Data Practices: Collecting, Analyzing and Evaluating Data

1. Part 1: Collecting and Structuring Data to answer a driving question

Use the flowchart below to 1) identify a driving question, 2) collect or locate data related to the question, and 3) structure the dataset to be analyzed to inform the question.

**Identify a Driving Question:**
- What do you wonder about this topic?
- What might someone learn from using the data you collect?

**Collect data related to the question:**
- What will you measure to inform your question?
- What tools can you use to collect this data? What are the affordances and limitations of using a computational tool?
- How frequently will you collect data? Are you collecting data on a certain day or during certain times?
- Is the data you are collecting able to be compiled easily (using the same units, avoiding open-ended responses to be able to group)?
- Could there be bias in your data (e.g. human error, missing data or questions that lead to certain responses)?

**Structure the dataset to be analyzed to inform the question**
- How might you structure your dataset to help you or someone else to answer a question related to your findings?
- Is there any data missing?
- How can you provide context or cues (e.g. titles, labels, colors) to help someone else understand your dataset?

**Locate data related to the question:**
- What data that has been collected can help to answer your question?
- What dataset(s) can you compile information from?
- If there is more than one data set being used, is the data comparable?
- Identify cases and attributes in the raw dataset(s). How will you modify or use to answer your question?
- Could there be bias in the instruments used to collect the data (e.g. questions that lead to certain responses)?
In this activity, a student explored existing data about the novel coronavirus (COVID-19). The student followed the algorithm above to:

1. **Identify a driving question:**

The student began by brainstorming questions they had about the coronavirus, such as:

- Where in the world has the highest number of cases?
- How many cases have been diagnosed in the US?

2. **Locate data related to the question:**

The student identified a dataset related to their COVID-19 questions from Our World in Data. The raw data was downloaded from the Our World in Data website. Then, it was uploaded into a Google Sheet (see Raw Data Sheet 1).

3. **Structure the dataset to be analyzed to inform the question**

The data needed to be structured in a way that allowed it to be analyzed to answer the driving questions. First, the raw data was formatted into columns/rows. (See Raw Data Formatted Sheet 2.) Then, the rows and columns were flipped to make it easier to filter data for individual countries. (See Select Countries Formatted Sheet 4.)

### Part 2: Analyzing Data using data moves

Then, the student used data moves (Erickson et al., 2019) to answer the questions they had about the dataset. Examples of data moves are:

- **Filtering** is selecting only a subset of the data that is available to explore. For example, you may view only a portion of the data because only some of it is relevant to the question you are exploring.

- **Grouping** is used to make comparisons between different subgroups of a data set.

- **Summarizing** is computing a value (e.g. mean, median, mode) to summarize a group.

- **Calculating** is creating a new attribute based on values of one or more existing attributes. For example, if an attribute is distance in kilometers, you may want to convert that value to distance in miles.

- **Merging/joining** is combining more than one dataset together. You may add more cases, or add more attributes to existing cases.

- **Making a hierarchy** is grouping data using multiple levels.

In order to complete data moves, you should become familiar with a program that allows you to manipulate data. Google Sheets, Microsoft Excel, CODAP (Concord Consortium) are all great resources for using data moves to analyze your data.
**Question: How many cases have been confirmed in the US between Jan and Feb?**

**Data Move(s)**
- Dive deeper into the data set by combining moves

**Sketch or describe how you organized or illustrated the data:**
- Filter by country to view only confirmed cases from the US
- Find total number of cases at the end of February
- Find total number of cases at the end of January

**What does this tell you about the relationships/patterns within your dataset?**
- There have been approximately 60 cases diagnosed in the past month.

**What new questions do you have?**
- How long will the cases increase?

---

**Question: What country has the highest number of confirmed cases?**

**Data Move(s)**
- Dive deeper into the data set by combining moves

**Sketch or describe how you organized or illustrated the data:**
- Sorted from most recent date by highest to lowest

**What does this tell you about the relationships/patterns within your dataset?**
- The United States has the highest number of cases, followed by Spain, Italy, the United Kingdom, and Germany.

**What new questions do you have?**
- What continent has the highest number of cases? (See analysis below)
Question: What continent has the highest number of confirmed cases?

Data Move(s)
Dive deeper into the data set by combining data moves.

Sketch or describe how you organized or illustrated the data:

- Identified the 25 countries with the highest total number of confirmed cases
- Created a new column to show which continent each country is on
- Filtered each column by continent and add total confirmed cases

What does this tell you about the relationships/patterns within your dataset?
North America has the most cases diagnosed followed by Europe. There are fewer cases in South American and Asia.

What new questions do you have?
What measures would be effective to isolate the disease to a specific location or continent?

<table>
<thead>
<tr>
<th>Country</th>
<th>Continent</th>
<th>2020-05-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td></td>
<td>3,386,595</td>
</tr>
<tr>
<td>United States</td>
<td>North America</td>
<td>1,120,000</td>
</tr>
<tr>
<td>Spain</td>
<td>EuropeAfrica</td>
<td>310,502</td>
</tr>
<tr>
<td>Italy</td>
<td>Europe</td>
<td>293,200</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>Europe</td>
<td>182,200</td>
</tr>
<tr>
<td>Germany</td>
<td>Europe</td>
<td>182,406</td>
</tr>
<tr>
<td>France</td>
<td>Europe</td>
<td>120,979</td>
</tr>
<tr>
<td>Turkey</td>
<td>EuropeAsia</td>
<td>121,379</td>
</tr>
<tr>
<td>Russia</td>
<td>EuropeAsia</td>
<td>120,904</td>
</tr>
<tr>
<td>Brazil</td>
<td>South America</td>
<td>90,926</td>
</tr>
<tr>
<td>Iran</td>
<td>Asia</td>
<td>96,414</td>
</tr>
<tr>
<td>China</td>
<td>Asia</td>
<td>63,967</td>
</tr>
<tr>
<td>Canada</td>
<td>North America</td>
<td>56,716</td>
</tr>
<tr>
<td>Belgium</td>
<td>Europe</td>
<td>45,517</td>
</tr>
<tr>
<td>Peru</td>
<td>South America</td>
<td>42,834</td>
</tr>
<tr>
<td>Netherlands</td>
<td>Europe</td>
<td>41,294</td>
</tr>
<tr>
<td>India</td>
<td>Asia</td>
<td>30,920</td>
</tr>
<tr>
<td>Switzerland</td>
<td>Europe</td>
<td>25,739</td>
</tr>
<tr>
<td>Ecuador</td>
<td>South America</td>
<td>24,439</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>Asia</td>
<td>23,498</td>
</tr>
<tr>
<td>Portugal</td>
<td>EuropeAfrica</td>
<td>25,100</td>
</tr>
<tr>
<td>Mexico</td>
<td>North America</td>
<td>20,988</td>
</tr>
<tr>
<td>Sweden</td>
<td>Europe</td>
<td>20,662</td>
</tr>
<tr>
<td>Ireland</td>
<td>Europe</td>
<td>21,178</td>
</tr>
<tr>
<td>Pakistan</td>
<td>Asia</td>
<td>19,103</td>
</tr>
<tr>
<td>Chile</td>
<td>South America</td>
<td>18,436</td>
</tr>
</tbody>
</table>

Asia        280,957
North America 1,213,871
Europe        847,808
South America 184,952
EuropeAfrica 341,712
EuropeAsia  249,412
Explore the sources that reported the data. Use the flowchart below to consider error and/or bias that could be present.

This student found that Our World in Data is sourced from the European CDC. They stopped relying on data from the WHO in mid-March 2020 after finding multiple reporting errors. However, the total number of COVID-19 cases is likely higher than the number of confirmed cases reported on the infographic. This is because some individuals/countries have more access to testing than others.