HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH

# BST219: Core principles of Data Science
## Fall 2023 - MW 11:30am - 1:00pm, Kresge 502

## Instructor Information

**Faculty**
Dr. Dongdong Li
Department of Biostatistics, HSPH and Department of Population Medicine, HMS
Email: dongdongli@hsph.harvard.edu

**Teaching Assistants**
Luke Benz - lukebenz@g.harvard.edu

**Lecture**

- Mondays and Wednesdays 11:30-1:00pm EST
- Room: Kresge 502

**Office Hours**
All office hours will be held in-person **and** online via Zoom link posted on Canvas.

| Day | Time | Staff | Location |
|---|---|---|---|
| Monday | 3:30-4:30pm | Luke | Building 2, 4th floor, 401 |
| Tuesday | 2:30-3:30pm | Dongdong | Building 2, 4th floor, 437F (401 on August 29) |
| Wednesday | 1:00-2:00pm | Dongdong | Building 2, 4th floor, 437F (401 on August 30) |
| Thursday | 1:00-2:00pm | Luke | Building 2, 4th floor, 401 |

## Lab

Fridays 9:45-11:15am
**Note that there will not be labs every week**. We'll post announcements on the course website and canvas.
Lab Room (Zoom link on Canvas):

- (default) Sep 1 - Dec 13 except for Sep 29, Oct 6, Nov 3: Kresge LL6
- for Sep 29 and Oct 6: FXB G03
- for Nov 3: FXB G12

## Credits
5 credits

## Course 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 and effective data science communication.

## Pre-Requisites
None

## 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

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

## 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:
- 4 Homework Assignments (40%)
- 1 Take-home Midterm (25%)
- 1 Final project (35%)

## Classroom Participation:
Attendance and participation are not graded components of this course.

## 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. 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 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. Li 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 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

Students may ask questions about the assignments during lecture, but we ask that any questions about grading be directed to the TFs or Dr. Li 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 assignments should reach out to Student Affairs at StudentAffairs@hsph.harvard.edu. A staff member will work with you and Dr. Li 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 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, run statistical analyses, 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)
- Throughout this program, you will be using VDI to access certain applications (e.g. EndNote and JMP); in turn, your computer must meet the minimum hardware and software requirements displayed on the VDI page.
- 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.

**Lecture Schedule**

| Lecture | Date | Topics | Assignments |
|---|---|---|---|
| 1 | 28-Aug | Introduction to course, R, RStudio, RMarkdown | |
| 2 | 30-Aug | Introduction to Git, GitHub, and homework submission | HW1 Assigned |
| 3 | 4-Sep | Labor Day - No Class | |
| 4 | 6-Sep | Basic R, data types and vectors | |
| 5 | 11-Sep | Sorting, vector arithmetic, and indexing | |
| 6 | 13-Sep | Basic data wrangling | |
| 7 | 18-Sep | Basic plots and importing data | |
| 8 | 20-Sep | Programming basics | HW2 Assigned |
| 9 | 25-Sep | Introduction to ggplot2 | |
| 10 | 27-Sep | Gapminder | |
| 11 | 2-Oct | Maps and infographic | |
| 12 | 4-Oct | Data visualization principles | |
| 13 | 9-Oct | Indigenous People's Day - No Class | |
| 14 | 11-Oct | Data visualization principles - continued | HW3 Assigned |
| 15 | 16-Oct | Advanced data wrangling | |
| 16 | 18-Oct | Advanced data wrangling - continued | |
| 17 | 23-Oct | Date and times, web scraping | |
| 18 | 25-Oct | String processing | |
| 19 | 30-Oct | Regression | |
| 20 | 1-Nov | Regression - continued | |
| 21 | 6-Nov | Introduction to machine learning | HW4 Assigned |
| 22 | 8-Nov | Machine learning - continued | |
| 23 | 13-Nov | Machine learning - continued | |
| 24 | 15-Nov | Machine learning - continued | |
| 25 | 20-Nov | Machine learning - continued | |
| 26 | 22-Nov | Thanksgiving Recess - No Class | |
| 27 | 27-Nov | Machine learning - continued | |
| 28 | 30-Nov | Machine learning - continued | |
| 29 | 4-Dec | Machine learning - continued | |
| 30 | 6-Dec | Introduction to Shiny | |
| 31 | 11-Dec | Shiny - continued | |
| 32 | 13-Dec | Shiny - continued<br>Next steps in data science | Final project due |