Final Project

Groups should apply one or more analytics tools from class to a real-world dataset, to either examine the relationship between 2 or more variables, or develop a prediction/classification model. If you find an existing published study with data available, you may also replicate the study's main findings, or discuss extensions.

If you are stuck on picking a topic, please let me know right away. I also find it helpful to read news sites like The Economist (healthcare), Five Thirty-Eight, Slate, Vox, etc. They will often do a simple analysis and mention a dataset, and you could dive deeper. Examining changes in laws or policies often present interesting opportunities to analyze (eg, Did banning single-use plastic bags increase food poisoning from reusable bags, using ER visits or Google search trends? Did California's mandatory vaccination law in 2015 change compliance more in rich or poor neighborhoods? Do state laws requiring birth control coverage reduce teen pregnancy rates?)

Possible analytic tools

- Multiple linear regression
- Difference-in-differences
- Classification using logistic regression
- Regression discontinuity

The overall goal of the project is to get practice using R on a real dataset. With any of these tools, you should also present some part of the data using ggplot (eg, histogram, scatterplot, line plot, bar graph). You do not need to run every type of analysis, or examine all possible hypotheses. It's better to have a good understanding of a narrower topic/question.

Suggested datasets

Below are publicly available datasets, but you are welcome to use a different dataset of your choice. You may want to merge multiple datasets (eg, California vaccination rates, household incomes) to gain deeper insights.

- COVID cases from New York Times
  https://github.com/nytimes/covid-19-data
- National Health and Nutrition Examination Survey (NHANES)
  https://wwwn.cdc.gov/nchs/nhanes/Default.aspx
- World Happiness Report
  https://worldhappiness.report/ed/2019/
- Medicare comparisons (physicians, hospitals, nursing homes, etc)
  https://data.medicare.gov/widgets/xubh-q36u
- California childhood vaccination rates by school, or disease cases by county
  https://data.chhs.ca.gov/dataset/school-immunizations-in-kindergarten-by-academic-year
- Humanitarian Data Exchange (global health, disease outbreaks, refugees, etc.)
  https://data.humdata.org/
- Google trends
  https://trends.google.com/trends/?geo=US

Deliverables

- During week 10, each team will give a **10-minute presentation** with 2-3 minutes for questions. I suggest creating about 10-15 slides (Background, Data overview, Methodology, Results, Conclusions).
- Teams should also submit any relevant **R code and data**, and this will help us assess what you did.