Gov 1005: Data

David Kane

Fall 2019

Description

Data matters. Learning to think critically about data is a fundamental skill. How much money is donated to political campaigns? How do polls help us forecast elections? Does exposure to Spanish-speakers affect attitudes toward immigration? We need data to answer these questions – to describe, to predict, and to infer.

This course, an introduction to data science, will teach you how to *think with data*, how to gather information from a variety of sources, how to import that information into a project, how to tidy and transform the variables and observations, how to visualize, how to model relationships, how to assess uncertainty, and how to communicate your findings. 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 topic which students find interesting.

We use the R programming language, RStudio, Git, GitHub and DataCamp.

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

Logistics: Class meets in Tsai Auditorium from 12:00 to 1:15 on T/TH.

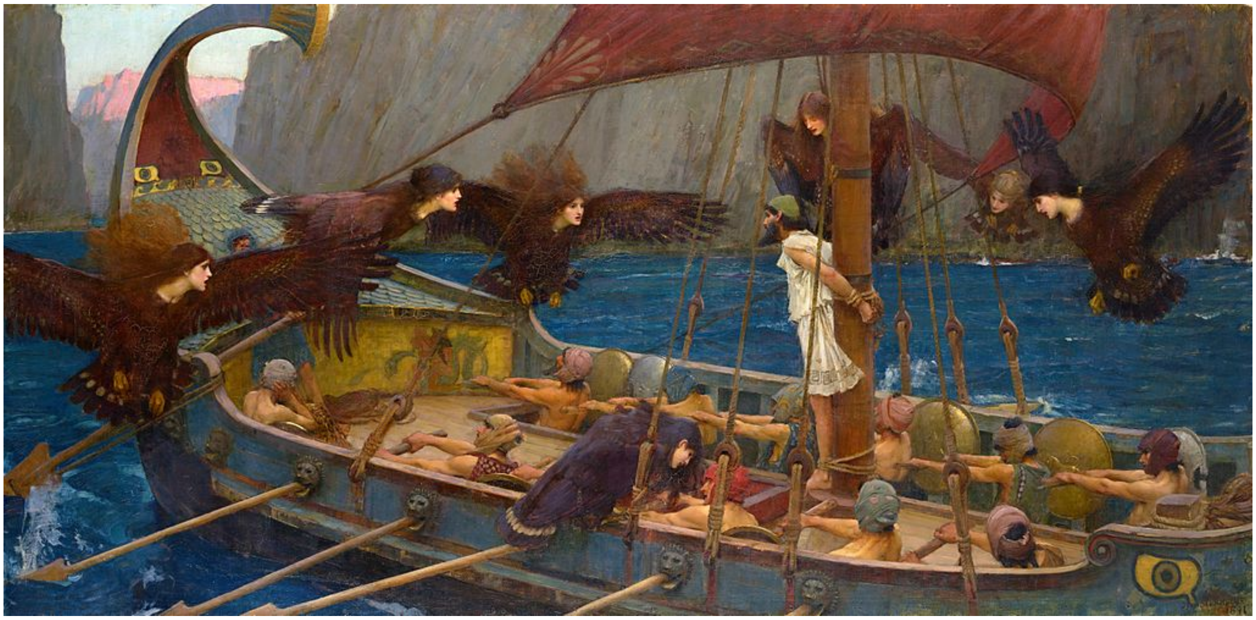

Figure 1: *Ulysses and the Sirens*, 1891, by John William Waterhouse. Homer's *Odyssey* recounts the decade-long journey home of Odysseus (known as Ulysses in Latin) after the Trojan War. Although Ulysses's ultimate goal is his kingdom of Ithaca, he does not shy away from adventure along the way. The Sirens use their enchanting voices to lure unwary sailors to their deaths. Ulysses wanted to hear their songs. He instructed his men to fill their ears with beeswax and to tie him to the mast.

Course Metaphor

The central metaphor for this class is *Ulysses and the Sirens.* You are Ulysses. Ithaca is the future you want. The Sirens are the many distractions of the modern world. *I am the rope.*

Course Staff

Preceptor David Kane; dkane@fas.harvard.edu; CGIS South 310; 646-644-3626; office hours Thursday from 1:30 to 4:00, generally held in Fisher Commons. Please address me as "Preceptor," not "David," nor "Preceptor Kane," nor "Professor Kane," nor "Mr. Kane," nor, worst of all, "Dr. Kane."

Teaching Fellows: Georgie Evans (georginaevans@g.harvard.edu), Sascha Riaz (riaz@g.harvard.edu) and Alice Xu (alicexu@g.harvard.edu). Georgie is your first stop for DataCamp questions. Sascha handles all Google sheets related issues. Ask Alice about all your Piazza and git/GitHub problems.

Course Assistants: Shivani Aggarwal (saggarwal@college.harvard.edu), Enxhi Buxheli (ebuxheli@college.harvard.edu), Claire Fridkin (clairefridkin@college.harvard.edu), Jack Luby (jluby@college.harvard.edu), Seeam Noor (seeamnoor@college.harvard.edu), Kodi Obika (cobika@college.harvard.edu), Dillon Smith (dillon_smith@college.harvard.edu), and Céline Vendler (cvendler@college.harvard.edu).

Course Philosophy

No Lectures: The worst method for transmitting information from my head to yours is for me to lecture you. There are no lectures. We work on problems together during class. You learn soccer with the ball at your feet. You learn about data with your hands on the keyboard.

R Everyday: Learning a new programming language is like learning a new human language: You will 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, in which you can't hide in the back row. Want to be left alone? Don't take this course.

Community: You will meet more Harvard students than you would in a normal course. *Awkwardness in the pursuit of community is no vice.* You will probably learn the names of more students in this course than in all your other courses combined.

Bayesian: The philosophy of this class is unapologetically Bayesian.

Professionalism: We use professional tools. Your workflow will be very similar to the workflow involved in paid employment. Your problem sets and final project will be public, the better to impress others with your abilities. High quality work will be shared with your classmates. We will learn the "full cycle" of how to draw inferences from data and communicate those inferences to others.

Millism: Political disputes are not the focus of this class but, when such topics arise, I will insist that we follow John Stuart Mills' advice: "He who knows only his own side of the case, knows little of that. His reasons may be good, and no one may have been able to refute them. But if he is equally unable to refute the reasons on the opposite side; if he does not so much as know what they are, he has no ground for preferring either opinion."

Teaching to Learn: My main goal is not to teach you how do X. That is easy! More importantly, in a few months, I won't be around to teach you Y. My goal is to teach you how to teach yourself, how to figure out X and Y and Z on your own. That is harder! Much of the pedagogy of the course — especially my insistence that you work on topics not covered in lecture — is driven by this goal. You will find it frustrating, at times. My apologies in advance.

Workload: The course should take about 10 to 15 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. Others will spend **much, much more**.

Course Policies

Late Days: Assignments (DataCamp, Problem Sets and Final Project Milestones) are always due at 11:55 PM, unless specified otherwise. An assignment is a day late if it is turned in any time after it was due (even 5 minutes after) but within 24 hours. After that, it is two days late, and so on. You have 5 *late days* in total. These may be used for any assignment, except for problem sets 3, 6 and 9; the four exams; and final project Demo Day. **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. 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. To claim a late day, you must e-mail the appropriate TF — DataCamp (Georgie), Problem Set (Alice), and Milestone (Sascha) — **before the assignment is due**. No need to ask for permission or to give us a reason. Just inform us that you are taking a late day. You may only use one late day on a given assignment. Hand it in after more than 24 hours and you get a zero on that assignment. **But you still must hand it in!** Everything must be completed. Late days accrue until you do. Each day late (beyond the five allowed) results in -1 point to your final score. This decrement is a *point* not *percentage* penalty. In other words, each additional late day used outside of the allotted five will drop your final class grade by one grade point.

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 are out of town or have a health issue. Simply put an X by your name in the Google absence sheet **and** send me an e-mail. Failure to do so will decrease your participation points, as will missing too many classes, even with notification. There is no need to put a reason in the sheet. An X is enough.

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. We require a signed letter (a piece of paper, not an e-mail) from your Resident Dean as documentation.

Organized by House: We use geography to create a community. During class, you will sit with students from your house, grouped with other houses near yours. You will work with different peers each class.

Monologues: I give brief monologues, designed to explain specific topics that have confused students in the past. I hope to never talk for more than 5 minutes straight.

Speakers: Data scientists, from both industry and academia, will speak with us. If there is someone you would like to meet, talk to me about it and we can invite them!

No Cost: Every reading/tool we use is 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.

Remind Me: In conversations outside a class, a student will often ask an important question or raise of issue of general interest. These topics should be brought to the attention of other students. I will ask you to "Remind me" about it. This means that, in the next class, **you must raise your hand** when I ask for reminders and then remind me! Couldn't I just write it down in my notes? Perhaps. But learning how to raise your hand/voice in a big class is a useful skill. This is your opportunity to practice.

Role of Teaching Fellows: The TFs run Study Halls, grade all assignments, keep track of late days, deal with emergencies and so on. Go to them first with any problems.

Role of Course Assistants: The CAs only run Study Halls. They are not involved in grading assignments and can make no commitments about how the TFs will grade. *Never ask a CA a question about grading.* Instead, ask on Piazza and a TF will respond, or come to a TF privately with your question.

Exceptions: There may be a reason why you can't adhere to class policies. For example, severe social anxiety may make being cold-called problematic. A learning disability may make take-home tests unfair. Whatever the situation, please seek me out for conversation. I am sure we can work out something! I will do whatever it takes to allow every Harvard student to participate (and thrive!) in this class.

Use your Harvard e-mail: Please use your official Harvard e-mail address for all aspects of this class, especially things like signing up for services like DataCamp, GitHub, 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 connect 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).

Plagiarism: If you plagiarize, you will fail the course. See the Harvard College Handbook for Students for details.

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 must not be identical, or even identical'ish.

Git and GitHub: Analyzing data without using source control is like writing an essay without using a word processor — possible but not professional. We will do all our work using Git/GitHub.

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 11:55 PM.

Readings: Assignments in a given week cover (approximately) the material that we will use that week, although DataCamp is a more precise guide to our in-class activities. 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 all assigned material.

Optional Activities: The syllabus includes background readings and DataCamp assignments which students may find interesting. You do not have to do them.

Waite Rule: We don't wear hats in the classroom. (Obviously, this prohibition does not apply to headgear of a religious nature.)

Computer Emergencies: We are not sympathetic about 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 push. You can then restart your work on a public computer (the basement of CGIS Knafel has machines with R/RStudio installed) or on your roommate's computer.

Github Classroom: We use Github Classroom to distribute problem sets and exams. 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 GitHub problems to arise the night before the assignment is due.

Speakers: We follow a *No Laptop Rule* during speaker presentations. Close your laptops. Put down your phones. If you want to take notes, use a pen. We do this because we respect the speakers, want to give them our full attention, and are thankful that they have taken the time to talk with us.

Tardiness: We begin on time and end on time. Do not start gathering your belongings until class is over, especially when we have a speaker.

Credit: Gov 1005 fulfills the QRD requirement. You may also get concentration credit. This is true, obviously, for Government. It is also true in Statistics, Psychology and Social Studies. I am happy to support students who want to petition other departments.

Announcements: You are responsible for any assignment/exam/deadline updates/changes which are either announced in class, promulgated via the course Canvas e-mail list or posted in the official pre/post class notes in Piazza. You are not responsible for every other random post on Piazza.

Grading

Solo Participation: 5 points. This category relates to things you do alone in class. Missing class (without notifying us) or missing too many classes will cost you points, as will a failure to participate in class activities. We keep track of this via Google sheets, so be sure to fill them out when requested. Note that I do not care if you know the answer when I cold-call you. This plays no part in your grade.

Group Participation: 5 points. This category relates to activities you do with other students. Helping your fellow students, especially on Piazza, is the best form of group participation, as is volunteering for a class role. Be a good class citizen. Help your classmates during Study Halls. Do not shirk on group projects.

DataCamp Lessons: 5 points. Grades are pass/fail only. 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. There are 8 weeks of DataCamp, so each week's assignments (whether one course or more) counts for 5/8th of a point.

Problem Sets: 25 points. The first problem set is worth 1 point. The remaining 8 are worth 3 points each. Problem sets after the first are distributed on Thursday and then due the following Wednesday at 11:55 PM. 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. We define "work with" very broadly, to include minor interactions. You would certainly list anyone you sat nearby during Study Hall, for example.

Exams: 35 points total. The four exams are take-home. The first is worth 5 points and the others are each worth 10 points. They are open-book and open-web. Because students have different schedules, you can complete the exam any time within a four-day window starting after exam distribution. Late exams earn zero points. You may not seek or receive help on the exam from a person, e.g., asking a roommate or posting at RStudio Community. You may use any written materials from the class, including problem set answers. If you have a question, ask on Piazza. Teaching staff (not other students) will answer it.

Final Project: 25 points. Students will present their projects publicly at the end of the semester. They will then have the opportunity to incorporate feedback before submitting the final version. There are eight milestones for the projects, worth either 1 or 2 points, depending on difficulty. All 8 milestones together count for 10 points. Demo Day counts for 5 points. The final project submission is worth 10 points.

Books

The texts for the class are *R for Data Science* (R4DS) by Garrett Grolemund and Hadley Wickham, *Statistical Inference via Data Science: A moderndive into R and the tidyverse* (MD) by Chester Ismay and Albert Y. Kim, and *Data Visualization: A practical introduction* (DV) by Kieran Healy. These resources may also be helpful. All are free.

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 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 graduate schools or potential employers. You must show this work publicly, both on the web (viewable by all) and in person at our Demo Day. You will host your final project using Shiny Apps, a free service provided by RStudio. Make use of free statistical consulting from the Harvard Statistics Department and from IQSS. Read this advice if you are working with data larger than 100 megabytes.

You may combine this project with a research paper or other assignment from a different class. You automatically have my permission. But you must get explicit permission from the instructor for the other class as well.

It is not enough to simply use an already-assembled data set. Instead, you must combine data from a variety of different sources. Looking at your data-munging code will confirm for us that you have made an actual contribution. Imagine that your roommate 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." Projects without at least 10,000 data points are unlikely to be interesting enough, but feel free to convince us otherwise. Projects must feature some statistical modelling, most commonly a regression.

The typical Shiny App will include three tabs. The "About" tab will provide background information about you and your data. The second tab will display your final model, and allow the user to change some of your assumptions and see the results. The third tab will be a detailed tour of the modeling choices you made and an explanation of why you made them.

All projects must include a 1 to 2 minute video in which you explain what you have found and a three page PDF which you must submit to this competition.

Consider scheduling an interview with Hugh Truslow (truslow@fas.harvard.edu), Head, Social Sciences and Visualization, Harvard University. No one at Harvard knows more about potential data sources. Visualization Specialist Jessica Cohen-Tanugi (jessica_cohen-tanugi@harvard.edu) is a great person to talk to about your graphics.

Possible Approaches

Most students will gather some data, estimate some models, and create a Shiny App. Good stuff! But there are other possible approaches:

Paper Replication

Read "Publication, Publication" (pdf) by Gary King. PS: Political Science and Politics, Vol. 39, No. 1 (Jan., 2006), pp. 119-125. King describes how to replicate the results of a published academic paper. See more details here. You will not be doing all of that! (Take Gov 1006 or Gov 2001 for that experience.) Instead, you will be creating a Shiny App which reproduces at least some of the key results of the paper and demonstrates what happens when changes are made in the modelling approach. How "robust" – to use Leamer's terminology — are the results?

Read "Making the Most of Statistical Analyses: Improving Interpretation and Presentation" (pdf) by Gary King, Michael Tomz and Jason Wittenberg. American Journal of Political Science, Vol. 44, No. 2 (April, 2000), pp. 347-361. This is one of the most cited articles in political science in the last 20 years. Just redoing the analysis/graphics of a published article by making use of these techniques would make for an outstanding final project.

Original Data Collection

Students interested in a topic about which there is no publicly available data are welcome to collect their own data. This must be something much more substantive than just asking 100 students outside Annenberg about their favorite salad. Two categories of data work best. First, pick a topic which you truly care about. Second, pick something Harvard-specific. This *Crimson* article and this spring 2019 project are great examples of the latter.

Work with Other Classes

You are welcome to use data from other classes/projects in the creation of your final project. This includes thesis work. You automatically have permission from us to do this, but you must also obtain permission from the instructor of the other class.

Others?

Interested in doing a project which seems different from what we describe above? Come talk to us! The best projects involve topics which students are passionate about. If you really care about X, then we are eager to help you create a final project about X.

Prior Projects

Consider all the final projects from past semesters. Click on the project title to explore the Shiny App. Click on the student's name to explore their Github repo. Note that, in prior years, the course had less of a statistical focus. So, these projects do not feature as much statistical modeling as yours will. Highlights:

Shivani Aggarwal: How Couples Meet. Visualizing the ways in which different kinds of U.S. couples meet and enter into relationships.

Neil Khurana: Harvard Dining. Archiving Harvard menus and exploring variations and repetition in meal choices.

Dasha Metropolitansky: First-Year Blocking Group Project. Harvard says it fosters a diverse community; trends in students' housing indicate otherwise. This was a group project. The other group members were: Adiya Abdilkhay, Ilkin Bayramli, April Chen, Alistair Gluck, Christopher Milne, Neil Schrage and Stephanie Yao. Read more about the project here and here.

Christopher Onesti: Course Enrollment Statistics. This project presents an inside look and trend visualization regarding fall and spring undergraduate course enrollment data at Harvard.

Margaret Sun: Beyond The Stage. Various insights into the music group BTS.

Ruoqi Zhang: Settling the Dust: Censorship & Environmental Activism in China, 2012. What does social media data tell us about environmental awareness and censorship in China, 2012?

Maclaine Fields: Harvard Volleyball. I analyzed setting, serving, receiving, digging, and attacking results and created plots that show the setting tendencies and serving trajectories of Harvard Volleyball and its opponents

Kemi Akenzua: Death Row Last Words. A closer look at the final words of people executed in Texas.

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 know how to describe, predict and infer – you will do Y well. The success (or failure) of the class can be measured by comparing Y with X.

Organization

Everything — DataCamp (Mondays), Problem Sets (Wednesdays), Milestones (Fridays) and Exams (Sundays) — is due at 11:55 PM, unless otherwise specified.

Rhythm of the Class

The class follows a steady weekly rhythm:

Sunday, 2:00 – 5:00 PM, Study Hall with Claire Fridkin, Dunster Dining Hall.
Sunday, 7:00 PM – 10:00 PM. Study Hall with Dillon Smith, Smith Center.
Sunday, 8:00 PM – 11:00 PM. Study Hall with Kodi Obika, Currier Dining Hall.
Monday 4:30 PM – 7:30 PM. Study Hall with Sascha Riaz, K108 in Knafel CGIS.
Monday 7:30 PM – 10:30 PM. Study Hall with Shivani Aggarwal, Science Center.
Monday 11:55 PM. DataCamp exercises due.
Tuesday 12:00 PM – 1:15 PM. Class.
Tuesday 5:00 PM – 8:00 PM. Study Hall with Alice Xu, Fisher Commons.
Tuesday 6:00 PM – 9:00 PM. Study Hall with Céline Vendler, Smith Center.
Tuesday 7:00 PM – 10:00 PM. Study Hall with Enxhi Buxheli, Lowell Dining Hall.
Tuesday 8:00 PM – 11:00 PM. Study Hall with Jack Luby, Winthrop Dining Hall.
Wednesday 2:00 PM – 5:00 PM. Study Hall with Georgina Evans, Fisher Commons.
Wednesday 7:00 PM – 10:00 PM. Study Hall with Seeam Noor, Eliot Dining Hall.
Wednesday 11:55 PM. Problem set due.

Thursday 12:00 PM – 1:15 PM. Class.
Thursday 1:30 PM – 4:00 PM. Office Hours with Preceptor, Fisher Commons.

Thursday evening. Problem set due next week will be distributed.
Friday 11:55 PM. Final project milestones are due.
Sunday 11:55 PM. Exams, if distributed, are due.

Key Dates

Part 1: Tools and Framework

DataCamp due Monday, September 9.
Problem Set #1 due Wednesday, September 11.
Final Project Milestone #1 due Friday, September 13.
DataCamp due Monday, September 16.
Problem Set #2 due Wednesday, September 18.
Final Project Milestone #2 due Friday, September 20.
DataCamp due Monday, September 23.
Problem Set #3 due Wednesday, September 25.
Exam #1 distributed on Thursday morning, September 26 and due Sunday, September 29.

Part 2: Sampling and Inference

DataCamp due Wednesday, October 2.
Final Project Milestone #3 due Friday, October 4.
Problem Set #4 due Wednesday, October 9.
Final Project Milestone #4 due Friday, October 11.
DataCamp due Monday, October 14.
Problem Set #5 due Wednesday, October 16.
Final Project Milestone #5 due Friday, October 18.
DataCamp due Monday, October 21.
Problem Set #6 due Wednesday, October 23.
Exam #2 distributed Thursday morning, October 24 and due Sunday October 27.

Part 3: Models

DataCamp due Wednesday, October 31.
Final Project Milestone #6 due Friday, November 1.
DataCamp due Monday, November 4.
Problem Set #7 due Wednesday, November 6.
Final Project Milestone #7 due Friday, November 8.
Problem Set #8 due Wednesday, November 13.
Final Project Milestone #8 due Wednesday, November 20.
Exam #3 distributed Monday, November 18 and due Tuesday, November 26.

Part 4: Projects

Thanksgiving is Thursday, November 28.

Tuesday, December 3 is last day of classes.

Final project due Friday, December 13. Students must participate in this competition. The pdf which you submit must also be available from your Shiny App. Place the url for that PDF in the Google sheet.

Exam #4 distributed Wednesday, December 4 and due Sunday, December 15.

Schedule

Part 1: Tools and Framework

Data science involves both inputs and outputs. We bring in data from somewhere to analyze and, once we have some answers, distribute our results. During Part 1, we will bring in data from R packages, downloaded text files and text files on the web. We will distribute our results as html files to the course staff, requests for help (from strangers) using reproducible examples and animated graphics posted to the web.

Week 1: September 2

Shopping Week

You are Ulysses. I am the rope.

Install R, RStudio and Git on your laptop. Start on the DataCamp assignments. They are due on Monday, September 9 at 11:55 PM. Sign up for a meeting with a member of the Course Staff. This will fulfill the first milestone, due September 13, for the final project. We will use RStudio Cloud on Tuesday and individual laptops on Thursday. Although it is not officially due till Monday, please try to do Introduction to the Tidyverse for Thursday's class.

Readings

R4DS: Chapters 1, 2, 3, 4, 6 and 8.
DV: Chapters 1 and 2.

Week 2: September 9

Visualization

You can never look at your data too much. – Mark Engerman

We will review some basic R operations including constructing vectors with c() and subsetting elements with []. We will mention useful functions, like slice() and pull(), which are not covered in the DataCamp assignments. We will learn how to create an R project in RStudio. The first problem set will be distributed on Tuesday, via Github Classroom, and completed during class. We will also learn how to recover from git mistakes. We will introduce the "potential outcomes" framework and review the fundamental problem of causal inference.

Readings

R4DS: Chapters 5 and 7.
DV: Chapter 3.

DataCamp

Remember: DataCamp assignments are due Monday at 11:55 PM.

Introduction to the Tidyverse
Introduction to Shell for Data Science only first chapter, "Manipulating files and directories"
Introduction to Git for Data Science only first chapter, "Basic workflow"
Visualization Best Practices in R
Communicating with Data in the Tidyverse, only third chapter, "Introduction to RMarkdown"

Assignments

Problem Set #1 due September 11 at 11:55 PM. We will complete and submit this problem set in class on Tuesday. Its purpose is to ensure that everyone has a working computer, understands Git/GitHub and can compile an R Markdown document.

Final Project Milestone #1 due Friday, September 13. Speak with any member of the course staff (Course Assistant or Teaching Fellow) about your final project. Bring your laptop. Most staff study halls are Sunday through Wednesday. Do not wait until Friday morning. Record the name of the person you met with in the Final Project Google Sheet. This is the first of three required meetings. No need to prepare for this meeting. But it is important to start thinking about what you want to do. This also provides for an opportunity to meet some of the course staff. Sign ups will be distributed via Piazza.

Optional: RStudio Essentials Videos. Most relevant for us are "Writing code in RStudio", "Projects in RStudio" and "Github and RStudio". Again, these are optional! But they are very useful for students who find find traditional lectures to be a helpful supplement to classroom practice. See also *GitHub Classroom Guide for Students*.

Week 3: September 16

Seeking Help

The best data science superpower is knowing how to ask a question. – Mara Averick

We will learn how to produce a **repr**oducible **ex**ample — a "reprex" — in order to help strangers to help us. We will discuss the slogan "no causation without manipulation."

Readings

R4DS: Chapters 9, 10, 11.
DV: Chapter 4.

DataCamp

Working with Data in the Tidyverse
Working with Web Data in R

Assignments

Problem Set #2 due Wednesday, September 18.

Final Project Milestone #2 due Friday, September 20. Github repo with Rmd (and knitted html) which discusses pros and cons two projects from past years. At least one project should be one which did extensive data gathering/cleaning. You should not select the same projects for commentary as your friends have. Students generally write about a paragraph for each project.

Causality, Chapter 2 of *Quantitive Social Science* by Kosuke Imai, especially pages 46 – 63.

Optional: RStudio Webinar on Reprex. Again, these are optional! But they are very useful for students who find find traditional lectures to be a helpful supplement to classroom practice.

Speaker: September 19: Mara Averick, RStudio. Lunch to follow.

Week 4: September 23

Animation

Workflow: you should have one. – Jenny Bryan

We will learn how to make engaging animations. We will discuss the meaning of "average effect" and related terms.

Readings

The Cognitive Style of Powerpoint by Edward Tufte.

MD: Chapters 1 through 5.
R4DS: Chapters 12, 13, 14, 15 and 16.
DV: Chapter 5.

DataCamp

Joining Data in R with dplyr
String Manipulation in R with stringr. Chapter 1, String basics
String Manipulation in R with stringr. Chapter 2, Introduction to stringr

Assignments

Problem Set #3 due Wednesday, September 25. You may not use a Late Day for this problem set.
Exam #1 due Sunday, September 29.

Optional: The Unix Workbench, chapters 1 – 6.

Speaker: September 26: Will Kurt, Data Science Manager at Wayfair. Lunch to follow.

Part 2: Sampling and Inference

Week 5: September 30

Sampling

Lot of points were taken off for small errors that I did not see as pedagogically important. – Gov 1005 student

Readings

MD: Chapter 7 Sampling.
R4DS: Chapters 17, 18, 19, 20 and 21.

DataCamp

Because of the exam, DataCamps are not due till Wednesday at 11:55 PM.

Introduction to Function Writing in R
Foundations of Functional Programming with purrr

Assignments

Final Project Milestone #3 due Friday, October 4. Speak with any member of the course staff (CA or TF) about your final project. This is the second of three required meetings. In addition to this meeting, you must create a rough Github repo, with at least some of your raw data *or* code which shows you working with data that is stored elsewhere *or* details of your plan to get the data, and a reproducible html which provides a brief description of the data: where you got it, what you have done with it so far and what you plan to do. You may change your project completely, all the way until Demo Day. But you are still responsible for meeting these milestones, even if you know you are going to pivot. Your data can not be from a single source. Typing **library(fivethirtyeight)** is not enough!

Optional: RStudio Webinar titled "How to Work with List Columns" by Garrett Grolemund. Background reading about anonymous functions in R.

Week 6: October 7

Confidence Intervals

Comment as a service to the dumbest possible version of your future self. – Alex Albright

Readings

MD: Chapter 8 Bootstrapping & Confidence Intervals.
R4DS: Chapters 25.

Assignments

Problem Set #4 due Wednesday, October 9.

Final Project Milestone #4 due Friday, October 11. Create a beautiful graphic, using ggplot2 or another package of your choice, which uses some of your data.

Optional: Two videos about permutation tests along with a traditional textbook treatment, pages 16-41ff. Also, this pretty animation.

Speaker: October 10: Angela Bassa, iRobot. Lunch to follow.

Week 7: October 14

Bayes

I stopped teaching frequentist methods when I decided they could not be learned. – Donald Berry

How many twins are there at Harvard? Does a subgroup of voters from a larger sample provide the best estimate for the entire subgroup population?

Readings

Chapter 1 in *Think Bayes* (pdf) by Allen Downey.
Chapter 2 in *Doing Bayesian Data Analysis* (pdf) by John Kruschke.
MD: Chapter 9 Hypothesis Testing.

DataCamp

Beginning Bayes in R. Note that there are a handful of exercises which use simple linear regression, which we don't cover for two weeks. If you are unfamiliar with this topic, feel free to just "Show Answer" on those questions.

Assignments

Problem Set #5 due October 16.

Final Project Milestone #5 due Friday, October 18. Create an Rmd/html which provides a draft of your About page. All of your data processing should be complete. Remember: You must gather data from two or more different sources. Learning how to source, clean and combine data is one of the goals of the project. On almost any topic, there are useful tables of information on Wikipedia. See here and here for advice.

Optional: Statistical Rethinking: A Bayesian Course with Examples in R and Stan (pdf) by Richard McElreath. Chapter 1.

Speaker: October 17: Stefanie Costa Leabo, Chief Data Officer, City of Boston. Lunch to follow.

Week 8: October 21

Shiny and Maps

Readings

"Let's Take the Con Out of Econometrics," by Edward E. Leamer. The American Economic Review, Vol. 73, No. 1 (March, 1983), pp. 31-43. link

"Causal effect of intergroup contact on exclusionary attitudes" by Ryan Enos. PNAS March 11, 2014 111 (10) 3699-3704. link

DataCamp

Spatial Analysis in R with sf and raster

Assignments

Problem Set #6 due October 23. You may not use a Late Day for this problem set.
Exam #2 due Sunday October 27.

Optional: "A Balanced Perspective on Prediction and Inference for Data Science in Industry" by Nathan Sanders, DV: Chapter 7, Introduction to Mapping with sf and How to Start Shiny video tutorial.

Speaker: October 24: Nathan Sanders, Chief Scientist at Warner Media Applied Analytics. Lunch to follow.

Part 3: Models

Week 9: October 28

Regression

Teach people to drive. Then later, if they want or need, they can learn how the internal combustion engine works. – Andrew Gelman

Readings

MD: Chapter 5 Basic Regression.
Additional material on basic regression by Céline Vendler.
DV: Chapter 6.

DataCamp

Note that, although the DataCamps this week focus on univariate regression, they both also uses examples from multivariate regression which we, officially, do not cover until next week. You may find it useful to look ahead at MD Chapter 6. Because of the exam, DataCamps are not due till Wednesday at 11:55 PM.

Modeling with Data in the Tidyverse
Bayesian Regression Modeling with rstanarm

Assignments

Final Project Milestone #6 due Friday, November 1. Speak with any member of the course staff (CA or TF) about your final project. This is the third of three required meetings. By 11:55 PM, you must have a working Shiny app, just to demonstrate that you can get something up and running. This does not have to be working for your meeting. This Shiny app does not have to use your data. It can just be the default example we practiced with during class. Add the name of the person you met with and the url of your Shiny App to the appropriate Google sheet.

Optional: Shiny tutorials.

Week 10: November 4

Multivariate Regression

Amateurs test. Professionals summarize.

Readings

MD: Chapter 6 Multiple Regression.
Additional material on multivariate regression by Céline Vendler.
DV: Chapter 8.

DataCamp

Machine Learning in the Tidyverse

Assignments

Problem Set #7 due Wednesday, November 6.

Final Project Milestone #7 due Friday, November 8. Cleaned up Github account. This is your chance to make use of the feedback you received during last week's meeting. Consider the GitHub homes of some prior students in the class. Note how professional they look! Make your GitHub look at least as good as theirs. If you want companies to treat you as a data scientist, you should present a professional visage to the world. Some observations:

- Have a bio. It may be as serious or as whimsical as you like, but it should mention your expertise in R. Perhaps the bio will include links to impressive graphics that you have put on-line, either at Rpubs or elsewhere. Start here.
- Note the photo. You do not have to use a photo of yourself — some people are shy — but you should use something other than the GitHub default.
- Provide at least an e-mail address so people can easily contact you. Providing other things — like a connection to LinkedIn — is fine as well, but the e-mail is the key item.
- Use the GitHub option to "pin" certain repos that you are proud of (and/or ones which you have cleaned up) at the top of the page. You should hide (or make private or delete) all junk/scratch repos, like the ones we create in class.

Optional: "The Bayesian New Statistics" by John K. Kruschke and Torrin M. Liddel.

Speaker: November 7: Alex Albright, Harvard University. Lunch to follow.

Week 11: November 11

Classification

Readings

Classical and Bayesian Logistic Regression by Céline Vendler.

Assignments

Problem Set #8 due Wednesday, November 13.

Optional: Video lectures of generalized linear models with binary data, parts 1, 2 and 3.

Speaker: November 14: Andrew Therriault, former Facebook and City of Boston. Lunch to follow

Week 12: November 18

Machine Learning

Put your work on the web. – David Sparks

Assignments

Hands-On Machine Learning with R by Bradley Boehmke and Brandon Greenwell: chapters 1 – 5.

Final Project Milestone #8 due Wednesday, November 20. Working rough draft of your final project. Demo Day is still three weeks away, and you can completely pivot if you want, but you must have a fairly complete version of your current project: a Shiny app with your About page, your data and your model.

Optional: Video: Intro to Machine Learning with R and Caret.

Speaker: November 21: David Sparks, Director of Basketball Analytics for the Boston Celtics. Lunch to follow.

Part 4: Projects

The main focus of the last two class meetings is the final projects. Note that we only have one meeting (on Tuesday) during each of the last two weeks.

Week 13: November 25

Wrap Up

Fitting is easy. Prediction is hard. – Richard McElreath

Readings

R4DS: Chapters 26, 27 28, 29 and 30.

Optional: Mastering Shiny by Hadley Wickham.

Week 14: December 2

Demo Day

A public portfolio of high quality work is better than a Harvard degree.

Last day of classes. Make memes, provide course feedback, discuss final projects and have fun!

Demo Day Sessions

Georgie: Tuesday, December 3 at 9:00 AM in Tsai.
Alice: Wednesday, December 4 at 9:00 AM in Tsai.
Sascha: Wednesday, December 4 at 1:30 PM in Belfer (next door to Tsai).

Class Room Seating

- Seating is organized, by campus geography, into several large "Groups" of 20 to 30 students: first years, Eliot House, Quadlings, et cetera. Details depend on enrollment.
- Students work in "Pairs" of two "Partners." Sometimes, this will be "side-by-side," each of you with a computer open, each writing code, but talking with each other throughout. Other times, we will "pair program," meaning just one computer open and both of you collaborating on a single project. You will work with a different partner every class.
- If you are the stronger student in a Pair, do not simply charge ahead. Instead, make sure that your Partner keeps up with you. Help each other! If you aren't talking with each other often, then you are doing it wrong. *There is no better way to learn than to teach.* The stronger student should type less and talk more.
- Pairs will always be grouped into larger "Circles," generally of 6 students. All these students will be from the same Group. Make sure to introduce yourself to everyone in your Circle at the start of class. There will be a quiz. Most Circles will include students you have worked with before, but please try to meet everyone in your larger Group.

Record the name of your Partner in the Google sheet for the day. Each person does this, even though doing so leads to duplication.

Assignment Details

Participation

There are several ways to earn group participation points in class.

Scribe: We need note-takers, two students for each day. They work separately, but will still be partnered with someone so they can participate in coding. After class, the two scribes get together and create one unified set of notes, which must be posted to Piazza before 11:55 PM that evening.

Welcome Committee: We organize a Welcome Committee of five students for each speaker. See here for the duties associated with this role.

Piazza: Answering your classmates questions on Piazza is the best way to earn participation points. Be a good class citizen! If you find a (meaningful!) typo in a problem set or exam, please post it to Piazza. The first student to do so earns many participation points.

Final Project Milestones

Final project milestones are always due at 11:55 PM of the designated date, which is always a Friday. You may use late days, except for Demo Day and the final due date. All submissions are made via a Google sheet, the url of which will be distributed on Piazza. There are no milestones due during exams periods. The milestones which occur in the week after an exam (Oct 4th and Nov 1st) are major milestones, requiring more work and, therefore, being worth two points. Other milestones are worth 1 point. All 8 milestones together count for 10 points. Demo Day counts for 5 points. The final project submission is worth 10 points and is due on December 13. Fill out the Google sheets correctly!

Study Halls

Study Halls (SH) are run by Course Assistants (CAs), undergraduates who have taken the class in the past. They are one of the most popular parts of the course. Teaching Fellows (TFs) also run Study Halls, although these will often have more of an office hours flavor. Students who make the most use of these resources do better in class, and enjoy it more, than students who do not. Course Staff (CS) is a term which incorporates both course assistants and teaching fellows.

Introductions

At every SH, the CS will ensure that everyone knows everyone else's name. This class is a community and community begins with names. The process starts with the first student arriving and sitting at the table. They and the CS chat. (It is always nice for the student to take the initiative and introduce themselves to the CS. Remembering all your names is hard!) A second person arrives and sits at the same table, followed by introductions. Persons 3 and 4 arrive. More introductions. Help your CS by introducing yourself, even if you are 75% sure they remember your name. Be friendly!

At this point, the table is filled. Another person arrives. Instead of that person starting a new table, CS gives the new student their spot and moves their belongings to a new table. No student ever sits alone. The CS hovers around the table until more students arrive and start filling out table #2. And so on. At each stage, students are responsible for, at a minimum, introducing themselves to the CS and, even better, to the other students. Best is when students who are already present shower newly arriving students with welcomes and introductions.

Help Us Help You

CS will, to the greatest extent possible, never just give you the answer. Something like "Use annotate()" might solve your immediate problem, but it does not set you up for success during the exams — when we won't be around to serve as your personal o**R**acles — much less for the rest of your life.

Instead, we will take the time to show you how to find the answer yourself. This starts with how to search for help, especially when you are not sure what you are looking for. This is more art than science, but adding certain strings — like "R", "tidyverse", or "ggplot" — to the search often helps. Then, we provide advice about which locations are the highest quality (anything to do with RStudio or tidyverse), which locations are less good than they initially appear (sthda.com, r-statistics.co, rdocumentation.org), and which are difficult to use (Stack Overflow). We then explain the best way to make use of what you find.

We also point you directly to the best resources, especially to *R for Data Science* by Garrett Grolemund and Hadley Wickham and to *Data Visualization: A practical introduction* by Kieran Healy. We won't say: "Just use starts_width()." Instead, we will ask, "Have you read Section 5.4 of R4DS, involving the use of select()?" Yes, this will require an extra five minutes of your time. **But every extra minute you spend reading a high quality reference is a minute well-spent.**

We also help you learn how to seek help from others. There is a good way to ask for help on Piazza or Stack Overflow — generally involving the use of reproducible example which highlights your precise problem — and a bad way.

Only if none of this works will we just tell you the answer.

Extra Help

Course assistants (but not teaching fellows) are available for one-on-one or small group meetings outside of their regularly scheduled Study Halls. These meetings may *not* be used to work on the next problem set.

That is what regular Study Halls are for. The most common purpose of these meetings is to review the questions/answers from past problem sets and exams, the better to set a solid foundation for students moving forward. A second purpose is to provide help for the final projects. Process:

1. E-mail a CA with whom you would like to work to see if they are available. Mention the material you hope to review and cc Preceptor. (You only cc me on the first e-mail, not on subsequent back-and-forths.)
2. If that CA is too busy, try another CA.
3. Arrange with the CA a mutually agreeable time/location for the meeting.
4. Do not blow off the meeting! The CA will tell me and I will be upset.
5. After the meeting, the CA will e-mail me with an update on material covered.

Social Events

Socializing with students outside of class is fun. Joining/inviting me is optional and has no influence on your grade in the class, i.e., **it earns you no participation points.** The three main options are:

Restaurant Lunches

My wife and I host students in groups of 4 for lunch throughout the semester, sometimes via the Harvard Class-Room-to-Table program and sometimes on our own dime. We organize this by House at the start and then open up spots to everyone later. Invitations to come. Dress is casual. Please be on time. The reservation will be under "Kane." Just go straight to the table.

House Lunches

I enjoy having lunch with you after class, either in the CGIS cafe, Annenberg or at your House.

Faculty Dinners

I enjoy attending faculty dinners, so feel free to invite me to yours. My only request is that you also invite the other students in the class who live in your House. It is often fun to take over a table with a group of 4 or 5 or . . .

Useful Links

Google sheets for daily partners, class roles, absences and final project milestones.

Overview of, and grading rubrics for, the problem sets and exams.

Possible data sources for final projects.

Detailed instructions for members of the Welcome Committee.

Technical advice which students should follow. Read this at least once before submitting Problem Set #2.

Follow this advice if you have computer problems.

List of my friends/acquaintences from the world of data science, all of whom are happy to talk with students in my class. Reach out to them!

Free R Books.

Acknowledgements

This course is inspired by STAT 545, created by the legendary Jenny Bryan. The pedagogical goals follow Don Rubin's vision. Some of the slides and exercises come from Data Science in a Box, by Mine Çetinkaya-Rundel. Some 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*), Kieran Healy (*Data Visualization: A practical introduction*), Chester Ismay and Albert Y. Kim (*Statistical Inference via Data Science: A moderndive into R and the tidyverse*) for making their books freely available. Thanks to Kosuke Imai for open sourcing several of the datasets from *Quantitative Social Science: An Introduction.* Slides were created via the R package **xaringan** by Yihui Xie. Many thanks to all the folks responsible for R, RStudio, Git, GitHub and DataCamp. This course would not be possible without their amazing contributions.