COGS 108 - Final Project (Individual)

Project Overview

The COGS 108 Final Project (Individual Option) will give you the chance to carry out a common practice used in data science interviews - the take-home analysis. This will require you to formulate a question around a topic and dataset provided to you during Finals Week.

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.
- Explore the dataset and 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.

You will work individually to accomplish this. You will have access to course materials and the Internet, but you will not be allowed to discuss this with other members in the class. You will have 5 days to complete the final project.

The basic project steps:

- Understand the dataset and topic provided to you. Formulate a question using these data that you believe can be solved with one or more of the techniques we have learned in class.
- After understanding the dataset and identifying the goal, set out a plan for answering this question.
- 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 your experience completing the individual final project.

Getting Started

The Project Proposal

The Project Proposal is completed individually. For students who opt for the individual final project, a topic will be provided to you. Your proposal will have to be written on the specified topic.

The Proposal template can be found here.

Your proposal must include the following sections:

NAME & ID: Be sure to include your name and ID where it's asked for in the project notebook

RESEARCH QUESTION: What is your research question? Include the specific question you'd set out to answer, given the topic. (1-2 sentences)

BACKGROUND & PRIOR WORK: Research and present background information and context of your topic and proposed question. Describe what information is currently known about the topic. Include references to other projects who have asked similar questions, approached similar problems, and/or have carried out work on the topic. Explain what others have learned in their projects.

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/or 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 the question you've proposed. 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 this project proposal, you do ***not*** have to find the actual dataset(s) needed to carry out this project.

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)

Project Proposal - Topic

General Topic: Health Inspections

Prompt: Imagine you work for a local city government and want to improve the restaurant health inspection process for the city government.

Task: Formulate a strong data science question (or set of questions) around this topic and carry out background research to better understand what others have done previously. Use the provided template, and inlcude this information in a strong Project Proposal.

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.

Final Project

The main product of the final 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 notebook submitted on GitHub in the Individual submissions Group repo (to which you will be given access during the quarter). **Change the name of the file to include your PID. (For example, if your PID is A1234567, your file would be named 'FinalProject_A1234567.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.

Final Project Sections - Instructions

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

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

NAME & ID: See proposal description.

RESEARCH QUESTION: See proposal descriptions.

BACKGROUND & PRIOR WORK: See proposal description.

HYPOTHESIS: See proposal description.

DATASET(S): What data will you use to answer your question? Describe the dataset(s) provided in terms of number of observations, what kind of features it contains, etc. If you are provided and use more than one dataset, 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 may 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 description.

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

Timeline

Dataset and topic will be made available to students by Friday at 11:59 PM of Week 10. Students will have 5 days to complete their individual final project. We do not anticipate it taking you 5 days straight; however, we know you'll have other finals to study for and take during this time. Final Jupyter notebook will be submited on GitHub.