|  |  |
| --- | --- |
| **Communicating with Data**  **Individual Project** | Image result for infographic |

## Objectives

**Students completing this series of exercises will learn to:**

* **Investigate existing data resources**
* **Access appropriate data resources**
* **Clean, configure and standardize a dataset**
* **Critically evaluate information derived from a dataset**
* **Analyze and visualize data for effective communication of science**

**Introduction**

**As part of the larger group research project, your team has assigned you a specific data resource to investigate. Your task is to access, clean, configure, and standardize a dataset for utility then create a relevant and scientifically valid data visualization to communicate science.**

**Activity I: Conduct a Data Review**

**A *data review* is a survey of existing data resources related to the specific subject, overarching idea, or research question you are interested in learning more about. The goal of a data review is to examine existing data and determine how the existing data can inform your research question or expand your knowledge on a specific topic.**

**Your data review is guided by your specific research topic. You are looking for specific research and data resources related to your topic. When you identify these resources, you will analyze this data and interpret the results of your analysis relative to your broader question.**

1. **Re-state the big idea or research question that your group has identified for your infographic.**
2. **Identify the individual research question or idea you have been assigned. Describe the topic or idea you are planning to cover and how that fits into the bigger overarching infographic topic.**
3. **Identify the types of data that can be used to address your question. Start by thinking broadly to identify any data related to any aspect of your question. For this question think big and list your ideas, key words, and potential resource for data types below.**
4. Identify 3-5 potential data sources that could be related to your research question. Your data resources can be a *dataset* (a set of data usually collected for a specific project or question; organized in rows or columns) or part of a *database* (a large collection of structured data often stored as multiple datasets that can be manipulated and accessed in various ways). This can be quite challenging. There is a lot of data out in the world and you need to think creatively. As you look for data you may also revise your question.

As a start, look at the [data resources](http://libguides.cmich.edu/lifesciencedata) we have identified. Fill in the table with your data sources and how this data could inform your research question (this is required for all students even if you have already identified a dataset/database).

|  |  |  |
| --- | --- | --- |
| Data Resource | Content or Data Type (what is in this data resource) | How could this data inform your research question? |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. Evaluate your potential data sources. For each of your potential data resources consider the questions below. Use the questions as a way to identify your best data resource. As you ask these questions, be thinking about the utility of each data source to address your research question or big idea. You could use multiple datasets to address your question but for this exercise you only need to select one. You are picking the best dataset for your question and the bigger infographic. Mentally go through the questions for each of the datasets you identified above and use this process to identify your final dataset/database.
2. Answer the questions below in writing for the single data set you are selecting for your project. Use complete sentences:
3. Is this single or multi-source data? (if multisource then the questions below need to be considered for each data set contributed to the whole before using the combined data).
4. What methods were used to acquire this data?
5. Has this data been gathered in a way that it is appropriate for your question?
6. Is this data verifiable? (not all data is; but knowing this can facilitate later or extended use)
7. Who gathered this data? Was this data acquired by experts? And educated novice?
8. Do you trust the accuracy of the data?
9. Is the dataset complete? Are there other sources of data that should be accessed to create a “complete” dataset?
10. Is this data relevant to your question/usable? How?
11. Is the data corrupt? Can you access and download the data for use?
12. Is the dataset complete and ready to use or will you need to configure, load, or aggregate the data yourself?
13. Are their duplicates in the data? How do you know?
14. Is the data gathered in a consistent manner?
15. Is their sufficient data to address your question with an acceptable level of error/variation?
16. Is this raw data or has it been manipulated in any way?
17. Do you have to enter, manipulate, or aggregate the data?
18. What is the resolution of the data? (think quantitative, spatial, temporal, and scale)
19. If data is aggregated how do the individual data sources impact the resolution of the overall data set?
20. Is this data linked to other data? What additional data can be associated with this data (think extended data)?
21. Is this presence only data?
22. Is this data open source? Do you have permissions to use the data?
23. Was this data gathered in an ethical manner and is it ethical to use the data?
24. Is this dataset continuous, categorical, discrete?
25. How will you use this data to address your question?

**Activity II: Clean, Configure and Standardize Your Dataset[[1]](#footnote-1)**

**Everyone’s dataset will be different, and everyone will need to work with the dataset to optimize it for use and analysis. You could be doing anything from entering data from paper field notes, downloading data from a larger database, cleaning a citizen science dataset, aggregating two or more data resources, creating shapefile from images, etc. Once your dataset is in the form you need for analysis (and this could be very different for each of you) you need to examine your data to make sure it is usable. For the purpose of this exercise follow the steps below:**

1. Inspect your data. Detect unexpected, incorrect, and inconsistent data. This might include running quick summary statistics and/or running a data visualization to “see” irregular data.
2. Clean your data. Fix or remove the anomalies discovered. This includes removing irrelevant data, duplicate data, and syntax and text type errors. This also means standardizing data, handling missing data points, and removing outliers.
3. Verify your data. After cleaning, inspect results to verify correctness.
4. Report your data. Generate a report including the changes made to the dataset and the quality of the currently stored data in the dataset. Document the steps you took clean, configure, and standardize your data. Use complete sentences.
5. Turn in both the cleaned dataset and the report. The dataset and report should be detailed to the extent that the instructor can fully review your process, decisions, methods and final dataset product.

**Activity III: Analyze and Visualize Your Data**

1. **Review the** July 9, 2019 Ecological Society of America Webinar titled, “[Sharing Science Clearly: Communicating Science through Data Visualization](https://register.gotowebinar.com/recording/4591639022969051404).” There are several data visualization resources noted in the webinar that may assist you in identifying the appropriate analysis and picking an appropriate data analysis and visualization.
2. Run your analysis and create your data visualization.
3. List your data resource and provide a proper citation (for details on citing data go to: <http://libguides.lib.msu.edu/citedata>).
4. Describe the question or content addressed with your data visualization (be specific and make sure you describe what you are trying to communicate in your data visualization).
5. Describe the statistical or data analysis conducted. What does the analysis show?
6. Provide a figure caption. The figure caption should be precise and concise but include sufficient information to orient the person examining the figure. The description following the caption should include a description of the data and source, specific reference to important elements of the figure, and some interpretation of the data. **Figure captions should be below the figure**.
7. Check off the items reviewed below (make sure to check each item so it is clear it was addressed):

* **Does the graph/figure help answer the question?**
* **Is the graph/figure type appropriate for your data?**
* **If axes are needed, are they drawn and scaled correctly?**
* **Are the graph/figure components (including axes) labeled clearly and correctly?**
* **Are units provided where needed?**
* **Are data plotted and/or represented accurately?**
* **If needed, is a legend present?**
* **Is the graph/figure neat and legible?**
* **Is a title and/or caption present?**
* **If appropriate, is a trend line shown?**

1. Explain the bigger overarching idea, question, or problem your working group is trying to address in your infographic.
2. Explain how your data, analysis, and visualization contribute to the overall infographic.

1. Materials come from: <https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4> accessed 15 March 2020 [↑](#footnote-ref-1)