DSST 289
Introduction to Data Science

Instructor	Erik Fredner (he/him)
Email	erik.fredner@richmond.edu
Classroom	105 The Refectory
Office	314 Carole Weinstein International Center
Office Hours	Wed. 14:30-16:30
Website	Blackboard
Updated	August 14, 2024

Land Acknowledgment

This class at the University of Richmond respectfully acknowledges the traditional custodians of the land we are on today, the Powhatan people, and pays respect to their elders past, present and emerging.

To learn more about the land on which the University of Richmond exists, I recommend students read the report "Knowledge of This Cannot Be Hidden" by Shelby M. Driskill and Dr. Lauranett L. Lee, which discusses both the University's geographic connection to the Powhatan people, as well as the presence of a burying ground for enslaved laborers on campus.

Accessibility

I strive to make this course accessible. If you encounter barriers to accessibility, please let me know as soon as possible.

Course Description

Topics will include techniques for collecting, organizing, analyzing, modeling, and presenting data. Applications to a variety of fields will be emphasized. Includes an extensive introduction to statistical programming. (See course catalog.)

Learning Goals

By the end of this course, students will be able to…

- Collect, manipulate, visualize, and explore data.
- Describe best practices for structuring tabular data.

- Articulate relationships between a dataset and analyses thereof.
- Use programming language documentation and cookbooks to solve problems.
- Understand key aspects of the R programming language and the `tidyverse`.
- Use the RStudio integrated development environment.

Prerequisites

- This course neither requires nor expects any prior experience with computer programming.
- While we discuss some statistical concepts, students only need experience with algebra.

Structure

This course has three units, each of which focuses on an aspect of data science:

Unit	Aspect
1	Visualization
2	Collection
3	Application

Materials

- Please bring a computer, pencil, and something to write on to each class.
- Class materials can be found on the course Blackboard site.
- You will submit assignments via Blackboard.

Inclusivity

I expect you to…

- Treat your classmates with respect.
- Support each other in your learning.
- Be patient.

If we spend time reviewing material that you already know, remember that it may be the first time some of your peers are learning this information. Students come to this class with different levels of knowledge. And everyone learns best in different ways.

Help

- We should have a lot of time during class to answer questions about course material.
  - I will also encourage you to collaborate with peers during class time.
- Come to office hours or schedule a meeting for any questions or personal concerns that cannot be addressed in class.
  - To schedule, email me with your availability at least one day before you'd like to meet.

- Note that I generally do not answer conceptual questions about homework before they are due. Just do your best, and we will discuss in class.

Grades

The tables below show the weights of each assignment and the associated ranges. I do not offer extra credit. I round fractional grades (e.g., 92.5% → 93%; 92.4% → 92%).

Assignment	Percentage
Participation	10%
Homework	10%
Exam 1	20%
Exam 2	20%
Exam 3	20%
Final Project	20%

Letter Grade	Range
A	93-100
A-	90-92
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
F	0-69

Schedule

- The schedule below outlines the major topics of the course.
- I reserve the right to change the syllabus as needed

  - I will inform you of any changes as far in advance as possible.

- **Bold** items are worth a large percentage of your grade.

Date	Topic
08-26	Intro
08-28	Tabular data
09-02	Grammar of graphics
09-04	Aesthetics and scales
09-09	Organizing data

Date	Topic
09-11	Summarizing data
09-16	Creating features
09-18	**Exam 1**
09-23	Creating data
09-25	Data feminism
09-30	Table joins
10-02	Table pivots
10-07	Review 1
10-09	Review 2
10-16	Review 3
10-21	Tidy models
10-23	Exam 2 review
10-28	**Exam 2**
10-30	Project assignments
11-04	Dates and times
11-06	Spatial data
11-11	Spatial joins
11-13	Time zones
11-18	**Exam 3**
11-20	Workshop 1
11-25	Workshop 2
12-02	Workshop 3
12-04	**Final projects**

Assignments

Homework

All class meetings, other than exam days, have a reading posted on our website. A few questions are included at the end of each reading. These must be completed before class. Please bring the written responses with you to class. You will upload completed questions to the course website.

You will be given an opportunity to begin working on homework during class. Your homework grade will be based on the proportion of assignments that you complete on time. If you put in a good-faith effort on the homework, it will be marked complete.

Participation

- Everyone may have up to two unexcused absences.

    – Additional unexcused absences will harm your participation grade.

- You may email me to request an excused absence.

    – Your request will be approved if it is for illness, hospitalization, death in the family, important religious holidays, or university activities (e.g., field trips, University-sponsored athletic events).

- You do *not* need to provide details.
    - Simply say, "I'm sick." You don't need to tell me, e.g., "I have strep throat."
- Students who receive excused absences may submit homework late.
- Please request an excused absence before missing class if possible.
    - If you request an excused absence *after* missing class, you must do so as soon as possible.
- If you cannot attend a scheduled exam or the final project presentation, please let me know as far in advance as possible.

Exams

Each of the three exams has two halves: a take-home open-book part and a closed-book in-class part.

The take-home will be distributed in advance of the in-class exam. Answers to the take-home should be submitted by the beginning of the in-class exam. A list of topics for the in-class exam will be posted on the course website.

Final Project

The final project will be due during the last week of class. The project will ask you to find or create a new dataset, and apply techniques learned throughout the class to analyze it. The project will take the form of a digital poster session and a one-page reflection. Detailed intructions will be provided later in the course.

Honor

This course is taught in accordance with the University of Richmond Honor Code, which can be accessed via The Honor Councils website. You are encouraged to collaborate on homework, but each student must contribute work to the group. On exams, cheating includes, but is not limited to, viewing another's work with or without their consent, or duplicating any portion of it. If you are found to have violated the Honor Code, you fail this course. If you ever have any questions about whether an action would be an honor violation, please ask.

Generative Artificial Intelligence

Generative artificial intelligence (GenAI) programs, especially large language models (LLMs), can be useful tools for coding. However, overreliance on LLMs or other resources where you can copy answers to programming problems directly (e.g., StackOverflow) will impede your learning.

Moreover, LLMs often answer questions incorrectly and/or using different methodologies than those we study in class. This is referred to as "hallucination". In order to be an effective user of these technologies, it is crucial for you to recognize when and how that happens.

Prohibited uses of GenAI

1. Submitting any model output, in part or in whole, as your own work. This includes code and writing.
2. Uploading any data used in this course (e.g., `.csv` files) to multimodal GenAI tools like ChatGPT.

3. Entering course materials (e.g., copy-pasting or paraphrasing homework questions) into GenAI tools.

Any of the above uses would be treated as violations of the honor code.

Permitted uses of GenAI

- If you get stuck on a problem that you find you cannot solve by consulting the course notes or other resources, you may consult GenAI tools.
- If you use GenAI tools like ChatGPT, you must **cite** your interaction with them.
    - This page explains how to share a link to a ChatGPT interaction.
    - An adequate citation would be: "I got this help from ChatGPT to solve this problem."

Communication

- I respond to email within 1 to 2 business days.
    - Do not expect responses over the weekend.
- If you have not received a response after 2 business days, feel free to write me again.
- For the most prompt response, schedule your email to arrive early in the morning (e.g., 7 or 8 AM).
    - Follow these instructions to schedule messages in Outlook.

Resources

The University of Richmond has many resources on campus that may help you succeed.

Weinstein Learning Center

The Weinstein Learning Center provides academic support services in several areas:

- Academic skills
- English language learning
- Quantitative resources
- Speech
- Technology
- Writing

Boatwright Library

Students may consult librarians to assist with their research, which may be especially useful for the final project. Use the Ask a Librarian service to reach librarians by email, phone, chat, text, or in person.

Counseling and Psychological Services

Mental health is crucial for academic success. Counseling and Psychological Services at the University of Richmond supports student success and enhances student well-being by providing comprehensive clinical services to currently enrolled, full-time, degree-seeking students.

Acknowledgments

Recommended Reading

The books below are not required reading. However, students who want to learn more about the main ideas in this course may want to consult some or all of them.

If you only look at one of these books, make it *R for Data Science* (2023).

Arnold, Taylor, and Lauren Tilton. *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text*. 1st ed. 2015. Quantitative Methods in the Humanities and Social Sciences. Cham: Springer International Publishing : Imprint: Springer, 2015. https://doi.org/10.1007/978-3-319-20702-5.

Bruce, Peter C., Andrew Bruce, and Peter Gedeck. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. Second edition. Sebastopol, CA: O'Reilly Media, Inc, 2020.

Cotton, Richard. *Learning R*. First Edition. Beijing ; Sebastopol, CA: O'Reilly, 2013.

Grolemund, Garrett. *Hands-on Programming with R*. First edition. Sebastopol, CA: O'Reilly, 2014.

Jockers, Matthew Lee, and Rosamond Thalken. *Text Analysis with R: For Students of Literature*. 2nd edition. Quantitative Methods in the Humanities and Social Sciences. Cham Heidelberg New York Dordrecht London: Springer, 2020.

Kuhn, Max, and Julia Silge. *Tidy Modeling with R: A Framework for Modeling in the Tidyverse*. Sebastopol, CA: O'Reilly Media, 2022.

Silge, Julia, and David Robinson. *Text Mining with R: A Tidy Approach*. First edition. Beijing ; Boston: O'Reilly, 2017.

Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Second edition. Beijing ; Sebastopol, CA: O'Reilly, 2023.