An Introduction To Statistical Programing In R:
A crash course on processing, analyzing, and visualizing data in R

Prepared for the 2018 BSOS Math Camp
(Draft version and subject to change)

Instructor: **Eric Dunford** | Ph.D. Candidate - Department of Government and Politics
Email: edunford@umd.edu
Website: ericdunford.com
Dates: **August 17, 2018** (Two 3-hour Sessions)
Time: 9am – 12pm & 1pm – 4pm
Location: TBD

Course Description

The rise of large-scale data collection has generated a need for fast, reliable ways to analyze information. Cleaning, processing, and visualizing data has become a growing necessity in social science research. The *R* statistical programing environment offers a reliable, open-source, and cost-free approach to data analysis. With a robust community of contributors, *R* allows for a wide range of different statistical analyses and data visualization approaches, making it the leading choice in data analytics.

This crash course will introduce incoming BSOS graduate students to the R programing environment in an effort to prepare them for their future methodological training in their respective programs. For many beginning users, manipulating and managing data in R can be frustrating. This course will focus on providing participants a basic knowledge of the R programming environment while simultaneously providing a practical data science toolkit that attendees can implement immediately. To this end, the course takes a 'Tidyverse' approach to R programming, which provides users an intuitive grammar for data manipulation. The goal is to provide a practical toolkit for analysis in R—without getting too bogged down in the nuts and bolts of functional programming—that they can then apply to their future academic training.

The course will take part across two 3-hour sessions. The first session will offer participants an underlying intuition of the R programming environment by focusing on the basics of (a) the syntax and environment, (b) data management, (c) graphics and presentation, and (d) modeling. Participants of the workshop will leave with an applied understanding of the R environment: specifically, importing and manipulating data, rendering graphics, presenting results, and implementing basic statistical models. The second session will focus on applied applications to different types of data science problems. This session intends to offer participants a general sense of what sorts of analyses and applications are possible in R.

All sessions will include example code to accompany the materials being presented. No prior statistical or computer programming knowledge is assumed.

Introduction to Statistical Programing in R

COURSE OVERVIEW

Morning Session

The basics: objects, data structures, and packages

- Getting started
  - An overview of *R* and the advantages to open source data analytics.
  - Installing *R* and *R* Studio
  - Understanding the *R* Studio GUI
- Data Types and Structures
  - What is an Object?
  - Differing types of data: integers, numeric, strings, factors.
  - Object structures: vectors, lists, matrices, arrays, and data frames.
  - Accessing information inside objects.
  - Object properties for different types of data
- Packages
  - What are packages?
  - Downloading, loading, and updating packages
  - Introduction to the *Tidyverse* suite packages.
- Importing, exporting, and joining data
- Operations
  - Using R as a calculator
  - Object-oriented calculations
  - Boolean vectors, conditional statements, and subsetting
- Basics of Cleaning Text
  - Toolkit for Dates and Text (*lubridate* & *stringr)*

Basics of Data Management, Manipulation, Modeling, and Presentation

- Piping: a readable logic for data manipulation
- Grammar of Data manipulation (*dplyr*)
  - Selecting, filtering, summarizing
  - Creating/renaming/mutating/deleting Variables
  - Grouping: summarizing/counting/ordering
  - Dealing with missing values and reshaping data (*tidyr*)
- Grammar for Graphics (*ggplot2*)
  - *ggplot2* logic: additive coding and plots as objects
  - Grouping: aesthetic features and facets
  - Building layers and customizing themes
  - Managing Legends and Colors
  - Gridding ggplots and mapping
- Modeling Basics
  - T-test to Linear models: understanding how models can be stored as objects and explored like any other data structure.
- Presentation
  - Introduction to R Markdown and why it is useful
  - Rendering Data Notebooks: HTML, Word, and PDF (using a latex distribution)

  - Generating publishable-quality tables (*stargazer*)

Afternoon Session

Wrap up

- Finish any lingering topics from the morning session

Applied Examples

- Text Mining
  - Curated example of using text-as-data and drawing out latent topics.
- Networks
  - Curated example of visualizing and analyzing network data.
- GIS
  - Curated example of visualizing and mapping spatial data.
- Web-scraping
  - Curated example of how to scrape and structure web content into analyzable data.

Additional Learning Resources for *R*

The true power of *R* lies in its robust community of users. Almost any question one might have in *R* can be answered with a simple Google search. This crash course seeks to give each attendee the base knowledge to understand how to program and analyze data in *R*. However, there will be many issues one runs into as each data problem is unique (but not new). Moreover, becoming proficient at R is a process that takes time. Below I've collated some outside sources that offer further instruction in *R*.

Codeschool (https://www.codeschool.com/courses/try-r) offers an easy and free course for learning the basic functionality of *R*. The course is interactive and fun, leaving the user with hands-on knowledge of programing in R.

Datacamp (https://www.datacamp.com/courses) offers great tutorials for free and offers modules to learn specific tasks in R. They also offer a great introductory course in R (https://www.datacamp.com/courses/free-introduction-to-r)

The makers of **R Studio** offer a list of resources and cheatsheets for programing in R: https://www.rstudio.com/online-learning/

The **R project** also has some helpful tips, links, and manuals: https://www.r-project.org/help.html

UCLA's Institute for Digital Research and Education (http://www.ats.ucla.edu/stat/r/) has several R primers that can be accessed for free. They typically contain reproducible examples and code