**Expository Data Analysis with R**

**ECOG 314 / ECON-181**

**Introduction to Data Exploration and Analysis with R**

**About this Course**

Conducting economics research requires both an understanding of theoretical concepts, and practical knowledge of empirical methods. Today most empirical work in conducted using statistical programming languages such as Stata and SAS In this course, students will learn how to use one such language, R, as a means of building their empirical toolkit. R has become one of the leading languages in data science and statistics and therefore a very valuable programming language to learn. R is a free software environment used by a variety of fields including economics.

This course presents the fundamental of data analysis using R, with emphasis on economics applications. This course is an opportunity for students to apply many of the concepts covered in earlier econometric courses. Through the course you will improve your financial literacy and general problem solving skills through lectures, homework, and projects. By the end of the semester your will be able to use your knowledge of R to clean a data set, combine data from different sources, uncover patterns and insights in data, make predictions using statistical methods and communicate your finding through the use of tables, plots, and clear and concise prose.

**Learning Objectives**

After completing this course, students will be able to use R as a data analysis tool to:

* Create, read, modify and store R datasets
* Use available R packages and write custom functions
* Create figures and plots
* Perform efficient dataset manipulation
* Perform and interpret multiple linear regression
* Create, manage and share reproducible project files using R markdown packages

**Course Structure**

* **Lectures** **–** Lectures will be held Fridays from 9:15 am to 12:15 p.m. Class time will include lectures and labs with Federal Reserve Board staff. **Class will meet in the Federal Reverse Board’s building at 1801 K-Street, NW. Washington, DC.** (*Metro reimbursement to and from Shaw/Howard stop to Farragut North to students, will be paid in a lump sum in May*).
* **Office Hours –** TAs will hold office once a week in the Howard University Economics department for two hours. In the weeks leading up to major deadlines we will have more than one session. The first session is Tuesday 1/23/2018.
* **Github Site –** All the lectures and homework will be posted to the [github](https://github.com/wampeh1/Ecog314_Spring2017) site.
* **Piazza Site –** We will use a [piazza](https://piazza.com/federal_reserve_boardhoward_univeristy/spring2017/ecog314econ181/home) site as our course wiki. This is a great format for asking questions where the TAs and instructors can respond collectively to make sure your questions are answered quickly. We encourage you to use this site to ask questions throughout the course. You will be asked to create an account and contribute to the site as part of our first homework.

William Ampeh, a Lead Technology Analyst at the Federal Reserve Board, developed the course content with a team of other Federal Reserve staff, and will lead the class sessions. Andrew Cohen, an Assistant Director at the Federal Reserve Board and Visiting Professor in the Economics Department will coordinate logistics for the course.

**Course Prerequisite**

All applicants must have completed a college level course in Econometrics with a grade of **B** **or higher**. No prior training in programming or data science is required.

**Computer**

A Windows or Mac laptop is required with the following minimum configuration: 4 GB RAM or higher; 320 GB hard disk; configured to allow the installation of R and RStudio software. A limited number of loaner laptops will be made available if needed, **for in-class use only**.

**Software**

R and selected R packages will be the primary software for this class. R is free. Substantial instruction will be provided in lecture notes and assignments, and additional instructions will also be available in the online reference materials.

Link to Download R: [Comprehensive R Archive](http://cran.us.r-project.org/)

RStudio is the recommended R integrated development environment

RStudio Download: See https://www.rstudio.com/products/rstudio/download/preview/

RStudio is easy to install and the installation does not require any instruction. However, the following links provide additional setup and navigation guidance:

<http://web.cs.ucla.edu/~gulzar/rstudio/index.html>

<http://dss.princeton.edu/training/RStudio101.pdf>

<https://support.rstudio.com/hc/en-us/sections/200107586-Using-RStudio>

**Textbook:**

**R 4 Data Science: r4ds.had.co.nz**

The R4DS textbook is freely available online and provides an excellent practical introduction to using R for data science. The book is written by Hadly Wickham, the author of many packages used in the course and provides insight into how to most effectively utilize R for data analysis.

**Grading**

Numerical class grades will be based on:

* Homework (35%)
  + Assignments will be due by midnight the Wednesday after the class they are assigned.
  + Late assignments will lose 10% of the total for each day late.
* Participation (10%)
  + Attendance – (no points given if late to class)
  + Good class citizenship (helping other students, contributing to the piazza site)
* Take-home Exam (20%)
  + This will be a take home exam given around the midpoint of the semester
  + This will be an individual assignment where you will not be able to ask questions of the TAs.
  + There will be no office hours while the mid-term is live.
* Final project check in and the final project (35%).
  + Separate syllabus to follow

The instructor reserves the right to amend weighting.

**Topics**

1. **Introduction to R and Rstudio**
2. **Creating communicative graphics with ggplot2**
3. **Data transformation with dplyr**
4. **Data cleaning – strings and dates**
5. **Control flow and function writing**
6. **Creating a replicable workflow with Rmarkdown**
7. **Regression**
8. **Advanced topics**