This version: January 6, 2020

Economics 5/Political Science 5D
Introduction to Social Data Analytics

M/W 4:00 - 4:50 PM in Center 212

Zack Goodman
zgoodman@ucsd.edu
Office: Sequoyah 108
Office Hours: 9:00 - 11:00 AM on Wednesdays

Overview

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 teach skills necessary to navigate the world of social data. We will learn basic principles of coding through the lens of popular social science data analytics softwares Excel, Stata, and R. While learning coding fundamentals, we will shed light on big social science questions and grapple with larger societal questions that the era of a society governed by data presents us.

Assessment

Your grade will be based on a combination of:

- **Homeworks (40%):** Four problem sets will be given throughout the quarter. Problem sets will contain analytical, computational, and data analysis questions. Each problem set will be counted equally toward the calculation of the final grade. The following instructions will apply to all problem sets unless otherwise noted.
  - Homeworks will be due on Monday's at 3:00 pm. Late submission will not be accepted under any circumstances.
  - Each student will be allowed to drop one problem set grade to accommodate for special circumstances.
  - Copies of the homework write-up and accompanying code should be turned in electronically via Canvas 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 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 at the top of your solutions submission.
  - 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 (20%):** A midterm will be given in class on May 3 covering the material in the first half of the class.
- **Final Project(30%):** Students will complete a final project in lieu of a final exam.
- **Pre Class Exercises (5%):** Short exercises will be given before each lecture. Students should complete these exercises before coming to class and upload their responses to Canvas.
- **Participation (5%):** Attendance in lecture and lab is mandatory. Attendance will be taken in lab, and attendance quizzes may be held in lecture. Students are strongly encouraged to ask questions and actively participate in discussions during lectures and sections.

Academic Honesty and Plagiarism

All of your 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.

Course Website and Piazza Forum

Syllabus and course materials. The syllabus, assignments, solutions, and other course materials will be posted on Canvas. Assignments will be turned in via Canvas.

`https://canvas.ucsd.edu/courses/12007`

Online Q&A. Throughout this class we will use the Piazza online discussion board. This is a question-and-answer platform that is easy to use and designed to get you answers to questions quickly. It supports code formatting, embedding of images, and attaching of files. We encourage you to ask questions on the Piazza forum for clarifications, questions about concepts, or about your projects in addition to attending recitation sessions and office hours. You can sign up to the Piazza course page either directly from the below address (there are also free Piazza apps for the iPhone and iPad):

`https://piazza.com/class/k4xna87tzdl4ju`

Using Piazza will allow you to see and learn from other students' questions. The instructors will regularly check the board and answer questions posted, although everyone else is also encouraged to contribute to the discussion. You can opt to send a question only to your class, or to send it to the wider group of students in Poli5D/Econ100. A student's respectful and constructive participation on the forum will count toward his/her class participation grade. *Do not email your questions directly to the instructors* (unless they are of personal nature) — we will not be answering your questions regarding course materials or problem sets through email.

Course Materials

Since 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)

Software

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

- Excel: All students will need to have purchased 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 Canvas 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/`

Dates to Remember

- January 22nd 3:00 pm: Problem Set 1 Due
- February 5th 3:00 pm: Problem Set 2 Due
- February 6th during lab: Midterm
- February 24th 3:00 pm: Problem Set 4 Due
- March 2nd 3:00 pm: Problem Set 4 Due
- March 13th 11:59 pm: Final Project Due

COURSE SCHEDULE

1 January 6: Course Introduction and Why Data Analytics?

Learning Objectives

- Understand the syllabus and expectations including evaluations and academic integrity
- Know where to find learning objectives for each lecture and how they relate to evaluations
- Recall what resources are available to students (book, software, Canvas, Piazza)
- Describe how social data analytics can be used to address important social problems

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`

2 January 8: Data Format and Intro to Excel

Learning Objectives

- Open Excel, save workbook, edit cells, autofill down column, apply filter, sort columns
- Identify observations and variables in an Excel workbook
- Discern the unit of analysis in a data table and demonstrate how to change it
- Implement statistical and logical functions
- Understand basic Boolean logic and use logical operators

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 Functions" http://www.excel-easy.com/functions/logical-functions.html
- "Count and Sum Functions" http://www.excel-easy.com/functions/count-sum-functions.html
- "Statistical Functions" http://www.excel-easy.com/functions/statistical-functions.html

3 January 9: Lab 1, Excel

Learning Objectives

- Finish a partially complete Excel file that demonstrates mastery of the following:
  - Generating new variables using functions
  - Freezing columns/rows, adding filters, and sorting
  - Implementing logic and Boolean operators
- Apply the following Excel functions: COUNTIF, AVERAGEIF, MATCH, VLOOKUP, and summary statistical operators

4 January 13: Functions in Excel

Learning Objectives

- Resize columns, paste values, use MATCH and VLOOKUP
- Classify kinds of variables (numerical, (un)ordered categorical, logical)
- Identify when a sample contains sampling bias and implications for external validity
- Use the RAND function to conduct a simple random sample

Course Materials

- "Lookup and Reference Functions" http://www.excel-easy.com/functions/lookup-reference-functio
html
- "Function Errors" http://www.excel-easy.com/functions/formula-errors.html
- "Random Numbers [in Excel]" https://www.excel-easy.com/examples/random-numbers.html

5 January 15: Plotting in Excel

Learning Objectives

- Create the following plots in Excel: scatter, histogram, bar, pie
- Add elements to plots: title, axis labels, trendlines, etc.
- Adjust axis ranges, bin sizes, and colors

Course Materials

- "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-functio html`
- "Function Errors" `http://www.excel-easy.com/functions/formula-errors.html`

6 January 16: Lab 2, Excel

Learning Objectives

- Finish a partially complete Excel file that demonstrates mastery of the following:
  - Creating the following plots: scatter, histogram, bar, pie
  - Adding elements to plots: title, axis labels, trendlines, etc.
  - Adjusting axis ranges, bin sizes, and colors

7 January 20: Dr. Martin Luther King Jr.'s Birthday Observed, No Class

8 January 22: Introduction to Stata and Reproducibility

Learning Objectives

- Locate and identify the essential parts of the Stata interface
- Create, edit, save, and load "log", .do, and .dta files
- Recall where to find syntax and other information on commands (`help`, StackExchange, etc.)
- Differentiate between the different data types in Stata, particularly different types of missing values

- Generate new variables and rename existing variables
- Use the following new functions/operators:
  - `help, set more off, cd, log, use, describe, sum, tab, list, in, gen, =, rename, recode, label`

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`

9 January 23: Lab 3, Stata

Learning Objectives

- Finish a partially complete .do file that demonstrates mastery of the following in Stata:
  - Cleaning data and generating variables using commands learned this week
  - Writing if statements and using Boolean operators

10 January 27: Data Cleaning in Stata and If Statements

Learning Objectives

- Assign values to variables using functions and logic statements (e.g. `mean` and `if`)
- Delete observations that meet certain criteria
- Use the following new functions/operators:
  - `egen, replace, if, keep, drop, missing, replace, sort, by, _n, _N, &, |, !, 1, 0, and ==`

Course Materials

- 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`

11 January 29: Graphics in Stata

Learning Objectives

- Create the following plots in Stata: scatter, line, bar, box, histogram
- Recall how to overlay multiple plots
- Add elements to plots: titles, legends, fitted-lines, etc.
- Interpret elements of plots after creating them (e.g. quartiles in box plots)
- Use the following new functions/operators:
  - `graph, twoway, scatter, line, bar, box, histogram`

Course Materials

- "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`

12 January 30: Lab 4, Stata

Learning Objectives

- Finish a partially complete .do file that demonstrates mastery of the following in Stata:
  - Merging data
  - Plotting bar graphs and scatter plots
  - Regression and interpretting coefficient estimates in a linear model
  - Generating residuals using `resid` and predicted values with `pred`

13 February 3: Regression in Stata

Learning Objectives

- Conduct basic regression analysis in Stata using `reg`
- Explain why one must be careful with linear form assumptions and out of sample extrapolation
- Distinguish causal effects from correlations between variables, and describe how naive regression is useful
- Analyze regression results and interpret key elements such as coefficient estimates and variance
- Construct a best fit line in a scatterplot and identify the slope, intercept, and residuals

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-regress`

14 February 5: Data wrangling in Stata

Learning Objectives

- Demonstrate appending and merging data
- Generate identifiers to differentiate between observations within a group
- Explain the difference between 1:1 and m:1 merges
- Collapse a dataset to a coarser unit of analysis
- Identify whether a dataset is long or wide and reshape it from one to the other

Course Materials

- "How to append files into a single dataset," StataCorp, `https://www.youtube.com/watch?v=AZGW8tohiqw`, 5 minutes.
- "How to merge files into a single dataset," StataCorp, `https://www.youtube.com/watch?v=niGZBRyyDuY`, 5 minutes.

15 February 6: Midterm

16 February 10: Introduction to R

Learning Objectives

- Locate and identify the essential parts of the RStudio interface
- Create, edit, and save .R and .RData files
- Generate objects and differentiate between datasets, numbers, strings, and functions
- Use the following functions:
  - `length, min, max, range, mean, sum, setwd, getwd, read.csv, load, write.csv, save, head, names, nrow, ncol, dim, summary, <-`

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

17 February 12: Analysis of Experiments by Subsetting Data in R

Learning Objectives

- Write logic statements in R and identify the relevant Boolean operators
- Generate subsets of data using logic operators and `$`
- Use the following new functions/operators:
  - `&, |, !, ==, sequence, class, as.class` (coercion), `is.class, c` (concatenate), `subset`

Course Materials

- Imai, 2.1-2.2

18 February 13: Lab 5, R

Learning Objectives

- Finish a partially complete .R file that demonstrates mastery of the following in R:
  - loading a dataframe, examining data, and calculating summary statistics
  - generating new objects including subsets of data using logical operators

19 February 17: For Loops and If Statements in R

Learning Objectives

- Describe how loops can reduce coding necessary to accomplish data analysis
- Construct for loops to accomplish simple tasks such as printing numbers 1 through 10 or calculating `n!`
- Define 'iteration' and give examples of how the iteration 'counter' can be used within a for loop
- Recall from the Excel lectures how to use the `if` operator and describe the syntax in `R`
- Use the following new functions/operators:
  - `if, for, else, in, print`

20 February 19: For Loops and If Statements in R

Learning Objectives

- Describe three ways how the iteration 'counter' `i` can be used within a loop:
  - As a number for calculations
  - As a subset index
  - As an element number of a vector (of numbers or strings)
- Build for loops that utilize the 'counter' `i` in all three ways
- Use the following new functions/operators:
  - `data, %%` (remainder function)

21 February 20: Lab 6, R

Learning Objectives

- Finish a partially complete .R file that demonstrates mastery of the following in R:
  - Using loops to repeat basic mathematical operations
  - Constructing if statements and for loops to create new objects

22 February 24: Visualizing Data in R

Learning Objectives

- Create the following plots in R: barplot, histogram, boxplot, line plots, and scatter plots
- Recall how to generate tables and which plots require tables as inputs
- Add elements to plots: titles, axis labels, ablines, text, colors, etc.
- Interpret elements of plots after creating them (e.g. quartiles in box plots)
- Use the following new functions/operators:
  - `barplot, hist, boxplot, plot, points, lines, table, ptable, abline, text,` and various plot parameters (e.g. `main, xlab, ylab,` etc.)

23 February 26: Regression in R

Learning Objectives

- Fit a linear model to data in R
- Produce regression results in R using `summary`

- Construct a best fit line in a scatterplot and add data labels
- Describe how regression can be used to determine the causal effect of treatment in an experimental setting
- Use the following new functions/operators:
  - `cor, lm, resid`

24 February 27: Lab 7, R

Learning Objectives

- Finish a partially complete .R file that demonstrates mastery of the following in R:
  - Perform regression analysis to determine linear relationships between variables
  - Interpret coefficient estimates and add best fit lines to scatter plots of the data

25 March 2: Functions in R

Learning Objectives

- Describe how functions can save time and space while writing code
- Construct functions that perform basic calculations, e.g. the mean of a subset of data
- Identify the inputs and output in a function
- Use the following new functions/operators:
  - `function, return`

Course Materials

- Chris Bosh on Functions, `https://www.youtube.com/watch?v=0eo0ESEX9DE`
- Imai, 1.3.4

26 March 4: Data Wrangling in R

Learning Objectives

- Download and install R packages (e.g. dplyr)
- Demonstrate appending and merging data
- Generate identifiers to differentiate between observations within a group
- Explain the difference between 1:1 and m:1 merges
- Collapse a dataset to a coarser unit of analysis
- Reshape data from wide to long and vice versa

27 March 5: Lab 8, R

Learning Objectives

- Finish a partially complete .R file that demonstrates mastery of the following in R:
  - Constructing functions that perform basic calculations on inputs including subsets of data
  - Constructing functions that produce plots
  - Constructing functions that have default inputs if none are provided

28 March 9: Creating beautiful plots with ggplot2

Learning Objectives

- Describe the three components of ggplot2: data, aesthetic mappings, and layers
- Demonstrate knowledge of ggplot2 by including at least one plot in your final presentation using this package

Course Materials

- Watch the first three videos in this playlist by DataCamp: `https://www.youtube.com/watch?v=YxKr2a-Y1WE&list=PLjgj6kdf_snaBCTJEi53DvRVgOuVbzyku`, 11 minutes total
- The other videos in the series are optional but may help spark some inspiration for your final projects.
- Check out example plots and code here: `http://r-statistics.co/Top50-Ggplot2-Visualizations-Maste`
  `html`

29 March 11: Zack's tips, tricks, best practices, and advice for the future

Learning Objectives

- Analyze data quicker and easier using some shortcuts
- Identify common pitfalls of data analysis including axis scaling, model dependence, missing value problems, etc.
- Recall other data analysis methods that one may learn and what they are used for
- Know where to find resources to further one's knowledge of data analysis

30 March 12: Lab 9, Final Project Workshop

- Use this time to ask your TA questions that pertain to your project
- If you have already submitted your project via Canvas you are excused from lab