Final Project

The 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.
- Work effectively to manage a project as part of a team.

To accomplish this, you will work in teams of 3 to 4 students to conceive of and carry out an analysis project. We will form your groups randomly based on the number of students enrolled in the class. Everyone who will be part of some group**.** You will find in your future careers the need to work on projects in groups frequently (even if you really, really don't want to). Larger Goal is to give your more hands-on practice to the concepts you will learn in the lessons and can apply that knowledge to new challenges.

The Project will be completed in multiple parts due separately:

Phase I: Propose a dataset and purpose of analysis due Friday, February 10, 2023, before midnight.

Phase II: EDA and Insights -TBA

Phase III: Modeling - TBA

Part I:

Propose a data analysis project. This can be almost anything that you choose. You might select a project from your field of study, your extracurricular interests, government, or public policy, or elsewhere. We strongly encourage you to discuss potential project ideas with your TAs and IAs, and/or with Professor. This will give us a chance to make sure you're on the right track even before you submit your draft.

You must use at least one dataset containing at least approximately 1000 observations (rows) (if your data are smaller but you feel they are sufficient, Talk to Professor) and 5-10 attributes (columns) with the mix of both numerical and categorical variables.

The main purpose of the proposal is for us to give feedback on whether the scope of the project is in the range of what we're expecting. On average we expect proposals to be about 1-2 pages long, though we know the lengths will vary.

Your report should contain at least the following parts. You are permitted to write additional sections as well.

Title and author(s). The title should describe your research questions. It should not be something like "CSCE 5310 Final Project", your title should self-explanatory.

Summary of research questions and results. Give a numbered list of one or more research questions. Each one should be a specific question with a specific answer, not merely a general topic or area of investigation. In 1-3 sentences per research question, state what you are trying to compute, and why.
After each research question, clearly state the answer you determined. Don't give details or justifications yet — just the answer.

Motivation and background. Explain the context and why the problem matters. This expands on the research questions that you already stated. Why are they worth computing? What difference would knowing the answers make? We require a problem with some kind of real-world motivation.

Dataset. Describe briefly the real, existing dataset that you used, including exact URLs. You may not use a dataset that has been used by other groups or are already used by someone. You may not repeat an analysis that you have performed in another class (though you might do something inspired by another class). The data must be real — neither you nor someone else may make up the data.
You do not have to turn in the dataset itself (it may be quite large, or it might be available only via the web rather than as a single download). Your program might access the data directly via the web. If your program needs to access the data through the file system, then ideally, when your program is run, it should automatically check whether the required data files are present, and if not download them before doing any additional work. Alternately, your report must include simple, clear, unambiguous instructions that anyone can follow to download the data themselves.

Do not use a dataset that cannot be shared with the course staff, such as one that contains confidential medical information or intellectual property.

What To Turn In

Your proposal should be in a pdf document named **group#_proposal.pdf**. Include clearly at the top of the document the **name**(s) and **SUID**(s) of all the group members submitting the proposal. Upload the pdf document to Canvas.