STA 3100: Programming with Data in R

Fall 2024
Course Meetings: M, W, F, Period 7 (1:55 PM-2:45 PM): Griffin-Floyd 100
Course Website

Instructors

- **Lead Instructor:** Josh Miles
  - **Email (preferred):** miles.j@ufl.edu or Canvas message
  - **Phone:** (352) 392-1941 (Department of Statistics Office)
  - **Office:** Griffin-Floyd Hall 100A (right behind classroom 100)
  - **Office hours**: Monday, 3 PM-4 PM — Thursday, 9:30 AM-11:30 AM, or by appointment
- **Teaching Assistant:** Xiaoda Qu
  - **Email:** quxiaoda@ufl.edu or Canvas message
  - **Office:** Griffin-Floyd Hall 234
  - **Office hours**: Monday and Wednesday, 11:00 AM-12:30 PM, or by appointment

Course Objectives

This course is intended to introduce students to computer programming and data science in the R language. In particular, the course will cover the following concepts in the R language:

- Objects, data types, and data structures
- Packages and functions
- Manipulating and visualizing data
- Producing analyses with integrated programming output
- Basic data analysis via statistical modeling

Prerequisites

Any one of the following is required to ensure basic statistical and mathematical background:

- STA 2023 with a minimum grade of B
- STA 3024 with a minimum grade of B-
- STA 3032 with a minimum grade of B-
- an AP statistics score of 4 or 5.

In addition, the course assumes no prior experience with R (nor any other programming language), and it assumes no knowledge of computer science concepts more broadly.

A note about learning to program

In my experience, learning to program is considerably different than the way we learn most academic subjects. Most of the time, we learn in a "bottom up" manner: we begin with the most basic things and gradually build up to more complex and specific subjects. For example, in algebra, you first learn the most basic things, like adding, subtracting, multiplying, etc., and then you learn to do more complex operations like logarithms and trigonometric functions.

Programming, on the other hand, is impractical to learn in this manner. This is because even the most basic programming tasks depend on numerous complexities that are not the focus, and often not immediately important, for most programmers. Consequently, we must take several principles for granted in order to learn to write any code. This is more of a "top-down" approach, and it adds an extra challenge to learning. As such, it's very normal to feel overwhelmed and uncapable in the beginning. I encourage you to remain patient and trust the process. Everything will make more sense with time. As we progress, we'll be able to understand more of the complexities that we have to take for granted at the beginning.

Materials

All student materials used in this course are free. The course will not use a textbook, instead following notes prepared by the instructor. The course will also make frequent use of free software and its documentation. Excluding projects, **assignments in this course can be completed without needing anything beyond the aforementioned materials**.

That said, you may find it useful (though not required) to consult various online tutorials, manuals, forums, etc. The volume of these kinds of materials for all sorts of projects in R is immense, and I will occasionally refer to some of them for additional learning or practice. Some of the most popular, streamlined, and thorough sources for learning R programming are listed below, but many other very good sources are available.

- RDocumentation
- R Markdown: The Definitive Guide
- R for Data Science, 2nd Edition

Computational Requirements

Students will need to have frequent and reliable access to a computer capable of running R code. All computers supplied by UFIT Academic Technology classrooms and laboratories, including those at the university libraries, have R and RStudio installed and can run any program of interest to this course without any trouble. However, because of data security policies, it's somewhat tedious (though possible) to efficiently manage an R installation on university computers. For these reasons, students with personal computers are encouraged to use them for programming tasks.

Students who wish to use personal computers will likely have no trouble with the computational requirements for R and RStudio. That said, please ensure that you have a sufficiently new operating system:

- For Windows users, Windows 10 or later
- For macOS users, macOS 10.13 (High Sierra) or later

Other hardware requirements (e.g. RAM, processor speed) will be satisfied if you meet the operating system requirements.

On Windows, R requires about 179 MB of space, and RStudio requires about 861 MB. On other operating systems, I suspect that each software will be similar in size to the corresponding sizes on Windows. You will

also need a handful of storage space to hold downloaded packages, datasets, and output that you produce. I suggest having at least 5 GB of storage available once you have downloaded R and RStudio.

Please contact the instructor if you have any concerns about the technology requirements of the course.

Attendance

Attendance at lectures will not be recorded nor graded. Students can miss lectures for any reason (except for your final project presentation, for which you must be present in class to present). Course notes will be posted on Canvas, along with any other supplementary material used in the lectures. The lectures, which will not be recorded, will follow the course notes closely but will often include extra detail and examples that are not in the course notes. The lectures will therefore be more helpful than the notes alone, and as such, attendance at lectures is encouraged.

Lectures will usually involve both a slide presentation and demonstrations of code writing and execution. Students who have laptops are encouraged, though not required, to bring them to follow along with the code writing, so that they can take their own notes in their code files and experiment with the code beyond what is shown in lecture. The code files will be posted after lecture for students' reference, though they likely will be of limited help for students who did not attend the corresponding lecture.

This course will also abide by the university attendance policies.

Tentative Schedule

The following is a rough outline of subjects that the course will cover, by week. The time spent on certain subjects will often be somewhat more or less than what is estimated here, and certain subjects will be reiterated and expanded upon throughout the course.

- Week 1: Syllabus, RStudio and R basics
- Week 2: Objects, data types, data structures, functions, packages
- Week 3: Reading datasets, basic dataset properties
- Week 4: Basic graphics and plots, basic dataset manipulation. **Practice 1, Homework 1 due**
- Week 5: Dataset manipulation, control structures
- Week 6: Introduction to RMarkdown
- Week 7: Introduction to LaTeX. **Practice 2, Homework 2 due**
- Week 8: RMarkdown continued, introduction to the `tidyverse`
- Week 9: Tidyverse data operations. **Practice 3, Homework 3, Project 1 due**
- Week 10: Tidyverse data operations
- Week 11: `tidyverse` graphics, introduction to statistical modeling
- Week 12: Statistical modeling. **Practice 4, Homework 4 due**
- Week 13: Statistical modeling continued
- Week 14: Miscellanea: data pulling operations, regular expressions
- Week 15: NO CLASS (Thanksgiving break)
- Week 16: Project presentations. **Practice 5, Homework 5 due**
- Finals week: Project presentations. **Final project due**

Assignments and Grading

The course will have 5 brief practice assignments, 5 homework assignments, and 2 projects (a midterm and final). The practice and homework are structured as 5 pairs of assignments, with each pair covering the same material at different levels of difficulty (practices more simple, homeworks more challenging). All assignment submissions will be completed via Canvas and will require access to R and RStudio.

Practice assignments Each practice assignment will be worth 3% of the final grade. They will contain short and simple exercises to help maintain and solidify your understanding of concepts in lecture before the corresponding homework is assessed. Each practice assignment will be issued 3-4 days prior to the corresponding homework assignment and will be due on the same day as this homework. While you will have 13-14 days to complete each practice, you are encouraged to do the practice early and to complete most or all of it prior to doing the corresponding homework. Accordingly, each practice will list a suggested deadline: if you complete the assignment by the suggested deadline, you will very likely be able to receive a grade and feedback prior to the due date for the corresponding homework.

For practice assignments, you are allowed to consult any online resources that you can find, and you may consult with other students for help. You may not, however, copy answers from other students, nor have them write any code for you at any point while completing the assignment. Accordingly, all students will need to submit their own copy of each practice. If you have questions about the resources allowed for practice, please direct them to the instructor. For questions about grades and feedback on practice assignments, please direct them to the TA (who will grade and provide feedback on the practices and homeworks).

Homework Each homework assignment will be worth 9% of the final grade. These will contain more thorough and challenging exercises than those in the practice, requiring you to put together multiple concepts at once to write realistic R programs. You will be given roughly 10 days to complete each homework. As mentioned above, the homework will be assigned concurrently with the corresponding practice. If you have questions about the grade or feedback that you receive on homework, please direct them to the TA.

Projects There will be 2 projects assigned in the semester: one to be submitted near Homecoming week, worth 15% of the final grade, and the other (which will be presented before the class) to be submitted on or shortly before finals week, worth 25% of the final grade. Projects can be done individually or in groups of 2 or 3 students, at the discretion of students. Each member will be graded individually based on their contributions.

Please direct questions about grades and feedback on the projects to the instructor (who will grade them).

Other Notes about Grading

- If you submit your practice assignment by the suggested deadline, we will make diligent effort to have it graded within a week (giving you 3 days or so to revise your homework based on the feedback on the practice). If you submit after the suggested deadline, you may receive feedback before the homework is due, but this is not guaranteed.
- Homework assignments should be graded within 2 weeks of their due date.
- If you use submission comments on Canvas, please only use them prior to receiving a grade on an assignment. If you comment on a submission after receiving a grade, it will most likely not be seen. Please use email or Canvas messages for these inquiries.

A statement about assignment and grading philosophy for this course This course is intended to give you a working knowledge of the real world process of writing code and producing useful output. It's also meant to get you comfortable using the huge wealth of resources at your side to help you succeed. Rarely, if ever, will you be asked in real life to write code that needs to be correct on the first try, with

no troubleshooting, in a handful of minutes. Accordingly, I find it unnecessary to assess students in such a manner. This course is also not intended to make you struggle for a good grade. I intend it to function as a means for you to expand your skills and resources as a programmer and data scientist.

Late Work This course implements a Base Policy (applies to all assignments), and homework-only Late Hours Provisions. Any circumstances regarding late assignments not described below will be handled according to the discretion of the instructor. Please direct all inquiries or concerns about late work policy to the instructor.

Base Policy The following applies to **all assignment types** (practice, homework, and projects):

1. If the student does not have an excused reason for late submission:
    - Assignments that are 0-24 hours late will be given a 20% penalty.
    - Assignments that are 24-48 hours late will be given a 40% penalty.
    - Assignments that are more than 48 hours late will not be accepted and will be given a 0% grade.
2. If the student does have an excused reason for late submission, the instructor will arrange an alternate deadline for the student in consideration of their circumstances. Beyond this deadline, the above late penalties will apply.

Special Provisions for Homework For homework assignments, you will receive additional "Late Hours" provisions that, by default, apply before the penalties of the Base Policy are considered. **These provisions are only for homework. Not for practice, and not for projects.** These provisions are as follows:

- Students will be given 3 days (72 hours) of late submission time. These hours act as a "bank" of time to "pay" for lateness on homework, regardless of the reason for late submission. If you use late hours on a homework, it's as though your submission deadline was extended by the number of late hours used.
- Late time is cumulative across homework assignments. For example, if you submit your first homework 24 hours late, you will have 48 remaining late hours to apply to any combination of your remaining homework assignments.
- Late hours will be calculated individually per homework, and will be rounded down to the nearest hour.
- If you have remaining late hours, they will be applied by default to a late assignment. If you wish not to use your late hour budget for a given assignment, please state so in a comment when you submit the assignment.
- Your late hour usage will be visible to you in Canvas.
- If you receive approval for a deadline extension, your late hours will apply in the same manner to the new deadline.

If you have used all of your late hours, or if you choose not to use late hours for a given homework, you will be graded according to the Base Policy penalties for any lateness not covered by your late hours.

A	94-100%
A-	90-93%
B+	87-89 %
B	83-86 %
B-	80-82%
C+	77-79%
C	73-76%
C-	65-72%
D+	57-64%
D	50-56%
F	$\leq$ 49%

Final Grades

- 5 practice assignments (3% each): 15%
- 5 homework assignments (9% each): 45%
- 2 projects (15% and 25%): 40%

The grading scale below will be used.

Final percentage grades will be rounded to the nearest percentage point when assigning letter grades. Any decision to curve grades will be made at the discretion of the instructor. In this case, grades would only be curved upward, and the scale below will be used for the curved grades.

Grade points will be assigned to letter grades according to the university's grading policy.

Important Dates

1. Monday, September 2: NO CLASS (University holiday)
2. Friday, October 18: NO CLASS (Homecoming)
3. Monday, November 11: NO CLASS (University holiday)
4. November 25-30: NO CLASS (Thanksgiving break)
5. Wednesday, December 4: Last day of regular class
6. Thursday, December 12, 3:00-5:00 PM: the scheduled final exam time for this course. I will likely use this time as one of the options for students to present their final projects. We will discuss this possibility closer to the end of the semester.

Accommodations

Students with disabilities who experience learning barriers and would like to request academic accommodations should connect with the disability Resource Center. It is important for students to share their accommodation letter with their instructor and discuss their access needs, as early as possible in the semester.

Course Evaluation

Students are expected to provide professional and respectful feedback on the quality of instruction in this course by completing course evaluations online via GatorEvals. Guidance on how to give feedback in a professional and respectful manner is available at https://gatorevals.aa.ufl.edu/students/. Students will be notified when the evaluation period opens, and can complete evaluations through the email they receive from GatorEvals, in their Canvas course menu under GatorEvals, or via https://ufl.bluera.com/ufl/. Summaries of course evaluation results are available to students at https://gatorevals.aa.ufl.edu/public-results/.

Academic Honesty

All instances of academic dishonesty will be handled according to the university's academic honesty policies (see pages 11-14, in particular).