Analysis and Modeling of Air Quality in the California 2020 Wildfire Season

Trevor Oldham

May 2023

Abstract

In the year of 2020 there were 4,304,379 acres burned in California during an especially dire wildfire season (1). California is known to have a seasonal pattern of large scale wildfires in a multitude of counties across the state. Despite the fact that California is one of the largest continental states, we seek to show that the presence of such intense wildfires has a marked effect on the air quality index reported on a county basis. In this report we will organize the data into a county by county basis and seek to determine the extent to which the wildfire season effects air quality and if the difference is large when compared to other states in the nation. We will then turn our attention to a subset of greenhouse gas emission data that was reported in the dataset and create a linear regression model that can predict the air quality index given the daily emissions of those gasses such as carbon monoxide, nitrogen dioxide, and ozone. Finally we will compared our models predictions over the state of California against predictions made for other states that do not have a perennial wildfire season and contrast the accuracy of our model in both cases.

Introduction

The Air Quality Index (AQI) is a quantitative scale that is calculated based on the five major components which are carbon monoxide, nitrogen dioxide, sulfur dioxide, ground-level ozone, and particulate matter. Of these substances, it is said that the most harmful are both ozone and particulate matter, whereas carbon monoxide, sulfur dioxide, and nitrogen dioxide rarely affect the AQI (2). Wildfires increase the local particulate matter and can spread poor air quality towards neighboring states as well, and the rising temperatures as a result of both wildfires and climate change increases the rate that harmful substances in the air are converted to ground-level ozone. Ozone is also a result of increased vehicle activity. In the first part of this paper we will rank the US states by air quality to determine under which metric would California be the worst, and we will then show that the level of ozone is the most impactful greenhouse gas in the calculation of AQI. In the second part of the paper we will create a more simple model that predicts the daily AQI by just ozone concentrations alone on that day, and further investigate the effect that traffic volumes contribute to the concentration of ozone and therefore the AQI as well.

This research is relevant because we must take careful considerations of the various gas emissions in order to be better stewards of our planet. The AQI is simply an index which calculates a quantitative scale to rank the quality of a given day in a given location. In this paper, we consider the data as it is presented using a convenience sample first in order to better understand the significant and non-significant factors which contribute to air quality. We do not consider the explicit equation for calculating AQI because we wish to find those relationships from the data as presented to us. After determining the most significant emissions we will suggest that the calculation for AQI can actually be simplified and even predicted based on a smaller subset of the emitted gasses. In the end we hope to arrive a reasonable estimation of air quality as determined first by advisory days and then second by concentrations of the most significant gasses. This will provide a reasonable framework for making suggestions for people and businesses to use in their

own efforts to reduce greenhouse gas emissions and improving their local air quality. As a whole we hope to learn if the particulate matter released from seasonal wildfires has any significant effect on the air quality and determine under what conditions a weather station would report higher AQI on a day to day or annual basis.

Analysis of AQI By State

The first dataset to analyze contains the yearly data of weather station measurements across the states and counties in the USA, including Mexico and the US Virgin Islands. using this data we will be interested in the columns which measure the days on which there was an advisory for each of the four greenhouse gases, and we are also interested in the median AQI for each weather station in the records. Grouping by state and aggregating the Median AQI by taking the maximum results in a sorted list that shows the ten best and worst states by AQI, among which California is the worst, reporting a maximum countywide AQI of 1928 and the US Virgin Islands is the best, reporting a maximum AQI of just 82.

The second dataset contains daily AQI measurements for each weather station in the records on any given date throughout the year of 2020, as well as the category of that day which is drawn from the options of Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. **Figure 1** shows the resulting chart which demonstrates the average AQI by category for the ten worst states as previously determined. A similar process reveals the same data for the ten best states, among which the average AQI does not ever surpass 120.

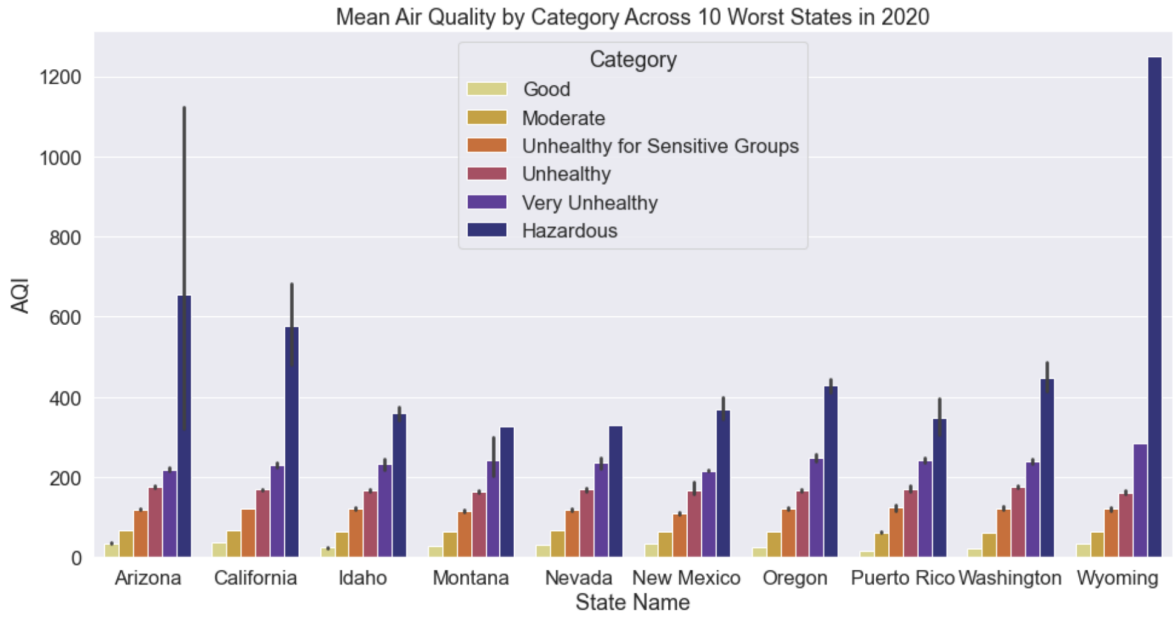

Figure 1: Mean Air Quality Index across 10 worst states. Each bar height represents the average reported value of a given category of a given state.

Next the daily AQI data is grouped into bins for each day in the records, and the mode is taken to find the most commonly reported category on that day in that state. California is taken to be the worst state and is compared to the US Virgin Islands which is determined to be the best state by AQI. **Figure 2** shows the resulting pie chart which visualizes the difference between the worst and best state. This is interpreted to the percentage of days in which each category was the most commonly reported value among all weather stations in the records. It is evident that in California, on 25% of the days among 2020, the most commonly

reported category was Moderate or worse, and even on 5.8% days the most commonly reported category was Unhealthy of Unhealthy for Sensitive Groups. The US Virgin Islands shows the contrasting results, which reflect the smaller likelihood of wildfires, smaller traffic volume, less industry, and coastal weather.

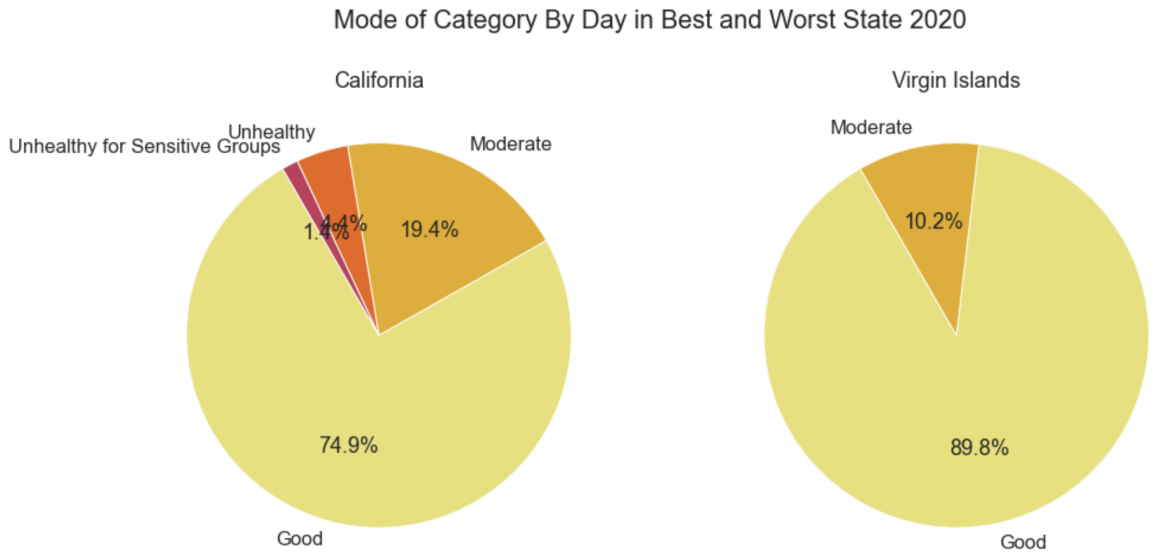

Figure 2: The most commonly reported category on each day in California and US Virgin Islands.

Next the annual AQI data is grouped by State and the median is taken among all counties in the state to determine the median AQI in that state during the year 2020. A similar process is followed to determine the average median AQI for each county in the state of California and the results are sorted to find the ten worst counties, among which were the counties of San Bernardino, Mono, Los Angeles, Riverside, and Kern. This data is plotted as a histogram in **Figure 3** and overlaid on top of the countrywide data by state to visualize the stark differences between California and the rest of the nation. This plot shows a serious shift to the right among the histograms bins as well as the kernel density estimator, suggesting that counties in California have much higher AQI than other states in the nation. This supports the claim that wildfires and traffic volumes are the most influential factor in determining the air quality index, and it also suggests that the rest of the country is a better place to live by this metric.

The data suggests that the air quality in California counties is on average much higher than the rest of the United States. But how does the air quality in California compare to itself over the years in which there were less active wildfires? Two other datasets were gathered from (3) which contain the annual AQI data for both 2021 and 2022. The data was processed by selecting the counties in California and grouping the data by average median AQI in those counties similar to **Figure 3** and plotted against the year 2020 to investigate the differences that a wildfire season has on the air quality. Surprisingly the data in **Figure 4** suggests that 2020 was in fact worse as measured by the kernel density estimator, yet it was not much worse - the histogram suggests that the air quality was close to equal in the three years of 2020, 2021, and 2022. This leads to a conclusion that California has poor air quality whether or not there was a serious wildfire season. There is even a second model spike towards the higher range of AQI in the year 2022, which corresponds to the worst counties and could be a result of either the wildfires or the traffic volume. Also, since we determined that the worst counties in California included San Bernardino and Los Angeles county, we infer that the traffic volumes of these metropolitan areas is also a significant contributor to the air quality of those counties, thus raising the average air quality across the state.

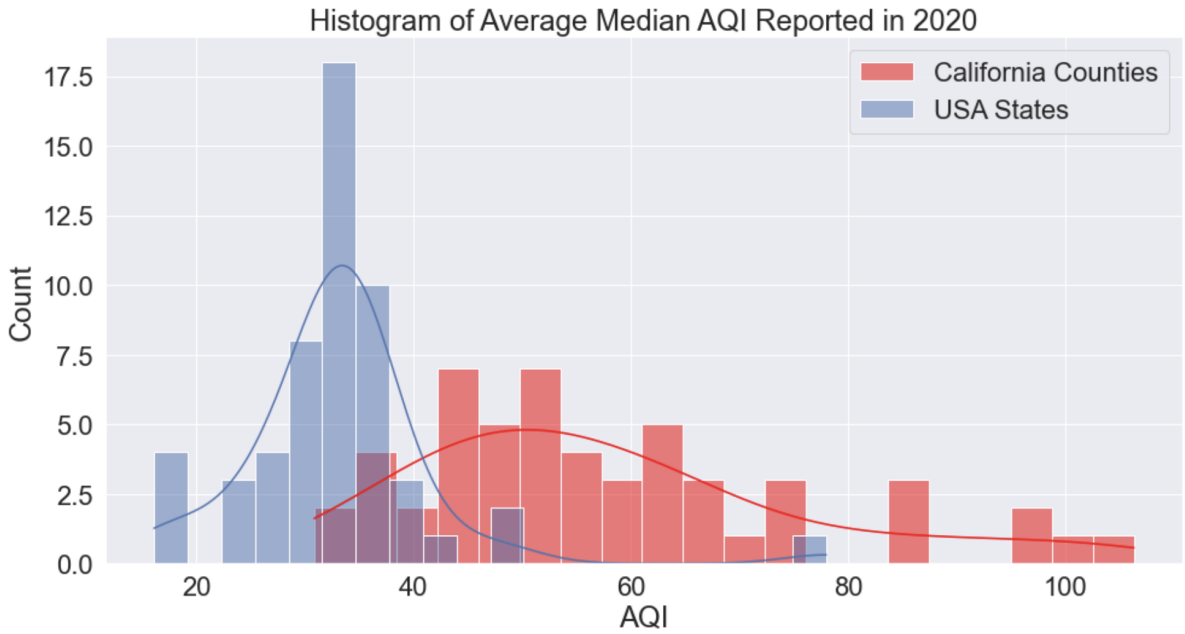

Figure 3: Histogram of Average AQI across California counties and USA States.

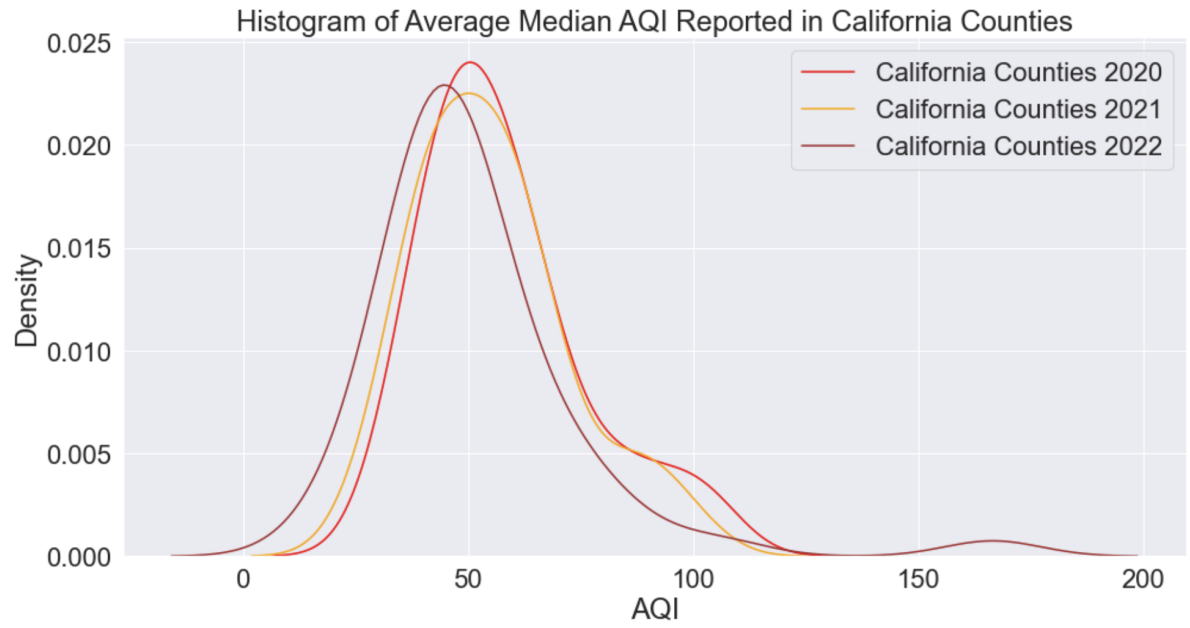

Figure 4: Kernel Density Estimator of average median AQI reported in California counties over the years 0f 2020, 2021, and 2022

Modelling AQI by Greenhouse Gas Measurements in 2020

By using the data in the annual report, a model is created to predict the median AQI of the year in a given county by the number of advisory days recorded for each of the four greenhouse gasses: carbon monoxide, nitrogen dioxide, ozone, and sulphur dioxide. Also included is the advisory days for particular matter below the threshold of 2.5 microns. This naive model can reveal whether these are appropriate features and how

effective they are in determining the AQI by county, and a better model will be fit as well using only the most correlated substances and training on the daily measurements of those gasses to predict the AQI on a given day, which should generalize better across all the counties among all the states. Once we determine that this model will underpredict the AQI in a county with heavy particular matter from a wildfire we will infer the extent to which particular matter accounts for changes in AQI, as well as determining which of the greenhouse gasses included is most correlated to AQI.

The annual data was reduced to those five features and indexed by county, then the duplicate values were dropped for counties that share a name with another county in the United States. Outlier records with AQI less than 10 or more than 60 were dropped so that the model can focus on the most average counties across the US and generalize to counties in that range. An sklearn Linear Regression (4) model was fitted to the county data and used to predict the median AQI for a given year using those features. The mean of this data shows that ozone advisory days were most common with the average county reporting 187 days. The least common was for carbon monoxide at just 0.28 days for the average county.

The data was transformed to mean-centered data and then split into a training set and a test set. The model was fit to the training set and performed a prediction on the test set, resulting in the data from **Figure 5**. As measured by the root mean squared error, the testing model slightly outperformed the training model and the residual was calculated and plotted against the test set. This data suggests that the model has under-predicted the AQI for counties in the test set with higher AQI. Notice that the model is only trained on counties that did not have extreme values for median AQI, so it is inferred that the model does not have enough training to predict values of AQI for counties in California, and thus would always under-predict the severity of the AQI in those counties, even as the training set is updated to contain outlier values. This pattern shows that the AQI in California is so much higher on a county basis than the rest of the US states that we need to consider it as outliers in order to train the model for an average county.

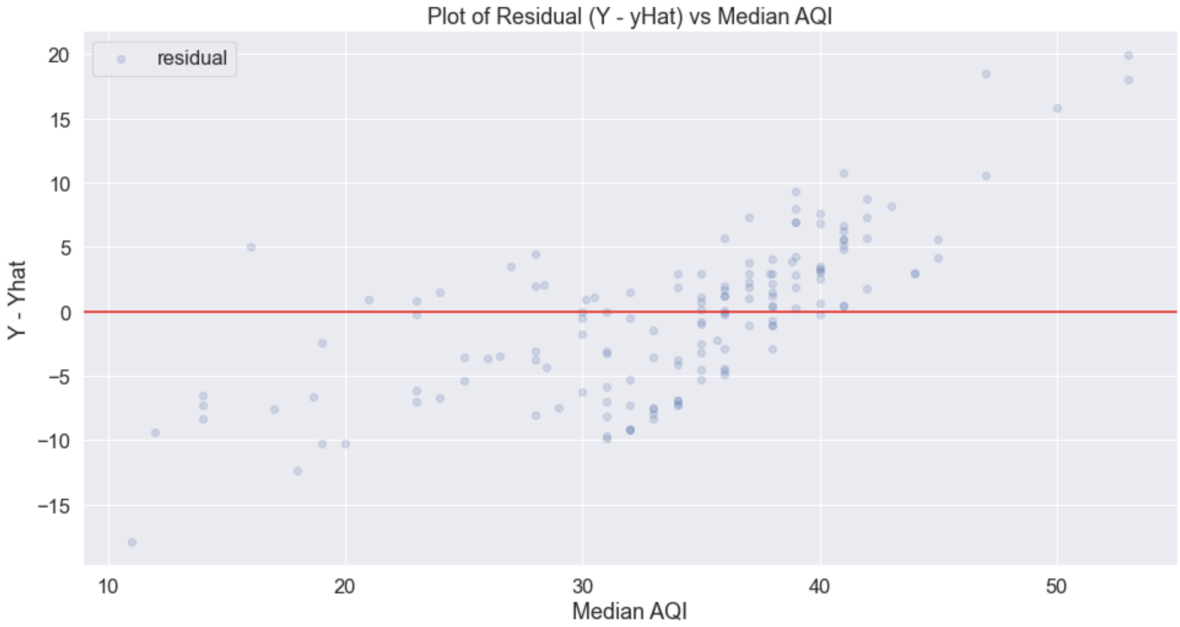

Figure 5: Plot of residuals from the linear regression prediction on US counties in the test set.

A heatmap is generated from the five features and plots the correlation of the features, revealing the extent to which each affects another. **Figure 6** shows that the only feature that is positively correlated to the median AQI is ozone. Also, ozone and particulate matter are negatively correlated with one another. To create a second model that is more effective at training and predicting on the daily measurements of gasses,

we can isolate the gound-level ozone as the most important factor in predicting AQI on a given day. This is corroborated by sources indicating ozone as the most important factor in determining the air quality index (2), to the point that some weather stations do not even record the others. This explains why the advisory days for those substances are zero for most of the counties in the US.

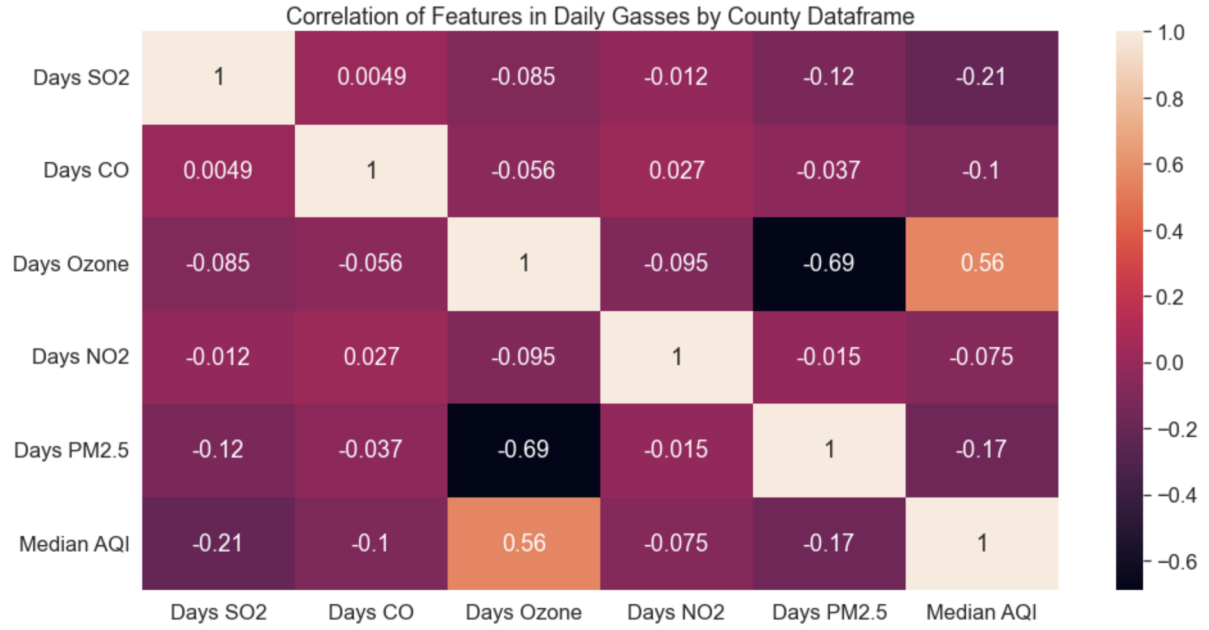

Figure 6: Heatmap of correlated features used in the linear regression model.

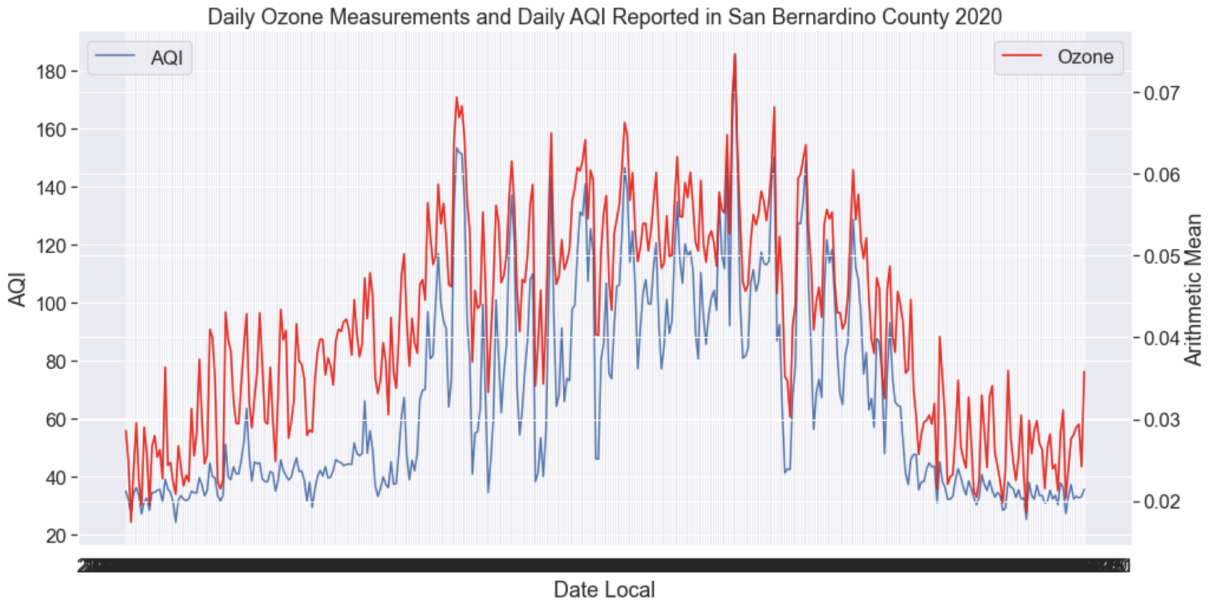

Figure 7: Daily ozone measurements and daily AQI from San Bernardino county in California. The vertial axis on the left represents AQI and the right axis represents the arithmetic mean of ozone concentration.

The data gathered from the project repository has a daily measurement of each greenhouse gas which is read

into a dataframe in which the records contain the state name, county name, AQI, and arithmetic mean of ozone concentration as measured at the weather station that day. The data are then reduced to the weather stations in California and then futher reduced to weather stations in San Bernardino County which was previously shown to be the worst county by AQI. The data are then grouped by the date and the average of AQI is taken along with average ozone concentration across the stations on that day. The ozone concentration on the given day will be the single input to our model, and the training and test sets will be the picked among the 366 weather stations in San Bernardino county. **Figure 7** shows the plot of mean ozone concentration and mean AQI across the county for that day, indicating that there is a correlation between ozone and AQI as well as a very clear seasonal variation throughout the year. The second simplified model was created with an sklearn Linear Regression class which takes as input the arithmetic mean of the ozone concentration from a weather station on a given day and predicts the AQI on that day using ozone as the lone feature. In order to model a non-linear relationship which was observed through exploratory data analysis we chose to transform the AQI measurement by taking the natural logarithm of the measurement. The results of the new model are shown in **Figure 8**.

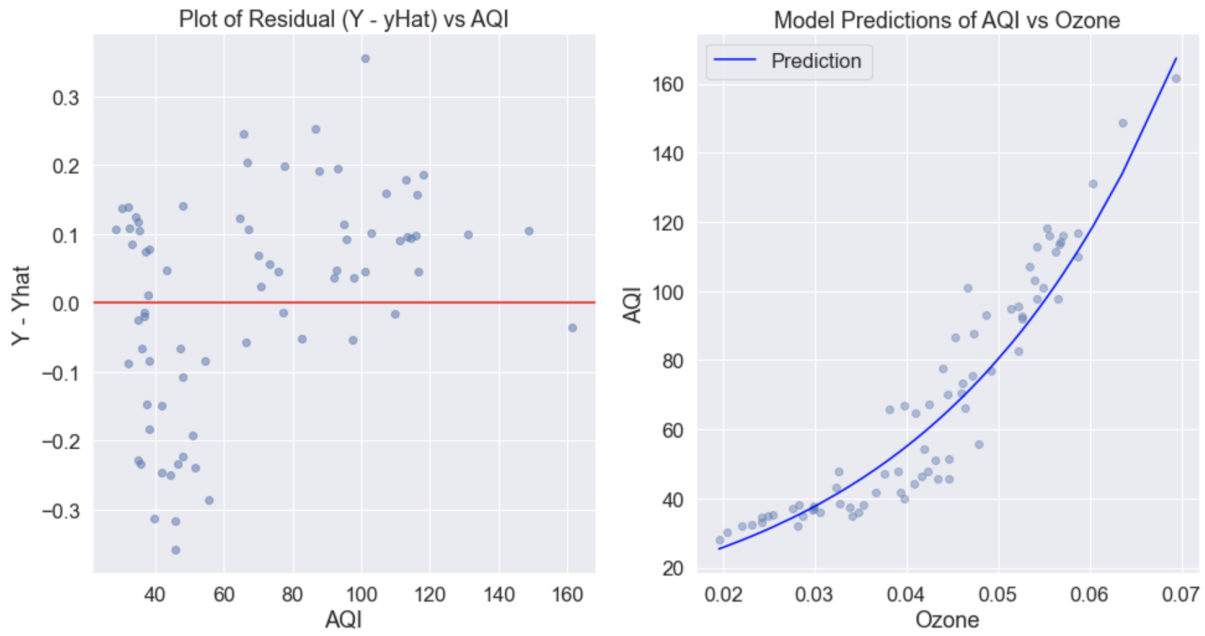

Figure 8: Predictions from the simplified model and residuals plotted over the test set data from San Bernardino county.

Results and Discussion

After analyzing the AQI data across counties and states in the US it becomes clear that California is among the worst as determined by many different metrics. A comparison of the median AQI across California counties suggests that the state has a much higher AQI in general than the rest of the states. Data also suggests that days in California during the 2020 wildfire season had a marked effect on the AQI when compared to 2021 and 2022. However, we have determined that the fire season has less of an effect on the AQI than previously inferred, and that the counties which report higher average AQI have a tendency to be counties with high commuter traffic. Examining the greenhouse gas advisory days allowed us to determine that ozone measurements are most responsible for the air quality as measured in AQI. Without considering the equation that calculates the air quality index, we were able to single out ozone as the major contributor and train a naive model which would predict the AQI given the count of advisory days in a given county. Such a model

was of less use when implemented to predict counties especially in California which are typically in the higher range of AQI. It was inferred that traffic pollution, and not wildfires, are the major component in production of ground-level ozone and a new model was fit draw conclusions about the relationship between heavy traffic and ozone levels, which also reveals seasonal patterns as a result of heavy spring and summer traffic volumes.

The second model in **Figure 8** is able to predict the AQI on a given day just by considering the ozone concentration measured at the weather station that day. This comes as no surprise because we already know that the air quality index is calculated directly from the concentration of ozone, carbon monoxide, sulphur dioxide, nitrogen dioxide, and particulate matter. However, this model is able to demonstrate the relationship between ozone concentration and AQI is not linear, and furthermore the model demonstrates that ozone concentration alone is enough to accurately predict the AQI at a given weather station. Drawing on the weather stations within San Bernardino county as a training set allows us to model the greenhouse gas emissions for a heavy polluting county, which was determined to be the worst county in California by mean AQI. By reconsidering the initial research question about how wildfires affect the AQI across the state of California we come to a surprising conclusion. The effect of wildfires does in fact shift the kernel density estimator to the right when compared to previous years with less wildfire activity, but the wildfire season is not a factor when examining the worst counties in the state, which have been proven to be mostly affected by ozone from polluting activities. The reasons for this are two-fold. Firstly, wildfires are isolated in a single county or a small group of counties and only affect a small percentage of the weather stations in the state. Second, the human polluting activity in populous states like San Bernardino county and Los Angeles county have emissions that last all year, and even increase during the spring and summer. The model suggests that the most effective course of action is reducing the air quality index is to reduce the emission of ozone.

According to the Arizona Department of Environmental Quality, commuter traffic - including boats and construction vehicles - is responsible for up to half of greenhouse gas emissions, overshadowing the emissions from heavy industrial activity. Furthermore, while cars do not directly emit ozone, the ozone is formed through a series of chemical reactions in the air from the hydrocarbons and nitrogen oxides that are emitted from vehicle exhaust (5). Some suggestions from this source are that because individual vehicles account for half of all emissions, especially in the summer, there are many ways for an individual to reduce their own emissions by carpooling, limiting their driving during rush hour traffic times, postponing or combining errands, and limiting the amount of time spent idling a car. The original research question about the wildfire season was shown to be less of an issue that originally thought, and the data actually suggest that the most populous states are worse day by day when measuring the AQI. Further research could be to look into the heaviest wildfires over the years, and single out that specific county to see if there is a spike in particulate matter and AQI during the wildfire season, but for now we see it sufficient to say that when considering counties by AQI, we must first look at the population and traffic volumes to predict the daily AQI trends, rather than looking at the wildfire activity. This paper further shows that each person has a responsibility and the capability to improve the air quality by consuming less fuel in their personal vehicles and limit vehicular activity to days in which there is no ozone advisory.

References

(1) “2020 Incident Archive.” Cal FIRE, https://www.fire.ca.gov/incidents/2020/.

(2) ”Understanding the air quality index (AQI).” Minnesota Pollution Control Agency, https://www.pca.state.mn.us/air-water-land-climate/understanding-the-air-quality-index-aqi.

(3) “Download Files — AirData — US EPA.” Epa.gov, 2015, aqs.epa.gov/aqsweb/airdata/download_files.html.

(4) “Sklearn.linear_model.LinearRegression — Scikit-Learn 0.22 Documentation.” Scikit-Learn.org, 2019, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

(5) “Cars and Pollution — Ozone — Air Division — ADEQ.” State.ar.us, 2019, www.adeq.state.ar.us/air/planning/ozone/cars