Math 241 - Data Science

Useful Information

Professor: Kelly McConville

Email: mcconville@reed.edu

Kelly's Zoom Office Hours: Tuesdays 2:45 - 4:30pm, Wednesdays 2-4pm

Course Assistant: Simon Couch

Simon's Google Meet Office Hours: Mondays 4:15 - 6:15pm

Course Websites:

- rstudio.reed.edu for course work and materials.
  - Handouts, Data, and Labs can be found in our class shared folder: `/home/courses/math241s20`
- Slides Repo for reviewing the lecture slides.
- Gradescope for submitting work.
- Slack for class correspondence.
  - **Make sure to check daily.**
- Our Reed Data Science Blog for sharing our work.

Learning Objectives

In this course, you will hone your ability to tell a compelling and accurate story with data. The Learning Objectives are:

- Develop the ability to effectively write about data for a non-technical audience.
  - Be able to write both short (500 word) to medium (1000 word) length data stories.
- Gain data acumen for a variety of data types, including:
  - Spatial data
  - Text/strings
  - Factors
  - Dates
  - Relational databases
  - Data on the web
- Develop coding habits that align with best practices in the field.
- Understand the current ethical issues around data and be able to reason through these issues using the ASA Ethical Guidelines.
- Create data visualizations, sometimes with interactivity, of multivariate data.
- Learn to disseminate data by creating data R packages.
- Acquire a reproducible and collaborative workflow for data analyses.

How does Math 241 fit into the statistics curriculum?

Math 241 is an applied statistics course with a heavy emphasis on building data acumen, creating data visualizations, and telling data stories. In contrast, Math 141 is an introductory course that focuses deeply on statistical inference (i.e., hypothesis testing and confidence intervals) and Math 243 largely focuses on statistical modeling. Therefore, while inference and modeling are important tools when extracting knowledge from data, they are not the focus of Math 241. Of course, you are likely to use concepts and tools from Math 141 in your projects.

Learning Tools

RStudio Server

- We will use R/RStudio extensively. R is the statistical software package and RStudio is the user interface for R.
- The Reed College RStudio server is located at rstudio.reed.edu.
- You will need access to a computer to complete the coursework. I will occasionally ask you to bring a laptop to class. Departmental laptops are available for rental (see me for more information), along with the Lottery Mac program.
- If you prefer a local version or if the server is down, you can download R and RStudio for free to your personal computer.
- Work done in any program other than R will not receive credit.
- When we are using computers during class time, you should **not** use your computer for any activity that is not directly related to class.

Course Materials

We will use the following textbooks:

- Primary textbook: Modern Data Science with R by Baumer, Kaplan, and Horton
  - Denoted as MDSR in the Schedule.
- Secondary textbook: The Curious Journalist's Guide to Data by Stray
  - Free online and in pdf form.
  - Denoted as CJGD in the Schedule.
- Secondary textbook: R for Data Science by Grolemund and Wickham
  - Free online
  - Denoted as RDS in the Schedule.

Avenues for Help

It is very important to stay on-top of the material since we will move at a fairly brisk pace. If you feel yourself falling behind, please seek out help.

- We will use class time to actively engage with the material.

- Come visit me during my office hours or Simon during his office hours.
- Form study groups that will help you learn the material.

Course Components

Your grade will be based on your performance on the following key components of the course:

- **Homework**: Most weeks you will receive a lab assignment.
  - Each assignment is due **before** the next week's Thursday class.
  - Don't wait until the night before a lab is due to work on it! They are designed to encourage consistent application and practice and are not structured to be completed in a single evening.
  - I do not answer emails on the weekend. Please get questions answered during the week!
  - **No late homework is accepted**.
  - The lowest homework grade will be dropped.
- **Projects**: We will have two mini projects and one final project.
  - Since data science is a collaborative field, all projects will be done in groups.
  - The mini-projects will involve writing blog posts for the website.
    - You have the right to request that your blog post not be included on the course blog or that your name be removed before it is posted.
- **Participation**: You are expected to come to class or to watch the videos of the lectures. Participation includes asking and answering questions in-class or on Slack, participating in class discussions, not checking email during class, etc.
  - **Cellphone use during class is prohibited.**
  - To receive full participation points, all students must contribute to **2** content related discussions on Slack. Examples include:
    - Asking a lab or project related question that is content related ("When is X due?" does not count).
    - Answering or at least providing guidance on a lab or project related question
    - Showcasing a neat function you found with a `reprex`

Accommodations

If you are a student with a documented disability in need of accommodations, I encourage you to reach out to Reed's Disability Services Office, or its director, Theresa Lowrie, to make the necessary arrangements. If you already have accommodations, in place, please submit your accommodations to me through the DSS portal, and then come to discuss your accommodation needs with me in person during office hours or by appointment.

Classroom Climate

Reed College values diversity and inclusion. I am committed to a climate of mutual respect, free of discrimination based on race, ethnicity, gender identity, religion, sexual orientation, disability, and other identities, in and out of the classroom. I expect everyone in this class to strive to foster a learning environment that is equitable, inclusive, and welcoming. If you experience any barriers to learning, please come to me or a college administrator with your concerns.

Academic Honesty

I encourage you to collaborate on assignments but **every piece of work you do must be your own.** Copying and pasting other people's work or code is not acceptable. The Honor Principle must guide your

conduct in this class. The following section from the Guidebook to Reed College summarizes the expectations for this class:

> Reed College is a community of scholars. The fundamental ethical principle governing scholarship is that one should never claim or represent as one's own work that which is not one's own. Proper academic conduct requires that all work submitted for academic purposes – including, but not limited to examinations, laboratory reports, essays, term papers, homework exercises, translations, and creative work—be entirely the work of the person or persons who submit it, and that, in the case of work based upon experiment and observation, the experimental results and observations be reported faithfully. The principle thus requires that no one claim authorship to the work of another and that no one falsify or misrepresent empirical data. This principle should be clear to every scholar, although determining its application in particular circumstances may require careful thought and guidance.

Schedule

The following is a tentative schedule for the material of the course and the important project dates.

Week 1 (Jan 27th)

Topics:

- What is Data Science?
  - Reading: MDSR Ch 1.1
- Data Journalism
  - Reading: CJGD Executive Summary and Introduction
- Reproducible workflow
- Hand drawn data visualizations
- Data Visualization Principles
  - Reading: MDSR Ch 2.1 - 2.4

Week 2 (Feb 3rd)

Topics:

- Data Stories
  - CJGD Communication
- Coding Style
  - Reading: Tidyverse Style Guide
- Reproducible Examples
  - Reading: reprex Vignette and Tutorial
  - Package: `reprex`
- Data Viz: Grammar of Graphics
  - Reading: MDSR Ch 3.1 - 3.3
  - Package: `ggplot2`

Week 3 (Feb 10th)

Topics:

- GitHub

  - Reading: D.3 Projects and Version Control, Ch 1 of Happy Git and GitHub for the useR
- R Projects
- R Packages
  - Data packages
  - Packages: `devtools`, `usethis`
- Data Wrangling
  - Reading: MDSR Ch 4
  - Package: `dplyr`

Projects:

- Receive Mini-Project 1: Packaging your data!

Week 4 (Feb 17th)

Topics:

- Data Wrangling
  - Reading: MDSR Ch 5.1 - 5.2
  - Package: `tidyr`
- Data Ingesting
  - Reading: MDSR Ch 5.5
  - Packages: `readr`, `rvest`, `httr`

Week 5 (Feb 24th)

Topics:

- More Data Ingesting

Projects:

- **Feb 27th**: Submit and give 5 minute lightning talk about Mini-Project 1.

Week 6 (Mar 2nd)

Topics:

- Writing Functions
  - Reading: MDSR Appendix C.1 - C.3
- Refactoring Code
  - Reading: Code Smells
- Quantification
  - CJGD Quantification

Projects:

- Receive Mini-Project 2: Telling a Data Story!

Week 7 (Mar 9th)

Topics:

- Iteration
  - Reading: RDS Iteration
  - Package: `purrr`

- Dates and Times
  - Reading: RDS Dates and times
  - Package: `lubridate`

Week 8 (Mar 16th)

Topics:

- Factors
  - Reading: RDS Factors
  - Package: `forcats`
- Strings
  - Reading: RDS Strings
  - Package: `stringr`

Spring Break

Week 9 (Mar 30th)

Topics:

- Text Analysis
  - Reading: MDSR Ch 15.1 - 15.3
  - Package: `tidytext`

Projects:

- April 2nd: Mini Project 2 is due
- Receive Final Project

Week 10 (Apr 6th)

Topics:

- Ethics
  - Reading: MDSR Ch 6, ASA Ethical Guidelines
- Spatial Data and Mapping
  - Reading: MDSR Ch 14.1 - 14.6
  - Packages: `leaflet`, `sf`, `ggspatial`

Week 11 (Apr 13th)

Topics:

- Animation
  - Reading: gganimate
  - Package: `gganimate`
- Interactivity
  - Reading: MDSR Ch 11.1, 11.3
  - Packages: `plotly`

Projects:

- Project meetings with professor or course assistant.
- No class on April 16th.

Week 12 (Apr 20th)

Topics:

- Creating Interactive Web Apps
  - Additional resource: Mastering Shiny by Wickham
  - Packages: `shiny`

Week 13 (Apr 27th)

Topics:

- Database querying
  - Reading: MDSR Ch 12.1 - 12.6

Final Exam

Projects:

- Make sure that the complete and final version of your group's final project is in your GitHub repo by May 11th at noon.