This version: March 29, 2020

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

UC San Diego

Spring 2020 Online[1]

## Instructional Team

| Instructor | Virtual Office Hours | Zoom Meeting ID |
|---|---|---|
| Rachel Schoner | Monday 9 - 10AM | 365-985-274 |
| Arushi Kaushik | Tuesday 1 - 2PM | 919-591-953 |
| Chelsea Swete | Tuesday 3 - 4PM | 335-585-973 |
| Zack Goodman | Wednesday 9 - 10AM | 843-500-763 |
| Shane Xuan | Wednesday 3 - 4PM | 138-768-343 |

| Teaching Assistant | Virtual Office Hours | Zoom Meeting ID |
|---|---|---|
| Rachel Stephens | Mondays 11 - 12PM | 813-919-734 |
| Sabareesh Ramachandran | Tuesday 9 - 10AM | 588-379-313 |
| Sujith Chappidi | Thursday 12 - 1PM | 889-270-364 |
| Camila Navajas | Thursday 5 - 6PM | 903-996-114 |

## 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:

[1] This course will be held entirely online in response to the university's efforts to limit the spread of COVID-19.

  - Homeworks will be due one week after they are posted. Late submissions will not be accepted under any circumstances.
  - Each student will be allowed to drop their lowest homework grade to accommodate for special circumstances (i.e. three of four homeworks will count towards the final grade).
  - Copies of the homework write-up and accompanying code should be turned in via Canvas by the due date.
  - Although it is permissible to discuss conceptual questions with other students enrolled in class, each student must submit their own writeup of the solutions that shows their independent work on the assignment. In particular, one should neither copy someone else's answers or code nor share their answers or code with anyone. 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. Solutions/code that appear overly similar between students will be reported to the Academic Integrity Office.

- **Final Project(40%):** Students will complete an independent project that demonstrates mastery of the material taught during the quarter. The project will be due on Friday, June 12th at 11:00AM, but updates will be due throughout the quarter with homework submissions. See the final project prompt for specific details on the final project. Late submissions will lose a letter grade for every day (or part thereof) late. No submissions more than three days late will be accepted.

- **Class and Lab Exercises (20%):** Every week there will be two classes and one lab, each accompanied by an activity. It is expected that you attempt to complete the activities and submit them via Canvas, and they will be graded by effort. To accommodate special circumstances, the lowest two classes and lowest lab will be dropped.

  - *Classes*: There will be two classes per week that, each a recorded video accessible via Canvas. Each class will be accompanied by an activity that you should download from Canvas before beginning the video. While the video is running, you should follow along and complete the activity. After you've finished, you may upload your completed activity to Canvas for credit. The class activities will be due Thursday at 11:59PM every week so that you will be prepared for Friday's lab exercise.
  - *Lab*: There will be one lab per week held on Fridays. Each lab will involve completing an Excel workbook, Stata Do-file, or R script. Students are encouraged to attend a virtual lab session via Zoom Meeting where they may collaborate with other students on the exercises and ask for help from an instructor. During the virtual lab session, students will be assigned to "breakout rooms" in small groups with whom they may collaborate on completing the lab exercises. Each lab will build upon the materials taught earlier in the week, so it is expected that students complete the class exercises before attending a virtual lab session. Please "show up" (virtually) to the lab session on-time to facilitate assignment to breakout rooms. The lab will be due on Sunday at 11:59PM every week.

## Academic Honesty and Plagiarism

All of your graded work must be done by you. To be explicit, *sharing solutions or code are violations of academic integrity* and will be reported. 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. All assignments will be turned in via Canvas.

https://canvas.ucsd.edu/courses/15517

**Online Q&A.** Given the online format of this class, it will be immensely helpful to share questions and answers on a platform that all students may access. Throughout this class we will use the Piazza online discussion board. Piazza 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 images, attaching files, and customized email frequencies. We encourage you to ask questions on the Piazza forum for clarifications, conceptual questions, or figuring out methods related to your project. Please do NOT share code that would give away a solution to a homework assignment that has not yet been graded. You may join the Piazza page for our course directly from the below address (there are also free Piazza apps for the iPhone and iPad):

https://piazza.com/class/k87s0sglw3f2po

*Please do not email questions directly to the instructional team.* Others likely share similar questions, so it is preferred that you share your question publicly on Piazza. If your question is of personal nature (e.g. related to a grade), you may set the visibility of your post to be seen only by instructors and yourself.

## 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, Stata, and R. Excel and Stata both require licenses that are available for free to UCSD students. R is open-source and is free to everyone. Instructions on how to install the three software packages will be shared on Canvas.

**A note on recordings.** All lectures, labs, and office hours will be delivered online through Zoom. You can have access to the meeting links through the "Zoom LTI PRO" navigation tab in Canvas. You do not need to be physically on campus for any of the activities pertaining to this course. Lectures will be recorded before the week that materials are due, and you can watch the lectures at your own pace. Office hours and lab sessions will *not* be recorded. Please note that lectures and labs will be held either by Arushi, Chelsea, Rachel, Shane, or Zack, depending on which week it is. Each instructor will hold individual office hours each week, and you are free to attend any instructor's office hours as long as space permits. If you have personal questions related to the class, please reach out to your assigned instructor, which is listed on WebReg.

## Important Due Dates

- Both weekly class exercises due Thursdays at 11:59PM
- Weekly lab due Sundays at 11:59PM
- Homework 1 assigned April 5th, due April 12th at 11:59PM
- Homework 2 assigned April 19th, due April 26th at 11:59PM
- Homework 3 assigned May 10th, due May 17th at 11:59PM
- Homework 4 assigned May 24th, due May 31st at 11:59PM
- Final Project due Friday, June 12th at 11:00AM

## COURSE SCHEDULE*

*Subject to change.

## 1 Week 1, Class 1: 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 Week 1, Class 2: 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 observation 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 Week 1, 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 Week 2, Class 3: 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 Week 2, Class 4: 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 Week 2, 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 Week 3, Class 5: 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`

## 8 Week 3, Class 6: 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`

## 9 Week 3, 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 Week 4, Class 7: 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`

## 11 Week 4, Class 8: 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`

## 12 Week 4, Lab 4: Stata

### Learning Objectives

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

## 13 Week 5, Classes 9-10: 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.

## 14 Week 5, Lab 5: Stata

### Learning Objectives

- Finish a partially complete .do file that demonstrates mastery of the following in Stata:
  - appending, merging, collapsing, and reshaping data
  - working with data at different units of observation

## 15 Week 6, Class 11: 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

## 16 Week 6, Class 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

## 17 Week 6, Lab 6: 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

## 18 Week 7, Class 13: For Loops and If Statements in R part I

### 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`

## 19 Week 7, Class 14: For Loops and If Statements in R part II

### 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)

## 20 Week 7, Lab 7: 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

## 21 Week 8, Class 15: 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.)

## 22 Week 8, Class 16: 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`

## 23 Week 8, Lab 8, 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

## 24 Week 9, Class 17: 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

## 25 Week 9, Class 18: 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_snaBCTJEi53DvRVg0uVbzyku`, 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`

## 26 Week 9, Lab 9: 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

## 27 Week 10, Class 19: 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

## 28 Week 10, Class 20: tips, tricks, best practices, and advice for the future

### Learning Objectives

- 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

## 29 Week 10, Lab 10: 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