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)
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 not using any supplementary data, since our research questions can be answered and justified using the provided datasets.

Question 1:

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

3. What is your research question?
Tobacco: Multiple factors (strong prevention policy, tobacco sold…) might influence the percentage of chronic diseases

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

5. (Optional) If you're using more than one technique for this question, what other(s) are you using?
We currently don't plan to use more than one technique for this question.

Multiple Hypothesis Testing
6a. Describe your hypothesis tests (≥6), and how you plan to test them (2-3 sentences).
1. Higher tobacco sales are associated with a higher prevalence of chronic disease.
2. There has been a significant increase or decrease in the prevalence of chronic disease.
3. Certain factors (age, gender, race …) are associated with the prevalence of diseases.
4. Stronger prevention policies are associated with lower chronic disease prevalence.
5. Higher levels of health education in a population are associated with lower prevalence rates of chronic diseases.
6. Higher proportion of the population protected by smoke-free policy are associated with lower prevalence of diseases..

Question 2:

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

3. What is your research question?
Air Quality and COPD: Does worse air quality (PM2.5 concentration) cause a higher prevalence of COPD? (compare the prevalence of COPD as well as the PM2.5 concentration between 2011 and 2014 because PM2.5 concentration data is only from 2011 to 2014)

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

5. (Optional) If you're using more than one technique for this question, what other(s) are you using?
We currently don't plan to use more than one technique 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).
Treatment: air quality (PM2.5 concentration)
Outcome: prevalence of COPD
Units: percentage
Confounders: the geographical locations can serve as a confounder variable (geographical locations affect both air quality and prevalence of COPD)
Instrumental variable: socioeconomic status can be the instrumental variable but we may need extra dataset (we are still searching for that).