# Data Processing, Analysis, and Visualization with R

*Daniel Anderson, University of Oregon - College of Education*

The purpose of this course is to provide a basic foundation in programming with R. The course has three main components, as the title suggests: data processing/munging/wrangling, basic analysis, and visualization. These three components are routine in applied data analysis work. The focus will be on working with and programming in R, as opposed to the specifics of any given dataset/analysis. The course begins with a basic overview of the R syntax, data structures, overview of object-oriented programming (OOP), and the use and writing of functions, which are the heart of R. We apply these concepts to process "messy" data into workable formats, and analyze the data. We use data visualization to explore both raw data and the results of fitted models. We will also go over the basics of writing an R package, specifically personal R packages. We will be doing some very basic work with *git*, which is a version control software that can help when collaborating on code. You will also become familiar with *R Markdown*, which is pretty straightforward and is a great way to share snippets of code, put together a tutorial, or share the results of an analysis with practitioners. I will be asking for assignments to be completed with R Markdown.

This class was constructed with the philosophy that the only way to truly learn R and become proficient with it, is to dive in and practice, practice, practice. Class sessions will be highly interactive. During class, I will be asking you to write code with me concurrently, independently other times, and in small groups at other times. Weekly homeworks will also be assigned, which are intended to be brief but get you more practice. These can be completed independently or in small groups, and will be scored on a completion basis. The term project requires you complete the three main components of the course described above with a "real" dataset. This can, again, be completed independently or in small groups. I encourage you to find data that you are actually interested in using. If you do not have access to data, please contact me as soon as possible.

## Required Materials

Prior to the first class you should have the latest version of R downloaded and installed on your computer, available through the comprehensive R archive network (CRAN): https://cran.r-project.org. We will be working with various packages as well, but these can be installed as we work through the material. All readings will be provided as PDFs or are freely available online. Please also download and install git: https://git-scm.com/downloads. I also encourage you to create a *github* repository (https://github.com) for all your work in this class.

**Recommended Materials:** Prior to the first class, it is highly recommended that you spend some time exploring different text editors and/or integrated development enviornments (IDEs) for R. I use the text editor Sublime Text (http://www.sublimetext.com) which has some really nice features and, to me, is very aesthetically pleasing. By far the most popular IDE is RStudio, which is being actively developed and has numerous plugin-like features that make many tasks simpler. Most of this is just personal preference. The built in editor for Mac is also pretty nice, but on the Windows side is about as basic as can be. Getting a text editor/environment that you feel comfortable with can really ease the process.

**Resources:** One of the great things about R is that it is open-source. Like many open-source platforms, there is a tremendous community behind it. Online forums can be a great place to find answers to specific problems (e.g., http://stackoverflow.com/questions/tagged/r). There are also lots of sites that give basic tutorials (e.g., http://www.statmethods.net/interface/help.html). UCLA has also listed a set of resources that may be helpful (http://www.ats.ucla.edu/stat/r/). In this course, we will primarily rely on the following resources

- Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.
- Kabacoff, R, I. (2011). *R in Action: Data analysis and graphics with R*. Shelter Island, NY: Manning Publications.

- Peng, R. D. (2015). *R Programming for Data Science*. Victoria, BC: Lean Publishing.
- Wickham, H. (2015). *Advanced R*. Available online at http://adv-r.had.co.nz
- Wickham, H. (2015). *ggplot2*. New York, NY: Springer.

## Course Website

All the code for this course, as well as lectures and course readings, will be available in a github repository (https://github.com/DJAnderson07/CourseR). This is where the basic work with git will come in, as I'll be asking you to clone the repository onto your local machine and update it periodically as I push new content into the repository.

## Course Objectives

Upon completion of this course, the successful student will:

1. Understand the foundations of the R environment (i.e., OOP, functions, data structures, etc.).
2. Understand and be able to work efficiently with different data types (i.e., string, numeric, integer, factor).
3. Be able to read various data into R from various sources (e.g., .csv, .sav, .txt, etc.) and prep the data for analysis.
4. Understand the basic structure of plotting with the base graphics and the *ggplot2* package.
5. Be able to conduct basic analyses (e.g., simple and multiple linear regression, basic varying intercept models, etc.) with R, and explore the fitted model through visualizations.

## Course Schedule

### Week 1: Getting Started Introduction and overview of the R environment

- Objectives
  - Clone course github repository
  - Get a text editor or IDE for R
- Topics
  - Objects in R
  - Vectors and Matrices (brief intro)
  - Reading data into R
    - Setting and changing your directory
  - Subsetting
    - vectors
    - matrices
    - data frames
    - conditional subsetting
    - subsetting other elements with vectors
  - R packages
  - Getting help
  - The grammar of syntax
  - Introduction to *R Markdown* & *git*
- Lab

  - Reading data into R and subsetting
- Readings
  - R Markdown
    - http://RMarkdown.rstudio.com
    - http://RMarkdown.rstudio.com/authoring_basics.html
    - http://RMarkdown.rstudio.com/authoring_rcodechunks.html
  - Syntax style: http://adv-r.had.co.nz/Style.html
  - Kabacoff: Chapter 1
- Homework
  - Creating an *R Markdown* document

### **Week 2: Data structures and data types (part 1)** Homework 1 Due, Homework 2 Assigned

- Topics
  - Data frames, Matrices, Arrays, Vectors, Scalars
  - Matrix algebra versus element-wise algebra
  - Numeric (double), integer, character, factor, logical
- Lab
  - Coercions
- Readings
  - Kabacoff: Chapter 2

### **Week 3: Data structures (part 2) & routine functions** Homework 2 Due, Homework 3 Assigned

- Topics
  - Lists
  - Data frames as a special type of list
  - Recycling
  - routine functions
    - `rep()`, `seq()`, `:`
    - `table()`
    - `c()`, `cbind()`, `rbind()`
    - `ifelse()`
    - `rnorm()`, `rbinom()`, `dnorm()`, etc.
    - `length()`, `nrow()`, `ncol()`
    - `str()`, `head()`, `tail()`
    - `summary()`
    - `with()`, `attach()`
    - etc.
- Lab
  - Working with lists
- Readings
  - Wickham (http://adv-r.had.co.nz/Data-structures.html)

**Week 4: Loops** Homework 3 Due, Homework 4 Assigned

- Topics
  - For loops
  - Apply family of loops
- Lab
  - Batch processes with lists and loops
- Readings
  - Peng: Control Structures

**Week 5: Functions** Final Project Plan Due Homework 4 Due, Homework 5 Assigned

- Topics
  - Function components
  - Basics on writing packages
- Lab
  - Writing custom functions
- Readings
  - Wickham (http://adv-r.had.co.nz/Functions.html)
  - Peng: Functions
  - Parker: Writing R Packages (http://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/)

**Week 6: Basic analysis** Homework 5 Due, Homework 6 Assigned

- Topics
  - Correlation
    - plotting relations (brief)
  - Linear regression and the `lm()` function
    - adding the regression line to a plot
    - evaluating assumptions with plots
    - predicting new data
    - extracting coefficients
- Lab
  - Fitting and visualizing regression models
- Reading
  - Gelman & Hill: Chapter 3

**Week 7: Transforming & manipulating data** Homework 6 Due, Homework 7 Assigned

- Topics
  - Recoding variables
  - String functions
    - `grep()`, `substr()`, etc
  - Merging
    - Merging as a tool for producing new variables
  - Aggregation
  - Reshaping data
    - Long to wide and wide to long
- Lab
  - Data restrucuring

**Week 8: Plotting with Base Graphics** Homework 7 Due, Homework 8 Assigned

- Topics
  - Scatter plots
  - Controls:
    - line width, color, and type
    - point size, color, and type
    - x and y axis labels
    - titles and subtitles
  - Scatter plot matrices
  - Boxplots and extensions (violin plots and bean plots via alternate packages)
  - Histograms, density plots
  - Low-level functions
  - points, lines, text, axis
  - saving plots
  - Graphical Parameters: `par()`
    - `mfrow`, `mfcol`, `mar`, `oma`
- Lab
  - Overlaying plots
- Readings
  - http://rpubs.com/SusanEJohnston/7953
  - http://www.ling.upenn.edu/~joseff/rstudy/week4.html

**Week 9: Plotting with *ggplot2*** Homework 8 Due, Homework 9 Assigned

- Topics
  - `qplot()` (similar to base)
  - layering with `ggplot()`
  - faceting
- Lab
  - Cluster-level visualizations
- Readings
  - Wickham (ggplot book): Chapter 2 & 3 (especially 3)

**Week 10: Basic multilevel models and plotting** Homework 9 Due

- Topics
  - lme4 and the `lmer()` function
  - Varying intercepts
  - Varying slopes
  - Varying intercepts and slopes
  - Covariance matrices
  - Exploring multilevel models through plotting
- Lab
  - Fitting, visualizing, and interpreting varying intercept models
- Readings
  - Gelman & Hill: Chapter 12

**Week 11: Finals week**

- Final Project Due

## Homeworks

Homeworks are worth 10 points each (90 points total), and are graded on a completion basis. As long as you address each part of each homework, you will get full credit (regardless of correctness). Below is a brief overview of each homework. More details will be given at the start of each week (which will then be due one week later).

1. Creating an R Markdown document
   - The purpose of this homework is to get you some basic familiarity with *R Markdown*. You'll be required to produce a document with some of the basic features.
2. Working with data frames, matrices, arrays, vectors, and scalars
   - In this homework, you'll be transforming data into different structures and calculating basic descriptive statistics to understand the utility of the different structures. You will also be working with various data types.
3. Lists
   - The focus of this homework is on lists, which are the most flexible data structure in R. You will be storing different data types in lists, subset lists, etc.
4. Looping
   - This homework will focus on using various types of loops (specifically, for loops and the apply family of loops) to various data structures.
5. Functions
   - This homework will focus on writing custom functions to perform specific tasks. You will also be asked to compile these functions into a personal package.
6. Linear regression
   - The focus of this homework will be on the `lm()` function. We will be using stock data from R, but if you would prefer to use your own data, please let me know.

7. Transforming data
    - This homework will essentially be a micro version of the final project, but without focusing too much on plotting or analysis. In other words, you will only be focusing on the data processing portion, but you will move a “messy”, “real world” dataset from its raw form to a format appropriate for analysis and/or exploratory plotting.
8. Plotting 1
    - You will be asked to pull together many of your skills learned earlier in the course to produce publication quality plots with the base graphics.
9. Plotting 2
    - Homework 9 will be essentially equivalent to Homework 8, but with using the *ggplot2* package. The homework will also be more heavily focused on exploratory plots, rather than publication level plots.

## Final Project (100 points)

For the final project in this class, you will need to complete the three major components of the class with a “real world” dataset. If you do not have access to data, please get in contact with the instructor as soon as possible. More details will be provided later in the term, but the project will consist of the following steps: (a) processing the raw data to a format appropriate for analysis and exploratory plotting, (b) producing at least three exploratory plots of the analytic sample, (c) analyzing the data, and (d) plotting at least one aspect of the fitted model. Your plan must be approved by the instructor. The final project is due week 11 of the term (finals week), while a full description of your plan is due week 5.

The final project must be produced in accordance with the principles of reproducible research. In other words, You will need to use *R Markdown* (or something similar, e.g., LaTeX). The final paper should not include any code, but it should be accessible and reproducible. I encourage you to use *git* to make this all accesible and transparent. All of this should become more clear as the term progresses.

## Grading

### Points breakdown (300 possible)

- Participation: 60 points (20%)
    - 10 labs @ 5 points each
    - 10 points for general involvment (asking questions, coming prepared, etc.)
- Homeworks: 9 @ 10 points each = 90 points (30%)
- Final Project outline: 50 points (17%)
- Final Project: 100 points (33%)

**Letter grade breakdown**

- Below is a breakdown of letter grades, based on the number of points received.

| Lower point range | Grade | Upper point range |
|---|---|---|
| ≥ 93% (279 points) | **A** | |
| ≥ 90% (270 points) | **A-** | < 93% (279 points) |
| ≥ 87% (261 points) | **B+** | < 90% (270 points) |
| ≥ 83% (249 points) | **B** | < 87% (261 points) |
| ≥ 80% (240 points) | **B-** | < 83% (249 points) |
| ≥ 77% (231 points) | **C+** | < 80% (240 points) |
| ≥ 73% (219 points) | **C** | < 77% (231 points) |
| ≥ 70% (210 points) | **C-** | < 73% (219 points) |
| < 70% (210 points) | **F** | |