**Working with Datasets in R**

**Author:** Caitlin Hicks Pries, Dartmouth College

**Focus:** The goal of this lesson is to learn how to import datasets into R, understand variable types, make basic adjustments to columns in dataframes, and begin data visualization through graphing.

**Overview:** This lesson is part of a Global Change Biology course where one of my main learning objectives is to have the students develop and answer scientific questions by using R to access, organize, and graph actual longterm data sets.

In order to explore global change data and learn how various anthropomorphic activities are affecting ecosystems and species, we need to import datasets into R and make sure all the variables are in the correct format that will allow us to perform data visualization (aka, create graphs). In this lesson, we will learn how to import data into R, perform basic QA/QC (quality assurance/quality control) of the data, and create some basic exploratory graphs of the data from which we can begin to answer scientific questions. We will be working with a dataset from the PRISM Climate Group (http://www.prism.oregonstate.edu/) that includes air temperature and precipitation over the past one hundred years two towns in the United States (Boulder, CO and Hanover, NH).

**Learning objectives:**

* 1. You will learn how to import data into R and perform basic QA/QC.
	2. You will learn about different data and variable formats you will encounter in this course (and throughout your data ‘career’).
	3. You will learn how to manipulate date variables in R in order to get them into the format you need for data visualization.
	4. You will learn how to use piping in R (from tidyverse package) to subset the data and perform basic calculations with the data.
	5. You will be able to create simple graphs to visualize data in order to answer scientific questions.

**Lesson sequence:**

1. Prior class lecture and swirls (see below)
2. Introduction to lesson. This introduction should include an introduction to the database (PRISM) and dataset the students will be working with during the swirl, an introduction to the concepts of long vs wide-form data and a review of variable types (numeric, integer, character, factor, etc.). Discuss the scientific questions we will ask of the data and ask the students to sketch the graphs that they think will answer the question. These graphs will be the overall goal of the SWIRL lesson.
3. SWIRL lesson
4. Assessment activity

**Pre-lesson activities:** For R preparation during previous classes, we downloaded R and R studio, I gave a brief introduction to R studio, and the students completed three swirl lessons from the “R Programming: The Basics of programming in R” course. These were: 1: Basic Building Blocks, 7: Matrices and Dataframes, and 12: Looking at Data. For previous classes, we also read part of the Intergovernmental Panel on Climate Change report and discussed climate change. This content gave students the context for working with the climate data and what types of questions we might want to ask using the data.

**Post-lesson activities:** After this swirl lesson, as an assessment activity, the students completed a homework assignment wherein they had to make three graphs of the climate data and submit their R scripts to me. The homework and a key for the homework are available with this lesson.

**Implementation notes:** As expected the hardest part of getting students working in R was downloading the program and getting it to run on their computer. There are always bugs, so be prepared. Also, I ran into problems running my swirl lesson, if I did not start with a fresh, clean Global Environment in R Studio. The other big sticking point for several students was actually downloading the .csv data file from the web onto their computer and putting it into the appropriate folder. Thus, covering general R workflows and file management might be beneficial.

**Resources Used**:

Climate data was downloaded from PRISM (<http://www.prism.oregonstate.edu/>). This website has climate data for the contiguous United States. You may want to download data from the location of your school. However, that will change the answers to two questions in the YAML file (the mean temp and which city has more precipitation), so those would have to be changed.

The graphing part of this lesson was modified from the existing swirl lesson GGPlot2\_Part1 in the Exploratory\_Data\_Analysis course (<https://github.com/swirldev/swirl_courses/blob/master/Exploratory_Data_Analysis/GGPlot2_Part1/lesson>).