Each of you has been placed in a group based on the general topic you submitted for Part 1 of the project. As a group, you will need pick a specific causal question to answer. You will then analyze the data and produce a written report which will be due on **Dec 8** (the last day of classes). There will also be a planning check-in which will be due **Nov 25**. The initial two parts of the project are described below and details about the final report (and a R markdown template) will be posted later.

As a first step, your group should meet to discuss a plan for how you will complete the project. Your group can meet in-person or over Zoom. Since it's a group project, you may split up the work however you see fit, as long as everyone contributes relatively equally. You can certainly choose to do everything together, or split off into smaller groups to tackle different pieces–just make sure the end product is a cohesive representation of what your group worked on! You will first turn in a short check-in through canvas on **Nov 25**.

1 Project planning

Look for the "Project Plan" assignment on Canvas. It is a group assignment, so only one person from your group needs to submit for everyone in the group to receive credit. An RMarkdown template will be posted to the course website which you can compile to pdf and submit. It will have the following questions:

1. Has your group been able to get into contact with everyone in the group? (yes/no)
    - If the answer is "no," someone in your group should email the instructors to let us know what's going on.
2. List the name of the treatment and the name of the outcome your group is considering.
3. Data:
    (a) If your group is using the ADD Health dataset, have you identified the variable names for the treatment and outcome for your casual question? (yes/no)
    (b) If your group is not using ADD Health, have you picked a dataset (that includes your treatment and outcome variable)? (yes/no)
4. Write 1-2 sentences describing your group's plan for Data processing (see details in the corresponding section below). Your plan should include the following:
    - The "point people" in your group for this part, i.e. which members are responsible for taking charge and making sure this gets done
    - Since this needs to be done before your can conduct an analysis, include a tentative timeline
    - EXAMPLE: In our group, Sam and Christina will be taking the lead on the data cleaning and processing. They will check in with us via our groupchat by Monday Nov 18.
5. Write 1-2 sentences describing your group's plan for Analysis strategy (see details in the corresponding section below). Your plan should include the following:
    - State the method you will use and why it is suited for your causal question and data. For example, if using regression discontinuity, you should indicate the running variable and cutoff and confirm your dataset has this information.
    - Regardless of which method you are doing, list the "point people" in your group for this part, i.e. which members are responsible for taking charge and making sure this gets done and a timeline for getting this done
    - EXAMPLE: We will use matching. In our group, Shira and Filippo will work on identification (drawing the DAG and determining conditional exchangeability). They will work with Sam and Christina to make sure the variables are in the dataset to use in matching. They plan to finish

this by Nov 18th. Ezra and Touchdown will work on conducting the matching in R, starting Nov 18th.

6. State the next date your group is planning to check in on progress (either via a meeting or through a groupchat). You can use this internally as a deadline to ensure things are getting done, or to start drafting the final paper, etc.

2 Data preparation

The first thing you will need to do is gather and process relevant data into a form which is usable for the causal analysis. This includes, at minimum, the following steps:

1. Finding the relevant variables in the ADD health data and downloading the raw file
2. Read the file into R as a data table
3. Explore the data: are there missing values? are there outliers? What variables, besides treatment and outcome, will you need for your analysis?
4. Clean the data: handle missing values and outliers
5. Process the data: Do you need to modify the treatment or outcome variable? For example, if the treatment in the study was on a scale of 1-5, your group may need to turn this into a binary treatment. For example, your group might decide "treatment" is a value of 3 or less and "control" is a value of 4 or above, so you'll need to modify your dataset accordingly! For any variables you decide to dichotomize, you will need to do some processing.

Be prepared to discuss this in a paragraph (or two) for the final paper. Data preparation and analysis strategy will go together, so you may need to go back and forth between the two to complete both parts.

3 Analysis strategy

You will also need to think about an identification strategy for analyzing 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.

Specifically, you will need to:

1. Draw a DAG representing your causal question that includes at least three relevant variables besides treatment and outcome that are included in your dataset. This is like what you did for Task 2, but with the added piece that at least three of these extra factors should be in your dataset. (?)
2. Determine if conditional exchangeability holds i.e. does a sufficient adjustment set exist? If at first it doesn't hold, see if there is another variable in your dataset that you could add to the DAG so conditional exchangeability holds. In some cases, you might have to make some strong assumptions in order to make this work... that's okay, just be prepared to discuss this in the final paper!
3. Consider the other identification conditions: consistency and positivity. Do these seem reasonable in your setting? Why or why not?
4. Conduct a (reproducible) matching analysis in R. Similar to Question 3 in Problem Set 4, you should:
    - Pick a method we have discussed in class. For most groups, this will be either matching, IPW, outcome modeling with the parametric G-formula, or instrumental variables. If you are planning on using another identification/estimation procedure, please talk with us first to make sure it will be appropriate for your data.
    - Justify why your chosen method is suited to your causal question and data

- Discuss the assumptions required for your method and why they are reasonable for your setting. Note that even if your method does not require conditional exchangeability, you will need to discuss whether or not it might hold in your setting.
- Conduct a (reproducible) analysis in R and be prepared to discuss any choices you make, such as bias-variance trade-off, etc.