BST 260
Introduction to Data Science
Fall 2021
M/W, 9:45 - 11:15am EST, Kresge G1

Instructor Information

Heather Mattie
Office: Building 1 Room 421A
Phone: 617-432-5308
Email: hemattie@hsph.harvard.edu
Office hours: Wednesdays 8:30-9:30am or by appointment

- Will be held in-person in Heather's office as well as online via Zoom

Teaching Assistants

Rolando Acosta	racosta@fas.harvard.edu
Jonathan Luu	jluu@g.harvard.edu
Octavious Talbot	octavioustalbot@g.harvard.edu
Stephanie Wu	stephaniewu@fas.harvard.edu
Luli Zou	zou@g.harvard.edu

Office hours

Office hours will start the second week of the semester and will all be held in-person and online via Zoom. The Zoom links will be posted in the course Canvas site.

Day	Time	Teaching Staff Member	Location
Monday	1:00-2:00pm	Rolando	FXB G03
Monday	2:00-3:00pm	Jonathan	FXB G03
Tuesday	11:00am-12:00pm	Stephanie	Kresge 205
Tuesday	1:00-2:00pm	Octavious	Kresge 205
Wednesday	8:30-9:30am	Heather	Building 1 Room 421A
Thursday	1:00-2:00pm	Luli	FXB G03

Labs

Day	Time	Location
Wednesday	2:00-3:30pm	Fall 1: Kresge 200 Fall 2: Kresge 502
Thursday	3:45-5:15pm	Online Zoom link on course Canvas site

Credits

5 credits

Course Description

Unprecedented advances in digital technology during the second half of the 20th and beginning of the 21st centuries is transforming science, including health and biomedical research. Scientific fields that have traditionally relied upon simple data analysis techniques of smaller datasets have been transformed by technologies that continue to expand the possibilities of observing and deciphering massive amounts of data in an unprecedented way. This course includes concepts from Statistics, Computer Science and Software

Engineering. We will learn the necessary skills to manage, visualize and analyze data. We will learn concepts such as exploratory data analysis, statistical inference and modeling, machine learning, and visualization. We will also learn the necessary skills to develop data products including R programming, data wrangling, reproducible research, and communicating results.

Pre-Requisites

Must have basic R programming knowledge and statistics knowledge at the level of Stat 100 or above.

Learning Objectives

This class focuses on methods for learning from data, in order to gain useful predictions and insights. Separating signal from noise presents many computational and inferential challenges, which we approach from a perspective at the interface of computer science and statistics. Through real-world examples of wide interest, we introduce methods for five key facets of an investigation:

1) data munging/scraping/sampling/cleaning in order to construct an informative, manageable data set;
2) software engineering skills for accessing data as well as organizing data analyses and making these analyses sharable and reproducible;
3) exploratory data analysis to generate hypotheses and intuition about the data;
4) inference and prediction based on statistical tools such as modeling, regression, and classification;
5) communication of results through visualization, stories, and interpretable summaries.

Course Readings

None. Instead, students are encouraged to read the lecture documents and other resources available on Canvas site and the course GitHub repository.

Canvas Course Website: https://canvas.harvard.edu/courses/94161

Slack: We have a course Slack workspace you can join

Grading, Progress and Assessment

The final grade for this course will be based on:

- 5 Homework Assignments (40%)
- 1 Take-home Midterm (25%)
- 1 Final project (35%)

Homework Assignments (40%)

All homework assignments will involve writing code and communicating results. Students must submit the RMarkdown file and knitted html file associated with each assignment in their individual repository. A private repository for each assignment will be created for each student and will only be visible to the student and course teaching staff.

Each student is given two late days per homework assignment. A late day extends the individual homework deadline by 24 hours without penalty. No more than two late days may be used on any one assignment. Late days are intended to give students flexibility: students can use them for any reason, no questions asked. Student don't get any bonus points for not using late days. Also, students can only use late days for the individual homework deadlines - all other deadlines (e.g., project milestones, midterm exam) are hard.

Although each student is given late days, we will be accepting homework from students that pass this limit. However, we will be deducting 10% (10 points) for each extra late day.

Due to the unpredictable nature of COVID-19 students in need of extra time to complete assignments should reach out to Student Affairs at StudentAffairs@hsph.harvard.edu. A staff member will work with you and Dr. Mattie to accommodate you. You can also contact Student Affairs if you have a learning disability that requires accommodations. We will ensure you are accommodated as needed.

The TAs must be able to knit submitted RMarkdown files. The penalty for not being able to knit a file while grading increases for each subsequent homework – see breakdown below. To avoid this, students should be sure to include relative paths to files, images, etc. rather than absolute paths (paths specific to your computer). Examples of how to include paths will be given in lecture and lab sessions. Students may also double check with the teaching staff before submitting assignments.

- 0 points for HW1
- 5 points for HW2
- 10 points for HW3
- 15 points for HW4
- 20 points for HW5

Students may ask questions about the assignments during lecture, but we ask that any questions about grading be directed to the TAs or Dr. Mattie outside of lecture and lab sessions via email.

Take-home Midterm (25%)

A take-home midterm will be distributed in the form of an RMarkdown file in October (date TBD) to test comprehension of course material. The exam will consist of multiple-choice questions that may or may not require writing code, coding questions and short answer questions. All code used and text answers must be submitted using the RMarkdown file. Students will have 1 week to work on the exam and must submit the exam via Canvas by 11:59pm on the deadline (TBD). Students are encouraged to use lecture slides and code, lab material, homework assignments and the Internet to work on the exam, but may not work or consult with other students. The teaching staff will be available to answer any questions concerning the exam.

Due to the unpredictable nature of COVID-19 students in need of extra time to complete the midterm should reach out to Student Affairs at StudentAffairs@hsph.harvard.edu. A staff member will work with you and Dr. Mattie to accommodate you. You can also contact Student Affairs if you have a learning disability that requires accommodations. We will ensure you are accommodated as needed.

Final Project (35%)

Students will work on a month-long data science project. The goal of the project is to go through the complete data science process to answer questions you have about a topic of your own choosing. You will acquire the data, design your visualizations, run statistical analyses, and communicate results.

Project Team

Students will work in teams of 4-5. Students will work closely with other classmates on this project. Canvas or Slack can be used to find prospective team members. In general, we do not anticipate that the grades for each group member will be different. However, we reserve the right to assign different grades to each group member based on peer assessments (see below).

Project Milestones

There are a few milestones for the final project. **It is critical to note that no extensions will be given for any of the project due dates for any reason, except for COVID-19 related emergencies or other unforeseen emergencies. Late days may not be used.** Projects submitted after the final due date will not be graded. Students who anticipate any issues should send an email to the teaching staff **at least one week in advance**.

Key Dates and Deadlines	Description
November 1	Form a team and submit a project proposal
November 1 - 19	Project review meeting with assigned TA
December 12 by 11:59pm	RMarkdown and compiled HTML due
December 12 by 11:59pm	Peer assessment due
December 12 by 11:59pm	Project webpage and screencast due
December 15	Project screencasts shown during lecture

Deliverables

There are several deliverables that will be graded individually to make up a final project score.

Team Registration and Proposal

Students start by filling out a google form to define your teams and project proposal. This form should be filled out by 11:59pm on Monday November 1, 2021. The title and other projects may be changed at a later date if needed. Each team will only need to submit one form. Based on the proposal, a TA will be assigned to each team and will guide students through the rest of the project. Students will schedule a project review meeting with their assigned TA within the following three weeks (November 1-19, 2021). Students should ensure all of your team members are present at the meeting.

RMarkdown and HTML Files

An important part of the project is the RMarkdown and associated HTML file. These will detail the steps taken in developing a solution(s), including how students collected the data, alternative solutions tried, statistical methods used, and insights. Equally important to the final results is how the team got there! The RMarkdown and HTML files are the place you describe and document the space of possibilities explored at each step of the project. We strongly advise you to include many visualizations.

The RMarkdown file should include the following topics. Depending on the project type, the amount of discussion devoted to each will vary:

- **Overview and Motivation**: Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.
- **Related Work**: Anything that inspired the project topic, such as a paper, a website, or something discussed in class.
- **Initial Questions**: What questions are being answered? How did these questions evolve over the course of the project? What new questions were considered in the course of the analysis?
- **Data**: Source, scraping method, cleanup, etc.
- **Exploratory Analysis**: What visualizations were used to look at the data in different ways? What are the different statistical methods considered? Justify decisions made and show any major changes in ideas. How were these conclusions reached?
- **Final Analysis**: What did the team learn about the data? How did the team answer the questions? Justify your answers? **Note that 1 type of analysis per team member is required. A Shiny app counts as a type of analysis.**

As this will be the only chance to describe the project in detail, make sure the RMarkdown file and compiled HTML file are standalone documents that fully describe the process and results. The RMarkdown and HTML files are due Sunday, December 12 by 11:59pm. For instructions on how to submit, please see **Submission Instructions** below.

Code

We expect students to write high-quality and readable R code in the RMarkdown file. Students should strive for doing things the right way and think about aspects such as reproducibility, efficiency, cleaning data, etc. We also expect you to document your code.

Peer Assessment

It is important to provide positive feedback to people who truly worked hard for the good of the team and to also make suggestions to those perceived not to be working as effectively on team tasks. We ask students to provide an honest assessment of the contributions of the members of the team, including themselves. The feedback provided should reflect personal judgment of each team member:

- **Preparation**: were they prepared during team meetings?
- **Contribution**: did they contribute productively to the team discussion and work?
- **Respect for others' ideas**: did they encourage others to contribute their ideas?
- **Flexibility**: were they flexible when disagreements occurred?

Teammates' assessment of individual contributions will be considered as part of the overall project score. The peer assessment is due Sunday, December 12 by 11:59pm. For instructions on how to submit, please see **Submission Instructions** below.

Project Website

Students will create a public website for the project using Google Sites or Github Pages or any other web hosting service of their choice. The website should effectively summarize the main results of the project and tell a story. Consider the audience (the site is public) and keep the level of discussion at an appropriate level. The RMarkdown file, HTML file and data should be linked from the project GitHub Repository (see below) to the website as well. Be sure to also embed the main visualizations and screencast in the website.

The final project website is due Sunday, December 12 by 11:59pm. For instructions on how to submit, please see Submission Instructions below.

Project Screencast

Each team will create a **two-minute screencast with narration** showing a demo of the project and/or some slides. Information about how to prepare these screencasts can be found here. Please make sure that the sound quality of the video is good. Upload the video to an online video-platform such as YouTube or Vimeo and embed it into the project webpage. We will show some of the videos in class.

We will **strictly enforce the two-minute time limit for the video**, so please do not run longer than that. Use principles of good storytelling and presentations to get key points across. Focus the majority of the screencast on the main contributions rather than on technical details. What do you feel is the best part of your project? What insights did you gain? What is the single most important thing you would like your audience to take away? Make sure it is up-front and center rather than at the end.

The final project screencast is due Sunday, December 12 by 11:59pm. For instructions on how to submit, please see Submission Instructions below.

Submission Instructions

How to submit RMarkdown and HTML files (due Sunday, December 12)

1. Create a GitHub repository that includes the data used for the final project, the RMarkdown file and the compiled HTML file. If the data are too big to fit in the repository, make the data accessible somewhere online (google drive, downloadable link, etc). Inside the RMarkdown file at the top, include instructions on where to access the data. If we cannot access your work or links because these directions are not followed correctly, we will not grade your work. Be sure to include a detailed README file so we (and others) can navigate the files in the repository.
2. There should only one GitHub repository per team, but make sure the names of all group members are inside the RMarkdown file at the top as well as the README file.
3. Email your TA instructions on where to access the data and the location of your GitHub repository.

How to submit the Peer Assessment (due Sunday, December 12)

Each individual team member needs to fill out this google form for the peer evaluation. Your individual project score will take into account the self and peer assessment.

How to submit the Website and Screencast (due Sunday, December 12)

Fill out this google form to submit the links to the website and screencast. *If we cannot access the website or screencast, we cannot grade it*.

Grading

The final project is graded in two parts:

1. Final Project Part I (worth 10% of total course grade). This portion represents the Team Registration and Final Project Proposal which is due November 1 by 11:59pm.
2. Final Project Part II (worth 25% of total course grade). This portion will be split into two sub-portions:
    - 80% of the Final Project Part II will be based on the RMarkdown and HTML files in the GitHub repository. This includes the quality of the data analysis and R code, the complexity and level of difficulty of the project, and completeness and overall functionality of the analysis. This sub-portion (and peer assessment) is due Sunday, December 12 by 11:59pm.
    - 20% of the Final Project Part II will be based on the website and screencast and the quality of their storytelling aspects. This sub-portion is due Sunday, December 12 by 11:59pm.

Individual project scores will also be determined by peer evaluations.

Example Final Projects

Here are some examples of successful final projects. **Note**: These projects came from another course we taught on Data Science similar to this one except the previous course used Python, not R.

1. **Predicting Hubway bike/dock availability** (Website, Screencast)
2. **Across the Bay 10K Race** (Website, Screencast)
3. **Predicting Citation Counts of arXiv Papers** (Website, Screencast)

Course Schedule & Assessment of Student Learning

Lecture	Date	Topics	Assignments
1	30-Aug	Introduction to course, R, RStudio, RMarkdown	HW1 Assigned
2	1-Sep	Introduction to Git, GitHub and homework submission Introduction to data wrangling, tidy data, importing data, reshaping data, combining tables	
3	6-Sep	Labor Day – No Class	
4	8-Sep	Data wrangling continued	
5	13-Sep	Dates and times, web scraping	
6	15-Sep	String processing	
7	20-Sep	String processing continued	
8	22-Sep	Introduction to ggplot2	HW2 Assigned
9	27-Sep	Maps and infographics	
10	29-Sep	Data visualization principles	
11	4-Oct	Data visualization principles continued	
12	6-Oct	Discrete probability	HW3 Assigned
13	11-Oct	Indigenous Peoples' Day – No Class	
14	13-Oct	Continuous probability	
15	18-Oct	Bayesian inference	
16	20-Oct	Election forecasting	
17	25-Oct	Introduction to machine learning	HW4 Assigned
18	27-Oct	Machine learning continued	
19	1-Nov	Machine learning continued	Project proposals due
20	3-Nov	Machine learning continued	
21	8-Nov	Machine learning continued	
22	10-Nov	Machine learning continued	
23	15-Nov	Machine learning continued	
24	17-Nov	Introduction to Shiny	HW5 Assigned
25	22-Nov	Shiny layouts and reactivity	
26	24-Nov	Thanksgiving Recess – No Class	
27	29-Nov	Shiny layouts and reactivity continued	
28	1-Dec	Shiny case study	
29	6-Dec	Advanced RMarkdown	
30	8-Dec	Text analysis	
31	13-Dec	SQL and R	
32	15-Dec	Project screencasts Next steps in data science	

- Please note, session topics, assignments, assignment deadlines, and activities may be subject to change during the course.