GEM RUBY
FINAL CAPSTONE DELIVERABLE

AirWise: Air Quality Insights and State/County Comparisons

TABLE OF CONTENTS

I. INTRODUCTION

Air pollution is a major concern for public health, and its impact is visible across the world. In the United States, the Environmental Protection Agency (EPA) has been responsible for monitoring and enforcing air quality standards. Despite these efforts, air pollution remains a serious issue in many areas. The Air Quality Index (AQI) is a measure of air pollution that provides information on the quality of air in a specific location. The AQI considers five major pollutants and provides a rating system that ranges from good to hazardous.

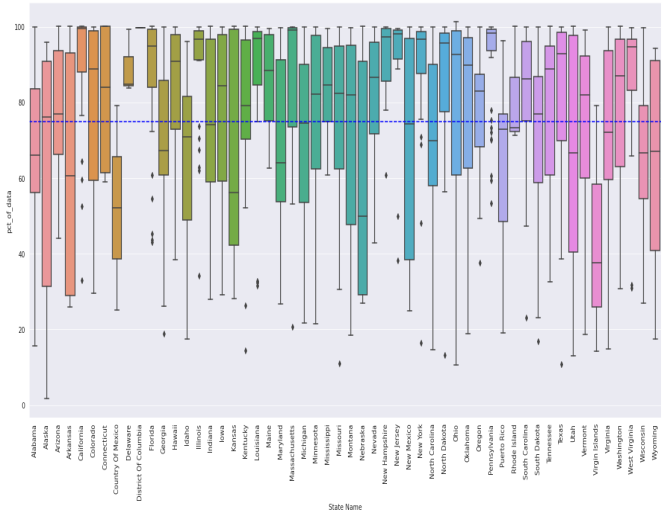
The goal of this capstone project was to develop a tool that allows users to compare air quality across different states and counties. AirWise, provides a user-friendly interface that is powered by Tableau public, that allows users to select their designated County/State and compare it to other County/State based on the AQI. AirWise uses data from the EPA website, including latest AQI readings and air quality forecasts. Below are the main purposes of this tool:

(1) Predicting air quality levels that can assist businesses that emit air pollutants in their daily operations to regulate their output based on the forecasted AQI for the week and ultimately optimizing their operations to reduce emissions and avoid potential penalties for noncompliance.
(2) Predicting air quality levels that can help individuals identify areas of high risk to their health due to poor air quality, allowing them to take necessary precautions to mitigate those risks. Additionally, real-estate developers can use AQI information to promote areas with good air quality or avoid areas where AQI may be poor, providing valuable information to potential buyers and investors.

AirWise provides individuals and businesses with a unique way to make informed decisions about where to live or conduct business or to mitigate code violations based on existing air quality information. It aims to raise awareness about the impact of air pollution on public health and encourage individuals and businesses to make more informed choices about their environment.

In the following sections, this report will detail the application of various data science techniques used to develop an AQI prediction model. These techniques include data cleaning, exploratory data analysis, and modeling. The results of the project will showcase the data science and data analysis techniques to predict and compare AQI levels across different locations.

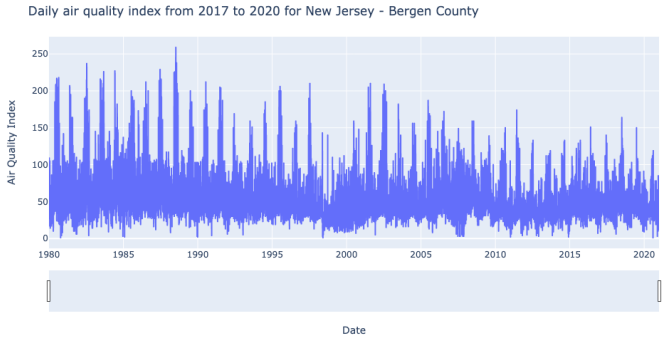
II. EXPLORATORY AND DATA ANALYSIS (INCLUDING DATA EXTRACTION AND DATA WRANGLING)



The data set used in this project was obtained from the United States Environmental Protection Agency (EPA) website, and it contains AQI data by state and county with corresponding dates, categories, and parameters used to determine the AQI. It was collected by the EPA and was provided to the public via csv file (in a zip format). The initial steps in the data cleaning and exploratory analysis involved a thorough review of all state and county combinations to determine the number of missing AQI values. The results were visualized using a box plot that displays the non-missing AQI values for each state and county combination.

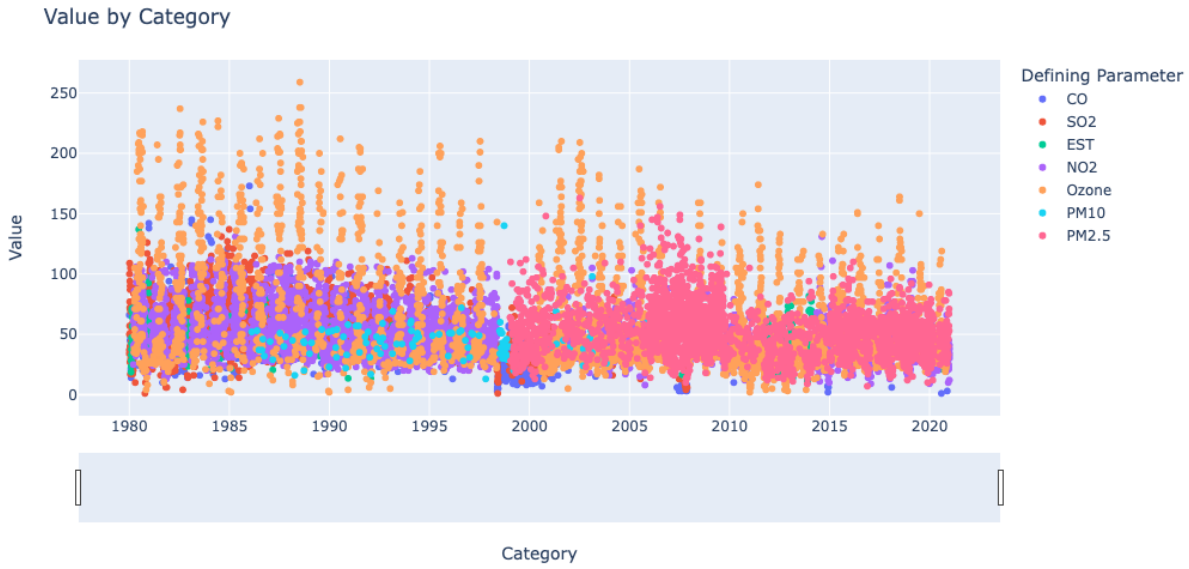
To address missing AQI values, the data set was re-indexed by date for each State/County combination and then missing AQI values were interpolated using linear interpolation. This approach allowed for a more accurate representation of AQI trends over time. Further analysis was then conducted on the complete data set.

Further analysis involved a detailed examination of the AQI trends in Bergen County, New Jersey. The review revealed significant fluctuations in AQI values over the years, but a general downward trend was observed since the 1980s, largely due to the efforts of the EPA and the Clean Air Act program. These findings highlight the importance of continued efforts to monitor and regulate air quality to promote healthy living and sustainable development.



After conducting a review of the time series graph for one county, an exploration of the parameter used to calculate AQI was conducted. It was discovered that a new parameter, PM2.5, became a prominent AQI measurement after 2000. PM2.5 refers to fine particulate matter, which is 2.5 micrometers or smaller in diameter and is produced by a variety of sources, including wildfires, power plants, and motor vehicles. Exposure to PM2.5 is associated with a range of adverse health effects, including respiratory and cardiovascular illnesses.

Based on this discovery, it was determined that the model should be trained only on data from 2000 onwards to avoid any potential issues with the dataset during modeling. As such, any data from before 2000 was removed from the analysis and to further reduce the dataset to help train the model. The graph below showcases the AQI value measured by category.

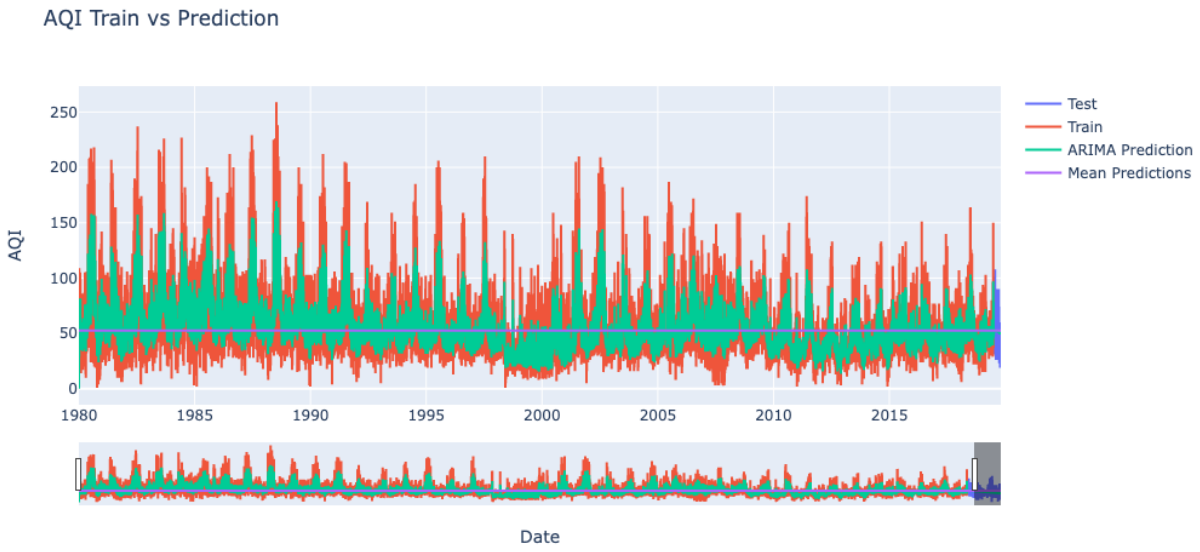


Please refer to notebook III, "Combine and Feature Extraction," where data was pulled from the EPA website to understand each component of the AQI. Although these components were not included in the modeling due to limited data, we plan to further explore the data for future use.

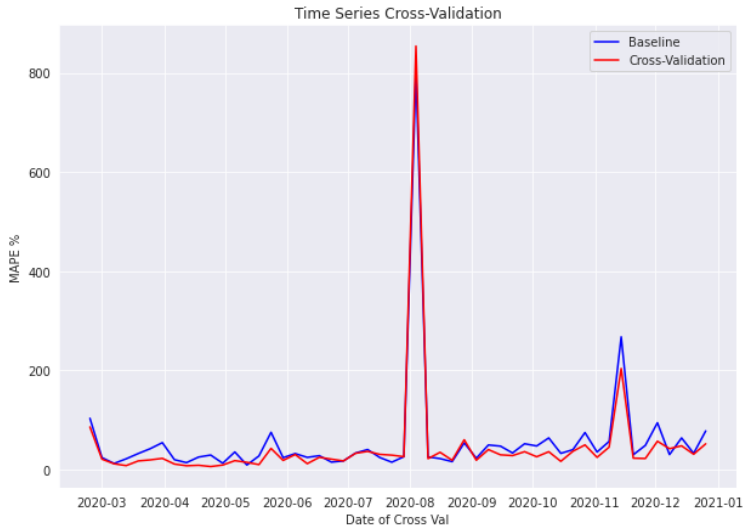
III. INSIGHTS AND MODELING USING SARIMAX

After cleaning the data set and conducting exploratory data analysis, we tested for stationarity using the Augmented Dickey-Fuller test to ensure that an autoregressive model was appropriate. Our analysis of the decomposed values revealed that for one county, AQI displayed a downward trend, with the highest peaks occurring during the summer months due to weather patterns. The residuals appeared to be stationary, except during 2000 when a new parameter, PM2.5, was introduced. Based on our evaluation, we found that the SARIMAX model was the best fit for our data and produced the best fit predictions.

We, then, we conducted a hyperparameter search using the “pmdarima” library to identify the best model for time series AQI prediction. The SARIMAX model is a variant of the ARIMA model that incorporates exogenous variables in addition to autoregressive and moving average terms. This made it particularly suitable for our AQI prediction problem, as other relevant features can be added such as weather and AQI pollutant components.



In addition to the hyperparameter search, we also evaluated our model using time series cross-validation and other statistical metrics such as mean absolute error, root mean squared error, and R-squared. In the best fit model, the scores were as follows using a rolling Cross-Validation period of 7-day forecast.

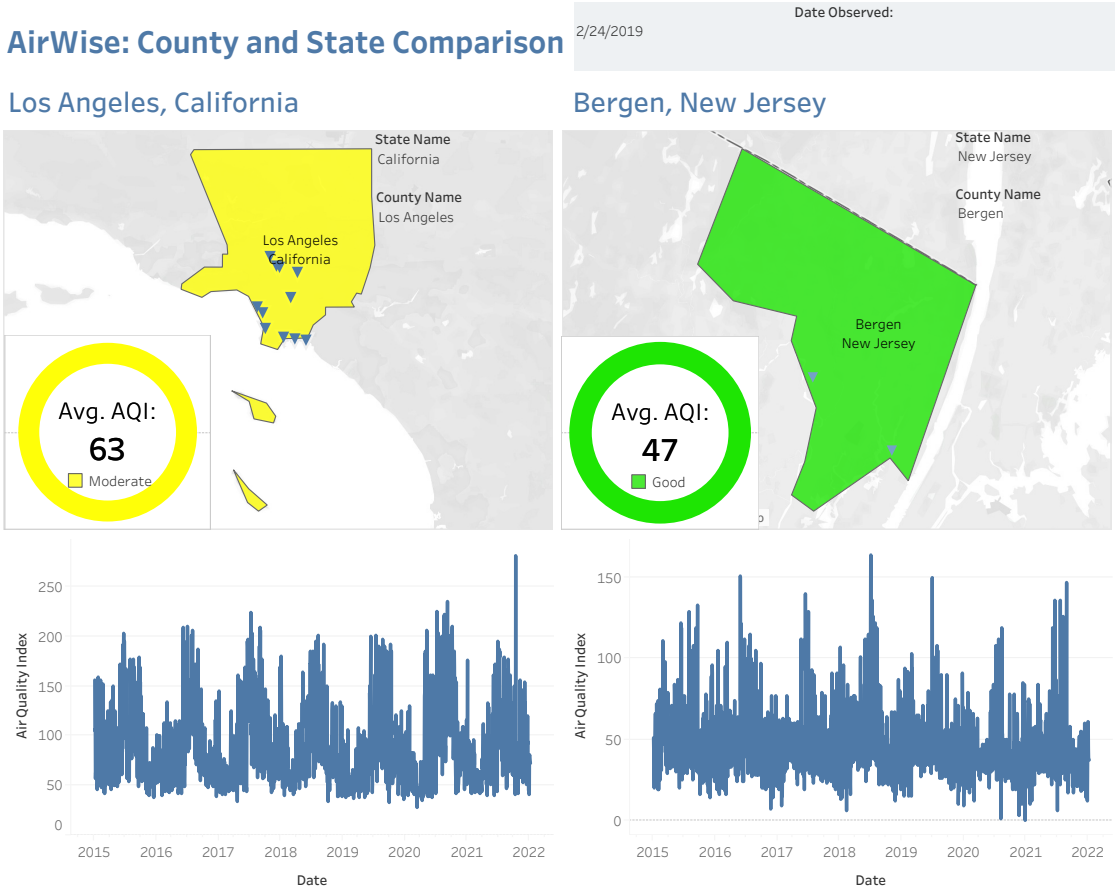


IV. CONCLUSION

After conducting extensive analysis and modeling, our results showed that the performance of our model was superior to the baseline average. However, there is still room for improvement as our model's score was higher than the desired threshold. To ensure that our model's forecast is reliable, we have included a sample of the prediction module:

index	lower AQI	upper AQI
2020-12-26 00:00:00	3.2837402030027363	70.15225113552563
2020-12-27 00:00:00	4.268028041604907	77.646379987092
2020-12-28 00:00:00	4.973952002062653	78.80443637587013
2020-12-29 00:00:00	4.903743231479176	78.85199343215356
2020-12-30 00:00:00	4.833662908757496	78.89942816506829
2020-12-31 00:00:00	4.763710420045527	78.94674118681048

The predictive model was able to forecast AQI values for a specific state and county combination within the acceptable range (the range is within two categorical intervals for AQI), as demonstrated by the sample prediction above. With these results, we were able to proceed with the development of the AirWise visualization tool, which allows individuals and businesses to compare historical and forecasted AQI values for different state and county. The dashboard includes a comparison of two counties from different states and provides users with an easy-to-use interface for exploring and comparing AQI data.



A supplementary dashboard was added to the project to display the locations of facilities with GHG reports, allowing individuals and businesses to identify potential areas of low AQI. Upon analysis, it was noted that there are more facilities located on the east coast, but the AQI values were relatively lower on the west coast. This observation can be attributed to the difference in weather patterns between the two regions, which were not incorporated into the current model. This factor will be considered in future analysis to gain a better understanding of the factors influencing AQI values across different regions.

US Map: County with Facility Locations (GHG Emissions)