S&DS 361: Data Analysis

Spring 2024 Syllabus

Instructor: Brian Macdonald

- Email: brian.macdonald@yale.edu
- Office: Kline 1141 (turn left out of the elevator)
- Course Meeting: Tue/Thu, 9:00-10:15am, Location TBD
- Office hours: Monday 1:30-2:30 on Zoom and Thu 10:30-11:30, in my office.
- Teaching Fellows, ULAs, and course manager, and office hours:
  - TF: TBD
  - TF: TBD
  - ULA: TBD
  - ULA: TBD
  - ULA: TBD
  - ULA: TBD
  - Course Manager: TBD
- Midterms: TBD
- Final exam: TBD
- Schedule: See shared spreadsheet

Course Objectives

- You will gain experience and comfort with working with data. You will explore, visualize and analyze data with R, and learn best practices for writing code that is organized and commented, and executes fully reproducible analysis. You will also gain experience producing production-ready code able to be run in automated overnight processes, as well as other tools that are useful for many industry jobs and that prepare you for S&DS 425 Statistics Case Studies and/or S&DS 491/492 Senior Project.
- You will gain experience exploring and visualizing data, including what questions you want to ask about the data, what aspects of the data you want to highlight in your visualizations, and how to implement your ideas in `R`. We'll focus on best practices in planning visualizations, sketching them by hand, and creating them using `ggplot`, `plotly`, `leaflet`, `gganimate`, and other visualization packages. We'll create static, interactive, and animated visualizations and we'll develop interactive web apps using `shiny`.
- You will learn data analysis techniques like linear, logistic and poisson regression, other generalized linear models, mixed effects regression, regularization methods, maximum likelihood, splines, Monte Carlo simulation, resampling methods, model selection, model diagnostics. We will focus on getting hands-on experience with determining what question we want to ask about the data, brainstorming the most appropriate approach(es) to answering the question, implementing those approaches in R, and assessing and interpreting the results of the analysis. We'll build the theoretical understanding that is necessary to appropriately apply these techniques as well.

In short, you will gain experience in all aspects of the data analysis workflow, including

1. Defining the problem, question, or goal

2. Choosing data, acquiring data, assessing the quality of that data
3. Cleaning and wrangling data
4. Exploring and visualizing data
5. Analyzing data and building predictive models
6. Interpreting, visualizing, and communicating the results
7. Developing recommended courses-of-action

although we'll focus more on #1, #4, #5, and #6 than on the others.

Prerequisites

This course assumes that you are comfortable with R, introductory to probability and statistics, theory of statistics, linear algebra, and calculus. None of the above are a 100% must, on the other hand, we will be relying on all of them and it might be difficult to keep up with this course if you haven't taken the appropriate prerequisite courses.

Here are some more details about specific topics you'll want to have experience with.

- **R**. R will be used very heavily in this course. We will assume you know the basics of programming in R and have had a course that uses R extensively (e.g. 106, 220, 230, etc.) or similar hands-on experience outside of the classroom.
- **Probability and statistics**. A course like S&DS 238 or S&DS 241 that covers discrete and continuous distributions and maximum likelihood estimation. At minimum, you should be comfortable with introductory-level probability and statistics topics like with discrete and continuous random variables, probability density functions and probability mass functions, expected value, variance, covariance, correlation, the central limit theorem, Bayes theorem.
- **Linear regression**. We will be reviewing linear regression but we will go fast and will not assume that this is the first time you have seen it.
- **Linear algebra**. We will assume you have taken linear algebra and have seen the matrix form of least squares regression, among other topics.

Preparing for the start of class

Computing. The most important thing is the installation of the newest versions of R and RStudio IDE. You will also want LATEX (MacTeX on the Mac and MikTeX on Windows). For Windows, make sure you get the complete distribution of MikTeX the basic version will not work. The `tinytex` package is another option if you have issues with other versions of LaTeX.

Please also install the packages listed in the first assignment. We will use those throughout the course. We may end up needing other packages too, which we'll install on the fly. But we will definitely use the packages at that link, so take the time to download them now.

Books.

We'll be using our course notes for most of the class but we'll also use the following books as references. All of them have free PDF or HTML versions online at the links below. If you'd like, you can order a hard copy of any of them as well.

- Beyond Multiple Linear Regression by Roback and Legler. This will act as our primary book for a lot of the semester.
- Regression and Other Stories by Gelman, Hill, and Vehtari. We'll use this as a supplement because it has more details on certain topics.
- An Introduction to Statistical Learning, 2nd edition, by James, Witten, Hastie, Tibshirani. This has some material not in the other books.

Assignments, Exams

There will be (roughly) weekly problem sets that focus on both theory and application of various methods of analyzing data. There will be two midterms and one final. There may be an occasional quiz on reading material or other topics. We will have in-class exercises and discussions in small groups.

- Problem Sets: 20%
- Midterms: 22.5% each
- Final: 35%

Submitting assignments. You will submit assignments on Gradescope. The pull request assignment from the first problem set is the only thing you will submit on GitHub. On Gradescope there is a place to submit your PDF only, and a place to submit other files (usually just an .Rmd file, but sometimes we'll have other files).

Late Policy on Problem Sets. You can submit 2 Problem Sets up to 48 hours late, no questions asked. Other than that, you need a dean's extension to submit late without penalty. Without a dean's extension, there is a 10% penalty per day.

If you are unable to submit the assignment on Gradescope (because e.g. the deadline has passed), please ask the course manager to open the submission back up for you. You can email the assignment to the course manager as evidence of when you completed the assignment, but you will eventually still need to upload it to Gradescope. You assignment will not be graded until it is uploaded to Gradescope.

Dropping Lowest Grades. I will drop the lowest grade from among problems sets that you made a *reasonable attempt* at completing. A reasonable attempt is one where you tried every problem, completed part of every problem, but got stuck somewhere on a couple of the problems and did them incorrectly. If you do not make a reasonable attempt, the grade will not be dropped. Examples:

Will not be dropped: - If you do not turn in a problem set and get 0%, that grade will not be dropped. - If you complete only half of the problem set and get a 50%, that grade will not be dropped.

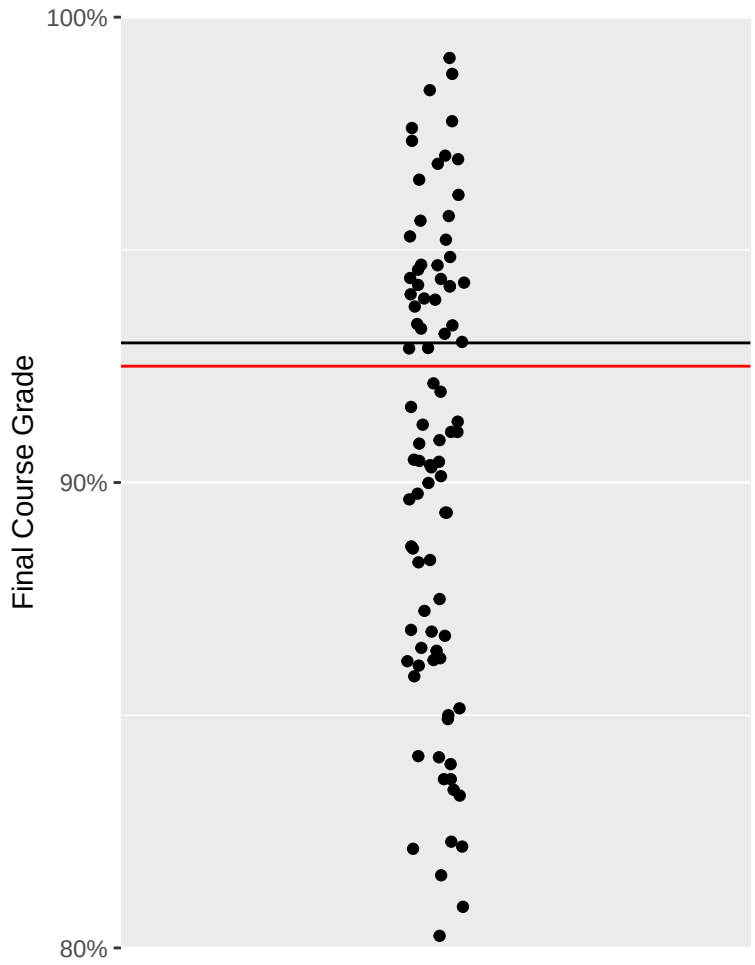
Will be dropped: - If you made a reasonable attempt at every problem but had trouble with a couple of them, and ended up getting a 82%, and that is your lowest grade on a problem set, that grade will be dropped.

Curving Final Course Grades. I do not plan on curving final course grades. I may lower some cutoff lines based on where there is a gap in grades. For example, in the situation below I would lower the cutoff for an A from 93 (black line) to 92.5 (red line) because a few students are clustered around 93 and there is a gap in the grades around 92.5.

```
library(ggplot2)
library(scales)
set.seed(1)
df = data.frame(x=rep(0,100),
                y=rbeta(100, shape1=9*2, shape2=1*2))
df = df[!(df$y<.928 &  df$y> .922),]

ggplot(df, aes(x=x, y=y))+
  geom_jitter(width=0.1)+
  labs(y='Final Course Grade',
       x=NULL)+
  scale_y_continuous(breaks=c(.8, .9,  1),
                     limits=c(0.8, 1),
                     expand=c(0,0),
                     labels=percent)+
```

```
  scale_x_continuous(breaks=NULL, limits=c(-1,1))+
  geom_hline(yintercept=.93)+
  geom_hline(yintercept = .925, color='red')
```



Expectations

Class time will be a mix of live-coding in R, in-class small group activities, interactive lecturing and work on the board, and possibly other types of activities from time to time. You are not necessarily expected to ask/answer questions in front of the whole class (though you are encouraged to do so, the more interaction the merrier!), but you are expected to participate in small group discussions/activities and smaller discussions with me during class.

Recordings. Some classes might be recorded, or at least partly recorded. Students who have valid excused absences can request access to the recording of the class they missed. In-class activities involving individual or small group work or small group discussions will not be possible to record. Please contact our course manager to request access to course recordings with an excused absence.

Office hours. You are expected to attempt problems before seeking help in office hours. When you ask for help, clearly state what you have attempted, where you got stuck, and what specific question you have.

Ed Discussion and GitHub Issues. You will be able to ask questions on Ed Discussion and GitHub issues. The same guidelines as office hours apply here: You are expected to attempt problems before seeking

help, and when you ask for help, clearly state what you have attempted, where you got stuck, and what specific question you have. Use minimial reproducible examples when possible.

Academic Integrity

> Academic integrity is a core institutional value at Yale. It means, among other things, truth in presentation, diligence and precision in citing works and ideas we have used, and acknowledging our collaborations with others. In view of our commitment to maintaining the highest standards of academic integrity, the Graduate School Code of Conduct specifically prohibits the following forms of behavior: cheating on examinations, problem sets and all other forms of assessment; falsification and/or fabrication of data; plagiarism, that is, the failure in a dissertation, essay or other written exercise to acknowledge ideas, research, or language taken from others; and multiple submission of the same work without obtaining explicit written permission from both instructors before the material is submitted.

The statement above was provided by the Yale Graduate School. In this class, we have the same expectations of undergraduate students. We encourage students to work together on most everything unless noted otherwise. But "working together" is difficult to define when code (programming) is involved. Many students can benefit from constructive collaborations. However, if you benefit from "working together" then you need to be able to explain and discuss your solution should questions arise. "I don't remember how I did this" would not be sufficient.

Specific requirement: any collaboration on homework must be acknowledged up front. Something like "Lastname, Firstname - Homework 1. Worked with John S. and Jane D." would be fine. This is very simple and encourages constructive collaboration if it aids the learning process. But not acknowledging such a collaboration is a considered a violation of academic integrity which we are obligated to report to Yale College rather than arbitrate ourselves.

Diversity, Equity, Inclusion, and Academic freedom

This class strives to be an inclusive learning community. We can all learn from perspectives of students and faculty with difference experiences and backgrounds. Aspects of data science can be subjective (e.g. the choice of the most appropriate question to ask, interpreting and communicating results, etc), and it is possible that some material may have overt or covert biases. A community that promotes open communication, academic curiosity, and freedom of expression and values a diverse set of experiences and backgrounds is essential for a comprehensive understanding of all aspects of a data science problem and possible solutions. I expect students to commit, along with me, to creating and maintaining such an environment, including time both in and out of the classroom.

Accessibility

I am committed to creating a course that is inclusive in its design. If you encounter barriers to learning or inclusion in this course, please let me know as soon as possible. Together we can develop strategies that can enable you to succeed in the course. Students are also encouraged to contact Student Accessibility Services to discuss a range of options to removing barriers in the course, including official accommodations.

Academic Support Resources

See this page for details on resources available to different student populations.

- Writing Tutoring (Yale College)

- Graduate Writing Lab (Yale Graduate Students)
- STEM Tutoring (Yale College)
- Language Tutoring (All Students)
- Academic Strategies Program (Yale College)
- Student Accessibility Services (All students)

Mental Health and Wellness Resources are summarized further down that page.

- Yale College Community Care (YC3)
- College Care Clinicians (Yale College)
- Community Wellness Specialists (Yale College)
- Yale Well (All Students)
- Yale Mental Health and Counseling (All students)