# Advanced Topics: Data to Manuscript in R (**01:185:**412:02)

Dr. Michelle Hurst

Fall 2023

Syllabus last updated on 2023-07-30.

## Logistics

In-Person Lectures

Tuesdays and Fridays 10:20am - 11:40am (Period 2)

Livingston Campus, Tillett Rm 207

Course website/Canvas site is TBD

## Contacting the Professor

michelle.hurst@rutgers.edu

Email Guidelines: You can email me at any time. I typically respond to emails within 24-48 hours, not including the weekend.

Student Support Hours: TBD (will be set prior to the beginning of the semester)

## Description

This course tackles the basic skills needed to build an integrated research report with the R programming language. We will cover every step from data to manuscript including: Using R's libraries to clean up and re-format messy datasets, preparing data sets for analysis, running statistical tools, generating clear and attractive figures and tables, and knitting those bits of code together with your manuscript writing. The result will be a reproducible, open-science friendly report that you can easily update after finishing data collection or receiving comments from readers. Never copy-paste your way through a table again! The R universe is large, so this course will focus specifically on: The core R libraries, the tidyverse library, and R Markdown. Students will also learn about the use of GitHub for version control.

## Course Materials and Technology

Classes will be interactive lectures. My expectation is that you will have a laptop computer to use for in class activities throughout the lecture. If you are unlikely to have a computer to use, let me know and we'll find a solution.

All other software and readings will be freely available to you - you *do not* need to purchase a textbook.

### Software

All software is freely available, with some additional free upgrades because of your educational affiliation.

R: https://www.r-project.org/

RStudio: https://posit.co/download/rstudio-desktop/

Github (https://github.com/), with a free educational account (which you can request here: https://education.github.com/)

If you're not familiar with git, github, or terminal, then you'll also need an IDE for working with github, such as Github Desktop (https://desktop.github.com/)

**Readings**

All readings are freely available online and will be pulled from the following sources:

Tidyverse Style Guide

Douglas et al.'s An Introduction To R

Rafael A. Irizarry's Introduction to Data Science

Wickham & Grolemund's R for Data Science aka R4DS, 1st edition

Garrett Grolemund's Hands-On Programming with R

Nicholas Tierney's RMarkdown for Scientists

Frederik Aust & Marius Barth's papaja: Reproducible APA manuscripts with R Markdown

Posit/RStudio Cheatsheets

Barbara Sarnecka's The Writing Workshop: Write More, Write Better, Be Happier in Academia

Yihui Xie, J. J. Allaire, Garrett Grolemund's R Markdown: The Definitive Guide

## Learning Goals

The learning goals of this course are meant to align with the learning goals of RuCCS undergraduate major. Specifically, after fully participating in this course you will:

1. Be familiar with the R programming language and statistical environment
2. Be familiar with version control using git
3. Be comfortable searching for answers and approaching new problems in R
4. Be able to work with data in various forms and think analytically about the relation between *data format* and *statistical analysis*.
5. Know the structure of a scientific report
6. Have a working and reproducible draft of a scientific report

What will you *not* learn in this course?

1. This is not a statistics or a mathematics course. Although you will be doing and writing about statistics, my assumption is that you have taken previous courses in research methods and statistics. We will be focused on how to implement these mathematics and statistical concepts in R and how to communicate about them using visualizations and in writing. We will *not* be learning about the mathematics or statistics themselves.
2. This is not, strictly speaking, a general programming course. R is a specialized language for statistical computing. We will be using R for working with data, visualizing data, and writing about data (by integrating R with Markdown, a markup language for text). On the other hand, there will be some programming basics and learning R might make it easier to learn another programming language later.

## Learning Strategies

The best way to learn any programming language (or even natural language!) is to practice. The assessments throughout the course are designed to help you do that. When learning about data, it can also be helpful to work with data you care about. Your final scientific report is meant to be based on your thesis project (if you're doing one). In the first few weeks of the course, we will talk about what data would be appropriate to use and how to go about finding data.

## Grading

In Class Mini Assignments, Total = 30

Take Home Assignments, Total = 30

Scientific Report, Total = 40

**Bonus of up to 3

## Assessments

### In Class Mini Assignments (30%)

Throughout class, there will be **required** in class activities.

You will do these activities in one continuous **R Notebook File**, which you will **push** back to github at the end of each class (more on all of this the first week of class!)

These activities will be "graded" based on completion (i.e., submitting it to github). If you miss a class, forget your laptop, or for any other reason aren't able to actually do and/or submit the activities *during class*, you must still submit the in class activities before the beginning of the next class with a brief explanation for why it's late (e.g., "forgot my laptop", "had to run to my next class, and submitted it when I got home", "I missed class, but caught up from the slides"). These late activities will still be given full credit.

For grading purposes, every class will have a required activity and you only need to submit 24 of them (out of 28) to get full credit, any additional missing assignment will be a deduction of 1.25 points each. But, like any programming language, how much you practice will directly relate to how quickly and effectively you learn. These in class activities are meant to give you that regular practice and I highly recommend you plan on doing all of them. If you join the class late into the semester, contact the professor about catching up.

### Assignments (30%)

The course is organized into major topics and each topic has an assignment. These assignments are more lengthy and more challenging than the in class activities. Each will have about 3 to 5 problems and might include a mix of programming problems and open-ended word problems. I do not recommend you wait until the week they are do to begin working on them. Instead, I recommend working on them throughout the section, as we cover the relevant sub topics. Each assignment is worth 10% and they are to be submitted via TBD.

Assignments will be graded based on the following criteria:

- coding style and understandability: can I easily follow what you are doing and why?
- accuracy and efficiency of the code: is the code doing what you want it to do? Is the code fairly efficient and well thought out, or is there a much better or faster way to do that same thing? It's OK if you don't do something in the *most* efficient way, but redundancies or not using a more efficient function we've explicitly discussed in class would be graded as less efficient.
- actually answering the question: is your solution actually solving or addressing the problem posed?

### Scientific Report (40%)

One major goal of the course is to provide the tools for writing a reproducible scientific report. Although this is unlikely to be your final thesis (if you're doing one), it will provide a structured first draft that you can continue adding to as you collect more data or do more analyses during the rest of the academic year. One of the benefits of our approach to data wrangling and reporting is that you can easily add more data and re-run the code without having to do it manually all over again! There will be two preliminary graded assignments, followed by the final project.

**Proposal (5%)**: A paragraph or two describing the data you plan on using for your final report, including: where you will get the data from, what kinds of variables it has, and what you want to do with it (e.g., research questions, analyses, visualizations). This will be graded based on completion. If your initial submission is not deemed appropriate (i.e., the data you plan on using will not be appropriate for the class, the plan is not complete), you must set up a 1-on-1 meeting with the professor (within two weeks) to discuss your plan in more detail and come up with alternatives as needed. Within one week of this meeting, you must resubmit the written proposal. Assuming the re-submitted written proposal is deemed acceptable and aligns with the conversation you had with the professor, you will get a complete grade.

**Outline (10%)**: An outline of your reproducible scientific report. This should include an outline of what will be included in each of the major sections (Intro, Method, Results, Discussion) of the report. Using bullets and lists is perfectly acceptable, as long as it's clear what you are going to do and how you're going to organize it. Half of this grade will be completion (i.e., you get 5% just for submitting it), and the remaining will be graded based on clarity of the introduction structure and completeness of the method and results.

**Final Scientific Report (25%)**: This is a reproducible final scientific report that includes: Title Page, Abstract, Introduction, Method, Results, and Discussion. It must also include specific elements covered in class. A more detailed specification of what is required and how it will be graded will be provided within the first 8 weeks of class.

### Bonus Points

You will have the opportunity to gain bonus points (1% each added to your overall grade, for a maximum of 3%) by submitting candidate problems for the "Problem Solving Live!" classes.

During the three "Problem Solving Live!" lectures (see schedule), we will (as a group) work through one or more of the submitted problems.

Submitted problems could be related to your research project (e.g., a data cleaning task you thought you'd have to do manually, but want to automate), a hobby or interest (e.g., you saw data about something and want to figure out how to visualize it), or something else entirely (e.g., you're organizing your family's holiday gift giving and want to randomly pair people together). Get creative about the kinds of problems you suggest. If you work in a research lab, ask the graduate students or postdocs if there are tasks they do manually or in Excel that could be done in R instead!

To submit a bonus problem, include everything needed for us to work on it (e.g., if we need data, include the data - but make sure you have permission to share the data and it's not identifiable!), a description of the problem you're trying to solve (including any peculiarities about the data or problem that we need to know), and a brief (e.g., one sentence) description of how you think we can solve it or why you haven't been able to solve it yourself. This can be brief, and does not have to demonstrate a complete understanding! I just want to know that you've thought about it.

Your submissions can be related to your final project, but should not be your entire final project. For example, I will not accept submissions like, "wrangle this data so that we have output in the form of XXX". Instead, it could be a much smaller piece of the larger, overall wrangling goal.

Where to submit these problems is TBD.

You'll get 1% (up to a max of 3%) for each submitted problem that meets the above requirements, regardless of whether it's chosen to be used during a Problem Solving Live! class.

## Course Outline

This outline is subject to change, in particular the readings might be reorganized and replaced between now and the first day of class! A full schedule is provided at the end of this document.

### Part 1: Introduction

The first part of the course will cover an introduction to important concepts (e.g., reproducibility, version control) and software (e.g., R, git) that motivate the structure of the course and will be used throughout.

There will not be a formal assignment for this section. Instead, we'll be focused on getting set up with software and basic concepts.

| date | topic | readings |
|---|---|---|
| Tue Sep 05 | intro to course | |
| Fri Sep 08 | intro to git and notebooks | Xie et al, Chapter 3.2 |

### Part 2: Basics of Base R

The second part of the course will cover basics of base R. This is the part of the course that will be most similar to learning programming, including different kinds of objects, structures, and the logic/flow through a program.

The assignment for this section will be maybe available on Monday September 11th and due on Friday September 29th.

| date | topic | readings |
|---|---|---|
| Tue Sep 12 | intro to R | Tidyverse style guide Chapters 1-5 |
| Fri Sep 15 | operators, functions, packages | Douglas et al., Intro to R: Chapter 2 |
| Tue Sep 19 | classes/types, data strutures | Douglas et al., Intro to R:Chapter 3.1 and 3.2 |
| Fri Sep 22 | control structures and logic | Douglas et al., Intro to R: Chapter 7.3, 7.4, 7.5 |

### Part 3: Data Wrangling

The third part of the course will move into Tidyverse, a set of packages that focus on data wrangling and summarizing with tidy data.

The assignment for this section will be made available on Monday September 25th and due on Friday October 20th.

During this section of the course we will also talk about what data you can use for your Scientific Report final project. Your proposal describing the data you will be using is due on Tuesday October 10th.

Finally, this section will include a "Problem Solving Live!" lecture period, where we'll jointly solve student-submitted problems. Problems must be submitted the week before this session (i.e., by Friday Oct 6).

| date | topic | readings |
|---|---|---|
| Tue Sep 26 | introduction to tidyverse | R4DS, Chapter 18 |
| Fri Sep 29 | importing and exporting data | R4DS, Chapter 10, 11 |
| Tue Oct 03 | tidyr and dplyr, part 1 | R4DS, Chapter 12 |
| Fri Oct 06 | tidyr and dplyr, part 2 | R4DS, Chapter 5 |
| Tue Oct 10 | tidyr and dplyr, part 3 | |
| Fri Oct 13 | Problem Solving Live! | Review cheatsheets for tidyr, dplyr |
| Tue Oct 17 | stringr, forcats, lubridate | R4DS, Chapters 14, 15, 16 |
| Fri Oct 20 | catch-up/flexible/work period | |

### Part 4: Data Visualization

The fourth part of the course will focus on data visualization. We will cover principles of data visualization and how to plot various kinds of figures using the ggplot2 package.

The assignment for this section will be made available on Mon Oct 23rd and due on Fri Nov 10th.

Finally, this section will include a "Problem Solving Live!" lecture period, where we'll jointly solve student-submitted problems. This could be data wranging or about plots. For example, have you seen a cool data visualization you'd like replicate? Problems must be submitted the week before this session (i.e., by Friday Oct 27).

| date | topic | readings |
|---|---|---|
| Tue Oct 24 | intro to data visualization with ggplot2 | R4DS, Chapter 3 |
| Fri Oct 27 | building layers | |
| Tue Oct 31 | making it pretty | |
| Fri Nov 03 | Problem Solving Live! | Review ggplot2 cheatsheet |

### Part 5: Data Communication

The fifth part of the course will focus on writing scientific reports that are reproducible. We will cover each aspect of a scientific report and principles of how to write well. The bulk of this section, though, will focus on using RMarkdown and the papaja package to write reproducible documents that integrate R code and text.

The "assignment" in this section is the final paper. First, you'll submit an outline of your scientific report (due Fri Dec 1). The due date of your final project is TBD, but will likely be due at the time of our scheduled final exam (instead of a "final exam").

| date | topic | readings |
|---|---|---|
| Tue Nov 07 | writing a scientific report | The Writing Workshop, Chapter 5 |
| Fri Nov 10 | intro to markdown and papaja | RMarkdown for Scientists, Chapter 6, 15 |
| Tue Nov 14 | text and r code | TBD |
| Fri Nov 17 | figures and tables | TBD |
| Tue Nov 21 | NO CLASS | |
| **WED Nov 22 | cross-referencing figures and tables | TBD |
| Tue Nov 28 | statistical objects | TBD |
| Fri Dec 01 | references | TBD |
| Tue Dec 05 | references | TBD |

### Buffer and Bonus

The last week of the semester is being kept open to either (1) provide some buffer time for us to shift things around and (2) to cover other topics that come up during the course but that we don't have already built in. For example, we could do a class on Data Simulation. If you have suggestions for other topics, let me know!

## Policies

### Attendance and Participation

I am expecting you to attend and participate in class. Participation will count toward your grade via completion of the in class assignments and I have already built in ways to get full credit for missed classes (see details in the In Class Mini Assignments section above). I do not need to know why you are absent, but letting me know you will be absent in advance is helpful and appreciated.

### Disability Accommodation

Rutgers University is committed to the creation of an inclusive and safe learning environment for all students, and welcomes students with disabilities into all the University's educational programs. The Office of Disability Services (ODS) is responsible for the determination of appropriate accommodations for students who encounter barriers due to disability. Once a student has completed the ODS process (registration, initial appointment, and submitted documentation) and reasonable accommodations are determined to be necessary and appropriate, a Letter of Accommodation (LOA) can be requested and will be sent to the student and instructor. This should be done as early in the semester as possible as accommodations are not retroactive, and a discussion should occur about how the accommodations will be implemented. More information can be found at www.ods.rutgers.edu. You can contact ODS at (848)445-6800 or via email at dsoffice@echo.rutgers.edu.

### Civility and Community

This course has a lot of ties to the "Open Science Movement" in psychology, which is not known for always being a welcoming space (e.g., see this discussion of "open science" being closed to many). Here, we will work to counteract that. We will be thoughtful, courteous, and supportive to others. We will strive for "open" science to mean accessible and transparent in more ways than just throwing code on a repository.

### Academic Integrity

Rutgers University takes academic dishonesty very seriously. By enrolling in this course, you assume responsibility for familiarizing yourself with the Academic Integrity Policy and the possible penalties (including suspension and expulsion) for violating the policy. As per the policy, all suspected violations will be reported to the Office of Student Conduct. Academic dishonesty includes (but is not limited to):

- Cheating
- Plagiarism
- Aiding others in committing a violation or allowing others to use your work
- Failure to cite sources correctly
- Fabrication
- Using another person's ideas or words without attribution, including re-using a previous assignment Unauthorized collaboration
- Sabotaging another student's work

If you are ever in doubt, consult your instructor.

Note that for the purposes of this class, you *can and should* use search engines and trusted websites (e.g., stackoverflow) to help work through coding aspects of your assignments. I also recommend however, understand how the code works before implementing it. If you are implementing a custom function from another source, include a link to that source in your script.

### Student Support and Mental Wellness

- Student Success Essentials: https://success.rutgers.edu
- Student Support Services: https://www.rutgers.edu/academics/student-support

- The Learning Centers: https://rlc.rutgers.edu/

- The Writing Centers (including Tutoring and Writing Coaching): https://writingctr.rutgers.edu

- Rutgers Libraries: https://www.libraries.rutgers.edu/
- Office of Veteran and Military Programs and Services: https://veterans.rutgers.edu

- Student Health Services: http://health.rutgers.edu/
- Counseling, Alcohol and Other Drug Assistance Program & Psychiatric Services (CAPS): http://health.rutgers.edu/medical-counseling-services/counseling/
- Office for Violence Prevention and Victim Assistance: www.vpva.rutgers.edu/

## Full Schedule

Table 1: Course Outline

| date | topic | readings | assignment |
|---|---|---|---|
| **intro and set up** | | | |
| Tue Sep 05 | intro to course | | |
| Fri Sep 08 | intro to git and notebooks | Xie et al, Chapter 3.2 | |
| **foundations of (base) R** | | | |
| Tue Sep 12 | intro to R | Tidyverse style guide Chapters 1-5 | |
| Fri Sep 15 | operators, functions, packages | Douglas et al., Intro to R: Chapter 2 | |
| Tue Sep 19 | classes/types, data strutures | Douglas et al., Intro to R:Chapter 3.1 and 3.2 | |
| Fri Sep 22 | control structures and logic | Douglas et al., Intro to R: Chapter 7.3, 7.4, 7.5 | |
| **tidyverse** | | | |
| Tue Sep 26 | introduction to tidyverse | R4DS, Chapter 18 | |
| Fri Sep 29 | importing and exporting data | R4DS, Chapter 10, 11 | Assignment 1 (Base R) |
| Tue Oct 03 | tidyr and dplyr, part 1 | R4DS, Chapter 12 | |
| Fri Oct 06 | tidyr and dplyr, part 2 | R4DS, Chapter 5 | |
| Tue Oct 10 | tidyr and dplyr, part 3 | | Scientific Report: Proposal |
| Fri Oct 13 | Problem Solving Live! | Review cheatsheets for tidyr, dplyr | |
| Tue Oct 17 | stringr, forcats, lubridate | R4DS, Chapters 14, 15, 16 | |
| Fri Oct 20 | catch-up/flexible/work period | | Assignment 2 (core tidyverse) |
| **ggplot2** | | | |
| Tue Oct 24 | intro to data visualization with ggplot2 | R4DS, Chapter 3 | |
| Fri Oct 27 | building layers | | |
| Tue Oct 31 | making it pretty | | |
| Fri Nov 03 | Problem Solving Live! | Review ggplot2 cheatsheet | |
| **markdown** | | | |
| Tue Nov 07 | writing a scientific report | The Writing Workshop, Chapter 5 | |
| Fri Nov 10 | intro to markdown and papaja | RMarkdown for Scientists, Chapter 6, 15 | Assignment 3 (ggplot2) |
| Tue Nov 14 | text and r code | TBD | |
| Fri Nov 17 | figures and tables | TBD | |
| Tue Nov 21 | NO CLASS | | |
| **WED Nov 22 | cross-referencing figures and tables | TBD | |
| Tue Nov 28 | statistical objects | TBD | |
| Fri Dec 01 | references | TBD | Scientific Report: Outline |
| Tue Dec 05 | references | TBD | |
| **buffer and bonus** | | | |
| Fri Dec 08 | bonus! | | |
| Tue Dec 12 | bonus! | | |