GSND 5345Q Fundamentals of Data Science

Spring (Jan-Feb), 2024

COURSE DESCRIPTION:

This class is an introduction to the ethics and essential computational tools and skills for data science. The course will cover command-line coding, literate programming, software development, version control, data wrangling and management, and visualization. The standards for open science, reproducibility, and ethical and responsible computing will also be discussed. Students be expected to use R and GitHub throughout this course.

COURSE OBJECTIVES:

Students who take this course will:

1. Gain experience with the fundamental tools and skills for data science
2. Develop an advanced understanding of the R programming language
3. Understand the principles and concepts surrounding reproducibility and open science
4. Discuss the ethical issues and potential bias in data and machine learning
5. Learn how to effectively plot and visualize data (know what to do and not to do!)

PREREQUISITES

An introductory course in statistics, biostatistics, epidemiology, or equivalent experience in statistical analysis is recommended (but not required). Programming experience in R is also recommended (again not required). Students without this experience will be encouraged to utilize the asynchronous resources provided at the end of this syllabus to obtain these skills before or during the course. Please contact Dr. Johnson to obtain a list of the required proficiencies.

COURSE FORMAT:

This class will be taught virtually using a synchronous remote modality, although students will be provided a classroom to gather for each lecture. A co-instructor will be present in the classroom for each lecture. Class will occur Mondays and Wednesdays from 10:00am-11:50am. Courses may also be recorded and made available for students who need to miss classes due to personal reasons, illness, or research related needs.

ZOOM LINK AND CLASSROOM:

Zoom Meeting ID for all sessions is 95798895550, with the password: 922124, or use the following direct link (the link is also available though the course GitHub page): https://rutgers.zoom.us/j/95798895550?pwd=VEd0Nk44MlFXemV6aXhFNXRwc0IvQT09.

Room XXX will also be available for the students to congregate for each lecture, with a co-intructor present.

FACULTY AND STAFF:

W. Evan Johnson, Ph.D.
Email: w.evan.johnson@rutgers.edu
Cell Phone: (801) 472-6951

Co-instructor (TBD)
Email: xxxx@rutgers.edu

OFFICE HOURS:

Instructor: Dr. Johnson will be available virtually by appointment only. Email or text him any time to set up a time to meet!

Teaching Assistant: The course TA will be available for in-person help and support on XXXX from XX:XX-XX:XX. XXXX will also be available (in person or virtually) by appointment on a limited basis.

GitHub REPOSITORY:

The course GitHub repository is located at: https://github.com/wevanjohnson/FDS2024. This page will contain all information in this syllabus plus more. Homework assignments and other information pertinent to this course will be posted on this web site, which will be updated frequently, so you should visit it regularly.

COURSE TEXTBOOKS:

We will use multiple text resources in this class. None are required, all are freely available online or can be purchased in hard-copy. Many of my materials are adapted from these resources (thanks to the authors for these):

1. *Modern Data Science with R*, 2nd edition, By Benjamin S. Baumer, Daniel T. Kaplan, Nicholas J. Horton, Chapman and Hall/CRC, 2021. https://mdsr-book.github.io/mdsr2e/
2. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*, 1st edition, By Rafael A. Irizarry, Chapman and Hall/CRC, 2020. https://rafalab.github.io/dsbook/
3. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, 2nd edition, By Hadley Wickham, Garrett Grolemund, O'Reilly, 2017 https://r4ds.had.co.nz
4. *Mathematical Foundations for Data Analysis*, By Jeff M. Phillips: https://mathfordata.github.io.

EVALUATION METHODS & COURSE GRADING

Assessment/Evaluation:

This course is a hands-on, project-based course. You will be graded based on homework assignments/mini projects (7 problem sets, each worth 100 points). There will be no final exam. Homework assignments and mini projects will be usually assigned at the beginning of each week and will be due by Wednesday of the week after the material is covered. The last homework assignment will include a presentation the last week of class. However, please plan to be flexible on due dates based on the material covered in class.

Course Grading:

Grade Scale:

≥	90%	85%	80%	75%	70%	60%
Grade	A	B+	B	C+	C	F

ATTENDANCE:

This course is being taught through a synchronous remote modality through Zoom. Attendance is mandatory; lecture recordings will only be available to students with university approved absences or pre-approved special circumstances. If you are sick or have any other justified reason to miss a lecture, please reach out to Dr. Johnson in advance and you will be reasonably accommodated.

WORKLOAD:

This is an 8-week, 2.0 credit class in the begining of Spring 2024. In general, you should expect four hours of in class each week, and two hours outside of class for every hour in class.

OTHER HELP:

I **strongly** encourage you to contact early me if you have difficulty with the material. This course builds on material from prior lectures, so do not fall behind! My job is to help you understand the material as well as possible, and I am flexible with meeting times.

ACADEMIC INTEGRITY:

You are expected to have read and follow the guidelines at the university's academic integrity website (http://academicintegrity.rutgers.edu). These principles forbid plagiarism and require that every Rutgers University student to:

- Properly acknowledge and cite all use of the ideas, results, or words of others
- Properly acknowledge all contributors to a given piece of work
- Make sure that all work submitted as his or her own in a course or other academic activity isproduced without the aid of unsanctioned materials or unsanctioned collaboration

- Treat all other students in an ethical manner, respecting their integrity and right to pursue their educational goals without interference. This requires that a student neither facilitate academic dishonesty by others nor obstruct their academic progress (reproduced from: ttp://academicintegrity.rutgers.edu/academic-integrity-at-rutgers/).

Violations of academic integrity will be treated in accordance with university policy, and sanctions for violations may range from no credit for the assignment, to a failing course grade to (for the most severe violations) dismissal from the university.

COURSE TOPICS AND OUTLINE (BY WEEK)

Introduction to Data Science (Week 1)

- 1/3/24: What is Data Science; Keeping the “science” in data science

Data Science Ethics (Week 2)

- 1/8/24: Data ethics and violations; Data science oath
- 1/10/24: Ethical and responsible computing; Open science and reproducibility

Essential Tools for Data Science (Week 3)

- 1/15/24: Martin Luther King Jr. Day (No Class)
- 1/17/24: The terminal and Unix; High performance computing

Essential Tools for Data Science (Week 4)

- 1/22/24: Git and GitHub
- 1/24/24: Introduction to R and Rstudio

Advanced data wrangling in R (week 5)

- 1/29/24: RMarkdown; Data Structures
- 1/31/24: The tidyverse; Tidydata wrangling

Advanced R Tools (week 6)

- 2/5/24: Creating R packages
- 2/7/24: R/Shiny

Data Visualization (Week 7)

- 2/12/24: General plotting principles; Ggplot2
- 2/14/24: D3, plotly, other advanced plotting tools

Final Project Presentations (Week 8)

- 2/19/24: Final student presentations
- 2/21/24: Final student presentations

ADDITIONAL (ASYNCHRONOUS) MODULES

Learning R:

- RStudio Education
- R Programming (Coursera/Johns Hopkins)
- Data Science R Basics (edx/Harvard University)
- R Training Course (LinkedIn)
- R Programming A - Z: R for Data Science (Udemy)
- Programming with R (Pluralsight)

Here are some resources to learn basic statistics (and in some cases R simultaneously):

- Data Analysis with R Specialization (Coursera/Duke University)
- Introduction to statistis (Coursera/Stanford)