**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:

* Integrate the principles of data organization into their analyses;
* Easily produce static and interactive graphics;
* Implement analyses in a reproducible way;
* Use Github to publish and disseminate analyses;
* Develop usable software packages in R;
* Collect data from online sources using web-scraping.

**INSTRUCTOR**

Jeff Goldsmith, PhD

Assistant Professor of Biostatistics

Email: <ajg2202@cumc.columbia.edu>

**TEACHING ASSISTANTS**

Zixuan Chen <zc2361@cumc.columbia.edu>

Julia Wrobel <jw3134@cumc.columbia.edu>

Alan Wu <aw3013@cumc.columbia.edu>

Emily Zabor <ecz2108@cumc.columbia.edu>

**CLASS SESSIONS**

Mondays 1:00 - 2:20 Hammer Health Sciences Building, Room 301

Wednesdays 1:00 - 2:20 Hammer Health Sciences Building, Room 312

**DISCUSSION BOARD, OFFICE HOURS, EMAIL**

There are several ways to get help answering course related questions. The course website through courseworks 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, in-person office hours (Tuesday 3:00-4:00 and Thursday 4:00-5:00 in ARB Room 638, and by appointment) may be more appropriate. 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*](http://r4ds.had.co.nz/) by G. Grolemund and H. Wickham

[*Exploratory Data Analysis with R*](https://leanpub.com/exdata) by R Peng

[*R Programming for Data Science*](https://leanpub.com/rprogramming) by R Peng.

[*R Packages*](http://r-pkgs.had.co.nz/r.html) by H. Wickham

[*Advanced R*](http://adv-r.had.co.nz/) by H. Wickham

**ASSESSMENT AND GRADING POLICY**

Student grades will be based on:

## Homework Assignments 50%

Midterm Project 20%

Final Project 30%

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 acceptable (and, in some cases, required), but all submissions must be completed independently and clearly indicate the submitter’s understanding of the material. Late homework will not be accepted. Unclear or disorganized homework may 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 and presentation. This will be a group project, and group members will collaborate on the project using Github.

**SOFTWARE USE**

We will use R and R Markdown; R Studio is recommended.

**COURES WEBSITE**

The course website contains lecture materials, homework assignments, supplementary materials, helpful links, and project information. It can be accessed at [www.jeffgoldsmith.com/DSI](http://www.jeffgoldsmith.com/DSI).

**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](http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram) * [‘Janitor Work’](http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?partner=rss&emc=rss&smid=tw-nytimesscience&_r=0) vs [‘Data Carpentry’](http://www.mimno.org/articles/carpentry/) * [‘What have you tried?’](http://mattgemmell.com/what-have-you-tried/) and a [follow-up](http://mattgemmell.com/hindsight/) 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: Base R** | |
|  | Learning Objectives:   * Explain variable assignment, data types, and basic functions in R.   Required Reading:   * The [swirl](http://swirlstats.com/) R package * *R Programming for Data Science:*   + R Nuts and Bolts   + Subsetting R Objects   Homework:   * Assignment 1 |

|  |  |
| --- | --- |
| **Lecture 3: 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   + 6) Workflow: scripts   + 8) Workflow: projects * [R Studio Code Diagnostics](https://support.rstudio.com/hc/en-us/articles/205753617-Code-Diagnostics) * [BEH Commandments for Variable Names](https://beh.columbia.edu/2016/08/03/commandments-for-variable-naming-and-data-management/) * [Using R Projects](https://support.rstudio.com/hc/en-us/articles/200526207-Using-Projects)   Homework:   * Assignment 1 |

|  |  |
| --- | --- |
| **Lecture 4: 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 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** | |
|  | 3Learning 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: 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 10: Case study** | |
|  | Learning Objectives:   * Pull together skills learned through this point * Produce a complete analysis and written summary |

|  |  |
| --- | --- |
| **Lecture 11: 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](http://happygitwithr.com/)   Homework:   * Assignment 4 |

|  |  |
| --- | --- |
| **Lecture 12: Team data science** | |
|  | Learning Objectives:   * Collaborate using code as a means of communication; contribute to shared repositories; implement code reviews.   Required Reading:   * [Happy Git and GitHub for the useR](http://happygitwithr.com/)   Homework:   * Assignment 4 |

|  |  |
| --- | --- |
| **Lecture 13: 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 14: Resampling** | |
|  | Learning Objectives:   * Use loop and apply functions to resample from a dataset. Explore statistical properties of resampling and compare to traditional inferential techniques   Required Reading:   * *R Programming for Data Science:*   + Simulation   + Loop functions   Homework:   * Assignment 5 |

|  |  |
| --- | --- |
| **Lecture 15: Writing R functions** | |
|  | Learning Objectives:   * Create simple R functions to abstract common processes.   Required Reading:   * *R Programming for Data Science:*   + Functions * *R for Data Science:*   + 19) Functions   Homework:   * Assignment 5 |

|  |  |
| --- | --- |
| **Lecture 16: Writing R functions** | |
|  | Learning Objectives:   * Implement complex R functions using multiple arguments and control structures   Required Reading:   * *R Programming for Data Science:*   + Scoping Rules of R * *R for Data Science:*   + 19) Functions   Homework:   * Assignment 6 |

|  |  |
| --- | --- |
| **Lecture 17: Writing R packages** | |
|  | Learning Objectives:   * Develop a simple R package, including documentation and dependencies.   Required Reading:   * *R Packages*   + Introduction   + Package Structure   Homework:   * Assignment 6 |

|  |  |
| --- | --- |
| **Lecture 18: Writing R packages (documentation, dependencies, devtools)** | |
|  | Learning Objectives:   * Create a complex R package including several functions. Deploy on GitHub.   Required Reading:   * *R Packages*   + Object Documentation   Homework:   * Assignment 6 |

|  |  |
| --- | --- |
| **Lecture 19: Interactive graphics** | |
|  | Learning Objectives:   * Use Shiny to implement a simple interactive graphic.   Required Reading:   * [Shiny Tutorial I](http://shiny.rstudio.com/tutorial/) * [Shiny Tutorial II](http://rstudio.github.io/shiny/tutorial/#reactivity)   Homework:   * Assignment 7 |

|  |  |
| --- | --- |
| **Lecture 20: Interactive graphics** | |
|  | Learning Objectives:   * Develop a complex graphic, including multiple tabbed panels and several user inputs, focusing on several aspects of a single dataset.   Required Reading:   * [Shiny Tutorial I](http://shiny.rstudio.com/tutorial/) * [Shiny Tutorial II](http://rstudio.github.io/shiny/tutorial/#reactivity)   Homework:   * Assignment 7 |

|  |  |
| --- | --- |
| **Lecture 21: GitHub Pages** | |
|  | Learning Objectives:   * Publish a personal website using GitHub Pages.   Required Reading:   * [GitHub Pages](https://pages.github.com/) |

|  |  |
| --- | --- |
| **Lecture 22: Regular expressions (tidytext)** | |
|  | Learning Objectives:   * Match patterns, subset, and analyze text.   Required Reading:   * *R for Data Science:*   + Strings * [Tidy Text with R](http://tidytextmining.com/)   Homework:   * Assignment 8 |

|  |  |
| --- | --- |
| **Lecture 23: Web scraping (APIs; rvest, httr)** | |
|  | Learning Objectives:   * Gather data from online sources (i.e. “scrape”) using APIs, rvest and httr.   Homework:   * Assignment 8 |

|  |  |
| --- | --- |
| **Lecture 24: Debugging code** | |
|  | Learning Objectives:   * Use built-in features to identify and solve problems in complex code environments.   Required Reading:   * *R Programming for Data Science:*   + Debugging * *Advanced R:*   + Performance   Homework:   * Assignment 9 |

|  |  |
| --- | --- |
| **Lecture 25: Profiling code** | |
|  | Learning Objectives:   * Explain common sources for slow code execution. Identify and implement solutions for speeding code.   Required Reading:   * *R Programming for Data Science:*   + Profiling R Code * *Advanced R:*   + Profiling   Homework:   * Assignment 9 |

**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-L9** | 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** | |
| **L11-L12** | Assignment 4 covers simulation and looping. Students will conduct simulation experiments to explore basic statistical properties, and will illustrate these graphically and in words. |

|  |  |
| --- | --- |
| **Assignment 5** | |
| **L13-L14** | 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** | |
| **L15-L18** | Assignment 6 covers writing R functions and R package development. Students will create R functions to achieve specific goals, taking into account specifications for argument names and function scope. Packages will be evaluated for use, clarity of documentation, and dissemination via GitHub. |

|  |  |
| --- | --- |
| **Assignment 7** | |
| **L19-L20** | Assignment 7 covers interactive graphics using Shiny. Students will create a tabbed panel with several user inputs to illustrate important features of a single dataset. |

|  |  |
| --- | --- |
| **Assignment 8** | |
| **L22-L23** | Assignment 8 covers text analysis and web scraping. Students will scrape a text-based dataset from an online source (e.g. twitter) and conduct relevant exploratory analyses and create relevant graphics (static and / or interactive). |

|  |  |
| --- | --- |
| **Assignment 9** | |
| **L24-L25** | Assignment 9 covers code debugging and profiling. Students will be provided a poor code example and will be expected to identify and correct coding issues that lead to errors, and to improve execution time of the implemented methods. |

**MAILMAN SCHOOL POLICIES AND EXPECTATIONS**

Students and faculty have a shared commitment to the School’s mission, values and oath. http://mailman.columbia.edu/about-us/school-mission/

*Academic Integrity*

Students are required to adhere to the Mailman School Honor Code, available online at http://mailman.columbia.edu/honorcode.

*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 s/he has been notified of your recommended accommodations by Lillian Morales (lm31@columbia.edu), the School’s liaison to the Office of Disability Services.

*Student Affairs*

The Office of Student Affairs (OSA) supports the needs of students who experience life challenges, which may disrupt their successful completion of a Public Health degree. Students' needs may manifest in such areas as their physical, mental, and/or emotional health; economic, family, and/or social stressors; difficulties resulting from adjustment to graduate-level work and/or transitioning to academia after time away from school; as well as other barriers to students' success. Students in need of support should reach out to OSA by phone (212-342-3128), [email](mailto:msph-osa@cumc.columbia.edu), or as a walk-in during office hours (8:00 a.m. – 6:00 p.m.; located on the 10th floor of ARB). Students may also directly access the resources and services of Student Health Services, Mental Health, Services, the Center for Student Wellness, and other supportive offices throughout CUMC directly through the offices’ websites, links to which can be found on the [Health and Wellness page](https://www.mailman.columbia.edu/people/current-students/student-resources/health-and-wellness) of the Mailman website.