HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH

BST 219: Core Principles of Data Science
Fall 2025
T/Th 9:45-11:15am, Kresge G2

Instructor Information
Faculty
Dr. Heather Mattie
Lecturer on Biostatistics, Co-Director of the Health Data Science Master's Program,
Director of Community Engagement and Development
Department: Biostatistics, HSPH
Office: Building 1, 4th Floor, Room 421A
Email: hemattie@hsph.harvard.edu
Office Hours: Mondays 2-3pm (in person), Wednesdays 1-2pm (Zoom)
Note: email is **by far** the best way to contact me. I rarely check Canvas messages.

Teaching Fellows
Carmen Rodriguez Cabrera - crodriguezcabrera@g.harvard.edu
Claire Chu – clairechu@hsph.harvard.edu
Sajia Darwish – sajiadarwish@g.harvard.edu

Office Hours

Day	Teaching Staff	Time	Location
TBD	Heather	TBD	Building 1, 4th floor, room 421A
TBD	Carmen	TBD	TBD
TBD	Claire	TBD	TBD
TBD	Sajia	TBD	TBD

Lab: TBD, Zoom

Credits
5 credits

Course Purpose and Description
Modern technology has led to the generation of unprecedented amounts of data, prompting the need to train researchers to leverage data for decision-making in public health and medicine. This course assumes no prior knowledge and serves as a gentle, practical introduction to data wrangling, visualizing, and modeling data using the R statistical programming language. We also emphasize the importance of reproducible research, effective data science communication, and the risk of algorithmic bias.

Pre-Requisites
None

Course Learning Objectives
Upon successful completion of this course, you should be able to:

- Write reproducible code using the statistical programming language R
- Clean and wrangle data for downstream analysis
- Perform exploratory data analysis, including visualizations
- Apply machine learning models for regression and classification
- Interpret and communicate key results

Course Readings
Required: students are encouraged to read the lecture documents and other resources available on the course Canvas site and the course GitHub repository.

Suggested:

1. R for Data Science (2nd edition, 2023; open access)
2. Storytelling with Data (available via Harvard Hollis)
3. ggplot2: Elegant Graphics for Data Analysis (3rd edition; open access)
4. Happy Git and GitHub for the useR (mainly chapter 12; open access)
5. An Introduction to Statistical Learning with R (aka ISR; free download)
6. R for Health Data Science (open access)

Course Structure
This course will be held synchronously and in-person. We encourage students to attend class and participate, but it is not required, and all lectures and lab sessions will be recorded and available on the course Canvas site.

The final grade for this course will be based on:

- 5 Homework Assignments (60%)
- 1 Take-home Midterm (15%)
- 1 Final group project (25%)

Participation
Attendance and participation are not graded components of this course but are highly encouraged in order to get as much as possible out of this course.

Homework Assignments (60%)
All homework assignments will involve **writing code and communicating results**. Students must **submit the RMarkdown file (.Rmd file) and knitted html file (.html file)** associated with each assignment in their individual GitHub 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. Students do not 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 firm.

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.

The TFs 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 Dr. Mattie via email.

Students may work together on assignments but all responses to text questions must be in their own words and not copied from another student's assignment or from a generative AI tool like ChatGPT.

In this course, we recognize the growing presence and impact of generative Artificial Intelligence (AI) tools (e.g., ChatGPT, Gemini, etc.) in academic and professional environments. These tools can be powerful aids in the learning process when used responsibly. The following guidelines are designed to ensure that generative AI is used ethically and effectively in your course assignments:

Permissible Use

1. Assistance: You may use generative AI tools to assist in brainstorming ideas, drafting outlines, generating code snippets, or seeking clarification on complex topics.
2. Supplemental Learning: These tools can serve as supplemental resources to explain concepts and provide additional examples that may aid in your understanding.

Impermissible Use

1. Plagiarism: You must not directly copy and submit AI-generated content as your own work. This includes text, code, images, or any other type of content that has not been adequately personalized or appropriately cited.
2. Substitution for Learning: Relying on AI tools to complete assignments in lieu of engaging with the course material is discouraged. Your primary goal should be to develop a deep understanding of the subject matter.

Transparency and Citation

1. Citation: Cite AI tools appropriately in your bibliography or reference sections in accordance with the citation style prescribed for the course.

Academic Integrity

1. Originality: Ensure that your submissions reflect your own understanding and synthesis of the material. AI tools should not replace your critical thinking and analytic skills.
2. Integrity: Any use of AI tools should uphold the principles of academic integrity outlined by Harvard TH Chan's academic policies.

By adhering to these guidelines, you will be able to harness the benefits of generative AI while maintaining the highest standards of academic integrity and personal learning. If you have any questions regarding the appropriate use of these tools, please feel free to reach out to Dr. Mattie for clarification.

Take-home Midterm (15%)

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 their individual GitHub repository 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. Students may not use any form of generative AI on 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 (25%)

Students will work in small groups on a month-long data science project. The goal of the project is to go through the complete data science process to answer an assigned prompt. You will be given a dataset and series of questions to answer. You will design your visualizations, provide summary statistics, build machine learning models, and communicate results. A full description is available on the course website.

Technical Information

Assistance

Canvas

If the issue is Canvas-related (e.g., you can't figure out how to use something or a feature seems broken), first try the documentation located under the Help menu found on the left-hand side of each Canvas page. If the issue is not covered there, contact Instructure directly, also via the Help menu. You can e-mail, text, or speak live with them at any time day or night. If you cannot access Canvas to view the Help menu, you can reach Instructure by phone at +1 (844) 326-4466.

Zoom
For help with Zoom video conferencing, first check the variety of video tutorials and online help at https://support.zoom.us. In addition, you may contact the Helpdesk by emailing helpdesk@hsph.harvard.edu or calling +1 (617) 432-HELP (4357).

Harvard-Specific Issues
If the issue seems Harvard-specific (e.g., HUID or myHarvardChan authentication, email not working, etc.), contact the Helpdesk at helpdesk@hsph.harvard.edu or +1 (617) 432-HELP (4357).

Other
If you are unsure where to turn, but think the issue is related to technology or the course lecture videos, contact the Helpdesk as noted above.

Technical Requirements

- Reliable, high-speed internet connection
- Your laptop must meet the minimum technical requirements found on the Student Guide page.
- Modern and updated web browser (e.g., a recent version of Firefox or Chrome)
- Web camera and microphone (integrated into computer or USB peripheral)
- Please contact helpdesk@hsph.harvard.edu with questions.

Please note that while it is possible to access most of the course materials via mobile and wireless devices, video conferencing and other bandwidth-intensive sessions will have the greatest reliability on a wired high-speed connection.

Harvard Chan Policies and Expectations
Inclusivity Statement
Diversity and inclusiveness are fundamental to public health education and practice. Students are encouraged to have an open mind and respect differences of all kinds. I share responsibility with you for creating a learning climate that is hospitable to all perspectives and cultures; please contact me if you have any concerns or suggestions.

Bias Related Incident Reporting
The Harvard Chan School believes all members of our community should be able to study and work in an environment where they feel safe and respected. As a mechanism to promote an inclusive community, we have created an anonymous bias-related incident reporting system. If you have experienced bias, please submit a report here so that the administration can track and address concerns as they arise and to better support members of the Harvard Chan community.

Title IX
The following policy applies to all Harvard University students, faculty, staff, appointees, or third parties: Harvard University Sexual and Gender-Based Harassment Policy.
Procedures For Complaints Against a Faculty Member
Procedures For Complaints Against Non-Faculty Academic Appointees

Academic Integrity
Each student in this course is expected to abide by the Harvard University and the Harvard T.H. Chan School of Public Health School's standards of Academic Integrity. All work submitted to meet course requirements is expected to be a student's own work. In the preparation of work submitted to meet course requirements, students should always take great care to distinguish their own ideas and knowledge from information derived from sources.

Students must assume that collaboration in the completion of assignments is prohibited unless explicitly specified. Students must acknowledge any collaboration and its extent in all submitted work. This requirement applies to collaboration on editing as well as collaboration on substance.

Should academic misconduct occur, the student(s) may be subject to disciplinary action as outlined in the Student Handbook. See the Student Handbook for additional policies related to academic integrity and disciplinary actions.

Accommodations for Students with Disabilities
Harvard University provides academic accommodations to students with disabilities. Any requests for academic accommodations should ideally be made before the first week of the semester, except for unusual circumstances, so arrangements can be made. Students must register with the Local Disability Coordinator in the Office for Student Affairs to verify their eligibility for appropriate accommodations. Contact Colleen Cronin ccronin@hsph.harvard.edu in all cases, including temporary disabilities.

Religious Holidays, Absence Due to
According to Chapter 151c, Section 2B, of the General Laws of Massachusetts, any student in an educational or vocational training institution, other than a religious or denominational training institution, who is unable, because of his or her religious beliefs, to attend classes or to participate in any examination, study, or work requirement on a particular day shall be excused from any such examination or requirement which he or she may have missed because of such absence on any particular day, provided that such makeup examination or work shall not create an unreasonable burden upon the School. See the student handbook for more information.

Grade of Absence from Examination
A student who cannot attend a regularly scheduled examination must request permission for an alternate examination from the instructor in advance of the examination. See the student handbook for more information.

Final Examination Policy
No student should be required to take more than two examinations during any one day of finals week. Students who have more than two examinations scheduled during a particular day during the final examination period may take their class schedules to the director for student affairs for assistance in arranging for an alternate time for all exams in excess of two. Please refer to the student handbook for the policy.

Course Evaluations
Constructive feedback from students is a valuable resource for improving teaching. The feedback should be specific, focused and respectful. It should also address aspects of the course and teaching that are positive as well as those which need improvement.

Completion of the evaluation is a requirement for each course. Your grade will not be available until you submit the evaluation. In addition, registration for future terms will be blocked until you have completed evaluations for courses in prior terms.

Course Schedule

Lecture	Date	Topics	Assignments
1	2-Sep	Introduction to course, R, RStudio, RMarkdown	HW1 Assigned Due Friday, September 19th
2	4-Sep	Introduction to Git, GitHub and homework submission	
3	9-Sep	Basic R: data types, vectors, sorting, indexing	
4	11-Sep	Basic R: basic plots and programming basics	
5	16-Sep	Data wrangling: `dplyr` function, importing data, reshaping data	HW2 Assigned Due Friday, October 3rd
6	18-Sep	Data wrangling: combining data frames, parsing dates and times	
7	23-Sep	Data visualization: Introduction to `ggplot2`, fine-tuning visualizations	HW3 Assigned Due Friday, October 24th
8	25-Sep	Data visualization: data visualization principles, data visualization case study: vaccines	
9	30-Sep	Data visualization: Gapminder case study, time series plots, transformations	
10	2-Oct	Data visualization: Comparing distributions, smooth density plots	
11	7-Oct	Maps: the maps package, plotting global life expectancy, plotting US state murder rates	
12	9-Oct	Maps: COVID-19 case study	
13	14-Oct	Summarizing data: numerical summaries, frequency tables, the `tableone` package, reporting missing data	
14	16-Oct	Managing missing data: types of missingness, simple imputation, advanced imputation, survey skip patterns	
15	21-Oct	Hypothesis testing: tests for one group, tests for two groups, tests for more than two groups, nonparametric tests	HW4 Assigned Due Friday, November 7th
16	23-Oct	Introduction to linear regression	
17	28-Oct	Linear regression continued	
18	30-Oct	Introduction to logistic regression	
19	4-Nov	Logistic regression continued	
20	6-Nov	Introduction to machine learning	Midterm assigned November 7th – 14th
21	11-Nov	Veterans' Day – No Class	
22	13-Nov	Machine learning continued	
23	18-Nov	Machine learning continued	HW5 Assigned Due Friday,

			December 12th
24	20-Nov	Machine learning continued	
25	25-Nov	Thanksgiving Recess – No Class	
26	27-Nov	Thanksgiving Recess – No Class	
27	2-Dec	Machine learning continued	
28	4-Dec	Machine learning continued	
29	9-Dec	Machine learning continued	
30	11-Dec	Machine learning continued	
31	16-Dec	Introduction to Shiny	
32	18-Dec	Shiny continued Next steps in data science	Final Project Due

Note: all lecture topics and assignment dates are subject to change as they rely on the pacing of the course.