MA415/MA615

Data Science in R

Version: 2016-09-06

Haviland Wright
64 Cummington Mall, Rm 233
Office hours: Th 2:30pm - 5:00pm

hav1126@bu.edu
skype(text please): havilandw

Class: CGS 527, MWF 3:00pm - 3:50pm
Discussion: CGS 527, W 2:00pm - 2:50

Teaching Assistant:
Xiaoyi Zhang, iamxyz@bu.edu
Office Hours: Wednesday 10:00 - 11:00

NOTE: This a new course. The syllabus will change during the semester. See Blackboard for the latest version.

Description

The purpose of this course is to provide you with the instruction and experience you need to develop and execute data analytic workflows using R. R is an open source programming language written by and for statisticians. It is supported by a broad community of developers who extend the language by writing packages that implement methods for data acquisition, exploration, manipulation, presentation, graphical visualization, statistical computations, and many other useful things. The R community also provides documentation, working examples of package applications, and active online discussions about how R can be optimized and enhanced for data analysis, presentations, and reproducible research. R can be accessed and used in many ways. In this course you will learn to use R in a development environment made with open source tools that support code development, revision control, R package management, and applications for creating documentation, presentations, and final deliverables. The integrating component of this environment is RStudio.

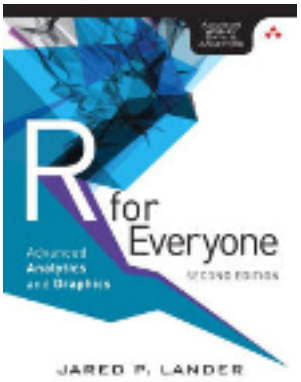
By the end of the course, you will have hands-on experience with data acquisition, visualization, cleaning, and organization; analysis and modeling; report preparation, dynamic web presentations, and delivery of reproducible research. You will have produced maps , analyzed data from relational databases, and used R to analyze data sets that exceed the size limits of R. Through the course exercises you will revisit much of what you learned in introductory probability and statistics, seeing how to put that knowledge to work.

This course is organized in two sections. The first section is focused on the R language itself and common R packages used for data acquisition, organization, and description. Assignments will include situations that are common in actual practice, including difficult-to-read data files, data coding issues, missing value coding, data presentation choices, and so on. You will be individually responsible for submitting assignments in this section of the course and will complete a project in which you prepare a data set for an initial descriptive presentation.

In the second section of the course, the focus will broaden, and you will use R as the core of an integrated environment for organizing complex data sets, preparing data analyses, and estimating parameters of statistical models. You will use R-based tools to assemble presentations, websites, and documents that contain text, mathematics, maps, and plots. You will practice methods for database access and maintenance and for loading and analyzing datasets that are larger than R can normally load. You will learn to deliver analytic results in self-contained modules that support reproduction of the analysis. In this second section of the course, you will work in teams and use online tools to organize and coordinate collaborative projects. Assignments will be submitted as team reports and presentations. There will be a final project.

Textbooks

Jared P. Lander, *R for Everyone* , Edition 2
Pearson Education, 2014



This book covers most of the topics that we will address in the course. Things are moving fast in the R universe, including *R for Everyone* which has been updated to Edition 2.

Hadley Wickham
Advanced R CRC
https://github.com/hadley/adv-r/
http://adv-r.had.co.nz/

This is a book authored by one of R's most active developers. Wickham is the author of the ggplot2, ggvis, dplyr, tidyr and many other packages that you will use in this course. This fact alone makes his newly released book worth having. But there's more. Wickham's clear explanations of the technical underpinnings of R will make R easier to understand and use. The book is available online and in hard copy.

References

Scott Chacon and Ben Straug
Pro Git , 2nd Edition
ProGit
https://git-scm.com/book/en/v2

In-Class Response System

We will use the Turning Technologies in-class response system during the semester. You will need to buy a license for the system unless you already have one for another class. Instructions are on Blackboard. You DO NOT need to buy a clicker. You may use the ResponseWare App that works on your phone.

Policies

For graduate students, registered for MA615, course policies are governed by Graduate School of Arts & Sciences. See GRS Policies, especially the GRS Academic Conduct Code.

For undergraduates, registered for MA415, course policies are governed by College of Arts & Sciences. See CAS Policies, especially the CAS Academic Conduct Code.

Attendance, Assignments, Grades

Assignments for this course will be posted and submitted on Blackboard. Grading will be weighted as follows: Homework 20%, Peer Assessment 10%, Midterm Project 30%, Final Project 30%, quizzes & attendance 10%.

Four kinds of assignments are planned:

- **Reading** assignments should be completed before class to encourage active classroom discussions.
- Readings will be accompanied by short **quizzes** that must be completed on Blackboard before the beginning of class. Graduate students will have required supplemental reading and reading quizzes.
- **Homework** assignments will focus on coding, data analysis, and written explanations. For some cases, graduate students will be given extended assignments to gain exposure to applications they are likely to need in other courses. Topics could include, for example generalized linear models, model diagnostics, and regularization. Keep in mind that homework will be evaluated on the basis of statistical content in addition to the quality of submitted code.
- Although homework will be graded as described in the previous bullet, some homework will also be submitted for **peer assessment** to provide code review practice.
- The **midterm** and **final projects** will integrate key topics in the course.
  - The midterm project will focus on preparing a dataset for analysis and modeling.
  - The final project will emphasize R functionality for combining statistical methods with graphical presentations that contribute insight and clarity.

Project assessment will focus on five criteria:

Organization: The work is well-organized and is presented in a form that is easy to read and understand. Analysis combines numeric, graphical, and verbal presentations.

Planning: The approach taken reflects a strategy formed by understanding the problem to be solved.

Execution: A minimum amount of code has been used. The code runs without error, reproducing the product that has been submitted

Clarity: The code is easy to read and understand. The code and associated comments form a seamless whole.

Curiosity: The work submitted reflects a process of exploration with many trials to make the analysis as clear as possible.

		Day				
Month	**Week**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Sept	1	5	6	7	8	9
		L A B O R D A Y		**Class 1** Intro, Why R?, R community, R-related websites		**Class 2** R studio, packages, git, projects, github, R notebooks, Submitting on Blackboard. R Markdown & LaTeX
	2	12	13	14	15	16
		Class 3 Variables and basic math Reference sheets		**Class 4** Basic R operations, data types, Missing data		**Class 5** Data structures, Reading data
	3	19	20	21	22	23
		Class 6 Data Structures Reading Data		**Class 7** Reading different kinds of data and data files Missing data, Null data, NaN		**Class 8** Visualization – built-in graphics, ggplot2, ggvis
	4	26	27	28	29	30
		Class 9 Visualization		**Class 10** R functions Code review procedures.		**Class 11** R Functions Flow control Vectorization
Oct	5	3	4	5	6	7
		Class 12 Debugging		**Class 13** Group Manipulation Apply, dplyr, Data.table		**Class 14** Data reshaping Tidy Data
	6	10	11	12	13	14
		C O L U M B U S D A Y	**Class 15** Text Regular expressions Reading data from the Web	**Class 16** Web scraping with R		**Class 17** Midterm Project Presentations Read, clean, organize, describe
	7	17	18	19	20	21
		Class 18 Midterm Project Presentations Read, clean, organize, describe		**Class 19** Github for teams & open source Probability & statistics in R		**Class 20** Probability & statistics in R
	8	24	25	26	27	28
		Class 21 Hands-on methods tour: EDA in R		**Class 22** Hands-on methods tour: Linear models in R		**Class 23** Hands-on methods tour: Linear models in R

		Day				
Month	**Week**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Nov	9	31	1	2	3	4
		Class 24 Hands-on methods tour: Time Series		**Class 25** Hands-on methods tour: Cluster analysis		**Class 26** Docs, Slides, & Sites in R Markdown / LaTeX / sweave / knitr
	10	7	8	9	10	11
		Class 27 Docs, Slides, & Sites in R Markdown / LaTeX / sweave / knitr		**Class 28** Mapping, Census and other geospatial data.		**Class 29** Mapping, Census and other geospatial data.
	11	14	15	16	17	18
		Class 30 Dynamic presentations Shiny, ggvis		**Class 31** Dynamic presentations Shiny, ggvis		**Class 32** Database access with SQL in R
	12	21	22	23	24	25
		Class 33 Database access with SQL in R		THANKSGIVING BREAK		
Dec	13	28	29	30	1	2
		Class 34 Working with databases in the cloud		**Class 35** Working with databases in the cloud		**Class 36** Working with big data sets: FF, spark
	14	5	6	7	8	9
		Class 37 Working with big data sets: FF, spark		**Class 38** Working with big data sets: FF, spark		Class 39 Final Projects
	15	12	13	14	15	16
		Class 40 Final Projects				