**PROJECT PROPOSAL**

TEAM 3

**U.S. Pollution Data**

**Data Structure**

We have chosen to work with the US Pollution Dataset. The US EPA has recorded the four main pollutants (Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide, and Ozone) for every day between 2000 and 2016.

The link to the dataset is :

<https://www.kaggle.com/sogun3/uspollution>

Dataset shape: 1.746.661 x 29 (rows x columns)

This dataset has 29 columns and more than 1.4 million observations.

The columns are as follows:

State Code,County Code,Site Num,Address State, County City, Date Local, NO2 Units , NO2 Mean, NO2 1st Max Value, NO2 1st Max Hour, NO2 AQI,O3 Units, O3 Mean, O3 1st Max Value, O3 1st Max Hour, O3 AQI , SO2 Units, SO2 Mean,SO2 1st Max Value, SO2 1st Max Hour,SO2 AQI, CO Units, CO Mean, CO 1st Max Value, CO 1st Max Hour, CO AQI

The following are the different data types of the 29 columns:

| DECIMAL | 10 |
| --- | --- |
| INTEGER | 10 |
| STRING | 8 |
| OTHERS | 1 |

## **Problem Statement**

Air quality has developed into a significant environmental health hazard for industrialized humans, serving as the primary cause of lung cancer and other illnesses. The pertinent statistics for NO2, SO2, CO, and O3, include the average, maximum, and lowest values, as well as the air quality index. These statistics pertain to 49 states in the United States. We can determine the air quality in these cities over the last decade based on the data, and then draw conclusions about the air quality trends in these cities. Finally, we can conduct predictive analysis based on the data to forecast the air quality in these cities over the next two years, in order to determine whether to formulate corresponding air protection policies for specific pollutants.

## **Methodology**

In this instance, we may do trend analysis and develop linear models in the United States of America. By doing trend analysis, we can determine which cities exhibit trends but not whether they are declining or growing. To get acceptable results, we're going to construct a 12-month moving average for each city and compare the first and last numbers in the 12-month moving average. Additionally, we will develop a linear regression model and forecast the next two years for each city with a trend. We shall illustrate our forecast using a plot.