Awash in Data

Tim Erickson

3/4/23

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

2 Lessons

3 Lessons Overview

> Big overview: First, students learn CODAP basics in a brief introductory lesson. Next, you (the teacher) demonstrate additional CODAP skills using a dataset about age and height. Then students practice those skills, and expand on them, using a more engaging and complex dataset about gender, race, education, and income. After two cycles of this (i.e., two sets of increasingly-complex skills) students do a *short* project on a topic they choose.

The first "Part" of this book—the next few chapters—contains lessons and assignments for a quick introduction to data science. I created these lessons for students in a high school, as part of a class called "Applied Mathematics," which includes a number of topics including consumer-ish math. This introduction occupied a couple of weeks of class time, roughly four seventy-five minute blocks. Although the students varied widely in their level of preparation, they were generally polite, attentive, and well-intentioned.

We ran this most recently in the Spring of 2020, under quarantine, when the class was "virtual," run over Zoom. The lessons here roughly parallel what we did in class that time, when students were isolated in their homes and groupwork was possible, but harder than "before." The discussion suggestions in the lessons work fine, but not as well as in an in-person classroom. Therefore I think these lessons might also work, with sensible modifications, for self-study. For example, you will not be turning in your projects—unless you *want* to send them to me; I'd love to see them!

Likewise, if you are a teacher with a class of your own, of course you should modify what you see here so that it makes sense for your students and your situation. For example, our school uses Google apps extensively, so students all know about setting permissions and submitting links to Google docs via email. In previous years, I had insisted that students make **pdf**s and email those in. What works for you depends on your situation.

You may want to modify assignments, add new ones, skip others. Go for it. I will try to describe, here and in the commentary that accompanies each element in this Part, my rationale for the order and content I chose.

Class preparation: We (the teacher-of-record and I) set up a Google doc as a "landing page" for the class. It had material that students would need: links to class-specific instructions, links to the CODAP software and to this book, as well as links to various data sets. Before each class, we edited the doc so that today's most important links were at the top, with an

agenda of what we would do. That way, all past links and agendas were still in the document, at the bottom, in reverse chronological order.

For the first lesson, for example, we had a link to this book and to a blank CODAP document, with advice to bookmark the page. We also had a link to an "800 children and teens" document, which we easily got to in the first session.

Those data links are also embedded in this book, so if your students use this book (for ours, this was just a resource), or you are doing self-study, there's no need to make an extra set of links.

3.1 The "getting started" lesson

Here is a link to the lesson chapter

Critical links:

- a blank CODAP document, https://codap.concord.org/app
- this book https://codap.xyz/awash

This introduces CODAP basics, especially how to make graphs. This takes students only a few minutes, and, as the commentary suggests, could even be homework.

In an actual classroom, circulate and make sure that everybody has found the relevant page and can make graphs. Lead a brief discussion as suggested in the commentary.

3.2 800 children and teens, part one

Here is a link to the lesson chapter

Additional critical link:

- age and height data: http://codap.concord.org/app#shared=31797

The dataset has information about 800 USA-ians, aged 5 to 19, from a national health survey (NHANES). The attributes (a.k.a. variables) include age, gender, height, weight, armspan, upper arm length, and pulse. We will eventually focus on age, gender, and height—but we don't tell students that at first.

Eventually, we focus on how height changes with age—and how it also depends on gender. That presents a problem that often makes students feel "awash": they have three attributes but only two axes on a graph. The obvious solutions—multiple graphs, color the points—don't work out very well.

Data move: filtering

We show students a different approach, using a data move: *filtering*. We look at only a small part of the data, just the 10-year-olds (where the girls are taller!). Then it's easy to compare the boys and the girls in a graph, now that age is not a factor.
Getting to the filtering is the critical part of the lesson, and sets up the groupwork described in that chapter: each group does this filtering move for different ages, pooling the class data so students can enter and plot the *mean* ages for each gender at each age. For logistics, we created a shared Google *sheet* where students entered the means.

3.3 A First Assignment

Here is a link to the assignment

- New link—Census data: https://codap.concord.org/app#shared=22176

This assignment introduces a new dataset from the US Census. It has income data for 1000 25-to-44-year-olds, along with gender (of course) but also race, Hispanic status, and educational attainment.

The assignment asks students to explore the data, and then to come up with a *claim* and a graph to address the claim. This should be simple—simple enough that they could conceivably do the data analysis using the live illustration embedded in the assignment.

This assignment is written so it's entirely done in CODAP (no Google doc at this point), so it can be done and turned in in class (although some students might not finish). It could even be done orally, though we did it as an assignment to turn in.

Many students will use this dataset to compare incomes, and that will often lead into social justice issues. We love that kind of topic, and students are often intrigued and motivated that they are exploring it with real data about real people.

3.4 800 children and teens, part two

Here is a link to the lesson chapter

Data moves: grouping and summarizing

This chapter introduces *grouping* cases in the dataset by reorganizing the table and *summarizing* those groups by calculating summary (or *aggregate*) values, in this case, mean heights. These are our second and third core data moves. Conceptually, it's the

most subtle—and powerful—part of this intro to data science. Your goal, as a teacher, is to get as many students as possible to understand this.

The session may begin with the students entering the mean heights by gender and age that they worked on in groups. They can then make the summary graph. If they made this graph already (e.g., in the previous session), remind them of it.

> Tip
>
> Remind students that that process used the *filtering* data move.

Now students will learn how to have the computer do all that. This is perfect for a demo. Get their close attention and go through it *slowly*, sometimes backing up as students have questions. There are two main parts: making the groups; and then making the aggregate calculation. The text in the chapter has all sorts of questions embedded in it; use them to guide how you show students this technique.

You will probably end the session debriefing the first assignment (with the Census data) and previewing the second one.

3.5 A Second Assignment

Here is a link to the assignment

This is a more complex version of that first assignment. Notice these differences:

- Logistical issues:
  - It's supposed to be a Google doc, not just a CODAP doc. (If you want a different format, go ahead!)
  - Students put a link to their CODAP doc in the Google doc.
  - They need to learn how to get a graph into the Google doc.
  - If possible, they should use that grouping move they just saw, and, in addition, calculate summary values for the groups.
- They need to enhance their investigation and give it more nuance. We refer to this in the text as "dig deeper."
- They need to pay more attention to communication, for example in the size and choice of their graphics and the coherence of their narrative.

  Students may be perplexed about what we mean by "dig deeper." An early section in the chapter contains an example of the kind of thing they could do—exploring the obvious issue of income disparity between males and females—and then shows what "digging deeper" might look like.

The other difference is that we restrict the topic: We give them just two claims to choose from. Most students become invested in the Census data. It's engaging, and quickly brings social issues into math class. But they need an alternative; ours is data from BART, the Bay Area Rapid Transit. There are links in the assignment to orient you and those students to the data.

The next class session

In our most recent class, some students "got" this assignment and were done in a flash. For others, the whole grouping-and-aggregation thing was alien and complex. That's not surprising: it *is* conceptually the hardest part of this unit.

So you cannot expect students to have completed even this simple task by the beginning of the next session. We devoted pretty much the whole session to working on this task, giving students help individually and in small groups, and getting them to help each other.

For students who were moving faster, we presented the "small project" task (below) and let them get started. It might be better, however, to give them other useful things to do. In an in-person class, these students could be resources for their peers, or we could have given them other enriching problems to solve. There are some in this book (xxx) that were not available at that time.

3.6 A Small Project

Here is a link to the assignment

This is the capstone project for this unit, and is really just a small extension beyond the second assignment students have just completed.

The key differences are:

- Students pick their own topics.
- They can use the more extensive, national Census data that includes different states and different years.
- For students using BART data, they can play the embedded "secret meeting game."

Some consequences are:

- Students might be more invested in the topics because they chose them themselves.
- They will have ideas for other things they want to do with the data.
  - Some will be too hard, and students may need to find alternatives.
  - Some will be possible, but may not have been covered in class.

Data move: linking

As an example of this last possibility, some students, independently, wanted to connect the Census data to data that wasn't in the data set, that is, to link our data to external data. That's another data move we call *joining*, so we've added a chapter about it even though it's not really part of a simple introduction to data science.
Here is a link to that chapter

4 Summary

In summary, this book has no content whatsoever.

References

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.