STAT 433 (Fall 2021); UW Madison

Data Science with R

Prerequisite:	STAT 333 or 340	**Credits:**	3.0
Classes:	TTh 4:00-5:15PM	**Room:**	INGRAHAM 22
Professor:	Karl Rohe	**E-mail:**	karlrohe@stat.wisc.edu
Office Hours:	W 12:30-1:30 and by appointment	**Office:**	MSC 6110
Webpage:	http://pages.stat.wisc.edu/∼karlrohe/ds.html		
Zoom office:	https://uwmadison.zoom.us/j/99698146393 (if possible, please keep your camera on)		

TA:	Yongsu Lee	**Office:**	MSC 1245J
E-mail:	yongsulee@stat.wisc.edu		
Office Hours:	Tu 2:50pm-3:50pm. https://uwmadison.zoom.us/j/2293406944 (Appointment-Based only)		

COVID precautions: I'm super excited to be back in person and bummed that we have to wear masks. I have three children at home who are not yet old enough to be vaccinated. My youngest just turned 1 year old. I would feel safer coming to class if you wear a mask during class, zoom to class if you have any symptoms, talk to your doctor about getting vaccinated, and generally take good care of yourself.

Course Objectives: Data Science (DS) is applied statistics in the age of the internet. This has led to two major changes from previous forms of applied statistics: easy sharing of software (e.g. CRAN, github, etc) and easy sharing of data (e.g. API data requests). Data Science (or applied statistics) is an iterative (back and forth) performance of four different types of activities (data collection, data wrangling, data analysis, communication) that require five different types of stances (scientist, coder, mathematician, methodologist, skeptic). In the age of the internet, the skills required to do these steps is rapidly evolving. In class, we will practice the four activities and the five stances using R. We will share software and data. Group projects will synthesize these things into a performance of DS.

The overarching objective is to develop agile and reproducible code to quickly iterate through the pipeline (i.e. the four steps). As you develop your pipeline, you will necessarily iterate forward *and backward* through the pipeline, developing the separate pieces in non-consecutive order. In order to quickly iterate, we need to develop the ability to *think* and *code* in concise/agile syntax. The base R syntax is excessively broad; the tidyverse (which we will learn) and higher levels of programing more generally aim to streamline 80% of the concepts into short syntax. With agile syntax, it will be easy to update code, incorporate new pieces, etc.
You will develop:

- A broad set of computational tools for managing data in R; but not the broadest!

- A broad set of statistical/machine learning tools in R; but not the broadest!

Because 80% of problems are very similar, we will focus on doing these with agility. Zen tip: in "agility" (not "speed"), our aim is to make the software and the coding "transparent." This focuses our concentration on our grand objectives (not on the coding). Analogously, experienced car drivers can navigate complicated directions while driving a car. Due to driving experience, the car has become transparent. Driving the car is like walking or breathing or riding a bike (or "navigating" a familiar path!); it does not require our conscious thought. The coding aspect of data analysis should become similar and agile code is the bridge to get there.

Text:

(r4ds) *R for Data Science* by Garrett Grolemund and Hadley Wickham. Access at r4ds.had.co.nz

(islr) *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. You can download the text at www.StatLearning.com. If you enjoy reading words printed on tree carcass (I do), you can order this book as well. A second edition is coming Summer 2021.

Topics: I am hopeful that we will include all topics (at some resolution): importing data, dplyr, `%>%`, tidying, ggplot2, playgrounds and relational data, shiny, writing a thesis statement, prediction vs inference, unsupervised vs supervised learning, network analysis, PCA, topic modeling, nonlinear methods such as splines and generalized additive models, random forests.

Course Projects: There is one primary project (presentation and written document with a group) and perhaps one smaller (group?) project (shiny app).

Exams: There will be a midterm and a final.

Grading: Projects 41%; in class projects 39%; two exams, each 10%. Over 93.0 is an A. Over 90.0 is at least an AB. Over 85.0 at least a B. Over 80.0 at least a BC. Over 75.0 is at least a C. Over 70.0 is at least a D.

In class projects: This category is a catchall for homeworks, exercises, quizzes, and attendance. Because we will often be "working" in class (i.e. not lecturing), you need to (i) come to class (ii) prepared and (iii) participate. When `R` code is required for handing in, it needs to be integrated into the text. You are expected to use Rmarkdown; it enables you to keep your notes clean while doing the work and then requires little effort to create the "final" document. See `http://rmarkdown.rstudio.com`. Unedited computer output will not be graded. There will be periodic reading quizzes (sometimes announced beforehand) to check that you have properly prepared for class. Each class should have at least one of these things. You are allotted one "miss" for personal reasons.

Academic Honesty: You are permitted, in fact encouraged, to talk to other students, your teaching assistant, or me about homework. However, you may not present other people's work as your own. If you work with other students solving problems, make sure that you write up your own solution independently. It is not acceptable for one student to write a solution for another student to copy. – On exams, your work is to be entirely your own.

Academic integrity and data ethics:

By virtue of enrollment, each student agrees to uphold the high academic standards of the University of Wisconsin-Madison; academic misconduct is behavior that negatively

impacts the integrity of the institution. Cheating, fabrication, plagiarism, unauthorized collaboration, and helping others commit these previously listed acts are examples of misconduct which may result in disciplinary action. Examples of disciplinary action include, but is not limited to, failure on the assignment/course, written reprimand, disciplinary probation, suspension, or expulsion. For detailed information, please see `https://conduct.students.wisc.edu/academic-misconduct/`.

The members of the faculty of the Department of Statistics at UW-Madison uphold the highest ethical standards of teaching, data, and research. They expect their students to uphold the same standards of ethical conduct. Standards of ethical conduct in data analysis and data privacy are detailed on the ASA website, and include:

- Use methodology and data that are relevant and appropriate; without favoritism or prejudice; and in a manner intended to produce valid, interpretable, and reproducible results.
- Be candid about any known or suspected limitations, defects, or biases in the data that may affect the integrity or reliability of the analysis. Obviously, never modify or falsify data.
- Protect the privacy and confidentiality of research subjects and data concerning them, whether obtained from the subjects directly, other persons, or existing records.

By registering for this course, you are implicitly agreeing to conduct yourself with the utmost integrity throughout the semester.

Diversity and inclusion:

Diversity is a source of strength, creativity, and innovation for UW-Madison. We value the contributions of each person and respect the profound ways their identity, culture, background, experience, status, abilities, and opinion enrich the university community. We commit ourselves to the pursuit of excellence in teaching, research, outreach, and diversity as inextricably linked goals.

The University of Wisconsin-Madison fulfills its public mission by creating a welcoming and inclusive community for people from every background–people who as students, faculty, and staff serve Wisconsin and the world. `https://diversity.wisc.edu/`

Accommodations for students with disabilities:

The University of Wisconsin-Madison supports the right of all enrolled students to a full and equal educational opportunity. The Americans with Disabilities Act (ADA), Wisconsin State Statute (36.12), and UW-Madison policy (Faculty Document 1071) require that students with disabilities be reasonably accommodated in instruction and campus life. Reasonable accommodations for students with disabilities is a shared faculty and student responsibility. Students are expected to inform me of their need for instructional accommodations by the end of the third week of the semester, or as soon as possible after a disability has been incurred or recognized. I will work either directly with the student or in coordination with the McBurney Center to identify and provide reasonable instructional accommodations. Disability information, including instructional accommodations as part of a student's educational record, is confidential and protected under FERPA.

Complaints:

If you have a complaint about a TA or course instructor, you should feel free to discuss the matter directly with the TA or instructor. If the complaint is about the TA and you do not feel comfortable discussing it with him or her, you should discuss it with the course instructor. Complaints about mistakes in grading should be resolved with the instructor in the great majority of cases. If the complaint is about the instructor (other than ordinary grading questions) and you do not feel comfortable discussing it with him or her, contact the Director of Undergraduate Studies, Professor Cecile Ane, `cecile.ane@wisc.edu`.

If your complaint concerns sexual harassment, please see campus resources listed at `https://compliance.wisc.edu/titleix/resources/`. In particular, there are a number of options to speak to someone confidentially.

If you have concerns about climate or bias in this class, or if you wish to report an incident of bias or hate that has occurred in any statistics class, you may contact the Chair of the Statistics Department Climate & Diversity Committee, Professor Karl Rohe (`karlrohe@stat.wisc.edu`). If you would prefer someone who is not the instructor for this course, you may contact anyone on the Statistics Department Climate & Diversity Committee (Jessi Cisewski Kehe `jjkehe@wisc.edu`, Vivak Patel `vivak.patel@wisc.edu`, or Kris Sankaran `ksankaran@wisc.edu`). You may also use the University's bias incident reporting system, which you can reach at `https://doso.students.wisc.edu/services/bias-reporting-process/`.

University level rules, rights, and responsibilities for students:

See: `https://guide.wisc.edu/undergraduate/#rulesrightsandresponsibilitiestext`.