Introductory Data Science Pipeline Activity – Yellow Fever and Global Precipitation

Author: Mary Mulcahy, Associate Professor of Biology, University of Pittsburgh at Bradford, 2024 BEDE faculty mentoring network participant
Date: 2024_05_07
Version: 001

Overview

This activity is an introduction to one version of the **data science pipeline** with the intent of inspiring students to consider their own research questions that could be answered using this process. In this activity, the data science pipeline is defined as a series of steps (start to finish) for using existing data, especially publicly available data, to answer new research questions. The data sources available to students and researchers are wide and varied: a repository or open data source (such as the NEON data portal datasets, World Health Organization Data Collections, Genbank, or ArcGIS Living Atlas Layers, among many others), a crowd-sourced project such as iNaturalist, or data collected by other researchers, often available from an already-published study. Completing most of these data science pipeline steps manually is often impossible because they require many small steps, and the datasets are often too large to handle without help from a computer, programming, and automation. Thus, many data science methods use coding (often R or Python) to enable scientists to accomplish many small tasks on a large dataset efficiently.

This activity introduces only a narrow view of data science. Data science is a much broader field of using modern techniques and technology to work with data and is not restricted to these ordered steps. This activity focuses on using pre-existing datasets, but data scientists work with all types of data and are frequently collecting their own data through carefully designed hands-on research projects. Although this activity oversimplifies the data science pipeline, working through the steps can give a beginner some familiarity with the common terminology and a frequent approach to answering questions.

Learning Objectives

- 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.

What is the Data Science Pipeline?

In this activity, we will treat the Data Science Pipeline as a series of seven steps displayed in the table below, ordered from left to right ("Data Pipeline", School of Data). This definition is deliberately narrow, and it is important to be aware that the data science process can be defined differently.

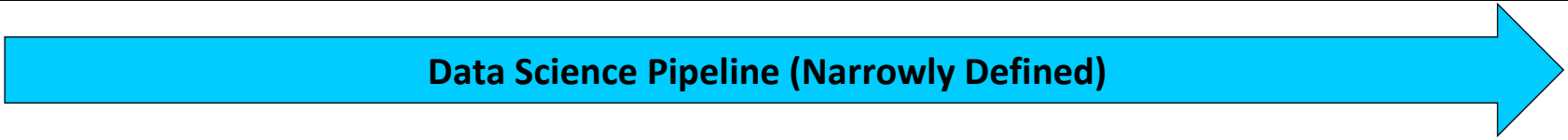
Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
Develop a research question and hypothesis.	Find data to answer your question, often using pre-existing datasets.	Pull, Get, or Download the data.	Verify the Data	Clean and Wrangle the Data.	Analyze/Explore the Data.	Present or share your conclusions
						

Image and Concept Credit: Data Pipeline Images. Creative Commons License CC BY 4.0 - https://schoolofdata.org/methodology/

Please answer the following questions

1. **STEP 1: DEVELOP A RESEARCH QUESTION AND TESTABLE HYPOTHESIS.** In the first step of this narrowly defined data science pipeline, a data scientist comes up with an interesting question to explore. In today's activity, your research hypothesis has been chosen for you, but we hope this activity will be a starting point for you to develop your own data science research questions. Your research question is: "Are the number of cases of yellow fever associated with global average precipitation?". The claim or hypothesis is that yellow fever cases will be strongly positively associated with precipitation.
    a. **How is yellow fever transmitted to a person?** [Your answer to this question will help you to develop a reason for the hypothesis.]

    b. **Given its mode of transmission, what is one <u>reason</u> why yellow fever cases might increase with more precipitation?**

2. **STEP 2: FIND DATA.** You will be using two datasets to answer the question. Please locate these datasets: one dataset on global precipitation and one on the number of yellow fever cases. For the precipitation dataset, locate the dataset using this direct link: https://www.epa.gov/climate-indicators/climate-change-indicators-us-and-global-precipitation - make sure to find the data for the average global precipitation (not the average precipitation for just the United States). At the time these directions were written, the download link was depicted by this symbol: If you have difficulty getting the precipitation data, try this link. For the yellow fever dataset (which WHO calls a "data collection"), locate the data on the World Health Organization (WHO) site: https://www.who.int/. Read the menu tabs at the top of the main WHO page to find your way to the WHO Data Collections. Since yellow fever is preventable by vaccine, it is found in a WHO data collection related to immunizations (you can try searching the data collections with the term "immunization" or just scroll to find it). To restate, you will be following this path to find the data: WHO main page, Data Collections, WHO Immunization Data Portal, Yellow Fever]. Data sites are constantly updated and restructured. You may have to be creative to find the datasets if these directions become outdated. [Data scientists often have to be very creative to find data!]
   Paste the link to the WHO yellow fever dataset below:

3. **STEP 3: GET/PULL THE DATA.**
   Please download the two datasets identified in the previous step. The term get or pull is similar to download, though the term pull often refers to a specific line of code used by programmers to bring a data set into a programming environment. The precipitation and yellow fever files do not have the same file extension. A file extension is the set of letters at the end of the file after a period mark (the "suffix"). The file extension .csv means comma separated values, and one advantage of this file type is because it is "interoperable", able to be imported into and utilized in many different coding and software systems. The file extension .xlsx means that the file is a Microsoft Excel workbook file. One of the advantages of the MS Excel workbook is that it has multiple sheets, but one disadvantage is that it is less interoperable than a .csv file. Fewer programs can open the .xlsx file than .csv.
   View the files you downloaded in the download folder and determine the full names of the raw downloaded files and their extensions. If you can't see the extension, try right clicking on the name of the file, and viewing "properties".

   a. The file name of the raw downloaded precipitation file with extension is: _____

   b. The file name of the raw downloaded yellow fever file with extension is: _____

4. **STEP 4: VERIFY THE DATA.**
   Verifying the data refers to the step where scientists confirm that the data are not corrupted, and that the entire dataset has been downloaded intact. With the small data files that we are using today, verification is as simple as opening the file and confirming that the files contain the data that we expected and that the files open properly.
   a. Open the files and view the data. Do the data appear to be intact?

b. The yellow fever dataset has multiple sheets which can be viewed by clicking on the tabs at the bottom of the open workbook. What are the names of the four tabs?
Tab 1: _____
Tab 2: _____
Tab 3: _____*
Tab 4: _____

***Use the internet to explore and come up with a brief and paraphrased definition of the term "metadata" in data science.**

_____

c. These two files differ in how the data are displayed in the spreadsheet environment. Read a bit about long versus wide format: https://libguides.princeton.edu/R-reshape . Make sure the files are open so that you can look at how the data are arranged. **Which file, the precipitation or yellow fever file is in long format? This is not a trick question – it should be obvious which file is in which format, just by the words "long" and "wide".**

5. **STEP 5: CLEAN AND WRANGLE THE DATA. (And "KEEP RAW DATA RAW")**
Cleaning data refers to activities such as fixing incorrectly formatted data, making sure data are consistently formatted, corrected mistakes in data, dealing with missing data or duplicated data, among other activities. **Wrangling data** might include merging two datasets or changing a dataset into a more usable format, such as converting the dataset from wide to long format. When wrangling data, you might need to group data in a particular way. For example, maybe the data is provided per month, but you really want averages for each year. Before you begin cleaning and wrangling data, it is very important to make a copy of the original dataset in case any problems arise. Creating well-structured folders and informative, and consistent file and folder naming systems to store your data files, metadata, codes, and analyses is critical for success in data science, and should be decided before you begin a project. As you grow in your data science and data handling skills you will learn some of the best practices for organizing data. **Whatever system you decide to use, be sure to leave the raw files alone and make all your changes in your working files only.** Why do you leave the raw files unchanged? They are the backup files in case anything goes wrong. If there is ever any future question of what changes were made, the raw file can be compared to the final cleaned file to confirm that the changes were appropriate.

The cleaning and wrangling step often takes the most time. For this reason, we have already cleaned and wrangled the data and provided it to you in this link: Shared Merged Data

a. Look at the shared data link. In order to merge the two datasets, it was necessary that there be a common variable found in both datasets. **What variable was used for the merger of these two datasets?**

b. **Compare the original files and the newly merged shared dataset. State a couple of ways that the original data has been modified/cleaned/wrangled in the newly merged dataset.**

Curious about file naming conventions? Here are a couple of excellent tips and resources for file naming: https://datamanagement.hms.harvard.edu/plan-design/file-naming-conventions and https://authors.library.caltech.edu/records/mmnpf-cez11 -

6. **STEP 6: ANALYZE OR EXPLORE THE DATA.**

Data scientists frequently explore their data visually. The word "graph" just doesn't encompass all the possible creative ways that data can be summarized, and data scientists typically call any visual summary of data a "visualization". We will use an open science educational tool called Common Online Data Analysis Platform (CODAP) to obtain a visualization of our data very quickly. Open this link to see the data in Codap: Shared Merged Data. Please click on the graph icon and drag or select the variables and make a scatterplot. It is pretty straight-forward to make a scatterplot in Codap, but this video will help if you have trouble: https://serc.carleton.edu/details/files/520956.html We are predicting that yellow fever is dependent on precipitation, and not the other way around (precipitation is not affected by the number of cases of yellow fever). You probably noticed that the precipitation values are actually "anomaly" values that tell the inches of precipitation above or below the average for the entire time period. These precipitation anomaly values are just the number of inches of precipitation above or below the longer-term average precipitation.

Which variable is the independent variable that goes on the X-axis? _____
Which variable is the dependent variable that goes on the Y-axis? _____

It may be tempting to think you see a pattern in the scatterplot. Instead of just informally looking for a pattern, we will use statistics to help assess the presence of a pattern in a quantitative and precise way. To do this, we will construct a linear equation (a "best fit" line) and calculate a statistic called the r-squared value which will give us a way to measure the nature and strength of the relationship between the variables. The r-squared statistic gives us a numerical value for how well the equation of the line (also called the "model") approximates the real data points. If the data points are haphazardly and widely scattered around the best fit line, r-squared value will be close to zero, and this suggests that there is not a strong linear relationship between the X and Y variables. If the data points all fall exactly on the best fit line, the r-squared will be close to 1, and we could conclude that there is a strong linear relationship between the X and Y variables. Visit this simple description to learn more about the statistic and how it is calculated. We can also examine the slope of the linear equation to see if there is a positive or negative association between the variables. You can add the best fit line to your Codap graph by clicking on the ruler icon to the right of the graph and then making sure the "Least Squares Line" is checked.

What was the r-squared value for the best fit line for Precipitation and Yellow Fever? _____
Keeping in mind that r-squared values range from 0 to 1, is the above value high or low? _____

7. **STEP 7: PRESENT OR SHARE THE CONCLUSIONS.**
   a. **Use your analysis to write a conclusion. Did the data analysis (including the r-squared value) support the original hypothesis that yellow fever cases are closely linearly related to global average precipitation? Explain your conclusion.**

   b. **If you could spend more time on this project, how might you refine the research question or project? For example, are there different environmental factors besides global precipitation that should be explored in relation to yellow fever cases? Are there entirely different factors (not environmental) that might be interesting to explore in the context of rates of yellow fever?**

   c. **Propose a different data science question that you think would be interesting to explore (don't worry – you are just proposing a question, and you don't need to work out the details or determine if it is doable).** Be creative! The data science process can be used to ask a myriad of questions: health, environmental, social justice. . .

Congratulations! You just completed the data science pipeline!

Review Questions:

1. What is the first step in the data science pipeline shown in this activity?
   a. Get the data.
   b. Verify the data.
   c. Find the data.
   d. Develop a research question.

2. A scientist has a dataset with a variable called date in several different formats. The scientist corrects all the date values so that they are in the same format (yyyy-mm-dd). This process would be best described as:
   a. verifying data
   b. cleaning data
   c. wrangling data
   d. merging data

3. Before a scientist begins cleaning and wrangling the data, they should:
   a. visualize the data.
   b. analyze the data.
   c. save the raw data.
   d. publish the data.

4. Which of the seven steps in this activity *often* takes the most time?
   a. Get the data.
   b. Verify the data.
   c. Clean/Wrangle the data.
   d. Analyze the data.

5. The best example of something that might be found in the metadata is:
   a. file naming convention and description of data sources
   b. the data values in long format
   c. the data values in wide format
   d. precipitation and yellow fever cases
   e. all of the above

6. Which file type is most interoperable?
   a. .xlsx
   b. .csv
   c. The two file types shown in a and b are equally interoperable.

References:

Common Online Data Analysis Platform "Codap" [Computer software]. (2022). Concord, MA: The Concord Consortium.

Data Pipeline Images. Creative Commons License CC BY 4.0 - https://schoolofdata.org/methodology/

Mulcahy, Mary. (2022). Project EDDIE: Biomes, vegetation structure, and canopy height. https://serc.carleton.edu/eddie/teaching_materials/modules/biomes_vegetation_structure.html

For More Information:

Hamrick, P. N., Aldighieri, S., Machado, G., Leonel, D. G., Vilca, L. M., Uriona, S., & Schneider, M. C. (2017). Geographic patterns and environmental factors associated with human yellow fever presence in the Americas. *PLoS neglected tropical diseases*, *11*(9), e0005897.

File Naming Conventions: https://datamanagement.hms.harvard.edu/plan-design/file-naming-conventions

File Naming Convention Worksheets: https://authors.library.caltech.edu/records/mmnpf-cez11

Navigating Codap for the Biomes Module Video Tutorial - https://serc.carleton.edu/details/files/520956.html - although this video is not designed for this activity, this video provides an overview of how to use codap with a shared data link and how to obtain the r-squared value.