HED 696C: Data Management and Manipulation in R

The University of Arizona

Fall 2022

Karina Salazar
Assistant Professor
Center for The Study of Higher Education
Office Hours: Wed 2:00-4:00pm; via Calendly
E-mail: ksalazar@email.arizona.edu

Class Room: COE 311
Class Hours: M 4:15pm - 6:45pm
Class Website: ksalazar3.github.io/HED696C_RClass
Class Discussion: d2l.arizona.edu

Course Description

This course has two foundational goals: (1) to develop core skills in "data management," which are important regardless of which programming language you use, and (2) to learn the fundamentals of the R programming language.

Data management consists of acquiring, investigating, cleaning, combining, and manipulating data. Most statistics courses teach you how to analyze data that are ready for analysis. In real research projects, cleaning the data and creating analysis datasets is often more time consuming than conducting analyses. This course teaches the fundamental data management and data manipulation skills necessary for creating analysis datasets.

The course will be taught in R, a free, open-source programming language. R has become the most popular language for statistical analysis, surpassing SPSS, Stata, and SAS. What differentiates R from these other languages is the thousands of open-source "libraries" created by R users. R is one of the most popular languages for "data science," because R libraries have been created for web-scraping, mapping, network analysis, etc. By learning R you can be confident that you know a programming language that can run any statistical modeling technique you might need and has amazing capabilities for data collection and data visualization. By learning fundamentals of R in this course, you will be "one step away" from web-scraping, network analysis, interactive maps, quantitative text analysis, or whatever other data science application you are interested in.

Students will become proficient in data manipulation tasks through weekly "problem sets" and a final project. Class will begin each week with a discussion of challenges encountered while completing the problem set. The rest of class time will be devoted to learning and practicing new material. The instructor will provide students with lecture notes, and also data and code used during lecture. Therefore, students can follow along by running code from their own computers.

Course Learning Goals

1. Understand fundamental concepts of object oriented programming
    - What are the basic object types and how do they apply to statistical analysis
    - What are object attributes and how do they apply to statistical analysis

2. Become familiar with Base R approach to data manipulation and Tidyverse approach to data manipulation
3. Investigate data patterns
    - Sort datasets in ways that generate insights about data structure
    - Select specific observations and specific variables in order to identify data structure and to examine whether variables are created correctly
    - Create summary statistics of particular variables to diagnose errors in data
4. Create variables
    - Create variables that require calculations across columns
    - Create variables that require processing across rows
5. Combine multiple datasets
    - Join (merge) datasets
    - Append (stack) datasets
6. Manipulate the organizational structure of datasets
    - summarize and collapse by group
    - Tidy untidy data
7. Automate iterative tasks
    - Write your own functions
    - Write loops
8. Learn habits of mind and practical strategies for cleaning dirty data and avoiding errors when creating analysis variables variables

Prerequisite Requirements

1. Students must have taken at least a one-semester introductory statistics course.
2. Students should have some very basic experience using statistical programming software (e.g., SPSS, Stata, R, SAS).
3. [General computer skills] Students should be able to download files from the internet, rename these files, save them to a folder of your choosing, and open this folder.
    - During this course we will often be downloading datasets, opening .Rmd files and .R scripts, changing directories to the folder where we stored the data, and then opening the dataset we just downloaded. Therefore, it is important that students feel comfortable doing these tasks.

Course Format & Modality

This course is designed for the instructor and students to collectively contribute to the process of learning how to code in R. Weekly class sessions will largely be based on instructor-led lectures and in-class practice via group/buddy or individual coding (instructor will provide prompts/problem sets).

This course is designated as a hybrid "flex in-person: synchronous + zoom". These class sessions proceed as synchronous meetings, with some students in the classroom and some students attending via Zoom, with the instructor presenting content and facilitating in-class discussion.

Hybrid formats, while accommodating to various needs and preferences, can be challenging. Particularly when we are group/buddy coding. Given our classroom technology, I will pair students for these in-class assignments to other students in the same modality (in-person with in-person,

online with online). It may be that in some instances you are the only participant online or in-person. If that is the case, you will be responsible for completing the in-class activities on your own.

Course Readings

Course readings will be assigned from:

- Wickham, H., & Grolemund, G. (2018). *R for Data Science*. Retrieved from http://r4ds.had.co.nz/ [FREE!]
- Xie, Y., Allaire, J. j., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Retrieved from https://bookdown.org/yihui/rmarkdown/ [FREE!]

Required Software and Hardware

Software

Instructions on downloading software can be found on D2L.

Please install the following software on your laptop

- R
- RStudio
- TinyTex

Hardware

- Please bring in laptop with above software installed each week

Course Website and Resources

Course Website can be found TBD. We will use this website to download course materials such as lecture slides in pdf and .Rmd formats, data, weekly problem sets, and other class resources.

We will only use D2L when necessary to submit assignments or to post discussion questions.

Discussion and Homework Questions

We are using D2L as our class discussion forum where students can ask homework questions/comments to share with the instructor and the entire class. If you're stuck on a homework question or are experiencing problems with R more generally odds are others are too. Posting questions and concerns on D2L is the easiest way for us to all benefit from each others knowledge. When asking questions on D2L, please include as many details to replicate the "error." Always indicate the homework assignment and question number that's causing you issues, insert your code and provide screenshots to your posts.

I strongly encourage all questions related to course content to be posted on the D2L discussion forum for each week. I will do my best to reply to all posts within 24 hours. I also encourage you all to share your thoughts/answers on posts by your classmates. Writing out explanations to student questions will improve your own knowledge and will benefit your classmates. Sharing different ways to get at the "right" answer will be beneficial for all.

Assignments & Grading

Your final grade will be based on the following components:

- Weekly problem sets (60 percent of total grade)
- Final Project (25 percent of total grade)
- Attendance and participation (15 percent of total grade)

Weekly problem sets (60 percent of total grade)

Problem sets are due by 4:15PM each Monday (right before the class meeting). Late submissions will not receive points because we will discuss solutions during class. The lowest grade across all problem sets will be dropped from the calculation of your final grade.

In general, each problem set will give you practice using the skills and concepts introduced during the previous lecture. For example, after the lecture on joining (merging) datasets, the problem set for that week will require that students complete several different tasks involving merging data. Additionally, the weekly problem sets will require you to use data manipulation skills you learned in previous weeks.

Students can work on problem sets with classmates and I highly encourage you to do so. However, each student will submit their own assignment. You are encouraged to share ideas and get help from your classmates. However, it is important that you understand how to complete the problem set on your own, rather than copying the solution developed by group members.

A general strategy I recommend for completing the problem sets is as follows: (1) attempt the problem set on your own; (2) talk/meet with classmates to work through the problem set, with a particular focus on areas group members find challenging.

Final Project (25 percent of total grade)

Final Project (20%), Class Presentation (5%)

Students will complete a final project that incorporates many of the skills learned throughout the semester on a "real world" research task. The final project can be completed via three different options: 1) The final project can be fulfilled by completing and/or making progress on a research data task you are currently working on for your thesis/dissertation or for your job; 2) You can complete a guided online tutorial/workshop on a data related topic or task (e.g., building maps, machine learning, connecting to API's) but it must use R (i.e., I won't accept SPSS or Stata workshops); 3) I will also provide a final assignment to fulfill the requirement, which will be similar in format to weekly problem sets but will not provide as detailed "guidance" in how to complete the project.

If you are *not* fulfilling the requirement via the instructor created final assignment, then you will need to receive approval on your final project idea by November 14, 2022. I highly recommend you begin thinking about the project early in the semester and meet with me individually to discuss.

We will discuss details of the final assignment in class on October 24, 2022, including instructor provided examples and approved tutorials for Option #2. After November 21, 2022, class lectures will continue to introduce new topics but no weekly problem sets will be assigned. You are expected to use that time working on the final project that is due on December 13, 2022.

Students will also create and give a 10-15 minute presentation for their final projects Presentations will be scheduled for the last day of class: December 5, 2022.

Attendance and Participation (15 percent of total grade)

Students are expected to participate in weekly class sessions synchronously (via Zoom or in-person). This is a lecture and activity based class. It requires your *active* participation. Please come to each class session prepared to discuss the readings, ask questions, and practice coding. However, I understand there is still much uncertainty and complexity in the semester ahead of us. If you cannot attend class sessions for professional, personal, or health reasons, please just let me know ahead of time (if possible).

Course Policies

Class Recordings

Class sessions will be recorded to provide an "asynchronous" option for students that may need to miss some class sessions. Recordings are also a helpful resource for students that are able attend weekly class sessions but need to return to a topic to complete the weekly problem set. Class sessions will be recorded and accessed via D2L only. Students may not modify content or re-use content for any purpose other than personal educational reasons. Per university policy and FERPA, all recordings are subject to government and university regulations. Therefore, students accessing unauthorized recordings or using them in a manner inconsistent with UA values and educational policies are subject to suspension or civil action.

Classroom environment

We all have a responsibility to ensure that every member of the class feels valued, safe, and included.

With respect to the course material, learning coding/programming and the essential skills of data manipulation is hard! This stuff feels overwhelming to me all the time. So it is important that we all create an environment where students feel comfortable asking questions and talking about what they did not understand.

With respect to creating an inclusive environment, be mindful that what you say affects other people. So express your thoughts in a way that doesn't make people feel excluded.

Accessibility and Accommodations

At the University of Arizona, we strive to make learning experiences as accessible as possible. If you anticipate or experience barriers based on disability or pregnancy, please contact the Disability Resource Center (520-621-3268, https://drc.arizona.edu/) to establish reasonable accommodations.

Academic Honesty

Academic Integrity at the University of Arizona is the principle that stands for honesty and ethical behavior in all homework, tests, and assignments. All students should act with personal integrity and help to create an environment in which all can succeed.

Violations of the UA Code of Academic Integrity are serious offenses. As your instructor, I will deal with alleged violations in a fair and honest manner. As students, you are expected to do your own work and follow class rules on all tests and assignments unless I indicate differently. Alleged violations of the UA Code of Academic Integrity will be reported to the Dean of Students Office and will result in a sanction(s) (i.e., loss of credit on assignment, failure in class, suspension, etc.)

Students should review the UA Code of Academic Integrity which can be found at: https://deanofstudents.arizona.edu/policies/code-academic-integrity

Course Schedule

Students in the course are likely to have varying levels of experience with R. Because it is difficult to anticipate our pace as a class, the following schedule should be treated as a guide. Topics will likely carry-over into the following week(s). We may also end up cutting later topics if, as a class, we need additional time to cover a previous topic thoroughly. For this reason, readings will be assigned on a week-to-week basis.

Work and course requirements are subject to change at the discretion of the instructor with proper notice to the students.

Week 1, 8/22/2022: Course introduction; Getting started with R

Week 2, 8/29/2022: Objects in R and Missing Data

Week 3, 9/5/2022: No Class, Labor Day

Week 4, 9/12/2022: Introduction to using tidyverse to investigate data patterns

Week 5, 9/19/2022: Introduction to using "base R" to investigate data patterns

Week 6, 9/26/2022: Pipes and variable creation

Week 7, 10/3/2022: Processing across rows

Week 8, 10/10/2022: Augmented vectors, Survey data, and exploratory data analysis

Week 9, 10/17/2022: Guidelines for investigating, cleaning, and creating variables

Week 10, 10/24/2022: Acquiring data

- Discuss Final Projects

Week 11, 10/31/2022: Tidy Data

Week 12, 11/7/2022: Joining multiple datasets

Week 13, 11/14/2022: Working with Strings and Date/Time Variables

- Final Project Option Selection Due
- If selecting Option 1 (own research project) & Option 2 (online tutorial/workshop), instructor approval prior to selection is needed

Week 14, 11/21/2022: Accessing object elements

Week 15, 11/28/2022: Introduction to looping and Functions

Week 16, 12/5/2022: Last Day of Class

- Students Present on their Final Projects
- Final Project due by 11:59pm 12/13/2022