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

Dataset 1: Chronic disease and Air Quality

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).

We are using data from *census.gov* for county populations in order to convert the county-wise ozone measurements into a weighted average for each state. We will also use the Census data for the average population age in each state. We are also using the eGRID dataset mentioned in the supplemental section of the spec for CO2 emissions.

Research Question #1

3. What is your research question?
Is there a significant association between diabetes and…

- Soda consumption among high school students
- Obesity among adults aged >= 18 years
- Overweight or obesity among adults aged >= 18 years
- Physical exercise
- Gender [categorical]
- Ethnicity [categorical]
- Poverty percentage
- Current lack of health insurance among adults aged 18-64 years
- Median daily frequency of fruit consumption among adults aged >= 18 years
- Hypertension (high Blood Pressure)
- Computer use among high school students
- Television viewing among high school students

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

Multiple Hypothesis Testing

6a. Describe your hypothesis tests (≥6), and how you plan to test them (2-3 sentences).

For each variable listed in Q3:

Null Hypothesis: The variable of interest does not have a significant effect on Diabetes.
Alternate Hypothesis: The variable of interest does have a significant effect on Diabetes.

For the categorical variables, we plan to test the effect using A/B test and for numerical variables, we plan to test the effect using correlation. We will correct for these multiple hypothesis tests using FWER and FDR.

Research Question #2

3. What is your research question?

What is the causal effect (if any) of Ozone levels on COPD prevalence?

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

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).
The treatment in this case is the level of ozone in each unit, i.e each US state in the data. The outcome is the incidence of COPD in each state. We have no confounders that affect *both* treatment and outcome. However we will be controlling for other variables that affect COPD, namely: average population age, incidence of asthma, and PM2.5 levels each measured per state. We have two instrumental variables: CO2 levels and population density, both measured by state as well.

Since we are evaluating the effect of a continuous variable (Ozone), we will compute the marginal effect of treatment: i.e, if treatment(ozone) was to increase by one unit, how would the outcome(copd) change. Instead of using a simple difference in means, we will use a regression type of formula which allows us to directly estimate marginal effects — our regression will be weighted, where weights are the inverse densities found by Inverse Propensity weights (IPW) technique.