Reproducible Research in R

{.table .table-condensed .table-striped .text-left}

Instructor	Tad Dallas
Location	Coker 202
Times	T & TH 4:25 - 5:40pm
Office Hours	T 3:00 - 4:20pm

Overview

This course is designed for undergraduate students and early career graduate researchers regardless of prior experience. We aim to be accessible to those new to programming, but those who have been using R for years will find new material, best practices, and tools to enhance reproducibility in scientific research. The course is project-focused and centered around modules aimed at teaching programming skills while also exploring scientific data and questions. A series of short tutorials will introduce relevant technology, but most concepts will be first introduced in reading outside of class, leaving class time to focus on the more complex examples encountered in the modules.

Approach

This course will use a flipped classroom model, with new material introduced in reading assignments prior to class while class time will focus on applying these skills to explore interesting data sets. If you do not do the reading, you will quickly find yourself struggling to keep up. Students are expected to come to class with the conceptual background in the topic of the lecture, as the lectures will focus on skill-building and the analysis of biological data. Students will be expected to work collaboratively in and out of class, and course content and grading will emphasize communication and reproducibility of an analysis as much as scientific or technical completeness. That being said, there are numerous ways to programmatically solve the same problem, and I do not expect to see the same code from multiple people. The Course Syllabus provides an overview of the modules and topics covered as well as links to weekly reading, assignments,

and any lecture material. This syllabus is preliminary and always subject to change.

Texts

There is no required text, but we will use some material from Grolemund and Wickham's *R For Data Science* and Wickham's *Advanced R*. Additional reading material will be linked from the syllabus. Please be sure to review the relevant reading prior to each class session.

Course design

This website, and the modular structure of the course, was inspired by Carl Boetigger's ESPM 157 course at Berkeley (https://espm-157.carlboettiger.info/). I not only used his website code, but borrowed some of the readings and topics for tutorials. Without his willingness to keep his course materials open access, and without the open source tools to build the website, this course would have to be created from scratch. The content would surely have suffered. The focus of this class is on reproducible science, but reproducibility and access to tools are inextricably linked. The most reproducible MatLab code is still only reproducible on machines that have access to MatLab. This means that reproducible science and aspects of open science (e.g., development and use of open source tools) are quite related.

Policies

Prerequisites

You should be generally numerically literate, and some prior programming experience will be helpful. Students new to programming might find Hands on Programming with R to be helpful. Students with significant experience in programming and statistical analysis should find themselves well prepared but should find plenty still to learn in each lesson.

Instructional Methods

As a flipped classroom, students are provided with either reading or video material that they are expected to view/read prior to class. Classes will involve brief refreshers on new concepts followed by working on exercises in class that cover that concept. While students are working on exercises the instructor will actively engage with students to help them understand material they find confusing, explain misunderstandings and help identify mistakes that are preventing students from completing the exercises, and discuss novel applications and alternative approaches to the data analysis challenges students are attempting to solve. For more challenging topics class may start with 20-30 minute demonstrations on the concepts followed by time to work on exercises.

Assignment policy

Data science is about analyzing real-world data sets, and so a series of projects involving real data are a required part of the course.

All assignments are due by immediately **before** the start of class on the day indicated.

Assignments should be submitted as instructed.

Grading

Grades will be assigned using the following weighted components:

Undergraduate students

{.table .table-striped}

component	weight
HW Exercises	50%
Participation	10%
Group project	40%

Graduate students

{.table .table-striped}

component	weight
HW Exercises	50%
Participation	10%
Project	40%

Homework exercises are set up to be the same for both undergraduate and graduate students. This may change depending on how much folks struggle. The point is not to overwhelm, but to get everyone as much experience and excitement as possible without overburdening.

Details of grading criteria will not usually be announced in advance. It is expected that students in this course will have a wide range of prior experience and ability, and grading will aim to reflect learning and effort in the course. It is certainly possible for all students to receive high grades in this course if all of you show mastery of the material and completely attempt all assignments.

Project details

The project will be a demonstration of your learning and allow you to use the new tools you have acquired in the course. This will be done in small groups (3-5) for undergraduates and will be a solo project for graduate researchers.

The final project will be an R markdown document that reads in and analyzes a data source. I can provide data, if necessary, but would prefer if data came from the student (either from their research or on something they are passionate about).

You will actively develop your work on Github, structuring your directory as we discussed in class. The final product will be an R markdown file that is entirely reproducible. This means I will clone your directory, and run your files on my local machine. **I must be able to reproduce your analyses**.

The final project will consist of 3 parts. First, you will develop a project proposal. Second, you will do the proposed project, and the final product will be a versioned and end-to-end reproducible analytical workflow that could potentially lead to a

peer-reviewed publication. Lastly, you will briefly present your work to the class (15 minute scientific presentation).

This means that I will expect not only a well-written and clear manuscript, but also documentation to easily reproduce all analyses, figures, and compile manuscript text.

Final project proposal Please prepare a short proposal on your final project idea by February 1. The proposal should include:

- Title & description of the project
- Team members names
- A description of the data required, and how it will be obtained (e.g. URL/DOI to data source)
- 3 questions / analysis tasks you will perform on the data; in the spirit of the assignments we have been doing.

Please create your proposal in a markdown file called `proposal.md` in the root directory of the final project repo.

Project development You will develop the project throughout the semester, using tools and approaches that you learn during the course. Stay focused on the goal of the course... reproducibility. If I can't clone your GitHub repo and reproduce your analyses with a fresh install of R, that's not great.

Final product You will present your final project as a 15 minute presentation at the end of the semester. It should provide context and information about the 'what, why, and how' of your research question, and go through your findings, preferably demonstrating your ability to visualize data.

Make-up policy

I will use the latest commit in the GitHub repository for a given assignment as the submission. Commits made after this point will not be considered. In cases of emergency, or with prior approval, I am happy to consider late assignments, but the overall grade will be docked by 20% and will be accepted within 3 days of the original due date. Try to plan ahead to get the assignments done. Some of the assignments may seem fairly straightforward, but may actually be more challenging than they seem.

Attendance Policy

The lab-based, hands on course design really depends on students being in class, for every session. I expect students to make every effort attend every class. I cannot accomodate scheduling conflicts that would cause a student to regularly miss part of class. However, I recognize that now and again an occasional absence will be unavoidable. Please notify the instructor beforehand if possible. I will not require any note or explanation and trust you to make the right decisions for your own education, but advanced notice may help ameliorate the disruption.

Keep up with the reading assignments while you are away and we will all work with you to get you back up to speed on what you miss.

Academic Honesty

Cooperation has a limit. You should not copy your code or answers directly with other students. Feel free to discuss the problems with others, but write your own solutions. Penalties for cheating are severe – they range from a zero grade for the assignment or exam up to dismissal from the University, for a second offense.

Rather than copying someone else's work, ask for help. You are not alone in this course! If you invest the time to learn the material and complete the projects, you won't need to copy any answers.

Disability services

My goal is to help you learn. Students who have any difficulty (either permanent or temporary) that might affect their ability to perform in class can reach out to the University of South Carolina Disability Services staff.

More information on registering a disability is available at U of SC Disability Services.

Support

You are not alone in this course; your student colleagues and the course instructor are here to support you as you learn the material. It's expected that some aspects of the course will take time to master, and the best way to master challenging material is to ask questions. Time will be set aside in each class to ask questions and discuss the material. You are encouraged to bring up related questions that arise in your research as well.

Code of Conduct

Our course is committed to providing a respectful and welcoming environment to all participants. Please review the Contributor Covenant Code of Conduct guidelines for respectful and harassment-free conduct.

To report an incident or request more information, contact the Office of Equal Opportunity.

Course Technology

Students are required to provide their own machines and to install free and open source software on those machines. Support will be provided by the instructor in the installation of required software. If you don't have access to a computer please contact the instructor and they will do their best to provide you with one.

Materials & Resources

All reading material required for this course will be made available through this website and links to related resources.