Final Project Guidelines

This page contains consolidated information for the final project.

Logistics and schedule may change slightly, so make sure to check this page for the latest info, as well as watch for announcements on Canvas and Piazza.

As you make teams and complete portions of the project, fill out your progress on our class spreadsheet.

Project Overview

The COGS 109 final project is meant to challenge students to apply data analysis techniques to real, large data sets to demonstrate their skills in data analysis, visualization, and communication.

The gist of the project is that you'll be doing some sort of novel analysis on a data set. **Your primary analysis should be regression or clustering**, but you may use additional analyses as well. The caveat is that you can't do an analysis that's been done before. For example, as there are documented examples of linear regression on the Iris data set, that wouldn't be a valid analysis for this project. We can further clarify on this in lab section if needed.

We ask that you work in groups of four for this project. Please use Piazza, Canvas, and lab section to find teammates. If you would like to work in a smaller group or on your own, please contact the TAs. Smaller groups will be held to the same expectations for the final project.

Project Objectives

- Identify the problems and goals of a real situation and dataset.
- Choose an appropriate analysis for the problem and the goals.
- Defend your choice of analysis by explaining how it helps to answer your research question.
- Implement one or more data analysis techniques for your dataset.
- Display data and/or results using appropriate visualizations, such as histograms, scatterplots, error calculations, etc.
- Interpret and evaluate the results of the analysis by stating whether the analysis was appropriate and whether the research question was answered.
- Communicate your results effectively to both experts and non-experts.
- Work effectively to manage a project as part of a team.

Analysis Pipeline (Outline)

This should be a bulleted outline of your project. This should be no more than a single page. Make

sure to include at least the following:

- A description of your dataset
  - What attributes are included in your dataset? What do they represent?
  - Provide a link to your dataset and mention any analyses that were previously performed on your data.
- How will you explore and visualize your data?
- Your group's research question
- Your planned analysis
  - What type of analysis? Any cross-validation? How will this model help answer your research question?
- Your planned report
  - How will you report your results? Will you report just a single model or something else?

Poster

For the poster, the recommended size is 36 inches by 56 inches, with text no smaller than ~28pt font (references can be smaller). You can use software like **Figma (https://www.figma.com/education/)** to collaborate on your poster remotely. Think of your poster as similar to what you might see at a research conference. The poster should include each of the following sections:

- Background
  - Dataset description: what the dataset consists of, description of variables, samples, and labels. The dataset should be included as a reference somewhere in the report.
- Methods
  - Description of methods chosen, why they were chosen, and how they work
  - This section may include supporting analyses in addition to the primary technique (regression or clustering)
- Results
  - Depiction of models being analyzed, including model equations (regression) or number of clusters (clustering)
  - Graphical representation of the results of your analysis
  - This section may include SSE values, the outcome of cross-validation methods, and/or plots depicting models
- Discussion
  - Interpretation of the results: Was the research question answered? If so, what is the answer and how do you know? If not, why not? Was the analysis technique appropriate for the dataset?
  - This section may include suggestions for future work or a description of data that could be analyzed to answer a particular question.

Written Report

The written report is not a "formal" report, but you should include all of the same sections that are in your poster (Background, Methods, Results, and Discussion). For your poster, you should omit Python code and excessive text so the content is easier for an audience to read. These should be included in your written report. If you choose to include your code inline (such as within a Jupyter notebook), make sure it is organized well. You may also attach your Python code at the end of your written report, but the same advice follows. Make sure your code is clear and easy to understand by using comments. Please cite any outside references used for your project.

Presentation

Details here are subject to change!

Your presentation may be done live over Zoom during finals week or pre-recorded and submitted by Friday of finals week. During your presentation you should present your poster. Presentations should be roughly **5 minutes.** If you present live, expect a couple minutes of questions.

Further Notes

Students may not perform an analysis that has been already been completed and published. For example, if there is a dataset on Kaggle that has a linear regression analysis posted online, students may not perform a linear regression project on that dataset.

If there are difficulties within a team (for example, you are unable to contact another student or you feel that a student's behavior is significantly affecting the group), please bring them to the attention of the TAs as early as possible. The project will be graded as a group grade except in specific circumstances.

Timeline and Deliverables:

Make sure to include a link to what is due on the Google Sheet in addition to submitting on Gradescope. Things that need to be turned in by the end of each week are in *italics*.

- Week 7
  - Start forming teams
- Week 8
  - *Analysis pipeline due*
  - *Finalize team and dataset on Google Sheet*
- Week 9
  - Continue working on project
  - Choose new dataset/analysis if previously not approved

- Week 10
  - Schedule presentation time if live presentation
- Finals Week
  - Submit all deliverables

Grading

- Analysis Pipeline: 10%
- Poster: 60%
- Written Report: 15%
- Presentation: 15%