Final Project

In the final project for this course, you will apply the techniques learned in this class to analyze a data set of personal interest to you. Your goal should be to create an original project that you would be proud to show off to a potential employer.

Requirements

You must work on this project with a partner (in a group of 2).

- The data that you analyze should be complex to collect or to clean in some way. All of the following would satisfy this requirement:
  - data that has to be scraped from a website or a REST API
  - textual data
  - geospatial data
  - data from multiple sources that has to be joined

A CSV file that you downloaded from Kaggle would *not* satisfy this requirement.

- Your analysis should tell a clear story through visuals. It is not enough to do data analysis; you must weave the analysis into a compelling story.
- You are encouraged to try fitting machine learning models, but only if it fits with the story you want to tell.

Then, you will turn your work into a poster. You don't have to print your poster. In addition to the poster, you must prepare a presentation to present in class.

You will also submit your poster, presentation, and code to Aybuzem.

Rubric

Criterion	**20 points**	**16 points**	**12 points**	**6 points**	**0 points**
Research Question	Interesting research question that could be the basis of a publication.	Clear, well-motivated research question.	Research question is fuzzy or not motivated.	Research question is not well defined.	No clear research question.
Data Collection	Data collection is extraordinarily complex.	Data collection meets the	Data collection was simplistic but	Superficial data collection (e.g.,	No data collection.

Criterion	20 points	16 points	12 points	6 points	0 points
		complexity requirement.	challenging in some way.	downloaded data set from Kaggle)	
Data Visualization	Unusually appealing and/or insightful visualizations.	Data visualizations were clean, labeled, and insightful.	Visualizations were technically correct, but not insightful.	Poor data visualizations that were incorrect (e.g., bar plot for a quantitative variable)	No visualizations were provided.
Data Analysis	Correctly applied a broad range of techniques from this class and perhaps a few beyond this class, in technically challenging situations.	Correctly applied a broad range of techniques from this class.	Applied techniques incorrectly, or applied only a limited set of techniques.	Data analysis was done, but the approach was fundamentally flawed.	No data analysis.
Storytelling	Weaved visualizations and analysis into a compelling story.	Visualizations and analyses told a coherent story.	Visualizations and analyses seemed scattered, with the main thread unclear.	Visualizations and analyses were not tied to a main thread.	No attempt to tell a story.
Real-World Application	Project generates insights with immediate real-world impact.	Project generates insights that clearly have the potential to be useful.	With some tweaking, project could have generated	The insights generated are not clearly useful.	No insights were generated from this project.

Criterion	20 points	16 points	12 points	6 points	0 points
			useful insights.		
Poster	Poster goes above and beyond.	Poster is clean, with a good balance of text and visuals.	Poster content is satisfactory, but a bit lacking in professionalism (e.g., too much text, blurry images).	Poster layout is sloppy.	No poster was made.
Presentation	Presentation was highly engaging and memorable. Fielded tough questions.	Gave a good summary of the poster and answered questions well.	Presentation was unclear, or speakers had difficulty answering questions.	Presentation was unclear, and speakers had difficulty answering questions.	Did not attend presentation session.
Peer Reviews	Completed required peer reviews and provided insightful feedback that even the instructors missed.	Completed required peer reviews and provided good feedback about each poster.	Completed required peer reviews, but provided perfunctory feedback.	Completed some, but not all peer reviews. Feedback was perfunctory.	Did not complete peer reviews.
Submission	Poster, presentation and code submitted on time, well-organized.				Did not submit poster or code.

Where to Find Datasets

The best data set is one that you are passionate about. I recommend that you start by finding a question you want to answer and then finding data to answer that question, rather than starting with a data set. That said, here are some helpful websites with large collections of data.

- Google Data Set Search https://datasetsearch.research.google.com/
- Reddit Datasets https://www.reddit.com/r/datasets/
- U.S. Government's Open Data https://data.gov/
- List of JSON APIs https://github.com/toddmotto/public-apis
- Project Gutenberg (good source of textual data) https://www.gutenberg.org/
- Data is Plural

Example Projects

Posters

- National Anthems Over the Ages: A Lyrical Timelapse
- "He Said, She Said": How Men and Women Converse in Movies

Github Repositories

- "He Said, She Said": How Men and Women Converse in Movies https://github.com/minako-m/datasci112_final_project
- Figure Skating https://github.com/abigailibarrola/data301-figure-skating
- Solar Panel Efficiency https://github.com/tmgerrit/Data301FinalProject