HED 696C: Data Management and Manipulation in R

The University of Arizona

Fall 2020

Karina Salazar
Assistant Professor
Center for The Study of Higher Education
E-mail: ksalazar@email.arizona.edu
Office Hours: Via Zoom by appt

Class Room: COE 311
Class Hours: Thur. 4:15pm - 6:45pm
Class Website: ozanj.github.io/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 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 based on a "real world" research task created by the instructor. 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 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 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
- MikTeK/MacTeX/TinyTex

Hardware

- Please bring in laptop with above software installed each week

Course Website and Resources

Course Website can be found here. 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.

Discussion and Homework Questions

We are using D2L as our class discussion forum where folks 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 code, screenshots, and text 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 (70 percent of total grade)
- Final Project (20 percent of total grade)
- Attendance and participation (10 percent of total grade)

Weekly problem sets (70 percent of total grade)

Problem sets are due by 4:15PM each Thursday (right before the class meeting). Late submissions will not receive points because we will discuss solutions during class. The two lowest grades 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 other students. 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 do 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) after lecture, do the reading associated with that lecture; (2) try doing the problem set on your own; (3) talk/meet with classmates to work through the problem set, with a particular focus on areas group members find challenging.

Final Project (20 percent of total grade)

Students will complete a final project that incorporates many of the skills learned throughout the semester on a "real world" research task created by the instructor. The final project will be a similar format to weekly problem sets but will not provide as detailed "guidance" in how to complete the project.

We will discuss details of the final paper in class on November 19, 2020. After November 19, 2020, 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 10, 2020.

Attendance and Participation (10 percent of total grade)

This course is designated as a "flex in-person: synchronous + zoom". These class sessions proceed as synchronous meetings, with some students in the classroom (we will be meeting remotely until the University notifies us that in-person meetings may commence) and some students attending via Zoom, with instructors presenting content and facilitating in-class discussion. The option to attend weekly class sessions remotely in a synchronous online format (via Zoom) will be available throughout the entirety of the semester.

However, given the uncertainty and complexity of the semester ahead of us, I do understand that you may not be able to synchronously attend class sessions for professional, personal, or health reasons. Thus, you may earn your weekly attendance and participation points one of two ways:

- Attend weekly class sessions and participate in class discussion. You can attend either in-person (if option is available based on University-wide policy and instructor's preference) or via Zoom.
- If you need to skip a class session(s), you can earn your weekly attendance and participation points by posting on the class R-resource discussion board on D2L. R capabilities are infinite, so the purpose of the class R-resource discussion board is to build a repository of resources for class members. Posts can be based on any task, topic, or resource. Posts can build on topics covered in the class or can be based on new topics/capabilities not covered in the class. Posts should explain the topic/capabilities, explain why you think this topic/capability is important and/or interesting, and provide a few "real world" problems or tasks you think this topic/capability would be helpful for. Include links to any external resources. Please submit your post within one week of the missed class session to receive full participation points. See D2L for instructor examples of posts.
    - Weekly synchronous lectures will be recorded via Zoom. You are still responsible for watching lectures on your own time and submitting the weekly problem set on time (one week after the missed class session/prior to next class session). If you are unable to submit the homework assignments after a missed class, incomplete assignments will be dropped as part of the two lowest grades. However, I would encourage you to review

the problem set and solutions missed as lectures and assignments build on skills from previous weeks.

Course Policies

COVID-19

Given the ongoing uncertainty of starting a new academic year in the middle of a global health pandemic, we will practice the following principles:

- Your health and safety are top priority. Please take care of yourselves and your loved ones.
- The course will remain flexible. We will adjust expectations, assignments, and/or objectives if necessary.
- Everyone's circumstances are different and may change throughout the semester. We will make individual accommodations if necessary.
- This is a no judgement and no guilt zone for needing support, flexibility, leniancy. Please extend this same grace to your classmates and your instructor.
- We will prioritize supporting and looking out for each other. This includes wearing masks and social distancing if/when we meet in-person.

If you feel sick, or may have been in contact with someone who is infectious, please stay home and self-quarantine. I also encourage you to self-report via https://health.arizona.edu/SAFER. Per University recommendations https://covid19.arizona.edu/, monitor your symptoms and seek emergency care immediately if your illness is worsening. If seeking medical care, call the doctor's office or emergency room ahead and tell them about your symptoms.

Campus Health is testing for COVID-19. Please call (520) 621-9202 before you visit in person.

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 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 does not't make people feel excluded.

Online Collaboration/Netiquette

You will communicate with instructors and peers virtually through a variety of tools such as discussion forums, email, and web conferencing. The following guidelines will enable everyone in the course to participate and collaborate in a productive, safe environment.

- Be professional, courteous, and respectful as you would in a physical classroom.
- Online communication lacks the nonverbal cues that provide much of the meaning and nuances in face-to-face conversations. Choose your words carefully, phrase your sentences clearly, and stay on topic.

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 and Required Reading

In the below schedule, I lecture on a topic, and then you do the reading about that topic and are required to complete a problem set about that topic. However, if you would prefer to the reading about a topic **prior** to me lecturing about that topic, feel free to do so.

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

Module 1, 8/27/2020: Course introduction; objects in R

- Reading (after class):
  - Wickham and Grolemund 2018 (W&G) 1; W&G 2; W&G 4; W&G 20.1 - 20.3

Module 2, 9/3/2020: Objects in R and Missing Data [continued]

Module 3, 9/10/2020: Introduction to using tidyverse to investigate data patterns

- Problem set due (before class): Yes
- Reading (after class):
  - W&G 5.1 - 5.4

  - Xie, Allaire & Grolemund (XAG) 3.1 LINK HERE
  - Spend 15 minutes studying the "R Markdown Reference Guide" LINK HERE

Module 4, 9/17/2020: Introduction to using "base R" to investigate data patterns

- Problem set due (before class): Yes
- Reading (after class):
  - Wickham, H. (2014). *Advanced R*. Retrieved from https://adv-r.hadley.nz/: 4.1-4.3
  - Nicholls, A., Pugh, R., & Gott, A. (2015). *Sams Teach Yourself R in 24 Hours*. (pg. 236-237) [on D2L]

Module 4, 9/24/2020: Pipes and variable creation

- Problem set due (before class): Yes
- Reading (after class):
  - W&G 5.5 (creating variables)
  - XAG 3.3 (R Markdown, creating PDF documents) LINK HERE
    - note: sections 3.3.5 through 3.3.8 will feel somewhat cryptic and are not required for this course; so just do the best you can with those

Module 5, 10/1/2020: Processing across rows

- Problem set due (before class): Yes
- Reading (after class):
  - W&G 5.6 - 5.7 (grouped summaries and mutates)
  - XAG 4.1 (R Markdown, ioslides presentations) LINK HERE and 4.3 (R Markdown, Beamer presentations) LINK HERE

Module 6, 10/8/2020: Augmented vectors, Survey data, and exploratory data analysis

- Problem set due (before class): Yes
- Reading (after class):
  - W&G 15.1 - 15.2 (factors) [this is like 2-3 pages]
  - [OPTIONAL] W&G 15.3 - 15.5 (remainder of "factors" chapter)
  - [OPTIONAL] W&G 20.6 - 20.7 (attributes and augmented vectors)
  - [OPTIONAL] W&G 10 (tibbles)

Module 7, 10/15/2020: Guidelines for investigating, cleaning, and creating variables

- Problem set due (before class): Yes
- Reading (after class): TBD

Module 8, 10/22/2020: Tidy Data

- Problem set due (before class): Yes
- Reading (after class):
  - W&G chapter 12 (tidy data)
  - [OPTIONAL] Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10), 1-23. doi:10.18637/jss.v059.i10
    - This is the journal article that introduced the data concepts covered in W&G chapter 12 and created the packages related to tidying data
    - Link to article here: LINK

Module 9, 10/29/2020: Joining multiple datasets

- Problem set due (before class): Yes
- Reading (after class): W&G 13

Module 10, 11/5/2019: Acquiring data

- Problem set due (before class): Yes
- Reading (after class): W&G 11

Module 11, 11/12/2020: Working with Strings and Date/Time Variables

- Problem set due (before class): Yes
- Reading (after class): W&G 19

11/19/2020: Introduction to Final Project

- Problem set due: Yes

11/26/2020: No class (Thanksgiving)

- Problem set due: No

Module 12, 12/3/2019: Accessing object elements, Intro to looping and Functions

- Problem set due (before class): Yes
- Reading (after class): W&G 20.4 - 20.5; 21.1 - 21.3; W&G 19

12/10/2020: No class (Reading Day)

- Final Project due by 11:59pm 12/10/2020