HED 696C: Data Management and Manipulation in R

The University of Arizona

Spring 2025

Karina Salazar
Assistant Professor
Center for the Study of Higher Education
Email: ksalazar@arizona.edu
Office: COE 313

Class Hours: Mon 7:00 - 9:30pm
Class Modality: Hybrid
Classroom: COE 432/Zoom
Office Hours: Wed 1:00 - 4:00pm; Calendly

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

Course structure consists of weekly asynchronous course materials and weekly synchronous meetings. Each week we will focus on a particular topic (e.g., creating variables; writing functions). For each weekly topic, students will complete a problem set. Problem sets will be completed in groups and focus on practical application of concepts/skills from the topic of the week.

Asynchronous Course Materials. Asynchronous course materials will focus on the topic for that week (e.g., processing across rows). Course materials will consist of three types of resources:

1. Detailed lecture slides (PDF or HTML) with sample code
2. Pre-recorded video lecture of the instructor working through these slides
3. The ".Rmd" file that created the PDF/HTML lecture slides.

  - The .Rmd file will contain all "code chunks" and links to all data utilized in the lecture. Thus, students will "learn by doing" in that they will run R code on their own computer while they work through lecture materials on their own.

Synchronous Meetings. Synchronous class meetings will range from 1-1.5 hours and will vary being in-person and Zoom. Attendance during the entire period is required.

During synchronous class time, we will spend time collectively as a group going over lecture materials and/or working through lecture materials/problem sets in partner groups or on your own. The synchronous workshops are a great time to ask questions about course material or practical applications.

How Get The Most Out of This Class

In just a few words, the keys to success in this class are: **start the material early, work through as much as you can, ask for help, and help others.**

Some general tips:

- Work through weekly asynchronous lecture materials as soon as you can The weekly asynchronous lecture materials (lecture PDF/HTML, lecture .Rmd file with code, video lecture) are the core of this course. Lecture materials are designed for you to run the code on your computer as you work through the lecture. Therefore, treat each lecture as an active learning experience/coding workshop rather than passively reading slides.
- Start the weekly problem set early so that have time to seek help on questions you are struggling with during the synchronous meetings
- If you can't figure something out, ask for help!
  - Discuss with your problem set group
  - Ask a question on D2L
  - Come to office hours
- Be supportive of your classmates; let's create a classroom environment where we all help each other succeed!

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

Hardware

- Please bring in laptop with above software installed each week

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 lecture component or homework assignment question 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 (75% of total grade)
- Final Project (10% of total grade)
- Attendance and participation (15% of total grade)

Weekly Problem Sets (75% of total grade)

Problem sets are due by 11:59PM each Wednesday (two days after synchronous classes). 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 in the asynchronous lecture and build off skills learned in previous weeks. Students can work on problem sets in partners or individually. 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) look over/ attempt the problem set on your own before synchronous meetings; (2) talk/meet with classmates to work through the problem set, with a particular focus on areas group members find challenging.

Final Project (10% of total grade)

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

Students will complete a final project that incorporates some of the skills learned throughout the semester on a "real world" research task. The final project can be completed via two 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; or 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).

We will discuss details of the final assignment in class in advance of the due date, including instructor provided examples and approved tutorials for Option #2. The final project is due on May 13, 2025.

Students will also give a 10-15 minute presentation for their final projects to share with the class new skills learned. Presentations will be scheduled for the last day of class.

Attendance and Participation (15% of total grade)

Students are expected to participate in weekly synchronously class sessions. These sessions will be a lecture and activity based meetings. It requires your *active* participation. Please come to each class session prepared to discuss the asynchronous material, ask questions, and practice coding. If you cannot attend asynchronous sessions for professional, personal, or health reasons, please just let me know ahead of time (if possible).

Course Policies

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

Artificial Intelligence

This course will likely vary considerably from other courses in regards to the use of AI tools.

In principle, you are encouraged to use AI tools to help you with the coding components of this course. This includes using AI-generated code, or code that is based on or derived from AI-generated code, as long as this use is properly documented in the comments of your R script: you need to include the prompt and the significant parts of the response. This course is helping you learn the foundational skills needed to manage and manipulate data in real research settings. One of those skills is learning how to seek out help/resources to complete specific tasks and to debug code. Coders create online communities and open-source programming resources specifically for these reasons. AI tools can provide those same resources such as giving immediate code suggestions, debug code, and help with troubleshooting.

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.

1/20/2025: *MLK Holiday, No Class*

Week 1, 1/27/2025: Course introduction; Getting started with R

- To get everyone started with R, our first class session will meet in-person the full 2.5 hour class session

Week 2, 2/3/2025: Investigating Data Patterns in R

- Synchronous meeting: In-Person

Week 3, 2/10/2025: Investigating Data Patterns in R cont…

- Synchronous meeting: Zoom

Week 4, 2/17/2025: Introduction to Tidyverse: Pipes, dplyr, and Variable Creation

- Synchronous meeting: Zoom

Week 5, 2/24/2025: Tidyverse, Processing Across Rows

- Synchronous meeting: In-Person

Week 6, 3/3/2025: Attributes and Data Class

- Synchronous meeting: Zoom

3/10/2025: Spring Break

- No Class

Week 7, 3/17/2025: Strings and Dates

- Synchronous meeting: Zoom

Week 8, 3/24/2025: Visualizations with ggplot2

- Synchronous meeting: Zoom

Week 9, 3/31/2025: Tidy Data

- Synchronous meeting: In-Person

Week 10, 4/7/2025: Joining Data

- Synchronous meeting: In-Person

Week 11, 4/14/2025: Data Quality

- Synchronous meeting: Zoom

Week 12, 4/21/2025: Introduction to Data Science Techniques; Review Final Projects

- Synchronous meeting: Zoom

Week 13, 4/28/2025: Accessing object elements; Using BibDesk as Citation Manager

- Synchronous meeting: Zoom

Week 14, 5/5/205: Final Projects

- Synchronous meeting: In-Person
- Students Present on their Final Projects
- Final Project due by 11:59pm 12/13/2025