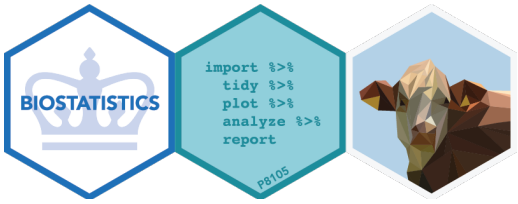

P8105: Data Science I

COURSE DESCRIPTION

Contemporary biostatistics and data analysis depends on the **mastery of tools for computation, visualization, dissemination, and reproducibility** in addition to proficiency in traditional statistical techniques. The goal of this course is to provide training in the elements of a complete pipeline for data analysis. It is targeted to MS, MPH, and PhD students with some data analysis experience.

LEARNING OBJECTIVES

Students who successfully complete this course will:

- Utilize best practices for project organization;
- Implement analyses in a reproducible way;
- Use GitHub to publish and disseminate analyses;
- Integrate the principles of data organization into their analyses;
- Easily produce static and interactive graphics;
- Collect data from online sources using web-scraping.

INSTRUCTOR

Jeff Goldsmith, PhD
Associate Professor of Biostatistics
Email: <ajg2202@cumc.columbia.edu>

TEACHING ASSISTANTS

Amy Pitts (ajp2257)
Derek Lamb
Yuxuan Du
Ru Jin Lim
Ruiqi Xue
Shaolei Ma
Gustavo Garcia-Franceschini (geg2145)
Peng Su
Xiaoting Tang
Lu Qiu
Yifei Liu

CLASS SESSIONS

Tuesdays and Thursdays 10:00 - 11:20; (Alumni Auditorium)

SLACK, DISCUSSION BOARD, OFFICE HOURS, EMAIL

There are several ways to get help answering course related questions. Slack channels will be created for each topic, and are a way to ask and answer questions in real time during class sessions. The Courseworks page for the class includes a discussion board, and we encourage students to proactively use this as a way to get help and to help others.

For more complex issues, office hours are more appropriate. These will be held:

- Tues-Fri 11:30-12:50 in ARB 657
- Mon & Wed 4:30-5:50 via zoom (link available via Courseworks)
- Tues & Thurs 7:30-8:50 via zoom (link available via Courseworks)

Email should be used to address questions regarding course structure or policy; content-related questions will generally be referred to the discussion board or office hours.

PREREQUISITES
Experience in R programming (or programming in another language) and data analysis is **recommended but not required**. A laptop with R installed is required and should be brought to every class session.

RECOMMENDED REFERENCES (note: there are no required texts for this course)
The Internet (stackoverflow; google; blog posts; twitter)
R for Data Science by G. Grolemund and H. Wickham
Exploratory Data Analysis with R by R Peng
R Programming for Data Science by R Peng.
Advanced R by H. Wickham

ASSESSMENT AND GRADING POLICY
Student grades will be based on:
Homework Assignments..... 30%
Midterm Project..... 35%
Final Project..... 35%
Questions regarding the grading of HW assignments must be raised within a week of the assignment being returned.

Homework assignments will be due following the completion of each course topic. Only electronic submissions will be accepted. Collaboration on homework assignments is governed by the course policy on collaboration; acceptable use of LLMs / ChatGPT is covered by this course policy. Late homework will not be accepted. Unclear or disorganized homework will have points removed, even if the content is correct.

The midterm project will focus on demonstrating proficiency in the topics covered in the first half of the course (R, R Markdown, data wrangling, exploratory analysis, and plotting). Collaboration on the midterm project is **strictly prohibited**.

The final project will consist of a complete analytic pipeline, starting with getting data and ending with a polished report, website, and screencast. This will be a group project, and group members will collaborate on the project using GitHub.

SOFTWARE USE
We will use R and RStudio.

COURSE WEBSITE
The course website contains lecture materials, homework assignments, supplementary materials, helpful links, and project information. It can be accessed at www.p8105.com.

COURSEWORKS
Individual- and course-level activity data are collected and maintained in CourseWorks, Panopto and other educational technology tools, and may be analyzed or monitored by the course faculty, teaching team, and/or the Office of Education to improve course experience and student support. Details about the information collected can be found here.

COURSE STRUCTURE

Class sessions will be lectures, delivered using a mix of static content and live demonstrations.

COURSE SCHEDULE

Lecture 1: What is data science?

Learning Objectives:

- Define "data science" and its role in public health research

Required Reading:

- "50 Years of Data Science" by David Donoho
- The Data Science Venn Diagram
- 'Janitor Work' vs 'Data Carpentry'
- 'What have you tried?' and a follow-up by the author
- *R Programming for Data Science:*
  - History and Overview of R
  - Getting Started with R

Homework: Assignment 0 (for details on all assignments, see below)

Lecture 2: Best Practices

Learning Objectives:

- Use best practices for coding, including commenting and human-readable naming structures.

Required Reading:

- *R for Data Science:*
  - 4) Workflow: basics, scripts, projects
- R Studio Code Diagnostics
- BEH Commandments for Variable Names
- Using R Projects

Homework: Assignment 1

Lecture 3: Writing with data

Learning Objectives:

- Implement basic analyses using R Markdown and R Notebooks. Export analysis reports into several formats.

Required Reading:

- *R for Data Science:*
  - 27.1 – 27.4) R Markdown
  - 29.1 – 29.5) R Markdown Formats
  - 30) R Markdown Workflow

Homework: Assignment 1

Lecture 4: Version control and dissemination

Learning Objectives:

- Create local and remote Git repositories, and integrate with R Projects. Use commits for version control.

Required Reading:

- Happy Git and GitHub for the useR

Homework: Assignment 1

Lecture 5: Data import

Learning Objectives:

- Read data into R from a variety of sources
- Parse variable types

Required Reading:

- *R Programming for Data Science:*
  - Getting Data In and Out of R
- *R for Data Science:*
  - 11) Data Import

Homework: Assignment 2

Lecture 6: Data manipulation

Learning Objectives:

- Clean and organize data using dplyr verbs and piping.

Required Reading:

- *R Programming for Data Science:*
  - Managing Data Frames with the dplyr package
- *R for Data Science:*
  - 12.1 – 12.5) Tidy Data
  - 18) Pipes

Homework: Assignment 2

Lecture 7: Tidy data and relational datasets

Learning Objectives:

- Explain principles of "tidy" data.
- Use relational databases; merging datasets

Required Reading:

- *R Programming for Data Science:*
  - Getting Data In and Out of R
  - Managing Data Frames with the dplyr package
- *R for Data Science:*
  - 11) Data Import
  - 12.1 – 12.5) Tidy Data
  - 18) Pipes

Homework: Assignment 2

Lecture 8: Data visualization

Learning Objectives:

- Create effective graphics using ggplot using the grammar of graphics. Implement best practices for effective graphical communication.

Required Reading:

- "A Layered Grammar of Graphics" by Hadley Wickham
- *R for Data Science:*
  - 3) Data Visualization
  - 28) Graphics for Communication

Homework: Assignment 3

Lecture 9: Data visualization

Learning Objectives:

- Create effective graphics using ggplot using the grammar of graphics. Implement best practices for effective graphical communication.

Required Reading:

- "A Layered Grammar of Graphics" by Hadley Wickham
- *R for Data Science:*
  - 3) Data Visualization
  - 28) Graphics for Communication

Homework: Assignment 3

Lecture 10: Exploratory analysis

Learning Objectives:

- Conduct exploratory analyses using dplyr verbs (group_by and summarize).

Required Reading:

- *R for Data Science:*
  - 7) Exploratory analysis

Homework: Assignment 3

Lecture 11: Case study

Learning Objectives:

- Pull together skills learned through this point
- Produce a complete analysis and written summary

Lecture 12: Reading Data from the Web

Learning Objectives:

- Gather data from online sources (i.e. "scrape") using APIs, rvest and httr.

Lecture 13: Strings and factors

Learning Objectives:

- Edit / manipulate strings; take control of factors

Lecture 14: Websites

Learning Objectives:

- Publish a personal website using GitHub Pages.

Required Reading:

- GitHub Pages

Homework: Assignment 4

Lecture 15: Plot.ly and dashboards

Learning Objectives:

- Create interactive graphics using plot.ly
- Design an data dashboard using flexdashboard

Homework: Assignment 4

Lecture 16: Writing R functions

Learning Objectives:

- Create simple R functions to abstract common processes.

Required Reading:

- *R Programming for Data Science:*
  - Functions
  - Scoping Rules of R
- *R for Data Science:*
  - 19) Functions

Homework: Assignment 5

Lecture 17: Simulating data

Learning Objectives:

- Simulate datasets in R. Use loops, apply functions, and map functions.

Required Reading:

- *R Programming for Data Science:*
  - Simulation
  - Loop functions

Homework: Assignment 5

Lecture 18: Simulation

Learning Objectives:

- Use loop and apply functions to simulate data. Explore statistical properties of usual estimate methods using simulations.

Required Reading:

- *R Programming for Data Science:*
  - Simulation
  - Loop functions

Homework: Assignment 5

Lecture 19: Linear and generalized linear models

Learning Objectives:

- Review fundamentals of linear and generalized linear models. Fit models in R and tidy results for further analysis.

Required Reading:

- *Introduction to Statistical Learning with R*
  - Chapter 3.1-3.3
  - Chapter 4.1.-4.3

Homework: Assignment 6

Lecture 20: Cross validation

Learning Objectives:

- Use cross validation to assess predictive value of a model. Implement CV using tools for iteration.

Required Reading:

- *Introduction to Statistical Learning with R*
  - Chapter 5.1

Homework: Assignment 6

Lecture 21: Bootstrapping

Learning Objectives:

- Implement the bootstrap to obtain inference in non-standard cases using tools for iteration.

Required Reading:

- *Introduction to Statistical Learning with R*
  - Chapter 5.2

Homework: Assignment 6

Lecture 22: Extra topics

Learning Objectives:

Required Reading:

Homework: Assignment 7

Lecture 23: Extra topics

Learning Objectives:

Required Reading:

Homework: Assignment 7

ASSIGNMENTS

Assignment 0	
L1	Assignment 0 covers the installation of software and creation of accounts.

Assignment 1	
L2-L4	Assignment 1 covers basic R coding, including variable assignments, data manipulation, and the use of basic functions. Submissions will use the R Markdown format to ensure reproducibility, and best practices for clarity.

Assignment 2	
L5-L7	Assignment 2 covers data input and output; principles of data cleaning; and implementation of data cleaning using dplyr.

Assignment 3	
L8-L10	Assignment 3 covers exploratory data analysis. Students are expected to produce reasonable summaries of data, including both tables and graphics, and accompany these with clearly-written text describing the results.

Assignment 4	
L14-L15	Assignment 4 covers dashboards and websites. Students will develop a professional website including their contact information and highlighting their work / CV. This website will also link to a dashboard.

Assignment 5	
L16-L18	Assignment 5 covers simulation and looping. Students will conduct simulation experiments to explore basic statistical properties, and will illustrate these graphically and in words.

Assignment 6	
L19-L21	Assignment 6 covers linear models.

Assignment 7	
L22-L23	Assignment 7 covers extra topics.

MAILMAN SCHOOL POLICIES AND EXPECTATIONS

Students and faculty have a shared commitment to the School's mission, values and oath. mailman.columbia.edu/about/mission-history

Academic Integrity

Students are required to adhere to the Mailman School Conduct and Community Standards, which includes the Code of Academic Integrity. Columbia Mailman and Columbia University take academic integrity very seriously. This instructor and course are no different. Should any student be suspected of an academic integrity violation, there will be a report submitted to the Center for Student Success & intervention/Student Conduct. After these offices conduct their process, if a student is found responsible for violating an academic integrity policy (see Standards & Discipline/Academic Violations and Student Honor Code & Professional Guidelines), they will be assigned a grade penalty, with a possible outcome being a 0% on the assignment.
Please review the university, school, and course policies, as you are responsible for behaving according to the outlined expectations.

Personal Support

Students sometimes experience life challenges that require additional support and connection to resources. If you are experiencing difficult circumstances, please reach out for help and support. Student Support Services in the Office of Student Affairs is poised to connect with students, provide resource referrals, and provide ongoing, non-clinical support. They are a good place to start if you do not know where to turn. If you would like to connect with this support resource, please contact Sarah Tooley (st3146@cumc.columbia.edu).

Disability access

In order to receive disability-related academic accommodations, students must first be registered with the Office of Disability Services (ODS). Students who have or think they may have a disability are invited to contact ODS for a confidential discussion at 212.854.2388 (V) 212.854.2378 (TTY), or by email at disability@columbia.edu. If you have already registered with ODS, please speak to your instructor to ensure that they have been notified of your recommended accommodations by Meredith Ryer (mr4075@cumc.columbia.edu), Assistant Director of Student Support and Mailman's liaison to the Office of Disability Services.

Bias Incidents

Our community at Columbia University's Mailman School of Public Health is committed to creating an inclusive working, learning, and living environment where all are respected. The occurrence of bias related incidents, involving conduct, speech, or expressions reflecting prejudice are an opportunity for learning and growing as a community.

As part of our efforts to create as inclusive a community as possible, when bias incidents occur at Columbia, we provide an opportunity for those involved to engage in education, advocacy and conversation. In this way, we work to address the incident and minimize the potential for future occurrences. Our community's tools to address bias include a reporting process and the Bias Incident Resource Team, plus resources within schools and various offices. You can access information about the Bias Reporting Process and FAQs here.