Air Quality Analysis: 2000 - 2021

Christian Shadis
Worcester State University
Big Data Analytics Capstone
Spring 2022

Introduction

Air Quality Index (AQI) is a method of evaluating measurements of the four major pollutants (Carbon Monoxide, Sulfur Dioxide, Nitrogen Dioxide, and Ground-level Ozone) for the safety of breathing the air. Measurements of CO, SO2, NO2, and O3 are taken, converted to respective Air Quality Indexes, and then aggregated into a total Air Quality Index measurement. The higher the AQI, the more dangerous the air is.

AQI Range	Safety Level
$0 \leq AQI \leq 50$	Good
$50 < AQI \leq 100$	Moderate
$100 < AQI \leq 150$	At-Risk
$150 < AQI \leq 200$	Unhealthy
$200 < AQI \leq 300$	Very Unhealthy
$300 < AQI$	Hazardous

Dataset

The US Environmental Protection Agency released measurements of the air pollutants and the resulting Air Quality Index for various places in the US daily between 2000 and 2021. There are approximately 608,700 measurements recorded in the dataset from 47 different states and 148 different cities. The features of the dataset include a date (with separate year, month, and day columns), a location of the measurement (address, county, and state columns), and corresponding pollution measurements; for each pollutant, the average and maximum densities (in parts per million) are recorded along with the corresponding pollutant AQI and the time of day of the maximum measurement.

Preprocessing

The dataset was already very tidy with one observation per row, one feature per column, and no missing values. To examine overall air quality, each pollutant based AQI was added together, resulting in an overall AQI column. This column was then discretized into the bins listed in the table above. There were also several cases in which multiple observations for the same location and day occurred with varying measurement values. This is likely due to readings being taken multiple times a day. To correct this issue, the data was grouped by address and date, then the measurements were aggregated by mean and duplicates were removed. Now one observation exists for each location and day, containing the average readings from each of that day's observations.

Data Visualization

Time Series

The average and maximum measures of each pollutant were recorded every day from 2000-2021 at 1630 North Main St in Los Angeles, California, the city with the highest number of observations. The average measures are indicated by the blue lines, and the maximums by the greens. Some pollutants have yearly cyclic patterns, so each pollutant is visualized first over the 21-year span, and second over the course of 2020. The single-year plots are overlayed with a color-coded smoothed conditional means regression curve to show the overall trend of the data.

All plots were made using the ggplot package in the R Programming Language.

CO

The average CO density in the air has remained steady in Los Angeles over the past 21 years, but the maximum amounts of CO have noticeably decreased in that same span. CO concentration was higher during the cold months of the year than the warm in 2020. It is possible that the burning of fossil fuels for warmth is behind this seasonal shift, but that would require further research to confirm.

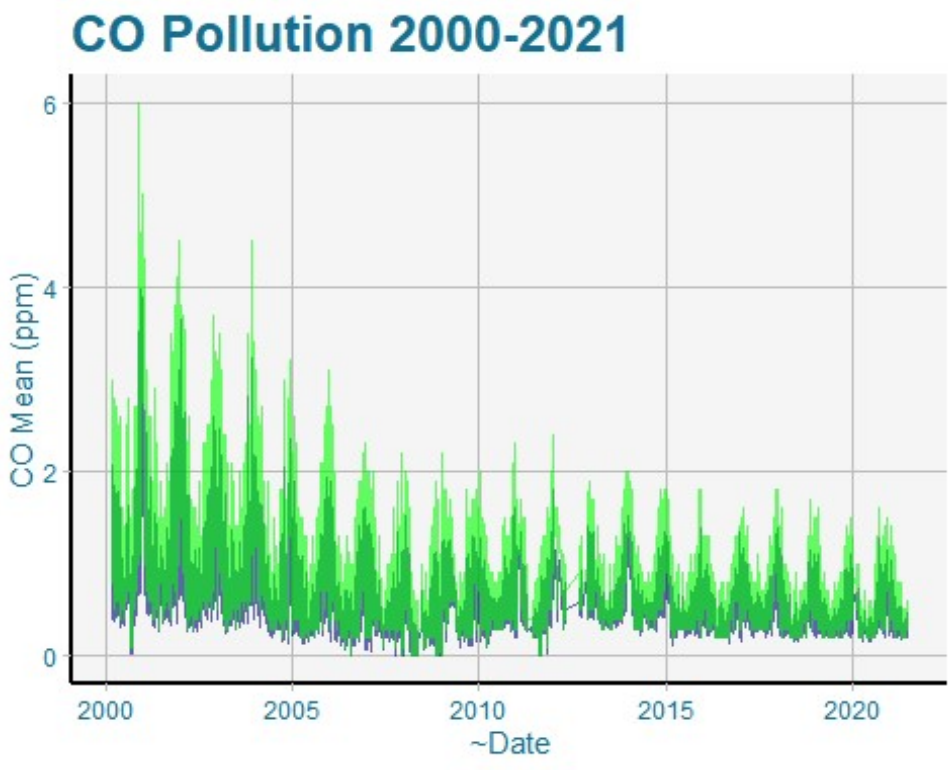

Figure 1: mean = blue, maximum = green

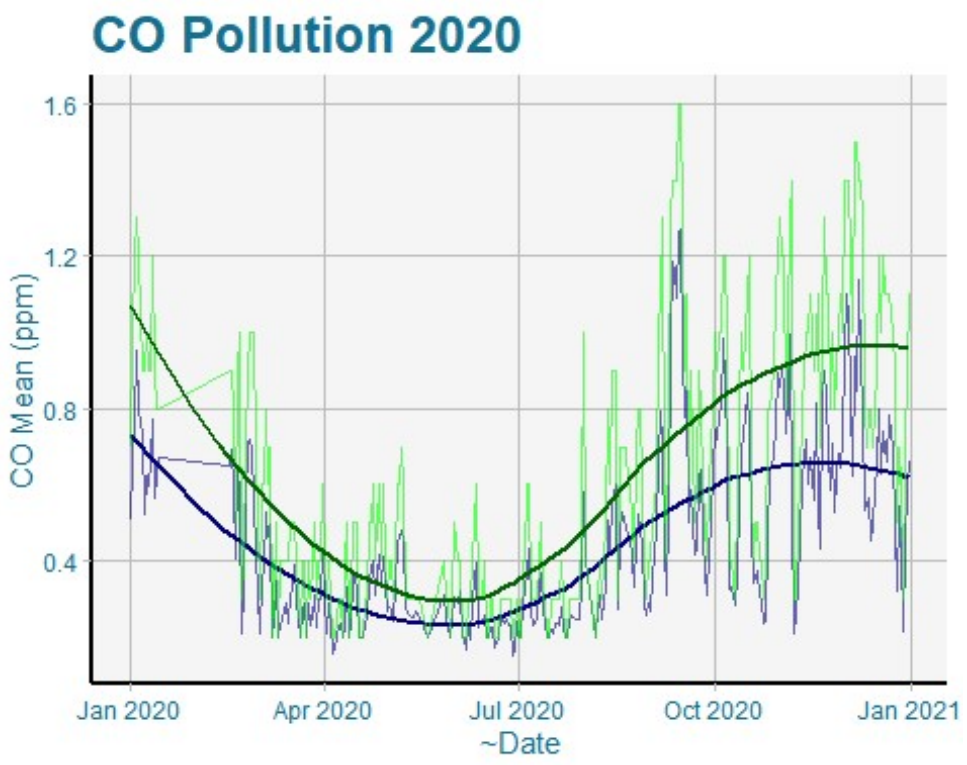

Figure 2: mean = blue, maximum = green

NO2

Similarly, the average density of NO2 has remained approximately the same while the maximum readings have become lower. There was also a trend of lower measurements during the summer than the winter.

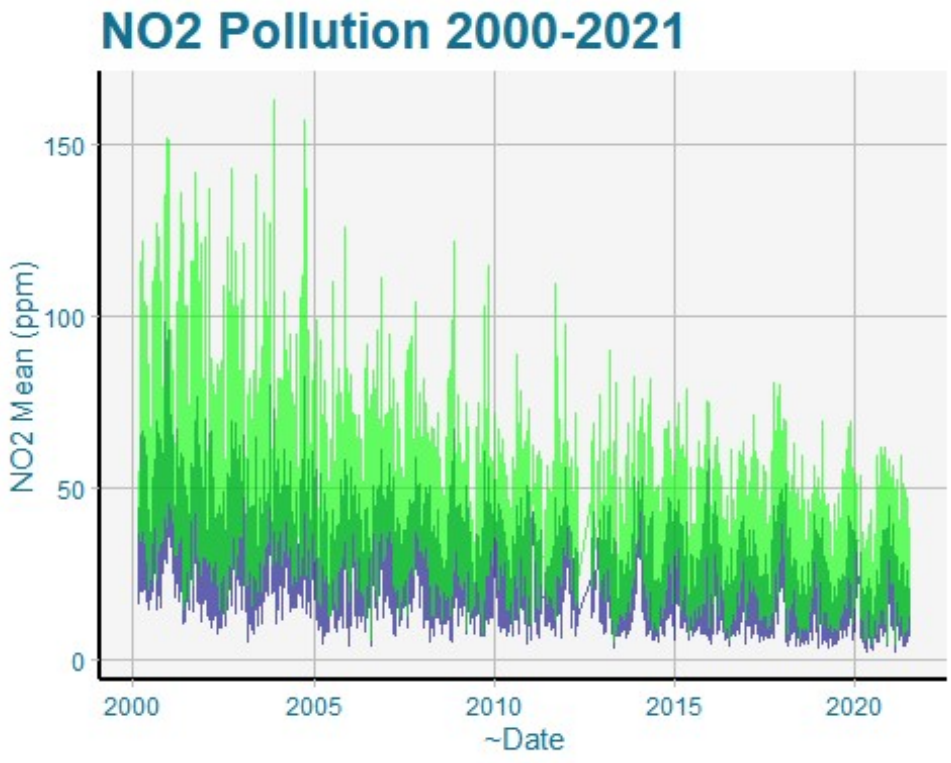

Figure 3: mean = blue, maximum = green

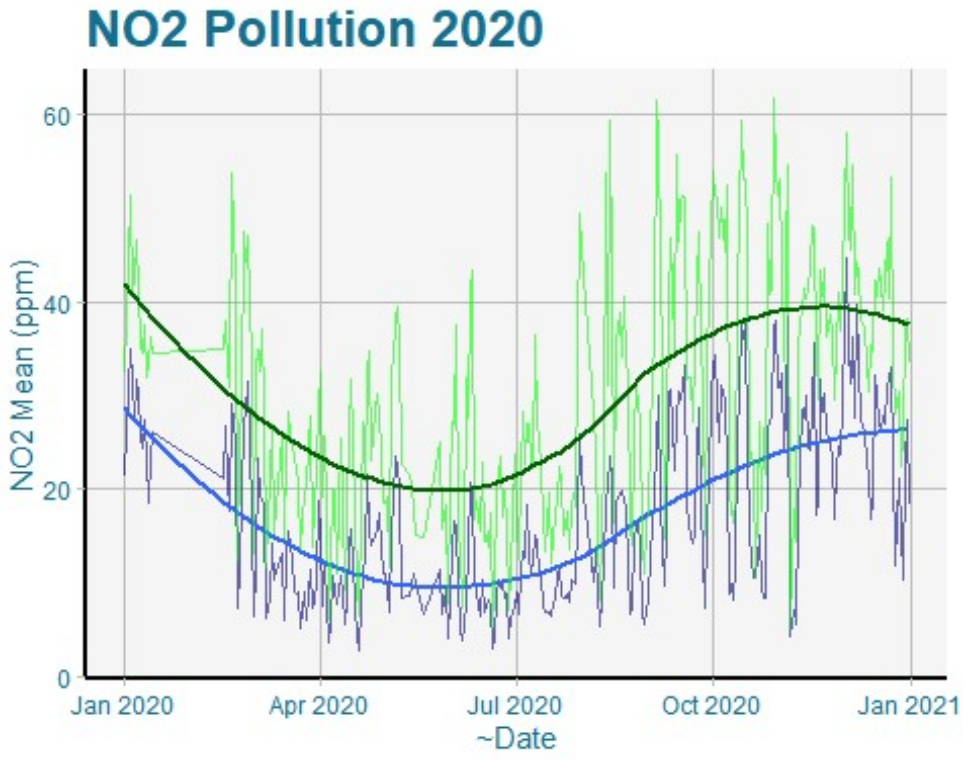

Figure 4: mean = blue, maximum = green

O3

O3 is unique to the other pollutants in that it shows no trend at all over the 21-year span. There seems to be a yearly cycle in which O3 measurements tend to be higher in the summer than the winter, but there has been no consistent increase or decrease in O3 measurements.

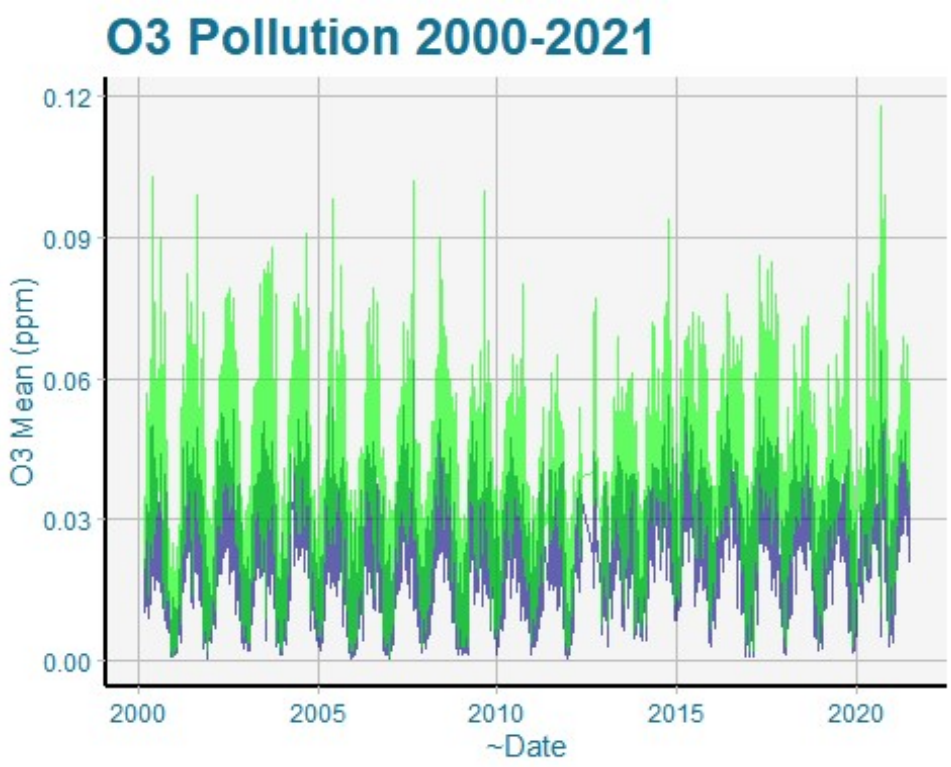

Figure 5: mean = blue, maximum = green

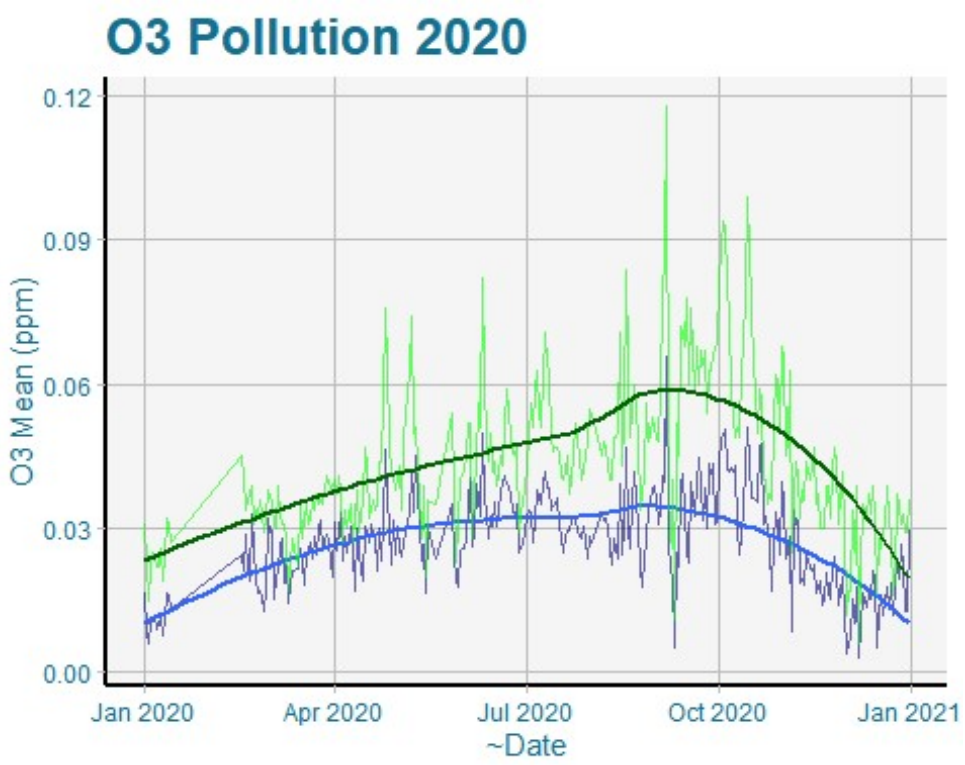

Figure 6: mean = blue, maximum = green

SO2

SO2 is unlike its counterparts in that there does not seem to be a yearly cyclic trend. Measurements do not tend to be higher during one time of year than another, but the average amount of SO2 seems to have decreased over the past 21 years.

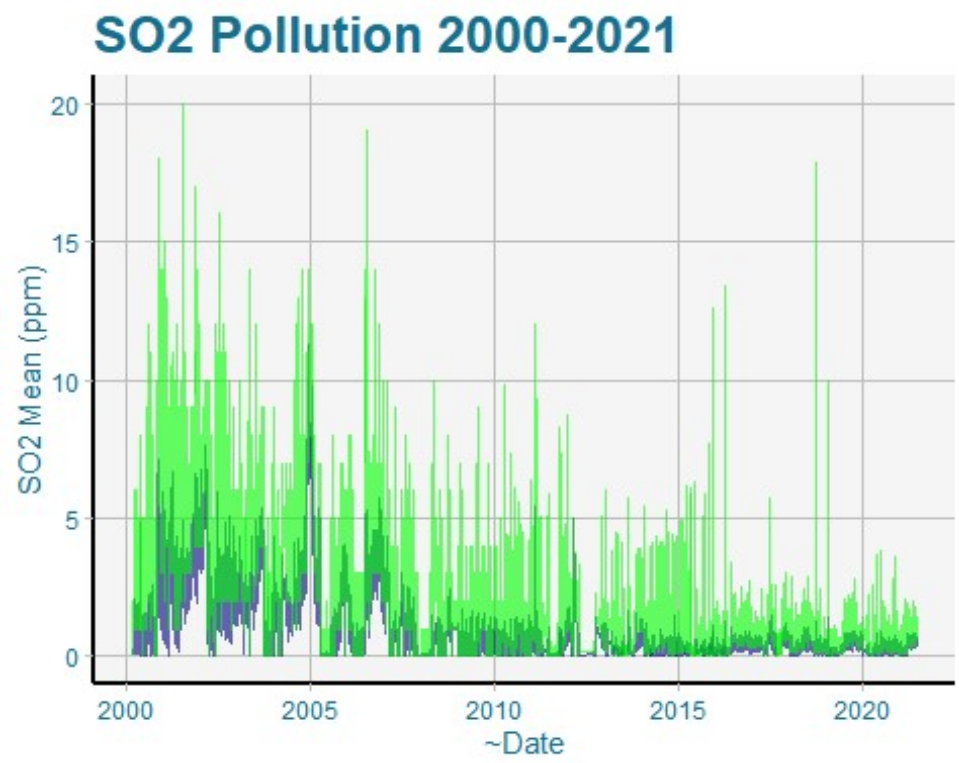

Figure 7: mean = blue, maximum = green

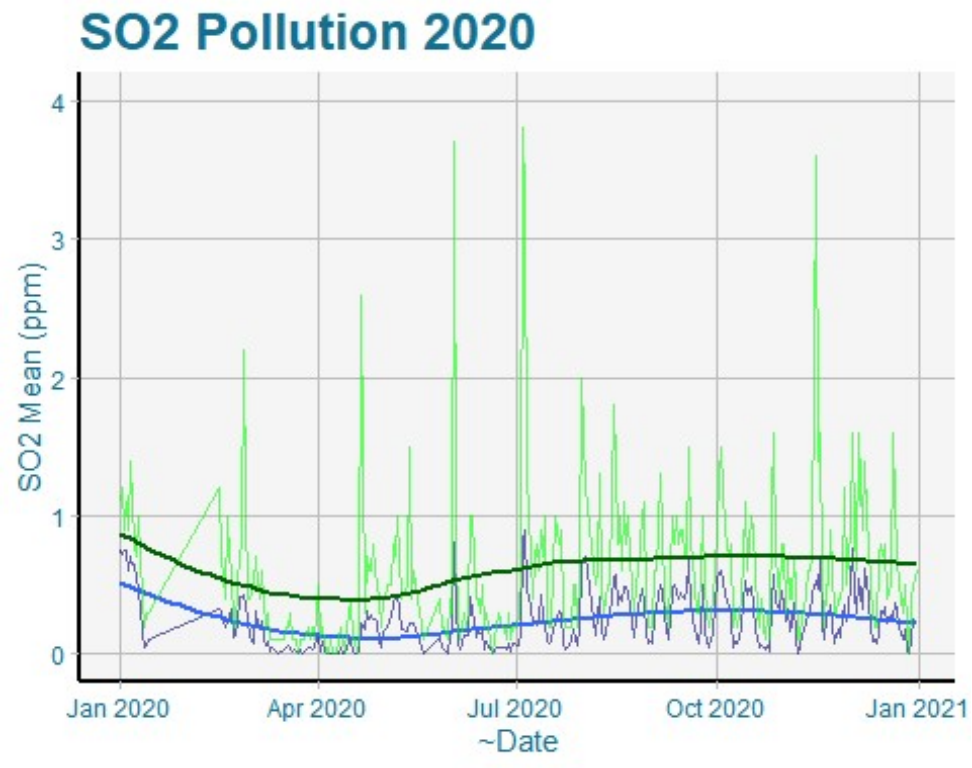

Figure 8: mean = blue, maximum = green

Mapping

Nationwide pollutant data labeled by state is ideal for exploring geographic trends across the country. The R package Leaflet is an intuitive tool for creating interactive maps viewable in a web browser. An R script was written for this project which creates a separate Leaflet map for each year provided by the data. The map allows the user to select which pollutant they would like to see or the total AQI.

There do not seem to be any noticeable regional trends. However, when examining maps several years apart, the overall air quality seems to have improved. Consider the following maps displaying mean total AQI in the years 2000, 2010, and 2020. The main conclusions to draw are that air quality measurement has become more widespread throughout the country over the years, and that air quality seems to have increased. The map of 2000 is far redder than the 2020 map, which is on the exact same color scale. The map has gotten noticeably greener throughout the years.

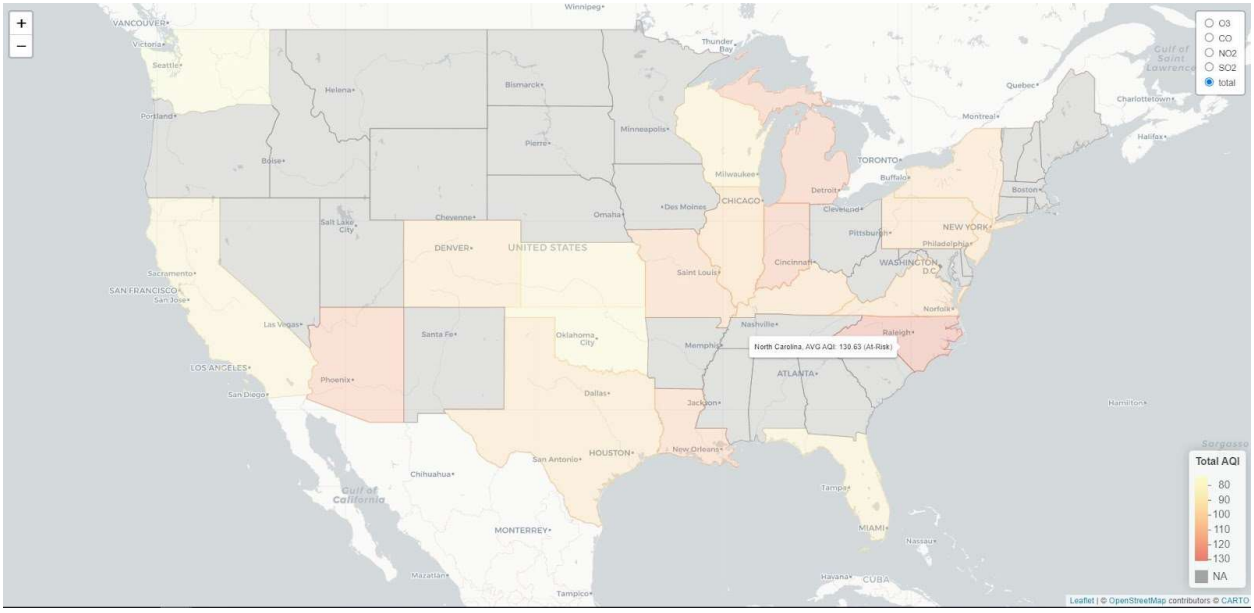

Figure 9: AQI, 2000

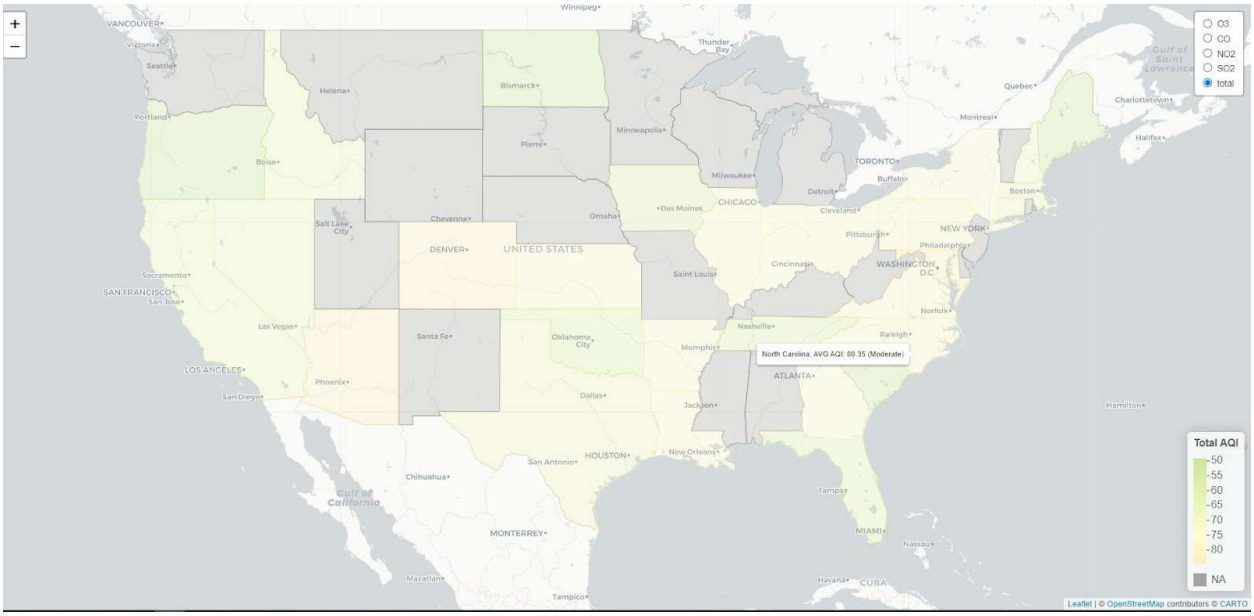

Figure 10: AQI, 2010

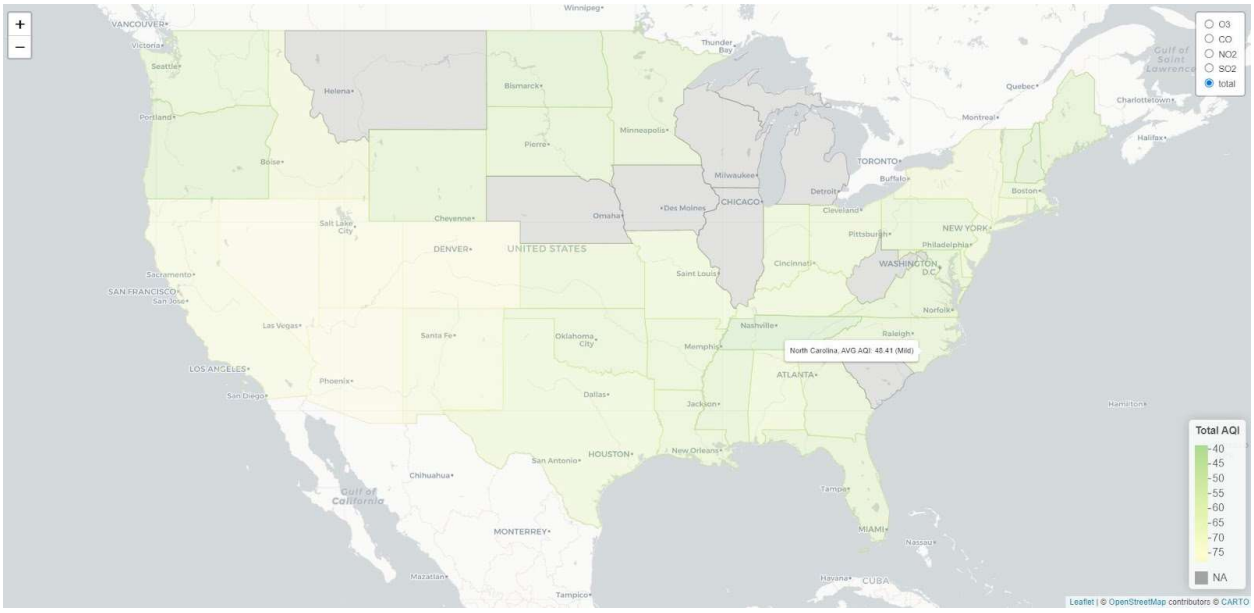

Figure 11: AQI, 2020

Statistics

Summary Statistics

Measure	AQI	O3 Ave	O3 Max	NO2 Ave	NO2 Max	SO2 Ave	SO2 Max	CO Ave	CO Max
Min	1.00	0	0	-4.63	-4.40	-2.51	-2.30	-0.44	-0.40
Q1	49.00	0.02	0.03	4.98	11.20	0.19	0.60	0.18	0.20
Median	65.00	0.03	0.04	9.54	21.80	0.66	1.70	0.26	0.40
Mean	72.18	0.03	0.04	11.74	23.61	1.52	4.21	0.34	0.48
Q3	86.00	0.04	0.05	16.30	33.70	1.77	4.00	0.42	0.60
Max	425.00	0.11	0.14	140.65	269.20	321.63	351.00	7.51	15.50
St. Dev.	33.72	0.01	0.02	9.08	15.41	2.50	7.98	0.28	0.45

Distribution Change Over Time

A fundamental question to answer is whether the air quality has been increasing with time. Have climate initiatives, greener technology, and greater awareness of the environment resulted in any change in AQI? Plotting all daily AQI measurements by year in a boxplot provides visual intuition that it has. It is especially evident in the difference between 2003 and 2018 – notice that the 3rd quartile of AQI in 2018 is lower than the *median* of 2003. This means that 75% of all measurements in 2018 were better than the median AQI in 2003. Furthermore, the spread of measurements has decreased and there are far fewer outliers in the Very-Unhealthy range, and *none* in the Hazard range. All of these are strong indicators that air quality truly has increased over the past 21 years.

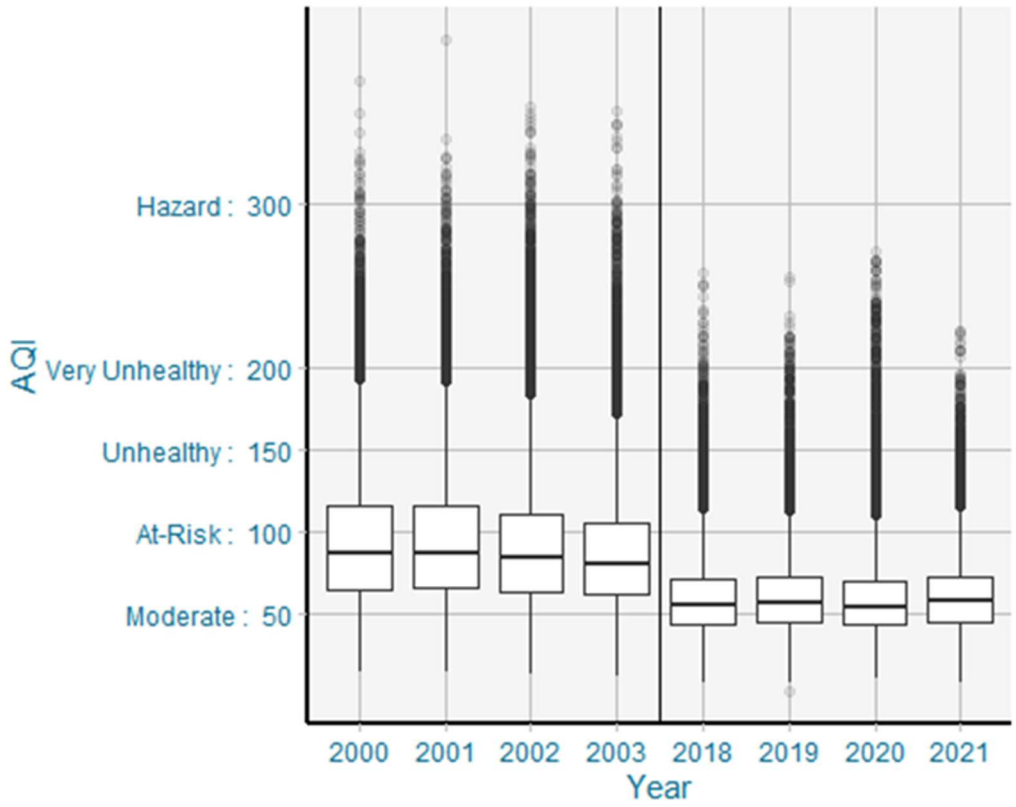

Figure 12: Average total AQI by Year

Test: Sulfur Dioxide in NY vs. TX

In Biology 101, a professor mentioned offhand that acid rain (consequence of SO2) was a common problem in the Adirondack region of New York. This led to the question: is SO2 pollution higher in New York than it is in Texas, the state with the most fossil fuel power plants in the US? Answering this question provides valuable insight into whether pollution control is a regional or a national issue.

A Mann-Whitney U-test was used to determine whether New York had a higher average SO2 density than Texas during the 21-year span. The Mann-Whitney U-Test is analogous to the two-sample T-test but has no requirement for the data to be normalized.

Null Hypothesis: $\mu_{NY} == \mu_{TX}$

Alternate Hypothesis: $\mu_{NY} > \mu_{TX}$

W-statistic: 5.7×10^{8}

P-Value: 2.2×10^{-16}

At the 1% significance level, we reject the null hypothesis as we do have enough evidence to conclude that New York has a higher average SO2 density than Texas, despite Texas emitting far more SO2 than New York.

Correlation

A natural question to ask about these pollutants is whether the presence of one of them is correlated with the presence of another. To answer this question, a correlation map between all numeric attributes was produced.

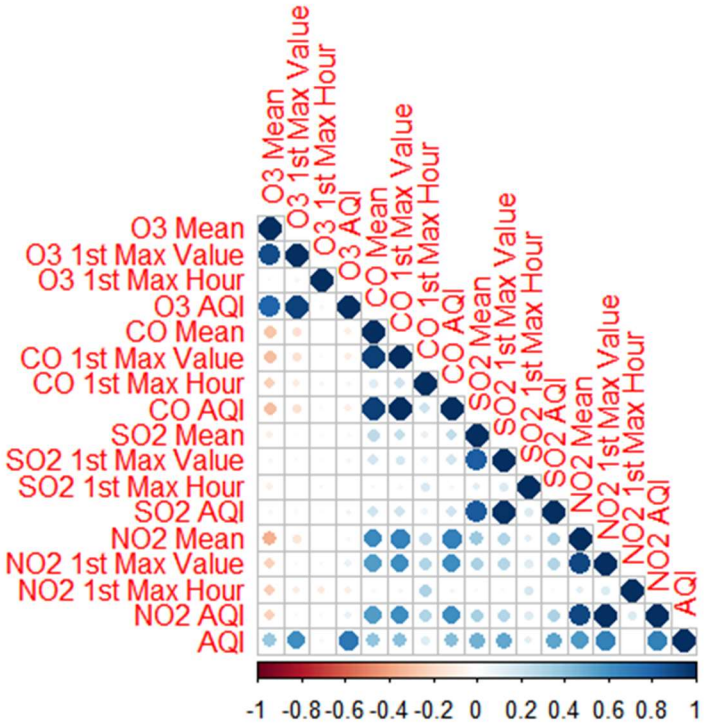

Figure 13: Correlation Map

As one would expect, overall AQI will increase with any of the pollutants due to their positive correlation. However, there does not seem to be any strong correlation between the four pollutants, except for CO and NO2 which have a mild positive correlation. Thus, to reduce AQI, all pollutants must be accounted for independently.

Machine Learning

Clustering

The first machine learning technique used on the dataset was unsupervised learning using both hierarchical clustering and K-means clustering. The goal of clustering is to see if any geographic patterns appear in the pollution data.

The first step in this process was to aggregate the data by year: For each state and each year, the average values of each pollutant and AQI were calculated. Then, the state labels were removed, and all numeric data was standardized and passed to a distance matrix, which was then used to create the hierarchical model with complete linkage. The standardization process was subtracting the mean from the observation and dividing by standard deviation.

The unlabeled data was also passed to fifteen different K-Means models each with a different number of clusters, and the total within sum of squares (statistic describing homogeneity of the clusters) was calculated, resulting in a scree ("elbow") plot. According to the plot, the ideal number of clusters is approximately three. Thus, the K-means and hierarchical models were created with three centers each, resulting in the following clusters:

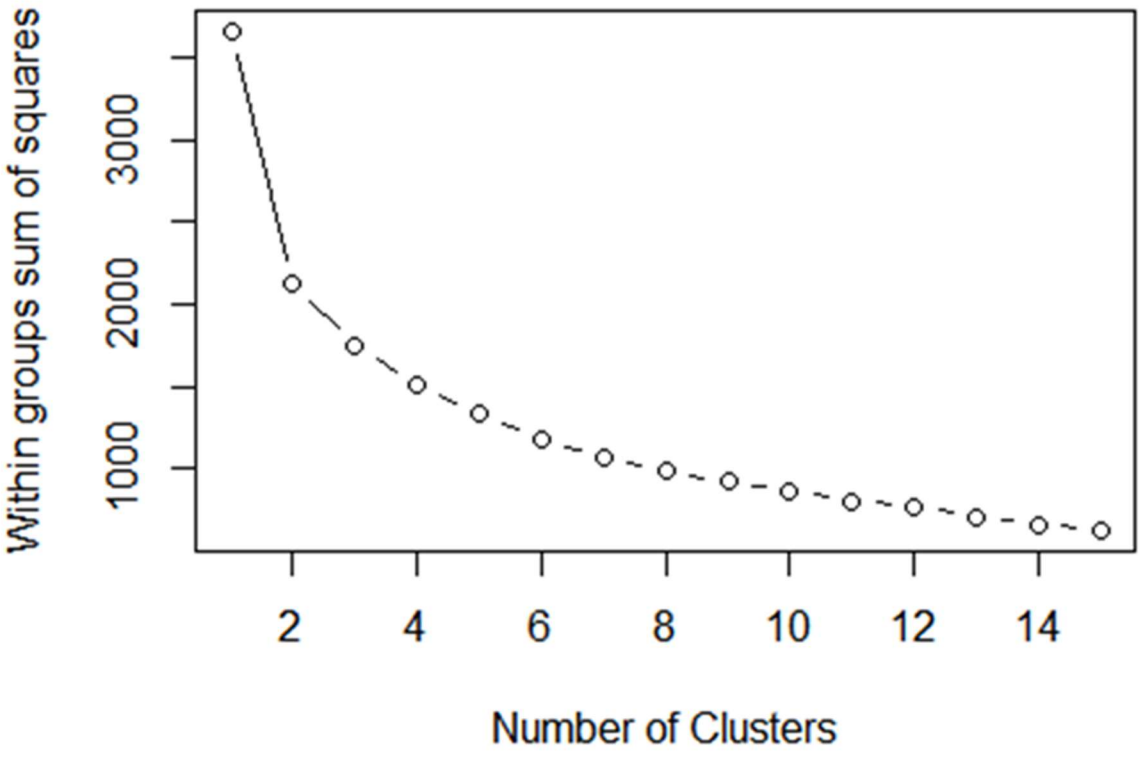

Figure 14: Scree plot - cluster homogeneity by number of centers

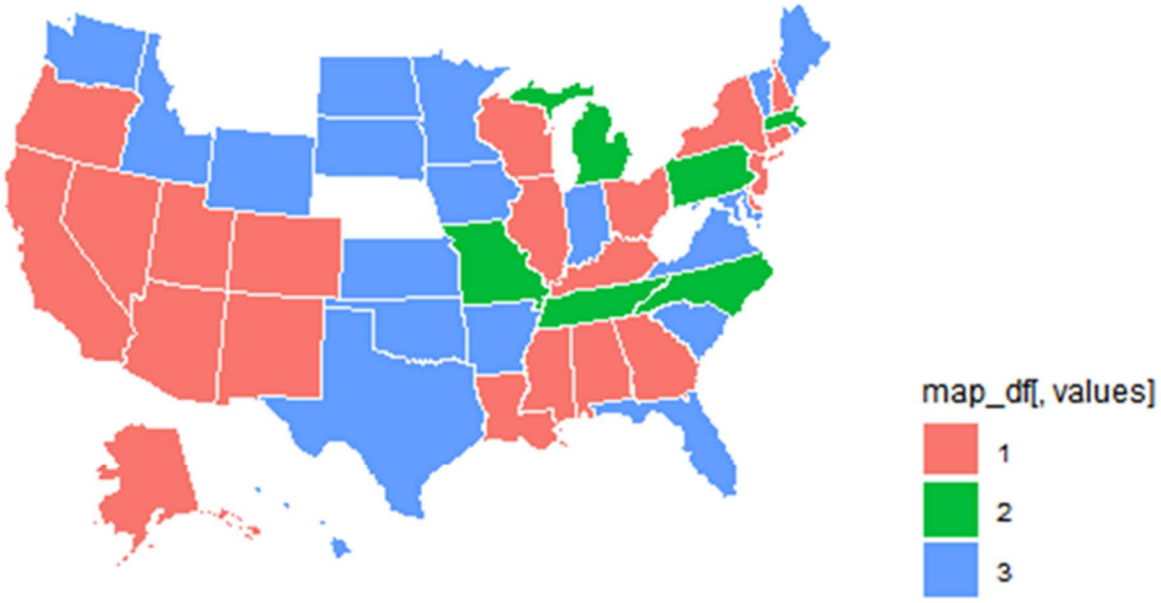

Figure 15: Clusters created by K-Means model

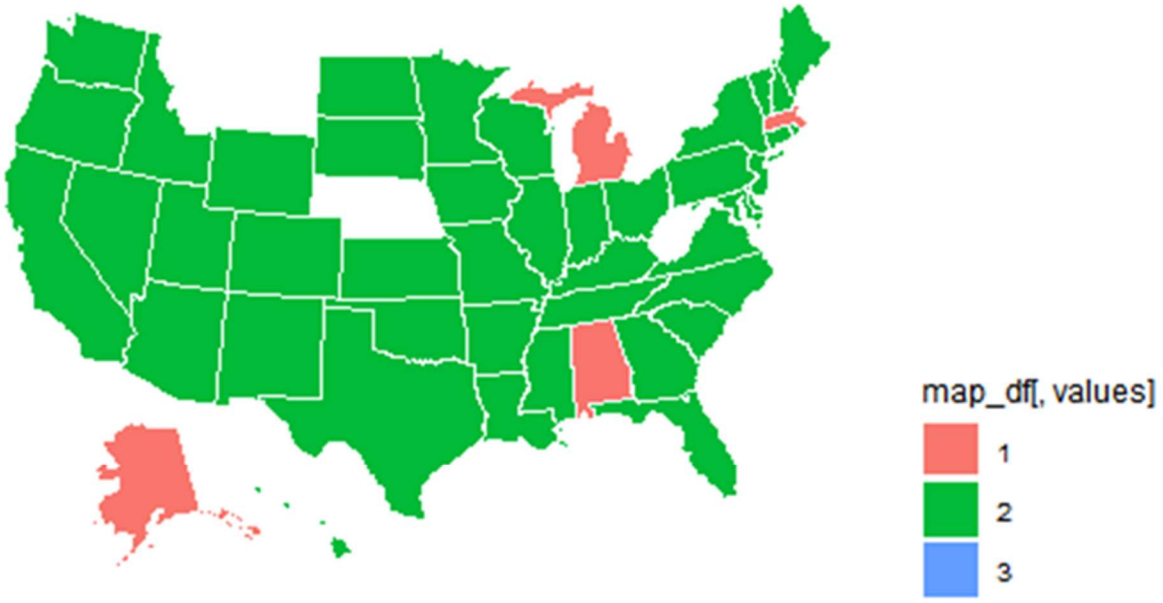

Figure 16: Clusters created by hierarchical clustering model

The K-means model did a better job at creating balanced clusters than the hierarchical model, but neither resulted in any obvious patterns. Bordering states tend to belong to the same cluster, which indicates that neighboring states have similar air qualities, but no other geographic patterns are evident.

Regression

AQI is determined by the amounts of each pollutant in the air, but is it more affected by the average density of pollutants or the maximum density of pollutants? Multiple Linear Regression modeling was used to explore the efficacy of predicting AQI based on mean and maximum pollutant densities.

Two linear models were created: one in which AQI is predicted by the mean densities of each pollutant, and one in which AQI is predicted by the maximum densities of each pollutant. The models were trained on a subset of the data using an 80/20% split before making predictions on the remaining test data.

The adjusted R^2 of the max-value model was 0.9376. Therefore, 93.76% of the variation among AQI in the training data is explained by the additive relationship between the maximum measurements of each pollutant. Similarly, the adjusted R^2 of the mean-value model was 0.7678. Thus, 76.78% of the variation among AQI in the training data is explained by the additive relationship between the average

measurements of each pollutant. From this information, we expect maximum densities to be a better predictor of AQI.

Next, the models were evaluated on the remaining test data. Predictions were made from each model and compared to the true AQI. The max-density model performed exceptionally, as 98.9% of predictions fell within one standard deviation of true AQI. 96.8% of predictions fell within ½ standard deviations of the true AQI, and 82.2% of predictions fell within ¼ of a standard deviation of the true AQI. Predictions made using the max-density model were extremely close to their true values.

The mean-density model performed more poorly: 96% of predictions fell within one standard deviation of AQI, 78.6% fell within ½ standard deviation, and only 47.4% fell within ¼ standard deviation. While it is a good predictor within one standard deviation, the mean pollutant densities are a less-effective predictor of AQI than the maximum densities.

Classification

The burning of fossil fuels is a principal cause of the emissions of the four major air pollutants. Power plants produce incredible amounts of emissions every year, and where there are people, there must be power. This leads to the question: Can we determine whether a state is urbanized based on the average air pollution?

For the purposes of this study, an urban state is one whose urban population is greater than 70% of its total population[1] as of May 2022. The data was augmented with a logical column containing TRUE if the population is mostly urban and FALSE otherwise. For several different train/test splits, a logistic regression model was fit to predict whether a measurement was taken in an urban or rural state based on the average densities of each of the four major pollutants. The results can be found in the table below:

% Split	Classification Accuracy
50/50	62.3%
75/25	61.2%
80/20	62.1%
90/10	***62.4%***

The regression model did not do very well at classifying urbanism, as even the best results were barely over 50% which would be achieved by random guessing. Perhaps the classification would be more successful at differentiating between whether the measurement was taken in a rural town or urban city, but this hypothesis will be tested in future work.

[1] US Census, https://www.census.gov/

Future Work

There is still plenty of work to do in the future for this project. First, I would like to refactor the Leaflet mapping to be one single map with a yearly slider instead of separate maps for each year. I would like to deploy this interactive map on a Shiny server and link it to my portfolio website. I would also like to work with a climate scientist to get a better understanding of the scientific reasons for the data patterns seen in this study. Additionally, I would like to add all United States fossil-fuel powerplants as an overlay group on the Leaflet maps to examine the relationship between air quality and the distribution of high-emission power plants. Lastly, I would like to examine AQI on a city-wide level rather than a state-wide level, as it is possible that there is greater AQI disparity between rural and urban cities than between different states. This will remain a living document on my portfolio website to be updated with further work and results.

Conclusions

There is often hopeless language used in the discourse surrounding climate change from those who have been told that we are nearing the point of no return. But the data shows that there is hope. There have been substantial changes over the past twenty years alone, and if top minds continue to work, it will only get better.

It seems that air pollution is a nationwide issue rather than a statewide one, since the pollution emitted by states like Texas does not stay there – it travels elsewhere, such as New York. The problem must be addressed on a national level rather than a local or state level.

Air pollution has improved over the past 21 years. All pollutants other than O3 are recording lower maximum measurements than in the past, and overall AQI seems to be improving with time. As shown in the Leaflet maps, more and more states are monitoring their air quality, and the measurements are proving healthier as the years go by.

From this data we cannot conclude why air quality is improving, whether it is climate initiatives and regulation, public awareness, EV vehicles, some combination, or something else entirely. However, it is certain that improvements are being made and we should do everything we can to continue the trend.

Bibliography

Soetewey, A. (2020, June 7). *Wilcoxon test in R: How to compare 2 groups under the non-normality assumption*. Stats and R. Retrieved May 10, 2022, from https://statsandr.com/blog/wilcoxon-test-in-r-how-to-compare-2-groups-under-the-non-normality-assumption/

Tiseo, A. (2021, July 2). *U.S. Power Plant SO2 emissions by facility 2020*. Statista. Retrieved May 10, 2022, from https://www.statista.com/statistics/1248106/so2-most-polluting-power-plants-united-states/

US Environmental Protection Agency. (n.d.). *Sulfur Dioxide Basics*. EPA. Retrieved May 10, 2022, from https://www.epa.gov/so2-pollution/sulfur-dioxide-basics

R Packages and Documentation

corrplot - https://www.rdocumentation.org/packages/corrplot/versions/0.92/topics/corrplot
ggmap - https://cran.r-project.org/web/packages/ggmap/ggmap.pdf
htmlwidgets - https://cran.r-project.org/web/packages/htmlwidgets/htmlwidgets.pdf
leaflet - https://www.rdocumentation.org/packages/leaflet/versions/2.1.0
leaflet.extras - https://cran.r-project.org/web/packages/leaflet.extras/index.html
leaflet.extras2 - https://cran.r-project.org/web/packages/leaflet.extras2/index.html
leaftime - https://cran.r-project.org/web/packages/leaftime/index.html
secr - https://cran.r-project.org/web/packages/secr/secr.pdf
sf - https://cran.r-project.org/web/packages/sf/sf.pdf
spData - https://cran.r-project.org/package=spData
sp - https://cran.r-project.org/package=sp
tidyverse - https://www.tidyverse.org/
tmap - https://cran.r-project.org/web/packages/tmap/tmap.pdf
usmap - https://cran.r-project.org/web/packages/usmap/usmap.pdf
xts - https://www.rdocumentation.org/packages/xts/versions/0.12.1