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

Instructor: Eric Dunford
Email: edunford@umd.edu
Website: ericdunford.com
Dates:
- ▪ **Session 1: October 6 (9:30am-11:30am)**
- ▪ **Session 2: October 12 (1pm-3pm)**

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 all types of statistical analysis and data visualization, making it the leading choice in data analytics.

This short course will take place across two sessions. Each session will be approximately 2 hours in length and will offer an introduction to the *R* statistical programing language. All sessions will include slides and example code to accompany the materials being presented. Attendees need only bring a personal computer to follow along. No prior statistical or computer programming knowledge is needed to benefit from the course, but the assumption is that most users are familiar with statistics and STATA and are looking to expand their programming toolkit.

The course will offer attendees an underlying intuition of the R programming environment by focusing on (a) data management, and (b) graphics and presentation. Attendees of the workshop will leave with an applied understanding of the R environment: specifically, importing and manipulating data, rendering graphics, presenting results, and implementing a range of statistical models.

For many beginning users (especially those that come with a background in STATA), manipulating and managing data in R can be frustrating. This course will focus on providing attendees 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.

Required Materials

Attendees are encouraged to have a personal computer with them for each workshop to follow along with the examples in class.

Introduction to Statistical Programing in R

COURSE OVERVIEW

Session 1: Introduction to R: 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)*

Session 2: **Basics of Data Management, Manipulation, and Presentation**
- Review of basic concepts covered in Session 1
- 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
- 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*)
- Overview of Modeling Packages and Model Output in R

Introduction to Statistical Programing in R

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 short 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). 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