UCSB Data Hack 2024

Syllabus

Camilo Abbate and Michael Topper

Winter 2024

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) after the second week of class. 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, LaTeX for pdf generation and GitHub for version control.

Course Structure

Class will be held once a week on Friday's from 3:00-5:00pm in the North Hall 1110. Class sessions will consist of lectures, in-class group activities, and Q&A. Please bring your laptops to class. Tablets and chromebooks are discouraged (ok for notes), given that RStudio is not

usually available on these devices. Surface Pros are fine. Check the CANVAS site to install R and RStudio before the first class.

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 by a panel of both industry professionals and professors based on a rubric created by both Michael and Camilo. Prizes money will tentatively follow the following structure:

- 1st place: $500
- 2nd place: $200
- 3rd place: $100

While the rubric has not yet been created, be assured that presentation flow, clarity, visualizations, motivation, and comprehensiveness of argument are all important aspects of any presentation.

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 Webscrapping is the topic for Week 6, we can easily bring it before if groups have a desire/need for it.

- **Week 1 (January 12): Introduction to R and Piping**
  - Topics (R): R and RStudio. Libraries. 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 (January 19): More Data Cleaning**
  - Topics (R): Grouping, further filtering with conditions, cleaning dates with `lubridate`.

  - Topics (Organization): Using RMarkdown for reproducible documents.
- **Week 3 (January 26): Visualization with ggplot2**
  - Topics (R): ggplot2 grammar of graphics, facet wraps, aesthetic mappings, reasons to use ggplot2. Plotting heat maps, using fill, colours. If time allows: `usmap` and `sf` packages.
- **Week 4 (February 02): Merging and reshaping data**
  - Topics (R): Reshaping data with `tidyr`, `pivot_wider` and `pivot_longer`. Merging using left join, right join, inner join, rbind, cbind and alternatives.
  - Topics (Organization): How to save/organize and structure your datasets.
- **Week 5 (February 09): Regressions + Presentation of Models**
  - Topics (R): Regressions with the `fixest` package, `modelsummary` and `kableExtra` for presentation purposes.
  - Brief introduction to LaTeX.
  - Check-in: Does every group have their topic/path to data? Should have data source identified by this point.
- **Week 6 (February 16): Functions and Webscraping**
  - Topics (R): Using functions to manage coding tasks and improve readability. For-loops and if-statements. Functional programming with `lapply`/`map`. Extracting information from websites: `rvest`, css selectors, `RSelenium` and alternatives.

  - In-class activity: Design a function, scraping data from a webpage.
- **Week 7 (February 23): Coding with AI and Git/Github**
  - Using AI to help you code (Chatgpt and Copilot).
  - Topics (Organization): `git` and `github` for version control and collaboration.
- **Week 8 (March 01): Optional Class: Office Hours or another topic of interest**
  - This topic will be presented based on student interest.
  - Check-in: issues with projects? Dedicated feedback to each team/progress reports.
- **Week 9 (March 08): Presentations + Reception**