## Teaching Notes: Introduction to Data Science Pipeline

### By ***Mary Mulcahy***

*University of Pittsburgh at Bradford*

*Mary.mulcahy@pitt.edu*

**Context for Use**

Department: **Biology Program**

Audience Level: **Undergraduate, Biology Majors or Nonmajors**

Instructional Setting: **Lecture**

Activity Length: **30 minutes or assign as homework.**

**Resource Learning Goals:**

* Define the terms: clean, pull, verify, wrangle, merge, interoperable, file extension, and long and wide format.
* Recognize the difference between long and wide data format.
* Review metadata associated with a downloaded data file.
* Describe the ordered steps of the data pipeline process, but also recognize that the phrase “data science pipeline” means different things to different scientists and that these steps don’t always occur in order.
* Create a graph with appropriate axes with units under guidance.
* Use an r-squared value to assess the strength of a relationship between two variables.
* Explain why statistical analysis may be needed to interpret a data pattern.

**Changes Made (If any)**

My original activity was very similar but focused on canopy height, and it was well-received by students in both a nonmajors and a majors biology course. We did the activity during lecture in a computer room, and that reduced the chances of students encountering technical challenges, but this activity can be completed by students independently as homework. I modified the activity to focus on Yellow Fever because I felt the canopy height activity was overly complicated and cumbersome to complete.

**Teaching Notes**

*This is an introductory activity.*

This is an introductory activity intended for students with very little experience with data science methods.

*What was I hoping for when I created this activity?*

I really wanted to give the big picture of the entire process that a data scientist might use from beginning to end. The reason I wanted the entire process covered is that I am using this activity in preparation for students to explore their own very simple course project, and I felt like students needed to see how each step (finding data. cleaning data, wrangling data, etc.) fit into a bigger picture of the scientific method. I also feel like this activity is a good scaffolding activity to use in a class to give students who may have less data science background a chance to catch up with the terminology without overwhelming them.

*In what ways, if any, did you modify your teaching practice with this activity?*

When I first found the “data pipeline” and the icons found here (<https://schoolofdata.org/methodology/> ), I felt like it gave me the structure that I wanted to use to organize my data-learning activities in a structured and fun way. I wanted students to see the forest for the trees – and to see how data science methods are used to answer an interesting question.

The yellow fever question is rather mediocre, but the dataset is nice and small, and I wanted to present an authentic question that allowed students to the steps of data science within the context of answering that questions, from beginning to end.

I would love to see new versions of this activity focusing on other questions, especially questions that support bringing inclusion, diversity, equity, and access into the curriculum. I would also love to see complementary modules/activities that could be used in sequence to allow students to dive more deeply into individual steps. I know that many such activities already exist.

*Merged CODAP dataset*

For some reason the link to the codap file couldn’t be added to the Qubeshub links area – but the .cvs file is uploaded, the link is found in the student handout, and here’s the link again:

[Shared Merged Data](https://codap.concord.org/app/static/dg/en/cert/index.html#shared=https%3A%2F%2Fcfm-shared.concord.org%2FVAaAqilJumu9hy6txXr2%2Ffile.json)

https://codap.concord.org/app/static/dg/en/cert/index.html#shared=https%3A%2F%2Fcfm-shared.concord.org%2FVAaAqilJumu9hy6txXr2%2Ffile.json

*Answer Key*

The answer key is provided as a file in the resource.