Final Project for Pathways in Data Science – Summer 2022

Day-by-day todos

Thursday, July 14: Brainstorming

- Find your groups!
- Come up with specific research questions
  - Good: How does time spent on videogames impact high school performance?
  - Bad: How much do people like dogs?
- Brainstorm possible datasets

Friday, July 15: Data Exploration

- Submit 1-2 paragraph tentative proposal, including at least one exploratory figure

Monday, July 18:

- Present two hypotheses and potential analysis methods you'll use to investigate them

Tuesday, July 19 – Wednesday, July 20:

- Work!

Thursday: Presentations

- 5 – 10 minutes

Project outline

The goal of the project is to go through the complete data science process to answer questions you have about some topic of your own choosing. You will acquire the data, design your visualizations, run a statistical or machine learning analysis, and communicate the results.

Your project should consist of a question or hypothesis that can be answered/analyzed using one of the methods we learned in this class. As a reminder, we covered:

- Linear regression
- Decision trees and random forests
- K-nearest neighbors
- K-means clustering

You are also welcome to try variations of the methods that we covered, including the ones that we may have brainstormed in class. However, if you do use a technique that we did not cover in class, you must demonstrate an understanding of the mathematics behind it in your writeup.

Deliverables

- **Project proposal:** one paragraph discussing the project goals. This will not be used for the project score - it is a way to check what the instructors think about your ideas.

Please indicate in the proposal the source(s) of the data used for the project. The proposal file should also have the group number and the names of the students in the group.

- **Presentation:** each group will present to the instructors and the rest of the class on the last day.
- **Notebook:** a high-quality and readable Jupyter notebook containing the code you used to do your analysis
- **Report:** a write-up describing your project. Your report should be no more than **4 pages** (1 inch margins, single spaced, 11pt font, but figures can exceed 4 pages in an appendix) and should include the following sections:
  - Overview: Summary of projects goals and the motivation for it
  - Related work: Anything that inspired you, such as a paper, a newspaper/magazine article, etc.
  - Initial questions: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?
  - Data: Source, scraping method, cleanup, etc.
  - Exploratory analysis: What visualizations did you use to look at your data in different ways?
  - Predictive analysis: Clearly state what predictive model you decided to use and why you decided to use it. For the method you chose to use, report appropriate error metrics
  - Conclusion: What did you learn about the data? How did you answer the questions? How can you justify your answers?

Data Examples

The following are some of the sources we use for data. You can use data there or you can use them for inspiration for project ideas. Using multiple datasets could enhance the analysis.

Google Dataset search: https://datasetsearch.research.google.com/
https://blog.google/products/search/discovering-millions-datasets-web/
CDC: https://data.cdc.gov/browse
500 cities: https://www.cdc.gov/500cities/index.htm
UN: http://data.un.org/
Kaggle: https://www.kaggle.com/datasets
AWS: https://registry.opendata.aws/
FEC: https://www.fec.gov/