FISH 6002: Data Collection, Management, and Display - Assignments

Dr. Brett Favaro

10% of your course grade is earned by participation. Just show up, be yourself, and participate!

The other 90% will be earned by completing assignments. These fall into three categories:

- Explo**R**ation - Partner presentation (20%)
- Major assignment (50%)
- Minor assignments (30%)

This document outlines the details and grading criteria for each assignment.

ExploRation - Partner presentation

R is an ever-expanding piece of software, with thousands of package and millions of users across the world submitting content. R can be used to take data from all sorts of databases. It can be used to read text from a PDF. It can even be used for desktop publishing.

In this assignment, your job is to explore R, identify something interesting that isn't otherwise being taught in the course, and tell us about it.

Deliverable:

In partners, you will prepare a 10-15 minute presentation (plus 5-10 minutes for audience questions) that accomplishes these objectives:

1. What is it?
2. What does it do?
3. How does it improve on other methods?
4. Demonstrate how to do it
5. Share **reproducible** code with the class

Your "trick" can be serious (e.g. how to extract data from a specific database, how to read text from a PDF, etc.), silly, or anything in between. The key should be that it builds your ability and mastery of the R language, and expands the horizons of your classmates.

Code that is shared for this assignment must be fully reproducible and extensively commented. Any data needed to execute it should be included with the project.

Timeline

Week 1: Identify your partner By Week 3: Select topics. Review with instructor. Starting week 4: One group per week will present.

Value: 20% of course grade

Grading scheme:

1-3 will net you 10%, while #4 and #5 are worth 5% respectively.

Major Assignment

The major assignment for this course will proceed in three parts: Data collection, management, and display.

Your task is to take data from a published source, document it thoroughly and prepare it as 'tidy data', and then redisplay it.

Part 1: Collection

Select a science paper related to fisheries (if you need inspiration, select one from https://sites.google.com/a/uw.edu/most-cited-fisheries/). DFO reports are acceptable as well, and they often contain data tables directly in the paper. Pick wisely - this paper is going to be the basis for the duration of your major assignment.

Your *first task* is to obtain the core data from this paper. It is up to you to figure out how to do this. You may contact the author directly, you may use a program like GraphClick or DataThief to pull data from figures, or you may find a paper that drew from a specific database such that you could recreate the database query the authors used to produce their paper. This assignment has two deliverables:

1. Produce a one to two-page summary indicating where the data came from, the volume of data collected, and how you obtained it. Describe what the data include (time period covered, subject, geographical range, etc.) This should be comprehensive enough that your summary gives enough information to understand the dataset.
2. Collect raw data files in CSV format representing the entire dataset. They need not be tidy, just complete.

This dataset should contain at least 500 unique values (i.e. be too big to easily manipulate manually, thus requiring computer code to do so effectively).

Value: 10%

Grading:

- Summary /6
- Raw data: /4

Due: End of Week 4.

Part 2: Manipulation

Your *second task* is to tidy the data you have collected. Organize the data into a coherent spreadsheet that is ready for analysis in R. Produce CSV files containing both the long and wide-format data. Do all manipulations in R, and share your code.

Note that this will require pulling data into R, manipulating it, and then outputting it as new CSV files - all things we will talk about in class.

Part 2 has three deliverables:

1. Two CSV files - one long-format and one wide-format, containing the data you collected from Part 1. These data should meet the criteria of being 'tidy data'
2. R code, fully commented, that shows how you conducted your manipulations. I want to see evidence of:

- Error detection and cleanup
- At least three dplyr operations (e.g. mutate, join, count)
- Manipulation of data into long and wide formats

Even if your data are very clean, you need to demonstrate to me how you checked for that and verified it to be the case.

Value: 15%

Grading:

- CSV files
- long format /2.5
- wide format /2.5
- R code
- comprehensiveness (i.e. operations conducted in R) /5
- commenting (i.e. operations are explained) /5

Due: End of Week 8.

Part 3: Display

Your *third task* is to use R to produce three figures and one table that summarize the data that you have collected visually. You may use ggplot, base plot, or any other R mechanism to do so, but you should try to make the graphs as informative and beautiful as possible. We will talk at length in class about how to make figures clean, readable, and informative, and you will be assessed on that basis. However, I encourage creativity here, and would like you to challenge yourself to make truly beautiful data visualizations.

Part 3 has four deliverables:

1. At least three graphs that summarize key aspects of your dataset. You may recreate (and improve upon) figures already in the paper, or start from scratch. You may produce a single multipanel plot with three figures if you think it would be more effective than doing them separately. **You must use at least three different graphing elements, e.g. boxplots, dots, violin plots, lines, etc.)**

These plots should be formatted for publication - at least 300 DPI, in PNG or TIFF format, following Figure prepration directions from Plos One.

2. At least one data table, formatted according to Plos One requirements, that summarizes something from the dataset
3. R code, fully commented, that shows how you produced your table and figures

- Error detection and cleanup
- At least three dplyr operations (e.g. mutate, join, count)
- Manipulation of data into long and wide formats

Even if your data are very clean, you need to demonstrate to me how you checked for that and verified it to be the case.

Value: 25%

Grading:

- For each graph:
- Effectiveness /3
- Aesthetic appeal, including high DPI output /2
- Data table:
- Effectiveness /2.5
- Data table formatted for publication /2.5
- R code
- Comprehensiveness (i.e. operations conducted in R) /2.5
- Commenting (i.e. operations are explained) /2.5

Total Value: 50% of course grade

Due: End of Week 11.

Minor Assignments

I define a minor assignment as something that takes < 1 week to complete. You will be given directions on these in class. Expect there to be three minor assignments, each worth 10% of the course grade. I will do my best to spread them out.