UCSB Data Hack 2022

Syllabus

Camilo Abbate and Michael Topper

Spring 2022

Effective policy is backed by data. This course will focus on developing the tools necessary to complete a research project which will result in a final presentation where students will convince a panel of professors/industry leaders why policy makers should focus on a particular issue.

Students will be *randomized* into small groups (3-5) at the beginning of the quarter. Each group will be responsible for one final presentation at the end of the quarter which will provide evidence for the following question: what issue should be prioritized by California's policy makers and why? This issue can be social, environmental, or political, but it is necessary that the issue's importance be grounded by data. Hence, while it is important to choose a topic of interest, it is also important to choose a topic of feasibility. As an example, rising crime rates may be a topic of particular concern. Data is readily available through police departments and Open Data Portals provided by major cities. Using this data your team would work together on cleaning/analyzing the data to create a presentation that could convince a policy maker to prioritize the rising crime rates. Some questions you may want to answer are: why are crime rates rising? Where are crime rates rising (e.g., urban/suburban/rural)? What is a possible solution? What is the cost/benefit analysis of your solution?

Learning Objectives

Completion of the course will result in a strong background in the R programming language for uses such as cleaning, analysis, presentation, and data collection methods. No prior R experience is assumed. Moreover, students will learn tools/methods that aid the workflow of research such as file organization/reproducibility, Zotero for literature reviews, and Freedom of Information Acts for data collection.

Course Structure

Class will be held once a week on Friday's from 2:00-4:00pm in the North Hall Computer Lab. Class sessions will consist of lectures, in-class group activities, and Q&A.

Homework

Homework assignments will be completely optional. However, it is recommended that you complete the homework assignments for particular skills that you want to become more fluent in. For example, if you find webscraping particularly useful/interesting, it would be optimal to complete the homework assignment. Homework assignments will not be graded, although solutions will (likely) be provided.

Grading

This class is pass or no pass.

Final Presentation

The final presentation will take place during the last class session (Week 10). Each group will have approximately 10-15 minutes to present their topic. Students will be judged based on a rubric created by both Michael and Camilo. A prize will be given to the top performing group.

Schedule

Note that topics can be excluded/extended based on student interest. In particular, if there are topics that are not listed here that students find interesting/want to know more about, please let us know and we can try to accommodate. Also note that after Week 3, we are open to changing the course schedule to accommodate skills needed for data collection. For instance, while Spatial Data is the topic for Week 4, we can easily skip to Webscraping if groups have a desire/need to gather their data using this method.

- **Week 1: Introduction to R and Piping**
  - Topics (R): Data types (characters/strings/logicals), importing data (excel/csv emphasis), common statistics (mean/sd/var), introduction to piping with mutating variables/summarizing/filtering/renaming.
  - Topics (Organization): Rprojects for file path management/collaboration, naming variables (e.g. snake_case and informative), the importance of READMEs, coding etiquette (leaving spaces, commenting etc.).
- **Week 2: Continuation of Cleaning**
  - Topics (R): Grouping, further filtering with conditions, cleaning dates with `lubridate`.
  - Topics (Organization): Using RMarkdown for reproducible documents.
- **Week 3: Visualization with ggplot2**

  - Topics (R): ggplot2 grammar of graphics, facet wraps, aesthetic mappings, reasons to use ggplot2.
  - Coding Check-in: Have us evaluate your code! Give us an example of a homework assignment or small project you completed and let us give feedback.

- **Week 4: Spatial data**
  - Topics (R): `sf` package, `usmap`, raster vs. vector data, heatmaps, `mapview`.
  - Check-in: Does every group have their topic/path to data? Should have data source identified by this point.

- **Week 5: Regular Expressions and Scraping Documents**
  - Topics (R): Using regular expressions along with extract, str_detect, str_replace, separate etc. PDF scraping techniques using `pdftools` and `tabulizer`.
  - In class activity: Scraping a PDF document using the tools discussed.
  - Quick presentation on the DO's and DON'Ts of visualizations.

- **Week 6: Functions, Packages, and Programming**
  - Topics (R): using functions to manage coding tasks and improve readability, if-statements, for-loops, creating packages to manage functions.
  - In class activity: create a package yourself!

- **Week 7: Webscraping**
  - Topics (R): `rvest`, css selectors, brief mention of `RSelenium`. For-loops and if-statements. Functional programming with `lapply/map`. Reshaping data with `tidyr`.
  - In-class activity of scraping data from a webpage and creating an effective visualization. Prize is a coupon for 1 redeemable handshake.

- **Week 8: Merging Data + Regressions + Presentation of Models**
  - Topics (R): merging using left join, right join, inner join, regressions with the `fixest` package, `modelsummary` and `kableExtra` for presentation purposes.

- **Week 9: Git/Github (optional)**
  - This topic will be presented based on student interest.
  - Check-in: issues with projects? Dedicated feedback to each team/progress reports.

- **Week 10: Presentations + Reception**