Spencer Rubin
Capstone2: Final Report

Ground Level Ozone Air Quality Analysis

Problem Statement:

Air quality throughout the U.S. fluctuates based on various environmental, climatic, and anthropogenic features, and varies at the both spatial and temporal scales. Air quality measurements include particulate matter, nitrous oxides, sulfurous oxides, and ozone, all which have an impact on human health depending on threshold, quantity, and exposure limits, and that are monitored by either the state or federal EPA. Understanding what features are influencing air quality will allow environmental regulators, city planners, and government agencies opportunities to improve poor air quality by understanding the influence of the changes at a temporal scale in the U.S. Specifically, ground level Ozone (O3), a direct byproduct of anthropogenic forces combined with natural meteorologic features, plays a major role in human health relating to individual respiratory systems.

How does ground-level air quality fluctuate between various urban areas throughout the U.S.?

The EPA 8-hour Ozone threshold is set at 0.07ppm (2019) for attainment, although concentrations of 0.06ppm are shown to have health impacts on the human respiratory system, including asthma and other respiratory disease. By classifying health concerns (above or below 0.06ppm) and the EPA Ozone threshold (0.07ppm), we can further understand county-level concerns for air quality health.

In this project, we utilized each county's total population estimation in 2019 as the primary feature for modeling Ozone health concerns.

Data Wrangling:

The following counties were included in the dataset:

- Los Angeles County, CA
- Riverside County, CA
- Orange County, CA
- San Diego County, CA
- Maricopa County, AZ
- Dallas County, TX
- Harris County, TX
- Queens County, NY
- Miami-Dade County, FL
- Cook County, IL

  ** Kings County, NY, was not included since there are not full datasets available in this county through the EPA*

The total population estimation data for 2019 was acquired for each county from https://www.census.gov/. The range of total population ranged from approximately 2 million to 10 million, suggesting significant variance in county total population.

The 8-hour Ozone measurement data was acquired from https://www.epa.gov/. With multiple measurement stations throughout each county, data was grouped and summarized for each day for each county to acquire the "Max 8-Hour Ozone (ppm)". A dataframe was created from the Maximum 8-hour Ozone concentration data for the entire year (daily) of 2019, by county, that had a final shape of 3,606 rows (observations by county by date) and 2 rows (county category, max 8-hour Ozone (ppm)). It was important to acquire maximum daily Ozone since the maximum exposure is what will result in the onset of respiratory disease and illness.

Exploratory Data Analysis:

Differences in county-level Ozone concentrations measures and grouped by day were analyzed to visualize and understand the distribution of Ozone between counties and the range of observed Ozone concentrations across all counties. Maximum Ozone concentration measurements ranged between as low as < 0.01ppm and as high as > 0.10ppm, which indicate a large, dynamic range of possible Ozone concentrations relative to health concerns throughout the ten chosen counties. Although, county-level distribution and summary statistics suggest that each county is quite different with regards to number of days in 2019 above both the health-concern threshold of 0.06ppm and the EPA attainment level of 0.07ppm.

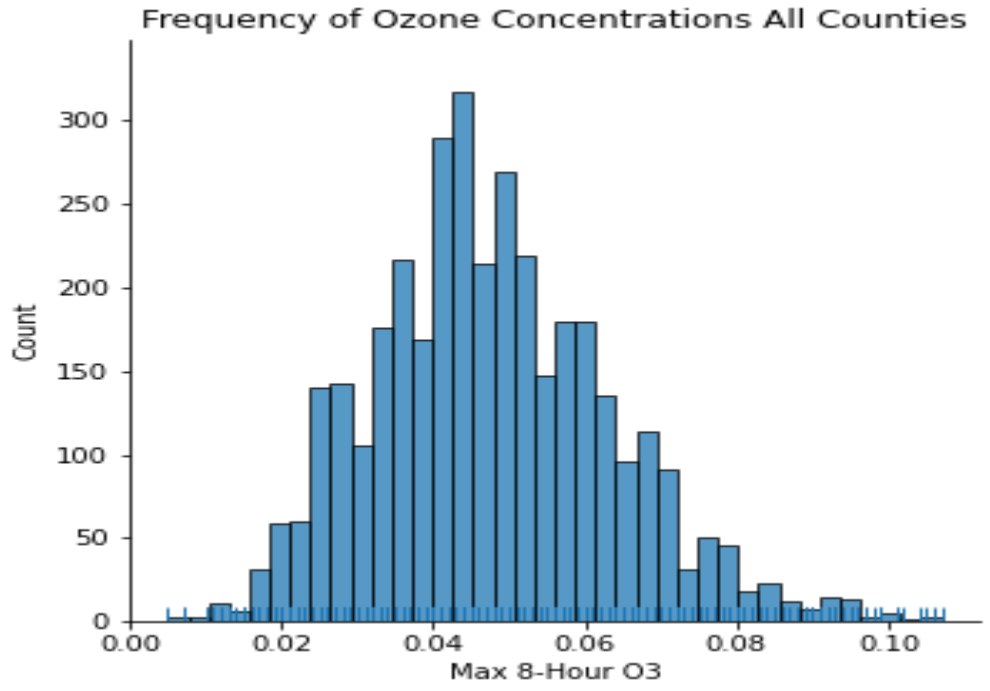

Figure 1: Distribution of the frequency Maximum Ozone (ppm) across all ten counties each day in the year 2019. The distribution exceeds the 0.06ppm level for a significant portion of the observations.

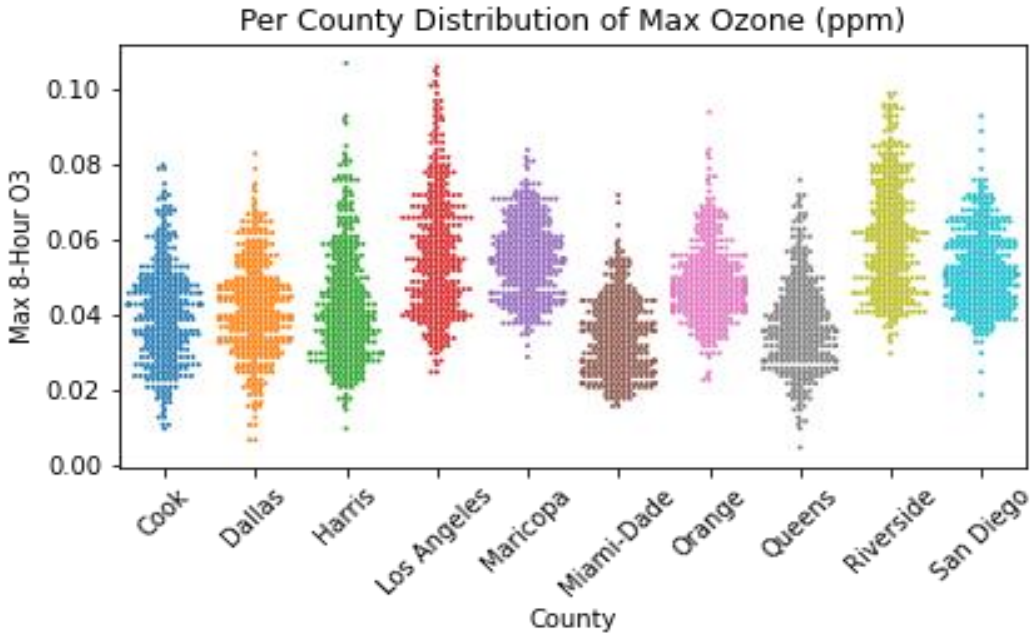

Figure 2: County level distribution of Maximum Ozone (ppm) for the year 2019. Los Angeles and Riverside counties have a notably higher distribution of Ozone concentrations above 0.06ppm and 0.07ppm.

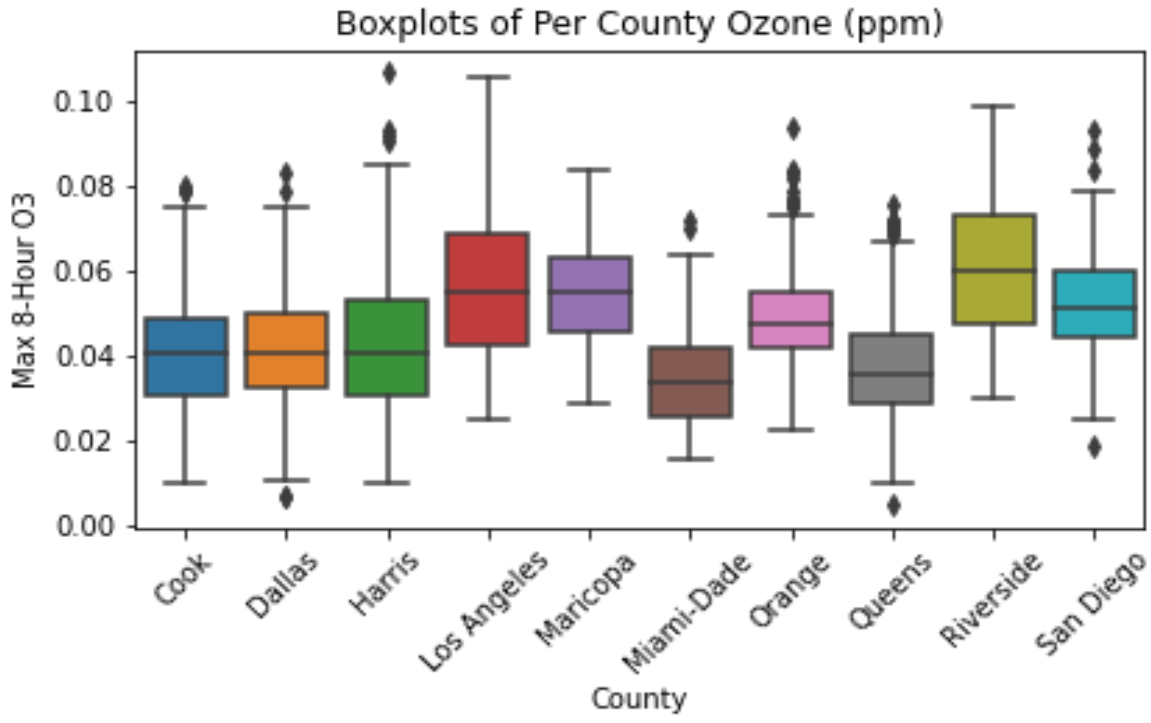

Figure 3: Boxplots showing the distribution of summary statistics of per county Ozone concentrations in 2019. The mean across all counties appears to be less than 0.06ppm, although the 25th percentile for various counties is near or above 0.06ppm.

Preprocessing and Feature Selection:

In order to classify counties as being of concern for air quality health related to Ozone, I added total county population (by estimation from the U.S. Census) for 2019, and then categorized each county as either being above or below the median total population level of 3.26 million. I used the median since Los Angeles County, CA, was an outlier in terms of total population in comparison to the other nine counties. I then created two categorical columns – "ozone_concern" and "ozone_threshold" – to categorize days with observed maximum Ozone concentrations below or above 0.06ppm and 0.07ppm respectively. County labels were categorized using pandas get_dummies().

The dataset was balanced further by utilizing the SMOTE package as part of imblearn, to create a more balanced sampling dataset for subsequent modeling efforts. The Max 8-hour Ozone (ppm) was scaled using sklearn's StandardScaler().

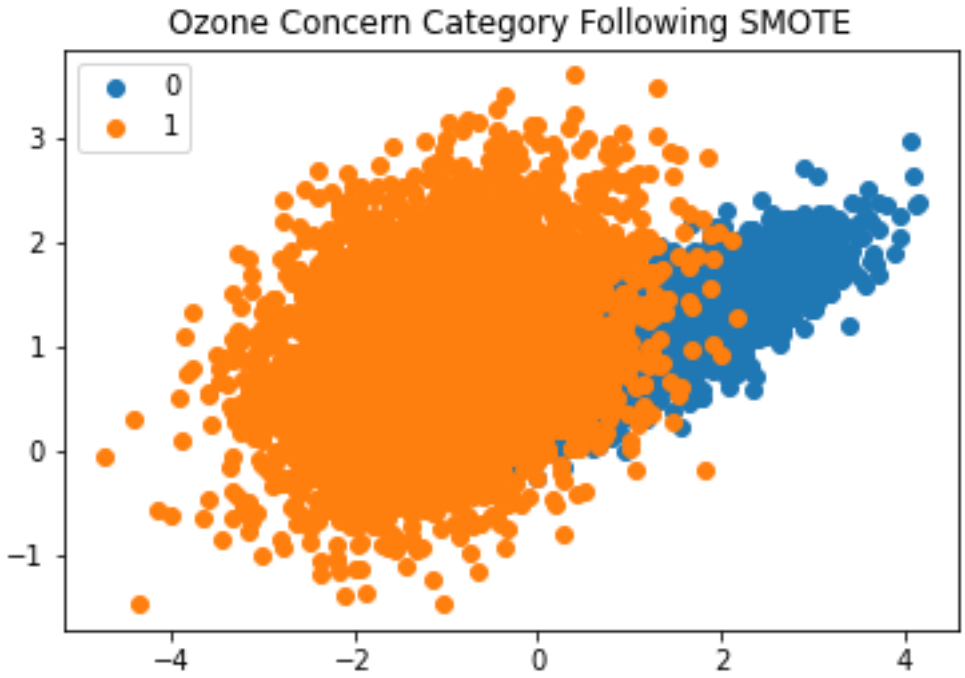

Figure 4: Distribution across all counties of 0: No Ozone Concern and 1: Ozone Concern relating to the 0.06ppm concentration, following dataset balancing utilizing SMOTE.

Model Selection:

I tested three separate supervised learning classification models. This included Logistic Regression, Decision Tree, and Random Forest. The model metric I utilized for model selection was precision, to understand how well each classification model predicted the county level Ozone health concern related to each county's total population.

Performance of the Logistic Regression classification model resulted in strong performance. The accuracy was 0.89, suggesting that 89% of the classification was predicted as true. Figure 5 shows the ROC curve for the performance of the Logistic Regression model. Using sklearn's Gridsearch cross validation marginally improved the Logistic Regression performance, by selecting C: 0.1, with a rounded accuracy score still at 0.89.

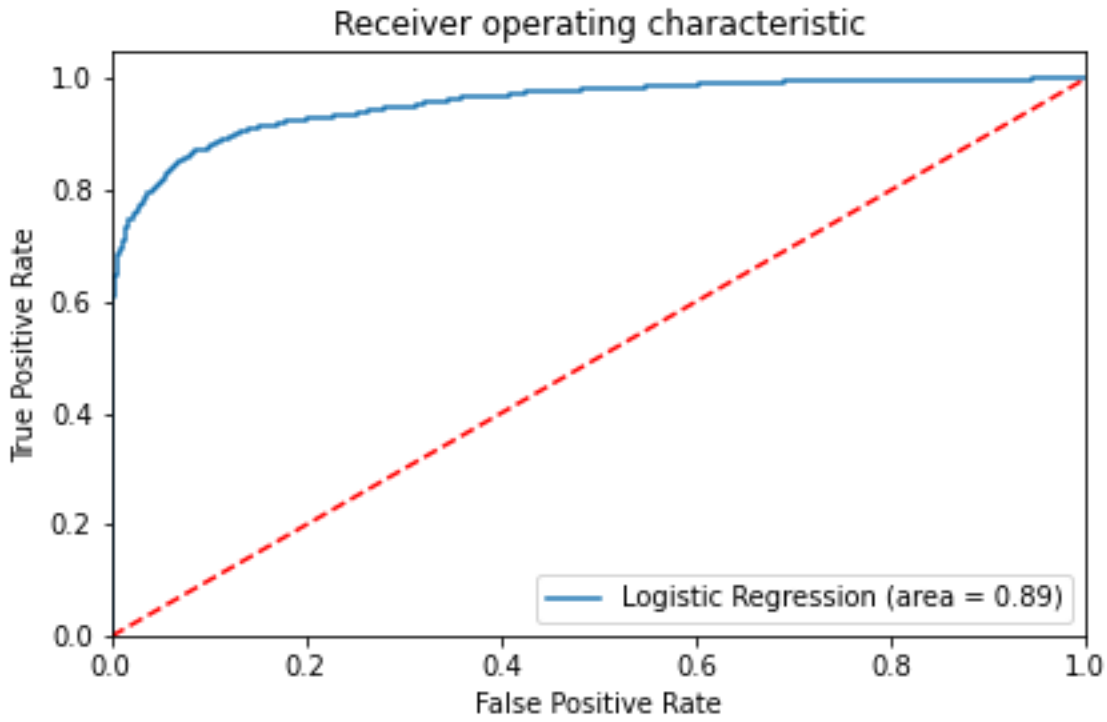

Figure 5: Receiver Operator Characteristic curve for the Logistic Regression classifier model.

Performance for the Decision Tree Classifier did not suggest any improvement, with the accuracy score at the same as the Logistic Regression Classifier of 0.89.

Utilizing the Random Forest Classifier, I was able to obtain a slight improvement over the Logistic Regression Classifier with an accuracy score of 0.90. Although, the marginal improvement does not suggest much of a better performance.

Conclusions:

All three tested models provide a high level of accuracy when classifying county level Ozone health concern relative to population. Counties with higher population are more likely to have a greater amount of days exceeding both 0.06ppm and 0.07ppm, resulting in a greater public concern regarding air quality health. The Random Forest Classifier performed with the highest accuracy and may be considered the best performing model.

Additional features should be considered for tuning and further adapting this classification system. Ground level ozone is not just a direct result of total population (e.g. car exhaust, power plants), but a result of the combination of temperature, sunshine, population density, and industrial output. Although I utilized ten of most populous counties in the U.S., these are not necessarily the most population dense. I also believe incorporating geographic, geologic, and meteorologic features will fine tune county to Ozone concentration. Lastly, since Ozone concentration increases with temperature and sunlight availability, and the distribution of counties across the U.S. varies greatly in this distribution (e.g. Los Angeles County, CA, compared to Cook County, IL), there may be potential for both a timeseries analysis and geographic distribution analysis in future studies.