BIO 410/510 Final Project Workflow

Please submit to Canvas before class on Wednesday 11/14

Points: 30

Final project objective

The objective of the final project will be to complete a fully reproducible workflow that uses data to address your chosen question. The project must illustrate all of the following tasks:

- Some form of data access / reading into R
- Integration of multiple datasets to address the question
- Data tidying preparation using tidyr, including data joins
- Use of dplyr to manipulate and summarize the data in relevant ways
- Initial data visualization with ggplot2
- Final, publication-worthy visualization with ggplot2
- RMarkdown writeup, with final submission as both the .Rmd file and a nicely formatted PDF document that includes code and results
- Overall clean and clear presentation of the workflow, code, and explanation

Intermediate step: Literature review and workflow plan

To progress toward the final project, please prepare a literature review, identify a dataset and develop a workflow plan. A description of each is below. We will peer-review these prior to the final project to help each other with the workflows.

I. Literature review

I expect the literature review will be around 5 well cited paragraphs that do the following:

1. Introduce the problem and explain why

- Set the stage for the problem
- Put the concept and question into context
- Lots of big-picture citations (such as reviews) in the first paragraph

2. Past work and data available on the project

- Who has addressed this problem, and what did they do it?
- What are the data available to address this problem?
- How has the data available and/or methods changed recently?

3. Purpose of the study

- Further refine your approach (e.g., what data will you combine, how will you address the question)
- Justify why this is needed now (e.g., visualization to test a new dimension of the question or better convey an old one)

4. Hypotheses/questions

- List these clearly and in a logical order
- Make hypotheses directionally using predictions (e.g. "I predict N will reduce plant diversity" rather than "I predict N will change plant diversity")

5. Literatured cited

- At least 10 relevant, peer-reviewed citations
- Citations are scientifically formatted (e.g., follow conventions for the journal *Ecology*)

This is a nice reference on scientific writing: Turbek, Sheela P., Taylor M. Chock, Kyle Donahue, Caroline A. Havrilla, Angela M. Oliverio, Stephanie K. Polutchko, Lauren G. Shoemaker, and Lara Vimercati. "Scientific Writing Made Easy: A Step-by-Step Guide to Undergraduate Writing in the Biological Sciences." The Bulletin of the Ecological Society of America 97, no. 4 (October 2016): 417–26. https://doi.org/10.1002/bes2.1258.

II. Dataset identification

Please identify the datasets you will be using to address your question. I expect at least two datasets that will be combined to address your questions. Provide the data source (including web url) and sufficient metadata to convey who collected the data, how it was collected, and what each column contains. In addition, please provide an overview of the structure of the datasets (give as much information about the data as the str() function would return).

III. Workflow plan

In English, please describe the workflow you will use tidy your raw data, manipulate and summarize it in relevant ways, and visualize it. Please include any QAQC steps (e.g., "remove non-species such as 'miscellaneous litter' from the species column") as well as aggregation steps (e.g., "count the number of entries by plot and year to calculate species richness"). The goal here is to develop a logic to your workflow before you code.