Instructor: Jose Aveldanes
Email: jose_aveldanes@berkeley.edu

ATDP: Data Science with R and Python
Meetings: Tuesdays and Thursdays
8:30-12:30 AM & 1:00 PM-4:00 PM

May 27th, 2022

Greetings! Welcome to Data Science with R and Python! My name is Jose Aveldanes, and I will be your instructor for this summer. I am a Ph.D. student in Sociology and Demography at UC Berkeley where I also teach Sociology and Data Science. I am looking forward to meeting you all and going on this journey with you through the trials and tribulations of Data Science. To prepare you for the course, I have included a pre-class assignment.

In this course we will be learning data science in both learning the technical skills and broader implications of data science. By "broader implications" I mean to suggest that--in addition to gaining the technical tools to model—we will also dive into the ethics and social problems related to data science. Data is incredibly powerful in our modern era; it shapes how we interact with others (e.g., social media), the movies and music we listen to (e.g., in Netflix suggestions), marketing (i.e., via ads), and resource-allocation (e.g., in social policy)to list a few examples.

This class will focus on gaining exposure to the technical aspects of modelling in addition to thinking deeply about model building and interpretation. Students will work in a mix of R and Python lab exercises and self-reflection essays and a final project that displays the skills they have learned. They will learn key concepts of data science (e.g., algorithms, machine learning, different types of analyses, data types). We will consider the steps requisite for building a statistical model and explore their limitations and ethical implications. We will investigate misleading data, how bias skews calculations, and learn skills to be able to recognize the kinds of questions you can answer with data science. Through lab exercises, class discussion, and writing exercises, students will learn the skills to create basic statistical models and the skills to think critically about the power of these skills in shaping different demographics in the U.S. They will end the course by engaging in a final project that showcases their skills by investigating a question of their choice.

Looking forward to seeing you all in class!

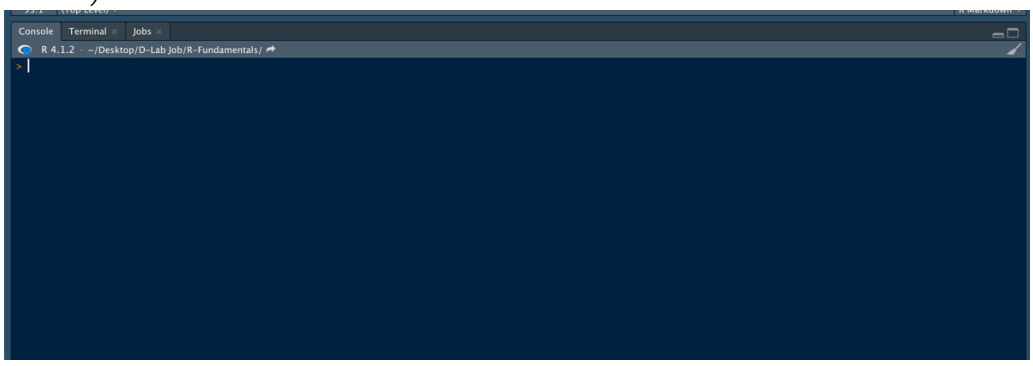
All my best,

Jose Aveldanes

Pre-Class Assignment

1. Go to The Comprehensive R Archive Network (**CRAN**) and download the R language as it corresponds to your computer system (e.g., Linux vs. Mac vs. Windows). System requirements are having a computer that has **at least 256 MB of RAM, a mouse, and enough disk space for recovered files, image files, etc**. The administrative privileges are required to install and run R-Studio utilities. You just also have a network connection for downloading data.
2. Go to the R Studio IDE (Integrated Development Environment) website and download the R Studio IDE. Below is the button you should click next after you follow the link I have provided above.

3. Install the R packages "dplyr" and "tidyverse." You will do this if and only if you have first installed the R Language from **CRAN** and then R Studio from the R Studio IDE website. In order to do this you must do the following:
   a. Launch R Studio after and only after you have installed R
   b. On the bottom left you should see a window that says "console" (see below).



4. Type install.packages("dplyr"). Do the exact same for tidverse (i.e. type install.packages("tidyverse"). Type exactly what is in the gray boxes, no more no

less. If you have done this correctly you should come across the following screen in the same console that you typed into in the previous step:

```
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.1/dplyr_1.0.9.tgz'
Content type 'application/x-gzip' length 1307837 bytes (1.2 MB)
==================================================
downloaded 1.2 MB


The downloaded binary packages are in
```

5.
6. If this works for both dplyr and tidyverse, you have done everything you need to do before the start of class in terms of software requirements. I will also provide a similar assignment for Python but that will be for another day.
7. The penultimate step of this pre-class assignment is for all of you to **read the Preface and Chapter 1** of the textbook for the course. If you don't understand all the math, that is okay. We will discuss the conceptual parts of the chapter as well as make sure we know the key terms. Write a one-page self-reflection of why you are interested in data science and how you might apply the reading for the first chapter to your interests. See you soon!
8. The final step is to ready the "About the Book" and "Introduction" to *Fairness and Machine Learning* and write down some questions you might have about fairness as it relates to statistical learning.

Data Science with R and Python

Instructor: Jose Aveldanes
Office: Social Sciences Building
Email: jose_aveldanes@berkeley.edu
Office hours: By appointment

Day: Tuesday & Thursday
Times: (Section 1) 8:30AM-12:00 AM
(Section 2) 1:30 PM-4:00 PM
Place: Zoom

Course Summary:

Welcome to Data Science with R and Python! I am thrilled to be a part of your introduction to the world of data science. While not an introductory course, per se, there are some aspects of the course that you might be getting introduced to for the first time. I have worked arduously to make sure this course is both intellectually rewarding and a source of community in these difficult times. I have selected reading material that will introduce you to the kinds of critical data scientists ask and how they might answer them. The assignments are designed to train you in practical data science skills, and to engage you socially and intellectually throughout the semester. If you ever wonder why we are engaging with a specific text, or why we are completing a specific activity, ask! I intend to make this a transparent classroom.

Breaking Down Sections:

Like most courses at Berkeley, Data Science with R and Python is split into two parts: lectures and lab sections. Everyone enrolled in the course must attend lecture and lab. You are expected to come to lecture, and lab having done the readings. It is imperative that you do this, otherwise it will be easy to get lost. The lectures and lab serve as a way to solidify your understanding.

Section Objectives:

- ✓ Discuss and clarify readings
- ✓ Ask questions and develop new ones
- ✓ Apply course concepts to lab sections
- ✓ Build and reinforce your understanding of the theoretical and substantive approaches in data science
- ✓ Practice critical thinking and technical reasoning
- ✓ Discuss/debate work with fellow students

Assigned Readings:

All readings are posted to the Canvas site. It is your responsibility to download all readings from Canvas. *Students are expected to read all the assigned readings before the start of class (as outlined on the syllabus) and be prepared to participate in the class discussions about readings.* Readings for this course come from a variety of materials and the course textbook. The textbook for this course is

Canvas Site Information: This course has enabled open enrollment. Students can self-enroll in the course at this URL: https://canvas.instructure.com/enroll/GL4MWA. Alternatively, you can sign up at https://canvas.instructure.com/register and use the following join code: GL4MWA

Grades are based on the following:

1. Attendance (15% of final grade)
2. Participation (15% of final grade)
3. Labs (30% of final grade)
4. Final Project (20% of Exam)
5. Reflection Memos (20% of Exam)

Attendance (15% of final grade): Attendance is required at all sections. Please make sure to arrive on time. You have **no discretionary section absences** to utilize throughout the course. Each absence will result in a penalty of 3 percent of your attendance grace.

Please talk to me if this is an issue and we can find a solution contingent on the situation. *I know that attendance is not possible for everyone for many reasons right now*. If you are unable to attend section (whether due to one-off instances of illness or ongoing scheduling conflicts due to work, home responsibilities, technological issues, other classes, etc.), **please send me an email as soon as possible.** We can find a work-around involving office hours, or meeting outside of class with other students, or another solution.

Participation (15% of final grade): This will primarily be group discussion. My classroom is a safe place to speak, and I will not judge you. I understand for some folks speaking up might be more intimidating relative to others. Participating in discussion section means coming prepared to engage with the course material; to ask questions of it, ourselves, and others; and to respectfully share, listen to, contemplate, and engage with new perspectives and with others' experiences. This also entails adhering to the discussion guidelines we will collectively craft together during the first weeks of section. We will work on this together.

Plagiarism:

Students who resort to plagiarism often do so out of a sense of hopelessness and overwhelming anxiety for any number of reasons. *Please*, if you find yourself in a situation where you are considering plagiarizing, realize this is a **serious** offense and will likely only make the situation worse. Instead, reach out to me to discuss whatever is going on and I will do my best to help you through the situation and work with you to find a solution as best I can. Plagiarism is extremely serious and the consequences are not worth it.

Please review UC Berkeley's policies on plagiarism here: https://sa.berkeley.edu/conduct/possible-outcomes/academic-misconduct.

Communication:

I will primarily use email to communicate with the class. **If you need to contact me, please email me directly at jose_aveldanes@berkeley.edu.** Please use proper etiquette when writing emails and include your name and Data Science with R and Python (ATDP) along with section in the subject line. I will respond in a timely manner (usually within 24 -48 hours), but do not expect email replies after 5 PM. If you email me after 5 PM on Friday or on the weekend, do not expect a response until Monday at the earliest.

Zoom Links

Please do not share this info outside of this class. Link:
https://berkeley.zoom.us/j/93644931697?pwd=RkdwcDJ2TkwzbFUvTVhWVlZSQkE0UT09

Meeting ID: 936 4493 1697
Passcode: 445064

Section Expectations and Etiquette:

- Give me and your peers your complete attention during class.
- Come to class prepared and ready to learn. Complete the reading assignments before coming to class, ask questions, and participate in discussions.
- If you have a grievance about the course, please express your grievance to me in writing.

Thematic Breakdown of Classes:

06/21/22 – Introduction, getting set up with R, R Studio, Python and Jupyter Notebook

Assignment: Introduction to Data Science in R

Reading: Chapter 1 from Practice Statistics for Data Scientists

06/23/22 – Frequencies, Distributions, Probabilities, Observations, and Variables (Part I: Lecture & Lab)

Reading: Chapter 2 from Practical Statistics for Data Scientists **through the section on Confidence Intervals**

06/28/22— Frequencies, Distributions, Probabilities, Observations, and Variables (Part II: Lecture & Lab)

Reading: Chapter 2 from Practical Statistics for Data Scientists **from section on Normal Distributions until the end**.

06/30/22—You can't run from testiny!

Reading: Chapter 3 (Statistical Experiments and Signficance Testing) from Practical Statistics for Data Scientists from section on **A/B Testing to t-tests**

07/05/22—Experiments, Observational Data, Confidence Intervals and Random Samples (Lecture & Lab)

Reading: Chapter 3 (Statistical Experiments and Signficance Testing) from Practical Statistics for Data Scientists from section on **Multiple Testing until the end of the chapter**

07/07/22—Modeling and its Discontents

Reading: Chapter 3 (Linear Regression) from *Introduction to Statistical Learning*

07/12/22 Statistical Learning: Applications Beyond Linear Regression (Part I: Lecture & Lab)

Reading: Chapter 4 with sections **4.1, 4.2, & 4,3** from *Introduction to Statistical Learning*

07/14/22— Statistical Learning: Applications Beyond Linear Regression (Part II: Lecture & Lab)

Reading: Chapter 4 (Classification) with sections **4.4, 4.5, & 4.6** from *Introduction to Statistical Learning*

07/19/22—Ethics for Data Scientists

Reading: Chapter 6 from *Bit by Bit* (Matthew Salganik)

Fairness and Machine Learning Chapter 6 by Barocas, Hardt, and Narayanan

07/21/22 –Review for end of class/final projects

No Readings

07/28/22 – Final Project Presentations

No Readings