

Economics 5/Political Science 5D

Introduction to Social Data Analytics

UC San Diego
Summer Session II 2025

Professor: Joshua Arons (jaarons@ucsd.edu)

Teaching Assistant: Precious Fasakin (pfasakin@ucsd.edu)

Class Schedule: Class: MW 11-12:50 in RWAC 0115, Lab: F 11-12:50 in RWAC 0115

Office Hours: Josh: MW 1-1:50 in RWAC 0115, [zoom](#) by appointment
Precious: Tuesday 10-11:30 by [zoom](#)

Overview

This course has three main goals. The first goal is to introduce you to interesting and important social science questions. Each chapter will highlight a different application, often highlighting research from faculty at UCSD. We will cover a wide range of topics, including how colleges promote intergenerational mobility, what motivates people to vote, and how we identify discrimination in labor markets, among others.

The second goal is to show you how data can be used to inform our discussion of these topics. We will be using a lot of different datasets in this course. Some come from experiments that have been run by researchers. Others come from administrative datasets collected by governments. For many of the datasets, we will only have time to scratch the surface of what is possible with this data.

The third goal is to give you the tools to perform data analysis. We will focus on three popular software: Excel (1 chapter), Stata (3 chapters) and R (4 chapters). While learning coding fundamentals in each of these programs, we will shed light on big social science questions. This course does not require any prior coding or data analysis experience, and lectures and materials are structured to help get you started using each software.

Lectures and Labs

There will be two lectures per week. All in-person lectures will be recorded and posted after the lecture via video podcast. In addition, all the lecture material, as well as bonus material, is covered in a series of videos on Canvas. In certain cases, I will ask you to watch a video before arriving at class or read a chapter from the online textbook. Make sure you do this reading so that you are prepared for the material in lecture.

In addition to lectures, each week there will be a lab. Each lab will involve completing an Excel workbook, Stata Do-file, or R script. If you attend in person, you will work on sections of the lab together with your classmates that attend in person.

Each week you will need to answer two short quizzes related to the lab. During the in person lab, the TA will discuss answers to the quiz. If you cannot attend in person, you can still get credit for the lab by completing the lab and then completing the quiz related to that lab.

Class will not be held on September 1st (Labor Day).

Course Materials

All of the course materials will be made available through [Canvas](#). This includes the software we will be using in the class, as well as a link to the [online textbook](#) for the course.

Assessment

Your grade will be based on a combination of:

- **Lab (15%):** Each week you will need to complete two online quizzes associated with the lab. If you attend lab, the answers for the lab will be covered during the lab itself. If you do not attend, you can still get credit by completing the lab on your own time and completing the associated canvas quiz. The lab quizzes will be due six days after the lecture to which they relate, so you will have labs due on Sunday and Tuesday evenings. Unlike the class quizzes, you will only have **one attempt**.
- **Quizzes (25%):** These are separate from the lab quizzes you will turn in. There will be quizzes that correspond to each lecture. You will need to complete them within four days of each lecture (so they will be due on Friday and Sunday evenings). In general, they will be a mix of multiple choice and occasional short answer questions. The quizzes randomly draw questions from a test bank and you will have **two attempts**.
Should you receive below your desired score, you will have the opportunity to study and raise your score. Four days before each exam, I will post a second, optional quiz for each topic on the exam that will be available for the next three days (answers posted one day before the exam). Should you take the quiz and score higher than on your previous quiz, you will receive the average of your two quiz grades (if you take it and score lower the second time, you will keep your first grade).
As an example, suppose you scored an 80% on the second quiz, which covers Stata basics and interpreting the distributions of data. Four days before the exam, I will post several quizzes. If you took the new quiz on Stata basics and interpreting distributions and scored 100% this time, your grade for the second quiz would be raised to 90%.
- **Midterm (20%):** The Midterm will be held in class on **Wednesday August 20th**. It will be a multiple choice and short answer test to assess your understanding of theoretical concepts from class rather than ability to code. There will be no questions about coding syntax.

- **Final (40%):** The final will be an open book/note/resource take-home exam due **Saturday September 6th**. It is meant to assess the coding skills you have learned in class with some application of concepts from class.

The course may be graded on a relative curve. The curve is a safety net, and can only improve your grade, not reduce it. If the curve comes into effect, letter grades will depend on your ranking in the class. In past courses, generally you can be confident you will receive an A of some type (A+, A, or A-) if you are in the top 30 percent. You can be reasonably confident you will receive at least a B of some type (B+, B, or B-) as long as you are in the top 70-75 percent of the class. These numbers are only meant as a guide to help you understand how well you are doing in the class and are not steadfast rules.

Important Due Dates

- Weekly labs are due 6 days after the class they correspond to at 11:59 PM.
- Quizzes are due 4 days after the class they correspond to at 11:59 PM.
- **Midterm:** Wednesday August 20th in class.
- **Final:** Take home exam due 12:00 PM on Saturday September 6th.

If you think you will not be able to complete an assignment on top for some reason, please email at least 12 hours ahead of the deadline for an extension.

Late Policy

Quizzes and labs can be turned in up to a week late for partial credit. Any quiz or lab turned in within 24 hours of the due date will receive a 25% reduction in maximum points. Any quiz or lab turned in after 24 hours but within a week will receive a 50% reduction in maximum points. Any quiz or lab not submitted within a week of the due date will earn a zero. The Canvas quiz will not remain open indefinitely, so please email (jaarons@ucsd.edu) if you need it to be reopened.

Academic Honesty and Plagiarism

All graded work must be done by you. If you are unfamiliar with the University's policy on academic integrity, please see

<http://senate.ucsd.edu/Operating-Procedures/Senate-Manual/Appendices/2>. There is a zero tolerance policy for academic integrity violations, and if you are found to have violated the University's academic integrity policy you will receive a failing grade in the course.

You are allowed to use online resources such as AI (i.e. ChatGPT) and forums (i.e. stackexchange.com) for coding, but if you choose to do so, you should exercise caution. These resources are very helpful for quicker coding, but will still require work on your part. An important part of using them is understanding what question you want to answer and how to modify the answer you've found on a forum or from AI for your purpose. For example if you try to just copy and paste the lab into ChatGPT and submit its output, you are likely to get the lab wrong.

Course FAQs

1. **What do I need to buy for the course?** Nothing. We will use an online textbook that was written for the course and is available through Canvas. The software we will be using (Excel, Stata, and R) is also available through Canvas with installation instructions on the homepage.
2. **I have a question about the course, where should I go to ask this question?**
 - a. First, check the course website and the syllabus. For example, if the question is: is there a quiz this week, you should navigate the Modules and find the given week for the course. Many questions can be answered by looking through the Plan for the Week on Canvas.
 - b. If you still can't find the answer to your question, you can ask it on the Class discussion board in Canvas. **This should be where you post most of your questions. Please, however, do not post questions that include code or partial answers to quizzes/labs.**
 - c. If you have a question related to the course that you don't think is relevant for other students or you have a particularly sensitive question that you don't want to be read by others, you can send an email to Josh (jaarons@ucsd.edu). I will try to reply to emails within 24 hours.
3. **I'm on the waitlist, what are my chances of getting off?** Waitlists at UCSD are automated, and I can't manually allow anyone off the waitlist. While we try to expand the class when there is excess demand, we are sometimes constrained by classroom size and TA availability.
4. **I think a question on my quiz or test was graded incorrectly.** You can submit a regrade request on the Midterm through gradescope. For all other requests, please email Josh (jaarons@ucsd.edu) and clearly document why you think a mistake was made. We will regrade that portion of the quiz or exam, meaning your grade could go up or down.

Course Schedule

The schedule below lays out what is covered each week both in terms of the empirical application as well as the coding and software.

Class 1: Class Introductions and Introduction to Excel

- Course expectations and the syllabus
- Empirical Application: Instructor Incentives and Student Performance, by Andy Brownback and Sally Sadoff (2020)
- Data tables
- Functions
- Pivot tables

Class 2: Introduction to Stata

- Empirical Application: Intergenerational Mobility Rates by College. Data comes from Opportunity Insights
- The Stata Graphical User Interface (GUI)
- Do-files
- Basic data analysis commands
- Interpret and constructing histograms

Class 3: Data Wrangling in Stata

- Empirical Application: Racial Discrimination in Traffic Stops. Data comes from the Stanford Open Policing Project
- Introduce concept of data wrangling
- Learn the append, merge, and collapse commands
- Bar charts in Stata
- Ways to improve data visualization in Stata

Class 4: Regression in Stata

- Empirical Application: Disrupting Education using Technology, by Muralidharan, Sing, and Ganimian (2019)
- Estimate and interpret linear regressions in Stata
- Introduce concept of fitted values and residuals
- Visualize and plot the results of regressions in Stata

Class 5 and 6: Midterm and Introduction to R

- Empirical Application: Resume Experiments, by Bertrand and Mullainathan (2004)
- Objects and variables in R
- Introduction to data frames
- Subsetting data frames in R

Class 7: Data Wrangling in R

- Empirical Application: The Rug Rat Race, by Garey Ramey and Valerie Ramey
- If statements
- For loops
- Introduction to the tidyverse package

Class 8: Data Visualization in R

- Empirical Application: China's War on Air Pollution by Greenstone, He, Jia and Liu)
- Histograms, scatter plots, and box plots in R
- Using dates in R
- ggplot2

Class 9: Linear Regression in R

- Empirical Application: The Butterfly Ballot
- Linear regression in R
- Plotting the regression line
- Predicted values and residuals