Data Workflow Challenge

Why, and what?

The other lessons in this workshop cover different tools that may be useful for you in your data analysis. The next step is to learn how to put these tools together in a way that will make your workflow more efficient, effective, and reproducible.

While you could theoretically do all of your work in the command line, you wouldn't be able to look back at your analysis to check, modify, or re-use it. Thus, it is a much better approach to save your work in scripts. It is often helpful to have a script that you can use as a scratchpad, to work out your approach and figure out how to code it, but it is a good idea to put your final code into its own script. This script should start with a (commented) description of what it does, along with your name and other useful information, like the date that you began working on it. It should also have comments throughout that give information on the different pieces.

In this challenge, you will have the opportunity to write scripts that will look like the kinds of scripts you might write to analyze your own data. This will allow you to:

- Practice using the tools that you have learned so far
- Learn how to put these tools together
- Create sample scripts that you can use as models for future work

---

The project

Imagine that you are studying nitrogen cycling in different ecosystem types, and you have conducted an assay to measure potential rates of nitrification - a microbial process that converts NH_4^+ to NO_3^- - in soils from a forest, an agricultural field, and a grassy meadow[1]. In this assay, you made a soil slurry, removed all the NO_3^-, added NH_4^+, and measured the production of NO_3^- over an 8 hour incubation at room temperature. (Typically you would collect samples at multiple time points, but we will just use an end point here.)

At the end of this assay, you ended up with sample solutions with different amounts of NO_3^- in them. To measure NO_3^- concentration, you added a reagent that colors the solution blue (Szechrome reagent), and then used a spectrophotometer to read absorption of light at a specified wavelength. You also did this for a series of samples at known concentrations, so that you could create a calibration curve.

What you have now is a csv file, *Nitrification_Absorbances.csv*, with sample type (sample or standard), sample ID, concentrations of the standards, and absorbance values. (You also have an inventory file, *Sample_Inventory.csv*, to match sample ID to sampling site and replicate number.) What you want is nitrate concentrations for all of your samples.

Furthermore, you know that you will have similar files to work with in the future, after you've done similar experiments with samples from other locations. So it's important to make sure that your code can be easily modified to be applied to other datasets! Fortunately, it also means that even though it will take time to set up the analysis properly now, it will be very easy to run it for future samples.

[1] *(When a biogeochemist runs the workshop, the data will involve nitrogen!)*

The challenge

We will work through pieces of this at a time, with breaks to talk about our approaches and any particular challenges.

Start by setting up a file structure for the project.

- Setup:
  - Create a project folder with an appropriate file structure, including subfolders for:
    - Raw data
    - Clean data

Then, write a script to read in all of the absorbance data and output sample concentrations to the clean data folder:

- Script 1: Raw data to clean data
  - Read in absorbance data (.csv file)
  - Plot the calibration data
    - Write (and run) a function to do this
  - Calibration curve
    - Again, use a function to do this!
    - Figure out how concentration (dependent variable) is dependent on absorbance (independent variable) for standards: run a linear regression
    - Print the adjusted r^2 value for the model fit
    - Extract the coefficients of the equation relating concentration to absorbance
  - Calculate sample concentrations
    - Write a function that uses the linear regression coefficients to calculate concentration based on absorbance
    - Apply this to only the samples, not the standards
    - Output a data frame with sample ID and calculated concentration
  - Write data to file
    - Output a .csv file with this data to the clean data folder

Then, write another script that uses this clean data and looks for differences between treatments:

- Script 2: Clean data to plotting and analysis
  - Read in the concentration data
  - Read in sample inventory with treatment data
  - Merge the clean data with the sample treatment information from the inventory
  - Make a plot to look at differences in NO_3^- concentrations among treatments
  - Run an ANOVA to assess differences in NO_3^- concentrations among treatments, and be sure to check model assumptions and run any relevant post-hoc tests