Econ 5 / Poli 5: Introduction to Social Data Analytics

Summer Session I 2019

Instructor:	Brandon Merrell, bmerrell@ucsd.edu
Lectures:	Tuesdays and Thursdays, 2:00–4:50 in SOLIS 110
Office Hours:	Tuesdays 11:45–1:00pm (and by appointment) in SSB 448
Online Content:	http://TritonEd.ucsd.edu
TA:	Duy Duc Trinh, ddtrinh@ucsd.edu

Description: As data about individuals, organizations, and governments become increasingly available, social data analytics are transforming the way we think about the economy, politics, and society. This course will introduce students to various skills that are necessary to navigating the world of social data. We will learn basic principles of coding through the lens of popular software, including Excel, Stata, and R. While learning coding fundamentals, we will shed light on important social science questions and grapple with larger issues that newly available data allow us to analyze.

Prerequisites: This course does not have any formal prerequisites. Although the class is relatively heavy on applied statistical techniques, the mathematical demands are relatively light; our focus will very much be on helping students understand the intuition of the methods we use and helping them interpret the outputs of these techniques rather than their internal workings.

Assignments and Requirements:

Problem Sets (35%): Three problem sets will be assigned during the quarter. The first two are worth 10% each, the last is worth 15%. Problem sets contain analytical, computational, and data analysis questions. The following instructions apply to all problem sets unless otherwise noted.

- The first two problem sets are due on Tuesday, July 16th and Tuesday, July 23rd via TritonEd. The final problem set is due on August 3rd. All problem sets are due at 2:00pm on the day of the deadline. Late submissions will not be accepted.
- Copies of the write-up and accompanying code should be submitted electronically via TritonEd by the due date.
- Working in groups is encouraged for conceptual and sometimes technical discussion, but each student must submit their own writeup of the solutions that shows their own independent work on the assignment. In particular, you should not copy someone else's answers or computer code. We also ask you to write down the names of the other students with whom you solved the problems together on the first sheet of your solutions.
- For analytical questions, you should include your intermediate steps, as well as comments on those steps when appropriate. For data analysis questions, include annotated code as part of your answers. You will lose points on your problem set if your code and write-up is not properly formatted and documented. All results should be presented so that they can be easily understood and code should run easily without errors.

Midterm Exam (30%): A midterm will be held in class on **July 25th** covering the material in the first half of the course.

Final Project (30%): Your assignment is propose an empirical research question and analyze data related to that question. Your grade will consist of a research proposal (5%), a paper (20%), and a short presentation (5%). This presentations are held on **August 3rd** from 3:00pm to 6:00pm. Students may work individually or in pairs.

Participation (5%): Students are strongly encouraged to ask questions and to actively participate in discussions during lectures and sections.

Rules and Policies:

Academic Dishonesty: Students must complete their own work unless the instructor has otherwise provided permission for collaboration. Any student who is caught cheating or plagiarizing will receive a failing grade for the course and will be reported to the Academic Integrity Office for administrative sanction. If you are unfamiliar with the university's policy on academic integrity, please see: http://senate.ucsd.edu/Operating-Procedures/Senate-Manuel/Appendices/2.

Late Assignments and Missed Exams: All problem sets are considered late immediately upon the due date. Assignments incur a 10-point penalty for each 24-hour period they are late. Example: a problem set earning 100/100 points would drop to:

- 90 points if submitted after Tuesday at 2PM, but before Wednesday at 2PM.
- 80 points if submitted after Wednesday at 2PM, but before Thursday at 2PM, etc.

If you know you will miss an exam for a legitimate reason, notify me at least a week in advance. Email is perfectly acceptable. If you cannot contact me in advance, you must do so as soon as possible. I will work with you to resolve reasonable problems, but it is your responsibility to arrange with me to take a makeup exam at least 48 hours before the grad submission deadline.

Attendance: Class attendance is not mandatory but will probably improve your performance on assignments. Some material is also easier to learn when you hear someone explain it and/or when you have an opportunity to discuss it with others.

Grades and Appeals: You will be graded solely on your academic performance. I use the following grading scale: "A-" = [90-$93.\bar{3}$), "A" = [$93.\bar{3}$-$96.\bar{6}$), "A+" = [$96.\bar{6}$-100], with other letter grades following analogous intervals. Students can appeal grades that they believe are incorrect. Grade appeals will consist of a single typed page that identifies the problem and presents a reasoned argument that the grade fits the appeal criteria. Regrade requests must be submitted within 24 hours of receiving the assignment grade.

Disability: Students who will request accommodations should register with the Office for Students with Disabilities (University Center 202; 858.534.4382) and provide me with documentation outlining appropriate accommodations. I am happy to meet with you during my office hours to discuss your needs.

Materials and Software: Because we will be learning Excel, Stata, and R, we will draw on a number of different resources. Many of these resources will be videos from YouTube, blogs, and some will be traditional textbooks. All are freely available online or have been provided by the authors. A few of the primary sources are listed below:

- Principles of Coding: We will rely on videos and exercises from the Hour of Code: https://code.org/learn
- Excel Easy Tutorial: http://www.excel-easy.com/
- Princeton Stata Tutorial: http://data.princeton.edu/stata
- UCLA Stata Resources: http://www.ats.ucla.edu/stat/stata/
- TextBook: *A First Course in Quantitative Social Science*, by Kosuke Imai (Princeton University Press)

This course will consist of three different statistical software programs commonly used by social scientists:

- Excel: All students will need to have access to Excel. Excel is also available in UCSD computer labs.
- Stata: Instructions for getting Stata through the Virtual Computing Lab are available on the TritonEd Website.
- R: an open-source statistical package. You can download it from the web here: http://cran.r-project.org/. RStudio is a useful tool for coding in R. You can download it from the web here: https://www.rstudio.com/

Course Schedule:

Meeting #1: Course Introduction and Web Scraping Demonstration (Tuesday, July 2nd)

Course Materials

- "Getting Started with Data," Hilary Mason. https://www.youtube.com/watch?v=GXjjMSn2Nws
- "Big data in the service of humanity: Jake Porway" https://www.youtube.com/watch?v=fZ3xXXeVrIQ
- *Statistical Modeling: A Fresh Approach*, Daniel Kaplan. Chapter 2, 2.1-2.4. http://www.mosaic-web.org/go/StatisticalModeling/Chapters/Chapter-02.pdf

No Class (July, 4th holiday)

Meeting #2: Functions, Boolean Logic, and Plotting in Excel (Tuesday, July 9th)

Course Materials

- “Introduction to Functions and Formulas” http://www.excel-easy.com/introduction/formulas-functions.html
- “Cell References” http://www.excel-easy.com/functions/cell-references.html
- “Logical” http://www.excel-easy.com/functions/logical-functions.html
- “Count and Sum” http://www.excel-easy.com/functions/count-sum-functions.html
- “Statistical Functions" http://www.excel-easy.com/functions/statistical-functions.html
- “Lookup and Reference Functions" http://www.excel-easy.com/functions/lookup-reference-functions.html
- “Function Errors" http://www.excel-easy.com/functions/formula-errors.html

Meeting #3: Reproducibility, Description, and Plotting in Stata (Thursday, July 11th)

Course Materials

- “Stata Tutorial: Introduction” http://data.princeton.edu/stata/
- “Introduction to the Stata Interface,” Alan Neustadtl, 15 minutes. https://www.youtube.com/watch?v=KkCKEK7lwuo&index=1&list=PLRYSxJ3XjgQM342QrBkzek8clHa5ue4Sd
- “Using the Stata Program Editor," Alan Neustadtl, first 7 minutes. https://www.youtube.com/watch?v=XmvWydFD2Y0&index=6&list=PLRYSxJ3XjgQM342QrBkzek8clHa5ue4Sd
- Bill Gates Explains If Statements, Hour of Code, https://www.youtube.com/watch?v=m2Ux2PnJe6E
- Data Management in Stata, http://data.princeton.edu/stata/dataManagement.html
- “The Beauty of Data Visualization,” David McCandless TED Talk, 20 minutes. https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization?language=en
- “Stata Graphics", http://data.princeton.edu/stata/graphics.html

Meeting #4: Data Wrangling and Regression in Stata (Tuesday, July 16th)

First problem set due.

Course Materials

- “Introduction to Residuals and Least Squares Regression," Khan Academy, https://www.youtube.com/watch?v=yMgFHbjbAW8, 7 minutes.
- “Simple and Multiple Regression in Stata,” Section 1.0 and 1.3 https://stats.idre.ucla.edu/stata/webbooks/reg/chapter1/regressionwith-statachapter-1-simple-and-multiple-regression/

Meeting #5: Logical Operators, For-Loops, and If-Statements in R (Thursday, July 18th)

Course Materials

- "Data Analysts Captivated by R's Power" *The New York Times* http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html
- Imai, 1.3.1-1.3.3
- Imai, 2.1-2.2
- Mark Zuckerberg on For Loops, https://www.youtube.com/watch?v=mgooqyWMTxk&
- Trina Roy on Counters, Hour of Code, https://www.youtube.com/watch?v=gxc9RCMvky0&
- Imai, 4.1

Meeting #6: Individual & Group Meetings (Tuesday, July 23rd)

Second problem set due.

Meeting #7: Midterm Exam (Thursday, July 25th)

Meeting #8: Visualization and Regression in R (Tuesday, July 30th)

Course Materials

- "Visualizing Ourselves with CrowdSourced Data," Aaron Koblin https://www.ted.com/talks/aaron_koblin
- Imai, 3.3, 3.6
- Imai, 4.2

Meeting #9: Homemade Functions and Data Wrangling in R (Thursday, August 1st)

Course Materials

- Chris Bosh on Functions, https://www.youtube.com/watch?v=0eo0ESEX9DE
- Imai, 1.3.4
- "Program or be Programmed," Excerpts and Video. Douglas Rushkoff. http://www.shareable.net/blog/program-or-be-programmed
- "Big Data, Inequality, and the Law" Latanya Sweeny (first 15 minutes ONLY) https://vimeo.com/146814921

Meeting #10: Project Presentations (Saturday, August 3rd)

Final problem set due.