**PollutionSight: Machine Learning Model Written Analysis**

* **Input Data:** Daily maximum value measurements of 4 main pollutants (NO2, O3, SO2, CO) and corresponding daily air quality index(AQI) readings for multiple states and counties 2000-1026
* **Target Output:** Pollutant and AQI levels of future dates
* **Goal:** Identify and raise awareness of areas in U.S. that have an trend of increasing air pollution so steps can be taken to improve air quality
* **Model Choice:** Multiple Linear Regression. Comparing the relationship between each pollutant and the AQI in a specific county from 2000-2016
* **Model Process**
  + **Cleaning Datasets:** Download original datasets from Data-World and EPA websites, load them into Jupyter Notebook as pandas data frames,drop unneeded columns, combine annual AQI data for each year 2000-2016, rename similar columns to match, merge pollutant and AQI data frames into master dataset
  + **Preparing Data for ML:** Use State and County Code Guide excel file to select desired State and County, filter through master dataset and make a new dataframe using the desired State and County Code, drop string columns of State and County name.
  + **Import and Split Data:** Import linear regression model from sklearn library, split desired county dataframe into X/y and train/test
  + **Scale/Normalize Data:** use StandardScaler, normalization, and MinMaxScaler to scale data
* **Using and Testing Model:** Pass data through model, use some method to test accuracy, use scaling option with best accuracy, graph training dataset and predicted values, repeat process with other counties and compare results.