Introduction to Data Science Using R

Preliminary Draft: September 1, 2016

Prof. Darin Christensen[1]

[1] 6341 Public Policy
darinc@luskin.ucla.edu
(310) 825-7196

1 *Course Information*

- **Location**: Public Affairs 4371
- **Schedule**: TR, 3-4:50p
- **Office Hours**: By appointment. Sign up at Public Affairs 6341.
- **Course Website**: https://piazza.com/ucla/fall2016/pubplc290dis3/home
- **Prerequisites**: This course assumes no prior knowledge of programming or statistics.

2 *Course Overview*

There is a growing demand that public policy be "data-driven." The White House, to take a prominent example, appointed its first-ever Chief Data Scientist in 2015. While ten weeks isn't enough time to prepare you for that job, this class introduces you to a set of common steps that data scientists use to transform a spreadsheet into a useful description or illustration. We cover acquiring, cleaning, merging, managing, summarizing, and visualizing quantitative data.[2] While these tasks can be accomplished using different software, this course demonstrates how to use R, a free and popular statistical program.

[2] If you're interested in effective data visualization consider taking Prof. Reber's *Arguing with Data* course.

3 *Learning Goals*

There are several goals for this course. By the end of the quarter, you should be able to:

1. Write down a recipe (i.e., a set of ordered steps) for transforming a dataset into a table or figure that describes a pattern;
2. Identify and apply functions/packages in R for accomplishing the steps of this recipe;[3]
3. Combine data from multiple (unrelated) sources; and
4. Apply best practices in writing code, including using clear naming conventions, commenting steps, and avoiding repetition. These conventions help ensure that code is error-free and legible to our colleagues and future selves.

[3] You won't be expected to memorize every R function, but rather identify sources for useful code snippets and understand how to adapt those snippets.

An important disclaimer: like any new language, R is challenging to learn, and you'll spend time fighting with the program. In those darker moments, try to remember two things. First, you are not alone; your peers and I can both relate and also offer help when you're hitting a wall.[4] Second, remember that the first goal of this course has nothing to do with R. Writing down the recipe that transforms data into insights requires careful, ordered reasoning (i.e., "algorithmic thinking"), not a specific piece of software.

[4] Our Piazza message board is a great place to post coding questions and get quick responses.

4 *Assignments*

- **Participation** (10%): There are several ways to participate in and outside of class. I encourage you to raise questions and make comments during our class meetings. In addition, I will reward students for assisting peers during in-class group work, as well as posting questions or responses to our Piazza discussion board.
- **Weekly Problem Sets** (30%): These assignments serve two purposes. First, they allow you to apply the reasoning and coding skills we cover in class. Second, some problems will help you get started on your final project. While there will not be much reading, you will need to set aside time for problem sets.

  All problem sets must be completed in R Markdown (discussed below). These assignments will be evaluated according to the following scale:[5]

  - ✓ : complete
  - ✓− : submitted, but incomplete
  - 0 : not submitted or less than half complete

[5] I will post detailed solutions online but cannot grade every problem. Please come to office hours or schedule an appointment if you have any questions.

- **Midterm Exam** (35%): The midterm exam offers an opportunity to assess your progress on the first two learning goals. As such, part of the exam will test your problem solving skills; another component will ask you to write R code.

  The mid-term will be conducted during class time. However, you can take the exam from the location of your choice. You will need R to complete the exam.

- **Final Project** (25%): You and a partner will write a quantitative report based on data of your choosing.[6] This project will allow you to practice the real-world tasks of acquiring, manipulating, and analyzing quantitative data and clearly conveying your approach and results. In addition to the written report, you will briefly present your findings during the last week of class. More detail about the format and rubric for the final project will be provided.

[6] This data needs to be approved by me via email before week 3.

5 Course Policies

5.1 Late Submissions

I don't want you to fall behind, so I will not accept late homework assignments. However, I understand that there are forces beyond your control, so I will drop your lowest homework grade.

Final projects will be docked a letter grade for every day that they are late. For example, an A project submitted in the first 24 hours after the deadline will receive an B. If that same project is submitted 24 to 48 hours after the deadline, the grade drops to a C.

5.2 Academic Accommodations

Students needing academic accommodations based on a disability should contact the Office for Students with Disabilities (OSD) at (310) 825-1501 or in person at Murphy Hall A255. When possible, students should contact the OSD within the first two weeks of the term as reasonable notice is needed to coordinate accommodations. For more information visit www.osd.ucla.edu.

5.3 Academic Integrity

Please review UCLA's rules related to academic integrity.[7] If you're feeling overwhelmed or are unsure about the collaboration policy, please speak with me; don't risk violating the honor code. Unfortunately, past MPP students have been investigated and punished by the Office of the Dean of Students.

[7] http://goo.gl/DCIwSN

Collaboration Policy:

- Problem Sets: You can work with peers (in groups of four or less). However, you must write every line of code and text that appears in any assignment that you submit. In short, never copy code or use other students' words to describe your approach or results.
- Midterm Exam: This is an individual assessment that helps us identify any knowledge gaps. I want to know what you know, so there's no collaboration.
- Final Project: You will work with a partner on your final project. You are responsible for ensuring that no written work is copied or summarized from other sources without appropriate attribution.

6 Resources

6.1 Mental Health

UCLA Counseling and Psychological Services offers services and programs in a confidential environment to promote mental health and wellness.

6.2 Academic Accommodations

Students needing academic accommodations based on a disability should contact the Office for Students with Disabilities (OSD) at (310) 825-1501 or in person at Murphy Hall A255. When possible, students should contact the OSD within the first two weeks of the term as reasonable notice is needed to coordinate accommodations. For more information visit `www.osd.ucla.edu`.

6.3 Software

- You will need to download and install R from this website: `https://cran.r-project.org/`.
- I *highly* recommend that you use RStudio, which can be downloaded here: `https://www.rstudio.com/products/RStudio/`. RStudio provides a nice interface for writing and executing R code and composing R Markdown documents.
- You will use R Markdown (`http://rmarkdown.rstudio.com/`) to write your homework assignments and final project. R Markdown allows you to create well-formatted documents that embed R code and outputs.[8]
- You may also use MS Excel or another spreadsheet program to peak at or enter data.

[8] I won't devote class time to teaching Markdown, but I can answer questions and point you to useful online resources.

6.4 Using R

The resource we will use most is *R for Data Science* by Wickham and Grolemund. An online version of the text is available here (`http://r4ds.had.co.nz/`) or you can purchase the book once it's published (`http://goo.gl/EUUNeh`).

There are many online and print resources for learning R. I've included a few below, but a Google search will turn up many others.

Online:

- Interactive Tutorials: Try R
- Examples: UCLA's Institute for Digital Research and Education

- Examples: Cookbook for R
- Crowd-sourced Solutions: Stack Overflow

Print:

- *R for Dummies* (`http://goo.gl/mel0fo`)
- *R Cookbook* (`http://goo.gl/Hh1REr`)

We are going to rely heavily on two R packages: `dplyr` for data manipulation (`https://goo.gl/Nc5gIV`), and `ggplot2` for visualization (`http://docs.ggplot2.org/current/`).

6.5 Data Sources for Final Project

Two general pieces of advice: first, if you have a topic and need help finding data, come to office hours; and second, look for datasets that contain several interesting variables (or can be merged with other datasets). If the dataset doesn't contain much information, then it's going to be difficult to write an interesting final project. Below are a few examples:

Domestic:

- New York City's Stop-and-Frisk Data (`http://www.nyclu.org/content/stop-and-frisk-data`)
- Los Angeles Open Data[9] (`https://data.lacity.org/`)

[9] Many cities have open data sites, including Chicago, NYC, and SF.

International:

- (Cross-National) Quality of Government Dataset (`http://qog.pol.gu.se/data/datadownloads/qogbasicdata`)
- World Development Indicators (`http://data.worldbank.org/data-catalog/world-development-indicators`)
- (Foreign) AidData (`http://aiddata.org/`)

7 Course Schedule

Legend: ⋆: Required; • Suggested.

Caveat: This schedule is approximate, and we may drop or reorder sections as the quarter proceeds.

0 *Before the First Class*

⋆ Complete the Try R interactive tutorials; and

⋆ Download, install, and open R *and* RStudio.

---

1 Introductions (1)

⋆ Chapters 1 and 2: *Introduction*. Garrett Grolemund and Hadley Wickham. *R for Data Science*. O'Reilly

• Mike Loukides. What is data science?, June 2010. URL https://www.oreilly.com/ideas/what-is-data-science

• Sharp Sight Labs. Why you should learn R first for data science, January 2015. URL http://www.r-bloggers.com/why-you-should-learn-r-first-for-data-science/

• Paul Curzon. A recipe for programming, 2014. URL https://teachinglondoncomputing.files.wordpress.com/2014/07/cs4fnissue16_pr_6_1.pdf

---

2 Description

2.1 Visualizing Data (2)

⋆ Chapter 3: *Data Visualisation*. Grolemund and Wickham.

⋆ Jonathan A Schwabish. An Economist's Guide to Visualizing Data. *The Journal of Economic Perspectives*, 28(1):209–34, February 2014

• H Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1), 2010

2.2 Manipulating Data Frames (2)

⋆ Chapter 5: *Data Transformation*. Grolemund and Wickham.

---

3 Wrangling

3.1 Importing Data (2)

⋆ Chapter 11: *Data Import*. Grolemund and Wickham.

3.2 Cleaning (or "Tidying") Data (1)

⋆ Chapter 12: *Tidy Data*. Grolemund and Wickham.

[Midterm Exam]

3.3 Working with Dates and Strings (2)

⋆ Chapter 16: *Dates and Times*. Grolemund and Wickham.
⋆ Chapter 14: *Strings*. Grolemund and Wickham.

3.4 Merging, Joining, and Appending (1)

⋆ Chapter 13: *Relational Data*. Grolemund and Wickham.

---

4 Coding

4.1 Functions (1)

⋆ Chapter 19: *Functions*. Grolemund and Wickham.

4.2 Iteration (2)

⋆ Chapter 21: *Iteration*. Grolemund and Wickham.
⋆ Chapter 20: *Vectors*. Grolemund and Wickham.

---

5 Skill Consolidation (1-2)

---

6 (Optional) Special Topics

6.1 Web Scraping

- *rvest* (`https://github.com/hadley/rvest`). Hadley Wickham.

6.2 GIS

- *GIS in R* (`http://www.nickeubank.com/gis-in-r/`). Nick Eubank.

---

7 Final Presentations (2)

[Final Project Due (7 December 2016 by 11:59 p.m.)]