executed by students presented implicit and explicit engagements with variability in ways that posed important instructional opportunities.

Our original study design called for a semi-structured, task-based clinical interview methodology (Maher & Sigley, 2014, p. 579). We sought to document students' naïve approaches to manipulating and analyzing complex datasets, and how such knowledge develops as they repeatedly analyzed public datasets with CODAP and a knowledgeable facilitator. We opted for a *repeated* version of task-based interview methodology (as described in e.g. Goldin, 2000; Wilkerson, Shareff, & Laina, Under Review) in order to capture longitudinal development. Given the novelty of the tasks and software (which we describe in more detail below, and in the Appendices), and the exploratory nature of this stage of work, interviews represented a more controlled environment for initial study than classrooms. In the interest of supporting student interest and agency, we selected datasets that reflected socioscientific topics about which youth are likely to have relevant knowledge and interest (e.g., health, local public transit patterns, ecology, local drought).

In total, participants completed three approximately one-hour-long interviews for the study. All protocol materials, including details about each dataset used and all task-based questions, are available in the Appendices A and B. In the first interview, all five pairs of participants watched a four-minute video that introduced them to the Common Online Data Analysis Platform ("CODAP"; Concord Consortium, 2014), which briefly reviewed the data manipulation capabilities of the software. A transcript of this video, along with key screens presented within the video, are available in Appendix C. They were then introduced to a public health dataset (U. S. National Health and Nutrition Survey or

“NHANES”; see Appendix A). Students were invited to ask questions about the dataset, and to pose any questions they may wish to investigate with it.

Next, we provided a list of five questions. The questions were designed to motivate particular data moves (e.g., sorting, grouping, filtering), in order to trace how learners’ use of each move develops over time and across datasets. For example, one question associated with the NHANES dataset asked participants to find the average pulse rate for teenagers, which requires a “grouping” data move so a summary statistic can be computed on observations for an entire age range. After participants investigated 2 or 3 of the provided questions, we invited them to revise and investigate their own questions. One major reason for returning to participants’ questions only after they pursued questions designed by the researchers was to allow them time to develop fluency with the dataset and at least a subset of available data moves.

During all interviews, the facilitators (all authors of this paper) observed learners’ actions and asked clarifying questions (e.g., “why did you choose to do that?”, “What is it that you are trying to do right now?”, “Why is that pattern exciting to you?”) when learners engaged in data moves, interpreted patterns, or expressed confusion, hesitation, or excitement during the interview. Facilitators reminded participants of the introductory instructional video, and offered assistance when students exhibited confusion about what was possible for them to do using the CODAP tool.

The second and third interviews followed a format similar to the first, but allowed participants to select from a set of three datasets. These dealt with the topics of climate and drought, public transit services, and ecosystem dynamics. They also had different structures and challenges: the transit dataset was so big that database queries were needed, the drought dataset was hierarchical, and the ecosystem data included missing observations. Details about each dataset and associated questions are in Appendix B.

Analysis

Our interviews were originally designed to elicit and capture participants' development and use of data moves during goal-oriented data exploration. Later, when we began to analyze our video data collaboratively with project advisors and other colleagues, we recognized the potential that our dataset had to also speak to students' engagements with variation and variability. This led us to pursue analysis in two phases.

Phase 1 – Identifying Data Moves. During each interview, students' conversations were audio and video recorded, and their actions on the computer screen were captured. Any written notes produced during the interview were collected. These materials were assembled using the Camtasia tool (TechSmith, 2002) into a consolidated file that synchronized all audio, video, and screen captured sources. A researcher that had not conducted a given interview then reviewed the consolidated file and produced preliminary notes identifying, alongside other meaningful actions such as goal setting, considering data context, or graphing, the execution of data moves (e.g., actions that modified the dataset's content, structure, or values as described in Erickson, Wilkerson, Finzer, & Reichsman, 2019).

Phase 2 – Characterizing Engagement with Variability via Data Moves. Next, we identified students' *engagements with variability* as an area worthy of deeper study. Given that much research has explored student reasoning about variability with graphs and other visualizations, we focused specifically on engagement with variability during the execution of data moves. We revisited the corpus of interview videos, and for each segment of video that captured the execution of a data move, identified students' engagement with variability during that segment. More specifically, we marked whether (a) the participants engaged with variation in the data as "unexplained" (e.g., not a result of an anticipated relationship), (b) if so, what (if anything) they identified as a possible source of variability, and (c) what specifically about the manipulation performed, resulting data, and/or researcher support may have facilitated attention to variability.

Results

Our research question was: What opportunities were there for participating youths to *engage with variability*—that is, to reason about potential unexpected causes for variation—when preparing public datasets for analysis using *data moves*? Table 1 summarizes how frequently participants engaged with variability while enacting data moves, and what types of data moves provoked such engagement.

	Interview 1	**Interview 2**	**Interview 3**
Pair 1 **Caleb & Emma**[2]	NHANES 12 data moves engagements: *sort*, *calc*	Yellowstone 5 data moves engagement: *calc*	BART 5 data moves engagement: *sort*
Pair 2 **Jenna & Ariana**	NHANES 9 data moves 0 engagements	Yellowstone 6 data moves 0 engagements	CA Reservoirs *Incomplete data; not analyzed*
Pair 3 **Ali & Rose**	NHANES 5 data moves 0 engagements	Yellowstone *Incomplete data; not analyzed*	CA Reservoirs 8 data moves engagement: *calc*
Pair 4 **Max & Leah**	NHANES 2 data moves engagement: *sort*	Yellowstone 3 data moves engagement: *filter*	BART 5 data moves engagements: *filter*, *calc*
Pair 5 **Betty & Sadie**	NHANES 8 data moves engagement: *calc*	BART 7 data moves engagements: *filter*, *filter*	Yellowstone 4 data moves 0 engagements

Table 1: Summary of data moves executed and engagements with variability, by participant group and interview.

[2] All participant names used in this paper are pseudonyms.

Even though this study was not designed to support or elicit reasoning about variability, we found evidence that all but one group, Jenna and Ariana, did explicitly engage with variability while executing a data move. In fact, three of the five groups did so multiple times, across multiple datasets. This emerged despite differences in how frequently each pair executed data moves during the interview.

There were also notable patterns in which types of data moves most frequently yielded engagement with variability across interviews. Of the 12 data moves that yielded engagements with variability among participant groups, five were *calculating* moves, four were *filtering* moves, and three were *sorting* moves. As we demonstrate below, certain moves highlighted particular features of variation that, in turn, led learners to consider different potential sources of variability in the data (e.g., natural variation; other causal factors). In contrast, moves that reduced the visibility of variation across observations within a dataset, such as grouping or calculating summary measures, did not lead participants to explicitly engage variability.

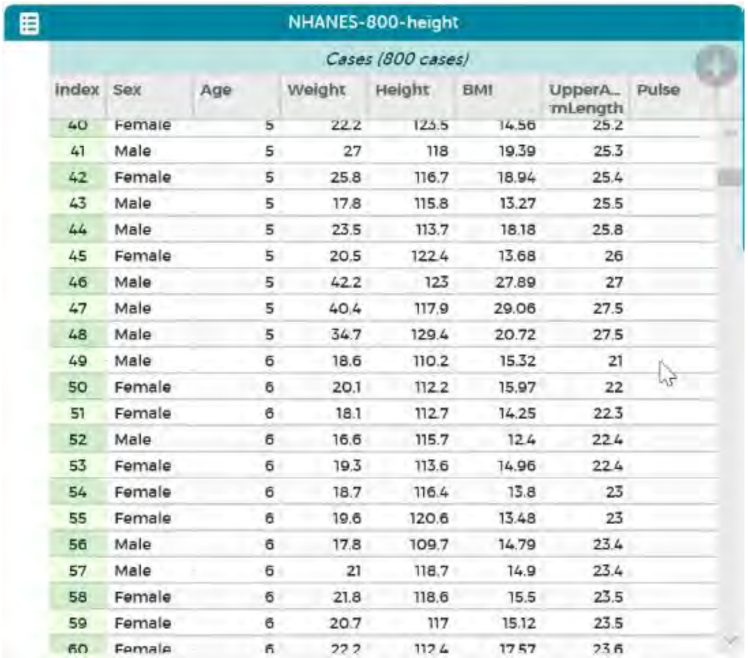
In the sections below, we present detailed examples of learner engagements with variability through detailed examples of learners performing a *sort*, *filter*, and *calculate* data move, respectively.

Pattern 1: Sorting within groups highlighted natural variation.

The act of sorting data led two groups of students to deeper engagements with variability in their datasets. In both cases, this emerged because sorting made the highest and lowest values of the measure or subset more visible, prompting participants to consider whether those extreme values still seemed reasonable for the dataset as a whole.

Consider the actions of one pair of participants, Max and Leah, working to address the question “At what age do children’s arms stop growing?” The task made use of the NHANES dataset that included observed children’s age, sex assigned at birth, weight, height, BMI, upper arm length, and pulse rate. Noting that they first wanted to find at what age the UpperArmLength measure stopped increasing, Max and Leah sorted the data by the UpperArmLength column, so records with the smallest UpperArmLength values appeared at the top of the table. They then sorted again by the Age column. This created a list of observations that increased by Age, and then by UpperArmLength among records with the same Age value. At the end a given age *n*, the largest UpperArmLength reported was followed by the shortest UpperArmLength reported for age *n+1* (Figure 4).

After sorting the table, Max began to scroll slowly through the values, presumably to get a general sense of when UpperArmLength would stop increasing. As he scrolled, the facilitator asked Max to share what he observed.



NHANES-800-height

Cases (800 cases)

index	Sex	Age	Weight	Height	BMI	UpperA... mLength	Pulse
40	Female	5	22.2	123.5	14.56	25.2	
41	Male	5	27	118	19.39	25.3	
42	Female	5	25.8	116.7	18.94	25.4	
43	Male	5	17.8	115.8	13.27	25.5	
44	Male	5	23.5	113.7	18.18	25.8	
45	Female	5	20.5	122.4	13.68	26	
46	Male	5	42.2	123	27.89	27	
47	Male	5	40.4	117.9	29.06	27.5	
48	Male	5	34.7	129.4	20.72	27.5	
49	Male	6	18.6	110.2	15.32	21	
50	Female	6	20.1	112.2	15.97	22	
51	Female	6	18.1	112.7	14.25	22.3	
52	Male	6	16.6	115.7	12.4	22.4	
53	Female	6	19.3	113.6	14.96	22.4	
54	Female	6	18.7	116.4	13.8	23	
55	Female	6	19.6	120.6	13.48	23	
56	Male	6	17.8	109.7	14.79	23.4	
57	Male	6	21	118.7	14.9	23.4	
58	Female	6	21.8	118.6	15.5	23.5	
59	Female	6	20.7	117	15.12	23.5	
60	Female	6	22.2	112.4	17.57	23.6	

Figure 4. Max and Leah inspect the difference between the highest observed upper arm length for a 5 year old and the lowest observed upper arm length for a 6 year old.

B*: What are you seeing?

M: [*scrolls*] Looks like there's not a huge amount of change, it goes up slowly.

L: But it does change.

M: This is all for age 5.

L: I know, so then

M: There's a decent amount of variation [*scrolling; pauses on Figure 4*].

L: Keep going down.

M: Looks like the highest for 5 is still decently higher than the lowest for 6.

L: Well keep going and see if there's any change when people get to their teenage years.

Here, both Max and Leah intended to use the sorting feature to more quickly identify at what Age the UpperArmLength attribute in the dataset stops increasing. When Max begins scrolling through the sorted data, they both note these values "change" and "[go] up slowly", as one might expect in patterns of growth. However while scrolling, Max encounters a dramatic decrease in arm measurements *across* ages 5 and 6 given how the data were sorted. At this point he states "there's a decent amount of variation" in the data. Eventually, these participants plotted UpperArmLength by Age, but the majority of their explicit discussion of variability occurred here, before any plot was created.

Though subtle, this episode demonstrates how sorting highlighted for Max both variation within age groups ("This is all for age 5"), and across them ("the highest for 5 is still decently higher than the lowest for 6"). These shifts in Max's attention to variation were complemented by a shift in Leah's scope of analysis as it relates to potential sources of variability for these data. Once Max articulated record-to-record differences in their sorted data as a “decent amount of variation” within age groups (referring to, we suspect, natural variation), Leah reoriented their attention to differences during adolescence development (“keep going…to their teenage years”).

Pattern 2: Filtering focused attention on potential sources of variability.

Whereas sorting drew participants’ attention to natural variation in our study, filtering often encouraged them to consider other potential sources of variability in data, such as confounding variables or patterns that might naturally emerge over time and space. Several groups, for example, used filtering moves with the BART dataset to limit variation in transit data due to known differences in travel patterns (e.g., weekend versus weekday) or train stop locations (e.g., a residential versus commercial area).

One such instance emerged as Max and Leah were exploring data related to the reintroduction of wolves to Yellowstone park. In the episode below, they had just begun working on the question, “Which variables do you think are the best indicators of the overall health of the ecosystem? Based on this, which year would you argue was the park the healthiest?” Particular to this investigation, the dataset included yearly records of animal and plant populations, along with percentages of plant life that had been eaten or “browsed” in two regions of the park (river and uplands). Soon into this inquiry, Max highlighted four rows of data, the years between 2002 and 2004, and inspected them quietly for some time. Eventually, the researcher Becca prompted Max to think aloud:

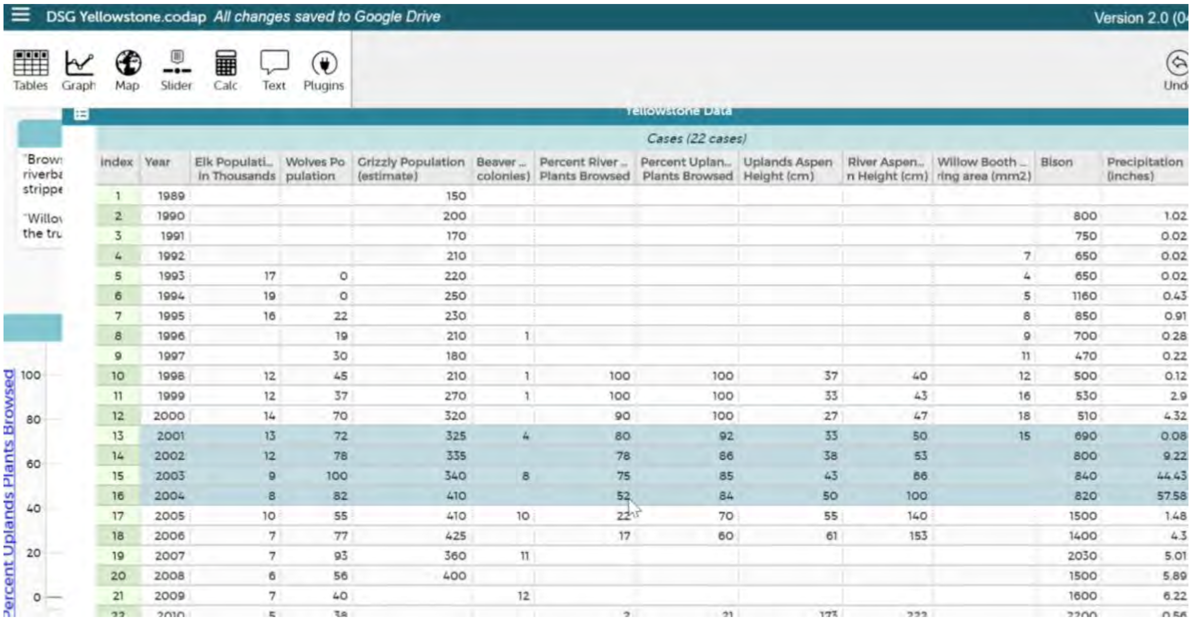
B*: It sounds like you’re saying one of the things you’re looking for is, like, non-drastic changes in animal population or in plant browsing. And, just watching you use the mouse, it looked like you had highlighted some area of time where that seemed to be happening. What years were that?

M: [*highlights three rows of table*] These. This is what I was looking at. This is where on both of these [*indicates percent river plants and percent uplands plants browsed*] it seems fairly steady. Well actually, it's more these four [*highlights 2001-2004; Figure 5*], and this [*indicates 52 percent river plants browsed in year 2004*] being a little bit of an outlier.

B*: What are you calling an outlier?

M: This [*indicates same value of 52*], because, in these three years [*2001-2003*] on both of these [*percent browsed columns*], it's not change, I mean this [*percent river plants browsed*] percent changes by a total of five percent, this [*percent uplands plants browsed*] percent goes down by, what, seven percent. So they're staying about the same, there's a pretty steady trend and it's not, it doesn't seem to be having like going from 75 to 52. You don't have, you know, a change of two, then three, then twenty. And I think that, that seems like it would be indicative of a healthy ecosystem because you wouldn't have any ecosystem that's not having huge changes. [...]

L: I do wonder how much temperature affects the wildlife. Because in some cases it seems like it's a really big. Like I know when you look at climate change statistics, like, it's a very like tiny increase, but it still matters a lot.



Yellowstone Data

Cases (22 cases)

index	Year	Elk Populati... in Thousands	Wolves Po pulation	Grizzly Population (estimate)	Beaver ... colonies)	Percent River ... Plants Browsed	Percent Uplan... Plants Browsed	Uplands Aspen Height (cm)	River Aspen... n Height (cm)	Willow Booth ... ring area (mm2)	Bison	Precipitation (inches)
1	1989			150								
2	1990			200							800	1.02
3	1991			170							750	0.02
4	1992			210						7	650	0.02
5	1993	17	0	220						4	650	0.02
6	1994	19	0	250						5	1160	0.43
7	1995	16	22	230						8	850	0.91
8	1996		19	210	1					9	700	0.28
9	1997		30	180						11	470	0.22
10	1998	12	45	210	1	100	100	37	40	12	500	0.12
11	1999	12	37	270	1	100	100	33	43	16	530	2.9
12	2000	14	70	320		90	100	27	47	18	510	4.32
13	2001	13	72	325	4	80	92	33	50	15	690	0.08
14	2002	12	78	335		78	86	38	53		800	9.22
15	2003	9	100	340	8	75	85	43	86		840	44.43
16	2004	8	82	410		52	84	50	100		820	57.58
17	2005	10	55	410	10	22	70	55	140		1500	1.48
18	2006	7	77	425		17	60	61	153		1400	4.3
19	2007	7	93	360	11						2030	5.01
20	2008	6	56	400							1500	5.89
21	2009	7	40		12						1600	6.22
22	2010	5	38			2	21	173	222		2200	0.56

Figure 5. Max filters records that he argues are reflective of a healthy ecosystem.

Using the filtered (highlighted, for the purpose of focusing attention on only these values) data as an anchor, Max and Leah negotiated whether the variation they observed within the “percent browsed” columns from year to year constituted signal or noise. Max argued that “fairly steady” indicators (1-3 percent change annually) reflect health, while “huge changes” (more than 5 percent; or a 20 percent change) signal a problem. Leah, attending to a smaller fluctuation in a different column within the highlighted range, countered that in climate even a “very tiny increase [in temperature] still matters a lot”.

As Max and Leah considered what counted as meaningful variation within this highlighted segment of data, their attention expanded to new columns in the table—moving from first the plant indicators to weather-related indicators. This attention to other measures also expanded their lens on what could be factors contributing to variability in the dataset as a whole (as a reflection of ecosystem stability). This conversation continued—after discussing whether changes in temperature and precipitation might meaningfully impact the system, Max and Leah then proposed even more potential sources for ecosystem stability or disruption, including competitive and symbiotic relationships among wildlife.

Pattern 3: Calculating revealed new patterns of variation or clarified existing ones.

Finally, participants engaged with variability when they used calculating moves to convert existing measures into ones that were more interpretable, or to reveal new patterns of variation. For example, the NHANES dataset that all participants analyzed during the first session reported patients' weights in kilograms. Two groups used calculation to convert these weights from kilograms to pounds, which are more commonly used in the U. S. Although this conversion did not impact the shape or relative distribution of data, it did change how these participants could reason about the data context. Because they were able to compare reported weights to their everyday experiences of weight reported in pounds, the participants were more comfortable examining what constituted an expected mean, or a reasonable range and shape for the data's distribution for a given age of patients.

One example of calculating to reveal new patterns of variation emerged as Ali and Rose worked with the dataset that reported weekly reservoir levels in the U. S. state of California during a multi-year drought. To address a question these participants posed themselves, "Did water levels drop long-term as a result of the drought?", they calculated the difference between the most recent (post-drought) and first (pre-drought) water levels recorded for each reservoir. At first, they scanned their new column values in the table:

R: There's negatives.

A: See, that's a good thing because we know it went down. But [*quickly indicating various values*] this one went up, and that one went up, and that one went down.

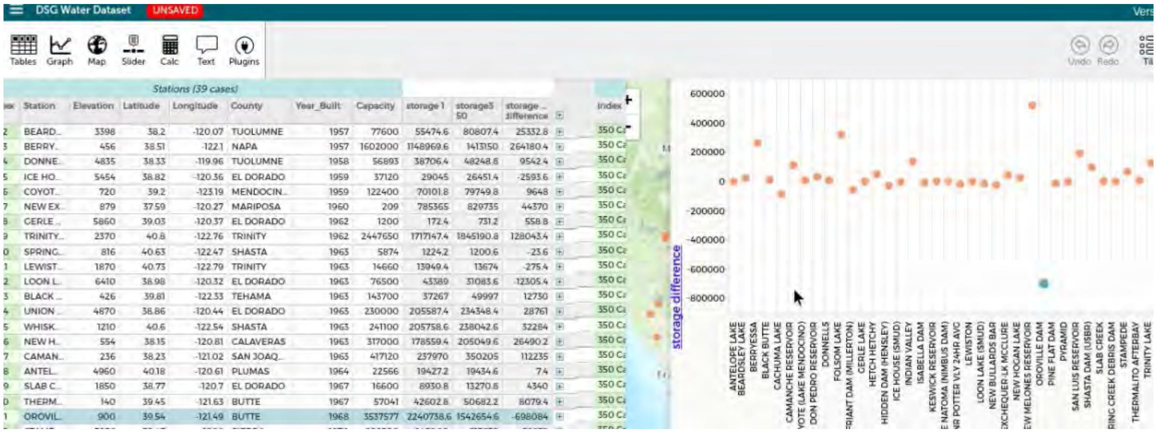
R: So now we can say, well the question was did the levels of the reservoirs drop?

A: No. Because clearly some of them went up and some of the went down.

R: This one didn't drop at all, it gained like 63. And that's such a small number. Look, 7.4. That didn't change. But this one changed a lot and got a lot bigger.

K*: Can you graph it? What would it look like?

R: [*creates graph featured in Figure 6*] Oh, a lot of them didn't change that much.



Stations (39 cases)

ex	Station	Elevation	Latitude	Longitude	County	Year_Built	Capacity	storage 1	storage3 50	storage ... difference
2	BEARD...	3398	38.2	-120.07	TUOLUMNE	1957	77600	55474.6	80807.4	25332.8
3	BERRY...	456	38.51	-122.1	NAPA	1957	1602000	1148969.6	1413150	264180.4
4	DONNE...	4835	38.33	-119.96	TUOLUMNE	1958	56893	38706.4	48248.8	9542.4
5	ICE HO...	5454	38.82	-120.36	EL DORADO	1959	37120	29045	26451.4	-2593.6
6	COYOT...	720	39.2	-123.19	MENDOCIN...	1959	122400	70101.8	79749.8	9648
7	NEW EX...	879	37.59	-120.27	MARIPOSA	1960	209	785365	829735	44370
8	GERLE ...	5860	39.03	-120.37	EL DORADO	1962	1200	172.4	731.2	558.8
9	TRINITY...	2370	40.8	-122.76	TRINITY	1962	2447650	1717147.4	1845190.8	128043.4
0	SPRING...	816	40.63	-122.47	SHASTA	1963	5874	1224.2	1200.6	-23.6
1	LEWIST...	1870	40.73	-122.79	TRINITY	1963	14660	13949.4	13674	-275.4
2	LOON L...	6410	38.98	-120.32	EL DORADO	1963	76500	43389	31083.6	-12305.4
3	BLACK ...	426	39.81	-122.33	TEHAMA	1963	143700	37267	49997	12730
4	UNION ...	4870	38.86	-120.44	EL DORADO	1963	230000	205587.4	234348.4	28761
5	WHISK...	1210	40.6	-122.54	SHASTA	1963	241100	205758.6	238042.6	32284
6	NEW H...	554	38.15	-120.81	CALAVERAS	1963	317000	178559.4	205049.6	26490.2
7	CAMAN...	236	38.23	-121.02	SAN JOAQ...	1963	417120	237970	350205	112235
8	ANTEL...	4960	40.18	-120.61	PLUMAS	1964	22566	19427.2	19434.6	7.4
9	SLAB C...	1850	38.77	-120.7	EL DORADO	1967	16600	8930.8	13270.8	4340
0	THERM...	140	39.45	-121.63	BUTTE	1967	57041	42602.8	50682.2	8079.4
1	OROVIL...	900	39.54	-121.49	BUTTE	1968	3537577	2240738.6	1542654.6	-698084

Figure 6. Ali and Rose executed 3 data moves to calculate the final (post-drought) minus initial (pre-drought) water levels for each reservoir, and graphed their results.

K*: What do you mean?

R: Like, this is zero [*indicating horizontal axis at y=zero*].

A: And zero means they didn't change.

R: And a lot of them are near zero.

A: And then these two [*Oroville, Pine Flat*] did drastically change in different ways.

R: Well then maybe we can say that they didn't change that much. I mean they didn't drop, the same levels. Because a lot of these remained similar levels or added some. And only a couple increased or decreased drastically.

Ali and Rose began by assessing whether their new calculated measure fit their expectations, noting that "negatives" are "a good thing because we know [water levels] went down" during the drought. Orienting to an expected negative value, however, highlighted for these participants the variation inherent in the newly calculated “storage difference” measure, as they began to notice that it featured several (unexpected) positive values. As they continued to explore, Ali and Rose also began to articulate the magnitude of variation present in their measure, distinguishing “small numbers” from those that indicate a water level “changed a lot”.

Ali and Rose then followed Kathryn’s recommendation to create a graph of their “storage difference” measure. Noting that most data points cluster “near zero”, they conclude that only two data points “drastically change in different ways” (indicating positive and negative change), and thus “maybe we can say [water levels] didn’t change that much.” While it was not a focus of their exploration (and was outside of the scope of their question), exploring the graph also gave Ali and Rose an opportunity to consider geographic space as a source of variability in this new measure. After they clicked on data points representing more extreme differences in pre- and post-drought water levels (in Figure 6, the extreme low value is selected), these points became highlighted on a map of California in CODAP. This led Ali and Rose to observe and question, in passing, whether location might have impacted water levels.

Summary.

These episodes demonstrate more precisely how data moves, as executed using CODAP as a material resource, helped learners engage with variability in the public datasets they explored. Sorting emphasized natural variation in datasets for both groups that used it. Filtering helped participants reduce expected variation by eliminating potential sources of variability in data (as groups did with BART data), and focused attention on how measures both within and beyond the dataset might contribute to the variability of a measure (as Max and Leah found with Yellowstone data). Calculating summary statistics hid variation within data from participants, reducing their focus on variability. However, calculating to make existing measures more interpretable (such as changing kilograms to pounds) allowed learners to use contextual knowledge to reason about patterns of variation, and related sources of variability. Similarly, calculating new measures (such as Ali and Rose's measure of pre- and post-drought water levels), revealed new patterns of variation, and related sources of variability, to explore.

Discussion

This paper begins to address what may appear to be a problem with current efforts to engage learners in analyzing large scale, public data. Though it is important to understand these types of data (Gould, 2017; Ridgway, 2015), they are constructed by others and are sparse and multidimensional in nature. This makes reasoning about sources of variability more challenging. We argue that one way to address this challenge is to engage learners with variability during the statistical inquiry cycle (Wild & Pfannkuch, 1999), in particular during *data preparation*—the phase of statistical inquiry whereby investigators purposefully evaluate and manipulate ("clean") data for analysis.

In this study, we conducted a secondary analysis of interviews designed to engage young adult learners in *data moves* (Erickson et al, 2019)—specific actions one can take to change the contents, structure, or values of a dataset—as they explored a variety of public datasets using CODAP. Our goal was to find out whether data preparation, and specifically the enactment of data moves, led participants to engage with variability in these datasets. We found that although all but one of the five groups interviewed did engage with variability at some point during the interview series, this happened only rarely—a total of 12 times out of 79 data moves across all groups.

Qualitative analysis of data further revealed that certain data moves—*sorting, filtering, and calculating*—led participants to engage with variability across a variety of datasets. Sorting emphasized natural variation, filtering highlighted interrelationships between measures, and calculating made existing patterns of variation more interpretable and revealed new patterns of variation to explore. These findings also highlight that there is considerable potential to better support students' engagement with variability in public datasets, through a more careful scaffolding of data moves and subsequent reasoning. Given that our study design did not explicitly support learners' attention to variability, there were likely many missed opportunities for learners to be supported in considering variability when they enacted a given data move (for example through guiding questions or suggested tasks). Calculating and filtering, as the moves that most frequently provoked emergent engagements with variability in the current study, may be high-leverage sites for designing more explicit curricular supports for exploring variability in public datasets.

This study complements literature which suggests that visualization (Gil & Gibbs, 2017) and modeling (Biehler, Frischemeier, & Podworny, 2018) are powerful ways to engage learners in reasoning about variability with public datasets. While we agree, this study establishes that additional engagement with variability can (and should, given statistical modeling's focus on *sources of variability*) occur during the reflective process of data preparation. Students executing data moves emphasized variation across specific sets of observations (e.g. at a case level), such as the interactions of variables within the ecosystem data. These are features that visual patterns such as shape and scatter (expressed at an aggregate level) may not emphasize. At the same time, we acknowledge that data moves and graphing are closely linked and mutually reinforcing. Just as data moves make some graphs possible, graphs reveal variation in ways that may encourage learners to further filter, group, or summarize their data.

This study was an exploratory, secondary analysis to examine the potential for data preparation to engage learners with variability, especially when working with public datasets for which other ways of reasoning about variability are limited. Our sample size was small and selective in that we worked with youth who agreed to travel repeatedly to a university to participate in a study about data and statistics. However as a theory-building and agenda-setting study, we've learned: data moves hold promise for engaging learners with variability; calculating and filtering are good candidates for supporting such engagement; and more explicit instructional support is likely to further deepen learners' engagement with variability during data preparation.

Conclusions

This work highlights the important role that data preparation can play in how learners reason about variability, particularly when learners are working with public datasets. Given learners' limited opportunities to meaningfully engage with the processes and events that are likely to introduce variability into such datasets, we argue that this is not only a potentially productive line of inquiry from a research perspective but that it is essential given calls for a more informed approach to data literacy across the curriculum.

Acknowledgements

We are grateful for the sustained and enthusiastic contributions from our youth participants. We also thank Susanne Schnell, Jennifer Noll, our anonymous reviewers, and the 2019 Statistical Reasoning, Thinking, and Literacy cohort for their insightful commentary and feedback on this research at various stages of formulation. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1530578. The larger project this research was a part of was conducted in collaboration with partners at the Concord Consortium.

References

Ainley, J., Gould, R., & Pratt, D. (2015). Learning to reason from samples: commentary from the perspectives of task design and the emergence of "big data." *Educational Studies in Mathematics*, *88*(3), 405–412. doi: 10.1007/s10649-015-9592-4

Aridor, K., & Ben-Zvi, D. (2018). Statistical modeling to promote students' aggregate reasoning with sample and sampling. *ZDM, 50*(7), 1165-1181.

Wilkerson, M. H., & Laina, V. (2018). Middle school students' reasoning about data and context through storytelling with repurposed local data. *ZDM Mathematics Education, 50*(7), 1223-1235. doi: 10.1007/S11858-018-0974-9

Wilkerson, M. H., Lanouette, K., Shareff, R. L., Erickson, T., Bulalacao, N., Heller, J., St. Clair, N., Finzer, W., & Reichsman, F. (2018). Data moves: Restructuring data for inquiry in a simulation and data analysis environment. Poster to appear in Proceedings of the International Conference for the Learning Sciences (ICLS 2018). London, England: ISLS.

Wilkerson, M. H., Shareff, R. L., & Laina, V. (In Progress). Learning from "interpretations of innovation" in the co-design of digital tools. Under consideration for inclusion in M-C. Shanahan, B. Kim, K. Koh, A. P. Preciado-Babb, & M. A. Takeuchi (Eds.), *The Learning Sciences in Conversation: Theories, Methodologies, and Boundary Spaces*. Routledge.

Ben-Zvi, D., Bakker, A,. & Makar, K. (2015). Learning to reason from samples. *Educational Studies in Mathematics, 88*(3), 291-303. doi: 10.1007/s10649-015-9593-3

Biehler, R. (1994). Probabilistic thinking, statistical reasoning, and the search for causes—Do we need a probabilistic revolution after we have taught data analysis. In J. Garfield (Ed.), *Research papers from ICOTS 4*, 20-37. University of Minnesota.

Biehler, R., Frischemeier, D., & Podworny, S. (2018). Elementary preservice teachers' reasoning about statistical modeling in a civic statistics context. *ZDM, 50*(7), 1237-1251.

Braham, H. M., & Ben-Zvi, D. (2017). Students' emergent articulations of statistical models and modeling in making informal statistical inferences. *Statistics Education Research Journal*, *16*(2), 116–143.

Cassel, B., & Topi, H. (2015, October). Strengthening data science education through collaboration. In *Workshop on Data Science Education Workshop Report* (Vol. 7, p. 27).

Chick, H. L., Pfannkuch, M., & Watson, J. M. (2005). Transnumerative thinking: Finding and telling stories within data. *Curriculum Matters*, *1*, 87-109.

Collins, A., & Ferguson, W. (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist, 28*(1), 25-42.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3-21.

The Concord Consortium. (2014). Common online data analysis platform. @*Concord, 18*(1), 16.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 1*3(1), 3-21.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, *16*(1), 44-49.

Erickson, T., Wilkerson, M. H., Finzer, W., & Reichsman, F. (2019). Data moves. *Technology Innovations in Statistics Education, 12*(1).

Gil, E., & Gibbs, A. L. (2017). Promoting modeling and covariational reasoning among secondary school students in the context of big data. *Statistics Education Research Journal, 16*(2), 163-190.

Goldin, G. A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education*. New York, NY, USA: Routledge.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., … Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, *10*(4), 271–288. https://doi.org/10.1177/1473871611415994

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning, 12*(3), 217-230.

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology innovations in statistics education*, *2*(1).

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259-289.

Lehrer, R. (2017). Modeling Signal-Noise Processes Supports Student Construction of a

Hierarchical Image of Sample. *Statistics Education Research Journal, 16*(2), 64-85.

Lehrer, R., Kim, M. J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning, 12*(3), 195-216.

Maher, C. A., & Sigley, R. (2014). Task-based interviews in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education*, 579-582. doi: 10.1007/978-94-007-4978-8

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1).

Pfannkuch, M., & Rubick, A. (2002). An exploration of students' statistical thinking with given data. *Statistics Education Research Journal, 1*(2), 4-21.

Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM: Mathematics Education*, *50*(7), 1113-1123.

Philip, T. M., Schuler-Brown, S., & Way, W. (2013). A framework for learning about big data with mobile technologies for democratic participation: Possibilities, limitations, and unanticipated obstacles. *Technology, Knowledge and Learning, 18*(3), 103-120.

Pratt, D., & Kazak, S. (2018). Research on uncertainty. In D. Ben-Zvi, K. Makar & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 193–228). Cham: Springer

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201-226). Springer, Dordrecht.

Ridgway, J. (2016). Implications of the data revolution for statistics education. Int*ernational Statistical Review, 84*(3), 528-549.

Saxe, G.B. (1991) *Culture and cognitive development: Studies in mathematical understanding.* Erlbaum (1991).

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA, USA: Addison-Wesley.

Watson, J., & Donne, J. (2009). TinkerPlots as a Research Tool to Explore Student Understanding. Te*chnology Innovations in Statistics Education, 3*(1). Retrieved from https://escholarship.org/uc/item/8dp5t34t

Wickham, H., Francois, R., Henry, L., & Müller, K. (2015). dplyr: A grammar of data manipulation. R package version 0.4, 3.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International statistical review*, *67*(3), 223-248.

Appendix A. First Interview Introductory Script including NHANES Data and Questions

Thank you for joining us! As you already know by now, this is a study of how people learn basic data science skills.

So over the course of this study, what I'll do is provide you with some datasets that we hope you will find interesting, and have you investigate that data to answer questions. These datasets come from the web, and so they will probably need to be transformed: organized, represented visually in different ways, or edited so that they can be explored more easily. The program we'll use, CODAP, which stands for Common Online Data Analysis Platform, has features that we hope make these kinds of transformations easy and we're here to show you how to use it. So if there is ever some kind of change you want to make to the data, let us know what it is you want to do and we'll help you figure out how. As you use the software in different ways to modify your data, we'll also keep a list here that you can refer back to as we go.

This definitely isn't a test, and we've tried to create an environment where you can really dig in. So don't worry about running out of time or anything, that's why we have a series of interviews. Our whole point is to understand how people learn to use these tools over time, as they explore things they find interesting.

We're going to start with a quick video that shows some things you can do in CODAP. Don't worry too much about remembering the specifics of how to do everything, since we'll be here. Our goal is just to give you an idea of the kinds of changes you can make to the data. You don't have to do everything you see in the video, and there are other things you can do that you won't see here. What we're most interested in is what is useful for you, depending on the questions you want to explore.

Questions asked about each dataset used in study

- Do you have any questions about this dataset or how it was made?
- Do you have any thoughts or comments about the columns (variables) that are included in this dataset?
- How would you describe the types of variables included in this data set?
- What are some patterns you would expect to see in this data?
- Okay, now, I'm curious about what kinds of questions you would be interested in exploring using this NHANES data.

NHANES Dataset

The NHANES data we have for you to explore has a lot of medical information about a random sample of 800 people. The records look like this, down here. You can see here that we have a lot of information about each person: their sex, age, weight, height, and more. Sometimes, there is data missing. This third record does not have a value for this person's pulse.

Sex	**Age**	**Weight**	**Height**	**BMI**	**UpperArmLength**	**Pulse**
Male	18	66.8	176	21.57	38.1	54
Female	17	58.7	164.7	21.64	35.5	50
Female	5	19.8	114.9	15	24.9	
Male	10	43	141.3	21.54	32.2	72

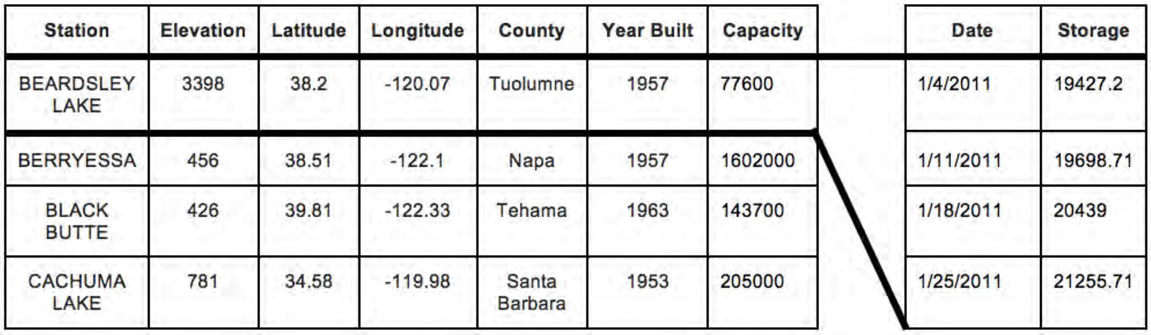
1. Based on this dataset, at what age do people's arms stop growing? [*question intended to prompt graphing*]
2. What is the average pulse rate for teenagers? [*question intended to prompt focusing, grouping, calculating*]

3. Many pulse rate readings are missing. Is there something special about the records whose values are missing? [*question intended to prompt focusing*]
4. Can you find an unexpected or surprising relationship in these data? [*question intended to prompt graphing*]
5. Are there any groups of people that you believe are overrepresented or underrepresented in these data? [*question intended to prompt filtering, grouping, searching*]

Appendix B. Choice Dataset Introductory Script and Questions

California Water Crisis/Reservoir Levels (2011-2017)

This dataset includes information about a number of reservoirs in California. It has two levels of information. One level describes each reservoir: its name, location in latitude and longitude, elevation in feet above sea level, the year it was built and its total water capacity. We also have the volume of water held in each reservoir for every week between January 2011 and September 2017.



Station	Elevation	Latitude	Longitude	County	Year Built	Capacity
BEARDSLEY LAKE	3398	38.2	-120.07	Tuolumne	1957	77600
BERRYESSA	456	38.51	-122.1	Napa	1957	1602000
BLACK BUTTE	426	39.81	-122.33	Tehama	1963	143700
CACHUMA LAKE	781	34.58	-119.98	Santa Barbara	1953	205000

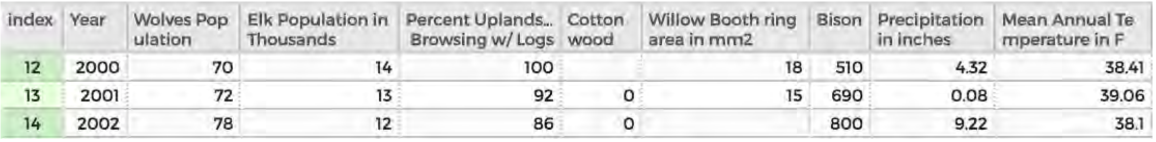
Date	Storage
1/4/2011	19427.2
1/11/2011	19698.71
1/18/2011	20439
1/25/2011	21255.71

1. During what period of time were most of the reservoirs in the dataset built? [*question intended to prompt graphing*]
2. Is there a relationship between the elevation of a reservoir and its capacity? [*question intended to prompt graphing*]
3. For just one reservoir of your choice, what month experienced the largest drop in stored water? [*question intended to prompt focusing, graphing, calculation*]
4. Which reservoir remained the fullest during the worst year of the drought? [*question intended to prompt focusing, graphing*]
5. Are there differences in how the drought affected reservoirs in northern vs southern California? [*question intended to prompt focusing, grouping*]
6. What was the worst week of the drought, taking into account *all* of the California reservoirs? [*question intended to prompt calculation*]

Yellowstone National Park Wildlife and Weather Data (1990-2010)

This dataset includes information about several animal species, plant life, and weather patterns over a span of 22 years in Yellowstone National Park. One of the reasons these data were collected was to observe the impact of the reintroduction of the grey wolf to the park. For reference, here is a map of the park and a sample of the data. Here 'Riparian Browsing' means the plant-life near a river that showed signs of being eaten. I have a sheet here that tells you details about all of the columns, if it's helpful. Note that the units for the variables vary, and not all data are present for each variable in each year.

MONTANA
BOZEMAN
BILLINGS
SOUTH DAKOTA
IDAHO
YELLOWSTONE
WYOMING
CASPER
NEBRASKA
UTAH
SALT LAKE CITY
COLORADO
CHEYENNE
DENVER
United States



index	Year	Wolves Pop ulation	Elk Population in Thousands	Percent Uplands... Browsing w/ Logs	Cotton wood	Willow Booth ring area in mm2	Bison	Precipitation in inches	Mean Annual Te mperature in F
12	2000	70	14	100		18	510	4.32	38.41
13	2001	72	13	92	0	15	690	0.08	39.06
14	2002	78	12	86	0		800	9.22	38.1

1. See if you can structure the data so that variables are grouped by similar categories (animals, plants, weather) [*question intended to prompt grouping*]
2. Is there a relationship between populations of carnivore species? [*question intended to prompt grouping, graphing*]
3. Are there differences in how the reintroduction of wolves affected species of carnivores vs. herbivores ? [*question intended to prompt focusing, grouping*]

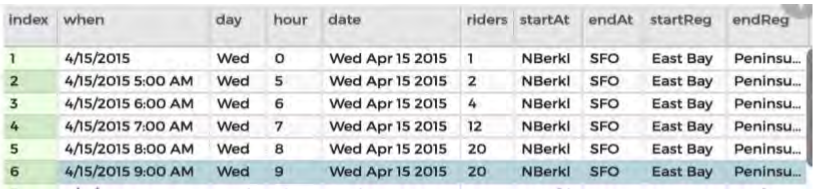
4. Which variables do you think are the best indicators of the overall health of the ecosystem? Based on this, which year would you argue was the park the healthiest? Least healthy? [*question intended to prompt focusing, grouping, graphing*]
5. In what year was there the least competition between wolves and bears for prey (bison and elk)? [*question intended to prompt focusing, calculation*]

Bay Area Rapid Transit Ridership Data (April – Sept 2015)

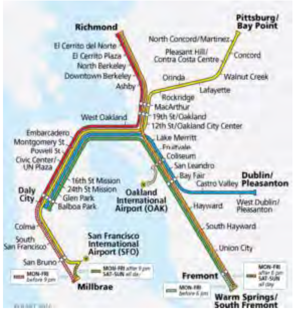
This data set includes transit rider information for the BART system in the Bay Area. This data was collected in 2015, between April 4, 2015 and September 30, 2015 and includes approximately 120 million BART rides. That's too much data to see all at once, so you need to request it using this interface here. Just as an example, I'm going to request data that shows rides from the North Berkeley station to SFO for the week of April 15.

Each row in this table represents one hour at a station, showing how many people entered (startAt) and exited (endAt) a specific pair of stations during that hour. There is a separate record from each of the other stations in the BART system. For each record (index), you can see the date and time it happened (when), the day of the week (day), the one hour time window out of 24 hours when it happened (hour), the date (date), the number of riders during that hour (riders), the station where the passenger entered the station (startAT), the station where the passenger left the system (endAt), and the broader region where travel started and ended (startReg, endReg).

To work with the data, you must first select a start date, then how much data you'd like to explore from that start date (1 day, 7 days, 30 days), and what station in the BART system you'd like to depart from and arrive at. This is an example of how the data is organized in CODAP:



index	when	day	hour	date	riders	startAt	endAt	startReg	endReg
1	4/15/2015	Wed	0	Wed Apr 15 2015	1	NBerkl	SFO	East Bay	Peninsu...
2	4/15/2015 5:00 AM	Wed	5	Wed Apr 15 2015	2	NBerkl	SFO	East Bay	Peninsu...
3	4/15/2015 6:00 AM	Wed	6	Wed Apr 15 2015	4	NBerkl	SFO	East Bay	Peninsu...
4	4/15/2015 7:00 AM	Wed	7	Wed Apr 15 2015	12	NBerkl	SFO	East Bay	Peninsu...
5	4/15/2015 8:00 AM	Wed	8	Wed Apr 15 2015	20	NBerkl	SFO	East Bay	Peninsu...
6	4/15/2015 9:00 AM	Wed	9	Wed Apr 15 2015	20	NBerkl	SFO	East Bay	Peninsu...

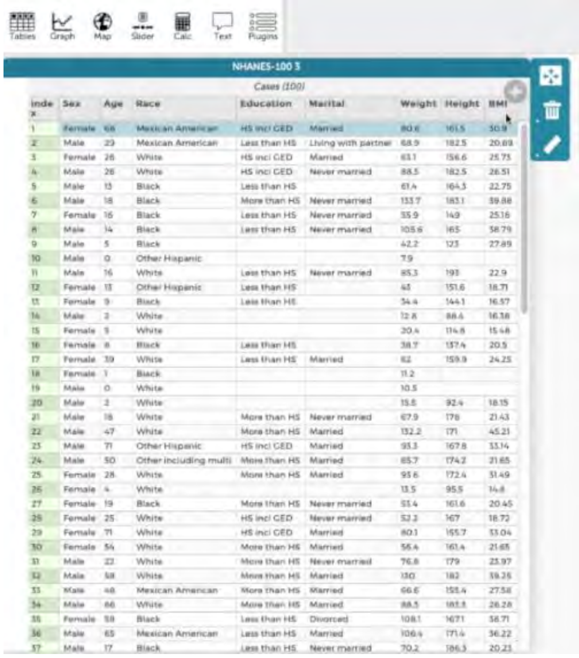


1. Is there a connection between time of day and ridership numbers? Share two examples that help you explain your thinking! [*question intended to prompt filtering, graphing, calculation*]
2. What is the busiest time of day to travel from North Berkeley Station to SFO Airport Station? [*question intended to prompt filtering, graphing, calculation*]
3. For travelers leaving SFO airport, what are some of the most popular and less popular destinations? [*question intended to prompt filtering, calculation*]
4. Which 3 stations show the biggest change between rush hour and non-rush hour ridership? [*question intended to prompt filtering, graphing, calculation*]
5. If you were in advertising and wanted to place your advertisement in the BART system where the most people would walk by, what station(s) would you choose? [*question intended to prompt filtering, calculating*]

Appendix C.

Transcript of the CODAP instructional video presented to students at the beginning of Interview 1 and available to students upon request throughout interview series. Note that we do not use the technical names of data moves (e.g., "filter") throughout. Instead, we use language that we have found to be more accessible to novice CODAP users.

This is a dataset from NHANES, or the National Health and Nutrition Examination Survey. This dataset has 100 records. Each record includes a person's self-identified sex, age, race, education level, marital status, weight, height, and BMI. As you can see, some data are missing for some people.



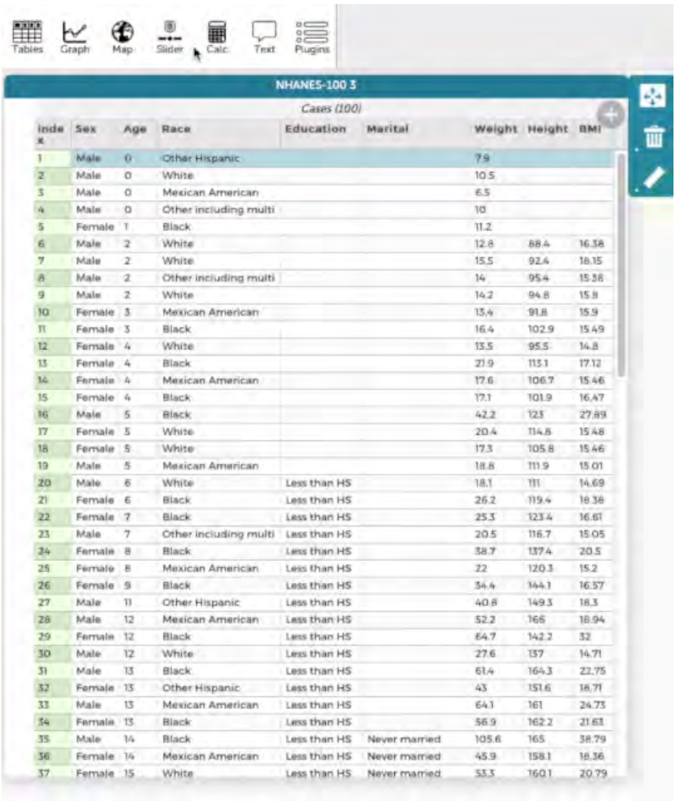
NHANES-100 3

Cases (100)

Index	Sex	Age	Race	Education	Marital	Weight	Height	BMI
1	Female	68	Mexican American	HS incl GED	Married	80.6	161.5	30.9
2	Male	29	Mexican American	Less than HS	Living with partner	68.9	182.5	20.69
3	Female	26	White	HS incl GED	Married	63.1	156.6	25.73
4	Male	26	White	HS incl GED	Never married	88.5	182.5	26.51
5	Male	13	Black	Less than HS		61.4	164.3	22.75
6	Male	18	Black	More than HS	Never married	133.7	183.1	39.88
7	Female	16	Black	Less than HS	Never married	55.9	149	25.16
8	Male	14	Black	Less than HS	Never married	105.6	165	38.79
9	Male	5	Black			42.2	123	27.89
10	Male	0	Other Hispanic			7.9		
11	Male	16	White	Less than HS	Never married	85.3	193	22.9
12	Female	13	Other Hispanic	Less than HS		43	151.6	18.71
13	Female	9	Black	Less than HS		34.4	144.1	16.57
14	Male	2	White			12.8	88.4	16.38
15	Female	5	White			20.4	114.8	15.48
16	Female	8	Black	Less than HS		38.7	137.4	20.5
17	Female	39	White	Less than HS	Married	62	159.9	24.25
18	Female	1	Black			11.2		
19	Male	0	White			10.5		
20	Male	2	White			15.5	92.4	18.15
21	Male	18	White	More than HS	Never married	67.9	178	21.43
22	Male	47	White	More than HS	Married	132.2	171	45.21
23	Male	71	Other Hispanic	HS incl GED	Married	93.3	167.8	33.14
24	Male	50	Other including multi	More than HS	Married	65.7	174.2	21.65
25	Female	28	White	More than HS	Married	93.6	172.4	31.49
26	Female	4	White			13.5	95.5	14.8
27	Female	19	Black	More than HS	Never married	53.4	161.6	20.45
28	Female	25	White	HS incl GED	Never married	52.2	167	18.72
29	Female	71	White	HS incl GED	Married	80.1	155.7	33.04
30	Female	54	White	More than HS	Married	56.4	161.4	21.65
31	Male	22	White	More than HS	Never married	76.8	179	23.97
32	Male	58	White	More than HS	Married	130	182	39.25
33	Male	48	Mexican American	More than HS	Married	66.6	155.4	27.58
34	Male	66	White	More than HS	Married	88.5	183.3	26.28
35	Female	58	Black	Less than HS	Divorced	108.1	167.1	38.71
36	Male	65	Mexican American	Less than HS	Married	106.4	171.4	36.22
37	Male	17	Black	Less than HS	Never married	70.2	186.3	20.23

This video will show you some of the things you can do with CODAP, to make answering questions using data a bit easier. You can organize data in the table however you'd like.

Here, we're sorting the data by age. Oh look! The data includes infants and children.

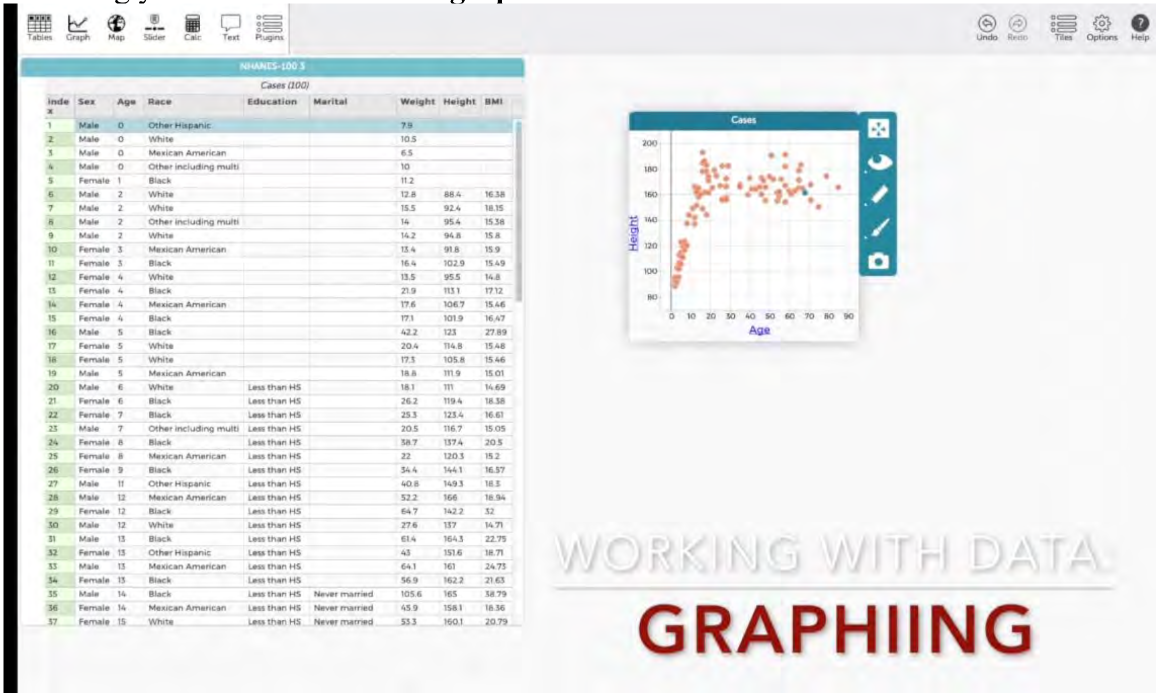


NHANES-100 3

Cases (100)

Index	Sex	Age	Race	Education	Marital	Weight	Height	BMI
1	Male	0	Other Hispanic			7.9		
2	Male	0	White			10.5		
3	Male	0	Mexican American			6.5		
4	Male	0	Other including multi			10		
5	Female	1	Black			11.2		
6	Male	2	White			12.8	88.4	16.38
7	Male	2	White			15.5	92.4	18.15
8	Male	2	Other including multi			14	95.4	15.38
9	Male	2	White			14.2	94.8	15.8
10	Female	3	Mexican American			13.4	91.8	15.9
11	Female	3	Black			16.4	102.9	15.49
12	Female	4	White			13.5	95.5	14.8
13	Female	4	Black			21.9	113.1	17.12
14	Female	4	Mexican American			17.6	106.7	15.46
15	Female	4	Black			17.1	101.9	16.47
16	Male	5	Black			42.2	123	27.89
17	Female	5	White			20.4	114.8	15.48
18	Female	5	White			17.3	105.8	15.46
19	Male	5	Mexican American			18.8	111.9	15.01
20	Male	6	White	Less than HS		18.1	111	14.69
21	Female	6	Black	Less than HS		26.2	119.4	18.38
22	Female	7	Black	Less than HS		25.3	123.4	16.61
23	Male	7	Other including multi	Less than HS		20.5	116.7	15.05
24	Female	8	Black	Less than HS		38.7	137.4	20.5
25	Female	8	Mexican American	Less than HS		22	120.3	15.2
26	Female	9	Black	Less than HS		34.4	144.1	16.57
27	Male	11	Other Hispanic	Less than HS		40.8	149.3	18.3
28	Male	12	Mexican American	Less than HS		52.2	166	18.94
29	Female	12	Black	Less than HS		64.7	142.2	32
30	Male	12	White	Less than HS		27.6	137	14.71
31	Male	13	Black	Less than HS		61.4	164.3	22.75
32	Female	13	Other Hispanic	Less than HS		43	151.6	18.71
33	Male	13	Mexican American	Less than HS		64.1	161	24.73
34	Female	13	Black	Less than HS		56.9	162.2	21.63
35	Male	14	Black	Less than HS	Never married	105.6	165	38.79
36	Female	14	Mexican American	Less than HS	Never married	45.9	158.1	18.36
37	Female	15	White	Less than HS	Never married	53.3	160.1	20.79

Let’s say you want to explore how the heights of people in this dataset are related to age. One thing you could do is create a **graph**.

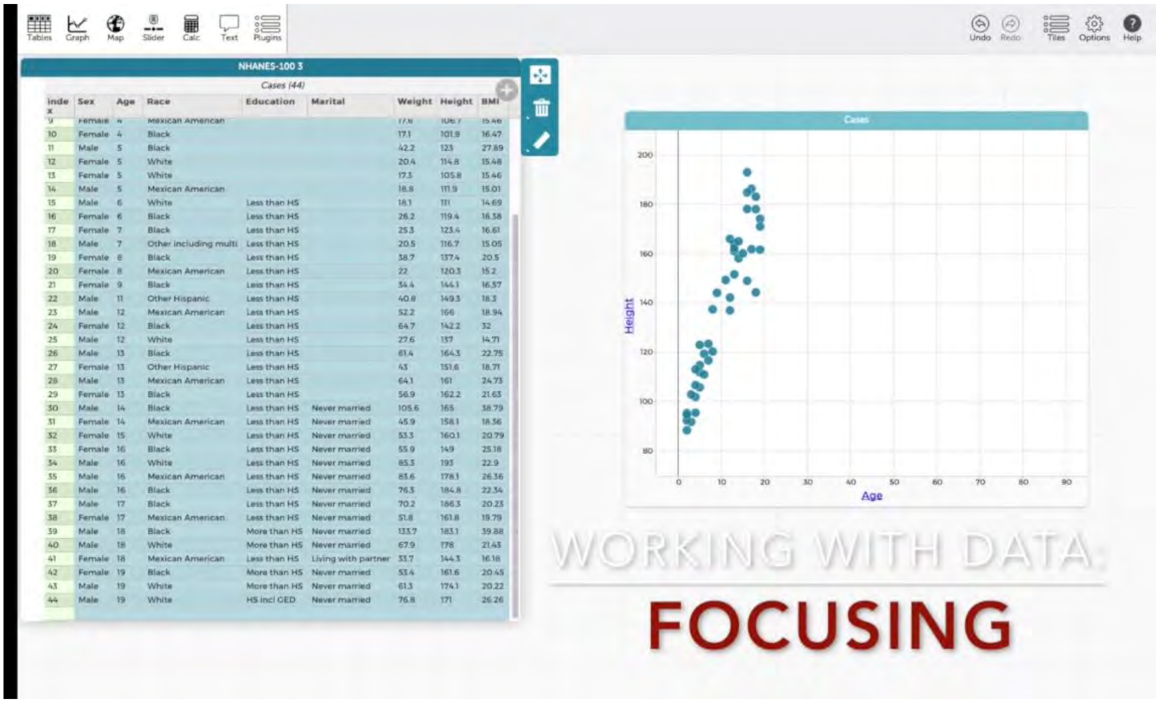


NHANES-100 3

Cases (100)

Index	Sex	Age	Race	Education	Marital	Weight	Height	BMI
1	Male	0	Other Hispanic			7.9		
2	Male	0	White			10.5		
3	Male	0	Mexican American			6.5		
4	Male	0	Other including multi			10		
5	Female	1	Black			11.2		
6	Male	2	White			12.8	88.4	16.38
7	Male	2	White			15.5	92.4	18.15
8	Male	2	Other including multi			14	95.4	15.38
9	Male	2	White			14.2	94.8	15.8
10	Female	3	Mexican American			13.4	91.8	15.9
11	Female	3	Black			16.4	102.9	15.49
12	Female	4	White			13.5	95.5	14.8
13	Female	4	Black			21.9	113.1	17.12
14	Female	4	Mexican American			17.6	106.7	15.46
15	Female	4	Black			17.1	101.9	16.47
16	Male	5	Black			42.2	123	27.89
17	Female	5	White			20.4	114.8	15.48
18	Female	5	White			17.3	105.8	15.46
19	Male	5	Mexican American			18.8	111.9	15.01
20	Male	6	White	Less than HS		18.1	111	14.69
21	Female	6	Black	Less than HS		26.2	119.4	18.38
22	Female	7	Black	Less than HS		25.3	123.4	16.61
23	Male	7	Other including multi	Less than HS		20.5	116.7	15.05
24	Female	8	Black	Less than HS		38.7	137.4	20.5
25	Female	8	Mexican American	Less than HS		22	120.3	15.2
26	Female	9	Black	Less than HS		34.4	144.1	16.57
27	Male	11	Other Hispanic	Less than HS		40.8	149.3	18.3
28	Male	12	Mexican American	Less than HS		52.2	166	18.94
29	Female	12	Black	Less than HS		64.7	142.2	32
30	Male	12	White	Less than HS		27.6	137	14.71
31	Male	13	Black	Less than HS		61.4	164.3	22.75
32	Female	13	Other Hispanic	Less than HS		43	151.6	18.71
33	Male	13	Mexican American	Less than HS		64.1	161	24.73
34	Female	13	Black	Less than HS		56.9	162.2	21.63
35	Male	14	Black	Less than HS	Never married	105.6	165	38.79
36	Female	14	Mexican American	Less than HS	Never married	45.9	158.1	18.36
37	Female	15	White	Less than HS	Never married	53.3	160.1	20.79

Here we see, as you might have expected, heights change dramatically between people at different ages as children and teens. I’m going to **focus** on just kids and teens.

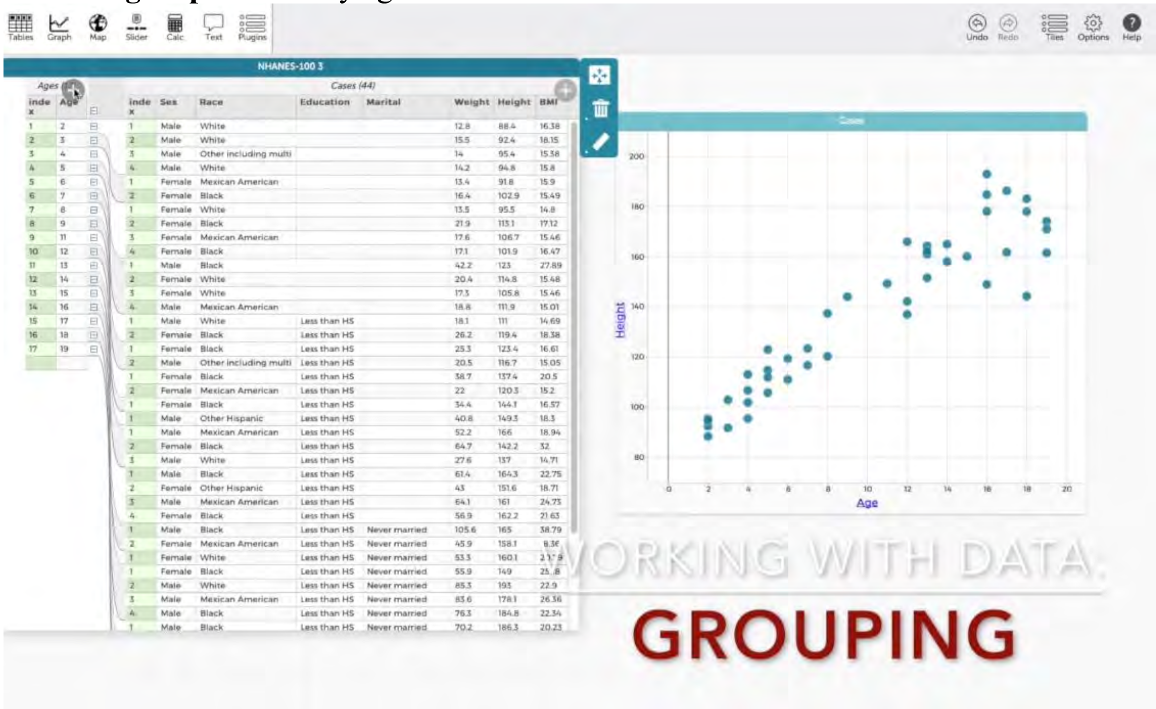


NHANES-100 3

Cases (44)

index	Sex	Age	Race	Education	Marital	Weight	Height	BMI
9	Female	4	Mexican American			17.6	106.7	15.46
10	Female	4	Black			17.1	101.9	16.47
11	Male	5	Black			42.2	123	27.89
12	Female	5	White			20.4	114.8	15.48
13	Female	5	White			17.3	105.8	15.46
14	Male	5	Mexican American			18.8	111.9	15.01
15	Male	6	White	Less than HS		18.1	111	14.69
16	Female	6	Black	Less than HS		26.2	119.4	18.38
17	Female	7	Black	Less than HS		25.3	123.4	16.61
18	Male	7	Other including multi	Less than HS		20.5	116.7	15.05
19	Female	8	Black	Less than HS		38.7	137.4	20.5
20	Female	8	Mexican American	Less than HS		22	120.3	15.2
21	Female	9	Black	Less than HS		34.4	144.1	16.57
22	Male	11	Other Hispanic	Less than HS		40.8	149.3	18.3
23	Male	12	Mexican American	Less than HS		52.2	166	18.94
24	Female	12	Black	Less than HS		64.7	142.2	32
25	Male	12	White	Less than HS		27.6	137	14.71
26	Male	13	Black	Less than HS		61.4	164.3	22.75
27	Female	13	Other Hispanic	Less than HS		43	151.6	18.71
28	Male	13	Mexican American	Less than HS		64.1	161	24.73
29	Female	13	Black	Less than HS		56.9	162.2	21.63
30	Male	14	Black	Less than HS	Never married	105.6	165	38.79
31	Female	14	Mexican American	Less than HS	Never married	45.9	158.1	18.36
32	Female	15	White	Less than HS	Never married	53.3	160.1	20.79
33	Female	16	Black	Less than HS	Never married	55.9	149	25.18
34	Male	16	White	Less than HS	Never married	85.3	193	22.9
35	Male	16	Mexican American	Less than HS	Never married	83.6	178.1	26.36
36	Male	16	Black	Less than HS	Never married	76.3	184.8	22.34
37	Male	17	Black	Less than HS	Never married	70.2	186.3	20.23
38	Female	17	Mexican American	Less than HS	Never married	51.8	161.8	19.79
39	Male	18	Black	More than HS	Never married	133.7	183.1	39.88
40	Male	18	White	More than HS	Never married	67.9	178	21.43
41	Female	18	Mexican American	Less than HS	Living with partner	33.7	144.3	16.18
42	Female	19	Black	More than HS	Never married	53.4	161.6	20.45
43	Male	19	White	More than HS	Never married	61.3	174.1	20.22
44	Male	19	White	HS incl GED	Never married	76.8	171	26.26

Now, let’s say you’re interested in when children seem to grow the fastest. One thing we can do is **group** the data by age.



NHANES-100 3

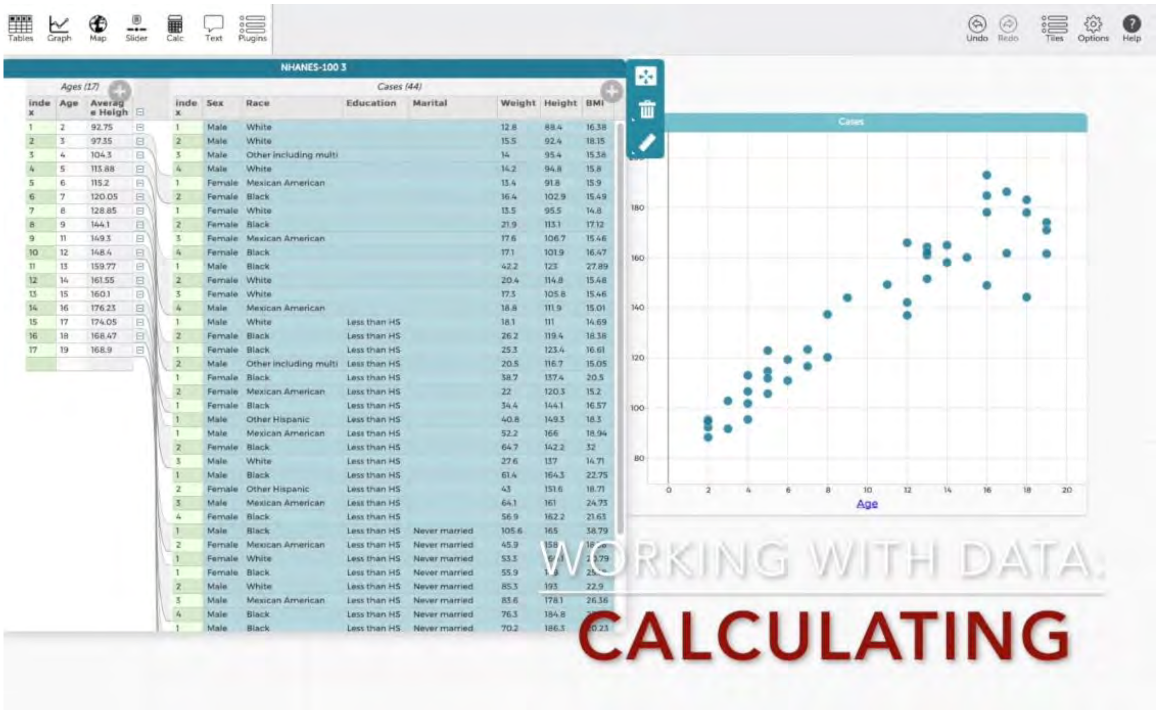
Ages (17)

index	Age
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
9	11
10	12
11	13
12	14
13	15
14	16
15	17
16	18
17	19

Cases (44)

index	Sex	Race	Education	Marital	Weight	Height	BMI
1	Male	White			12.8	88.4	16.38
2	Male	White			15.5	92.4	18.15
3	Male	Other including multi			14	95.4	15.38
4	Male	White			14.2	94.8	15.8
1	Female	Mexican American			13.4	91.8	15.9
2	Female	Black			16.4	102.9	15.49
1	Female	White			13.5	95.5	14.8
2	Female	Black			21.9	113.1	17.12
3	Female	Mexican American			17.6	106.7	15.46
4	Female	Black			17.1	101.9	16.47
1	Male	Black			42.2	123	27.89
2	Female	White			20.4	114.8	15.48
3	Female	White			17.3	105.8	15.46
4	Male	Mexican American			18.8	111.9	15.01
1	Male	White	Less than HS		18.1	111	14.69
2	Female	Black	Less than HS		26.2	119.4	18.38
1	Female	Black	Less than HS		25.3	123.4	16.61
2	Male	Other including multi	Less than HS		20.5	116.7	15.05
1	Female	Black	Less than HS		38.7	137.4	20.5
2	Female	Mexican American	Less than HS		22	120.3	15.2
1	Female	Black	Less than HS		34.4	144.1	16.57
1	Male	Other Hispanic	Less than HS		40.8	149.3	18.3
1	Male	Mexican American	Less than HS		52.2	166	18.94
2	Female	Black	Less than HS		64.7	142.2	32
3	Male	White	Less than HS		27.6	137	14.71
1	Male	Black	Less than HS		61.4	164.3	22.75
2	Female	Other Hispanic	Less than HS		43	151.6	18.71
3	Male	Mexican American	Less than HS		64.1	161	24.73
4	Female	Black	Less than HS		56.9	162.2	21.63
1	Male	Black	Less than HS	Never married	105.6	165	38.79
2	Female	Mexican American	Less than HS	Never married	45.9	158.1	18.36
1	Female	White	Less than HS	Never married	53.3	160.1	20.79
1	Female	Black	Less than HS	Never married	55.9	149	25.18
2	Male	White	Less than HS	Never married	85.3	193	22.9
3	Male	Mexican American	Less than HS	Never married	83.6	178.1	26.36
4	Male	Black	Less than HS	Never married	76.3	184.8	22.34
1	Male	Black	Less than HS	Never married	70.2	186.3	20.23

Now, we can **calculate** the average height of people for each age.



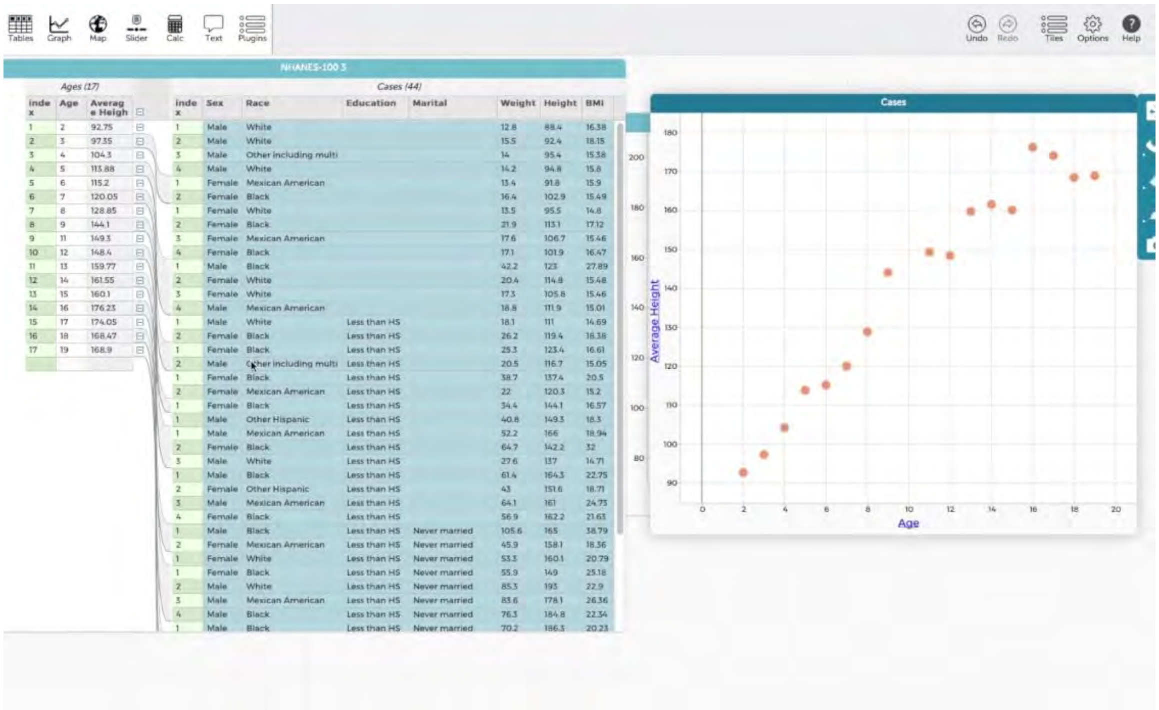
Ages (17)

index	Age	Average Heigh
1	2	92.75
2	3	97.35
3	4	104.3
4	5	113.88
5	6	115.2
6	7	120.05
7	8	128.85
8	9	144.1
9	11	149.3
10	12	148.4
11	13	159.77
12	14	161.55
13	15	160.1
14	16	176.23
15	17	174.05
16	18	168.47
17	19	168.9

Cases (44)

index	Sex	Race	Education	Marital	Weight	Height	BMI
1	Male	White			12.8	88.4	16.38
2	Male	White			15.5	92.4	18.15
3	Male	Other including multi			14	95.4	15.38
4	Male	White			14.2	94.8	15.8
1	Female	Mexican American			13.4	91.8	15.9
2	Female	Black			16.4	102.9	15.49
1	Female	White			13.5	95.5	14.8
2	Female	Black			21.9	113.1	17.12
3	Female	Mexican American			17.6	106.7	15.46
4	Female	Black			17.1	101.9	16.47
1	Male	Black			42.2	123	27.89
2	Female	White			20.4	114.8	15.48
3	Female	White			17.3	105.8	15.46
4	Male	Mexican American			18.8	111.9	15.01
1	Male	White	Less than HS		18.1	111	14.69
2	Female	Black	Less than HS		26.2	119.4	18.38
1	Female	Black	Less than HS		25.3	123.4	16.61
2	Male	Other including multi	Less than HS		20.5	116.7	15.05
1	Female	Black	Less than HS		38.7	137.4	20.5
2	Female	Mexican American	Less than HS		22	120.3	15.2
1	Female	Black	Less than HS		34.4	144.1	16.57
1	Male	Other Hispanic	Less than HS		40.8	149.3	18.3
1	Male	Mexican American	Less than HS		52.2	166	18.94
2	Female	Black	Less than HS		64.7	142.2	32
3	Male	White	Less than HS		27.6	137	14.71
1	Male	Black	Less than HS		61.4	164.3	22.75
2	Female	Other Hispanic	Less than HS		43	151.6	18.71
3	Male	Mexican American	Less than HS		64.1	161	24.73
4	Female	Black	Less than HS		56.9	162.2	21.63
1	Male	Black	Less than HS	Never married	105.6	165	38.79
2	Female	Mexican American	Less than HS	Never married	45.9	158.1	18.36
1	Female	White	Less than HS	Never married	53.3	160.1	20.79
1	Female	Black	Less than HS	Never married	55.9	149	25.18
2	Male	White	Less than HS	Never married	85.3	193	22.9
3	Male	Mexican American	Less than HS	Never married	83.6	178.1	26.36
4	Male	Black	Less than HS	Never married	76.3	184.8	22.34
1	Male	Black	Less than HS	Never married	70.2	186.3	20.23

In the **graph** of our new calculation, it looks like kids in our data grew fastest at age 8 and 15.



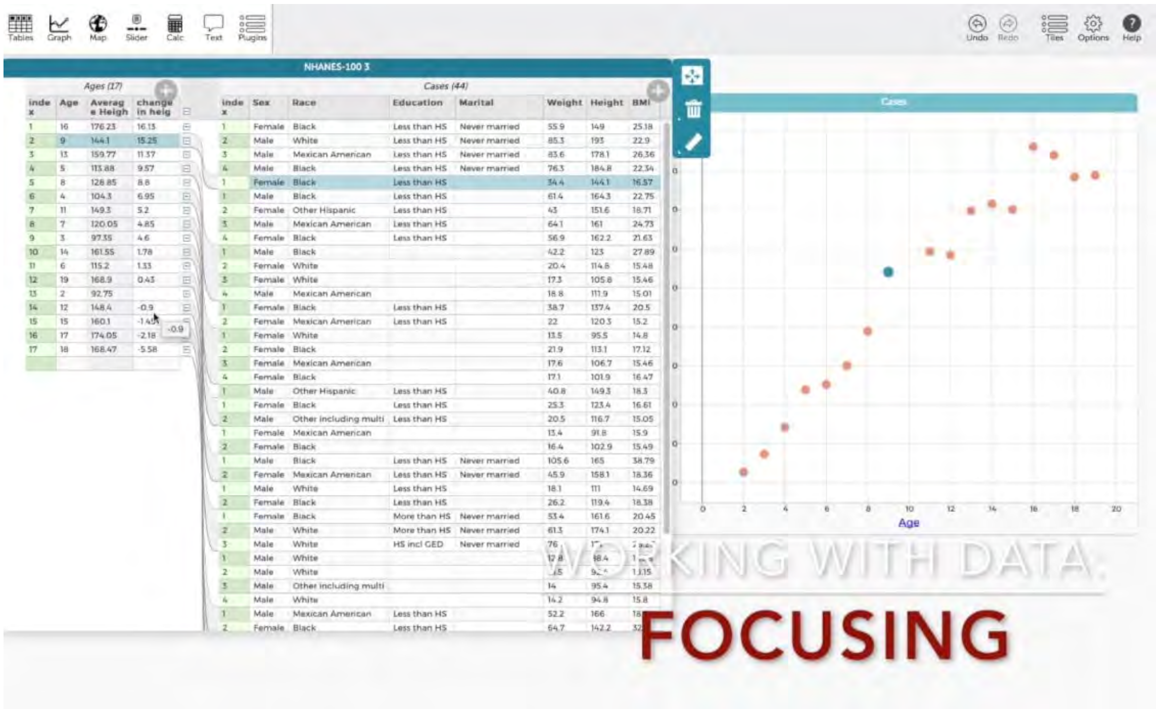
Ages (17)

index	Age	Average Heigh
1	2	92.75
2	3	97.35
3	4	104.3
4	5	113.88
5	6	115.2
6	7	120.05
7	8	128.85
8	9	144.1
9	11	149.3
10	12	148.4
11	13	159.77
12	14	161.55
13	15	160.1
14	16	176.23
15	17	174.05
16	18	168.47
17	19	168.9

Cases (44)

index	Sex	Race	Education	Marital	Weight	Height	BMI
1	Male	White			12.8	88.4	16.38
2	Male	White			15.5	92.4	18.15
3	Male	Other including multi			14	95.4	15.38
4	Male	White			14.2	94.8	15.8
1	Female	Mexican American			13.4	91.8	15.9
2	Female	Black			16.4	102.9	15.49
1	Female	White			13.5	95.5	14.8
2	Female	Black			21.9	113.1	17.12
3	Female	Mexican American			17.6	106.7	15.46
4	Female	Black			17.1	101.9	16.47
1	Male	Black			42.2	123	27.89
2	Female	White			20.4	114.8	15.48
3	Female	White			17.3	105.8	15.46
4	Male	Mexican American			18.8	111.9	15.01
1	Male	White	Less than HS		18.1	111	14.69
2	Female	Black	Less than HS		26.2	119.4	18.38
1	Female	Black	Less than HS		25.3	123.4	16.61
2	Male	Other including multi	Less than HS		20.5	116.7	15.05
1	Female	Black	Less than HS		38.7	137.4	20.5
2	Female	Mexican American	Less than HS		22	120.3	15.2
1	Female	Black	Less than HS		34.4	144.1	16.57
1	Male	Other Hispanic	Less than HS		40.8	149.3	18.3
1	Male	Mexican American	Less than HS		52.2	166	18.94
2	Female	Black	Less than HS		64.7	142.2	32
3	Male	White	Less than HS		27.6	137	14.71
1	Male	Black	Less than HS		61.4	164.3	22.75
2	Female	Other Hispanic	Less than HS		43	151.6	18.71
3	Male	Mexican American	Less than HS		64.1	161	24.73
4	Female	Black	Less than HS		56.9	162.2	21.63
1	Male	Black	Less than HS	Never married	105.6	165	38.79
2	Female	Mexican American	Less than HS	Never married	45.9	158.1	18.36
1	Female	White	Less than HS	Never married	53.3	160.1	20.79
1	Female	Black	Less than HS	Never married	55.9	149	25.18
2	Male	White	Less than HS	Never married	85.3	193	22.9
3	Male	Mexican American	Less than HS	Never married	83.6	178.1	26.36
4	Male	Black	Less than HS	Never married	76.3	184.8	22.34
1	Male	Black	Less than HS	Never married	70.2	186.3	20.23

But, it might be helpful to get real numbers. Here, I'm **calculating** the difference between the average heights for each year in age. This shows me how much larger the average height of people who are, for example, 17 are than the average height of people who are 16. If I sort by the differences in height per year, then I see that there are the highest differences in height for 16 year olds and for 9 year olds, which is what I expected from the graph.



NHANES-100 3

Ages (17)

index	Age	Average Height	change in height
1	16	176.23	16.13
2	9	144.1	15.25
3	13	159.77	11.37
4	5	113.88	9.57
5	8	128.85	8.8
6	4	104.3	6.95
7	11	149.3	5.2
8	7	120.05	4.85
9	3	97.35	4.6
10	14	161.55	1.78
11	6	115.2	1.33
12	19	168.9	0.43
13	2	92.75	
14	12	148.4	-0.9
15	15	160.1	-1.45
16	17	174.05	-2.18
17	18	168.47	-5.58

Cases (44)

index	Sex	Race	Education	Marital	Weight	Height	BMI
1	Female	Black	Less than HS	Never married	55.9	149	25.18
2	Male	White	Less than HS	Never married	85.3	193	22.9
3	Male	Mexican American	Less than HS	Never married	83.6	178.1	26.36
4	Male	Black	Less than HS	Never married	76.3	184.8	22.34
1	Female	Black	Less than HS		34.4	144.1	16.57
1	Male	Black	Less than HS		61.4	164.3	22.75
2	Female	Other Hispanic	Less than HS		43	151.6	18.71
3	Male	Mexican American	Less than HS		64.1	161	24.73
4	Female	Black	Less than HS		56.9	162.2	21.63
1	Male	Black			42.2	123	27.89
2	Female	White			20.4	114.8	15.48
3	Female	White			17.3	105.8	15.46
4	Male	Mexican American			18.8	111.9	15.01
1	Female	Black	Less than HS		38.7	137.4	20.5
2	Female	Mexican American	Less than HS		22	120.3	15.2
1	Female	White			13.5	95.5	14.8
2	Female	Black			21.9	113.1	17.12
3	Female	Mexican American			17.6	106.7	15.46
4	Female	Black			17.1	101.9	16.47
1	Male	Other Hispanic	Less than HS		40.8	149.3	18.3
1	Female	Black	Less than HS		25.3	123.4	16.61
2	Male	Other including multi	Less than HS		20.5	116.7	15.05
1	Female	Mexican American			13.4	91.8	15.9
2	Female	Black			16.4	102.9	15.49
1	Male	Black	Less than HS	Never married	105.6	165	38.79
2	Female	Mexican American	Less than HS	Never married	45.9	158.1	18.36
1	Male	White	Less than HS		18.1	111	14.69
2	Female	Black	Less than HS		26.2	119.4	18.38
1	Female	Black	More than HS	Never married	53.4	161.6	20.45
2	Male	White	More than HS	Never married	61.3	174.1	20.22
3	Male	White	HS incl GED	Never married	76	[illegible]	[illegible]
1	Male	White			12.8	[illegible]	[illegible]
2	Male	White			[illegible]	[illegible]	[illegible]
3	Male	Other including multi			14	95.4	15.38
4	Male	White			14.2	94.8	15.8
1	Male	Mexican American	Less than HS		52.2	166	[illegible]
2	Female	Black	Less than HS		64.7	142.2	[illegible]

WORKING WITH DATA:

FOCUSING

But, there's also something funny here: a lot of the numbers are negative. That makes me suspicious, I wouldn't want to say based on this data that young people shrink! Of course, these records are all for different people, so maybe it's not fair to talk about growth using this dataset after all.

We hope this video has helped show some of the things you can do with CODAP! You don't need to do all or even any of these things - we just want you to know about them in case they are helpful for your investigation. We know that learning these things take time, so remember your interviewer will always be around to help. We've also given you a sheet with keywords **- graph**, **focus**, **color**, **group**, and **calculate** - so that you can tell us exactly what it is you'd like to do, and to keep notes as you learn.