CS 6053 - Foundations of Data Science
Final Project

Brief Summary

You will be working in groups of 3 members with the goal of completing a causal inference-based data analysis using concepts learned during the semester in this course. This project will be evaluated based on the following 3 components:

Component	**Due Date**	**Percent of Final Grade**
Proposal	November 21st, 2023, 11:59 PM EST	5%
Data Communication/ Screencast	December 13th, 2023, 11:59 PM EST	10%
Write-up	December 13th, 2023, 11:59 PM EST	25%

Core Requirements

While the example problems considered in this course have been from a variety of domains and have utilized different models, a fundamental approach has been pursued each time:

1) Identify a question/estimand
2) Describe a scientific/causal model
3) Define a corresponding statistical model
4) Validate the model on simulated data with known parameter values
5) Analyze the real data

You will conduct this 5 step process on a real dataset of your choosing. There are obvious limitations about how in-depth a project can be when one only has part of a semester to complete the project. Therefore, the expectation is not that your end result will be something that would be a finished product for a company or publishable results. However, your group will need to complete each of the 5 steps described above.

There are a few additional requirements:

- Project must involve causal inference
- The causal model must include atleast 3 variables including a predictor variable, outcome variable, and one potential confound
  - It is ok to define a causal model that considers more than 3 variables
- Project code and analysis will be submitted on Gradescope
- A code walkthrough in the form of a screencast is required

  - Only group members voices are required (no need to include videos of yourselves presenting your work)
  - Each group member must be involved in the screencast
    - Introduce yourself when you begin presenting
  - Screencast should be a maximum of 15 minutes in length

Choosing a Project

The basic consideration for any project is a question of interest. A music streaming service might ask why certain customers cancel subscriptions after 6 months of using the service. A computational biologist might be interested in understanding why certain individuals tend to have severe reactions to a medication while others suffer no ill effects. A public health administrator might want to understand what messaging works best for encouraging individuals to return for a second dose of a vaccine sequence. The number of questions that can be answered (assuming data exists to help provide an answer) are endless.

You may not have any burning questions that you would like to tackle. If you do, this is a great opportunity to put what you have learned in this course into practice in order to investigate and, hopefully, provide some answers to a question of interest using data. If not, here are some possible data sources for a causal inference project:

Twins Dataset:
https://github.com/AMLab-Amsterdam/CEVAE/blob/master/datasets/TWINS/ReadmeTwins

National Study of Learning Mindsets:
https://github.com/grf-labs/grf/tree/master/experiments/acic18

Jobs Dataset:
https://paperswithcode.com/dataset/jobs

ChatGPT Advice:
https://github.com/petezh/ChatGPT-Advice

ICLR Reviews:
https://cogcomp.github.io/iclr_database/

News-Tweet Dataset:
https://github.com/bywords/NTPairs
Warning: There may be a good deal of data engineering work needed to get access to the data

Various R datasets:
https://cran.r-project.org/web/packages/causaldata/causaldata.pdf
Note: If you are unfamiliar with R and need help getting access to the data, ask for assistance

These are just some suggestions if you have no idea where to begin in identifying a dataset. Keep in mind that these datasets are often related to work that has already been published. Do not simply re-do the analysis that was completed by someone else. If you want to find inspiration from a previous analysis, feel free. But be sure to document what source you are using for that inspiration as well how your analysis builds on/differs from what was done previously.

If you already have your own question and want to use a data source other than one mentioned above, feel free. This project is your opportunity to tackle a problem of interest and use knowledge that you have gained from this course to do so.

Evaluation
Expectations for each component of the project are listed below

Proposal Requirements:

- 1 page (maximum)
- A clear question to be answered/estimand identified
- Identification and description of the data to be used
  - Each group must work on a unique dataset
  - Once a dataset has been selected, add it along with your group members names to this Google Doc to ensure no other group selects the same dataset
- Causal model in the form of a directed acyclic graph (DAG)
  - Ok if it changes in the final submission
- A proposed statistical model
  - Ok if it changes in the final submission
- Submitted in PDF format on Gradescope
- Feedback will be provided within a week after proposal due date

Screencast Requirements:

- Approximately 15 minutes
  - Exceeding this limit by more than 2 minutes will result in a grading penalty
- **Each group member must speak during the presentation**
- Required information
  - Clear explanation of the question/estimand investigated
  - Description of the data and the variables included in the causal model
  - Any scaling/transformations performed on the data
  - Brief discussion of the statistical model used and how it was validated
  - Answer to the question/summary of what was learned

Deliverable Requirements:

- At a minimum, a well-organized and detailed Jupyter Notebook (ipynb and pdf versions)

Please ask questions about requirements/expectations early and often. A discussion forum using edstem's Ed Discussion tool will be used to answer final project questions as they arise. You can find a link to this discussion forum under **Content** > **Final Project** in Brightspace.
Good Luck!!