Carnegie Mellon University

Programming R for Analytics 94-842

Location: virtual

Semester: Fall Mini**, Year:** 2024

Units: 6**, Section(s):** A1

Instructor Information

Name	Michael Simko – "Mike"
Contact Info	msimko2@andrew.cmu.edu

TA Information

TA name	TBD
TA Contact Info	
Office location	
Office hours	

Course Description

- An introduction to R, a widely used statistical programming language. Students will learn to import, export and manipulate different data types, analyze datasets using common statistical methods, design, construct and interpret statistical models, produce a variety of different customized graphical outputs, create scripts and generate reproducible reports. There will be a focus on using this experience to apply these skills in public policy areas.
- Prerequisites: 91-801 Statistical Methods for Managers, or 95-796 Statistics for IT Managers
- A good knowledge of statistics is preferred, but this course is for anyone wanting to learn the basics of the R language and how to use the tools R offers to be able to do basic data analysis, statistical calculations, create graphical output and generate reproducible reports using R Markdown.

Learning Objectives

- Use R with RStudio, find and understand R documentation, and write functional scripts.
- Import, export and manipulate various types of stored data and datasets.
- Produce statistical summaries of continuous and categorical data.
- Produce basic graphics using standard functions.
- Create more advanced graphics using ggplot2 and plot.ly packages.
- Perform basic statistical and hypothesis tests.
- Develop classification and regression models and interpret the results.
- Create reproducible reports in R Markdown.

Learning Resources

No textbooks are required for the course, but some good resources include:

- Garrett Grolemund and Hadley Wickham, R for Data Science
- Phil Spector, Data Manipulation with R
- Paul Teetor, The R Cookbook
- Joseph Adler, R in a Nutshell
- Winston Chang, The R Graphics Cookbook
- Norman Matloff, The Art of R Programming: A Tour of Statistical Software Design
- Yihui Xie, J. J. Allaire, Garrett Grolemund, R Markdown: the Definitive Guide
- Paul Johnson, R Markdown Basics

There are many resources online to help when learning the R language. A few that are particularly relevant for this course are listed below.

- R Style guide
- RStudio cheatsheets
- Tidyverse cheatsheet
- An Introduction to factors in R
- A brief introduction to apply in R
- R Markdown Reference Guide

Course Work

Your grade in this course will be determined by a series of 5 weekly homework assignments (25%), quizzes (20%) and a final project (55%).

Assignments

Weekly assignments will take the form of a single R Markdown text file: namely, code snippets integrated with captions and other narrative. All assignments are due **Tuesdays midnight (Pittsburgh/eastern US time).** Each homework assignment usually has **5 problems**, each of which may have several parts. Your score for each assignment will be calculated according to the scheme outlined in the rubric below.

Homework Rubric

Total: 5 points

Correctness: Each problem will be worth **2 points**. Deductions will be made at the discretion of the grader.

Knitting: **-0.5 point** deduction if the Rmd file you submit does not knit correctly (i.e., if there are errors and no HTML file is produced when the grader attempts to knit your Rmd file.). If your Rmd file fails to knit, you will be contacted by the grader and will be given 24 hours to resubmit your homework. You will need to trace the source of the error(s) and correct it(them).

Style: Coding style is very important. With the exception of Homework 1, you will receive a deduction of up to **1 point** if you do not adhere to good coding style.

- No deduction if your homework is submitted with:
    - good, consistent coding style

    - appropriate use of variables
    - appropriate use of functions
    - good commenting
    - good choice of variable names
    - appropriate use of inline code chunks
- **-0.5** if coding style is acceptable, but fails on a couple of the criteria above.
- **-1** if coding style is overall poor and fails to adhere to many of the above criteria.

Quizzes

There will be 2 short quizzes scheduled during the later weeks of class worth 10 points each. Dates and times will be announced in advance. The purpose of these quizzes is to assess understanding of various concepts that are central to the class. You will have access to the course materials, online resources, whatever you need to answer the questions. However, you may NOT collaborate with others and your answers must be your own. Your score on these quizzes will count for 20% of your final grade.

Final project

The final project for the class will ask you to explore a broad policy question using a large publicly available dataset of your choosing. This project is intended to provide students with the complete experience of going from a study question and a rich data set to a full statistical report. Students will be expected to (a) explore the data to identify important variables, (b) perform statistical analyses to address the policy question, (c) produce tabular and graphical summaries to support their findings, and (d) write a report describing their methodological approach, findings, and limitations thereof.

A separate rubric exists to more detail the final project requirements which will be on Canvas, and discussed during class time. While students may work in small groups to decide on appropriate statistical methodology and graphical/tabular summaries, **each student will be required to produce and submit their own code and final report**.

Regardless of grading basis, students must receive a score of at least 50% on the final project in order to pass the class.

Grade	Percentage Interval
A+	100-97% Outstanding
A	96-93% Excellent
A-	92-90% Very Good
B+	89-87% Good
B	86-83% Acceptable
B-	82-80% Fair
C+	79-77% Poor

C	76-73% Very Poor
C-	72-70% Minimal Passing
Fail	<70%

Course Grading

Your final course grade will be calculated according to the following breakdown.

HW Assignments	25%
Quizzes	20%
Final project	55%

Late submissions

Homework is to be submitted on the due date indicated (usually Tuesdays as mentioned above). **Late homework will NOT be accepted for credit**.

Collaboration

You are encouraged to discuss homework, quizzes and project work, with your fellow students and as a group during class sessions. However, **any work you submit must be your own**. You must acknowledge in your submission any help received on your assignments. **That is, you must include a comment in your homework submission that clearly states the name of the resource or reference from which you received assistance.** Submissions that fail to properly acknowledge help from other students or non-class sources **will receive no credit**. Copied work **will receive no credit.** Any and all violations **will be reported** to Heinz College administration.

All students are expected to comply with the CMU policy on academic integrity. This policy can be found online at http://www.cmu.edu/academic-integrity/.

What constitutes plagiarism in a coding class?

The course collaboration policy allows you to discuss the problems with other students, but requires that you complete the work on your own. Every line of text and line of code that you submit must be written by you personally. You may not refer to another student's code, or a "common set of code" while writing your own code. You may, of course, copy/modify lines of code that you saw in lecture or lab.

The following discussion of code copying is taken from the Computer Science and Engineering Department at the University of Washington. You may find this discussion helpful in understanding the bounds of the collaboration policy.

"[It is] important to make sure that the assistance you receive consists of general advice that does not cross the boundary into using code or answers written by someone else. It is fine to discuss ideas and strategies, but you should be careful to write your programs on your own."

"You must not share actual program code with other students. In particular, you should not ask anyone to give you a copy of their code or, conversely, give your code to another student who asks you for it; nor should you post your solutions on the web, in public repositories, or any other publicly accessible place. [You may not work out a full

communal solution on a whiteboard/blackboard/paper and then transcribe the communal code for your submission.] Similarly, you should not discuss your algorithmic strategies to such an extent that you and your collaborators end up turning in [essentially] the same code. Discuss ideas together, but do the coding on your own."

"Modifying code or other artifacts does not make it your own. In many cases, students take deliberate measures -- rewriting comments, changing variable names, and so forth -- to disguise the fact that their work is copied from someone else. It is still not your work. Despite such cosmetic changes, similarities between student solutions are easy to detect. Programming style is highly idiosyncratic, and the chance that two submissions would be the same except for changes of the sort made easy by a text editor is vanishingly small. In addition to solutions from previous years or from other students, you may come across helpful code on the Internet or from other sources outside the class. Modifying it does not make it yours."

"[I] allow exceptions in certain obvious instances. For example, you might be assigned to work with a project team. In that case, developing a solution as a team is expected. The instructor might also give you starter code, or permit use of local libraries. Anything which the instructor explicitly gives you doesn't normally need to be cited. Likewise, help you receive from course staff doesn't need to be cited."

Generative AI

Like any other online resource, generative AI (ChatGPT, Google Bard, Bing, etc) are allowed as additional tools in this class for learning and generating correct R code. However, your responsibilities as a student remain the same – ***you must follow the academic integrity guidelines of the university and the guidance for citations given above***. If you use any generative AI, you are required to cite the tool's contributions to your final work – using these resources without proper citation is considered plagiarism. Ultimately, ***you are responsible for the content that you submit***.

If you have any questions about any of the course policies, please don't hesitate to ask directly.

Course Policies

- **Attendance & Participation**: Regular class attendance is expected, either live and in person, or through video link, but, accommodations can be made for absences with good reason. Questions during class and discussions are ***strongly encouraged***. The best way to learn R is by creating code, making mistakes, and finding ways to produce results. Course time will be interactive with students encouraged to use their laptops/computers to connect and participate in the course as well as run R code and follow along with real time exercises. Course information in the form of R Markdown created slides will usually be shared before the start of class, and any questions about the material can be asked during the class session. There will be exercises during some of the classes and you will be asked to code along and participate by answering questions and offering code syntax and format.
- **Accommodations for students with disabilities**: If you have a disability and require accommodations, please contact Catherine Getchell, Director of Disability Resources, 412-268-6121, getchell@cmu.edu. If you have an accommodations letter from the Disability Resources office, I encourage you to discuss your accommodations and needs as early in the semester as possible. We will work with you to ensure that accommodations are provided as appropriate.
- **Statement on student wellness**: As a student, you may experience a range of challenges that can interfere with learning, such as strained relationships, increased anxiety, substance use, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns or stressful events may diminish your academic performance and/or reduce your ability to participate in daily activities. CMU services are available, and treatment does work. You can learn more about confidential mental health services available on campus at: http://www.cmu.edu/counseling/. Support is always available (24/7) from Counseling and Psychological Services: 412-268-2922.

- **Mobile Devices**: Mobile devices are discouraged during class time. However, with the reality of participating remotely, it may be necessary to use one. When live, and in person on campus, the use of phones in class is prohibited. If you need to take a call, you may be excused and leave the room, but avoid this whenever possible. Focus should be on the class during the sessions either when in person, or remote.

Course Schedule

Week	Theme/Topic	Learning Outcomes Addressed	Assignments Due
1	Introduction to R and RStudio	Foundational concepts of R: data types, function calls, RStudio basics	none
2	Importing, exporting, manipulating data and R Markdown	More detailed commands for wrangling datasets, data-frames, tibbles, packages	HW#1 - first R script
3	Data analysis and tidyverse	EDA, statistical tests, base-R graphics	HW#2
4	Basic and advanced graphics	More statistics, ggplot2 graphics	HW#3 and first quiz
5	Regression tools and modeling	Simple linear and multiple linear regression	HW#4
6	CARET package and classification	Classification models, metrics and measures	HW#5 and final quiz
7	Advanced Topics / Course summary / wrap-up / final questions and discussions	Any class requests, use of maps, text analysis, web scraping, others topics	none