Group Members: Jacqueline Telson, Emily Baytalsky, Uday Krishna Kalna, Benjamin Kaliczak

1. What dataset are you using?
   - Air quality: CDC Chronic Disease Indicators filtered for Asthma

2. If you're using any supplementary data, what are you using and why?
   - CDC Daily Census-Tract PM2.5 Concentrations
   - We are using this data set because we want to relate PM2.5 concentrations to asthma attacks in a causal way and a predictive way.

Q1-3. What is your research question?
   - Does an increase of PM2.5 cause an increase in Asthma cases?

Q1-4. Which of the four techniques will you primarily be using for this question?
   - Causal inference

Q1-5. (Optional) If you're using more than one technique for this question, what other(s) are you using? **N/A**

Q1-6b. Briefly describe the treatment, outcome, units, confounders (if applicable), and instrumental variables (if applicable). Briefly describe the technique you plan to use.
   - Use two states with similar population sizes, with the treatment of one being significantly higher PM2.5 levels than the other to predict increased rates of Asthma in the population. Confounders would be overall health, smoking levels, physical activity, weather, and allergens.

Q2-3. What is your research question?
   - Predicting asthma rates based on socioeconomic indicators of the demographics (age, sex, race etc.) using negative binomial regression and random forests.

Q2-4. Which of the four techniques will you primarily be using for this question?
   - Prediction with GLMs and nonparametric methods

Q2-5. (Optional) If you're using more than one technique for this question, what other(s) are you using? **N/A**

Q2-6d. 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.
   - We will use random forests for our nonparametric method because random forests are great with high dimensional data, and here we have multiple features we want to account for.
   - We will use negative binomial regression for our GLM because it is used for counts, and it does not restrict the mean to be equal to the variance.
   - We plan to use socioeconomic indicators of demographics (age, sex, race, etc.) for our features.