COGS 108 - Data Science in Practice
Final Project

Project Overview

The 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 these goals, you will work in teams of 4 to 6 students to conceive of and carry out an analysis project. Groups will be assigned during week 3 of the course.

Everyone 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 Gradescope for review (due Sunday, October 27 at 11:59 PM).
- 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 (this is the final product due the date of your final at 11:59 PM).
- Submit feedback about group and individual group members. This is done individually by every member of the group.

Project Teams

Teams will be assigned based on survey feedback.

Familiarize yourself with COGS 108 Team Policies document and note that no changes to teams will be made after week 7.

Getting Started

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

How to Find Datasets

The purpose of this project is to find a real-world problem and dataset that can be analyzed with the techniques learned in class. 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, email Prof. Voytek). You are welcome (and in fact recommended) to find multiple datasets! Think about Prof. Voytek's comments about the power of integrating heterogenous datasets.

- 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.
- You 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. *(Big credit here to Jeremy Karnowski from Insight Data Science)*.

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.

The Project Proposal

In addition to including your team name and team members at the top, the Project Proposal is a document that includes information about the following things:

1) **QUESTION & HYPOTHESIS**: Explain the specific data science question you're attempting to answer or problem you're trying to solve. Explain what you expect the answer to the question or solution to the problem and briefly why.

2) **BACKGROUND**: It will present the background and context of your dataset 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.

3) **ETHICAL CONSIDERATIONS**: Acknowledge and address any ethics and privacy related issues of your question(s), dataset(s), and/or analyses

4) **DATA**: Identify the source(s) of the data you will *actually* be analyzing. Include explanations of the datasets you've found or data that you will collect to answer your data science question. This should include information about the number of observations, variables included, and source of the data. You may well need more than one dataset for this project. Explain all the datasets you may use in this section. If ,in the process of doing this project you realize these data won't work or you'll need additional datasets, you're allowed to change from what you propose here. But, we want to see you've identified at least a dataset or two that will likely help you answer your question.

5) **TEAM EXPECTATIONS**: Read the COGS108 Team Policies and then state your team's specific expectations that all members agree upon.

To get started, one member of each team will first make a copy of the COGS 108 Project Proposal Word document template in the COGS10/Projects repo. When complete, save this Word doc as a PDF and submit the PDF to Gradescope. **Be sure all team members are added to the submission on Gradescope**.

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'll use for the project*. (Of course, for the Final Project you *will* need to actually do the analyses.)

10% of your course grade is from your project proposal, and 35% of the course grade is from your final project. Together, these will account for 45% of your final grade.

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 criticize 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 and 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 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 Details: Jupyter Notebook & Submission

The main product of the project is a single Jupyter Notebook, submitted in PDF form. You can find the Jupyter template in the COGS108/Project GitHub repo. You can work on your project how you wish (although we recommend you use version control and GitHub), but ultimately you will be graded on the one notebook for the whole group. The Final Project notebook is to be submitted electronically as a PDF on Gradescope and pushed to GitHub (details on this later). That is, **one person from your group will upload the PDF to Gradescope and add all members of the group on the Gradescope submission *and you'll have* push your final Jupyter Notebook to your Project Group's private GitHub repo (which we will create and assign to you)**.

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 the PDF of your notebook as uploaded to Gradescope and the notebook 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 may 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.

Grading

The Final Project is worth 45% of your grade (as noted on the course syllabus). 10% of this is from your Project Proposal. The other 35% 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
Introduction and Background	10%
Data Description	10%
Data Cleaning/Pre-processing	10%
Data Visualization	15%
Data Analysis & Results	25%
Privacy/Ethics Considerations	15%
Conclusion & Discussion	15%

Note: Individual grades *can* be adjusted based on the feedback provided in individual evaluations submitted. This means that team members in the same group can receive slightly 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 you 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 4: Project Proposal Due.

Week 5: Team Evaluations are released.

Week 6: You have met in a group at least twice and have begun exploratory data analysis.

Week 8: Analysis is mostly complete; group has met 3-4 times

Week 10: Time is provided to work on projects in sections.

Finals Week: Due Date for all projects and team evaluations. Submit final Jupyter notebook as a PDF to Gradescope *and* on GitHub. Be sure all group members are added on both the notebook and Gradescope submission.

Resources and 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 seek it early and often!
- Use Piazza
- Talk to your TAs, IAs, and Prof. Voytek—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 Class Projects

See Prof. Voytek's write-up of excellent class projects from the Spring 2017 instantiation 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, and Spring 2019. Note that these projects are of variable quality and that if you get inspiration or code from previous projects, this *must* be noted in your project, giving attribution to the former groups' work.

Example External Projects

Note these aren't civic/social good focused or things specifically completed for this course, but they are fun examples of what can be done with publicly available data.

- Most Trendy Names in US History
- The Largest Vocabulary in Hip Hop
- A Map of Where NFL Quarterbacks Throw the Ball
- Every Shot Kobe Bryant Ever Took

Dataset Resource List

Below 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

Final Project Section Instructions

Each of the following sections corresponds to a section in the *FinalProject.ipynb* Jupyter notebook found in the COGS108/Projects GitHub repo. One member of your team should make a copy of this notebook. Then, everyone should work on this copy of your team's project collaboratively. **GitHub and proper version control is the easiest way to do this successfully**.

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

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? (1-2 sentences)

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

BACKGROUND & PRIOR WORK: Why is this question of interest to your group? What background information led you to your hypothesis. Why is this important?

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. (2-3 paragraphs, including at least 2 references.)

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. You must use at least one dataset containing at least approximately 1000 observations (if your data are smaller but you feel they are sufficient, email Prof. Ellis). 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?
- 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.

There must be <u>at least three appropriate data visualizations</u> throughout these sections. Each visualization must include 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:

Briefly acknowledge and address any potential issues of ethics and privacy for the proposed project. In particular:

- Did you have permission to use this data, 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 did you handle issues you identified?

(1-2 paragraphs)

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)