S&DS 425: Statistical Case Studies

Syllabus, Spring 2018

Instructor: Xiaofei (Susan) Wang
E-mail: xiaofei.wang@yale.edu
Office: 24 Hillhouse Ave. Room 206
Meeting Time/Place: Tu/Th 9-10:15am, location TBD

1 Course Description

From a course description of the graduate version of Statistical Case Studies (S&DS 625):

> Statistical analysis of a variety of statistical problems using real data. Emphasis on methods of choosing data, acquiring data, assessing data quality, and the issues posed by extremely large data sets. Extensive computations using R.

The course involves the statistical analysis of a variety of problems which vary each time the course is taught. In previous years, we have studied nationalistic biases in the judging of Olympic diving; a job hiring discrimination lawsuit; the prediction of college basketball games; the prediction of the 2016 presidential election; and local real estate prices and patterns. We will emphasize methods of choosing data, acquiring data, and assessing data quality.

Unlike in most other undergraduate S&DS courses, there will not be much lecturing; during most class periods, there will be a particular data challenge that you will tackle individually or in small groups. You will be actively working during much of class time.

2 Intended Audience

This version of the course serves undergraduate students and can be taken in the senior year in lieu of writing a senior essay to fulfill the S&DS major senior requirement. It needs to be kept small; the following priorities will be considered for enrollment:

- S&DS seniors taking the course *for capstone credit*
- S&DS seniors taking the course *not for capstone credit*
- all other undergraduate students, at the discretion of the instructor

Undergraduate students from any department may be welcome if space is available but must seek permission (discussing their background in statistics and goals for the semester) at or before the first class meeting. This is not a class to be shopped casually

during the second week of the semester. The expectation is that you are self-motivated and will push yourself. The most important prerequisite is a willingness to get your hands dirty working with real data sets. There could be periods of two to three weeks where we delve deeply on a single data set. Some of the questions we seek to address will be open-ended. This course will entail a substantial amount of programming.

3 Prerequisites

Before taking S&DS 425, you should have prior coursework in probability, statistics, and data analysis (at the level of S&DS 230 or 361).

The course assumes you are well-acquainted with the **R** statistical programming environment. That means you should be able to read any R code and understand what it is intended to do, possibly with a bit of Googling.

4 Computing

You should use your own laptops for this course and will bring them to class every day. The ability to connect to a projector or display is required. In our classroom, this can be done via WiFi so that no adaptors are necessary.

This is a computationally intensive course, and you will be required to experiment beyond the examples given in class. Unedited computer output will not be accepted; you must carefully document and explain your work. Examples in class should serve as a basis for any required analysis, but you will be expected to use other sources for help: online help, web searches, and other references as the need arises.

In terms of preparation, the first important thing is the installation of the newest versions of **R** and **RStudio IDE**. You will also want LaTeX (MacTeX on the Mac and MikTeX on Windows). For Windows, make sure you get the complete distribution of MikTeX – the basic version will not work!

5 Grades

The homework (which can be open-ended and may sometimes be done in small groups), presentations and participation (both in class and in the extra meetings), and **an individual final project** constitute the entire grade for this class. Correctness of the submitted solution is less important than the preparation and process leading to the solution. A fair bit of emphasis will be placed on writing short reports and presentations. Students will be expected to take turns giving short presentations or demonstrations of their analyses without prior warning during any class.

The work required week-to-week may be substantial. But the class will not dribble into exam period – we'll wrap up during the last week of classes.

Details on the final project will be provided by the fourth week of class. It will be due Friday of the last week of classes. There are no exams.

6 Shopping this class and homework

All homework assignments (including the first one) are required of everyone, submitted on time. This will not be an easy course to slip into after the first class or two. Auditing is not permitted.

7 Academic Integrity

Academic integrity is a core institutional value at Yale. It means, among other things, truth in presentation, diligence and precision in citing works and ideas we have used, and acknowledging our collaborations with others.

You are free to work together on most everything unless noted otherwise. But "working together" is difficult to define when coding (programming) is involved. Many students can benefit from constructive collaborations. However, if you benefit from "working together" then you had better be able to explain and discuss your solution should questions arise. "I don't remember how I did this" would leave you on thin ice.

Specific requirement: any collaboration on homework must be acknowledged up front. Something like "Lastname, Firstname – Homework 1. Worked with John S. and Jane D." would be fine. This is very simple and encourages constructive collaboration if it aids the learning process.