CSCI E-5a: Programming in R

Spring 2022

Official Syllabus (Version 1.0)

Instructor: Theodore Hatch Whitfield, Sc.D.

Teaching Assistants: TBD

Time: Mondays, 5:10 PM – 7:10 PM, or on demand

Start Date: Monday, January 24, 2021

E-mail: twhitfield@fas.harvard.edu

Introduction

CSCI E-5a: *Programming in R* is an introductory course in computer programming using the R language. The course has a number of special features:

- The course is designed as an introductory course specifically intended for students with no previous programming experience.
- The course emphasizes both R-specific constructs as well as general principles that apply to all programming languages.
- All course materials use the modern R notebook format, which allows for integration of text, code, output, and graphics in a single document.
- The course format includes a weekly optional web conferencing session, but all course video is available on-demand and students may engage with the course in a fully asynchronous manner.
- Students will acquire a practical skillset for importing, managing, transforming, and summarizing complex datasets.
- Students will develop a repertoire of powerful visualization methods.

At the end of this course students will have a toolkit for working with data that they can use for their own projects, as well as for further coursework in statistics, machine learning, and data science.

Catalog Description

This course is an introduction to the R programming language, one of the most popular languages for modern data science. Intended for students with no previous coding experience, this course covers fundamental concepts such as variables, functions, flow of control, data structures, and data management. Special attention is focused on practical skills such as working with missing data, finding and repairing corrupted values, and summarizing variables. Visualization techniques are emphasized throughout the course, and students develop a repertoire of graphical tools such as histograms, scatterplots, line charts, bar plots, and stripcharts. Assignments are developed in the popular R notebook format, allowing for integration of code, output, and graphics, with an emphasis on robust and reproducible analysis.

Primary Learning Goals

Upon completion of the course, students will be able to:

- Import, manage, transform, and summarize datasets.
- Detect and repair damaged or missing data.
- Visualize data using a range of standard graphing techniques.

In addition, students will also develop an understanding of standard best practices for programming that will extend to other computer languages such as Python, Javascript, etc.

Intended Audience

CSCI E-5a is designed for beginning students in statistics, data science, and machine learning, as well as working professionals seeking to develop their analytics practice. No prior experience with computer programming is assumed.

What CSCI E-5a is NOT

CSCI E-5a does not cover certain topics:

- This is not a course in data analysis or statistical modeling.
- While the course is comprehensive in the sense that it covers all the fundamentals of the R language, it is not exhaustive, and we won’t discuss every detail.
- The course does not discuss related technologies such as Shiny or interfacing with database management systems.

Prerequisites

A strong command of high-school algebra and analytic geometry.

Course Format

The official listing for the course format is “Online (live or on demand) web conference”. This means that I will be running an optional web conferencing session each Monday night from 5:10 PM to 7:10 PM (Eastern Standard Time), every week from January 24 through May 2, except for March 14. These sessions will be conducted via Zoom, and will feature a combination of lecture, demonstration, and discussion. Students are not required to participate in the Monday night web conference, and there is no penalty for non-attendance.

Assessments

The course assessments will consist of three components:

- 11 problem sets
- 1 midterm assessment
- 1 comprehensive assessment

The final grade is calculated as a weighted average of the three components: the overall problem set score is 20%, the midterm assessment is 30%, and the comprehensive course assessment is 50%. Grades are calculated on an absolute scale, with 93 and above receiving an A, 90 to 93 an A-, etc. No scaling or curving is used.

CSCI E-5a is taught at the level of an introductory Master’s level graduate course. Thus, graduate students are expected to complete 100% of the assignments. Undergraduates are expected to complete 75% of the assignments, and any work above this will be pro-rated. For instance, for a problem set with 8 problems, a graduate student must complete all 8 problems to receive full marks for the assignment. However, an undergraduate only needs to complete 6 problems in order to receive full marks, and if an undergraduate successfully completes more problems these will count as extra credit.

Course Materials

The R notebooks, slides, and videos for the course form the full set of course materials, and all notebooks and slides will be available for download. No additional text or reference is required. However, some people like to have a professionally produced textbook, so I will make some suggestions for books that would be useful supplements to CSCI E-5a. One book that would be very useful as a companion volume would be:

- *R in 24 Hours*, by Andy Nicholls, Richard Pugh, and Aimee Gott (SAMS Publishing, ISBN: 978-0672338489)

This is an excellent reference text, and its coverage is quite thorough. You can purchase a PDF version of this text from the publisher’s website.

Another book that would be useful for the course is:

- *R for Data Analysis in easy steps*, by Mike McGrath (In Easy Steps, ISBN: 978-1840787955)

Despite its title, this book doesn't really have much to do with "data analysis". Instead, it consists of 2-page tutorials on various aspects of R programming, and while it's hardly a comprehensive reference it's a very valuable cookbook. Again, you can purchase a PDF from the publisher's website.

At a more advanced level, I recommend these two texts:

- *The Art of R Programming*, by Norman Matloff (No Starch Press, ISBN: 978-1593273842)
- *The Book of R*, by Tilman M. Davies (No Starch Press, ISBN: 978-1593276515)

These are excellent references at a slightly more sophisticated level than the previous two texts. Once again, PDFs are available for purchase at the publisher's website.

There is one more book that deserves mention:

- *R for Data Science*, by Hadley Wickham and Garrett Grolemund (O'Reilly Media, ISBN: 978-1439840955)

This is a very popular book, written by two prominent members of the R community. Personally, I think that it has a lot of excellent material in it, but I'm not sure that it's a great text for a beginner to learn from. The complete text is freely available online at:

https://r4ds.had.co.nz/

Schedule

Week 0: Course Orientation (Monday, January 10 – Monday, January 24)

In the two weeks before the semester starts, we'll review course policies and get our computing environment set up. Then we'll learn how to navigate the Canvas website, use RStudio, work with R notebooks, and submit assignments into the Gradescope grading system.

- Problem Set 0 assigned, due at 11 PM on Monday, January 24.

Week 1: Base R Graphics (Tuesday, January 25 – Monday, January 31)

We'll start the course by learning the fundamentals of R graphics. At the end of this lecture you'll be able to specify plotting regions and draw basic geometric shapes such as points, lines, polygons, and curves.

- Problem Set 1 assigned, due at 11 PM on Monday, January 31.

Week 2: Numeric Values (Tuesday, February 1 – Monday, February 7)

In this lecture, we'll learn the basic syntax for fundamental arithmetic operations, and review the concept of operator precedence. We'll see how to store the results of our calculations in variables, and we'll encounter the special values **Inf**, **-Inf**, and **NaN**. We'll then consider a variety of carefully selected examples of computations ranging from baseball statistics to Likert scales that we'll revisit throughout the course. We'll finish by learning our first data visualization, the classic pie chart.

- Problem Set 2 assigned, due at 11 PM on Monday, February 7.

Week 3: Vectors (Tuesday, February 8 – Monday, February 14)

This lecture focuses on numeric vectors, perhaps the most important data structure in R. We'll learn how to construct vectors and how to operate with them, as well as exploring some important vector functions. We'll begin our study of indexing techniques, including the powerful method of named vectors. And we'll finish by learning how to create stripcharts, a powerful graphical technique that enables us to visualize the values in a numeric vector in fine detail.

- Problem Set 3 assigned, due at 11 PM on Monday, February 14.

Week 4: Iteration (Tuesday, February 15 – Monday, February 21)

One of the most important concepts of computer programming is the idea of repetition, and in this lecture we'll learn how to automate this process using **for** loops. We'll learn about counter and accumulator variables, and see how to use these techniques to solve some difficult computational problems. We'll finish by learning about histograms, another graphical method for visualizing the values in a numeric vector in a more abstract, high-level manner.

- Problem Set 4 assigned, due at 11 PM on Monday, February 21.

Week 5: Logical Values (Tuesday, February 22 – Monday, February 28)

Now we turn to another important data type in R: logical values. We'll study the basic operations and functions for logical data, as well as techniques for summarizing this type of data. We'll see how to generate logical vectors from numeric data, and we'll study the powerful technique of indexing using logical values. In addition, we'll begin our study of missing data and the special value **NA**. We'll also learn about conditional branching using **if** and **if/else** statements. Finally, we'll learn how to visualize numeric data by using density curves, and how to superimpose these density curves on histograms.

- Problem Set 5 assigned, due at 11 PM on Monday, February 28.

Week 6: Categorical Data (Tuesday, March 1 – Monday, March 7)

Next, we'll study categorical data, which can be implemented in R using a variety of different methods. We'll learn how to perform a variety of operations on categorical data, how to summarize factors using tables, and how to visualize factors using one-way barplots.

- Problem Set 6 assigned, due at 11 PM on Monday, March 7.

Week 7: Midterm Review (Tuesday, March 8 – Monday, March 14)

We've covered a lot of material in the first six lectures, so it's a good idea to take some time to consolidate all of this. This lecture is devoted to a complete review of everything that we've studied since the beginning of the course, and at the end you should have a solid understanding of all the basic concepts as well as their applications. After that, it's time for the midterm exam. I'll discuss each problem on the exam, and then you're on your own!

- Midterm Exam released at 5 PM on Monday, March 7, due at 11 PM on Monday, March 14.

Week 8: Functions and Strings (Tuesday, March 22 – Monday, March 28)

In this lecture we'll study functions in greater detail, focusing on named parameters and default values. We'll also learn how to create our own functions, using local variables and the global environment. Next, we'll consider some special functions for working with string values. This week's data visualization method is the boxplot, which provides a high-level view of the distribution of a set of numeric values.

- Problem Set 7 assigned, due at 11 PM on Monday, March 28.

Week 9: Data Frames (Tuesday, March 29 – Monday, April 4)

Now we encounter the second major R data structure: the data frame. We'll learn how to construct data frames, and we'll meet some of the data frames that are built into R. We'll also investigate some useful data frame functions, and begin our study of data frame indexing. Then we'll study how to read data in from CSV and tab-delimited files, as well as from Excel spreadsheets. We'll study lists, which are another important R data structure, and we'll see how data frames are special cases of these lists. Our data visualization technique for this lecture is the scatterplot, which enables us to visualize the relationship between two numeric vectors, and we'll see how to fit a least-squares regression line to a scatterplot.

- Problem Set 8 assigned, due at 11 PM on Monday, April 4.

Week 10: Data Management (Tuesday, April 5 – Monday, April 11)

In this lecture, we'll learn how to manage data using operations on data frames. We'll see how to sort data, append rows and columns, and generate new columns in a data frame. We'll also study the process of merging two data frames. Finally, we'll learn how to create stratified stripcharts and boxplots, which enable us to visualize a range of numeric values across the levels of a categorical variable.

- Problem Set 9 assigned, due on Monday, April 11 at 11 PM.

Week 11: Summarizing Data (Tuesday, April 12 – Monday, April 18)

Now that we can work with data frames, we can investigate more sophisticated methods of summarizing data. We'll review how we use tables to summarize a single factor, and then generalize this to two-way tables that can summarize two factors together. We'll then study the **split()** function, the **tapply()** function, and the **aggregate()** function. Finally, we'll learn how to create two-way grouped barplots and multi-facet displays.

- Problem Set 10 assigned, due at 11 PM on Monday, April 18.

Week 12: The Tidyverse (Tuesday, April 19 – Monday, April 25)

The Tidyverse is a popular and modern collection of tools for working with data. In this lecture, we'll explore the **dplyr** and **tidyr** packages, and compare them with the data management functions in base R from the previous lecture. I'll also release a set of more challenging problems to give you some extra practice before the Comprehensive Course Assessment. Note: the material in this lecture and the next will not be examined in the Comprehensive Course Assessment.

- Practice Problems released at 5 PM.

Week 13: ggplot2 (Tuesday, April 26 – Monday, May 2)

The ggplot2 package is a modern approach to constructing complex data visualizations, but uses a very different approach to graphics than what we've seen with base R. This lecture is dedicated to helping you to understand the ggplot2 graphing paradigm, and to give you a strong foundation if you want to study this further.

Week 14: Comprehensive Course Review & Assessment (Tuesday, May 3 – Monday, May 9)

Can you believe it? We're already at the end of our course. We'll use our last lecture to review all of the course material, especially the material from the lectures after the midterm assessment. Then I'll step you through the Comprehensive Course Assessment, and you're on your own! Good luck!

- Final Comprehensive Course Assessment released on Monday, May 2, at 5 PM, due on Monday, May 9, at 11 PM.

Official Harvard Policies

Assessments

A letter grade will be given in accordance with Harvard Extension School's grading policy:

https://www.extension.harvard.edu/resources-policies/exams-grades-transcripts/grades

Accommodation Requests

Here's the official Harvard Extension School policy for accommodation requests:

*Harvard Extension School is committed to providing an inclusive, accessible academic community for students with disabilities and chronic health conditions. The Accessibility Services Office (ASO) (https://extension.harvard.edu/for-students/support-and-services/accessibilityservices/) offers accommodations and supports to students with documented disabilities. If you have a need for accommodations or adjustments, contact Accessibility Services directly via email at **accessibility@extension.harvard.edu** or by phone at 617-998-9640.*

Academic Integrity aka Cheating Policy

Here's the official Harvard Extension School policy on academic integrity:

You are responsible for understanding Harvard Extension School policies on academic integrity (https://extension.harvard.edu/forstudents/student-policies-conduct/academic-integrity/) and how to use sources responsibly. Stated most broadly, academic integrity means that all course work submitted, whether a draft or a final version of a paper, project, take-home exam, online exam, computer program, oral presentation, or lab report, must be your own words and ideas, or the sources must be clearly acknowledged. The potential outcomes for violations of academic integrity are serious and ordinarily include all of the following: required withdrawal (RQ), which means a failing grade in the course (with no refund), the suspension of registration privileges, and a notation on your transcript.

Using sources responsibly (https://extension.harvard.edu/for-students/supportand-services/using-sources-effectively-and-responsibly/) is an essential part of your Harvard education. We provide additional information about our expectations regarding academic integrity on our website. We invite you to review that information and to check your understanding of academic citation rules by completing two free online 15-minute tutorials that are also available on our site. (The tutorials are anonymous open-learning tools.)