STAT 209: DATA COMPUTING AND VISUALIZATION

SPRING 2018

Instructor. Colin Reimer Dawson (*he/him/his*)
Office. King 204
Email. `cdawson@oberlin.edu`
Website. `http://colindawson.net/stat209/`
Office Hours. M 3:30-5:00, W 9:15-10:45, F 3:30-5:20. ***Important: If your schedule conflicts will* all *of these times, let me know ASAP, so I can rearrange some things!***
Times and Locations. MWF 2:30-3:20, King 241, except that *we will meet in Peters 233 on Friday, 4/27. Mark your calendar!*

Course Overview

The main goal of this course is to equip you to create informative visualizations of datasets of various shapes, sizes, and sources. In realistic settings, before you can translate data into concise graphical representations, you need to be able to get it into a format that lends itself to visualization. The process of taking "wild" real-world data and "taming" it, which can involve pulling data from the web, merging data from multiple sources, cleaning data by removing or restructuring information, and creating new variables out of existing ones, is referred to as "data wrangling".

In the first half of the semester, we will discuss general principles of visualization, such as the "grammar" of data graphics, as well as basic data wrangling, both as practiced using professional-grade computing tools in the form of the R statistical language and the RStudio computing environment. Throughout, we will also develop some data science best practices and habits, such as documentation, reproducibility and version control.

In the second half, we will touch on some more advanced data science techniques that are helpful for wrangling and visualization, such as streamlining dealing with large datasets using database queries, basic machine learning (especially clustering and dimensionality reduction), and deal with some more specialized methods useful for particular kinds of data, such as text and geographic data.

Date: Last Revised February 4, 2018.

Learning Goals. After completing this course, students should

- Be able to recognize what statistical elements (quantities, categories, descriptive summary values) are captured by various graphical elements (such as color, size and position) in data visualizations encountered in the media or scientific publications
- Be able to create concise, informative, and accurate data visualizations by translating statistical elements to graphical elements, with the aid of standard statistical software
- Be capable of preparing "wild" data for visualization and analysis by merging, cleaning, filtering, recoding, and manipulating datasets, with the aid of standard statistical software
- Be able to produce transparent, and reproducible reports that integrate code, output, and verbal explanation and interpretation in a smooth and aesthetically pleasing way
- Develop good data science work habits, such as version control and documentation of code

Who This Course is For. Although this course is numbered at the 200 level, it has no prerequisites, and is intended to be accessible to anyone. The main requirement is that you have a willingness to write some code (but you need not have prior experience doing so!). Since the course is project-centered, ideally you should have some idea about what kinds of questions you'd like to use data visualization to help address. If you don't have any project ideas coming in, that's fine, but you should start thinking about some as soon as possible!

Materials

Textbook. The textbook is *Modern Data Science with R*, by Baumer, Kaplan and Horton. The paperback edition is about $70; there is an eBook only version available for slightly less (or hardcover for more).

Laptops. We will devote more than half of class time to working with data in groups. There is not a dedicated lab day for the class, so you should bring your laptop, if you have one, to the classroom. If you don't have a laptop, there will be some available to borrow for the semester through OCTET (`http://octet.oberlin.edu`).

Software. We will be using the R computer language throughout the course, with the aid of the RStudio environment. Oberlin has an RStudio server that allows you to access RStudio from your browser (which you may have done if you took another 100 or 200-level STAT course), which you can access at `http://rstudio.oberlin.edu`. If you do not already, you will have an account set up with your Obie ID as your user ID. It is also possible to download and install your own local copies of R and RStudio, though this will require you to manage a few more things on your own. To do this, go first to `cran.r-project.org` to download R, then to `www.rstudio.com` to download RStudio.

Miscellany

Communication With Me. Email is the best way to reach me outside of a face-to-face meeting. **It will be helpful if you include `[stat209]` at the start of your subject line**. You are welcome to address me by my first name, which is generally what I will use when signing emails. You should allow about 24 hours for me to respond to most things; it will often be less, but if you need a response by a particular time, be sure to ask me the day before. I usually do not respond to email on Saturdays.

Accommodations. If you have a disability of any sort that may require accommodations in order for you to do your best work in this class, please let me know as early as possible, and consult as well with the Office of Disability Services (ODS). By college policy, **all requests for accommodation require documentation from ODS.**

Honor Code. The Oberlin College Honor Code formalizes the idea that all work that you submit is your own and that you have given credit to the ideas and work of others when you incorporate them. You will be asked to write and sign the honor pledge on each graded assignment that you hand in. The honor pledge reads: "I have adhered to the Honor Code in this assignment."

What it means to adhere to the honor code depends on context. I describe what it means to follow the honor code on that assignment below.

More information about the honor code can be found on the web at the Dean of Students site:
`http://new.oberlin.edu/office/dean-of-students/honor/students.dot`

Assignments and Activities

Homework (15% of grade). I will periodically assign homework for you to work on outside of class, about every two weeks on average (more at the beginning).

Homework sets will involve a mix of computation and discussion of concepts. Most homework should be prepared as an RMarkdown document, and necessary files submitted to Blackboard (first two weeks) or via GitHub to the class repository (after we learn this). If these words don't make sense, don't worry! They will soon. You can miss one assignment before your grade is affected.

Late Policy: I will accept homework / revisions of previous submissions up until the point when I have a chance to sit down and start grading it, whenever that happens to be. Once I have downloaded the files associated with a particular homework, no additional revisions may be submitted for that assignment. However, this may sometimes be later than is ideal for your learning: you should make every effort to turn homework in by the posted deadline, so you are in the best position to build on the material as needed.

Labs (15% of grade). We will work on guided explorations ("labs") in class periodically. You'll work on these in pairs, with the goal that you can get at least close to finishing them during class. There will be a few questions interspersed throughout each lab for you to answer. Responses are due on paper or GitHub by the start of the next class period after we do a lab. These are graded for completion only, and no late work will be accepted. You can miss one before your grade is affected.

Projects (60% of grade). The main focus of the class is a set of four progressively more involved visualization projects, to be done in groups of 2 or 3. These are fairly open-ended, to allow you to explore questions of personal interest, though I may provide some suggested datasets that you can work with if you don't have something of your own that you want to do. The projects four projects each involve "illustrated writeups", by which I mean one or more data visualizations presented in the form of a short article discussing what they show (FiveThirtyEight or Vox posts are a good model for this). In order, the projects require you to

(1) **Project 1 (15%)** use a dataset that is fairly "clean" to start with, so that a minimum of "wrangling" is required.
(2) **Project 2 (15%)** merge multiple datasets with some non-trivial wrangling required to clean the data and combine information
(3) **Project 3 (15%)** employ a database to manage a dataset that is too large to hold comfortably in memory at one time
(4) **Project 4 (15%)** combine multiple elements from previous projects and/or explore a geographic or text dataset

Participation and Engagement (10% of grade). The remainder of the grade is based on consistent engagement during class and active participation in group activities.

Exams. There are no exams. **In lieu of a final exam, we will meet during finals week at the time of the scheduled final to present project 4 to the class.**

Honor Code. I encourage you to work on homework and labs collaboratively; however, **you must write your own solutions and code; you may not copy each other's words or commands**, and you should **indicate in your writeup what other students you worked with**. For group projects, you will of course collaborate with your group, and should turn in one set of files (via GitHub). However, **individual contributions should be recorded** in the form of individual commits to the submitted repository.

Key Dates

There are no collegewide holidays during the spring semester, so the only time when class will not be held is spring break. Tentative project due dates are (1) Friday, March 2, (2) Monday, April 7, (3) Monday, April 30, and (4) the day of the final presentations.

Approximate Topic Outline

See the "Schedule" tab on the course website