**PSYC GR6008 – R Programming for Behavioral Scientists**

**Instructor – Prof James P Curley – jc3181@columbia.edu**

I. Bulletin Description
II. The rationale for giving the course
III. A full description of the content of the course
IV. The reading list
V. Course grading

**I. Bulletin description**

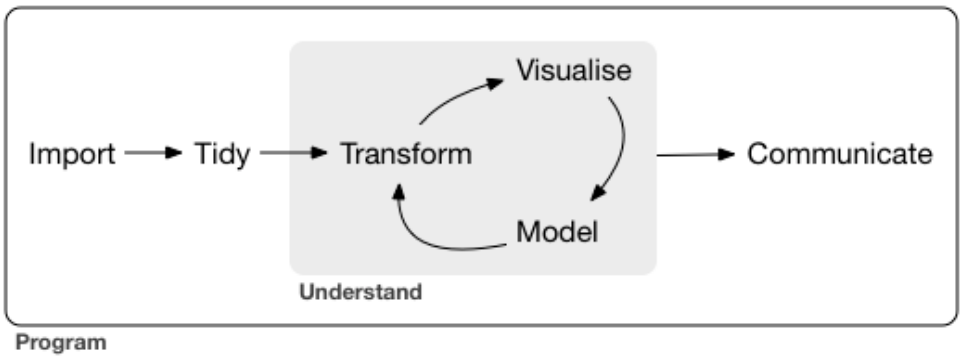
GR6008. R Programming for Behavioral Scientists
4 pts. Fall 2017. W 12:10 – 2 PM. Room 200C Schermerhorn Hall

Prerequisites: Open to Psychology Ph.D. students. Instructor permission required for students in other programs.

This seminar will provide students with R programming skills to be able to organize, analyze and visualize data. The key goal to be achieved will be to train students to think how to make their research collaborative and transparent with code such that all data analysis that they undertake will be fully reproducible. Using R and RStudio, students will learn the full process of designing and undertaking reproducible behavioral data analysis.

**II. The rationale for giving the course**

Learning how to effectively and accurately work with data is a fundamental facet of psychological and neuroscientific research. Beyond this, individuals in fields as diverse as journalism, business, linguistics, economics, epidemiology, and many others are utilizing data science methods and practices. This course aims to provide critical quantitative and programming skills not just to students intending to pursue academic research but also working with data in any field in the future. Many traditional research courses focus on the design of experiments or how to perform specific types of statistical analyses. However, there is a need for students to fully appreciate how to efficiently and accurately work with data from its initial collection to its final communication via data cleaning, visualization, analysis, and modeling. In this course, students will learn how to think both quantitatively and programmatically in how to manage, analyze and communicate data. This course provides the fundamentals of programming in the R programming language – the *de facto lingua franca* of modern data science. This course will follow the pedagogical approach for data science laid out by Grolemund & Wickham (R for Data Science, O'Reilly 2016 – see figure). Throughout students will have opportunity to use their own data or example datasets – to explore them, to generate questions and hypotheses, to work out how to carry these ideas out programmatically and how to effectively communicate quantitative information. Further, students will learn how to work collaboratively and produce reproducible output. Prof Curley has many years of experience of working with the R programming language and is an official RStudio instructor.




*R programming Pedagogical Framework*
*R for Data Science, Garrett Grolemund & Hadley Wickham, O'Reilly, 2016.*

**III. A full description of the content of the course**

Week 1. Introduction to R – Data Structures & Syntax

Aim: The first week will introduce the R programming language. Students will become familiar with basic programming concepts such as data structures, functions, and objects. Students will learn basic R syntax and become familiar with commonly used base R functions.

Before each class: Students will watch video tutorials (4 x 15 minutes each) by the instructor which will outline key concepts and provide worked examples of each topic to be covered. Students will also watch a video of the instructor outlining a mini-challenge assignment that students will complete and submit prior to the class. Prior to the first class students will also take a background skills and knowledge questionnaire, as well as a values questionnaire where students will describe their thoughts related to how important they value particular concepts such as reproducibility and data visualization.

In class: The instructor will lead a discussion about worked examples and mini-challenge from videos and troubleshoot questions. Students will complete more inline quizzes and debugging exercises to demonstrate their familiarity and understanding of concepts.

Reading: Chapter 3-6 of R. Cotton, *Learning R: A Step-by-Step Function Guide to Data Analysis,* O'Reilly, 2013.

Week 2. Data Carpentry

Aim: Students will learn how to import and export data from RStudio. Students will learn how to summarize data and how to reorganize and work flexibly with raw data.

Before each class: Students will watch video tutorials (4 x 15 minutes each) by the instructor which will outline key concepts and provide worked examples of each topic to be covered. Students will also watch a video of the instructor outlining a mini-challenge assignment that students will complete and submit prior to the class.

In class: The instructor will lead a discussion about worked examples and mini-challenge from videos and troubleshoot questions. Students will complete more inline quizzes and debugging exercises to demonstrate their familiarity and understanding of concepts.

Reading: Chapter 5, 11 & 12 of G. Grolemund & H.Wickham, *R for Data Science*, O'Reilly, 2016.

Weeks 3-4. Data Visualization

Aim: Students will learn how to visualize and explore data using the ggplot2 package. Students will become familiar with how to choose the most appropriate data visualization for different data types by adjust aesthetics and chart types. Students will learn Leland Wilkinson's grammar of graphics approach to data visualization.

Before each class: Students will watch video tutorials (4 x 15 minutes each) by the instructor which will outline key concepts and provide worked examples of each topic to be covered. Students will also watch a video of the instructor outlining a makeover challenge assignment that students will complete and submit prior to the class. This challenge will consist of a poor visualization of data that the students will be asked to improve or come up with different ways in which the data could be visually represented.

In class: The instructor will lead a discussion about worked examples from videos and troubleshoot questions. Students will also discuss the makeover challenges and how why they chose their particular

visualization. Students will complete more inline quizzes and debugging exercises to demonstrate their familiarity and understanding of concepts.

Reading:
Chapter 3 of G. Grolemund & H.Wickham, *R for Data Science*, O'Reilly, 2016.
Zev Ross blog post - http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2- cheatsheet-3/
W. Chang, *R Graphics Cookbook – Practical Recipes for Visualizing Data*, O'Reilly, 2012

Week 5. Self-learning exercise

Aim: An important aspect of learning to code is learning how to independently acquire new skills and to troubleshoot problems. Students can independently learn skills from blog posts, video tutorials, Q&A sites such as StackOverflow, package vignettes etc. Students will be set the goal of learning an R package previously unfamiliar to themselves and presenting some analyses and visualizations based on this package to the class. Students will also develop skills in orally communicating their newly acquired skills to other students.

Before each class: Students will select the R package either from a list provided by the instructor, or with the instructor's approval. Students are expected to complete their analysis and visualizations and upload them to GitHub prior to class.

In class: Each student will present their analysis and visualization to the class, describing what they learnt, how they found the information to learn the package, what problems they encountered and how they overcame them.

Week 6-7. Data Challenge Group Project

Aim: Students will work in groups of 2-3 and learn how to work collaboratively on the goals of a group project. Students will be given access to large raw datasets such as the NYC Open Data datasets (e.g. annual taxi rider trips), NBA or MLS player tracking data for one season, US city daily weather data. Students may also use data of their own choosing. Students will be assessed on their ability to take raw data, to tidy the data and put it into workable formats, to explore and determine potentially interesting research questions and to analyze and communicate their findings.

Before each class: Prior to the first class, students will be assigned to their teams and choose their raw datasets. Students should also submit some initial research questions related to their datasets. Prior to the second class students will continue to work on their projects and submit their final analyses and visualizations to GitHub.

In class: In the first class students will discuss their ideas and how they will proceed with working with the data with the instructor. The majority of class time will be spent working on their data projects. In the second class each team will discuss their research findings with an oral presentation to the instructor and other students.

Weeks 8-9. Communicating Data Interactively

Aim: Students will learn how to write code for visualizing and analyzing data that can be shared with other collaborators or users. Students will learn how to use RMarkdown and RNotebooks for producing standalone PDF or web-hosted data analysis and visualizations and how to use these methods for demonstrating their workflow to other users. Students will also learn the basic elements of the shiny R packages for generating interactive documents and applications, so other users can interact with their data.

Before each class: Students will watch video tutorials (4 x 15 minutes each) by the instructor which will outline key concepts and provide worked examples of each topic to be covered. Students will also watch a video of the instructor outlining a mini-challenge assignment that students will complete and submit prior to the class.

In class: The instructor will lead a discussion about worked examples and mini-challenge from videos and troubleshoot questions. Students will complete more inline quizzes and debugging exercises to demonstrate their familiarity and understanding of concepts.

Reading:
Chapters 27,29,30 of G. Grolemund & H.Wickham, *R for Data Science*, O'Reilly, 2016.
Zev Ross blog post - http://zevross.com/blog/2016/04/19/r-powered-web-applications-with-shiny-a-tutorial-and-cheat-sheet-with-40-example-apps/

Weeks 10-11. Reproducibility Exercises

Aim: Students will become familiar with the critical concept of open and transparent data analysis. They will understand why their code needs to be reproducible and that this is for the benefit of themselves, their co-workers or collaborators, reviewers and audience.

Before each class: Before the first class, students will retake the values questionnaire that they took at the beginning of the course. Students will also choose a recently published academic paper or piece of research from the internet that they will attempt to reproduce the analysis for. Students can choose from a list produced by the instructor or a piece of research of their own choice with the instructor's approval. Before the second class, students will attempt to reproduce their chosen paper's analyses and visualizations.

In class: The instructor will lead a discussion about the students current views about reproducibility in data science and academia. Students will present whether it was possible or not to reproduce the published findings in academic research papers and the implications of this exercise.

Reading:

Munafo MR et al., 2017, A manifesto for reproducible science, *Nature Human Behaviour* 1: 0021. doi:10.1038/s41562-016-0021

Markowetz F, 2015, Five selfish reasons to work reproducibly, *Genome Biology* 16:274. doi: 10.1186/s13059-015-0850-7

Week 12-13

R Package Development

Aim: Students will continue to develop best standard practices for data analysis and reproducible research. Students will learn how to develop simple R packages and how to submit their final package to GitHub and/or CRAN.

Before each class: Before the first class, students will watch video tutorials (4 x 15 minutes each) by the instructor which will outline key concepts and provide worked examples of each topic to be covered. Students will be assigned to pairs and will develop their idea for a package. Before the second class, students will submit their final package to GitHub and/or CRAN.

In class: The instructor will lead a discussion about the worked examples from the video tutorials. Students will discuss their package development with other students and troubleshoot with the instructor. During the second class students will orally present their final packages to other students.

Reading: Chapter 3-6 of R. Cotton, *Learning R: A Step-by-Step Function Guide to Data Analysis,* O'Reilly, 2013.

Hilary Parker blog - https://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/
Karl Broman blog - http://kbroman.org/pkg_primer/
H.Wickham, *R Packages*, O'Reilly, 2015.

Final Data Project

Aim: Students will produce a large final data analysis project based upon their own research or on publicly available datasets. Students are expected to produce a fully reproducible data project from raw data to final analyses and visualizations. Students will learn to communicate both orally and in writing the results of their data analyses to a target audience.

Before each class: Before the first class, students will develop their idea for their final data project and generate preliminary ideas about potential interesting research questions. Before the second class, students will work on their projects and submit the final project to the instructor as well as publishing online at Rpubs.

In class: In the first class the instructor will work with students and troubleshoot issues they have with progressing with their projects as well as guiding direction. During the second class students will orally present their research. Students can also choose to write up their research as a data journalism article to submit to fivethirtyeight.com, slate.com or wired.com.

**IV. Reading list**

Readings will be available as ebooks or PDFS on http://courseworks.columbia.edu. Specific readings are given with each week above.

The following books are general reference textbooks that will be made available to all students electronically:

G. Grolemund & H.Wickham, *R for Data Science*, O'Reilly, 2016.
H.Wickham, *R Packages*, O'Reilly, 2015.
H.Wickham, *Advanced R*, Chapman and Hall/CRC, 2014.
P.Dalgaard, *Introductory Statistics with R*, Chapman and Hall/CRC, 2008.
R. Cotton, *Learning R: A Step-by-Step Function Guide to Data Analysis,* O'Reilly, 2013.
W. Chang, *R Graphics Cookbook – Practical Recipes for Visualizing Data*, O'Reilly, 2012

**V. Course grading**

Grading will be as follows:

Discussion / participation 20%
Oral Presentations 20 %
Mini challenges 10%
Makeover challenges 10%
R package self-learning challenge 10%
Group Project challenge 10%
Reproducibility Challenge 10%
R Package development challenge 10%
Final Data Challenge 10%