Prediction of Unhealthy Air Pollutant Levels In Order To Better Manage Patient Care

Molly McNamara

Introduction

High air pollution levels have been consistently documented as a major environmental risk to health. The World Health Organization has stated that the majority of the world's population is living in areas that do not meet air quality guidelines and estimated that outdoor air pollution caused 3 million premature deaths in 2012. While programs do exist in many countries to combat air pollution with the aim of reducing levels of harmful pollutants, weather patterns and geographical features can also influence the concentration of air pollutants in many areas. Among other things, exposure to air pollutants can cause the development of respiratory diseases and exacerbate existing cases. Not only is air pollution a documented health risk, it is associated with an increased utilization of health care services (Environmental Health 2011, Epidemiology 2004, American Journal of Epidemiology 2001, Epidemiology 2005).

Given the health risks involved in elevated levels of air pollutants, it would be useful for health care providers and hospitals to predict the uptick of services during periods of poor air quality. Prediction of periods of increased patient illness could help with planning for hospital staffing needs as well as preventative care and messaging to reduce hospital visits. Staffing in particular is a challenge hospitals face, due to an anticipated shortfall in registered nurses. By 2025, researchers expect a shortage "more than twice as large as any nurse shortage experienced since the introduction of Medicare and Medicaid in the mid-1960s" (Health Affairs, 2009). Given the challenges facing the health care system in the United States, there is a tremendous financial incentive to better manage high risk patients.

The hypothetical client in this case is Kaiser Permanente, a large managed care organization that runs a number of hospitals and clinics with an emphasis on preventive care and population health management. Prediction of air quality incidents that may affect population health and cause a spike in emergency room visits and hospitalizations may help Kaiser staff their clinics accordingly and in advance rather than face nursing shortages that could impact quality of care, wait times or their quality metrics.

The Data

Multiple datasets were combined for this analysis. The primary dataset consists of daily levels of 4 primary air pollutants (Nitrogen Dioxide, Sulphur Dioxide, Carbon

Monoxide and Ozone) and their Air Quality Index values from major cities across the United States between 2000 and 2016. The data is sourced from the United States Environmental Protection Agency and downloaded from a compiled set at Kaggle as a CSV file.

Air Quality Index values equate to health concerns as follows, per the Environmental Protection Agency:

- 0 to 50 Good
- 51 to 100 Moderate
- 101 to 150 Unhealthy for Sensitive Groups
- 151 to 200 Unhealthy
- 201 to 300 Very Unhealthy
- 301 to 500 Hazardous

Additional data was compiled from climate data collected by the National Oceanic and Atmospheric Administration (NOAA) for 7 of the largest US cities in the pollution dataset from January 1, 2000 through December 31, 2015. The variables include elevation of the city, daily precipitation, air pressure, wind speed and the daily high and low temperatures. This information was ordered from the NOAA website, downloaded as multiple CSV files, in ~ 5-year blocks by city, and then appended to one another.

Data Wrangling

Pollution

The pollution dataset was imported into iPython notebook and determined to consist of 1,746,661 observations with 28 columns. Further evaluation revealed that there were multiple duplicate/triplicate observations for many days at the same collection site. When this duplicate data was removed (by grouping by date and site and then retaining the mean value for each variable), the dataset consisted of one row for each date for each site, for a total of 412,856 rows.

Columns that would not be used in analysis were dropped; specifically the columns for state and city code, address of testing site, and units of measure for pollutant levels were removed. The column for local date was converted to a datetime object. The dataset was evaluated for null/empty fields and these were replaced using fillna.

Boxplots were generated to evaluate the spread of each pollutant's Air Quality Index by state. Finally, the columns of the dataset were renamed to make them easier to work with going forward. A clean CSV file was saved.

Weather

The weather datasets were imported into iPython notebook and appended to one another. The final dataset was determined to consist of 1,247,655 observations with 90 columns. Hourly and monthly data columns were dropped so that the weather dataset would contain daily data to match the pollution dataset. Null values were handled using the ffill function.

The dataset was filtered to retain only the last daily timepjoint at 11:59pm, which captured the daily averages based on the data from the entire day. The column for local date was converted to a datetime object.

The columns of the dataset were renamed to make them easier to work with going forward and to match the nomenclature of date and city in the pollution dataframe. Finally, the pollution and weather dataframes were merged on a left join by City and Date. A clean CSV file was saved.

Exploratory Data Analysis and Statistical Findings

The pollution data comes from 141 cities in 46 US states (plus 3 cities in Mexico).

As seen in Figure 1, the four pollutants' Air Quality Index measured range from Good to Very Unhealthy. On average, based on the mean, they tend to be in the Good rating and so being able to predict those excursions into unsafe levels is the goal of this analysis.

Pollutant levels were further assessed visually for the 10 largest cities in the United States (by census data) to compare the distributions (Figures 2-5) and confirmed that each city has a unique profile. Their fluctuations were also evaluated over time and demonstrated seasonal/annual variability (Figures 6-9) for the largest 5 US cities (this number was reduced from 10 to improve visual clarity of the trends).

The pollutant Air Quality Index values did not appear to be normally distributed (though mean daily ozone levels are). This may be a function of many values close to 0 or some sort of natural limit.

The pollutant levels were assessed relative to the historical weather data to identify any correlations. There were relatively strong correlations present between pollutant levels and weather features such as average and maximum temperature and average sea level pressure. Figure 10 demonstrates an example of the correlative relationship between maximum temperature and Ozone Air Quality Index for the city of Phoenix.

Predictive Modeling

The data for 7 of the largest 10 cities in the US was used to develop the predictive model; this subset consisted of 15 years of data (January 1, 2000 through December 31, 2015) with both pollution and weather components. These cities were New York, Los Angeles, Houston, Phoenix, Philadelphia, San Diego, and Dallas. This constitutes a dataset of 56,073 observations with 27 columns. Categorical variables were created to translate Air Quality Index values into categories (Good, Moderate, Unhealthy, etc).

Initial Models

For each pollutant, a logistic regression model was built to predict Air Quality Index outcome based on a set of features (Average Temperature, Maximum Temperature, Minimum Temperature, Elevation, Average Relative Humidity, Average Dew Point Temperature, Sunrise, Sunset, Average Station Pressure, Average Sea Level Pressure, and Sustained Wind Speed). A train-test split of 80% train/20% test was used.

The initial round of regression analysis resulted in models with relatively high accuracies; however upon further evaluation of the confusion matrices, the models were extremely poor at discriminating the less populated classes.

Logistic regression was repeated with cross-validation and balanced class weighting. The accuracy of the cross validated models was lower but perhaps slightly better in terms of predicting the classes. The imbalance between the classes was still an issue.

Model Improvements

To begin with, all of the feature data was normalized to ensure consistent input. Then the Air Quality Index values were recategorized into less categories (Good versus Elevated) to attempt to better power the smaller classes.

Even with this recombination of pollutant categories, the dataset was simply underpowered with regard to elevated Carbon Monoxide levels. In over 55,000 observations, there were not enough to sufficiently power a predictive model. The analysis going forward focused on pollutants with a larger number of high levels observed.

Using the normalized feature data and new outcome categories, logistic regression with cross-validation and balanced class weighting was repeated and yielded improved results. While some accuracy was lost in each model, the prediction of Elevated pollutant levels was substantially better. ROC curves and AUC were also assessed for each model.

TABLE 1: IMPROVED MODEL METRICS

Pollutant	Accuracy	AUC
Ozone	0.79	0.89
Nitrogen Dioxide	0.72	0.81
Sulfur Dioxide	0.74	0.81

Hyperparameter tuning was then performed using the GridSearchCV function to test a range of lambda values. Using the tuned logistic regression parameters resulted in slight loss to AUC.

TABLE 2: HYPERPARAMETER TUNED METRICS

Pollutant	AUC
Ozone	0.88
Nitrogen Dioxide	0.79
Sulfur Dioxide	0.76

Feature Importance

Recursive feature elimination and random forest classification feature importance were both utilized to assess the most important features to these predictive models. The outcomes from both analyses were compared and contrasted to select the final features to be used in a reduced model:

- Ozone: Maximum Temperature, Average Temperature, Average Station Pressure, and Sustained Wind Speed
- Nitrogen Dioxide: Maximum Temperature, Average Temperature, Average Station Pressure, and Sustained Wind Speed
- Sulfur Dioxide: Average Relative Humidity, Elevation, Maximum Temperature, Minimum Temperature and Sustained Wind Speed

Final Modeling

Utilizing the abbreviated list of features identified as most important to prediction of the outcome variable, logistic regression with cross-validation was performed again to finalize the modeling. With a reduction from 11 to only 4-5 features, the reduced models still perform reasonably well without a drastic loss in accuracy or AUC. The final model metrics are summarized in Table 3 and the ROC curves in Figures 11-13.

TABLE 3: FINAL MODEL METRICS

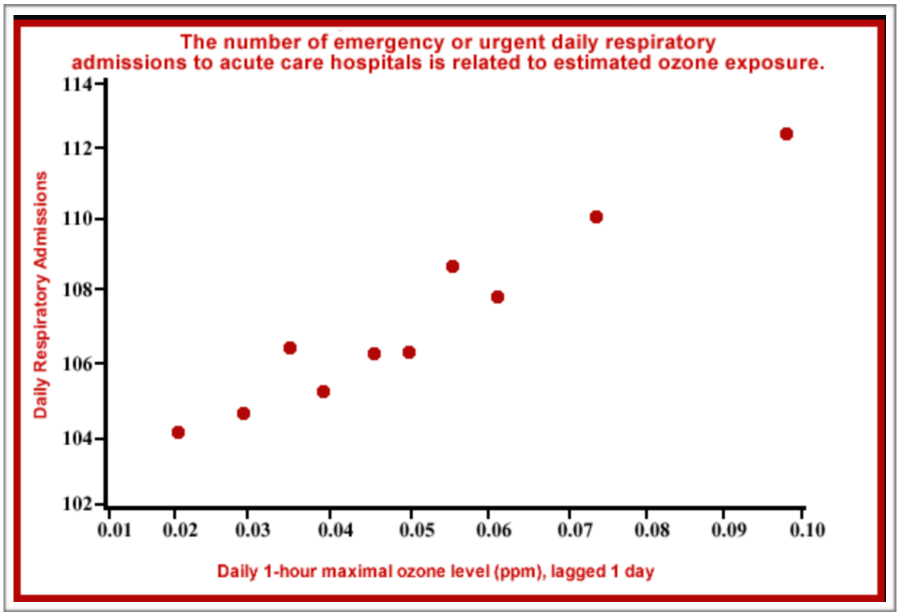
Pollutant	Accuracy	AUC
Ozone	0.76	0.87
Nitrogen Dioxide	0.70	0.75
Sulfur Dioxide	0.74	0.73

Discussion

Conclusions

Per the Environmental Protection Agency, nitrogen dioxide, carbon monoxide and sulfur dioxide are released into the air from burning of fossil fuels (such as operation of motor vehicles) and emissions from industrial plants; ozone is created when these and other air pollutants react in the presence of sunlight. Nitrogen dioxide, sulfur dioxide and ozone harmfully impact the human respiratory system while carbon monoxide affects the cardiovascular system.

The scientific literature strongly supports a relationship between elevated pollutant levels and increased hospitalizations, emergency department visits and mortality. As one example, the figure below shows the number of emergency or urgent daily respiratory admissions to acute care hospitals as related to estimated ozone exposure (Burnett et al., 1994; U.S. EPA, 1996).



Certain features of the weather have been demonstrated here to predict pollution levels as normal/good versus elevated. This could prove useful in managing at-risk populations during periods of elevated pollution levels.

Client Recommendations

Based on the analysis presented, recommendations could be made to Kaiser Permanente in several areas:

- Prevention
  - Population Health Management
    - As Kaiser Permanente prides itself on proactive communication with patients for reminders and health advisories, the organization's communications group should use pollutant forecasting to message members by region with recommendations for elevated pollutant days. In line with EPA recommendations, this may include advisories to avoid time or exercise outside. Given the links between air pollution and induction of new cases of respiratory disease, this is prudent even for the healthy as-yet-unaffected members of the population.
  - Management of high risk patients
    - For high-risk patients with known respiratory and cardiovascular diseases, or populations such as the elderly and infants, exposure to air pollution can trigger more serious episodes. Therefore, communications based on the pollutant forecasting are even more critical with these patients.
    - Depending on the severity of the pollutant levels, Kaiser Permanente should consider email and robocall messaging to provide information and recommended actions to patients already diagnosed with respiratory diseases, cardiovascular diseases, the elderly and the very young.
    - Kaiser Permanente should also consider using their existing online nurse practitioner and doctor visit system to reach out to their highest-risk patients to discuss any further medical recommendations, such as changes in medication or more aggressive case management.
- Staffing
  - Hospital and Emergency Room
    - The predictive modeling of pollutant levels shown here combined with the predicted increase in health service utilization from the literature could inform Kaiser Permanente staffing planning. The addition of medical staff to outpatient clinics and emergency rooms on days with an expected increase in pollution-caused illness would allow improved management of the spike in visits and perhaps reduce hospital admissions if care could be provided sooner in an outpatient facility.
  - TeleHealth

- Increased staffing of telehealth resources to communicate with patients by phone or the internet in periods of higher pollutant levels could allow management of some patients without utilization of in-person resources.

Taken together, these initiatives have the potential to increase cost savings while improving patient care. One recent example (American Journal of Accountable Care 2017) at Denver Health, a smaller managed care organization, evaluating the usage of population health strategies and predictive modeling to match resourcing with patient needs demonstrated $15.8 million savings over 26 months. A larger organization such as Kaiser Permanente could yield significant savings through increased usage of predictive modeling to address patient health needs.

Future Work

Two specific areas could help expand the utility of this analysis:

- Inclusion of emergency room visits and hospital admissions for respiratory diseases in order to better predict the actual increase in illness caused by pollutant levels. Unfortunately, the only available data through free sources was monthly/annual rates of hospital/ER visits, which do not provide the granularity necessary to make accurate predictions.
- It would be interesting to evaluate weather data leading up to any given day and determine how influential that is on pollutant levels, including if a rolling 3-day average is a better predictor and how large the window of influence is.

Code

All project code can be found in its Github repository here.

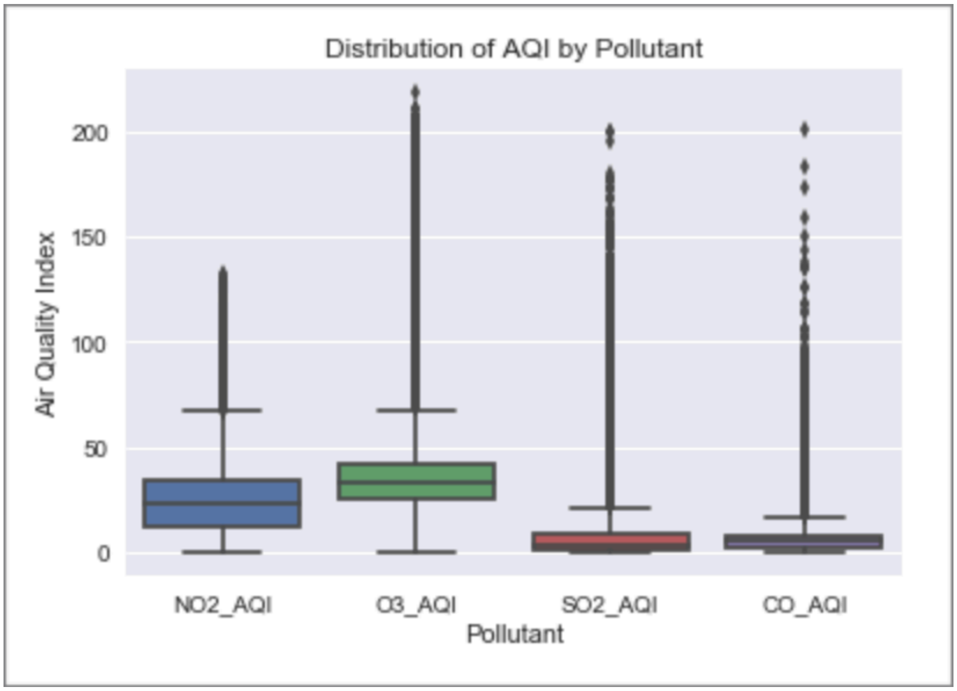


FIGURE 1: BOXPLOT OF AIR QUALITY INDEX BY POLLUTANT

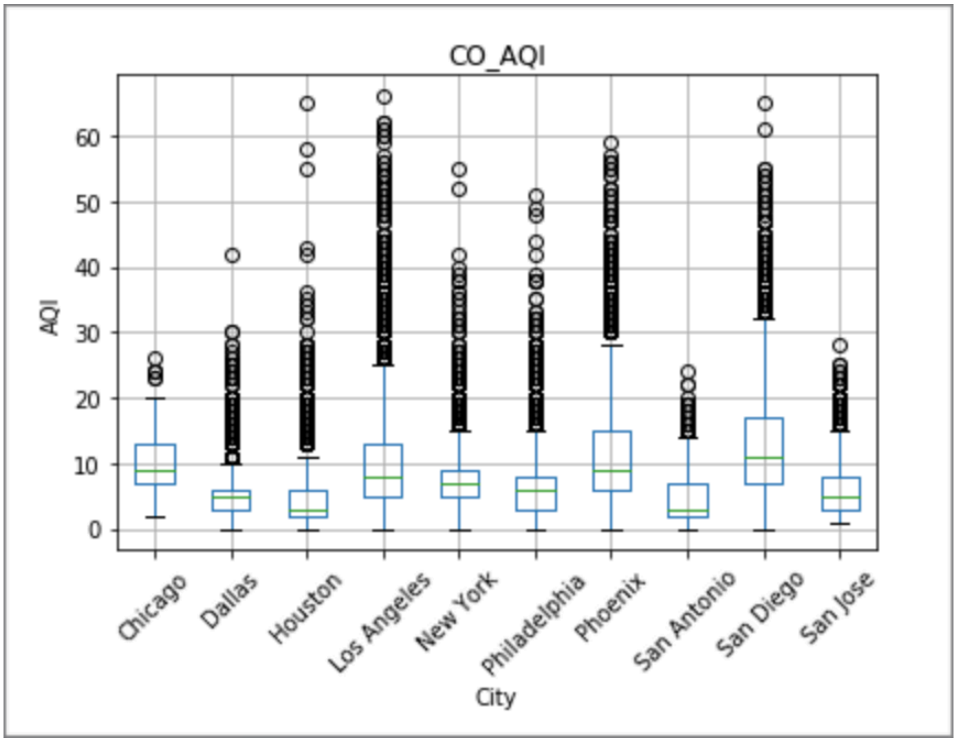


FIGURE 2: BOXPLOT OF CARBON MONOXIDE AIR QUALITY INDEX BY MAJOR CITY

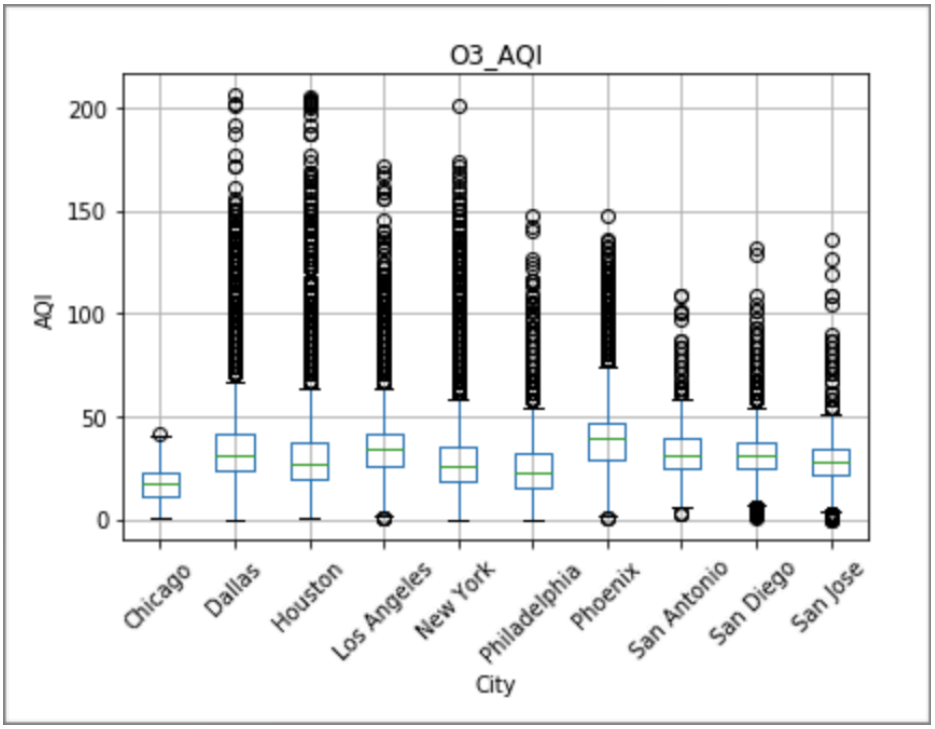


FIGURE 3: BOXPLOT OF OZONE AIR QUALITY INDEX BY MAJOR CITY

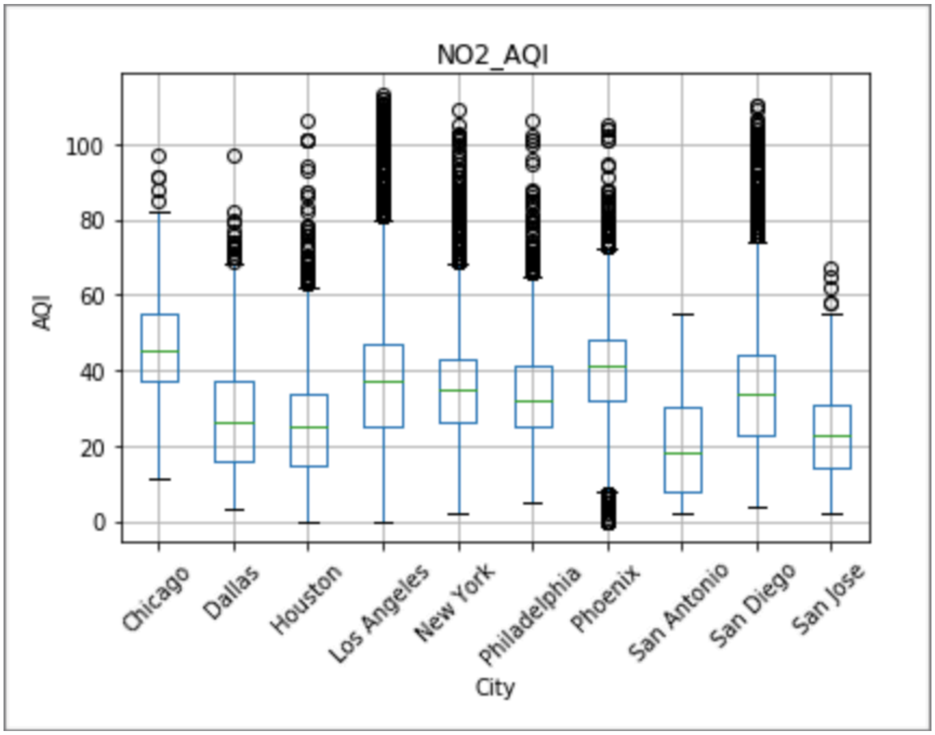


FIGURE 4: BOXPLOT OF NITROGEN DIOXIDE AIR QUALITY INDEX BY MAJOR CITY

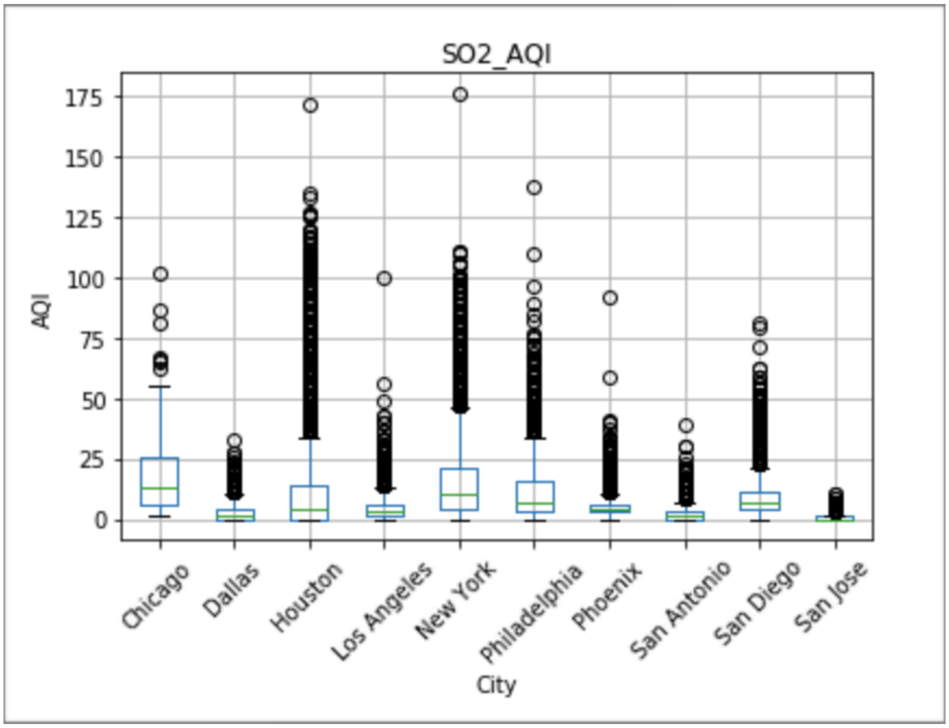


FIGURE 5: BOXPLOT OF SULFUR DIOXIDE AIR QUALITY INDEX BY MAJOR CITY

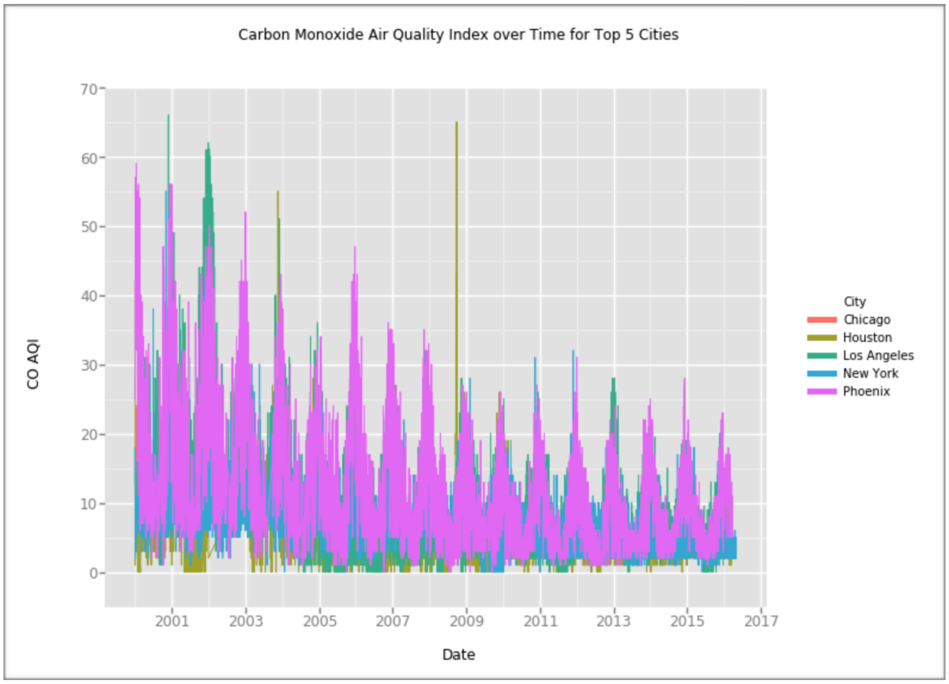


FIGURE 6: CARBON MONOXIDE AIR QUALITY INDEX OVER TIME BY MAJOR CITY

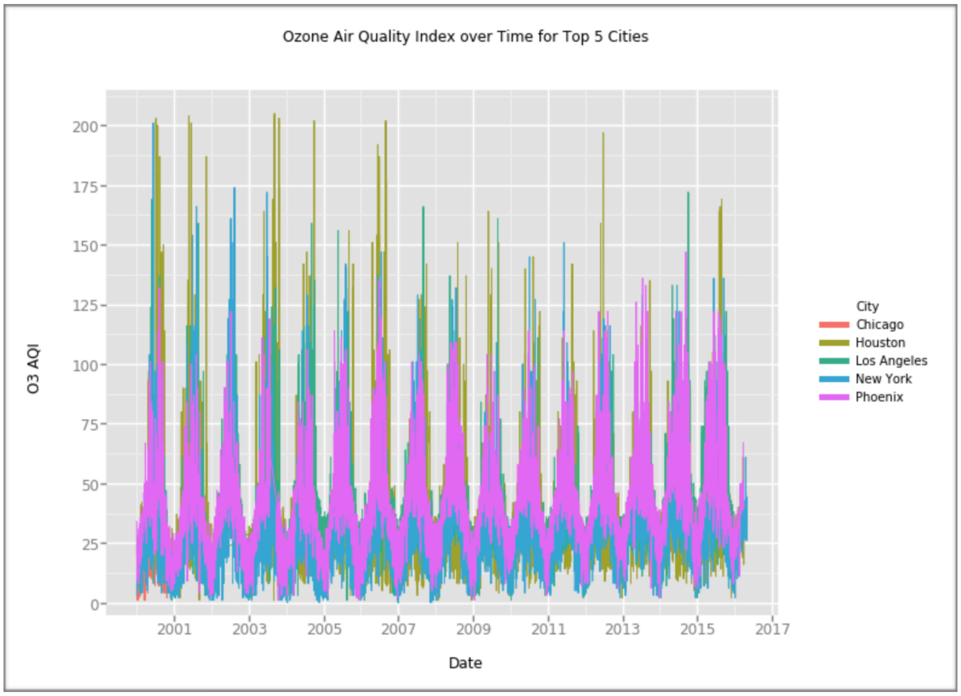


FIGURE 7: OZONE AIR QUALITY INDEX OVER TIME BY MAJOR CITY

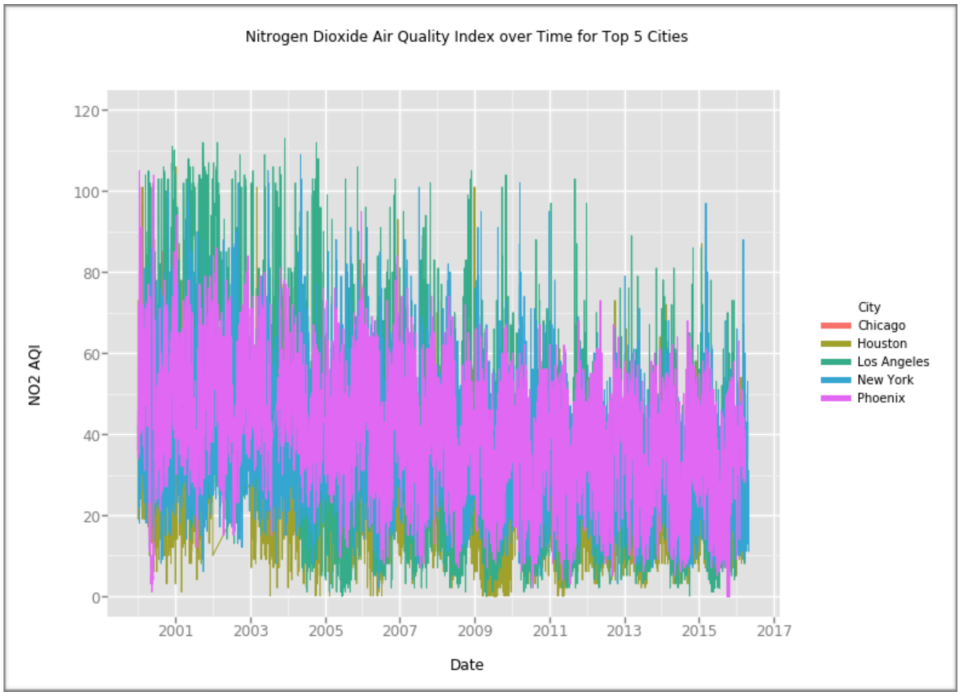


FIGURE 8: NITROGEN DIOXIDE AIR QUALITY INDEX OVER TIME BY MAJOR CITY

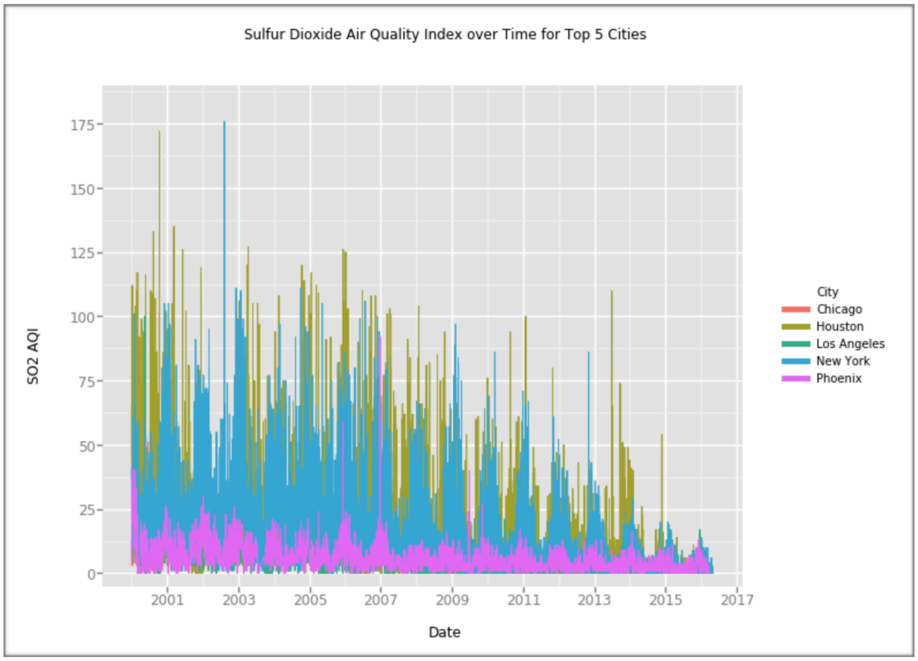


FIGURE 9: SULFUR DIOXIDE AIR QUALITY INDEX OVER TIME BY MAJOR CITY

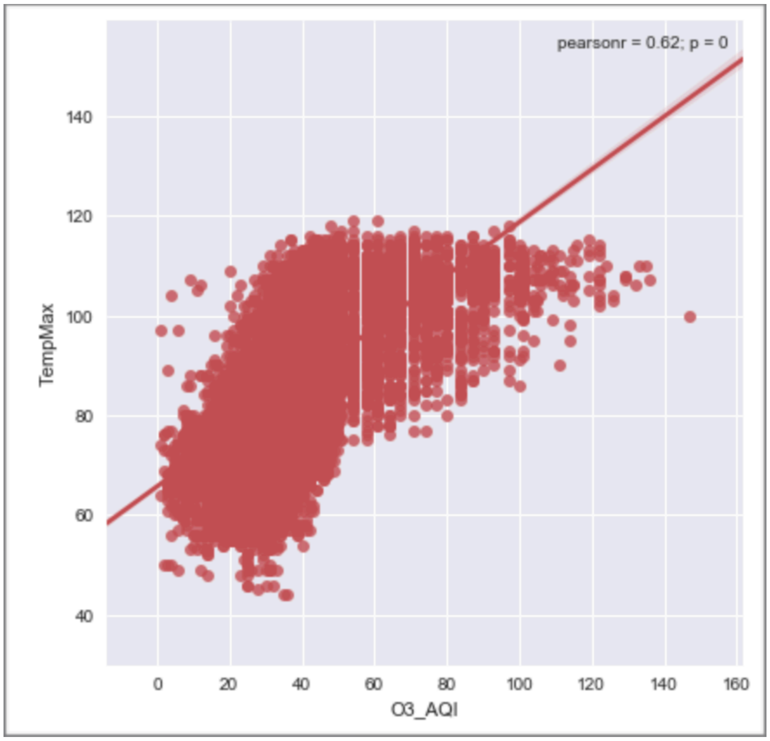


FIGURE 10: CORRELATION BETWEEN OZONE AIR QUALITY INDEX AND MAXIMUM TEMPERATURE FOR THE CITY OF PHOENIX

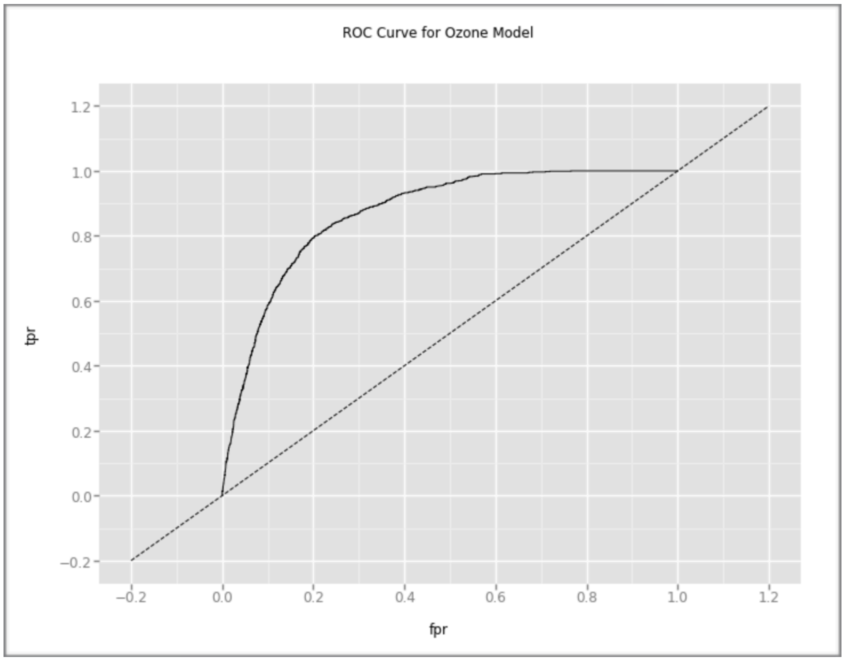


FIGURE 11: ROC CURVE FOR FINAL OZONE PREDICTIVE MODEL

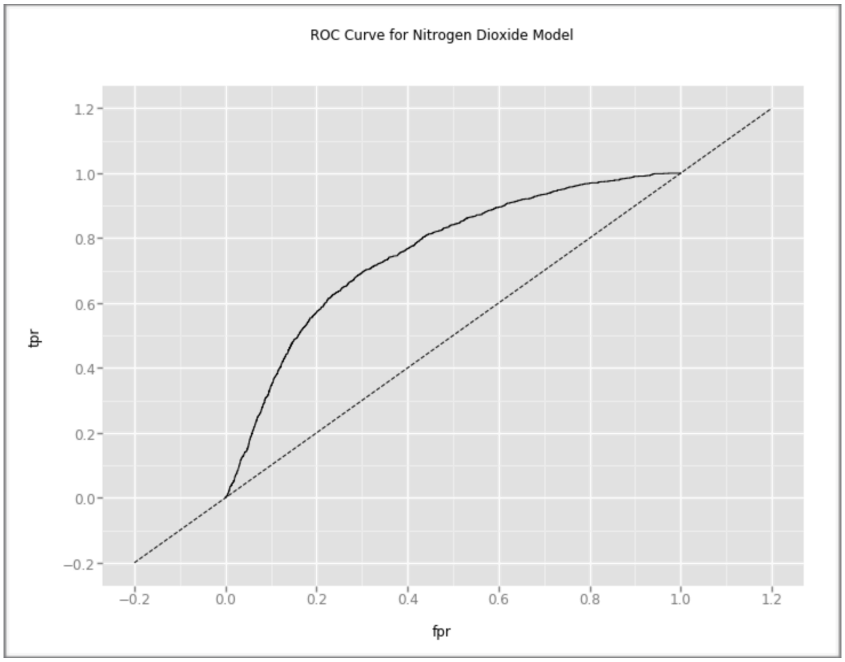


FIGURE 12: ROC CURVE FOR FINAL NITROGEN DIOXIDE PREDICTIVE MODEL

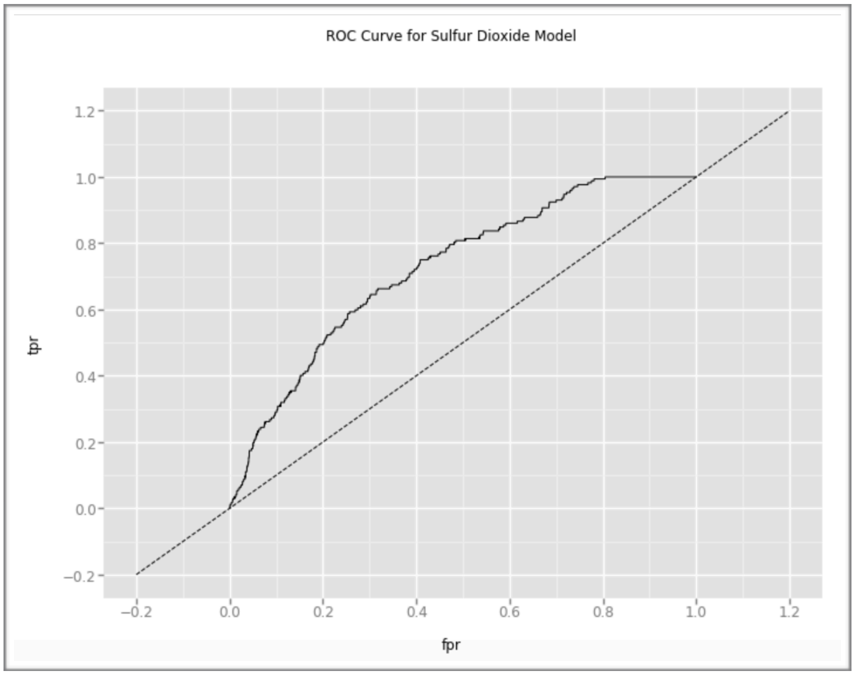


FIGURE 13: ROC CURVE FOR FINAL SULFUR DIOXIDE PREDICTIVE MODEL