# Predicting the Number of Ozone Exceedance Days for National Parks with Monitoring Sites

Gillian Tatreau
Department of Data Science, Bellevue University
DSC580: Applied Data Science
Amirfarrokh Iranitalab
7 January 2024

**Business Problem**

Ozone is a pollutant and major component of smog. It also poses a significant danger to living cells- damaging the tissues found in the respiratory tract and in the leaves of plants. It forms in the chemical reactions between pollutants, such as car emissions, industrial processes, factories, fossil fuels, combustion, and others (California Air Resources Board). More ozone is formed when an area experiences warmer temperatures and calm weather. Some possible reasons for an increase in the amount of ozone in an area might be increased temperatures, less rainfall, increases in the number of pollutants in the area, and wildfire. The NPS is dedicated to preserving the unique and fragile natural resources in the US; therefore, an increase in the amount of ozone in the air in the national parks poses possible dangers to both the ecosystem that the NPS is devoted to protecting and the tourists visiting the parks.

**Background/History**

The National Park Service was founded in 1916 to preserve the natural and cultural resources of this country for the enjoyment, education, and inspiration of visitors today and future generations. The NPS manages 425 sites, of which 63 are designated National Parks. Of those 63 National Parks, 45 are represented in the data used for this study.

An ozone exceedance day occurs on each calendar day when the daily maximum 8-hour average is greater than or equal to 71 ppb.

**Data Explanation**

The data comes from the National Park Service (National Park Service). The collection includes data from 2008-2022. For the years 2008-2015, there is one file that has data for national parks that experienced ozone exceedance days during the ozone monitoring season (April-October). For 2016-2020, there is also a file that includes data for all monitoring sites for the entire year. The data for 2021 only includes the months January-October and 2022 includes January-May. The data format for all the files is the same, with the only exception being the files for the entire year would have additional columns for the rest of the months of the year.

*Data Dictionary:*

- Park Code: 4 letter abbreviation of the national park (str)
- Unit: national park name (str)
- Site: location of the monitoring site (str)
- Apr: number of ozone exceedance days in April (int)
- May: number of ozone exceedance days in May (int)
- June: number of ozone exceedance days in June (int)
- July: number of ozone exceedance days in July (int)
- Aug: number of ozone exceedance days in August (int)
- Sept: number of ozone exceedance days in September (int)
- Oct: number of ozone exceedance days in October (int)
- Total: total number of ozone exceedance days from April-October (int)
- Max8hr: daily maximum 8-hour average ozone concentration for each site (int)
- 4thHi8hr: annual fourth-highest daily maximum 8-hour average ozone concentration for each site (int)
- Year: year the data was collected (int)
- Site ID: unique monitoring site ID, result of combination of Park Code and Site (int)

To prepare the data for use, the extra months present in the 2020 and 2021 data were removed, and this data was filtered such that only sites that had one or more ozone exceedance days were included, to be congruous with the other data. Column names were uniformly formatted before the data was combined into a single data frame to include all the data. Year was added so that the year data was retained within the larger data frame. Site ID was added to provide an ID for each unique site present in the data.

**Methods**

A combination of graphical analysis and statistics will serve to illustrate the relationships between sites. For predictions, ARIMA models will be built for every site that has data for all 14 years present in the data, as well as for generalized models for all the sites. As an attempt to improve model performance, the auto_arima() function from the pmdarima library was utilized, but it decreased performance for every model except for model1 (the median of the total number of ozone exceedance days).

**Analysis**

There are 72 unique Site IDs within the data, which corresponds to the fact that many of the national parks possess several monitoring sites each. Of these 72 unique sites, only 4 occur every year: Sequoia and Kings Canyon National Park- Ash Mountain, Yosemite National Park- Turtleback Dome, Sequoia and Kings Canyon National Park- Lower Kaweah, and Joshua Tree National Park- Black Rock. All four of these sites are in California. Of these 4 sites, both Sequoia sites (Ash Mountain and Lower Kaweah) are in the top 5 for total number of ozone exceedance days every year. There are 4 other sites that also frequently make the top 5 number of ozone exceedance days per year, which would be Joshua Tree National Park- Cottonwood Visitor Center, Mojave National Preserve- Kelso Mountains, Carlsbad Caverns National Park- Biology Building, and Great Smoky Mountains National Park- Look Rock. These 8 sites represent the national parks with the highest number of ozone exceedance days each year. 7 of these 8 sites are in the southwest region of the United States, and the Great Smoky Mountains is in the southeast region of the United States. There are not visible trends in the data for the total number of ozone exceedance days, daily maximum 8-hour average ozone concentration, and annual fourth-highest daily maximum 8-hour average ozone concentration across all 8 sites.

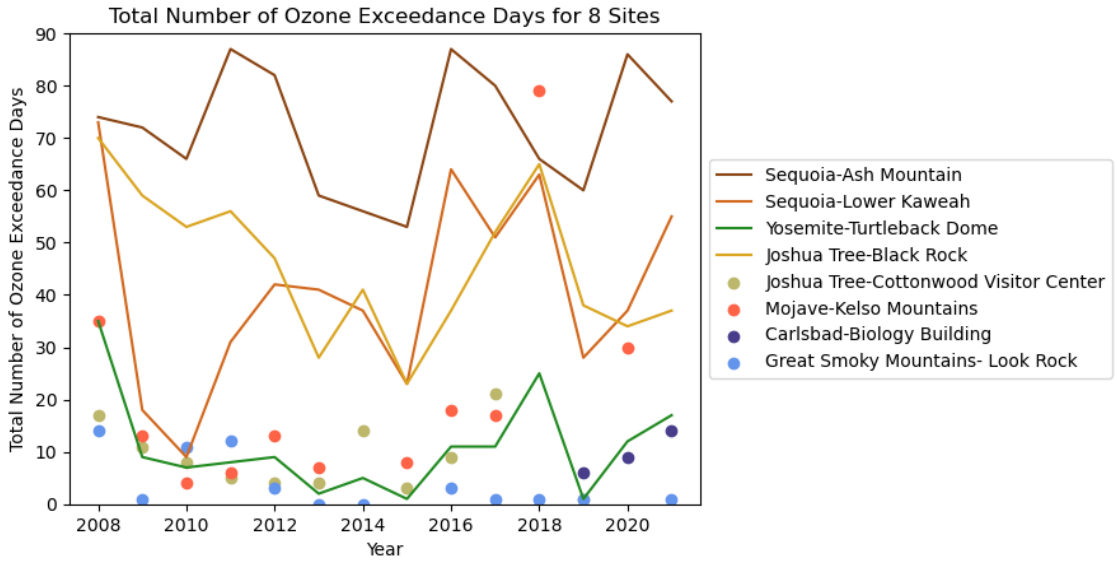



*Figure 1*. Total number of ozone exceedance days for each of the 8 sites.

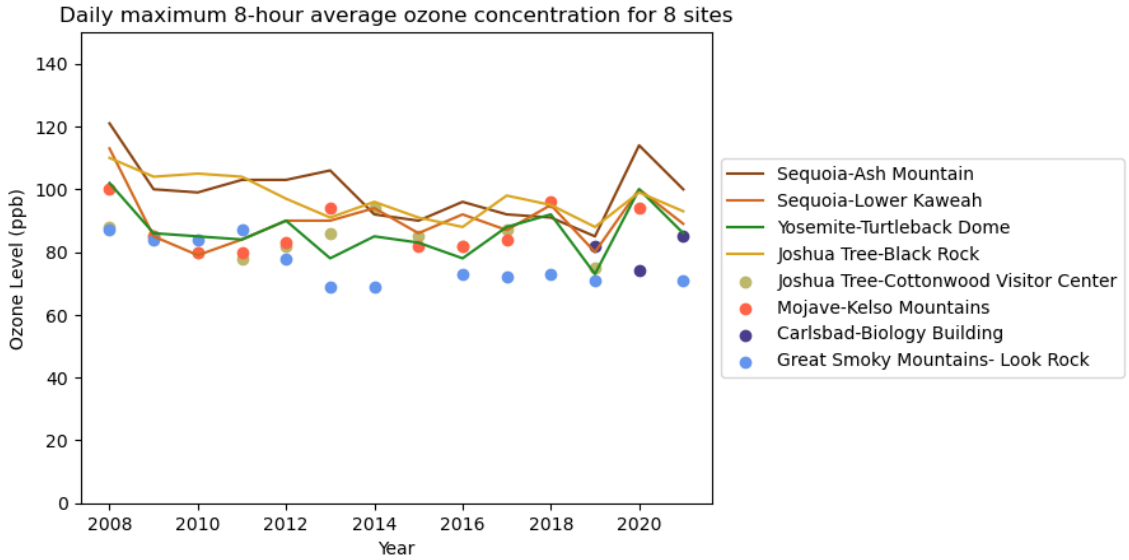



*Figure 2*. Daily maximum 8-hour average ozone concentration for each of the 8 sites.

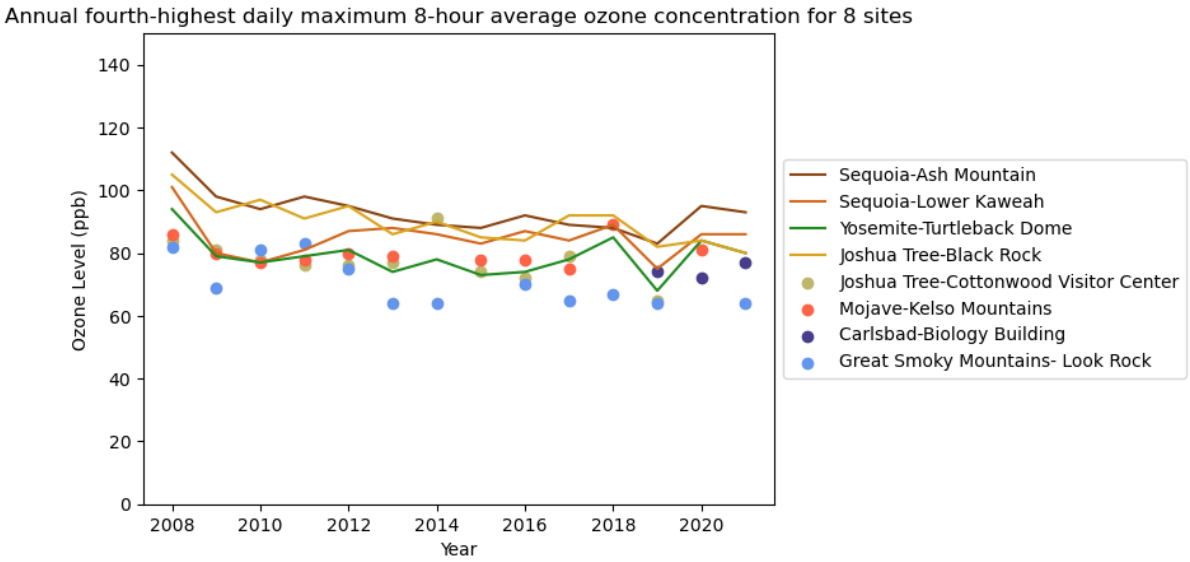



*Figure 3.* Annual fourth-highest daily maximum 8-hour average ozone concentration for each of the 8 sites.

Any trends in the data could be a reflection in the change in temperature. The average yearly temperature for Sequoia National Park is plotted against the total number of ozone exceedance days for both Sequoia sites (Figure 4). There appears to be a relationship between the temperature and the number of ozone exceedance days, as they follow similar trends. For example, there is a peak in all three plots that occurs between 2015 and 2017, and there is an increase in all three plots in 2020. The Ash Mountain site follows the temperature trends slightly more closely than the Lower Kaweah site.

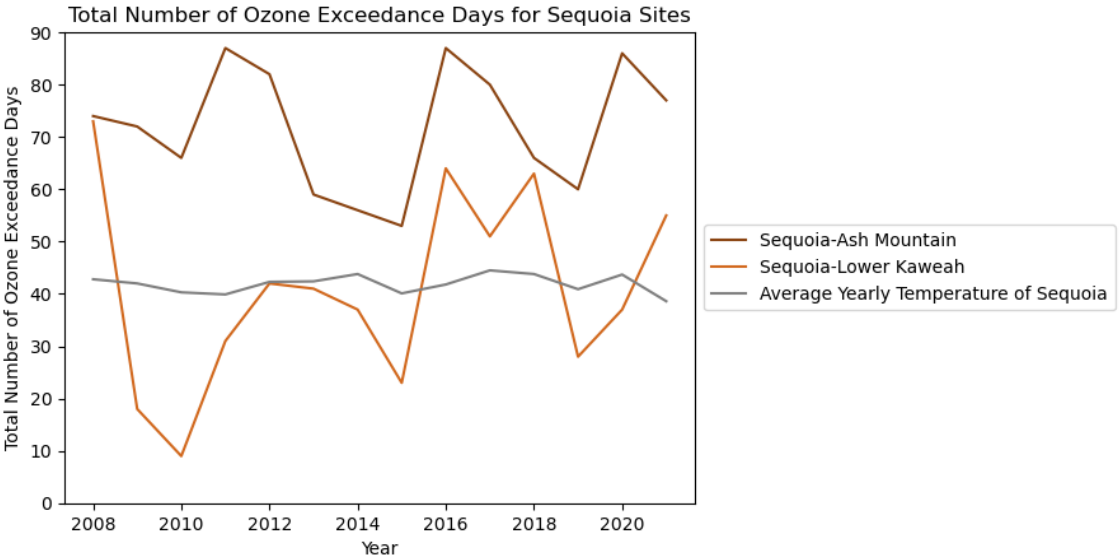



*Figure 4.* Total number of ozone exceedance days for both Sequoia monitoring sites and average temperature of Sequoia National Park.

The time series data appears stationary for all sites across the three values, without any obvious seasonality or other trends; however, after the augmented Dickey-Fuller test is applied, only the Max8hr and 4thHi8hr for Sequoia and Kings Canyon National Park- Ash Mountain; and Total and 4thHi8hr for Yosemite National Park-Turtleback Dome were proven statistically stationary. When the generalized data is examined closely, there are some very slight negative

trends visible in the average daily maximum 8-hour average ozone concentration and average annual fourth-highest daily maximum 8-hour average ozone concentration plots.

Models were created for each of the three values in the data: total number of ozone exceedance days, daily maximum 8-hour average ozone concentration, and annual fourth-highest daily maximum 8-hour average ozone concentration. For verification, the data for 2020 and 2021 were withheld from training the model and were used to calculate the RMSE of the model from predictions for those two years. The weighted mean absolute percentage error (WMAPE) was also used for additional verification of model fit.

*Sequoia and Kings Canyon National Park- Ash Mountain*

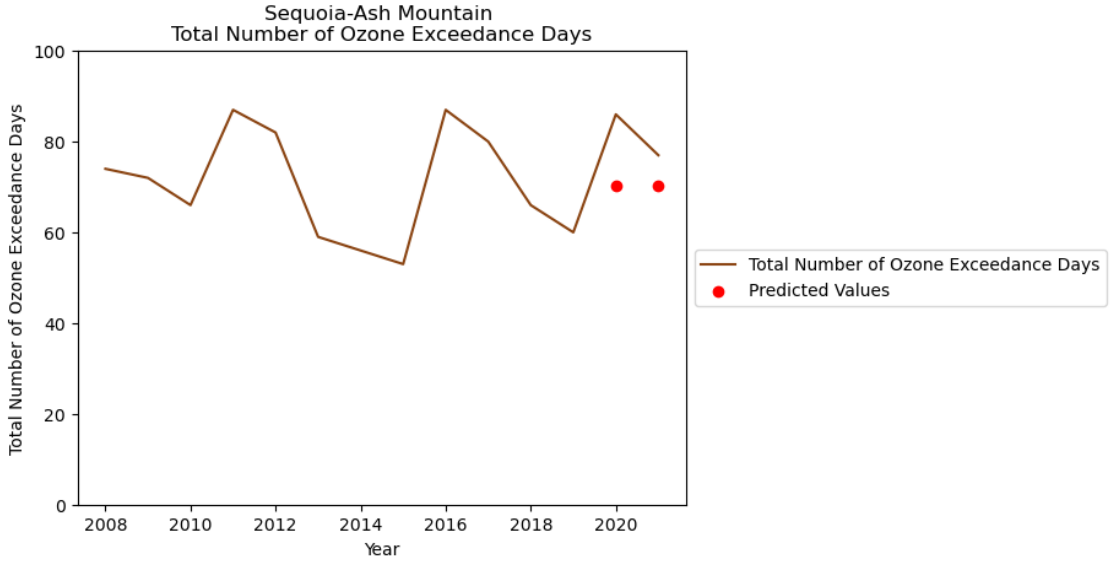



*Figure 5.* Results from model for the total number of ozone exceedance days at the Sequoia and Kings Canyon National Park- Ash Mountain monitoring site.

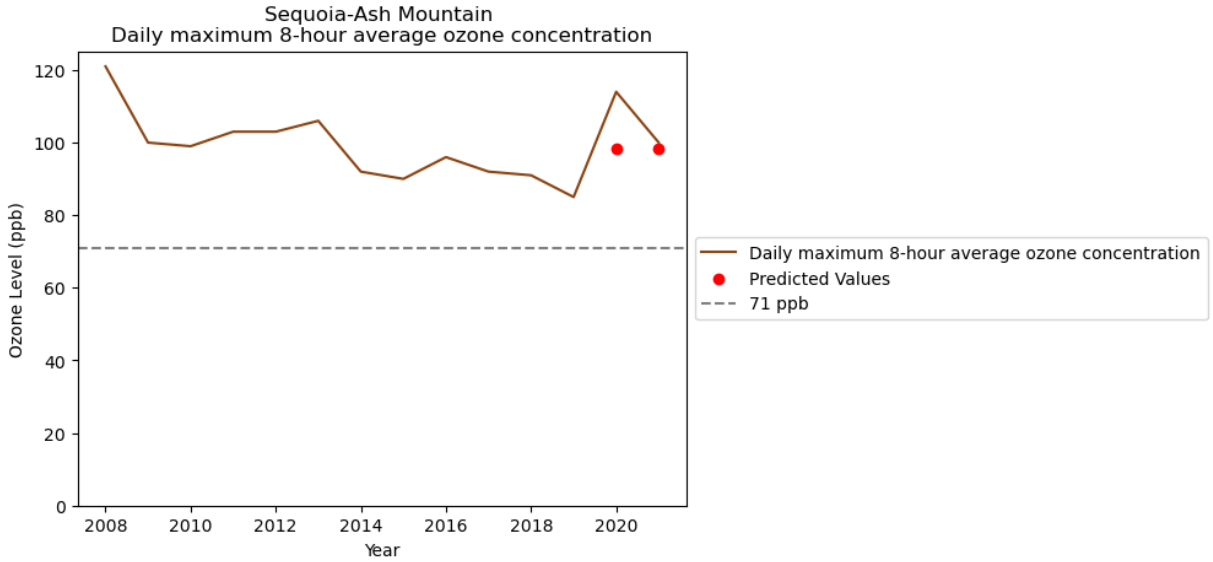



*Figure 6.* Results from model for the daily maximum 8-hour average ozone concentration at the Sequoia and Kings Canyon National Park- Ash Mountain monitoring site.

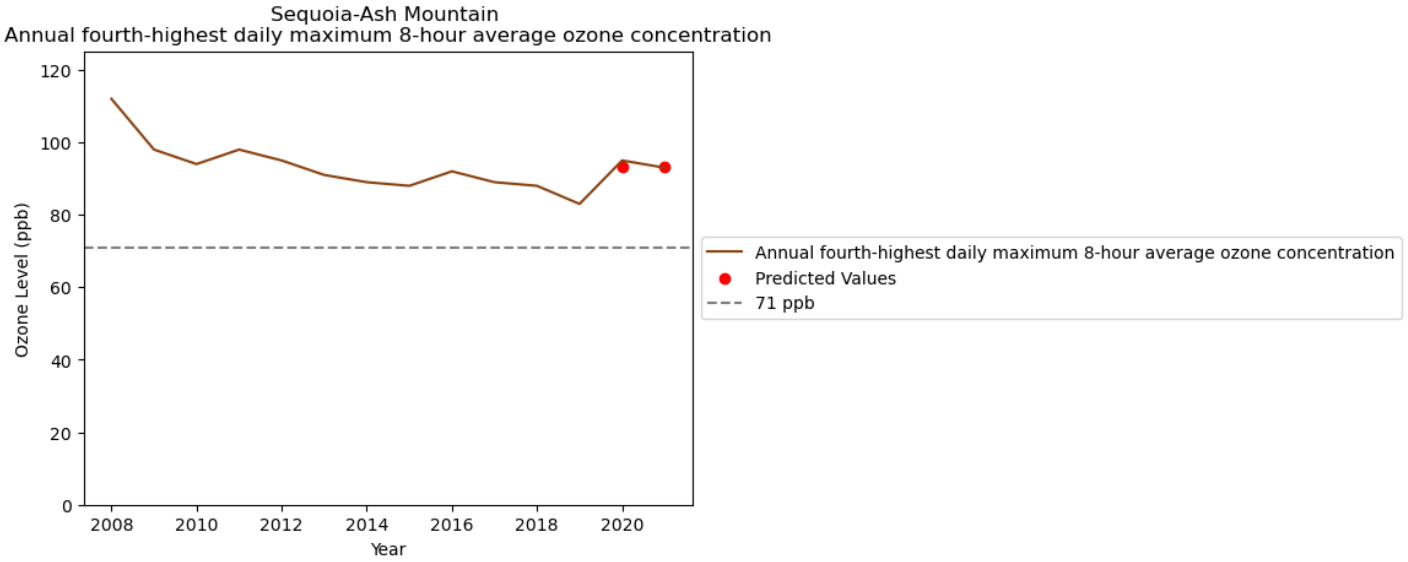



*Figure 7.* Results from model for the annual fourth highest daily maximum 8-hour average ozone concentration at the Sequoia and Kings Canyon National Park- Ash Mountain monitoring site.

Of the three models for the Sequoia and Kings Canyon National Park- Ash Mountain monitoring site, the model for the annual fourth-highest daily maximum 8-hour average ozone concentration performs the best with the lowest root mean square error value of the three. The other two models slightly underperformed and predicted values that were slightly lower than the observed values in 2020 and 2021.

*Yosemite National Park-Turtleback Dome*

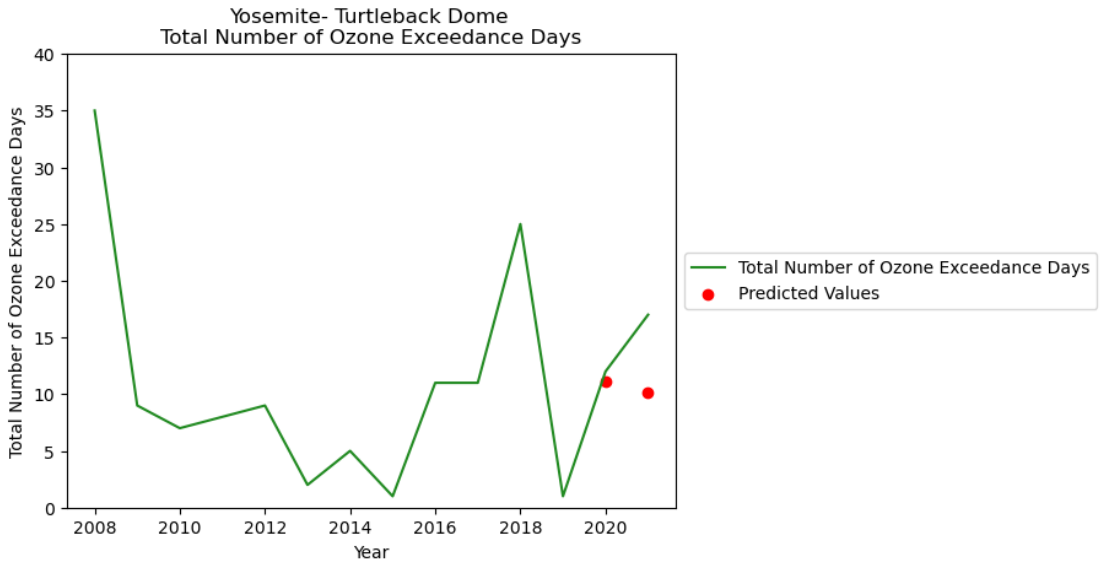



*Figure 8.* Results from model for the total number of ozone exceedance days at the Yosemite National Park-Turtleback Dome monitoring site.

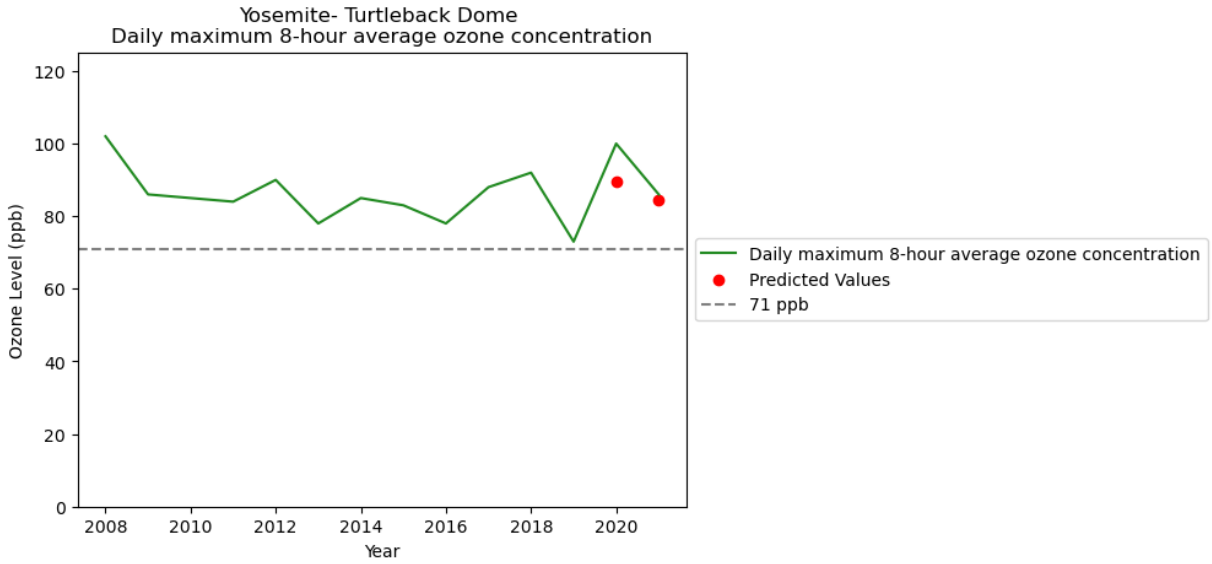



*Figure 9.* Results from model for the daily maximum 8-hour average ozone concentration at the Yosemite National Park- Turtleback Dome monitoring site.

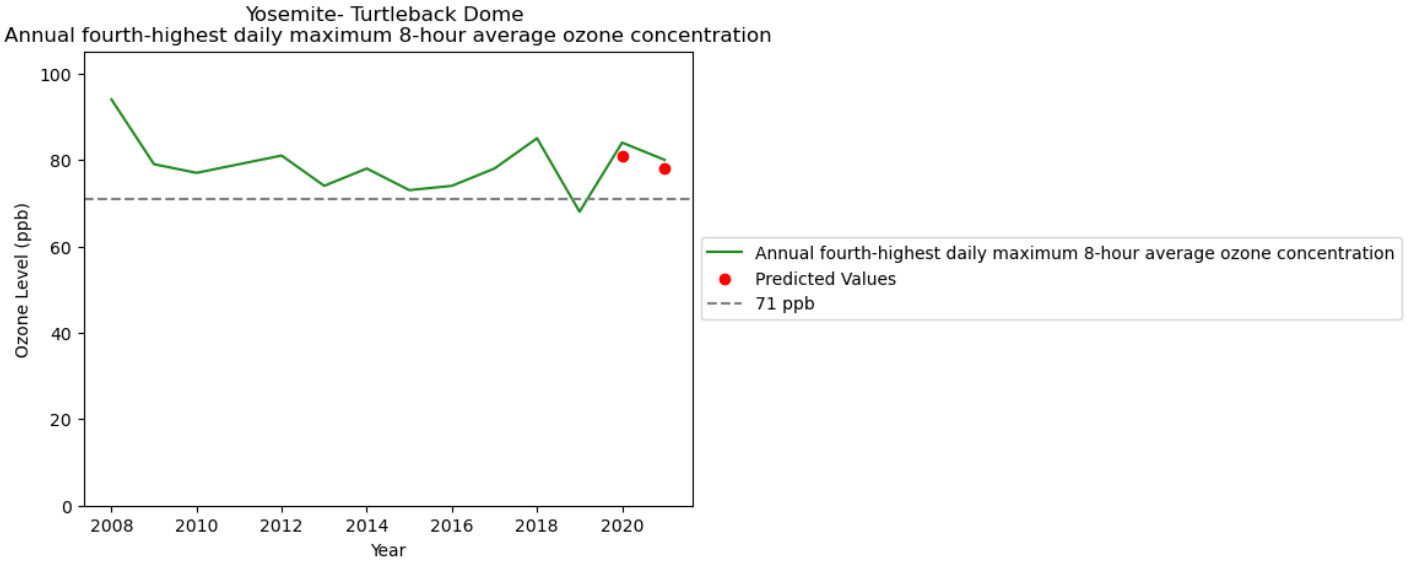



*Figure 10.* Results from model for the annual fourth-highest daily maximum 8-hour average ozone concentration at the Yosemite National Park- Turtleback Dome monitoring site.

Of the four sites with individual models, the Yosemite National Park- Turtleback Dome monitoring site is the only one in which the annual fourth-highest daily maximum 8-hour average ozone concentration dips below the 71-ppb line, which only occurs in 2019. 71 ppb is the determining value in deciding whether a day is considered an ozone exceedance day.

*Sequoia and Kings Canyon National Park- Lower Kaweah*

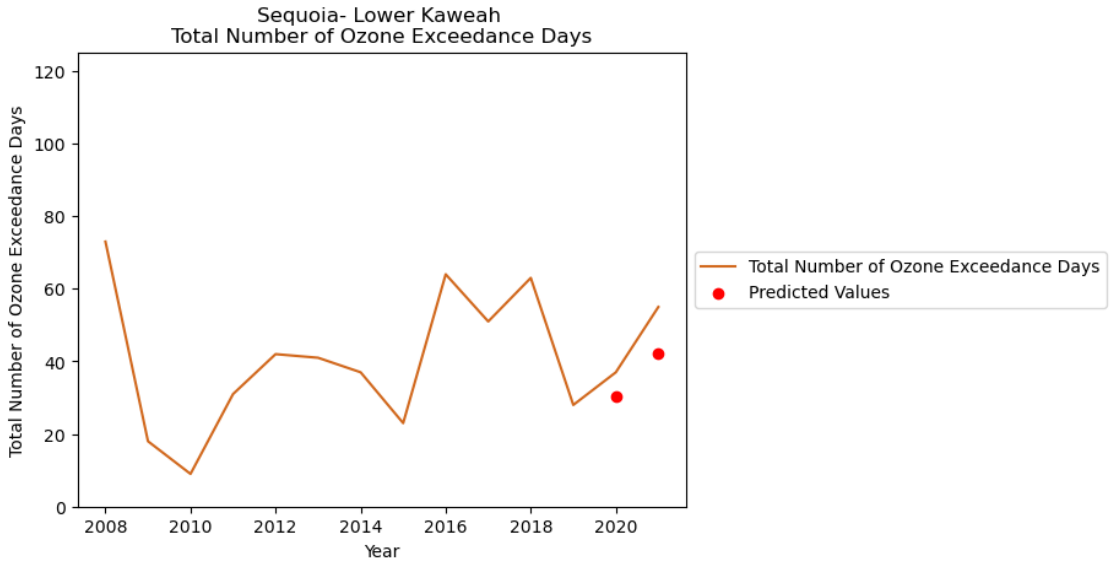



*Figure 11.* Results from model for the total number of ozone exceedance days at the Sequoia and Kings Canyon National Park- Lower Kaweah monitoring site.

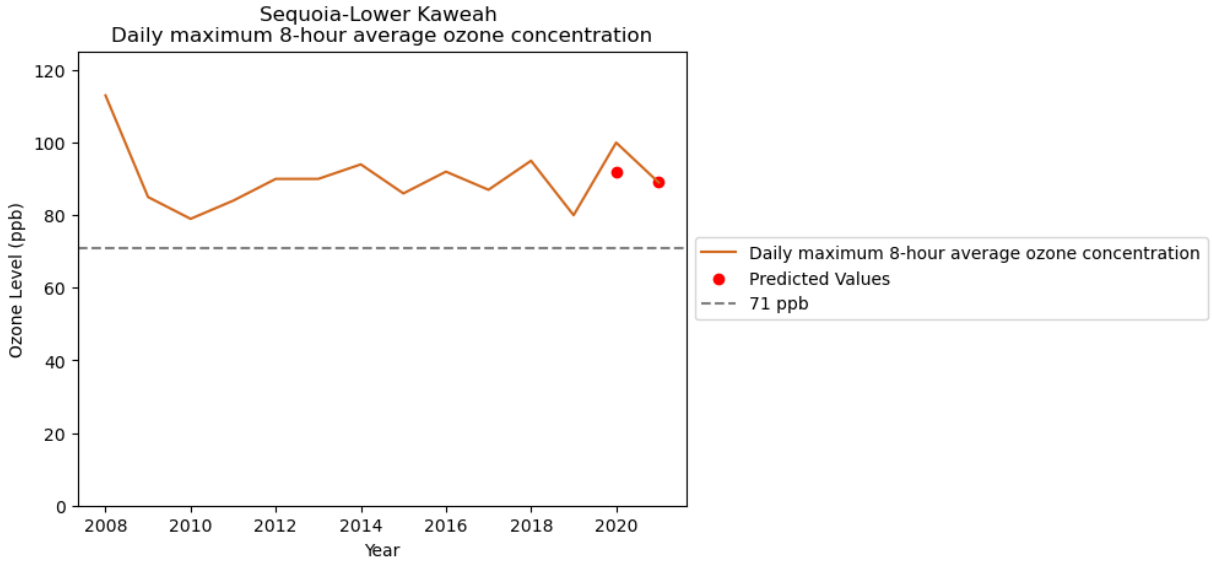



*Figure 12.* Results from model for the daily maximum 8-hour average ozone concentration at the Sequoia and Kings Canyon National Park- Lower Kaweah monitoring site.

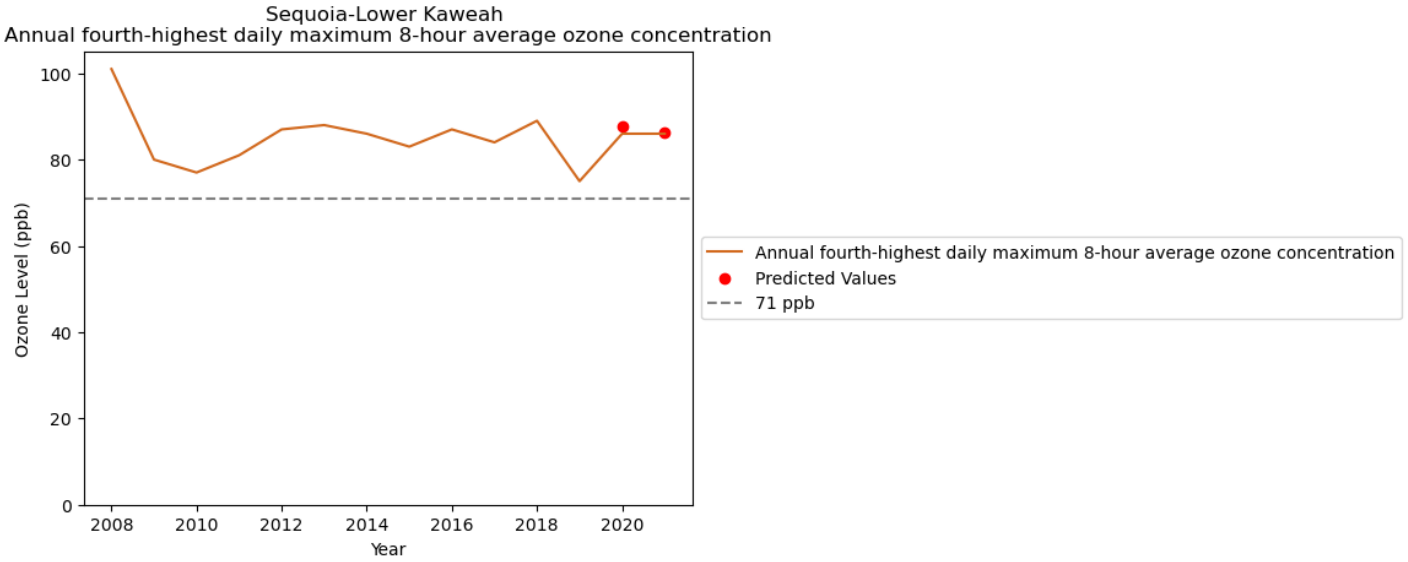



*Figure 13.* Results from model for the annual fourth-highest daily maximum 8-hour average ozone concentration at the Sequoia and Kings Canyon National Park- Lower Kaweah monitoring site.

There is a similar performance between both monitoring sites in Sequoia and Kings Canyon National Park, with the model for the daily maximum 8-hour average ozone concentration at the Lower Kaweah monitoring site slightly outperforming that for the Ash Mountain monitoring site.

*Joshua Tree National Park- Black Rock*

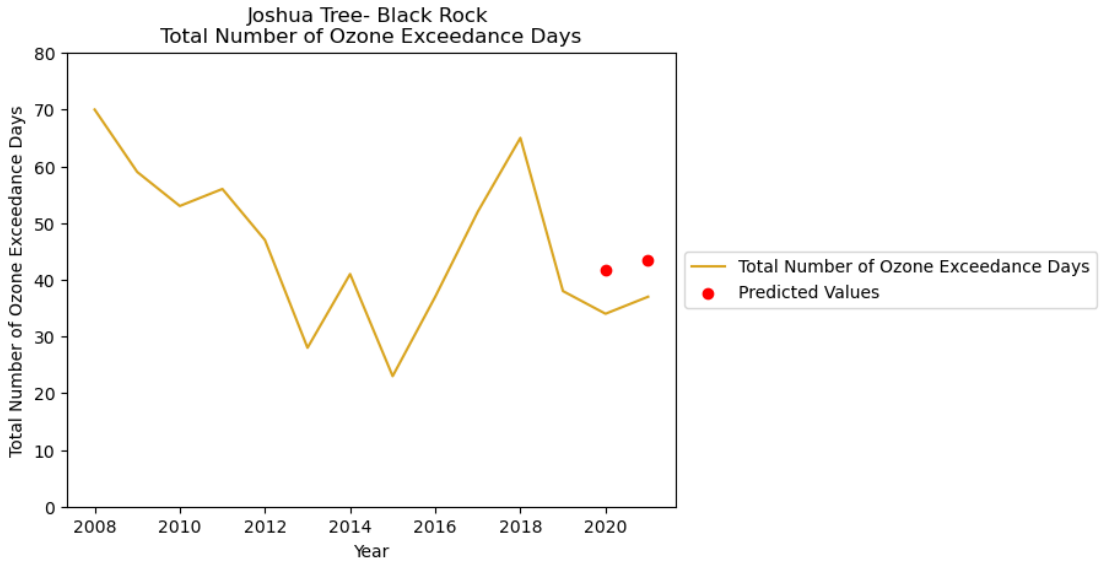



*Figure 14.* Results from model for the total number of ozone exceedance days at the Joshua Tree National Park- Black Rock monitoring site.

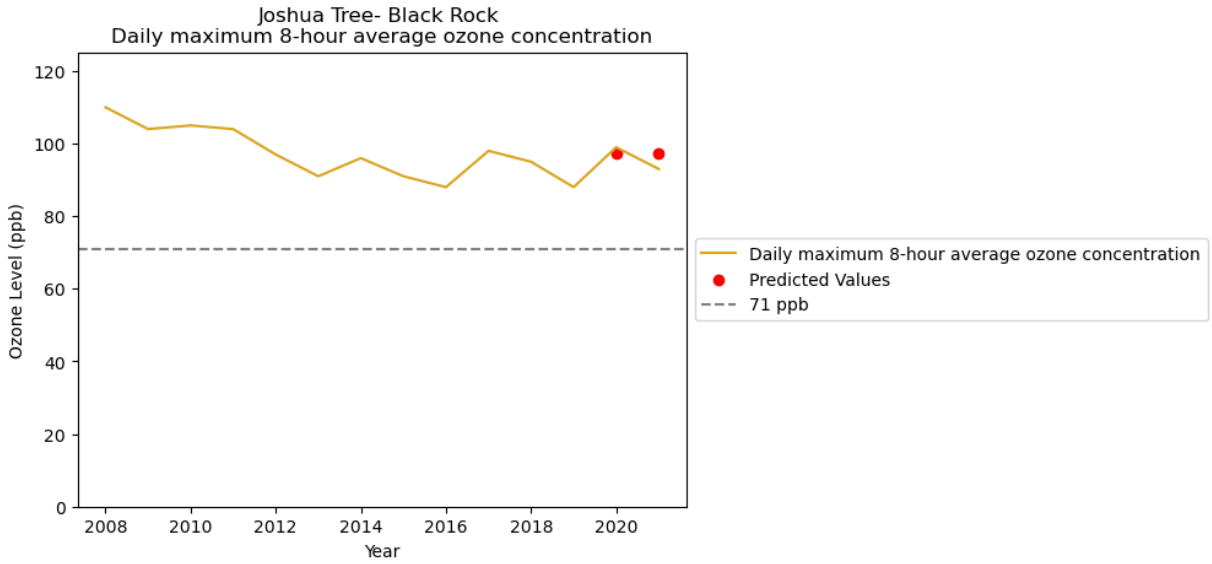



*Figure 15.* Results from model for daily maximum 8-hour average ozone concentration at the Joshua Tree National Park- Black Rock monitoring site.

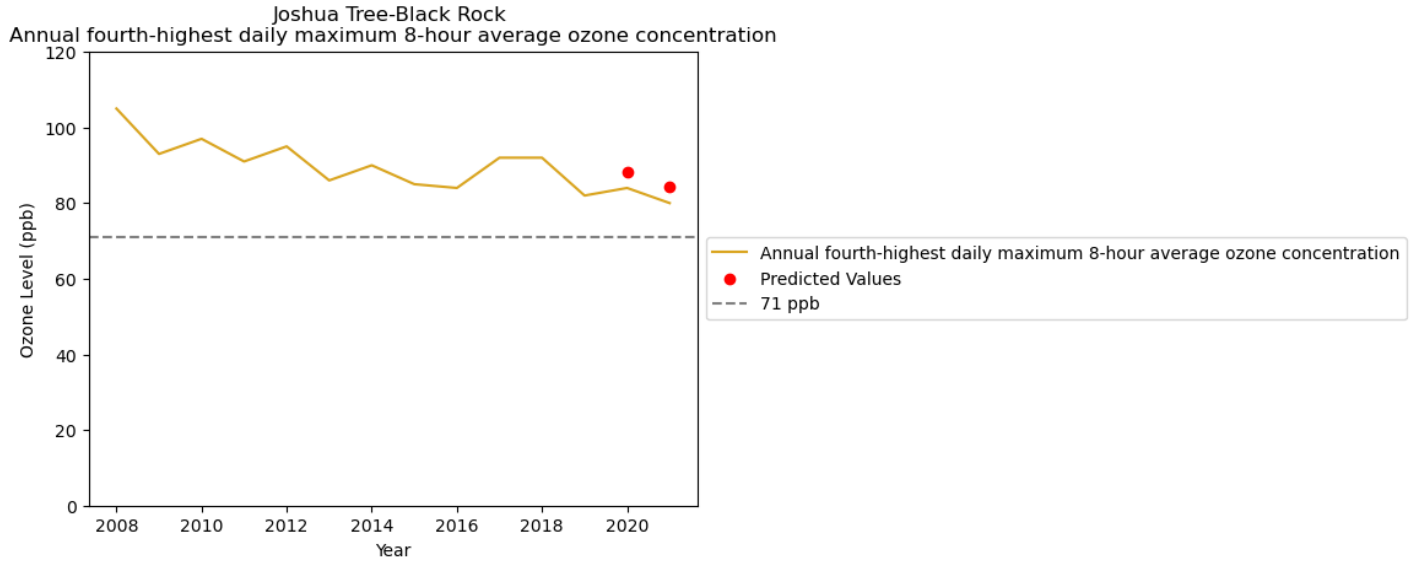



*Figure 16.* Results from model for the annual fourth-highest daily maximum 8-hour average ozone concentration at the Joshua Tree National Park- Black Rock monitoring site.

Of the four sites with individual models built, the Joshua Tree National Park- Black Rock monitoring site models are the only ones with predictions that are slightly higher than the observed values. It is also the only site in which the best performing model was that for the daily maximum 8-hour average ozone concentration rather than the annual fourth-highest daily maximum 8-hour average ozone concentration, as for the other three sites. In this case, overprediction would be preferrable to underprediction because overprediction might yield a more proactive responses that would be a greater benefit to the natural environment than a more limited response.

*Generalized Models*

Since only four of the sites occurred each year, it was important to generalize the data across all the sites for each year. The average of all sites for each year is used for the daily maximum 8-hour average ozone concentration and annual fourth-highest daily maximum 8-hour average ozone concentration values, while the median of all sites for each year was used for the total number of ozone exceedance days. The extreme spread of the data for months did not lend itself to a generalization by a statistic. The median is used for the total number of ozone exceedance days because of the presence of outliers in most years, while the average of the other two values is acceptable due to the low presence of outliers in either variable across all years (Appendix 1).

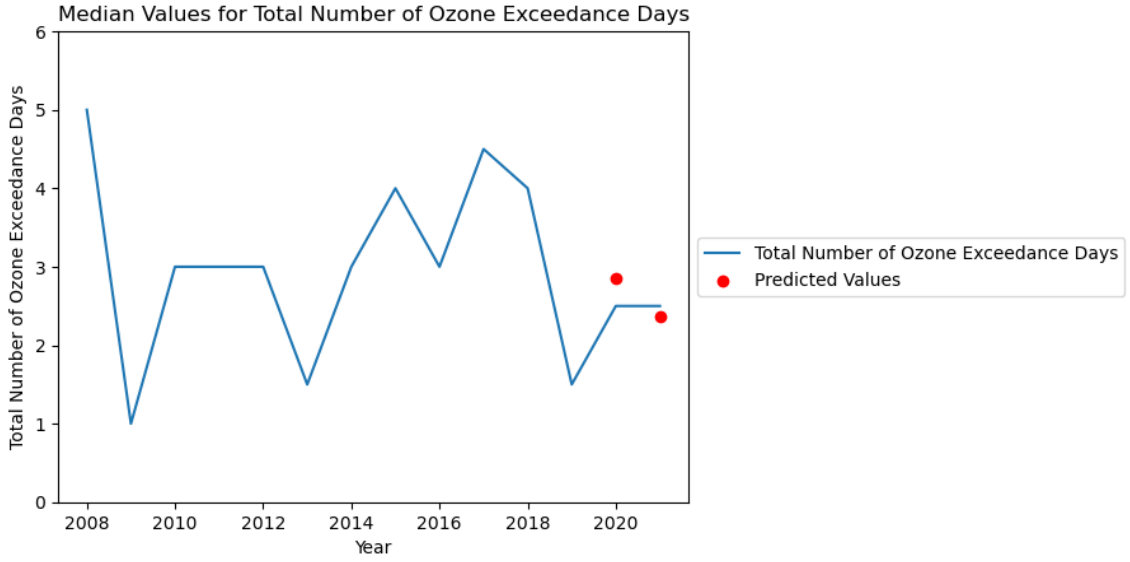



*Figure 17.* Predicted values from model1. RMSE = 0.265. WMAPE = 9.79%.

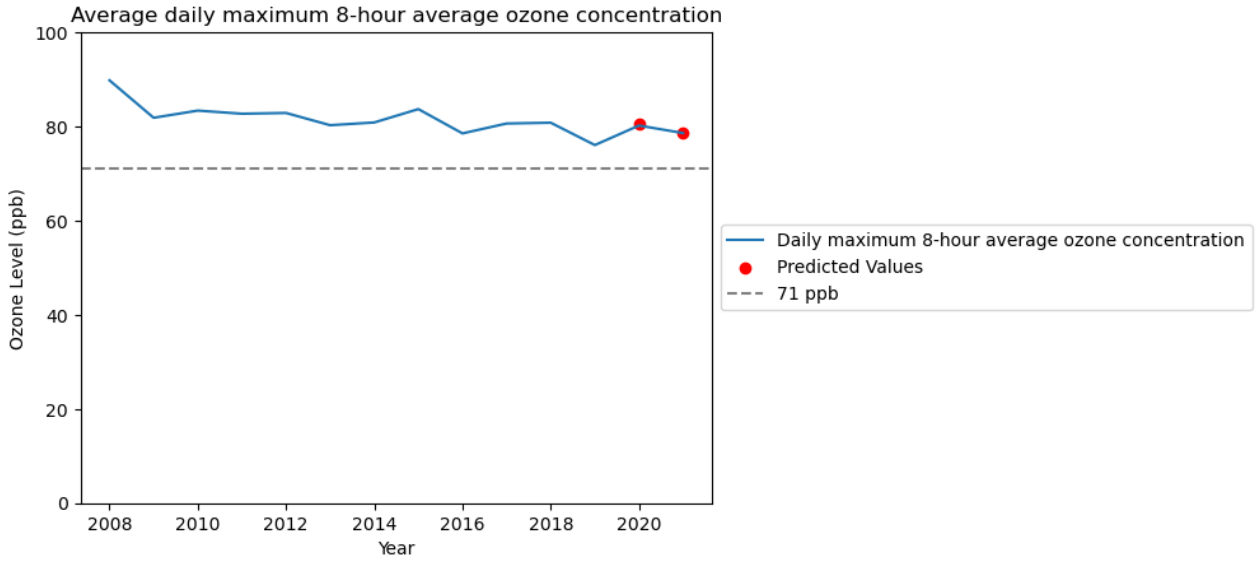



*Figure 18.* Predicted values for model2. RMSE = 0.318. WMAPE = 0.30%.

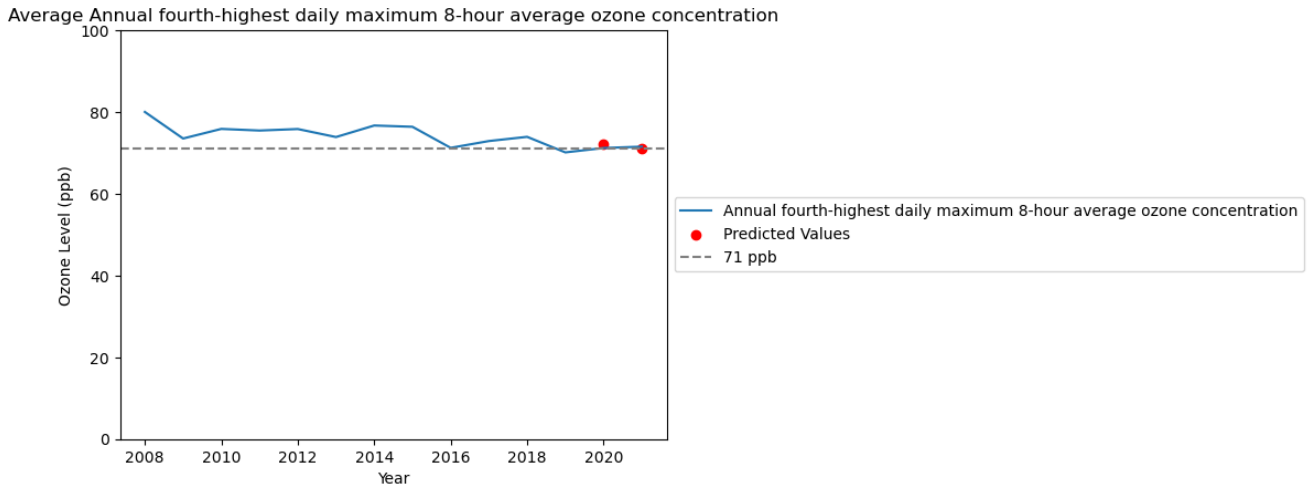



*Figure 19.* Predicted values for model3. RMSE = 0.800. WMAPE = 1.05%.

Of the three generalized models, the best performing model is for the average daily maximum 8-hour average ozone concentration.

## Conclusion

When we look at the data, it quickly becomes clear that the number of ozone exceedance days is not increasing at all, and that there is no recognizable pattern or trend within this data alone. The relationship between the monitoring sites is mainly regional. All 8 sites that regularly appear as part of the top 5 total number of ozone exceedance days are in the southern portion of the US, with 6 of those sites occurring in California alone. It is likely that these sites also are all heavily visited sites as well, which could be another factor in the total number of ozone exceedance days experienced by these sites.

The generalization of the data is necessary to obtain models that span most of the sites; however, the applicability of these models is somewhat questionable. The models for the average daily maximum 8-hour average ozone concentration and average annual fourth-highest daily maximum 8-hour average ozone concentration are built on averages of averages, which does not lend itself to many useful applications as it generally represents an overgeneralization. These models, however, do paint an illustrative story about the general air quality of the country as the national park sites are spread across the United States. When we compare the plot of the median of the total number of ozone exceedance days and the plots for the total number of ozone exceedance days for each of the four sites studied individually, all five plots have a peak between 2016 and 2018. These generalized models might not serve useful for decision-making at an individual park level, but they might still prove illustrative of the overall ozone concentration within the United States.

## Assumptions

Before spending time with the data, the assumption was that the data would exhibit strong patterns and there would be increases in between years for the values. It was also assumed that more sites would occur every year, and only a few would have missing years, while the vast majority of the monitoring sites only appear in the data a handful of times.

## Limitations/Challenges

The main challenge was creating accurate models on limited data- there were only 14 years' worth of data, which left 12 years of data to train and two for validation of predictions.

**Future Uses/Additional Applications**

In future, the most useful model is for that of the daily maximum 8-hour average ozone concentration. With additional work and additional data, models for each site of this value would be useful to give visitors and rangers information to guide those within sensitive groups about the possibility of high ozone concentration levels during their trip. This information could also be used to help park rangers and the NPS determine if limits must be placed on attendance during the summer, which is when peak ozone exceedance days occur, for the health of the visitors and to mitigate the presence of any additional pollution from additional vehicles to protect the wildlife from excessive ozone levels.

**Ethical Assessment**

The negative effects ozone has on both public health and the environment makes it imperative that we understand any underlying trends in data concerning ozone concentration levels to protect the natural resources and wildlife present in the national parks. According to the NPS's mission to preserve the natural resources within the National Park System, it is their responsibility to monitor changing ozone concentration levels within the park. The analysis from this project serves the environment itself and not the NPS as an organization.

References

*Average temperature in Sequoia National Park by Year*. Extreme Weather Watch. (n.d.). https://www.extremeweatherwatch.com/cities/sequoia-national-park/average-temperature-by-year

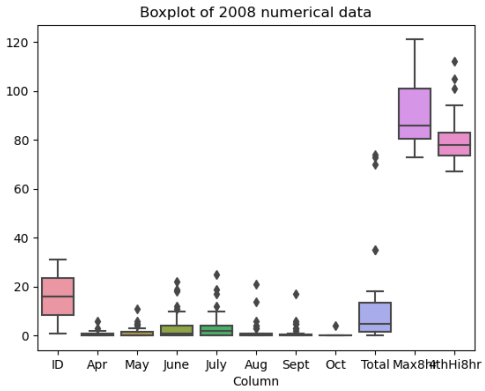
California Air Resources Board (n.d.). *Health Effects of Ozone*. Ozone & Health. https://ww2.arb.ca.gov/resources/ozone-and-health#:~:text=Ozone%20is%20formed%20in%20the,paints%2C%20and%20many%20other%20sources.

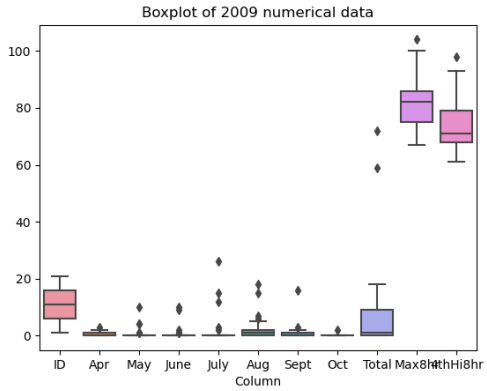
inderjalli. (2019, February 25). *wmape_group_example.py*. wmape_grouping. https://github.com/inderjalli/wmape_grouping/blob/master/wmape_group_example.py

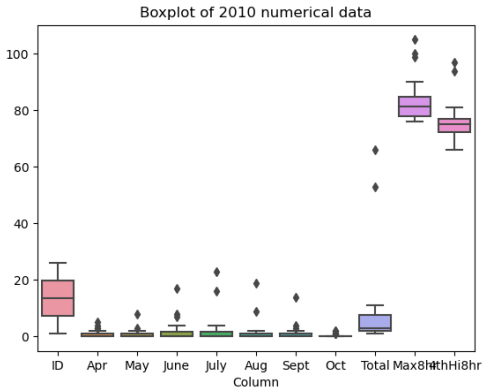
National Park Service (n.d.). *Ozone Standard Exceedances in National Parks*. NPS DataStore. https://irma.nps.gov/DataStore/Collection/Profile/4319

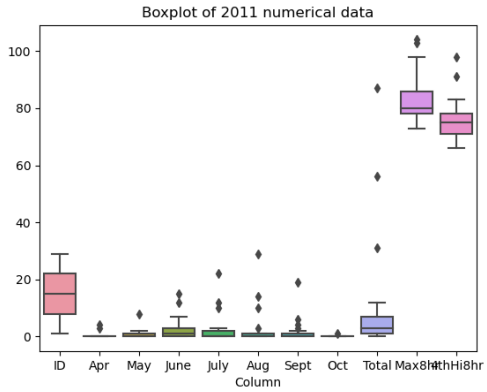
# Appendix 1

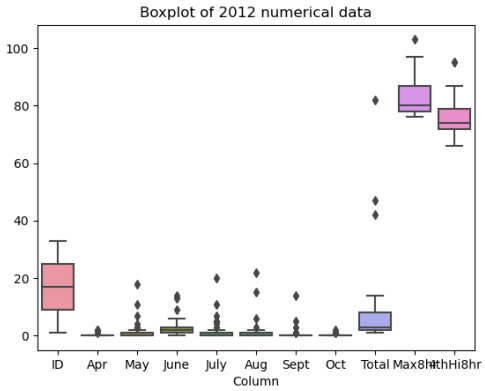
*Boxplots for each year in data*

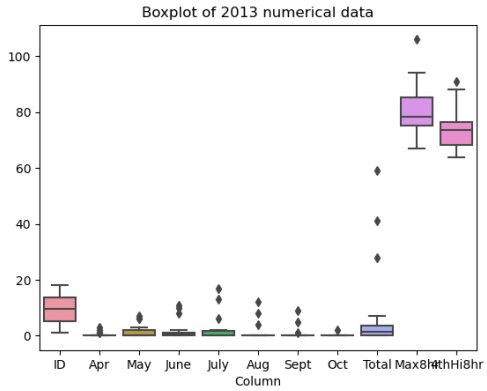


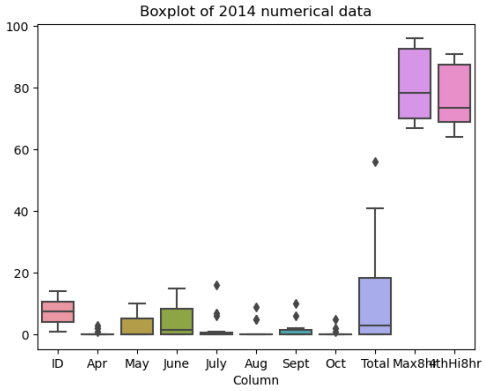



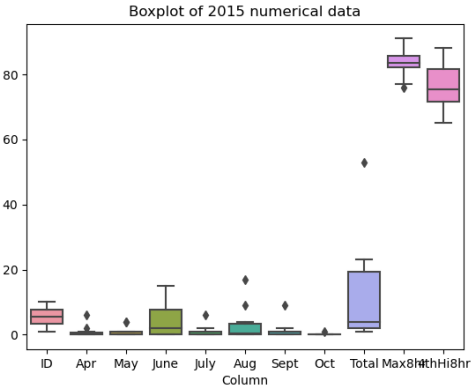



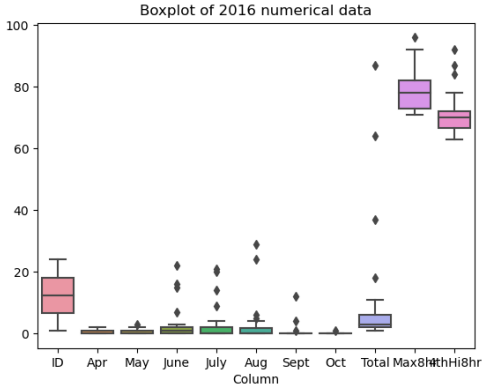



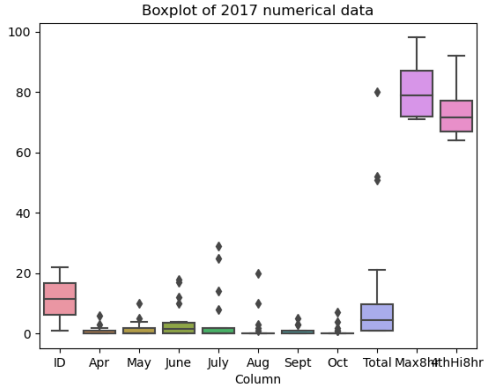



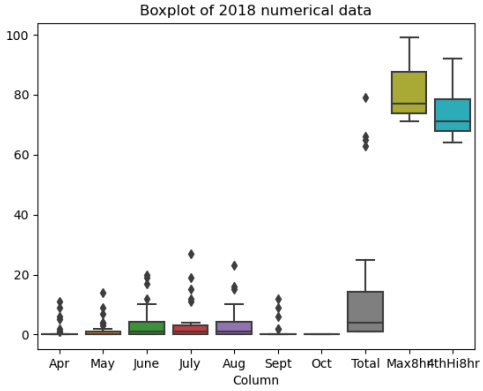



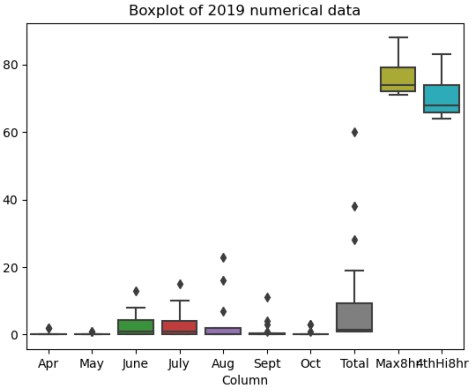

Boxplot of 2014 numerical data
ID
Apr
May
June
July
Aug
Sept
Oct
Total
Max8hr
4thHi8hr
Column

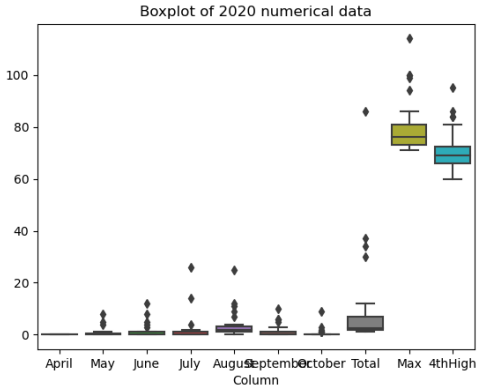

Boxplot of 2015 numerical data
ID
Apr
May
June
July
Aug
Sept
Oct
Total
Max8hr
4thHi8hr
Column

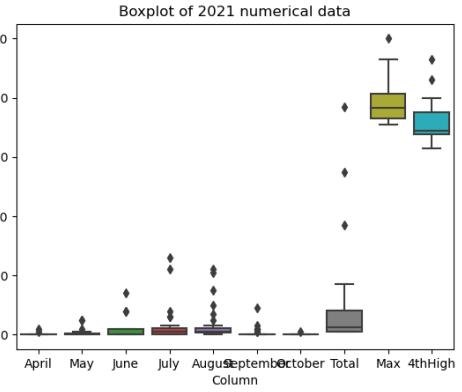

Boxplot of 2016 numerical data
ID
Apr
May
June
July
Aug
Sept
Oct
Total
Max8hr
4thHi8hr
Column


Boxplot of 2017 numerical data
ID
Apr
May
June
July
Aug
Sept
Oct
Total
Max8hr
4thHi8hr
Column


Boxplot of 2018 numerical data
Apr
May
June
July
Aug
Sept
Oct
Total
Max8hr
4thHi8hr
Column


Boxplot of 2019 numerical data
Apr
May
June
July
Aug
Sept
Oct
Total
Max8hr
4thHi8hr
Column


Boxplot of 2020 numerical data
April
May
June
July
August
September
October
Total
Max
4thHigh
Column


Boxplot of 2021 numerical data
April
May
June
July
August
September
October
Total
Max
4thHigh
Column

Appendix 2

*RMSE and WMAPE values for the Sequoia and Kings Canyon National Park- Ash Mountain monitoring site.*

| | RMSE | WMAPE |
|---|---|---|
| Total Number of Ozone Exceedance Days | 12.196 | 0.0% |
| Daily maximum 8-hour average ozone concentration | 11.271 | 0.0% |
| Annual fourth-highest daily maximum 8-hour average ozone concentration | 1.357 | 0.0% |

Appendix 3

*RMSE and WMAPE values for Yosemite National Park- Turtleback Dome monitoring site.*

| | RMSE | WMAPE |
|---|---|---|
| Total Number of Ozone Exceedance Days | 4.890 | 0.0% |
| Daily maximum 8-hour average ozone concentration | 7.419 | 0.0% |
| Annual fourth-highest daily maximum 8-hour average ozone concentration | 2.690 | 0.0% |

Appendix 4

*RMSE and WMAPE values for Sequoia and Kings Canyon National Park- Lower Kaweah monitoring site.*

| | RMSE | WMAPE |
|---|---|---|
| Total Number of Ozone Exceedance Days | 10.215 | 0.0% |
| Daily maximum 8-hour average ozone concentration | 5.803 | 0.0% |
| Annual fourth-highest daily maximum 8-hour average ozone concentration | 1.205 | 0.0% |

Appendix 5

*RMSE and WMAPE values for Joshua Tree National Park- Black Rock monitoring site.*

| | RMSE | WMAPE |
|---|---|---|
| Total Number of Ozone Exceedance Days | 7.101 | 10.74% |
| Daily maximum 8-hour average ozone concentration | 3.250 | 0.91% |
| Annual fourth-highest daily maximum 8-hour average ozone concentration | 4.278 | 2.56% |