# EDUC 263: Introduction to Data Management Using R

Ozan Jaquette

Fall 2018

Instructor: Ozan Jaquette
Pronouns: he/him/his
E-mail: ozanj@ucla.edu
Office Hours: Tues 3-4PM; and by appt
Office: Moore Hall 3038

Teaching Assistant: Patricia Martín
Pronouns: she/her/hers
E-mail: pmarti@g.ucla.edu
Office Hours: Wed 12-1PM; Thur 1:30-2:30PM
Location: Moore Hall 3120 (computer lab)

Class Room: Moore Hall 2120
Class Hours: Fridays 12 - 4 pm
Class Website: ozanj.github.io/rclass/
Class Discussion forum: piazza.com/class/jlo6477nzqo2j0

---

## 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" that you complete in groups of three. These problem sets will account for 80% of your grade for the course. Each week class will begin with one group will leading a discussion of challenges they 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, student 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

Required text

- Grolemund, G., & Wickham, H. (2018). *R for data science*. Retrieved from http://r4ds.had.co.nz/ [FREE!]

## Required Software and Hardware

### Software [FREE!]

PATRICIA - CAN YOU ADD TEXT HERE; PERHAPS PUT A LINK TO THE TO DO LIST?

Please install the following software on your laptop

- R
- RStudio
- MikTeK

### Hardware

- Please bring in laptop with above software installed each week

## Course Website and Resources

PATRICIA - CAN YOU ADD TEXT HERE; ABOUT PIAZZA, ABOUT PIAZZA DISCUSSION FORUMS; ABOUT THE WEBSITE

## Communication with Instructor and Teaching Assistant

Use Piazza discussion forums for all questions related to course content. All students can then benefit from my response. I will aim to respond within 24 hours of your post Monday through Friday and 48 hours on Saturday and Sunday. Email me directly if you have a question regarding any personal issue

I encourage students to answer questions your classmates post on CCLE discussion forums. Writing out explanations to student questions will improve your own knowledge and will benefit your classmates.

## Assignments & Grading

Your final grade will be based on the following components:

- Weekly problem sets (80 percent of total grade)
- Your homework group leads a discussion about how you completed the problem set (10 percent of total grade)
- Attendance and participation (10 percent of total grade)

### Weekly problem sets

Students will complete 10 problem sets. Problem sets are due by 12PM each Friday (right before the class meeting). Late submissions will not receive points because we will discuss solutions during class. The lowest grade will be dropped from the calculation of your final grade.

In general, each problem set will give your 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 will work on problem sets in groups of three people (groups assigned in week 2; same group throughout the semester). However, each student will submit their own assignment. You are encouraged to share ideas and get help from your group. 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. If I find compelling evidence that a student merely copied solutions from a classmate, I will consider this a violation of academic integrity and that student will receive a zero for the homework assignment.

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) meet with your group to work through the problem set, with a particular focus on areas group members find challenging.

### Group led discussion about problem set (10 percent of total grade)

Each week a student group (groups of three) will lead a discussion on the problem set. Groups have a lot of autonomy in how they want to approach this discussion (e.g., can have slides, handouts, or neither). However, my preference is that this is more of an open discussion than a presentation. In terms of topics to cover, the group can invite class members to share what challenges/problems they encountered while completing the homework, how they were able to overcome these problems, alternative ways to overcome data manipulation challenges, and what concepts or tasks remain confusing.

For now, we'll allot 40 minutes (at beginning of class) for these discussions. But this allotted time may increase to 50 minutes or decrease to 30 minutes.

### Attendance and Participation (10 percent of total grade)

Students are required to attend the weekly class meetings. Each unexcused absence results in a loss of 20% from your attendance/participation grade. Three or more unexcused absences will result in a failing grade for the course.

An excused absence is a professional opportunity that you discuss with me beforehand or a medical, or family emergency. Excused absences will not result in a loss of attendance points. However, you will be responsible for all material covered in that class and you will be expected to turn in homework assignments on time.

Students are expected to participate in the weekly class meetings by being attentive, by asking questions, by answering questions posed by classmates and by the professor. In addition to participation during class meetings, students can receive strong participation grades by asking questions and answering questions on Piazza.

# 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 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 creatin 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 and does not disparaging generalizations about a group.

As an instructor, I am responsible for setting an example through my own conduct. I hope to create an environment where students feel comfortable voicing concerns about the classroom environment. I will endeavor to present materials that are respectful of diversity: race, color, ethnicity, gender, age, disability, religious beliefs, political preference, sexual orientation, gender identity, citizenship, or national origin among other personal characteristics. The diversity of student experiences and perspectives is essential to the deepening of knowledge in a course. Any suggestions that you have about other ways to include the value of diversity in this course are welcome. During the quarter, I will also create forums where students can voice concerns anonymously.

## 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.
- It is expected that students may disagree with the research presented or the opinions of their fellow classmates. To disagree is fine but to disparage others' views is unacceptable. All comments should be kept civil and thoughtful.

## Academic accomodations

Students needing academic accommodations based on a disability should contact the Center for Accessible Education (CAE) at (310)825-1501 or in person at Murphy Hall A255. When possible, students should contact the CAE within the first two weeks of the term as reasonable notice is needed to coordinate accommodations. For more information visit www.cae.ucla.edu.

### Academic Honesty:

UCLA is a community of scholars. In this community, all members including faculty, staff and students alike are responsible for maintaining standards of academic honesty. As a student and member of the University community, you are here to get an education and are, therefore, expected to demonstrate integrity in your academic endeavors. You are evaluated on your own merits. Cheating, plagiarism, collaborative work, multiple submissions without the permission of the professor, or other kinds of academic dishonesty are considered unacceptable behavior and will result in formal disciplinary proceedings usually resulting in suspension or dismissal.

## Course Schedule and Required Reading

All reading is required reading. I have worked hard to keep reading load light, focusing only on essentials, because weekly problem sets will be time consuming.

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.

### Lecture 1, 09/28: Course introduction; objects in R

- Reading (after class): Grolemund and Wickham (GW) 1; GW 2; GW 4; GW 20.1 - 20.3

### Lecture 2, 10/05: Investigating data patterns

- Problem set due (before class): Yes
- Reading (after class): GW 5.1 - 5.4

### Lecture 3, 10/12: Variable creation, attributes, factors, and pipes

- Problem set due (before class): Yes
- Reading (after class): GW 5.5; GW 15.1 - 15.2; GW 20.6 - 20.7

### Lecture 4, 10/19: Processing across rows

- Problem set due (before class): Yes
- Reading (after class): GW 5.6 - 5.7

### Lecture 5, 10/26: Survey data and exploratory data analysis (for data quality)

- Problem set due (before class): Yes
- Reading (after class): GW 10 (note: this chapter is about "tibbles", not survey data/EDA)

### Lecture 6, 11/02: Tidy data

- Problem set due (before class): Yes
- Reading (after class): GW 12

### Lecture 7, 11/09: Joining multiple datasets

- Problem set due (before class): Yes
- Reading (after class): GW 13

### Lecture 8, 11/16: Acquiring data

- Problem set due (before class): Yes
- Reading (after class): GW 11

### Thanksgiving, 11/23: No class

### Lecture 9, 11/30: Writing functions

- Problem set due (before class): Yes
- Reading (after class): GW 19

### Lecture 10, 12/07: Accessing object elements and looping

- Problem set due (before class): Yes
- Reading (after class): GW 20.4 - 20.5; 21.1 - 21.3

### Finals Week, 12/14: No class

- Problem set due: Yes