COGS 108 - Final Project (Group)

Project Overview

The COGS 108 Final Project will give you the chance to explore a topic of your choice and to expand your analytical skills. By working with real data of your choosing you can examine questions of particular interest to you.

The broad objectives for the project are to:

- Identify the problems and goals of a real situation and dataset.
- Choose an appropriate approach for formalizing and testing the problems and goals, and be able to articulate the reasoning for that selection.
- Implement your analysis choices on the dataset.
- Interpret the results of the analyses.
- Contextualize those results within a greater scientific and social context, acknowledging and addressing any potential issues related to privacy and ethics.
- Work effectively to manage a project as part of a team.

To accomplish this you will work in teams of 4 to 5 students to conceive of and carry out an analysis project. ***Everyone who chooses this option must be part of a group.*** You will find in your future careers the need to work on projects in groups frequently (even if you really, really, really, really don't want to).

The basic project steps:

- Find a real world dataset and problem that you believe can be solved with one or more of the techniques we have learned in class.
- After selecting a dataset and identifying the goal, write out a proposed analysis plan using template provided and submit it through GitHub for review.
- Apply the techniques outlined and come up with a result for the dataset that you proposed.
- Assemble a Jupyter notebook that communicates your hypothesis, methods, and results. Submit this as your final project.
- Submit feedback about group and individual group members. This is done individually.

Taboo topics

While you are *encouraged* to pick a topic that helps our understanding of the world/society, you are free to pick *almost any* topic for your final project. That said, there are a few off-limit (or "taboo") topics. These topics are off-limits for three reasons: (1) these projects have been done a whole bunch of times on Kaggle already, which (2) limits the amount you can learn, and (3) makes these projects super boring to read and grade. We encourage you to pick a topic you care about, not one that is the easiest to find on the Internet.

The taboo topics in COGS 108 are:

- movie recommendation system
- youtube video comments / trending analysis

If you think you have a novel idea and *really, really* want to work on any of these topics, reach out to Prof Ellis, and she'll let you know if your idea will work.

The following datasets from kaggle are also off limits as your main dataset*:

- Trending Youtube Video Statistics
- 120 Years of Olympic history: athletes and results
- Spotify Song Attributes
- Wine Reviews
- Yelp Dataset

*If you use any of these to *support* your main analysis, that is fine. But, it cannot be your main source of information.

Getting Started

We strongly encourage you to discuss potential project ideas on Piazza, with your TAs and IAs, and/or with Prof. Ellis! This will give us a chance to make sure you're on the right track even before you submit your draft.

The Project Proposal

The Project Proposal is completed as a group. The Proposal template can be found here. Your proposal must include the following sections:

NAMES & IDs: Be sure to include each member's name and ID where it's asked for in the project notebook

- Names : Replace the lines to list each person's full name. Add lines as needed for your group size, and make sure each name is listed on a separate line.
- Group Member's IDs : Replace the lines below to list each person's student ID. Add lines as needed for your group size, and make sure each name is listed on a separate line.

RESEARCH QUESTION: What is your research question? Include the specific question you're setting out to answer. (1-2 sentences)

BACKGROUND & PRIOR WORK: It will present the background and context of your topic and question in a few paragraphs. Describe what information you all currently know about the topic, include references to other projects who have asked similar questions or approached similar problems. Explain what others have learned in their projects.

- Why is this question of interest to your group?
- What background information led you to your hypothesis.
- Why is this important?
- What has already been done on this topic? What is already known?

Find some relevant prior work, and reference those sources. Even if you think you have a totally novel question, find the most similar prior work that you can and discuss how it relates to your project.

References can be research publications, but they need not be. Blogs, GitHub repositories, company websites, etc., are all viable references if they are relevant to your project. It must be clear which information comes from which references. (2-3 paragraphs, including at least 2 references)

HYPOTHESIS: What is your main hypothesis and predictions? Briefly explain why. (2-3 sentences)

DATA: Here, you are to *think* about the *ideal* dataset (or datasets) you you would need to answer this question. Describe and explain the ideal datasets you would want to answer your question.

- What variables would you have?
- How would they be stored?
- How many observations would you have?
- What/who would the observations be? etc.)
- etc.

Note: For the project proposal, you do not have to find the actual dataset(s) needed for your project. For your final project, you will.

ETHICS & PRIVACY: Acknowledge and address any ethics & privacy related issues of your question(s), proposed dataset(s), and/or analyses. Use the information provided in lecture to guide your group discussion and thinking. If you need further guidance, check out Deon's Ethics Checklist. In particular:

- Did you have permission to use this data / use it for this purpose?
- Are there privacy concerns regarding your datasets that you need to deal with, and/or terms of use that you need to comply with?
- Are there potential biases in your dataset(s), in terms of who it composes, and how it was collected, that may be problematic in terms of it allowing for equitable analysis? (For example, does your data exclude particular populations, or is it likely to reflect particular human biases in a way that could be a problem?)
- Are there any other issues related to your topic area, data, and/or analyses that are potentially problematic in terms of data privacy and equitable impact?
- How will you handle issues you identified?

(1-2 paragraphs)

TEAM EXPECTATIONS: Read over the COGS108 Team Policies individually. Then, include your group's expectations of one another for successful completion of your COGS108 project below. Discuss and agree on what all of your expectations are. Discuss how your team will communicate throughout the quarter and consider how you will communicate respectfully should conflicts arise. By including each member's name above and by adding their name to the submission, you are indicating that you have read the COGS108 Team Policies, accept your team's expectations below, and have every intention to fulfill them. These expectations are for your team's use and benefit — they won't be graded for their details.

PROJECT TIMELINE PROPOSAL: Specify your team's specific project timeline. An example timeline has been provided. Changes the dates, times, names, and details to fit your group's plan.

If you think you will need any special resources or training outside what we have covered in COGS 108 to solve your problem, then your proposal should state these clearly. For example, if you have selected a problem that involves implementing multiple neural networks, please state this so we can make sure you know what you're doing and so we can point you to resources you will need to implement your project. Note that you are not required to use outside methods.

To reemphasize: for the Project Proposal you are not expected to have already done any analyses for the proposed project, but what you submit should be a plan for what you will answer and what data you would ideally have for the project. (Of course, for the final project you will need to actually find the data and do the analyses.)

Project Proposal - Style Guidelines

The proposal should be written as if to a fellow student. You may assume that your audience is familiar with the material we have covered as a class this semester.

This is a short proposal meant to give us time to assess and critique your Final Project (further described below), in order to give you time to improve upon it before your Final Project.

You will receive feedback on your project proposal, and you are fully expected to make the changes suggested by the Professor, TAs, IAs, and your classmates on this assignment before submitting your Final Project.

Remember to proofread your Project Proposal. Do not use overly flowery and/or vague language.

After the Proposal: Working on the Problem

Once you've settled on a problem and approach, it's time to actually find and analyze the data!

Note: It is very important that you get right to work on the problem and don't procrastinate. This is not a homework set — this is a large, complex problem that will take concerted effort to complete.

Final Project

The main product of the project is a single Jupyter Notebook, submitted on GitHub. You can find the template on GitHub, in the Projects directory. You will be graded on the one group notebook submitted on GitHub in your group's project repo (which will be created for you). **Change the name of the file to include your group's group number. (For example, if you were in group 001, your file would be named 'FinalProject__group001.ipynb'.)**

This single notebook should include all the code you used for all components of the project (cleaning, visualization, analysis). Because we won't be running the code in your notebook, it is important to make sure your notebook as submitted to GitHub has the code evaluated and outputs present (e.g., plots) so that we can read the project as is.

Submission must be successfully completed by 11:59 PM on the date of your final, and should be self-contained, so that we can evaluate your entire project from the notebook alone. Additionally, each individual group member must complete the team and individual feedback survey.

These notebooks will optionally be opened to the general public, so others may read what you've done! You will have the option to opt out of making your project public.

Final Project Sections - Instructions

Each of the following sections corresponds to a section in the FinalProject__group000.ipynb Jupyter notebookfound on GitHub in the `Projects` directory.

For sections included in your proposal, you can copy and paste into your final project, but be sure to edit these sections with feedback you received on your proposal or additional information you learned throughout the project.

PERMISSIONS: Specify whether you want your group project to be made publicly available. Place an X in the square brackets where appropriate.

OVERVIEW: Include 3-4 sentences summarizing your group's project and results.

NAMES & IDs: Same as proposal.

RESEARCH QUESTION: See proposal specifications.

BACKGROUND & PRIOR WORK: See proposal specifications.

HYPOTHESIS: See proposal specifications.

DATASET(S): What data will you use to answer your question? Describe the dataset(s) in terms of number of observations, what kind of features it contains, etc. (Typically students have datasets of ~1000 observations across their datasets. This is not a requirement, but is good to know around the scale of data we're expecting.) You are welcome (and in fact recommended) to find multiple datasets! If you do so, describe each one, and briefly explain how you will combine them together. Include the source of the dataset in the description here.

SETUP: Include packages used for analysis in cell provided

DATA CLEANING: What methods did you use to analyze your data? Briefly explain what steps you had to take before you were able to use the datasets you chose to answer your question of interest.

- How 'clean' is the data?
- What did you have to do to get the data into a usable format? (If you did nothing, how did you determine there was nothing to do?)
- What pre-processing steps that were required for your methods (for example, checking data distributions and performing any transformations that may be required)

DATA ANALYSIS & RESULTS: This section should include markdown text and code walking us through the following:

- EDA
  - What distributions do your variables take?
  - Are there any outliers?
  - Relationship between variables?
- Analysis (Note that you will likely have to do some Googling for analytical approaches not discussed in class. This is expected for this project and an important skill for a data scientist to master.)
  - What approaches did you use? Why?
  - What were the results?
  - What were your interpretation of these findings.
- Data Visualization - There must be at least three (3) appropriate data visualizations throughout these sections. Each visualization must included an interpretation of what is displayed *and* what should be learned from that visualization. Be sure that the appropriate type of visualization is generated given the data that you have, axes are all labeled, and the visualizations clearly communicate the point you're trying to make.

ETHICS & PRIVACY: See proposal specifications.

CONCLUSION & DISCUSSION: Discuss your project. Summarize your data and question. Briefly describe your analysis. Summarize your results and conclusions. Be sure to mention any limitations of your project. Discuss the impact of this work on society. (2-3 paragraphs)

Grading

The final project is worth 35% of your grade (as noted on the course syllabus). 8% of this is from your project proposal. 2% for your project survey. The other 25% is based on your project notebook that you submit by the day of your final exam.

Your project will be graded based on the rubric below. Make sure you address each rubric section in the notebook, in an organized manner, using cell Markdowns for textual descriptions.

The grading rubric for the Final Project is as follows:

Category	Percentage of Project Grade
Overview, Question, & Background	10
Data Description	10
Data Cleaning/Processing	10
Data Visualization	15
Data Analysis & Results	25
Ethics & Privacy	15
Conclusion & Discussion	15

Note: Individual grades can be adjusted based on the feedback provided in individual surveys submitted. This means that team members in the same group can receive different scores from one another, if evaluations suggest that contributions were not evenly distributed. To avoid this, work together as a group and ensure that you're contributing to the project.

Timeline

To make sure we are all progressing well toward the end of the project, use the following timing guidelines as a way to help your team set deadlines:

Week 1: This document is released
Week 2-3: Groups are formed; Topic decided
Week 4-5: Project Proposal Due

Week 6: You have met in a group at least twice and have begun EDA.
Week 7-8: Analysis should be underway
Week 9: Analysis is mostly complete; group has met ~3-4 times
Week 10: Project being edited; small improvements being made; Time is provided to work on projects in sections.
Finals Week: Due Date for all projects and team evaluations. Submit final Jupyter notebook on GitHub.

Advice

The main pieces of advice are:

- Start early
- Work consistently
- Be a good teammate
- Work as a team
- Seek advice when you are unsure, and see it early and often!
- Use Piazza
- Talk to your TAs, IAs, and Prof. Ellis - we're here to help!
- Choose a general interest domain, but then choose a dataset and decide on a problem, not vice versa. I promise it will go much better.
- Start early!!!!!

As far as resources go, it is okay to ask other teams what they are doing in terms of sources, presentation plans, and so on. As long as you are not using another team's work and claiming it as your own, collaboration with classmates is encouraged. If you find a good source of datasets, please share with everyone on Piazza!

Previous COGS 108 Final Projects

See Prof. Voytek's write-up of excellent class projects from the Spring 2017 instance of COGS 108 here, all of which received perfect scores.

Additionally, previous projects can be viewed from when this course ran in Spring 2017, Winter 2018, Spring 2019, Fall 2019, or Winter 2020. Note first, that these projects are of variable quality and second, that if you get inspiration or code from previous projects, this must be noted in your project, giving attribution to the former groups' work.

How to Find Datasets

The purpose of this project is to find a real-world problem and dataset (or likely, datasets!) that can be analyzed with the techniques learned in class and those you learn on your own. It is imperative that by doing so you believe extra information will be gained — that you believe you can discover something new!

You must use at least one dataset containing at least approximately 1000 observations (if your data are smaller but you feel they are sufficient, chat with Prof. Ellis). You are welcome (and in fact recommended) to find multiple datasets!

The best datasets are the ones that can help you answer your question of interest.

Your question could be just for fun: Using text mining of song lyric websites to identify the most commonly used phrases and sentiments by decade.

Your question could be scientific: Scrape data from animal taxonomies and Wikipedia to figure out if larger animals are more likely to be carnivores?.

Or, ideally, your question can be aimed at civic or social good, for example, use mapping, transit, and car accident data to identify which parts of San Diego are most in need of dedicated bike lanes.

To help you find datasets, we have collected a list of websites that have a considerable number of open source data sets and included them at the end of this document..

Eventually you will all have to decide on a problem to tackle, with each member of the team having a clear, delineated role in the project.

Dataset Resource List

Here, is a list of potential locations to find datasets and problems to investigate. If you have another dataset or search location, that is great!

- Awesome Public Datasets
- Data.gov
- Data Is Plural
- UCSD Datasets
- Datasets | Deep Learning
- Stanford | Social Science Data Collection
- Eviction Lab (email required)
- San Diego Data
- US Census
- Open Climate Data
- Data and Story Library
- UCSD behavioral mobile data
- Kaggle
- FiveThirtyEight
- data.world
- Free Datasets - R and Data Mining
- Data Sources for Cool Data Science Projects