***Instructions:*** *Make a copy of this document and fill in the relevant questions below. For each research question, you should copy-paste and answer questions 3, 4, 5, and one of 6a/6b/6c/6d, depending on the primary technique. We expect most proposals to be about 1-1.5 pages. Anything longer than 2 pages will* ***not*** *be reviewed. Submit your work to Gradescope, making sure that you include your group members in your submission.*

**1. What dataset are you using? (air/election/energy/NBA/other, fill in the blank for other and provide a link)**

We are using the provided air pollution and disease indicators dataset, specifically, the data sets that contain information about [COPD](https://data.cdc.gov/500-Cities-Places/500-Cities-Chronic-obstructive-pulmonary-disease-a/xzgm-d4ya) and [Asthma](https://data.cdc.gov/500-Cities-Places/500-Cities-Current-asthma-among-adults-aged-18-yea/7zpf-ctr4). For information about the demographics of the US cities, our group will use the ACS 5-Year Estimates from the US Census Bureau.

**2. If you’re using any supplementary data, what are you using and why (1-2 sentences)? If you aren’t, justify why you don’t need any (1-2 sentences).**

The IV being used depends on the minimum distance of cities to the nearest factory; we will use datasets involving the location of air pollution-causing factories in the United States. To adjust for confounders such as socioeconomic status, city population, and demographic factors, we will utilize datasets on a city-wide granularity.

We will also be using a data set that provides us the [location of power plants](https://www.propublica.org/datastore/dataset/toxic-air-pollution-hot-spots) which we will also integrate in our analysis.

**General Research Question (Duplicate this section and choose the relevant version of question 6 for each one)**

**3. What is your research question?**

What are the effects of pollution on rates of diseases such as COPD and asthma? Does the proximity to power plants(source of pollution) have a significant effect on the rates of diseases such as COPD and asthma?

**4. Which of the four techniques will you primarily be using for this question?**

Our group is using two-stage linear regression with instrumental variables to estimate the causal effect of pollution on respiratory diseases, adjusting for omitted variable bias. We are also using both parametric and nonparametric models to predict the rate of respiratory diseases of cities in the US. For the parametric model, we are using logistic regression. For the nonparametric model, we are using random forests and gradient-boosted trees.

**Causal Inference**

**6b. Briefly describe the treatment, outcome, units, confounders (if applicable), and instrumental variables (if applicable). Briefly describe the technique you plan to use (1-2 sentences).**

In our analysis, the treatment is air pollution measured by PM2.5, and the outcome is the rate of COPD/ asthma. In addition, the units are cities located in the US. The instrumental variable is the distance of cities to the nearest air pollution-causing manufacturing plants. Our primary technique is two-stage linear regression: the first stage regresses PM2.5 levels on the distance of cities to the closest manufacturing plants, and the second stage regresses the rate of respiratory diseases on the predicted PM2.5 levels.   
Then, our goal is to obtain the unbiased estimator for the average treatment effect by looking at the coefficient of the treatment in the second stage linear regression.

**Prediction**

**6d. Which nonparametric method will you use, and why? What kind of GLM will you use, and why? Briefly discuss what features you plan to use, or how you plan to come up with them (1-2 sentences).**

For the GLM requirement, our group is using a logistic regression model to predict the rate of respiratory diseases in US cities. Our group will use the logistic regression model because rates are restricted to the range [0, 1]. We are planning on using features that are known to have an effect on respiratory diseases such as demographics, pollution measures, concentration of factories, etc. Furthermore, our group is using the nonparametric models of random forests and gradient-boosted trees to predict the rates of respiratory disease.