Title: Air Quality Index Analysis of the World's Most Populated Cities
Name: Vaishnavi Deshpande

Air Quality Index (AQI) is a metric that is used by agencies around the world to represent the extent of air pollution. This project aims to understand major air pollutants and other ambient factors to determine how they contribute to the AQI of a city. This study can help governments take specific directed measures to reduce air pollution.

The World Air Quality Index Project (https://aqicn.org/ and https://waqi.info/) provides air quality data for 2000+ cities. The AQI values and chemical compositions are scraped from aqicn.org. The dataset contains information on more than 600 of the world's most populated cities. The population data for these cities are scraped from https://worldpopulationreview.com/world-cities. The combined dataset includes geographical factors (like latitude, longitude, and population), atmospheric factors (like dew, humidity, pressure, etc.), and chemical factors like (PM25 (particulate matter under 2.5mm), PM10 (under 10mm), Ozone, Carbon Monoxide, etc.).

A higher AQI indicates worse air quality. To facilitate better analysis, the cities are classified into three distinct categories indicating the level of air pollution. Cities with AQI >250 are assigned a maroon color, >150 are orange, and <150 are white. Figure 1 helps recognize the most polluted regions in the world. Each circle on the map represents one of the 628 cities from the dataset. The radius indicates the AQI value and the color indicates category. On the day that this data was scraped, the city of Laiwu, China had the highest AQI of 399. However, the most polluted city in the US (Fresno, CA) had an AQI of only 114. Further, this plot shows that the most polluted cities are in China, India, and Bangladesh. While cities in these countries make up only 43% of the dataset, it is interesting to note the massive extent of pollution arising from the three countries.
For 27% of the cities in the dataset (mostly cities in the US), the data source did not include measures for Carbon Monoxide, Nitrogen Dioxide, and Sulfur Dioxide. However, it did not impact the analysis as the AQIs for these cities are comparatively very low (<150).

This next part studies the dimensionality of this dataset by using Principal Component Analysis over 7 atmospheric and chemical components (Dew, PM25, PM10, CO, NO_2, SO_2, and O_3). The blue line in figure 2 specifies that 90% of the variance can be represented by one column and 95% of the variance can be represented by two columns. This is because values under PM25 have a much larger magnitude than the amount of SO_2, for example. Using a pipeline to preprocess the data using sklearn's StandardScaler and performing PCA, it is clear that at least 6 of the 7 are needed to capture at least 90% of the variance.

Given the concentration of pollutants, atmospheric factors, population, and location of a city, I have created a model that predicts the city's air pollution category as maroon, orange, or white. The model involves creating a sklearn pipeline consisting of StandardScaler followed by LogisticRegression. Stratifying on the air quality group of the city (maroon, orange, or white), I have split the dataset into a training dataset and testing dataset (75/25). The model has an accuracy of 80.25%. The recall score is 60.64%, and the precision score is 66.81.

Figure 3 helps understand the coefficients that contribute the most to deciding whether a country falls under the 'orange' category (250>AQI>150). The most important factor for classifying a city's air quality is PM25 with a contributing coefficient of 0.72, which is a dominant pollutant for 99% of the cities that are classified 'orange'. We also see that CO is the second-highest contributing factor (coefficient = 0.69). Further, dew has a high negative coefficient (-0.55), which means that a high dew concentration would likely not lead to an 'orange' prediction. A paper www.hindawi.com/journals/amete/2017/3514743/ by Yingying Xu and Xinyue Zhu validates this.

In conclusion, the most polluted cities are in China, India, and Bangladesh. Further, at least 6 out of 7 atmospheric and chemical columns are needed to represent the data meaningfully. PM25 and CO are the top two contributing factors in deciding whether a city falls under the 'orange' category. This information can help governments prioritize their measures to curb air pollution by focusing on specific factors.

Figure 1: Identifying the Most Polluted Regions in the World

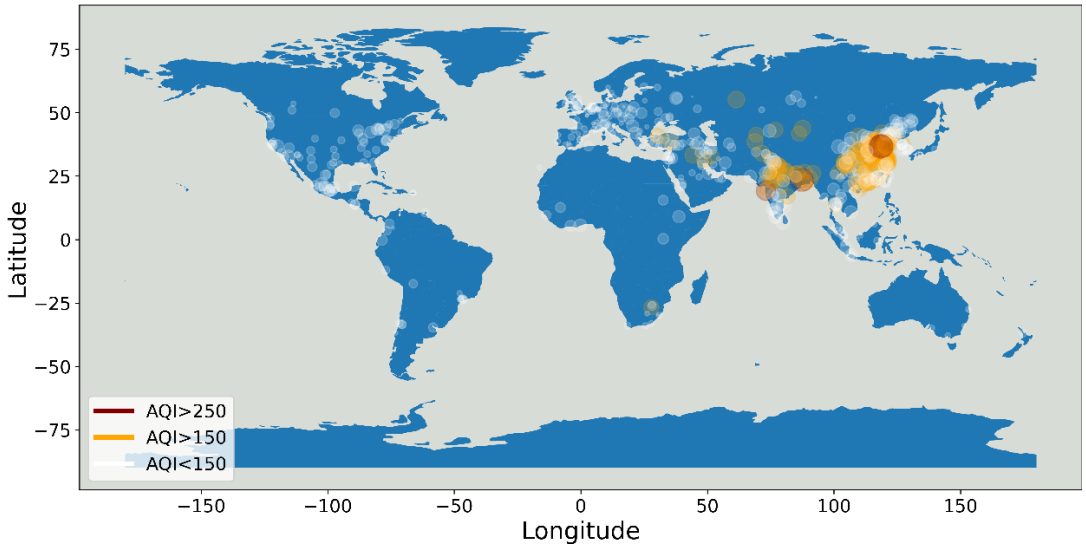

Figure 2: Principal Components of AQI values

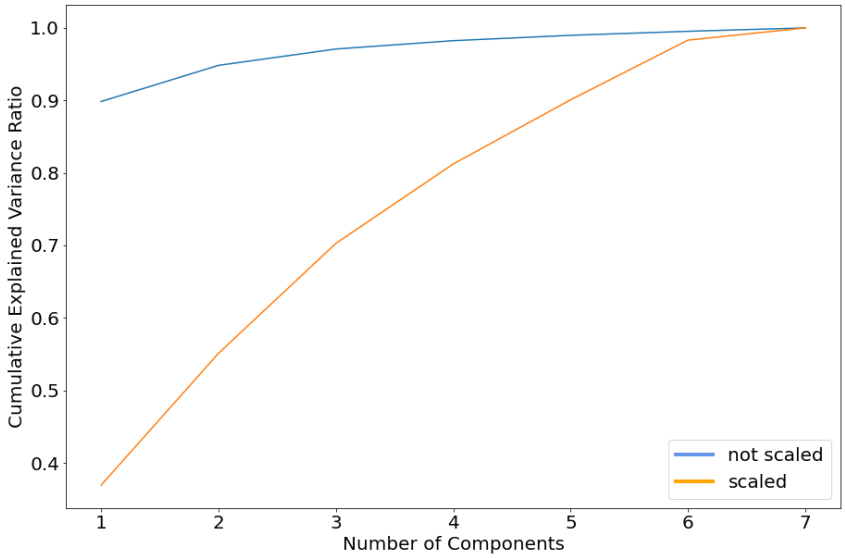

Figure 3: Logistic Regression Coefficients (Countries classified as Orange)