An Introduction To Statistical Programing In R:
A short course on processing, analyzing, and visualizing data in R[1]

Instructor: **Eric Dunford**
Email: edunford@umd.edu
Dates: April 3 – 7
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 analytic and consulting work. 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 5-day short course offers an introduction to the *R* statistical programing language. Each workshop will be approximately 2.5 hours in length. All workshops will include slides, PDF handouts, and example code and data to accompany the materials being presented. There will be hands-on examples and practice sets. Attendees need only bring a personal computer to follow along. No prior statistical or computer programming knowledge is needed to benefit from the course.

The goal of the course is to offer attendees an underlying intuition of the R programming environment by focusing on (a) data management, (b) basic statistical analysis, and (c) graphics and presentation. Attendees of the workshop will leave with an applied understanding of the R environment: specifically, importing and manipulating data, creating and deleting variables, rendering graphics and maps, managing and analyzing text, and common descriptive and statistical analytic methods.

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

Expectations
Each workshop will be accompanied by slides, documentation, and example code covering the day's materials. These documents should be reviewed and/or referenced if an attendee falls behind. In addition to these resources, practice sets will be distributed at the end of each workshop. Attendees are strongly encouraged to complete each practice set offline to help solidify the concepts of the day and to generate new issues/questions that might come up. Given the short duration of the course, it is expected that individuals who have trouble mastering the topics being covered should practice offline.

[1] I reserve the right to edit this syllabus.

Introduction to Statistical Programing in R

COURSE OVERVIEW

Day 1: Basics in R

- 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 *foreign, readstata13, readxl,* and *haven* packages for reading in different data types.
- Datasets
  - Importing and exporting basic data types.
  - Viewing data: print vs. GUI
  - Appending data.
  - Rdata

Materials: *Day One Handout*
Complete Survey of Topics of Interest (for day 5)

Day 2: Basics of Data Management and Manipulation

- Operations
  - Using R as a calculator
  - Object-oriented calculations
- Variables
  - Summarizing, tables, distributions
  - Creating and deleting variables
  - Dealing with missing values
- Data Management
  - Merging Data
    - Joining on variables, left and right joins, retention of all data structures
  - Introduction to the *dplyr* package and piping
  - *Dplyr* Methods: select, filter, arrange, group_by, summarize, mutate, and transmute
  - Applied examples for quickly summarizing and understanding data
- Cleaning Text
  - Introduction to the *stringr* package and regular expressions.
  - Cleaning text: punctuation, white space, and key phrase removal
  - Formatting dates and time
  - Splitting and joining strings

- Strings as variables

Materials: *Day Two Handout*

Day 3: Graphics and Maps

- Base Plots
  - histograms, scatterplots, barplots, and overlaying plots
  - Customizing plots
  - Exporting plots: JPEG, PDF, PNG, and vector formats
  - Gridding base plots.
- Introduction to the *ggplot2* package
  - Quick plots
  - *ggplot2* logic: additive coding and plots as objects
  - Grouping: colors and size
  - Building layers and Customizing themes
  - Scatterplot (with fits), histograms, barplot, boxplot, and violin plots.
  - Managing Legends and Color
  - Gridding ggplots
- Mapping
  - Introduction to Geo-Spatial Data: points, lines, polygons, and rastors
  - Maps using *ggplot2*
    - Color as description: managing color and Choropleth Maps
    - Matching data to maps
  - Maps using shapefiles
    - Loading shapefiles in R
    - Understanding shapefile objects
    - Plotting, assigning colors, and mapping geo-locations.
    - Overlaying data: example ethnicity and conflict in Nigeria
    - Names to points: using the Google API to recover geo-spatial coordinates

Materials: *Day Three Handout*

Day 4: **Analysis**

- Summary and non-parametric statistics
  - Five number summary, ranges, data summaries
  - Histograms using base plots
  - Correlations and Correlation Tests
  - Differences in means/proportions
  - Cross-tabulation
  - Scatterplots using base plots
- Regression
  - Linear and generalize linear regression in R
  - Understanding lm objects and extracting information
  - Fitting regression a line to a scatter plot
  - Fitting Loess curves to a scatter plot
  - Brief review of packages containing different regression methods
- Scaling

  - Summated Rating Model
  - Eigen Decomposition (to assess the dimensionality in the data)
  - Principal Component Analysis in R
  - Factor analysis
  - Reliability Scores
- Basics of Text Analysis
  - Introduction to the *tm* package
  - Word counts and Frequencies
  - Bag of words
  - Cosine similarity, topic models, and dimensionality
  - Natural Language Processing: installation and basic usage
- Presentation
  - Introduction to R Markdown and why it is useful
  - Rendering Data Notebooks: HTML, Word, and PDF (using a latex distribution)
  - Embedding code in reports
  - Generating HTML tables
  - Using HTML Slides to present results in R.

Materials: ***Day Four Handout***

Day 5: **Open**

- Catch-up/Review of topics covered over the last four days.
- Advanced Topics: based off a survey of topics attendees are interested in, the workshop will cover in greater depth areas of interest. Potential topics could include:
  - Web-scraping (for efficient recovery of online data)
  - Advanced Statistical Models/Manipulations
  - Loops and Functions (Designing scalable, reusable functions)
  - Building Packages (Have a set of functions/tasks that you do all the time? Why not build a package to streamline your workflow?)

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