CS 12200 Project Proposal, Winter 2021
Group name: **Development Data Scientists**
Samantha Koretsky, Noah Cohen-Harding, Rachel Li, Devesh Kodnani

The goal of our project is to create an interactive data visualization for users to explore cross-country data and the relationship between various quantitative indicators of global development. We plan on acquiring data for 10-20 development indicators from Our World in Data (ourworldindata.org), which has compiled and cleaned data from sources including the World Bank and the International Monetary Fund. These indicators could include air pollution deaths, SO2 emissions per capita, infant mortality, GDP per capita, contraceptive prevalence, and others. We will prioritize datasets that are aggregated by country and contain observations over time. Here is an example of a data set we would use, that contains the death rate from air pollution for each country from 1990 to 2017 (click download > CSV for the file we would use). Much of the data from our source is labeled with a country code and date of observation, which will be helpful when creating our database.

Our program will prompt the user to select 2-3 indicators from a drop-down menu of the 10-20 variables we have selected. Given these inputs, our program will generate interactive scatter plots to show correlations between the variables (with two variables on the axes and the third variable represented by the size of each country's dot on the scatter plot). There will be a regression line which documents the general correlation between the variables on the axes. A fourth axis, time, will be represented by a slider at the bottom of the graph – moving the slider will move each country's dot and adjust its size based on the observations for that year (assuming one is available for each of the variables selected).

Moreover, we will identify 5-10 countries that are reasonably representative of different regions and development levels (possibly including the United States, the United Kingdom, Russia, China, India, Saudi Arabia, Nigeria), as well as give the user the opportunity to choose one or two countries that they are especially interested in, to highlight and label on the scatter plot. We also want to make it possible for the user to click on certain dots – for example, outliers on the scatterplot – to view the country's name, and the values of the different variables represented on the scatterplot. We will also seek to add additional complications, such as color coding all of the dots on the scatter plot by continent, so that the user can see regional differences and trends among the variables. Finally, our output will include information about the variables selected alongside the plot, pulled from sources including Our World in Data and the World Bank (i.e., a short summary about how a certain indicator is measured, or broader global trends in the indicator, that provides interesting information).

To begin developing our project, we will scrape the necessary data from our source and organize it into a convenient database, which we plan to complete by the end of week 5. We will then

write functions for computing correlation for any 2-3 variables from our set over a given time frame, and potentially for completing other statistical tests as well. We plan to have these functional aspects completed by the end of week 6. The last portion of the project will be creating the interactive interface for the user, which will include the user selection of variables from a list and a slider for the time frame. We will also complete management and formatting of output for the plots and other information we plan to return. These tasks will be done by the end of week 8, which gives us one additional week to make final changes and deal with any outstanding issues.