Syllabus

NTRES 6100:
Collaborative and Reproducible Data Science in R
Cornell University, Spring 2022

Course Info *Note that the first two weeks of class (both lectures and lab sessions) will be taught remotely. Zoom links will be available from the class Canvas site.*

Lectures: Tuesdays and Thursdays 9:40am - 10:55am (January 25 - March 31, 2022), Ives Hall 111

Optional lab sessions: Thursdays (Academic Surge A 109) **or** Fridays (Fernow Hall G01) 12:25pm - 2:20pm

Instructor: Assistant Professor Nina Overgaard Therkildsen (nt246@cornell.edu)

TA: PhD Student Nicolas Lou (rl683@cornell.edu)

Office hours: Nina: by appointment; Nicolas: Tuesday 2:30-3:30 or by appointment (Zoom link)

Grading: S/U (2 credits / 3 credits with lab)

Course website: https://nt246.github.io/NTRES-6100-data-science/index.html

---

Course description As datasets grow larger and more complex across all areas of science, computational skills are increasingly in high demand. This course introduces a series of practical tools that enable researchers to spend less time wrestling with software or repeating error-prone manual data processing and more time getting research done in efficient and transparent ways that facilitate collaboration and reproducibility. We will work in R/RStudio, primarily with the tidyverse packages and with Git and GitHub integration. Topics covered include 1) tidy data formatting, 2) rearrangement, filtering, exploration, and visualization of complex datasets, 3) basic programming for building and automating custom tools, 4) tracking the history of file changes (version control) with Git and GitHub, 5) strategies for effective collaboration on data processing pipelines, and 6) using R Markdown to combine text, equations, code, tables, and figures into reports, websites, and presentations. The course emphasizes practical skill development and will be structured around hands-on (the keyboard) learning.

Learning outcomes By the end of this course, students will be able to:

- Describe strategies for ensuring that their data analysis is reproducible
- Demonstrate best practices for coding and project-oriented workflows in RStudio
- Import and clean messy data files using a variety of packages and functions in R
- Subset, reorganize, and merge diverse datasets in R
- Effectively explore and visualize patterns in complex datasets with ggplot in R
- Write simple functions/programs and data analysis pipelines in R
- Automate repeated analysis tasks in R
- Track the history of file changes (version control) and collaborate effectively on scripts with others with Git and GitHub

- Use R Markdown to combine text, equations, code, tables, and figures into reports, websites, and presentations

---

Prerequisites A basic working knowledge of R will be helpful, but no prior experience with the tidyverse packages or with Git, GitHub, or R Markdown is assumed. If you have never worked in R before, we recommend working through one or more of the following tutorials before the course:

- Jenny Bryan's STAT545 Chapter 2 R basics and workflows
- R Swirl interactive lessons
- Data Carpentry's Introduction to R for Ecologists

Course format The two weekly lectures will introduce new concepts and provide opportunities to practice through hands-on exercises. To participate effectively, you must have completed the assigned readings prior to class. Each Thursday, we will assign a problem set that applies the concepts covered in class in a new context to reinforce your learning. The problem sets are due the following **Thursday at 10pm**. We offer two optional lab sessions on Thursdays and Fridays for more opportunities to practice in groups and with TA support; the Thursday and Friday sessions are identical and you can attend either one of them.

Evaluation It takes practice to acquire and internalize data science skills, and what you get out of this course will be proportional to the effort you put in. Practice as much as you can. To pass, students are expected to attend all lectures (and lab sessions if enrolled), participate actively during class, submit at least **7** of the 9 problem sets with **demonstrated effort to complete all questions**, and give a brief (~2 minute) presentation at the end of the course about how you might adopt some of the course material in your own work. If you are unable to make a lecture or can not meet a problem set deadline, please let the instructor and TA know on Slack beforehand. If you are registered in one of the lab sessions (one extra credit), you are also expected to participate in lab activities in at least 7 of the 9 lab sessions.

Course materials All assigned readings are freely available online and will be linked to from the course website. We will draw from a variety of sources, primarily Grolemund and Wickham's R For Data Science and the STAT545 course developed by Jenny Bryan.

All students will need to bring a laptop to each session. Students who do not have their own laptop can arrange to borrow one from the Mann Library.

Please follow the instructions here to install the software we will need **prior to the first class**.

Code of conduct We are dedicated to providing a welcoming and supportive environment for everyone, regardless of background, identity and prior experience level. Everyone in this course will be coming from a different place with different experiences and expectations. We will not tolerate any form of language or behavior used to exclude, intimidate, or cause discomfort. This applies to all course participants (instructor, students, guests). In order to foster a positive and professional learning environment, we encourage the following kinds of behaviors:

- Use welcoming and inclusive language
- Be respectful of different viewpoints and experiences
- Gracefully accept constructive criticism
- Show courtesy and respect towards others
- Help each other - you may well learn something or reinforce your own skills in the process

---

Student accommodations In compliance with the Cornell University policy and equal access laws, we are available to discuss appropriate academic accommodations that may be required for student with disabilities. Requests for academic accommodations are to be made during the first two weeks of the course, except for unusual circumstances, so arrangements can be made. Students are encouraged to register with Student Disability Services to verify their eligibility for appropriate accommodations.

Tentative schedule *Subject to adjustment*

Lecture#	Day	Date	Topic	Assignment due dates
1	Tue	1/25	Intro to the course and R/RStudio	
2	Thu	1/27	Markdown and GitHub	
3	Tue	2/1	The Git workflow (version control)	
4	Thu	2/3	Collaborating with GitHub Part 1	Assignment 1
5	Tue	2/8	Collaborating with GitHub Part 2	
6	Thu	2/10	Plotting with ggplot part 1	Assignment 2
7	Tue	2/15	Data wrangling part 1 (dplyr::filter, mutate, select, arrange)	
8	Thu	2/17	Data wrangling part 2 (dplyr::summarize, group_by)	Assignment 3
9	Tue	2/22	Plotting with ggplot part 2 + good coding practices	
10	Thu	2/24	Tidy data	Assignment 4
	Tue	3/1	**NO CLASS** - February Break	
11	Thu	3/3	Data import, export, and conversion between data types	Assignment 5
12	Tue	3/8	Good coding practices, debugging strategies, and getting help	
13	Thu	3/10	Relational data	Assignment 6
14	Tue	3/15	Iteration (for loops) and conditional execution part 1	
15	Thu	3/17	Iteration (for loops) and conditional execution part 2	Assignment 7
16	Tue	3/22	Functions	
17	Thu	3/24	Factors in R	Assignment 8
18	Tue	3/29	Wrapping up and resources for learning more	
19	Thu	3/31	Student presentations, wrapping up and looking ahead	Assignment 9

Lab#	Date (Thu)	Date (Fri)	Topic
1	1/27	1/28	RMarkdown
2	2/3	2/4	RMarkdown and GitHub
3	2/10	2/11	Displaying data visualization on a website
4	2/17	2/18	Data exploration with the gapminder dataset
5	2/24	2/25	Data exploration with the Titanic dataset
6	3/3	3/4	Data cleaning and tidy data
7	3/10	3/11	Relational data and tidy data
8	3/17	3/18	Iteration and conditional execution
9	3/24	3/25	Functions and iterations
10	3/31	4/1	OPTIONAL: Bring your own project