HW 07: Data Importing and Wrangling 1

Graphical Analysis of Biological Data

By the end of this assignment, you should be able to achieve the following tasks in R:

- use R notebooks and R markdown;
- insert, write, and evaluate code chucks;
- import data stored in a variety of text file formats;
- make untidy data tidy;
- arrange, filter, and select data;
- use pipes;
- produce plots with `ggplot2`;
- use a typical workflow to wrangle and plot data; and
- confidently stage, commit, and push with Git.

These achievements belong to Learning Outcomes 2, 3, 4, 5, 6.

Click on any blue text to visit the external website.

Note: If you contact me for help or (better yet) open an issue in the public discussion forum, please include the code that is not working and also tell me what you have tried.

Preparation

- Open your `.Rproj` project file in RStudio.
- Create an `hw07` folder in the same folder as your project file.
- Create a `data` folder in the same folder as your `Rproj` file, at the same level as your other `hw` folders. Do *not* put this folder in your `hw07` or another homework folder.

  You will store all your data files that you use for this course in this folder. That allows everyone in class to have a common path to the data. You will use some data sets multiple times during this course so you want the data in a common folder accessible to all assignments.
- **Stage and commit regularly!** A good habit would be to stage and commit after you complete each section. You should also push each time but at least push your completed notebook to your remote repo.

Assignment

Your task here is simple. Write code chunks that

- import data,
- wrangle the data as specified, and
- graph the data.

Then, describe the results briefly. Identify patterns and trends, potential outliers, and other results you find interesting. Think like the scientist you are!

You may need to review your Introduction to R coding exercise to meet some requirements, especially factors and how to select one column from a dataframe, and how to change one value to another.

Remember to put all the data files in your `data` folder.

A few things:

- Create a new notebook file called `<lastname>_07.Rmd`. Edit the YAML file as you have for previous assignments.
- Load the `tidyverse` package.
- In this same code chunk, you can define a variable with the path to your `data` folder, as discussed in the notes.

  You only need to load `tidyverse` and define your path variable once at the top of your notebook. You do not do this for each chunk.
- Download this zip file of data. Unzip the file and place all of the files in your `data` folder. If you find yourself asking, "What data folder?", then (1) shame on you for not reading the notes and (2) go read the notes or the information above.
- Each data file has one of the following types.
  - Comma separated
  - Tab separated
  - Fixed width

  You must always open the file in a text editor to inspect the file. You should inspect the file to determine the separation type (comma, tab, fixed width), look for column headers or comments that need to be skipped, and get an overall sense of the data size and types, such as number of rows and columns, numeric and text variables, and so on. *You have to know the data you are going to use!*
- Commit early. Commit often. Push regularly but at least push your completed assignment.

Hint: `tbl$x <- factor(tbl$x, levels = c(...), ordered = TRUE/FALSE)`. Just sayin'....

Limpets

Let's begin gently.

Fenberg and Roy 2012 studied how human harvesting of the owl limpet (*Lottia gigantea*) affected its life history. The study sites are along the coast of California.

Requirements:

- Data file: `limpets.csv`.
- Are the data tidy?
- Use the `col_types` argument. The three column types are numeric, character, and character.
- Make the `Sites` column an ordered factor with these levels:
  - PBL, KNRM, VBG, WP, PF, DP, SIO, CTZ1, CTZ2, CNM
  - PBL is the northern most site (Pebble Beach). CNM is the southernmost site (Cabrillo National Monument).
- Make a boxplot of length (in millimeters) for each site, colored by protected status.
- Change the axis labels so they begin with capital letters. The y-axis should include the unit of measurement in parentheses.
- In your description, tell which two sites have outliers and whether the protected sites tend to have larger or smaller limpets.

Roseate Terns

Seward et al. 2018 studied metapopulation dynamics of roseate terns (*Sterna dougallii*) in northwestern

Europe to determine how abundance changed at nine sites. The number of individuals was counted at each site every year between 1992 and 2016.

- Data: roseate_terns.txt
- Are the data tidy?
- Use filter to remove sites with missing counts.
- Make a line plot of population size over time.
- Change the axis labels as appropriate (you have to start thinking about what is appropriate).
- Which population(s) obviously increased in size between 1992 and 2016? Which population(s) obviously decreased in size during that time?
- Some lines have breaks in them. That is, they are not continuous across all years. Why?

Blacklip Abalone

Warwick et al. 1994 studied the population biology of Blacklip Abalone (*Haliotis rubra*) from the north coast and Bass Strait Islands of Tasmania.

- Data: abalone.csv (ab-ah-LOW-knee; rhymes with bologne)
- Follow the instructions carefully. This exercise walks through a few steps of a "typical" analysis. Make a separate code chunk for each instruction.

Chunk 1: Import, remove the first column, then make a boxplot of height differences among the three types.

Chunk 2: The boxplot for height shows a female and a male outlier. Perhaps the samples contained two very large, old individuals. Make a scatterplot to see if height appears to correlate with rings. Rings is a measure used to estimate age. Based on the graph, are the extraordinarily large individuals really old individuals?

Chunk 3: Let's assume the outliers are coding errors so remove them by filtering. Filter the data to remove the two large individuals. Change Type to an ordered factor. Immatures must be first, as that makes sense in terms of age. The order of female and male after immature is up to you. Then, redo the scatterplot that you just made with the newly wrangled data.

What patterns emerge? Which type is the largest? Are all females and males larger than immatures?

Chunk 4: Are there really immatures with more than five rings with zero height? Srsly? - Print the records of the individuals with zero height. - Most likely, the two zero height values are mistakes made during data recording. This time, instead of filtering them, assign `NA` (missing data) to those two records. Repolot the data to ensure the two observations are not included in the graph.

Chunk 5: Make two scatterplots of your choice, between any two pairs of continuous variables that make sense to show as scatterplots. Color, shape, or both should distinguish the three types.

Darters

This will be your most challenging import. *Inspect the file!* Think through the problem. If necessary, jot some notes on paper to outline the information you need to consider, such as how many lines to skip, the position of columns, etc. Use chunks logically and appropriately to accomplish the tasks described below.

Taylor (unpublished) studied the microhabitat use of darters in the genus *Etheostoma* (Family Percidae) from the Niangua River watershed in Missouri.

- Data: darters.txt
- Column names and widths are included in the file. You can use whatever column names you want but adjust accordingly for info below.
- Make `riffle` an unordered factor with levels 1 and 2.
- Make `major_type` an ordered factor with levels s, fg, sg, lg, c

Do these four steps together with the pipe.

- Filter to remove rows with "zonale" and "tetrazonum".
- Remove `mintype` and `minsub` columns.
- Rename `majsub` and `majtype` to `major_substrate` and `major_type`, respectively.
- Arrange the data by `id`.

Data were collected from two riffles, separated by several hundred meters. The plots below explore differences between riffles. Use `facet_wrap()` to make pairs of plots separated by riffle.

Plot 1 Plot length as a function of depth. Map species to color and shape. What differences do you see between the two riffles?

Plot 1 chunk here.

Plot 2: Make a boxplot of length for each of the three species. Which riffle shows the greatest number of outliers?

Plot 2 chunk here.

Plot 3: Make a boxplot of length for major substrate types for each species for each riffle. This will actually be six plots in one! To do this, use `facet_grid(species ~ riffle)` in place of `facet_wrap()`. How does the plot change if you switch the order of the argument (`riffle ~ species`) in the `facet_grid()` function?

Plot 3 chunk here.