PSCI 107: Introduction to Data Science

Summer 2022

Course instructors

Richard J. McAlexander, Ph.D. – Postdoctoral Fellow, Perry World House. Email: rmca@upenn.edu

Course description

Understanding and interpreting large, quantitative data sets is increasingly central in social science and the business world. Whether one seeks to understand political communication, international trade, inter-group conflict, or a host of other issues, the availability of large quantities of digital data has revolutionized how questions are asked and answered. The ability to quickly and accurately find, collect, manage, and analyze data is now a fundamental skill for quantitative researchers.

The answers to a range of important questions lie in publicly available data sets, whether they are election returns, survey results, journalists' dispatches, or a range of other data types.

Becoming an effective Data Scientist requires two related, but distinct, skill sets: technical proficiency and theoretical knowledge of statistics. Most courses try to teach both at once. This course, instead, will focus primarily in the first: building your skills in data acquisition, management, and visualization. Leaving this course, students will be able to acquire, format, analyze, and visualize various types of data using the statistical programming language R.

A secondary learning goal of this class is to be able to write and talk about statistics in a concise and clear fashion. Being able to run all of the complicated statistics in the world is unhelpful if you can not explain (particularly to non-specialists) what you have found and why they should care. Too many high school and college classes emphasize long essays, when the primary skill you will need is to write short reports (or, let's be honest, emails) to quickly communicate an idea or finding. In this class, we will emphasize this type of writing.

While this course is not a statistics class, we will discuss (in non-technical terms) the fundamental nature of statistics, particularly the important concepts of uncertainty and causality. The expectation is that you take further courses to build on this knowledge. PSCI 207 "Applied Data Science" & PSCI 338 "Statistical Methods" are designed to be a direct follow-ups to this course.

While no background in statistics, political science, or computer science is required, students are expected to be generally familiar with contemporary computing environments (e.g. know how to use a computer, download new software, find the path to saved files etc.) and have a willingness to learn a wide variety of data science tools. Instructions will follow on software to be installed prior to the first class.

Assessment and grading

The grading assessments are designed to test two learning goals: technical proficiency in R, and the ability to communicate clearly about statistics.

Problem sets and the midterm will be graded anonymously. Please hand these assignments in on Canvas with your student number, not your name.

Course content.

The content of the course is going to consist of four different types of materials: pre-recorded video lectures (made by me). Assigned/recommended videos from other sources (these will most likely be math videos that I think will be helpful in intuitively understanding math concepts). Readings, sometimes describing a data visualization theory, sometimes describing a statistical concept. Live meetings with me and Shehroz: the best way for us to meet and communicate with you!

Books.

College is expensive but the books you need for this class are not! We will be using three main resources. They are all free!

Data Visualization: A practical introduction

R for Data Science

Introductory Statistics with Randomization and Simulation

- **Participation (5%)**

Given the distanced nature of the semester, we wanted to provide some incentive to stay connected with the class. At the same time, we understand the participation in the ‘normal’ sense is made more difficult by these conditions. There are multiple ways to participate in the class, and each can be used as a replacement for the others.

Participation can be: asking and answering questions in lecture and in recitations, asking and answering questions on the course Slack, attending office hours, or working with teaching staff on your final paper and presentation. If you are unable to attending any of the office hours or meetings, please be sure you participate in other channels.

- **Problem sets (40%)**

Five problem sets (roughly every two weeks)

Scored on a 1 to 12 scale. Getting all the questions “correct” will translate into a score of 10. Scores of 11 and 12 will be reserved for submissions that have all the correct answers, have code that is particularly cleanly and efficiently written, and have written explanations that clearly and concisely articulate the findings.

Students that average 10 out of 12 can expect to be in the B+/A- range for this component of the class.

• **Midterm 20%**

The midterm will be structured like the homeworks, but will be more in depth and require more written communication.

• **Final Presentation 10%**

Presentations will be about the same topic as your final paper (see below). These presentations will take place on April 22nd during your usual recitation period. Each presentation will be no more than 3 minutes long (with a strict cutoff). You will present exactly one slide with one figure on it that you think best tells the "story" of your final paper. The goal is to walk the audience through why you have a question they want to know the answer to, and why you have the data to answer it. This format is commonly used in "Three Minute Thesis" competitions. An example of this format is posted on Canvas.

• **Final Paper 25%**

This is going to be a report that you write that 1) describes the data, 2) visually presents the data, and 3) provides some analysis of the data (exploring relationships, patterns, etc). We expect that you begin working on this project midway through the semester. You will be graded on how well you communicate information about the data, the sophistication of the data analysis, and how well you present the data.

Computing

We will use R in this class, which you can download for free at www.r-project.org. R is completely open source and has an almost endless set of resources online. Virtually any data science job you could apply nowadays to will require some background in R programming.

While R is the language we will use, RStudio is a free program that makes it considerably easier to work with R. After installing R, you should install RStudio (www.rstudio.com). Please have both R and RStudio installed by the end of the first week of classes.[2]

Communication

We are going to establish a Slack channel for us to communicate. When coding and programming and working with data, one often encounters pitfalls and hiccups. Most experience coders turn to Google. But, this is a beginner course, and Googling code is a skill you will hopefully develop in time. Instead, we (and your fellow students) will be your Google! Ask questions on slack, it will ping my phone and also allow all other students to see what the question is (they likely have similar ones) and see what the answer is (they may be able to help!).

Course Schedule

Week 1 – What is Data Science? What is data? Writing Data

Week 2 – Basic R and Data Analysis

Week 2 – Datasets

Week 3 – Cleaning and Reshaping

Week 4 – Visualization

Week 5 – Linear Models (Bivariate, simple OLS)

Week 6 – Linear Models (Multiple)

Week 7 – Regression & Inference

Week 8 – Regression II

Week 9 – Regression III

Week 10 – Twitter?

Week 11 – Causal Inference