Data Management and Manipulation in R

Organizers:
Christina García, Ph.D., Spanish and Linguistics
Kelly G. Lovejoy, Ph.D., Spanish and Linguistics
Christopher G. Prener, Ph.D., Sociology

Spring, 2016

Contact Chris	**Seminar Meetings**
Office: 1918 Morrissey Hall	Select Tuesdays
Phone: 314-977-4276	11am to noon
Email: prenercg@slu.edu	Morrisey 3600 (GIS Lab)

Seminar Websites
Homepage: http://slu-data-science-seminar.github.io
GitHub Repositories: https://github.com/slu-data-science-seminar

Seminar Description

This seminar will provide an introduction to basic data management and manipulation using the programming language R. Using R for data analysis is becoming increasingly common because it is free, open-source, exceptionally flexible, and highly extensible - all features that set it apart from commercial statistical software. We will cover a number of topics, including the basic syntax of R commands, importing data into R, creating and modifying numeric and string data, and reshaping datasets. All of the examples will use sample data that comes pre-loaded with distributions of R. In order to get the most from this seminar series, we recommend selecting an additional dataset (preferably one that you are already familiar with) to practice these techniques. Programming is a 'use it or lose it' skill-set - the more you practice and work with these techniques, the more comfortable you will feel using R.

Software

We will focus on using the software application `RStudio` (link), which is an integrated development environment for R. It offers a number of advantages over using R's 'stock' programming environment. It can be downloaded for free from its developers.[1] Using `RStudio` requires that R is already installed on your computer. If you are a Mac user, you will also need to install `X11`, a.k.a. `XQuartz` (link).

[1] Be aware that there is also a paid version of `RStudio` - you do not need to pay for this software.

Resources

If you want a companion text for this seminar, we will be basing the seminar approach off of the following book:

1. Abedin, Jaynal and Kishor K. Das. 2015. *Data Manipulation with R*. 2nd edition. Birmingham, UK: Packt Publishing.

Since we are focused on using `R` for a small number of specific tasks, there is much about `R` and `RStudio` that we will gloss over or skip entirely. SLU's library has a wide range of books available on using `R`. Of those, there are three in particular that provide introductory-level material and are available as e-books:

2. Adler, Joseph. 2012. *R in a Nutshell*. Sebastopol, CA: O'Reilly. (link)
3. Cotton, Richard. 2013. *Learning R*. Sebastopol, CA: O'Reilly. (link)
4. Van der Loo, Mark P. J. 2012. *Learning RStudio for R Statistical Computing*. Birmingham, UK: Packt Publishing. (link)

We will also being posting a number resources on GitHub, a website that is designed for sharing and collaborating on the development of computer code. GitHub will be used to post examples of `R` scripts, a log of each seminar session, and other replication files. You do not need a GitHub account to download seminar resources. If you want to contribute to the development of these resources, however, you will need a (free!) GitHub account.

One of the great things about `R` is the wide variety of resources that are available online. Among them are `R` Bootcamp (link), *Introduction to Probability and Statistics Using* `R` (link), Quick-`R` (link), and UCLA's Institute for Digital Research and Education (link). If you have specific questions, a quick Google search or a search of the forums at Stackoverflow (link) will often identify helpful materials. Finally, `RStudio` maintains a curated list of web resources for learning `R` (link). Working with `R` and `RStudio` typically requires a significant amount of trial and error, more so than with other statistical packages. The resources online are often helpful starting places but will require additional adaptation to be applied to the specific problem you are trying to solve.

Tentative Schedule

1. January 19th - Intro to `R` and Scripting
2. February 2nd - Importing and Describing Data
3. February 23rd - Basic Manipulation with Numeric Data
4. March 15th - Basic Manipulation with String Data
5. April 5th - Advanced Data Manipulation with `dplyr`
6. April 26th - Reshaping Data