Syllabus

PSCI 107-900: Introduction to Data Science

11-Week Summer Session (May 26 - August 7, 2020)

Instructor:	Nicky Bell (he/him/his) belln@sas.upenn.edu
Office Hours:	Mondays 5:00-7:00pm (Eastern) and by appointment (both on Zoom) I'm happy to schedule appointments if office hours don't work for you; just ask!

Teaching Assistant:	Shehroz Malik (he/him/his) maliksh@sas.upenn.edu
TA Review Sessions:	Thursdays 7:00-8:00pm (Eastern) on Zoom. An asynchronous option is available for those who cannot attend live (see "TA Sessions" section below).
TA Office Hours:	Wednesdays 7:00-8:00pm (Eastern) on Zoom

Communication Preferences: I prefer to meet during office hours or by appointment. However, I am available by email, and I try to respond to emails by the end of the next business day (M-F). If you have a question about the course content, you should post your questions on the Piazza page (described below) and you will likely receive a quick response from your colleagues.

Course Description

This course introduces students to techniques for acquiring, analyzing, and visualizing "big data" using the free statistics program **R**. Although this is not a statistics course, students will learn basic principles of probability and statistics, such as hypothesis testing, simple linear regression, and experiments. Data science is best learned through practice, so this course focuses on real-world applications. In addition to weekly problem sets, students will complete two small research projects: one using publicly available data, and another in which students will design and run their own survey. Leaving this course, students will be able to obtain data in a wide variety of formats, test hypotheses using that data, and visualize and present their findings. This course uses examples and exercises from political science, but the techniques are applicable to many settings, including other social sciences, business, education, natural sciences, the humanities, and more. No background in statistics, computing, or political science is required.

This course fulfills the School of Arts & Sciences Quantitative Data Analysis requirement.

Learning Objectives

There are five major learning objectives in this course, and the activities and assignments have been designed to help you achieve these objectives.

1. You will be able to obtain large quantitative data sets and perform basic statistical analyses on that data using the statistics program **R**.
2. You will be able to visualize and present data in accurate and compelling ways.
3. You will understand the intuition behind hypothesis testing, linear regression, and experiments as statistical techniques and their application to problems in social science.
4. You will be able to design and conduct surveys according to best practices in survey design.
5. You will understand the pragmatic and ethical considerations of collecting, manipulating, and analyzing social science data.

Course Materials

This course will use two textbooks:

- Programming Skills for Data Science: Start Writing Code to Wrangle, Analyze, and Visualize Data with **R** (Addison-Wesley, 2019) by Michael Freeman and Joel Ross
- The Book of **R**: A First Course in Programming and Statistics (No Starch Press, 2016) by Tilman M. Davies

Both of these textbooks are available as "e-books" through Penn Libraries. Additional texts will be posted on Canvas. You are not required to purchase any materials for this course, however you must have access to a computer on which you can install **R**, RStudio, and Zoom (all available for free download and account set-up).

I hope that designing the course around these textbooks will be helpful to you in two ways. First, since each topic in this course builds on the skills and techniques learned previously, you will be able to easily reference back to earlier course material. Second, data science is best learned through practice, and after reviewing more than a dozen textbooks for this course, I believe that the end-of-chapter exercises in these books are among the best. I may assign these exercises as weekly problem sets.

How the Course Will Work

Video Lectures

Each Tuesday, I will post video lectures for that week's topics on Canvas, along with **R** script files (we'll talk about what these are) so that you can follow along on your own. I encourage you to watch these videos before reading the textbook chapters assigned for that week. I find that it is easier to learn **R** through "doing" rather than reading. By following along during the lectures, you will practice the techniques that we learn; the textbook will provide a review of what was covered in the video lectures and offer a few extensions that you can try on your own.

Please note that the video lectures will only be available for two weeks (until the second Tuesday), when they will be replaced by the next week's videos. Although this class meets "asynchronously," we will move through the material together on roughly the same timeline. This is necessary to prepare for the class projects. You should expect to spend about two hours per week on the video lectures.

TA Sessions

On Thursdays from 7:00-8:00pm (Eastern), the TA will hold a virtual recitation on Zoom. The TA will expand on the material covered in the video lectures by working through examples in **R**, and will also reserve time to answer questions related to the weekly problem sets. Attendance is required, though you do not need to attend these sessions "live." (It is strongly encouraged that you attend "live" so that you can ask questions as they arise.) If you cannot attend live, the session will be recorded and posted on Canvas. To receive credit for attendance, you must accurately complete a short quiz based on that week's TA session by Sunday at 11:59pm Eastern. You should expect to spend about one hour per week on the TA sessions.

Problem Sets

Each week, I will assign problem sets for you to complete on your own or in collaboration with other students. Problem sets are graded on completion, and are due each Monday by 11:59pm (Eastern). Please submit your completed problem sets via the "Assignments" page on Canvas. The assignment description on Canvas will indicate what you should submit to receive full credit. You should expect to spend about three hours per week on the problem sets.

The solutions to these exercises will be available to you on Canvas, though I strongly encourage you to complete the exercises without referring to the solutions. You may find that these problem sets are occasionally quite frustrating. Unfortunately, learning a new programming language is a process of trial-and-error. Many of your first attempts to write code will result in errors or outputs that you did not expect. However, figuring out why your code did not work is the best way to learn programming.

You are building muscle memory so that the next time you want to execute the same commands, you will intuitively know how to code it correctly. As we progress through the course, you will find yourself becoming a quicker, more competent programmer.

Course Projects

You will complete two projects as part of this course. More information will be provided during the course. You may work with a partner on these projects. I can assign partners based on shared interests and/or timezones, but if you know classmates with whom you would like to work, please let me know at the appropriate time.

Piazza

We will use Piazza (accessible through the course Canvas site) as a collaborative knowledge bank for questions about course content. To that end, neither I nor the TA will answer individual emails with questions about course content. Please post your questions to Piazza so that others with similar questions may also see the answer. More information about using Piazza will be distributed to the class.

Active engagement on Piazza is encouraged. You are welcome to respond to your colleagues' questions and also to distribute news articles, announcements, and support for Duke basketball. Answering your colleagues' questions is a good way to check your understanding of the material. I will monitor the group and clarify answers when necessary.

Career Talks

I am working to bring guest speakers to the class to talk about careers in data science. I will schedule these sessions according to the availability you indicated on your preclass survey as best as I can. The sessions will be recorded and posted on Canvas if you are not able to attend live.

Grading

Your course grade will be calculated as follows:

On-time Submission of Completed Problem Sets	20%
Project #1 (Analyzing a Dataset)	30%
Project #2 (Survey Experiment)	30%
Attendance at Weekly TA Sessions You may receive credit for attendance by viewing the recorded session and accurately completing a short quiz.	20%

Late assignments put you and the instructor at a disadvantage. For that reason, I will deduct five percentage points from your project grade for every 12 hours after the deadline for late project assignments. Late problem sets will not be accepted.

If your ability to complete any part of the course is affected by a medical issue, family emergency, or other reason, please notify me at the earliest opportunity. We will work together to devise whatever accommodations are necessary to ensure your well-being and success.

Course grades will be converted into letter grades according to the following rubric:

98-100 = A+ (4.0 GPA points)
93-97 = A (4.0 GPA points)
90-92 = A- (3.7 GPA points)
87-89 = B+ (3.3 GPA points)
83-86 = B (3.0 GPA points)
80-82 = B- (2.7 GPA points)
77-79 = C+ (2.3 GPA points)
73-76 = C (2.0 GPA points)
70-72 = C- (1.7 GPA points)
67-69 = D+ (1.3 GPA points)
60-66 = D (1.0 GPA points)

Academic Integrity

When we enter the classroom, we accept a responsibility to our colleagues that we will conduct ourselves with honesty and integrity in the pursuit of knowledge. That includes, but is not limited to, abiding by Penn's Code of Academic Integrity, which covers infractions such as cheating and plagarism. Violations of the Code of Academic Integrity will be referred to the Office of Student Conduct for further action.

I am available to answer any questions you may have about issues regarding academic integrity. It is not assumed that you have this knowledge before entering class.

Accessibility Policy

In compliance with Penn policy and equal access laws, I am available to discuss appropriate academic accommodations that you may require. Requests for academic accommodations need to be made during the first two weeks of the course, except under unusual circumstances, to arrange reasonable accommodations. Students must register with Student Disabilities Services (SDS) for accessibility verification and for determination of reasonable academic accommodations.

Course Outline

In the first part of the course, we will focus on gaining comfort and familiarity with the programming language **R**. The second part of the course will introduce you to basic statistical analyses using **R**.

Week of	**Topic and Readings**
May 27	Introduction to Data Science and **R** • O'Neil and Schutt (2013), Ch. 1: What is Data Science? • Interview with DJ Patil, Former U.S. Chief Data Scientist: Is there a data science code of ethics? • Davies, Ch. 1 and Appendix B • Recommended: Bion, et al. (2018): How R Helps Airbnb Make the Most of its Data
June 2	Wrangling Data I and II • Freeman & Ross, Ch. 11 • Recommended: Best Practices for Writing R Code (Software Carpentry) • Optional: For a review of introductory R, see Freeman & Ross, Chs. 5-7 and 9-10
June 9	Wrangling Data III and Managing and Sharing Data • Highly Recommended: Sandve, et al. (2013), "Ten Simple Rules for Reproducible Computational Research"
June 16	Visualizing and Presenting Data • Berinato (2016), Ch. 5: Getting to "The Feeling Behind Our Eyes" • Freeman & Ross, Chs. 16 • Optional: Freeman & Ross, Chs. 15
June 23	Elementary Statistics and Hypothesis Testing • O'Neil and Schutt (2013), Ch. 2: Statistical Inference, Exploratory Data Analysis, and the Data Science Process • Davies, Chs. 13 and 18.1, 18.2, and 18.5 • Reinhart (2015) Ch. 9: Researcher Freedom, Good Vibrations? **• Project #1 instructions distributed**
June 30	There will be no lectures and no problem set this week. The instructor will hold extended office hours for assistance with Project #1 **• Project #1 Due Friday, July 3 at 5:00pm (Eastern)**
July 7	Simple Linear Regression • Davies, Chs. 20.1-20.4
July 14	Multiple Linear Regression • Davies, Chs. 20.5, 21.1, and 21.3
July 21	Experiments • No assigned readings **• Project #2 instructions distributed**
July 28	Advanced Features of **R** • Readings TBA

August 4	There will be no lectures and no problem set this week. The instructor will hold extended office hours for assistance with Project #2 ● **Project #2 due Friday, August 7 at 5:00pm Eastern**

Last Updated: July 27, 2020
Subject to change.