The course project is an opportunity to engage with the course content via a real-world example. Over the course of the semester, you will walk through an entire causal analysis: starting with defining a causal question all the way to communicating the results of your analysis. Parts of this project will be done as individuals and parts will be completed as a group. We will assign parts of the project throughout the semester and we will give detailed instructions each time. We will also have plenty of check-ins along the way to make sure you're on the right track.

1 Project Overview

1.1 Individual tasks

You will work on the first two parts of the project on your own. This will require you to think of a causal question and consider what relevant variables might be included in your analysis.

- **Task 1:** Define a causal question: What is a causal question that you are interested in? Carefully define a treatment and outcome of interest. Make sure a data set exists which at least includes observations of both treatment and outcome.
- **Task 2:** Draw and reason about a causal graph: What is a causal graph which includes all variables relevant to your causal question? Given this causal graph, what is a sufficient adjustment set?

For these first two tasks, you have two options to choose from:

- You can ask a question which may be answered using data from the National Longitudinal Study of Adolescent to Adult Health (Add Health). This is a survey of roughly 20000 individuals which was first administered to high-school aged participants in 1994-1995. The survey followed up with the same participants at 4 subsequent time points, as recently as 2018 when the original participants were in their mid 20s or early 30s.
- You can ask a question which may be answered using data which you find on your own. We have provided several potential options below.

1.2 Group Selection

After the first two tasks, we will place students into groups of 4-7 students based on the following process

- We will take the proposed causal questions which use data from Add Health and group them into various thematic categories. This list will be made available to all students.
- Students who have proposed a causal question not using data from Add Health can also choose to have their causal questions included in the list
- All students will look through the list of causal questions/themes and express interest in several topics through a canvas survey
- Based on student responses, we will assign project groups of 4-7 students
- If a student proposes a causal question with an original data set (i.e., not the Add Health data) and that data set/question is selected for a group project, they will receive up to 5 extra credit points on the project[1].

1.3 Group tasks

After completing the first two parts, we will place each student into groups of 4-7 students.

[1] The total grade for the project cannot exceed 100%

- **Task 3:** Gather and process relevant data: Wrangle and process the data into a form which is usable for the analysis.
- **Task 4:** Select an identification strategy and analyze the data: Given assumptions which are reasonable for your causal question, choose an identification strategy and a method for analysis. This analysis should be made reproducible using an Rmarkdown file.
- **Task 5:** Communicating the results: Communicating the results will have two parts. Students will write a short report describing the causal question of interest, the assumptions made, the identification strategy and the analysis findings. You will also make a short presentation video (7-10 minutes)

2 Data

Students may either choose to answer a causal question using the Add Health Data or find a data set of their own.

2.1 Add Health

The National Longitudinal Study of Adolescent to Adult Health (often referred to as "Add Health") is a survey of roughly 20000 individuals which was first administered to high-school aged participants in 1994-1995. The survey followed up with the same participants at 4 subsequent time points, as recently as 2018 when the original participants were in their mid 20s or early 30s. The questions asked in the survey cover a range of topics including: Crime/Delinquency and Victimization, Demographic Characteristics, Education, Family, Medication and Substance Use and Abuse, Psychological Well-being and Cognition, Reproductive Health, SES, Labor Market and Occupation. You can find a more comprehesive list of topics and specific survey questions here. To get an idea of the types of questions people have used the data to answer, take a look at journal articles citing Add Health here.

2.2 Other potential data sources

We want you to be creative in the causal questions you are asking! If you have a causal question of interest but aren't sure if you can get data, feel free to ask us for pointers. We've provided a few ideas below to get you started.

- Detailed data is available for many major sports. See this page for a good list of potential resources. As examples, we have listed two applications of causal inference to sports questions below
  - Yam and Lopez (2019) consider data from the NFL and ask what is the causal effect of trying convert on fourth down more often?
  - Cummiskey et al (2024) consider data from major league baseball and ask what is the effect of bunting on the probability of scoring at least 1 run?
- State and federal governments have open data portals which provide a variety of data sets which may be of interest
  - The Environmental Protection Agency (EPA)
  - NY State
  - The Federal Govnernment makes has a data clearinghouse data.gov
  - The Federal Reserve makes many macroeconomic data sets available through their FRED website.
- Opportunity Insights is a team of economists who study inequality and social mobility. For instance, a recent paper by this group has asked "What is the causal effect of attending an Ivy League university on future earnings?" They have their data publicly available here. They also have interesting data on COVID 19.

- The Inter-university Consortium for Political and Social Research (ICPSR) holds a number of political and social science data sets. The datasets are grouped into the broad themes which you can see here. This includes topics like: Health and Medical Care Archive, National Addiction & HIV Data Archive Program, National Archive of Criminal Justice Data, and the Child and Family Data Archive.
- The Open Policing Project has a data set on traffic stops from various jurisdictions around the US. They have used this data to consider potential racial bias in policing. You can get their data here.