36-350 – Statistical Computing

Fall 2020

Instructor: Peter Freeman (`pfreeman@cmu.edu`)

Office Hours: TR 9:00 - 10:20 AM EST/EDT (Zoom link provided in Canvas announcement)

Useful Texts: *R for Data Science* by Grolemund & Wickham (`r4ds.had.co.nz/index.html`)
Advanced R by Wickham (`http://adv-r.had.co.nz/`)
R Cookbook by Teetor
Data Wrangling with R by Boehmke
Practical SQL by DeBarros

Web Site: `http://canvas.cmu.edu/courses/17620`

Statistical Computing is a one-semester course that will introduce you to the fundamentals of computational data analysis, as carried out in the `R` programming language, and to the fundamentals of working with relational databases, as illustrated with `PostgreSQL`. No previous knowledge of either is required.[1]

You will spend the bulk of your time in this course working with `R`. Using it, you will learn how to, e.g.,

- define and manipulate vectors, data frames, and lists;
- process strings and learn how to apply regular expressions;
- input and pre-process data (i.e., how to "wrangle data");
- write functions (with proper coding style);
- simulate univariate and multivariate data via random sampling;
- apply methods for, e.g., integration, approximation, optimization, and dimension reduction;
- work with date-and-time-based data; and
- apply debugging tools and unit testing.

Your work will be recorded using `R Markdown`, a platform that facilitates communication between you, your clients, and fellow statisticians (and allows you and others to easily replicate results!).

Later in the course, you will use `PostgreSQL` to learn how to work with the "CRUD" operations (Create, Read, Update, and Delete) and how to join elements of multiple tables. Note that in contrast with the `R` content in this course, the `PostgreSQL` content is relatively introductory, and one can expand upon this content by taking, e.g., 15-445 or 70-455, etc.

The work that you will carry out in Statistical Computing maps to some of the learning outcomes that Dietrich College has defined as vital to the full CMU undergraduate experience. Specifically:

- self-directed learning;
- critical thinking (statistical thinking, computational thinking); and
- complex problem solving.

[1] Since learning `Python` is not required for the core Statistics & Data Science degree, there is no `Python`-related content in this course.

Course Details

Software. At stated on the previous page, in this course we will be working with `R` and `PostgreSQL`. Both are free. If you have not done this previously, you should begin by going to `www.r-project.org` and downloading and installing the build of `R` that is appropriate for your computer, and then you should go to `www.rstudio.com` to download `RStudio`, a popular GUI for `R`. Later in the course, go to `www.postgresql.org/download` and download and install the build of `PostgreSQL` that is appropriate for your computer.

Classroom Time. *In Fall 2020, the course will be online only: there will be no classes, per se.* I will be available during class sessions on Zoom (meeting ID to be provided in a Canvas announcement). Note that the exercises will be based upon material that I introduce via slides created using `R Markdown`, with each week's slides made available to you on `Canvas` prior to Tuesday, along with Zoom recordings of me going through them. (*It is up to you to read the slides and watch the recordings! I will not lecture during class sessions.*) The slides will provide an introduction to the current topic(s), but are not meant to be exhaustive references...think of them as having just enough material to get you started. Lab exercises usually refer back to specific slides. You should supplement your reading of the slide presentations with reading documentation for `R` functions and with Googling. Just as you would do in "real life." (Class motto: "Stack Overflow is your friend.")

It is vital that you have access to a computer to complete your assignments. If you do not have access to a computer, please let me know immediately so that you and I can work out an accommodation.

Labs. Lab assignments are to be turned in each week by Thursday at 9:00 PM Pittsburgh time (unless otherwise noted). You will record your answers in an `R Markdown` format file and knit this file into a `PDF` document, which you will subsequently upload to `Gradescope`.[2] Don't leave submission to the very last minute: it will take you some time to map each lab questions to a specific page of your `PDF` document. If you are not starting at least 10 minutes early, you're risking the penalty. Your submitted lab report will be graded on completeness (and originality...meaning, don't submit a copy of someone else's lab in your name). You may discuss your work with others but your final written answers must be your own. Late lab submissions will be accepted with a grade penalty: 0-6 hours late, -20%. Any submissions after 11:59 (Pittsburgh time) PM on Thursdays will not be graded.

Homework. Homework assignments based on a particular week's topics are to be uploaded each week by Friday at 9:00 PM Pittsburgh time (unless otherwise noted). They are otherwise identical to the labs. Late homework submissions will be accepted with a grade penalty: 0-6 hours late, -20%. Any submissions after 11:59 PM (Pittsburgh time) on Fridays will not be graded.

Again: `R Markdown` files are not acceptable in place of `PDF` documents.

R Project. There will be a final `R` project, which you may think of as a "super-homework" that you will complete during Week 11. During this week, there will be no new material, nor will there be a lab or homework: all course time will be devoted to working through the project.

Quizzes. At the end of most weeks, there will be `Gradescope`-based quizzes. These quizzes are envisioned as "low-stakes assessments," i.e., they should be viewed as opportunities for you to review the note slides and for me to reinforce important points raised in class, and not something to obsess over. (You have other classes to obsess over, presumably.) The quizzes will be open-book/open-notes, and will involve multiple choice and/or short-answer questions. Given the low-stakes/open-resource nature of the quizzes, I would ask that you please do not confer with others in the class while taking them. More details (such as scheduling and the length of time allowed to take the quizzes, etc.) are TBD and will be posted as announcements on `Canvas`.

[2] If you cannot successfully set up PDF knitting in `R Markdown`, you should knit to HTML and then convert your file to PDF at, e.g., `html2pdf.com`. `R Markdown` files themselves are not acceptable for submissions!

⇒ **Very Important Note!** I will not grant last-minute accommodations because you are sick or because of a family emergency, etc., etc., *without the explicit consent of your academic advisor*! Contact your advisor first, and have your advisor contact me directly.

If you miss a final deadline (e.g., 11:59 PM Pittsburgh time on Thursday for labs, etc.) without previously contacting me, and/or without having your advisor previously contact me, then the grade for the assignment is zero!

Email. All course-related email should be sent to `pfreeman@cmu.edu`. I will respond to it if it is appropriate to do so. Sending email does not shift any responsibility from you to me; you are still responsible for completing your assignments. In particular, do not send complicated questions or requests via email; replies will not be given for email questions or problems requiring lengthy (more than a couple of sentences) responses. These types of communications should be done in person.

Piazza. `Piazza` will be used for discussion of class assignments. While you may make yourself anonymous to your classmates, your identity will be known to me. So . . .

(a) Do not provide answers to homework problems, or discuss test problems, etc., until after homework is turned in, or all makeup tests have been taken, etc. Violations of this rule will result in what I will call a "doubling penalty": the first time, a 1% final grade penalty; the second time, a 2% final grade penalty; etc. The penalties are cumulative, e.g., after the second violation your final grade will be reduced by 3% total, etc.

(b) Be considerate! This is *not* an Internet comment board (though you should be considerate on one of those as well, I guess). Inappropriate content will be removed. Repeated lack of consideration will lead to my implementation of the "doubling penalty," at my discretion.

Also, do not expect an immediate answer to your question, particularly between the hours of 10 PM and 8 AM (Pittsburgh time).

Cheating. Cheating or plagarism on labs, homework, or projects will be dealt with as allowed under CMU policies: `http://www.cmu.edu/policies/student-and-student-life/academic-integrity.html`.

Disability Resources. If you require a special accommodation, such as flexible deadlines, please visit the Office of Disability Resources (`http://www.cmu.edu/disability-resources/`) to obtain appropriate documentation. I will make no allowances without documentation.

Diversity, Equity, and Inclusion. It is my intention that *all* students be well-served by this course, regardless of gender, sexuality, disability, age, socioeconomic status, ethnicity, race, religion, and/or culture. If you observe that you or someone else in the class is experiencing unfair or hostile treatment, please let me know, or contact the Center for Student Diversity and Inclusion directly (`csdi@andrew.cmu.edu`; 412 268-2150).

Mental Health. Many of your syllabi will have verbiage about taking care of yourself.

My take on this is that you have to realize that in the greater scheme of things, your performance in this course is not as important as your physical and mental health. Use your time wisely during the day, and sleep at night. Sleep during the day too, if you need to. Twenty years from now, you won't remember your grade in this course; in fact, you may have forgotten that you even took a statistical computing course. But you will remember if you were generally happy, or if you were stressed beyond belief. Strive for the happy memories; don't take on more courses and more responsibilities than you can reasonably handle. For some of you, this is easier said than done, but do try to scale back if you need to.

If, however, you find that you are struggling and need support, feel free to seek me out. (Would I hold your struggles against you? No, of course not.) Or seek out other resources, such as CandPS (8-2922 or `http://www.cmu.edu/counseling/`) or the Re:solve Crisis Network (888-796-8226). If you or someone you know is in a life-threatening situation, however, forego these resources and call the police immediately (8-2323 on campus, 911 off campus).

Grading

Labs	25%
Homework	50%
Final R Project	15%
Quizzes	10%

Important Dates

Week	Date	What's Happening
13	24 Nov (T)	No Class: Thanksgiving

Final Grade. Your final grade will be assigned via straight scale, i.e., 90% and above is an A, 80% - 89.99% is a B, etc. Depending on circumstances, I may adjust this scale, but only be decreasing the threshold percentages. On the basis of my experience, I doubt there will be any adjustment (but one never knows for certain).

Preliminary Schedule (Subject to Change)

Week	What's Happening
1	R: Vectors, Lists, and Useful Vector Functions
2	R: Control-Flow Constructs, For/While, Functions
3	R: String Processing and Regular Expressions
4	R: Split-Apply-Combine, Tidyverse
5	R: Data I/O, Web Scraping, Base R Visualization
6	R: Random Sampling, Simulation, ggplot Visualization
7	R: Numerical Tools/Optimization, Bootstrap
8	R: Unit Testing, Debugging, Git+GitHub
9	R: Multivariate Normal, Eigendecomposition, PCA
10	R: Date/Time Functionality, Time Series, Periodogram
11	**R Project**
12	SQL: Creating/Populating Tables, Data I/O
13	Thanksgiving: No Class
14	SQL: Selection, Math/Statistics Functions
15	SQL: Selection, Joining Tables