The proposal document should address the following points. Use these points as headers in your document.

- **Basic Info.** The project title, your names, e-mail addresses, UIDs, a link to the project repository.
  - **Project Title: <u>Effects of Air Quality on COVID-19 Cases and Mortality</u>**
  - **Names:**
    - **Taryn Rahman - U6037359**
    - **Michael Lansford - U1130312**
    - **Anurag Gupta - U1317679**
    - **Isha Talegaonkar - U1317994**
    - **Soumyajit Saha - U1320949**
  - **Link to Project Repository:** **https://github.com/isha-talegaonkar/data-drangling-team-9**

- **Background and Motivation.** Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.
- COVID-19 has evidently been a major worldwide crisis affecting many lives through prolonged illnesses and mortality. While most individuals were cautious of the virus, those who were/are already immunocompromised were hyper aware due to potential correlations in worsening of COVID-19 symptoms and likelihood of contracting the virus. Another correlation was found between air quality and the virus. Essentially, the worse the air quality, the higher the likelihood of positive COVID-19 results and mortality rates.
- Our group is interested in this topic to learn more about the correlation between the quality of air from different air quality measures and COVID-19. It will be interesting to understand what populations are more susceptible to the virus based on their location/environment.

- **Project Objectives.** Provide the primary questions you are trying to answer with your data. What would you like to learn and accomplish? List the benefits of how the data could be useful.
- How does the air quality, considering different air quality measures (ie. ozone, NO2, etc) correlate with COVID-19 positive results and mortality in a given environment?
- We would like to learn how different concentrations in different air quality measures affects COVID-19 data (positivity and mortality). We may also like to learn more about each separate measure, and potentially if any has more of an effect on COVID-19 positivity and morality than others, or if this is environment based.
- Understanding the correlation between air quality and COVID-19 will evidently identify environments that are at higher risk. If data analysis as such are used in the real world, then proper/effective proactive measures can be taken depending on the population/location.

- **Data.** From where and how are you collecting your data? If appropriate, provide a link to your data sources.
- We are using the earth data from NASA that includes various air quality measurements data files: https://www.earthdata.nasa.gov/learn/find-data/near-real-time/hazards-and-disasters/air-quality
- We will use the provided COVID-19 data: https://github.com/CSSEGISandData/COVID-19_Unified-Dataset

- **Data Processing** - Do you expect to do substantial data cleanup? What quantities do you plan to derive from your data? How will data processing be implemented?
- To get a look at what the datasets look like, we generated a links list through GES DISC from 1st January 2020 through 31st December 2022. For exploratory purposes, we have downloaded the links list for Ozone. We wrote a python script to convert the datasets downloaded through the links list from the nc4 format to csv, using netcdf4 and xarray libraries.
- The air quality datasets are available for each separate day (single day). Once we compare the available timeframes in air quality to that of the COVID-19 datasets, we will decide on a timeframe, and similarly a region to focus our analysis on. This will require consolidating the datasets from the air quality data, for the chosen measures (ie. ozone, NO2, etc).
- From the air quality data, taking the ozone df for instance, we need variables that quantify air pollution/quality. This includes the pollutant itself (ozone) and the air density. From the COVID-19 datasets, we will consider features that include the region/location, and positivity and mortality rates.

- **Design -** How will you display your data? Provide some general ideas that you have for the design. Develop **one alternative prototype design for your data**.. Describe your designs and justify your choices of visual encodings.
- Since we will be comparing the air quality data along with the COVID cases for that particular region in that time period, we will be doing a frequency mapping for the covid cases for the day and the concentration of the various pollutants present in the air during that time period in that region. We will be making use of scatter plots to compare air quality measures with the numbers of COVID cases and box plots (to show concentration highs and lows for the various pollutants).

- **Must-Have Features.** List the features without which you would consider your project to be a failure.
- Air density
- Positive covid result (rate?)
- Mortality (rate)
- Specific Pollutant Measure (ie. for Ozone we need the O3 - ozone mass mixing ratio measurements).

- Region/Location codes/identifiers
- Time(frame)

- **Optional Features.** List the features which you consider to be nice to have, but not critical.
- Population density
- Diabetes
- Obesity
- Smoking
- COPD
- Hypertension
- Age
- Vaccinated population

- **Project Schedule.** Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

Week 1: Identify columns/variables from air quality and COVID-19 datasets that meet the must have features criteria. Begin cleaning and merging data for at least one set (air quality and/or COVID). We will still need to look at both data groups in week 1 for insight on available features (ie. which region(s) is/are best to choose from).

Week 2: Continue/complete cleaning and merging data, accounting for any primary/foreign keys across datasets. Finalize the processing.

This proposal is the first part of your process book. As a ballpark number: your proposal should contain about 3-4 pages of text.

You will schedule a project review meeting with a staff member during regular lecture times of the week marked in the schedule. Make sure all of your team members are present at the meeting.

The proposal will be submitted by uploading it to your team's GitHub repository.