Gov 1005: Data

David Kane

Fall 2018: M/W 1:30 – 2:45

Description

Data matters. How much money is spent on US political campaigns? How does the Chinese government use social media? How many seats will the Democrats control in the Senate after the next election? How does Harvard College decide whom to admit? We need data to answer these questions.

This course will teach you how to work with data, how to gather information from a variety of sources and in various formats, how to import that information into a project, how to tidy and transform the variables and observations, how to visualize and model the data for both analysis and prediction, and how to communicate your findings in a sophisticated fashion. Each student will complete a final project, the first entry in their professional portfolio. Our main focus is data associated with political science, but we will also use examples from education, economics, public health, sociology, sports, finance, climate and any other subject area which students find interesting.

We use the R programming language, RStudio, GitHub and DataCamp. Although we will learn how to program, this is not a course in computer science. Although we will learn how to find patterns in data, this is not a course in statistics. We focus on practice, not theory. We perform empirical analysis rather than write mathematical derivations. We make stuff.

Prerequisites: None. You must have a laptop with R, RStudio and Git installed.

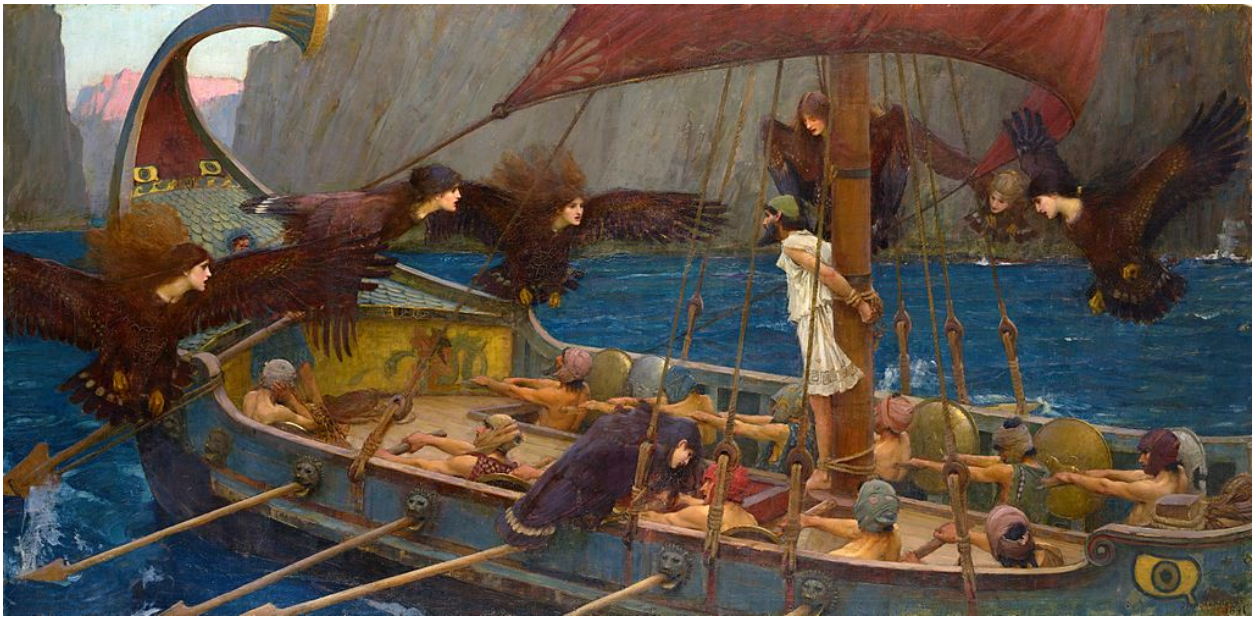

Figure 1: *Ulysses and the Sirens*, 1891, by John William Waterhouse. "This dramatic painting illustrates an episode from the journeys of the Greek hero Odysseus (in Latin, Ulysses) told in the poet Homer's Odyssey in which the infamous Sirens lured unwary sailors towards perilous rocks and their doom by singing in the most enchanting manner. Odysseus wished to hear the Siren's song and ordered his crew to lash him to a mast and block their ears in order to ensure their safe passage. Waterhouse has depicted each Siren with the body of a bird and the head of a beautiful woman, which came as a surprise to Victorian audiences, who were more used to seeing these mythic creatures portrayed as comely mermaid-like nymphs. He borrowed the motif from an ancient Greek vase that he studied in the British Museum." The next stop on Odysseus's journey was Thrinacia.

Course Metaphor

The central metaphor for this class is Ulysses and the Sirens. You are Ulysses. Thrinacia is a desirable internship/job next summer. The Sirens are the many distractions of the modern world. I am the rope.

Course Philosophy

- *No Lectures*: The worst method for transmitting information from my head to yours is for me to lecture you. There are (almost) no lectures. We work on problems together during class meetings. There are no sections. Instead, the Teaching Fellows work with individuals and small groups during Study Halls.
- *No Math*: Our focus is on practical skills for working with data. To make time for topics like git, we need to cut out material that might typically be included in a course like this. The biggest impact relates to (the lack of) math, probability and statistical theory. Fortunately, there are many Harvard courses which cover this material: GOV 50, GOV 1000, STAT 104, et cetera.
- *No Cost*: Every tool we use and reading I assign is available for free. You don't have to spend any money on this class. Some activities, like DataCamp and GitHub, have paid options which provide more services, but you never have to use them. Don't give anyone your credit card number.
- *R Everyday*: Learning a new programming language is like learning a new human language: You should practice (almost) every day.

- *Cold Calling*: I call on students during class. This keeps every student involved, makes for a more lively

discussion and helps to prepare students for the real world, where there will be no hiding in the back row.

- *Visitors*: We will have a variety of visitors in class, people performing professional data analysis, both inside and outside of academia, often using exactly the same tools that we use. If there is someone you would like to meet, talk to me about it and we can invite them!
- *Class Activities*: Awkwardness in the pursuit of learning is no vice. We will do a variety of class activities that will sometimes take you out of your comfort zones. You will meet and work with many more of your fellow Harvard students than you would in a normal class.
- *Professionalism*: We use professional tools in a professional fashion. Your workflow will be very similar to the workflow involved in paid employment. Your final project will be public, the better to interest employers in your abilities.

Course Staff

- Preceptor David Kane (dkane@fas.harvard.edu) CGIS South 310
- Teaching Fellow Albert Rivero (arivero@g.harvard.edu)
- Teaching Fellow Nick Short (nshort@g.harvard.edu)

Course Policies

Workload: The course should take about 10 hours a week, outside of class meetings, exams and the final project. This is an expected average across the class as a whole. It is not a maximum. Some students will end up spending much less time on the class. Others will spend more.

Use your Harvard e-mail: To the greatest extent possible, please use your official Harvard e-mail address for all aspects of this class, especially things like signing up for services like DataCamp, GitHub, Piazza and so on. Doing so makes it much easier for us to figure out who is doing what. This may not be easy if you already have an account with these services but, even in that case, you should be able to add your Harvard e-mail address to your account.

Piazza: All general questions — those not of a personal nature — should be posted to Piazza so that all students can benefit from both the question and the answer(s). Check out the class homepage and signup link.

Laptops: You must bring a laptop (or the equivalent) to class.

Plagiarism: If you plagiarize, you will fail the course. See the Harvard College Handbook for Students for details. Discussing ideas and work-in-progress with others is an important and desirable part of the research process, but in the end, a student's assignment must be their own effort, written by the student, and ultimately based on their own thinking. All written assignments must use appropriate citation practices.

Working with Others: Students are free (and encouraged) to discuss problem sets and their final projects with one another. However, you must hand in your own unique code and written work in all cases. Any copy/paste of another's work is plagiarism. In other words, you can work with your friend, sitting side-by-side and going through the problem set question-by-question, but you must **each type your own code**. Your answers may be similar (obviously) but they will not be identical.

R: You must use R and RStudio for this class. You are responsible for installing both on your laptop.

Git and GitHub: Analyzing data without using source control is like writing an essay without using a word processor — possible but not professional. Starting in week 3, we will do all our work using Git/GitHub. If Git is not already installed on your computer, please install it.

DataCamp: We make extensive use of lessons from DataCamp. All DataCamp courses are graded pass/fail. Each week's course(s) are due by Monday at 10:05 AM (except in the cases of holidays or exams). Class on Monday will assume the completion of this work.

R for Data Science (R4DS): Reading assignments from R4DS in a given week cover material that we will use that week. Some students prefer to do such readings ahead of time, the better to prepare for class. Some students prefer to do the readings after those classes, the better to reinforce the material. Some students prefer to never do the readings. No matter what path you select, know that, when constructing/grading the problem sets, exams and final projects, we will assume that you understand the material in R4DS. If you are struggling in the class, the best advice we can offer is to read R4DS cover-to-cover.

Optional Activities: The syllabus includes background readings and DataCamp assignments which students may find interesting. These are all optional, although I will sometimes refer to them during class discussions.

Computer Problems: If you are having problems with your computer, follow these steps. First, post the problem on Piazza, with details and screenshots. With luck, a fellow student will be able to solve it. (And students who help their peers with technical issues are guaranteed full participation points for grading purposes.) Second, if I and/or the TFs can't solve it, we will direct you toward the IQSS IT Client Support Services, located in the basement of CGIS Knafel. They are excellent! E-mail them with the details of your problem, mentioning your enrollment in this class, at help@iq.harvard.edu. Although I and the teaching fellows want to be helpful, we are not experts in troubleshooting computer problems. Third, once your problem is solved, tell us all the solution by responding to your own post on Piazza.

Pass/Fail: You may take this class pass/fail. Note, however, that to pass the course you must pass all segments: DataCamp, problem sets, midterms and the final project. You can not just do well on the first three, skip the final project at the end, and assume that you have enough points overall to pass.

Study Hall: We run three weekly study halls, all located in the Fisher Commons on the first floor of CGIS Knafel: Tuesday 9:00 to 12:00 (Nick), Tuesday 2:00 to 5:00 (Albert) and Thursday 8:30 to 11:30 (Preceptor). All are welcome! **This is 100% optional.** No attendance is taken. The Thursday session is part of a broader effort to provide help for all students using R in their Government classes, so you may also see guests from GOV 50, GOV 2000 and other courses. The Tuesday sessions are for GOV 1005 students only.

Missing Class: You expect me to be present for lecture. I expect the same of you. There is nothing more embarrassing, for both us, than for me to call your name and have you not be there to answer. But, at the same time, conflicts arise. It is never a problem to miss class if, for example, you have an interview or are out of town or have a health issue. Just e-mail me and the teaching fellows (all of us!) on the morning of the class you will miss. You do not need to ask for permission. You have it! But we need to be informed on that day, not weeks before or days after.

Late Days: Many assignments allow for late days. An assignment is a day late if it is turned in anytime after it was due (even 5 minutes after) but within 24 hours. After that, it is two days late, and so on. Each late day after the total allowed results in a 1 point penalty. These penalties accumulate and can result in a negative contribution to your final grade. **You should save your late days.** If you use them early in the semester for no particularly good reason and then, later in the semester, have an actual emergency, we will not be sympathetic. In general, we will not give you extra late days in such a situation. (That isn't fair to your classmates, and we are all about fairness.) We will just, mentally, move the late days you wasted so that they cover your actually emergency. You will now be penalized for being late earlier in the semester, when you did not have a good reason for tardiness.

Major Emergencies: We are not monsters. If you are hit with a major emergency — the sort of thing that necessitates the involvement of your Resident Dean — we will be sympathetic. Just put us (Preceptor and teaching fellows) in touch with your Resident Dean and we will work out something.

Role of Teaching Fellows: The TFs run all aspects of grading for the course, keeping track of late days, dealing with emergencies and so on. Go to them first with any problems. (Feel free to cc the Preceptor if you want to keep me in the loop but I am very respectful of TF authority on these matters.)

Computer Emergencies: We are very unsympathetic to computer emergencies. You should keep all your work

on GitHub, so it won't matter if your computer explodes. If it does explode, you will lose only the work after your last commit. You can then restart your work on a public computer (the basement of CGIS Knagel has machines with R/RStudio installed) or on your roommate's computer.

Github Classroom: We use Github Classroom to distribute problem sets and midterms. You will receive an e-mail with a link. Click on that link and a repo, with instructions, will be created. *Do this as soon as you receive the e-mail.* We don't want git problems to arise the night before the assignment is due.

Grading

Participation: 5 points. I expect you to participate, both in class and online. Helping your fellow students, especially on Piazza, is the best form of participation, as is volunteering for a Welcoming Committee for a speaker. Be a good class citizen.

DataCamp Lessons: 10 points. Grades are pass/fail only. **These are free points!** Given the level of the questions and the hints provided, it is essentially impossible not to get full credit as long as you make an honest effort. You have 5 free late days to turn in assignments late. That is, you can turn in 5 individual courses up to one day late or 1 course up to five days late or whatever. I recommend you "save" these late days for the end of the semester, when things get busy.

Problem Sets: 15 points. Problem sets are distributed after class on Wednesday and then due the following Wednesday at 10:05 AM. You are welcome to work on them with your friends but, first, you must personally type in every character in the work you submit and, second, you must list all the people you worked with. You have 2 free late days for problem sets. After those 2 late days, each additional late date counts as -1 point of the 15 points total for problem sets. Take more than 17 lates days, and further late days will make a negative contribution to your final grade.

Midterms: 20 points each. Midterms are take-home. They are open-book and open-web. Because students have different schedules, you can complete the midterm any time within a four-day window starting after midterm distribution. Late midterms will not be accepted.

Final Project: 30 points. Students will present their projects publicly during the last week of classes. They will then have the opportunity to incorporate feedback. The final version of the project is due at 10:05 PM on December 13.

Final Project

What topic in all the world do you find most interesting? As long as you can find data for that topic, use it for your final project. Do you love soccer or wine or NYC politics? The final project provides you with an opportunity to study that topic in depth. Your goal is to gather data and present it in an engaging fashion. We are not necessarily investigating specific hypotheses or trying to fit a statistical model, although you can do those things if you want. Instead, imagine that a friend of yours also cares about soccer/wine/politics/whatever. You are building something that would interest her, something that will make her say, "That is cool! Let's spend 30 minutes poking around with your data."

Your final project will be, for most of you, the first item in your professional portfolio, something so impressive that you will be eager to show it to potential employers. You must show this work publicly, both on the web (viewable by all) and in person at our Demo Day during the last week of classes. You will host your final project using ShinyApps, a free service provided by RStudio. Or, more ambitiously, you can create a blogdown site. Consider these suggestions for organizing your work. Make use of free statistical consulting from the Harvard Statistics Department.

Key Dates:

- Friday, October 19 at 10:05 PM. Precis are due. This is a one page html document which exists in a public repo which you have created from an Rmd file which is also in the repo. (You need only email Nick a copy of the html and a link to the repo.) It need only include a few sentences about the data that you plan to use along with perhaps one summary statistic — like the number of obseravtions — that is generated from that data. In other words, there must be a working Rmd document which is also in your repo and from which you generate the html. If you don't yet have access to actual data, you may just use some random data, even one of the csv files from the homeworks. (We do want to check that you know how to add data to a repo and push it to GitHub.) One point.
- Friday, November 2 at 10:05 PM. Very rough Shiny App (or other presentation mode) are due. Create a Shiny App which displays some of your data. And that is it! You still have another month to make this professional looking. The purpose is just to ensure that you are on track. But you must use real data for this. Again, e-mail Nick a link to confirm completion. Two points.
- Week of November 26. Make an appointment with a member of the course staff to discuss to your project and get suggestions. You must have a working Shiny App (or some other presentation medium) using real data and a plan for what else you will build. Your Shiny App must include a description of your data source and a link to the underlying code on GitHub. For these first three checkpoints, it is almost impossible not to get full points. We aren't evaluating anything yet. We are just ensuring that you are on track. Two points.
- Demo Day on December 5. Show off your cool work to the world! You will be spend 40 minutes presenting your work and 40 minutes looking at your classmates' presentations. Nick will provide organizational details. Ten points.
- December 13 at 10:05 PM. Final version. E-mail Nick with a link to your Shiny App (or to whatever webpage you have created to host your work) and a link to the GitHub repo which underlies it. Both must be public, unless you have discussed the issue with us ahead of time. Fifteen points.

You have a total of three late days that you may use for any of these deadlines, except for Demo Day. Use them wisely.

Resources

The text for the class is *R for Data Science* (R4DS) by Garrett Grolemund and Hadley Wickham. The primary resources below are also useful, but are not required reading. The secondary resources may also be helpful. All are free.

Primary

ModernDive: An Introduction to Statistical and Data Sciences via R by Chester Ismay and Albert Y. Kim
Happy Git and GitHub for the useR by Jenny Bryan
The Unix Workbench by Sean Kross
R Markdown: The Definitive Guide by Yihui Xie, J. J. Allaire, Garrett Grolemund
Data Visualization: A practical introduction by Kieran Healy

Secondary

Fundamentals of Data Visualization by Claus O. Wilke
Efficient R Programming by Colin Gillespie and Robin Lovelace
Handling Strings with R by Gaston Sanchez
Text Mining with R: A Tidy Approach by Julia Silge and David Robinson

Pro Git by Scott Chacon and Ben Straub
Hands-On Programming with R by Garrett Grolemund

Conclusion

If you had tried to complete a data analysis project before taking this class, you would have done X well. Now that you have taken the class – now that you have learned how gather information in various formats, how to import that information into a project, how to tidy and transform the variables and observations, how to visualize and model the data for both analysis and prediction, and how to communicate your findings in a sophisticated fashion – you will do Y well. The success (or failure) of the class can be measured by comparing Y with X.

Schedule

Rhythm of the Class

The class follows a steady weekly rhythm:

- Monday 10:05 AM. DataCamp exercises due, except for extensions because of holidays or midterms.

- Monday 1:30 PM – 2:45 PM. Class. Main focus of class will be interactive R session using material from DataCamp exercises you have just completed.
- Tuesday 9:00 AM to 12:00 PM. Study Hall in Fisher with Nick.
- Tuesday 2:00 PM to 5:00 PM. Study Hall in Fisher with Albert.
- Wednesday 10:05 AM. Problem set (distributed last Wednesday) due.

- Wednesday 1:30 PM – 2:45 PM. Class. In addition to continuing with the new R commands from Monday's class, we will review material from the previous problem set (or exam, if one was given last week). By "previous problem set," I mean the one you turned in last week, not the one you turned in that day. Other students who have taken a late day (or two) might still be working on that most recent problem set.

- Wednesday 4:00 PM. Problem set (or take-home midterm) distributed.

- Thursday 8:30 AM to 11:30 PM. Study Hall in Fisher with Preceptor.

- Friday 10:05 PM. Interim steps in the final project are due.
- Sunday 10:05 PM. Midterm exams, if distributed on Wednesday, are due.

Week 1: September 5. Introduction and Graphics

R4DS: Chapters 1, 2 and 3

Install R, RStudio and Git on your machine. You would also be wise to start on the DataCamp assignments which are due on Monday, September 10.

Optional

- How Obama's Team Used Big Data to Rally Voters by Sasha Issenberg

- An Extremely Detailed Map of the 2016 Presidential Election
- The Left Side of Steve Kerr's Brain by Marc Stein

Week 2: September 10 and 12. Working with R and RStudio.

DataCamp

Remember: DataCamp assignments are due Monday at 10:05 AM. If you have already decided to take the class, then these assignments are due Monday, September 10. Obviously, if you join the class at the end of shopping period, you have an extra time to finish these. But I will assume, in week 2, that students have completed them. The basic structure of the class is to learn skills on your own by Monday and then work on new projects together in class with those skills.

- Introduction to the Tidyverse (four hours)
- Data Visualization with ggplot2 (Part 1) Only do Chapters 1 through 4. (four hours)
- Working with the RStudio IDE (Part 1) Only do Chapter 1 Orientation this week. (one hour)

R4DS: Chapters 4, 6, 8, 26 and 27

Optional

- ModernDive: Chapters 1, 2 and 3

Week 3: September 17 and 19. Data Transformations

DataCamp

Remember: DataCamp assignments are due Monday at 10:05 AM.

- Working with the RStudio IDE (Part 1) — Chapters 2 and 3. (two hours)
- Data Manipulation in R with dplyr (four hours)
- Reporting with R Markdown (three hours)

R4DS: Chapter 5

Optional

- ModernDive: Chapter 4 and 5
- *Visual and Statistical Thinking: Displays of Evidence for Making Decisions* by Edward Tufte

Week 4: September 24 and 26: The Shell, Git, and GitHub

DataCamp

- Introduction to Shell for Data Science (four hours)

- Introduction to Git for Data Science (four hours)
- Working with the RStudio IDE (Part 2) — Only do Chapter 2 Version Control. (one hour)

The Unix Workbench, chapters 1 – 6. GitHub Classroom Guide for Students

Problem Set #1 due September 26 at 10:05 AM covering material through Week 3. This problem set will be distributed, collected and graded (pass/fail) "by hand." It should take less than one hour.

Speakers

- September 24: Hugh Truslow, Data Sciences and Visualization, Harvard Library
- September 26: Rafael Irizarry, Professor of Biostatistics, Harvard T.H. Chan School of Public Health

Optional

- *Happy Git and GitHub for the useR* by Jenny Bryan. Extremely useful as a reference.
- Excuse me, do you have a moment to talk about version control?
- A Quick Introduction to Version Control with Git and GitHub
- Consider installing SourceTree to examine the details of your git repository. This is probably overkill for what we have done so far but might prove useful when you starting working on group problem sets.

Week 5: October 1 and 3. Exploratory Data Analysis (EDA)

DataCamp

- Working with Data in the Tidyverse (four hours)
- Working with Web Data in R (four hours)

R4DS: Chapters 7 and 20

Problem Set #2 due October 3 at 10:05 AM covering material through Week 4. This problem set will be distributed, collected and graded using GitHub. It should take about 2 hours.

Speakers

- October 3: Cesar Hidalgo, Director, Collective Learning group at The MIT Media Lab

Optional

- Chapter 44 Web Scraping from *Introduction to Data Science* by Rafael A. Irizarry
- Exploratory Data Analysis (four hours)
- Exploratory Data Analysis in R: Case Study (four hours)

Week 6: October 10. Relationships

No class on Monday, October 8.

DataCamp

Because of the holiday, DataCamp exercises are not due until Tuesday, October 9 at 10:05 AM.

- Correlation and Regression (four hours)
- Building Web Applications in R with Shiny (four hours) If you have zero experience with programming, you may find it useful to read the chapter on functions from *R for Data Science* before starting this class. You also might check out these written Shiny tutorials. If you are finding Shiny a bit overwhelming, you have the option of just doing Chapter 1 of this course.

R4DS: Chapters 22, 23, 24 and 25

Problem Set #3 due October 10 at 10:05 AM covering material through Week 4. It should take about 2 hours.

First midterm distributed October 10 after class and due Sunday October 14 at 10:05 PM. Focus will be on tidyverse commands and graphics.

Speakers

- October 10: Natalia Urtubey, Executive Director, Imagine Boston 2030

Optional

- ModernDive: Chapters 6 and 7
- Causality, Chapter 2 of *Quantitive Social Science* by Kosuke Imai
- The Data Analytics & Technology Fair on Friday, October 12 from 1-4pm is a great chance to connect with 60+ for-profit, non-profit, and government organizations from across the data science and tech spaces. The current list of registered organizations can be found here.

Week 7: October 15 and 17. Tidy and Relational Data

DataCamp

Because of the midterm, DataCamp exercises are not due until Wednesday, October 17 at 10:05 AM.

- Cleaning Data in R (four hours)
- Joining Data in R with dplyr (four hours)
- Interactive Maps with leaflet in R Chapters 1 and 2. (two hours)

R4DS: Chapters 9, 10, 11, 12 and 13

No problem set due the week after the midterm.

Speakers

- October 15: Heidi Chen, Portfolio Manager, Acadian Asset Management

Optional

- Intro to SQL for Data Science

- Interactive Maps with leaflet in R Chapters 3 and 4.
- Importing & Cleaning Data in R: Case Studies

- Data Organization in Spreadsheets by Karl W. Broman and Kara H. Woo

Week 8: October 22 and 24. Text

DataCamp

- String Manipulation in R with stringr (four hours)

Optional

- Regular expressions

- String Manipulation in R with stringr
- Regular Expressions for Data Science in R

R4DS: Chapters 14 and 15

Problem Set #4 due October 24 at 10:05 AM covering material through Week 7.

Speakers

- October 22: David Sparks, Director of Basketball Analytics for the Boston Celtics
- October 24: Joe Harrington, Coordinator, Performance Science at Los Angeles Dodgers

Optional

- Naming Things by Jenny Bryan

Week 9: October 29 and 31. More Text

DataCamp

- Sentiment Analysis in R: The Tidy Way (four hours)

Speakers

- October 31: Jeremy Rogalski, Director of Hockey Analytics for the Boston Bruins

Optional

- Working with Dates and Times in R
- Categorical Data in the Tidyverse
- Intro to Python for Data Science
- "Rich State, Poor State, Red State, Blue State: What's the Matter with Connecticut?" by Gelman et al.

R4DS: Chapters 16

Problem Set #5 due October 31 at 10:05 AM covering material through Week 8.

Week 10: November 5 and 7. Functions

DataCamp

- Writing Functions in R (four hours)

R4DS: Chapters 17 and 18

Problem Set #6 due November 7 at at 10:05 AM covering material through Week 9.

Second midterm distributed November 7 after class and due Sunday November 11 at 10:05 PM. This midterm will be cumulative, including topics from the first midterm as well as relational data, strings, dates, times, factors, functions, iterating, and working with data from the web.

Speakers

- November 7: Huan Wang, Research Scientist, Office of the Vice Provost for Advances in Learning at Harvard University

Optional

- The Quartz guide to bad data
- Teaching Digital at the Harvard Kennedy School by David Eaves

Week 11: November 12 and 14. Modeling

DataCamp

- Modeling with Data in the Tidyverse (four hours)

R4DS: Chapters 19, 20 and 21

Speakers

- November 14: Mike Burke, Registrar, Harvard University

Optional

- The Technological Revolution Has Finally Hit the NFL—and the Vikings Are Ready
- *The Cognitive Style of Powerpoint* by Edward Tufte

Week 12: November 19. Model Basics

No DataCamp due Thanksgiving Week.

Problem Set #7 due November 21 at at 10:05 AM covering material through Week 10.

R4DS: Chapters 22 and 23

Speakers

- November 19: Daniel Koh, former Chief of Staff to the Mayor of Boston

No class Wednesday, November 21

Week 13: November 26 and 28. More Models

DataCamp

Because of the Thanksgiving, DataCamp exercises are not due until Wednesday, November 28 at 10:05 AM.

- Building Web Applications in R with Shiny: Case Studies (four hours)

R4DS: Chapters 24 and 25

Speakers

- November 28: Katherine Evans, Quantitative Analyst at Verily Life Sciences

Optional

- Building DashBoards with flexdashboard
- A Compendium of Clean Graphs in R

Week 14: December 3 and 5

R4DS: Chapters 28 and 29

No problem set or DataCamp during the week of final projects.

Final project *presentations* occur on December 5. Final project *submissions* are not due until 10:05 PM on December 13.

Technical Advice

Git

- If you have git problems, your first stop is *Happy Git and GitHub for the useR* by Jenny Bryan.

RStudio

- Under Tools -> Global Options -> General, set the "Save workspace to .RData on exit:" to "Never".
- Under Tools -> Global Options -> Code -> Saving, set the "Default text encoding:" to "UTF-8". This is especially important for Windows users from non-English locales.

Acknowledgements

This course is inspired by STAT 545, created by the legendary Jenny Bryan. Many of the slides and exercises come from Data Science in a Box, by Mine Çetinkaya-Rundel. Many of the in-class exercises are from *Teaching Statistics: A Bag of Tricks* by Andrew Gelman and Deborah Nolan. Kudos to authors like Garrett Grolemund and Hadley Wickham (*R for Data Science*) and to Chester Ismay and Albert Y. Kim (*ModernDive: An Introduction to Statistical and Data Sciences via R*) for making their books freely available. Thanks to Kosuke Imai for open sourcing several of the datasets from *Quantitative Social Science: An Introduction* and to Matt Blackwell and Xiang Zhou for sharing the data from their courses. Lecture slides were created via the R package **xaringan** by Yihui Xie. Many thanks to all the folks responsible for R, RStudio, Git and GitHub. This course would not be possible without their amazing contributions.