A Preliminary Analysis on the Relationship Between Local Air Quality and a Populations' Susceptibility to COVID-19

Ashlynn Steeves
University of Victoria
July 16, 2020

Table of Contents

Table of Figures

Summary

This report reflects the preliminary analysis that has been completed on the relationship between local air quality and a population's susceptibility to COVID-19. The relationship between symptoms associated with exposure to poor air quality and populations that are more susceptible to COVID-19 is explored in order to form a concrete hypothesis and build understanding as to why this is a question worth looking into.

The datasets used in this analysis and their contents are then described. These datasets include the John Hopkins COVID-19 Data Repository, the World Air Quality Index Project daily air quality reports and the United States Environmental Protection Agency county level annual air quality index datasets [1] [2] [3].

The majority of this report focuses on the various approaches that have been taken in order to approach the question regarding the impact of air quality on a populations' susceptibility to COVID-19. Initially, air quality data was specifically extracted and compared across five major Australian cities; problems arose fairly early in this analysis. Such issues included the minimal long-term air pollution experienced in these Australian cities and the proximity in air quality data across the cities. Next, pairwise city level comparisons were attempted for two major American cities. For this approach, issues primarily arose regarding the use of exclusively time series data.

The most recent approach is a county level analysis across the United States. This approach is based on county level annual summary statistics relative to local air quality. This data was combined with local COVID-19 death tolls and confirmed cases to define a feature space. Features were then normalized appropriately. Principal component analysis was then applied to the normalized feature space in anticipation for clustering. The shape of the reduced data indicated that there were likely issues with the normalization and/or the feature selection process. Moving forward these issues will be addressed by re-selecting the features used to describe local air quality.

Future work involves the inclusion of other variables that may be influencing a population's susceptibility to COVID-19 within the analysis. This analysis will continue to develop over the coming month.

Problem Definition

The COVID-19 pandemic is a widespread disease projected to claim the lives of hundreds of thousands of Americans [4]. COVID-19 is a highly infectious acute respiratory illness characterized by symptoms including but not limited to fever, fatigue, dry cough and shortness of breath [5]. As the disease continues to spread it is important to identify factors that may be impacting the health outcomes of populations' exposed to COVID-19.

There is strong and consistent evidence to suggest that individuals with certain underlying health conditions are at a higher risk for contracting COVID-19 [6]. Some of these underlying conditions include chronic obstructive pulmonary disease (COPD), serious heart conditions, hypertension and asthma [6].

Development of each of these pre-existing conditions has been linked to exposure to poor air quality. Research has shown that many respiratory disorders are closely associates with the inhalation of pollutants [7]. For example, major factors that increase the risk of respiratory disorders such as asthma and COPD include long term effects from traffic, industrial air pollution and fuel combustion [7]. Furthermore, a variety of cardiovascular effects such as hypertension, stroke, myocardial infracts and heart insufficiency have been observed following exposure to air pollutants [7].

This overlap has lead to the exploration within this report on the relationship between exposure to poor air quality and a populations' susceptibility COVID-19.

Air Quality Index

Air quality index (AQI) is a scale used for reporting and forecasting daily air quality. The scale focuses on health effects that may be experienced within a few hours or days after breathing polluted air and ranges from 0 to 500 where larger values represent worse air conditions [8]. AQI is subdivided into six categories, each category with its own heath and safety advisories. The six categories, their corresponding index values and their respective advisories are [9]:

- *Good (0-50).* Air quality is satisfactory, and air pollution poses little or no risk.
- *Moderate (51-100).* Air quality is acceptable. There may be a risk for some people, particularly those who are unusually sensitive to air pollution.
- Unhealthy for Sensitive Groups (101-150). Members of sensitive groups may experience health effects. The general public is less likely to be affected.
- *Unhealthy (151-200).* Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
- *Very Unhealthy (201-300).* Health alert: The risk of health effects is increased for everyone.

- *Hazardous (301-500).* Health warning of emergency conditions: everyone is more likely to be affected.

Air quality monitoring stations measure some of the most common ambient air pollutants such as ground level ozone (O3), particle pollution (pm10 and pm25), carbon monoxide (CO), nitrogen dioxide (NO2), and sulphur dioxide (SO2)[8]. For each of the measured pollutants at a given monitoring station, an air quality index is calculated, whichever index is reflective of the worst air quality is reported as the air quality index [10].

Datasets

This section outlines the publicly available datasets that have been used to perform the analysis described in this report.

John Hopkins COVID-19 Data Repository

The primary source of COVID-19 data for this analysis is obtained from the COVID-19 Data Repository by the Center for Systems Science and Engineering at John Hopkins University [1]. More specifically this analysis has used time series data pertaining to confirmed COVID-19 cases and deaths. Data is reported at a country level for all included countries besides Canada, Australia and China whose data is reported at a province/state level and the United States whose data is available at the county level.

World Air Quality Index Project

Used in early iterations of this analysis, the World Air Quality Index Project compiles and provides daily average air quality data for individual air quality monitoring stations around the world [2]. Datasets obtained from World Air Quality Index Project include the AQI value for all pollutants measured by a given monitoring station for all days with reported data. The measured pollutants as well as the available dates are not necessarily consistent across all air quality monitoring stations.

United States Environmental Protection Agency

As this analysis evolved so did the requirements for air quality data. The United States Environmental Protection Agency provides an assortment of publicly accessible pre-generated data files. For this analysis the annual AQI summary data by county was used. These datasets provide county level data for a wide variety of summary measurements relevant to local air quality, such as [11]:

- Number of days with measured air quality indices
- Number of days for which measured air quality index was classified as good
- Number of days for which measured air quality index was classified as moderate

- Number of days for which measured air quality index was classified as unhealthy for sensitive groups
- Number of days for which measured air quality index was classified as unhealthy
- Number of days for which measured air quality index was classified as very unhealthy
- Number of days for which measured air quality index was classified as hazardous
- Maximum recorded air quality index
- The value for which 90% of the rest of the measured air quality index values are equal to or less than
- Median air quality index
- Number of days where the reported air quality index was based on the carbon monoxide measurements
- Number of days where the reported air quality index was based on the nitrogen dioxide measurements
- Number of days where the reported air quality index was based on the ground level ozone measurements
- Number of days where the reported air quality index was based on the sulphur dioxide measurements
- Number of days where the reported air quality index was based on the particle matter 2.5 measurements
- Number of days where the reported air quality index was based on the particle matter 10 measurements

Past Approaches

As is common with many data analysis problems the approach to answering the question regarding the impact of air quality on a populations' susceptibility to COVID-19 has shifted a number of times over the course of the project. This section outlines some of the initial approaches that were taken to investigate this question.

Australian Wildfires

The idea for this analysis was initially sparked by an interest in the Australian wildfires and their impacts on COVID-19 spread throughout the country. This quickly evolved into looking at the impacts of poor air quality on COVID-19 spread, throughout Australia as wildfires can drastically impact air quality.

This brief analysis focused on 5 major Australian cities: Adelaide, Brisbane, Melbourne, Perth, and Sydney. For each city air quality data was obtained from the World Air Quality Index Project for all monitoring stations within city limits. The daily AQI was then averaged across all monitoring stations within a given city in order to obtain a true sense of the air quality on a given day and reduce the impacts

of potential outliers in the data. This air quality data, as well as a moving mean of this data was then plotted for all cities, as seen below in figure 1.

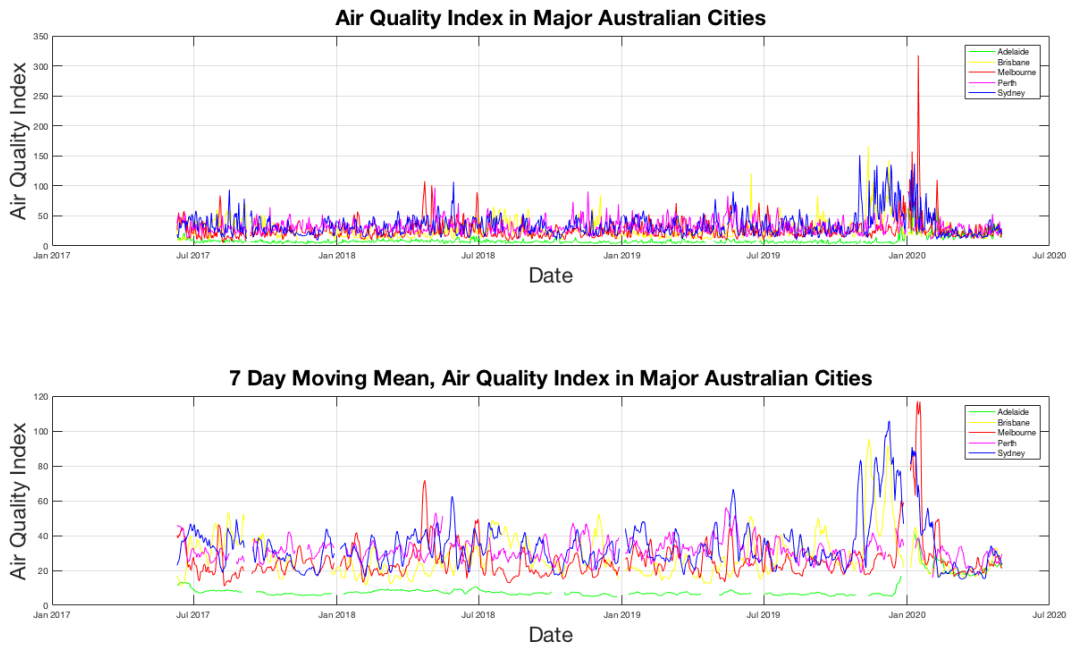

Figure 1 – Top: Air Quality Index in Major Australian Cities, Bottom: 7 Day Moving Mean of Air Quality Index in Major Australian Cities

This figure and others like it highlighted some key issues with this approach. First, many of the adverse effects of air pollution that have been hypothesised to make a population more susceptible to COVID-19 have stronger links to long term exposure to poor air quality. Analysis of these figures indicates that aside from the 2020 wildfires the AQI in any of these cities surpasses a moderate air quality index less than 25 times. This suggests that residents of these cities have not experienced long term expose to poor air quality. Furthermore, the variation in air quality between the cities is quite small consequentially making them difficult to separate in a data analysis context.

Pairwise Analysis

Following the inspection of the Australian data, the analysis shifted to focus specifically on the comparison of locations with significant differences in air quality levels. A variety of cities with few confounding variables between them were briefly inspected to determine their compatibility to this analysis. Ultimately, Los Angeles, California and Seattle, Washington were chosen as locations for initial analysis. Air Quality data was collected using the same methodology as for the Australian cities. Figure 2 below shows a plot of the averaged air quality data across Los Angeles vs. Seattle. Figure 3 shows the difference in daily air quality indices, where positive values represent a higher AQI in Los Angeles and negative values represent a higher AQI in Seattle.

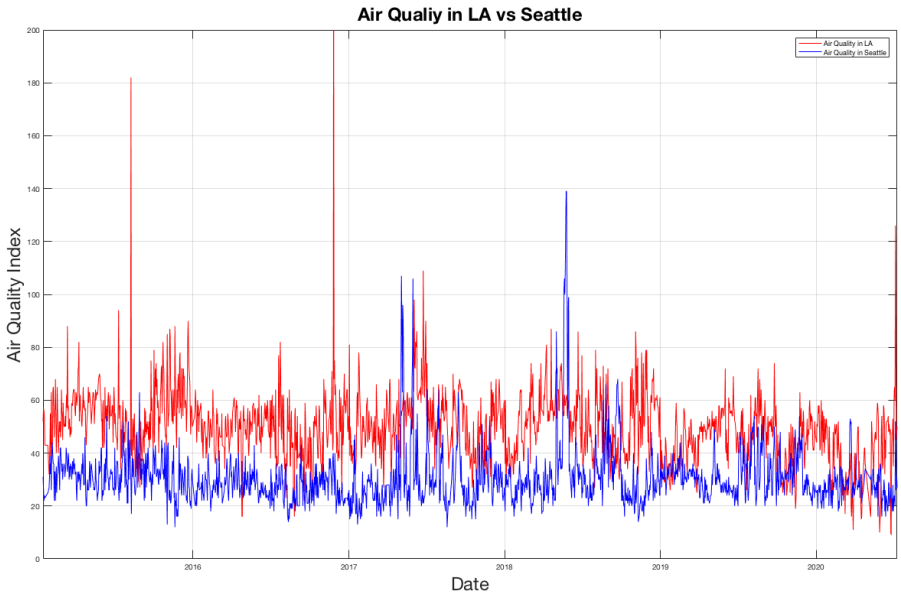

Figure 2 - Air Quality Index in Los Angeles and Seattle

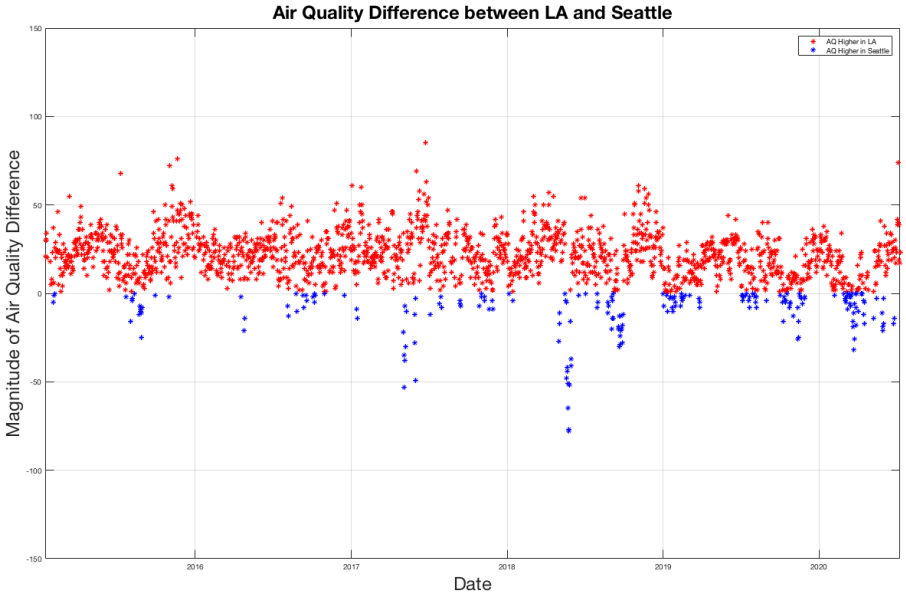

Figure 3 - Differences in Air Quality Index Between Los Angeles and Seattle

Following the initial analysis of the air quality data from these two cities their respective COVID-19 infection rates and death rates were analysed. At this point complications were arising from working exclusively with time series data and the appropriate methodology to move forward with this analysis was unclear. These complications lead to the discovery of a new source of air quality data, which introduced the shift in techniques used in future iterations of this analysis.

County Level Analysis

The most recent approach to analysing the relationship between local air quality and a populations' susceptibility to COVID-19 makes use of county level air quality data from the United States Environmental Protection Agency [3]. Annual summary datasets as described previously were obtained for 2014 through 2019. This analysis makes use the number of days for which AQI was reported and the number of times the recorded AQI fell into each of the six AQI categories. Pre-processing of this data involved removal of any counties who reported air quality measurements fewer than 180 days per year. The counties represented in this analysis are shown below in figure 4, note that there are also 4 counties represented in Hawaii and none in Alaska.

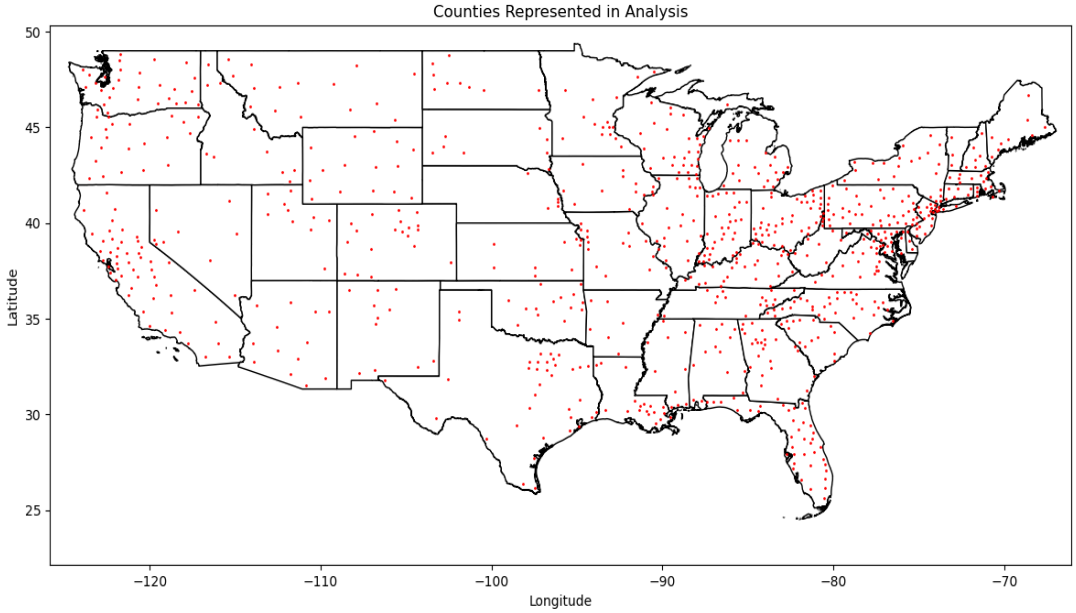

Figure 4 - Counties Represented in Analysis

Following the removal of counties with insufficient data, the number of occurrences of an AQI from each category was then divided by the total number of recorded days, turning this value into a percentage. These percentages were then averaged across the 6 years for which data was collected. These averaged rates as well as the total number of confirmed COVID-19 cases and death in a given county (both normalized by the counties population) were used to define the feature space for this analysis [1][12].

Feature level normalization was then applied in order to remove the variation of the respective feature scales. Normalization for all features followed the formula:

$$\tanh\left(\frac{arctanh(a)}{b}\boldsymbol{x}\right)$$

Where a represents the tolerance, b represents the maximum value within a given feature and x is the data. This method of normalization defines a curve where the maximum value within a given feature (b) is mapped to the tolerance value (a). The tolerance value also influences the rate of decay between 1 and 0. Figure 5 provides a visual representation of the impact of varying the tolerance value on the normalization.

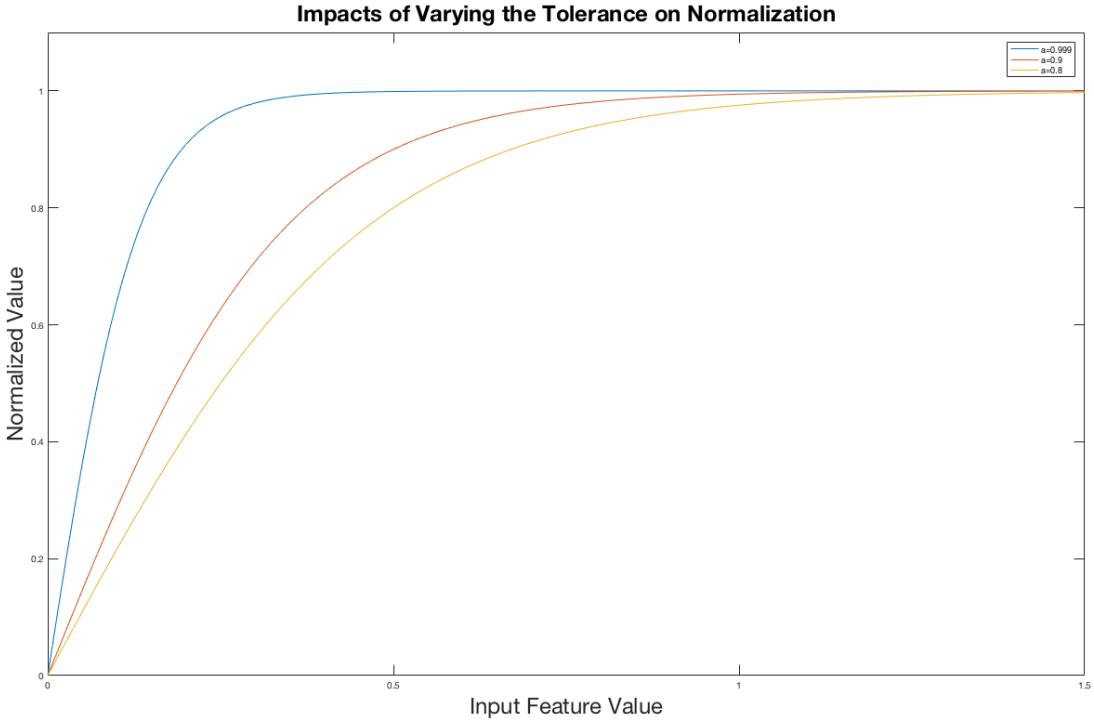

Figure 5 - Impacts of Varying the Tolerance on Normalization

Figure 6 provides an example of this normalization on the feature representative of confirmed COVID-19 cases, here a tolerance of 0.99 was used and the maximum feature value was 0.0772. Similar normalization histograms for each feature can be found in appendix A.

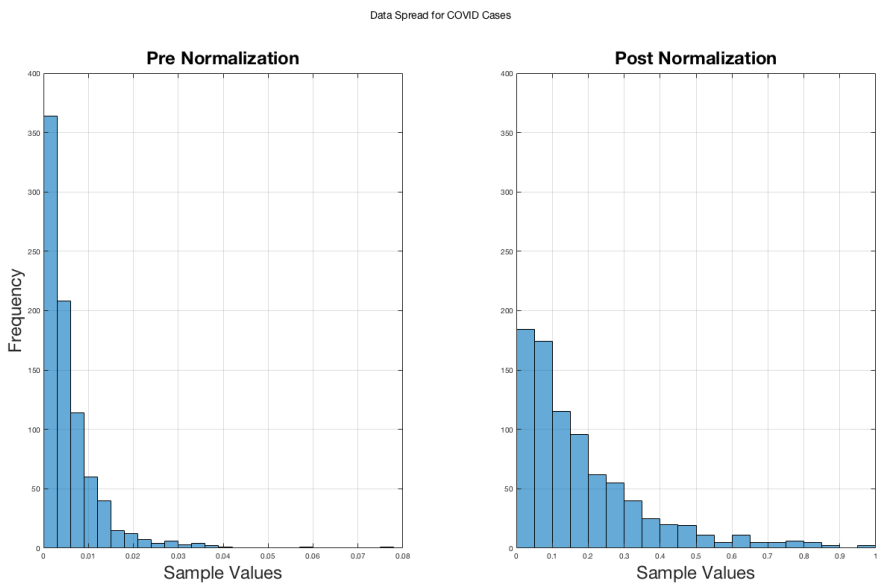

Figure 6 - Normalization of COVID-19 Cases

Once all 8 features had been normalized into a zero to one range principal component analysis was applied to the feature space as a method to reduce the dimensionality of the data in anticipation for clustering. Principal component analysis involves the projection of data onto new axes' defined by the eigenvectors of the covariance matrix of the original feature space. The eigenvectors that the data is projected onto are those with the largest correspond eigenvalues. The magnitude of these eigenvalues is representative of how much of the data is stored within the corresponding eigenvectors. Figure 7 shows the values of the eigenvalues generated for the normalized feature space.

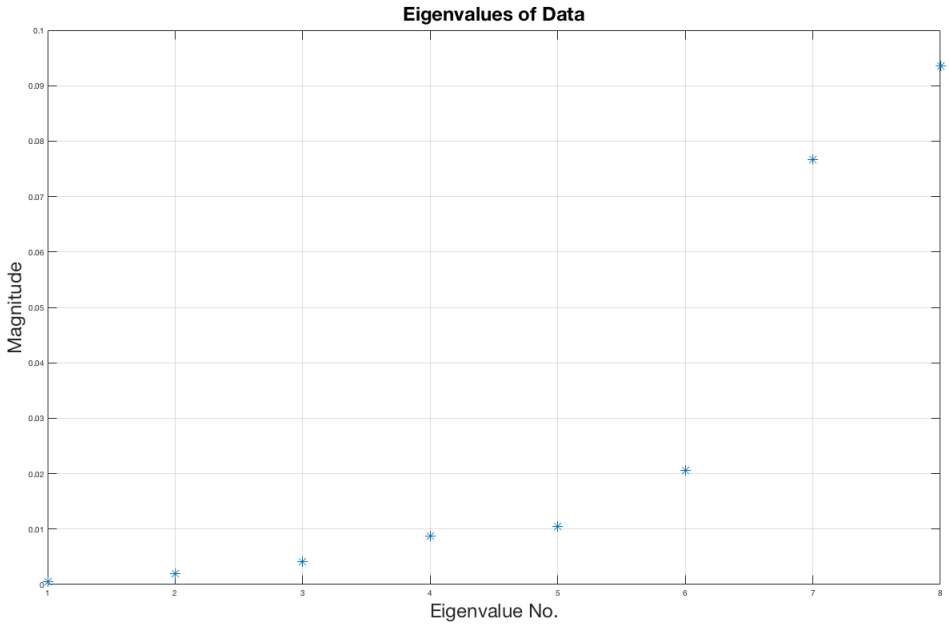

Figure 7 - Eigenvalues of the Normalized Feature Space

We can see in figure 7 that the magnitude of the eigenvalues drops off fairly quickly after the first two. Therefore, reduction of the data to two dimensions should not discard too much relevant information. Figure 8 shows the data after it has been reduced to two dimensions.

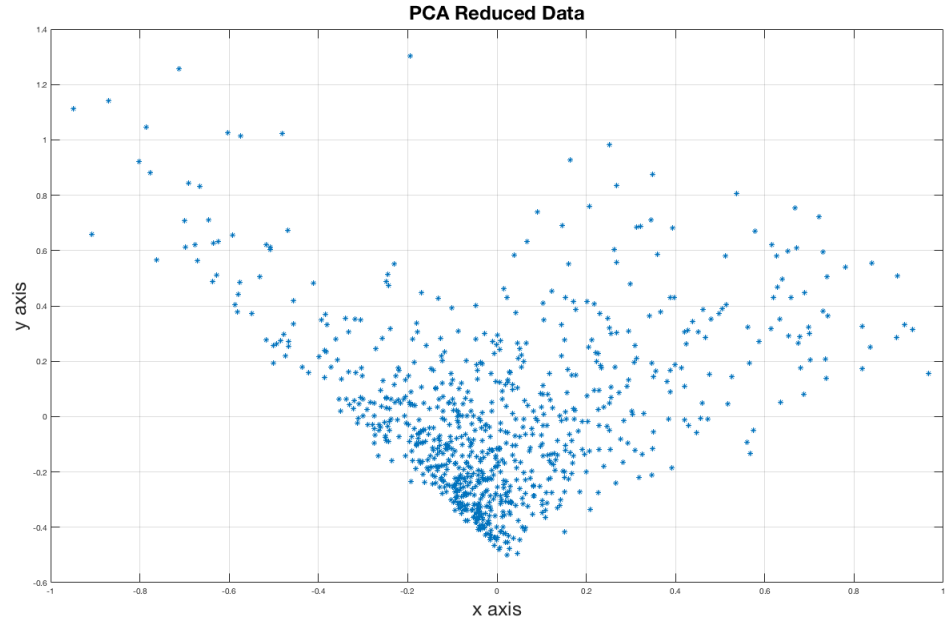

Figure 8 - PCA Reduced Data

The shape of the reduced data is very reminiscent of a rotated x-y axis. This indicates that there are likely issues with the normalization and/or the feature selection process. Moving forward in this analysis features will be adjusted to reflect the available data regarding county level median AQI, maximum AQI and 90th percentile data in hopes of remedying this issue.

Future Work

As this analysis progresses it is important to account for other variables within a county that might be impacting the populations' susceptibility to COVID-19. Some of these factors include, population density, environmental factors such as temperature and humidity, the average age of the population, socio-economic status, state politics and social distancing regulations.

Conclusions

As is common in many data analysis problems the approach taken to answer the question of whether air quality has an impact on a population's susceptibility to COVID-19 has shifted many times. High quality county level air quality has played a large role in the progression of this analysis, which will continue over the next month.

References

[1] Systems Science and Engineering (CSSE) at Johns Hopkins University, "CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE." https://github.com/CSSEGISandData/COVID-19 (accessed Jul. 14, 2020).

[2] W. A. Q. I. Project, "Air Quality Historical Data Platform." https://aqicn.org/data-platform/register/ (accessed Jul. 14, 2020).

[3] U.S. Enviromental Protection Agency, "Download Files | AirData | US EPA." https://aqs.epa.gov/aqsweb/airdata/download_files.html (accessed Jul. 15, 2020).

[4] U. of Washington, "COVID-19 Projections." https://covid19.healthdata.org/united-states-of-america (accessed Jul. 13, 2020).

[5] D. Wang *et al.*, "Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China," *JAMA - J. Am. Med. Assoc.*, vol. 323, no. 11, pp. 1061–1069, Mar. 2020, doi: 10.1001/jama.2020.1585.

[6] Centers for Disease Control and Prevention, "Evidence used to update the list of underlying medical conditions that increase a person's risk of severe illness from COVID-19 | CDC." https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/evidence-table.html (accessed Jul. 14, 2020).

[7] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and Health Impacts of Air Pollution: A Review," *Frontiers in Public Health*, vol. 8. Frontiers Media S.A., p. 14, Feb. 20, 2020, doi: 10.3389/fpubh.2020.00014.

[8] U.S. Enviromental Protection Agency, "Patient Exposure and the Air Quality Index | Particle Pollution and Your Patients' Health | US EPA." https://www.epa.gov/pmcourse/patient-exposure-and-air-quality-index (accessed Jul. 13, 2020).

[9] U.S. Enviromental Protection Agency, "AQI Basics | AirNow.gov." https://www.airnow.gov/aqi/aqi-basics/ (accessed Jul. 14, 2020).

[10] U.S. Enviromental Protection Agency, "Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI)." Accessed: Jul. 13, 2020. [Online]. Available: https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf.

[11] U.S. Enviromental Protection Agency, "AirData Download Files Documentation." https://aqs.epa.gov/aqsweb/airdata/FileFormats.html#_annual_summary_files (accessed Jul. 14, 2020).

[12] U. C. Bureau, "County Population Totals: 2010-2019," Accessed: Jul. 14, 2020. [Online]. Available: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html.

APPENDIX A – Normalization Histograms

This appendix includes normalization histograms for each feature in the feature space defined in the most recent iteration of this analysis. Note that the data for

‘Good AQI Days’ was not normalized as the un-normalized data was well spread out across the desired feature scale.

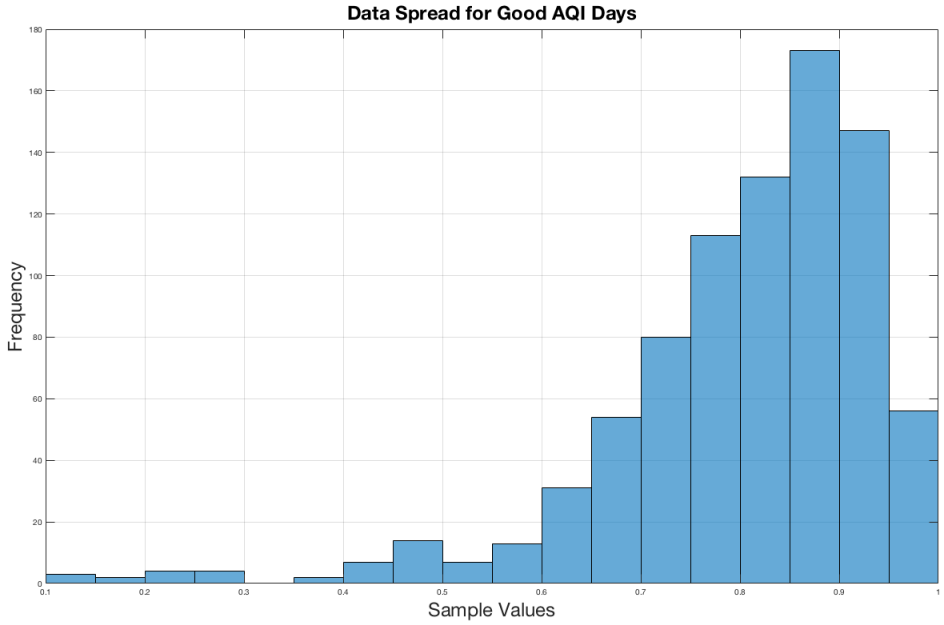

Figure 9 – Appendix A, Data Spread for Good AQI Days

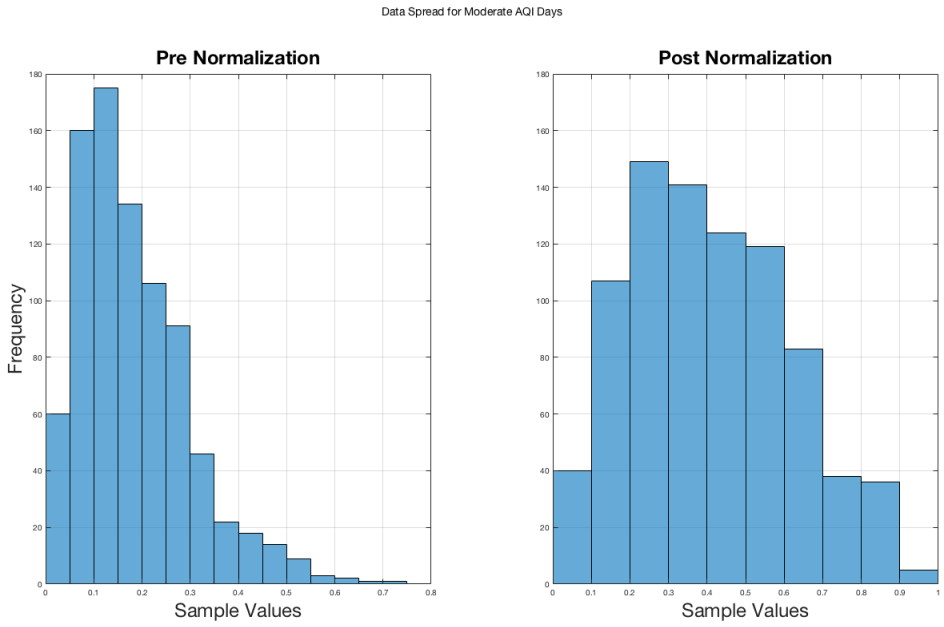

Figure 10 - Appendix A, Data Spread for Moderate AQI Days

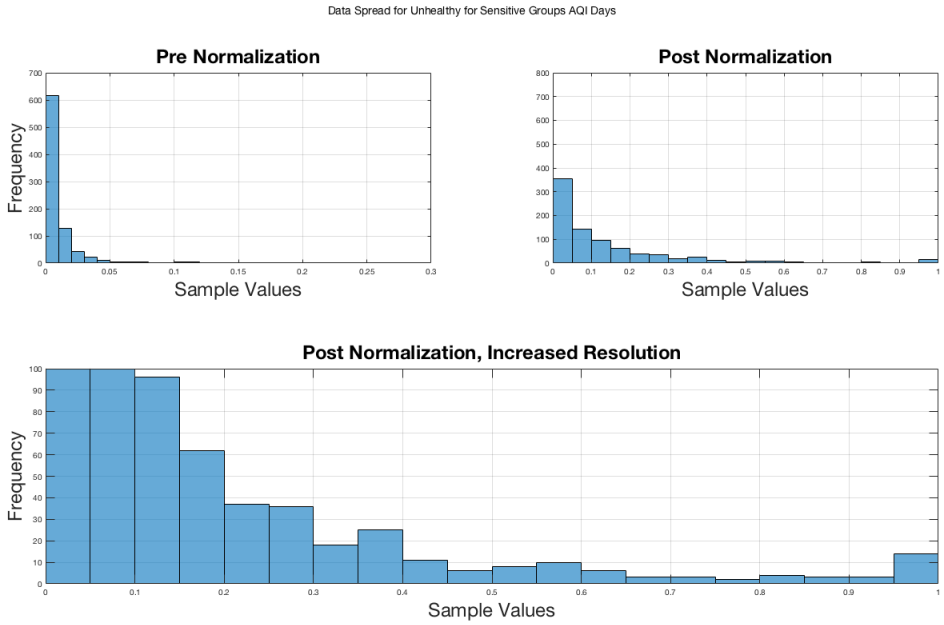

Figure 11 - Appendix A, Data Spread for Unhealthy for Sensitive Groups AQI Days

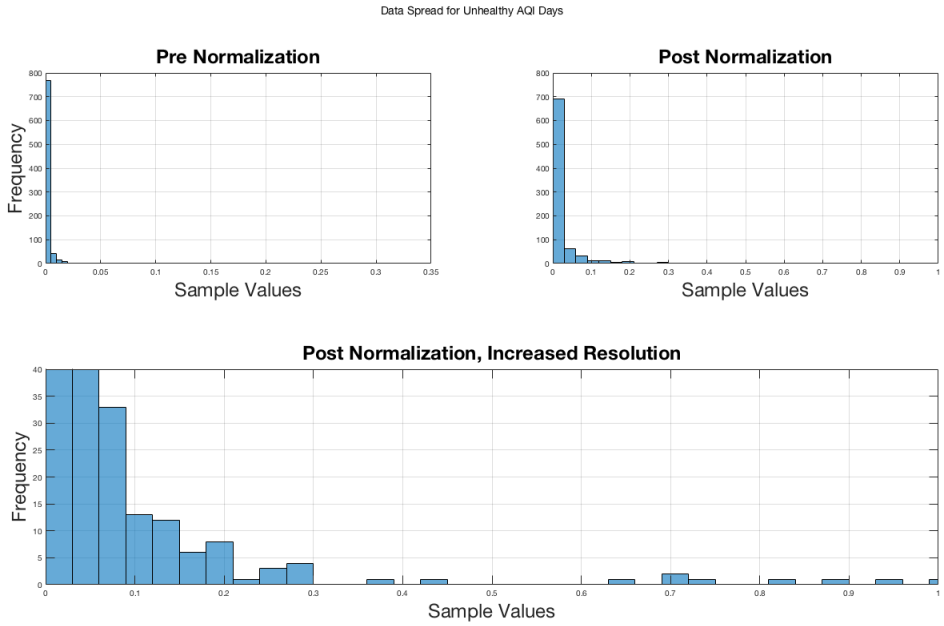

Figure 12 - Appendix A, Data Spread for Unhealthy AQI Days

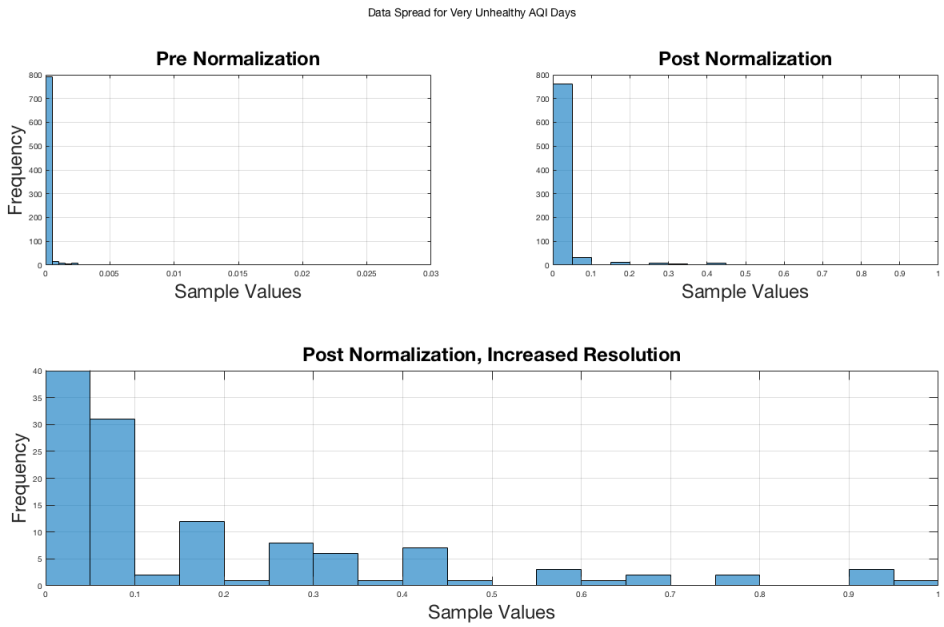

Figure 13 - Appendix A, Data Spread for Unhealthy AQI Days

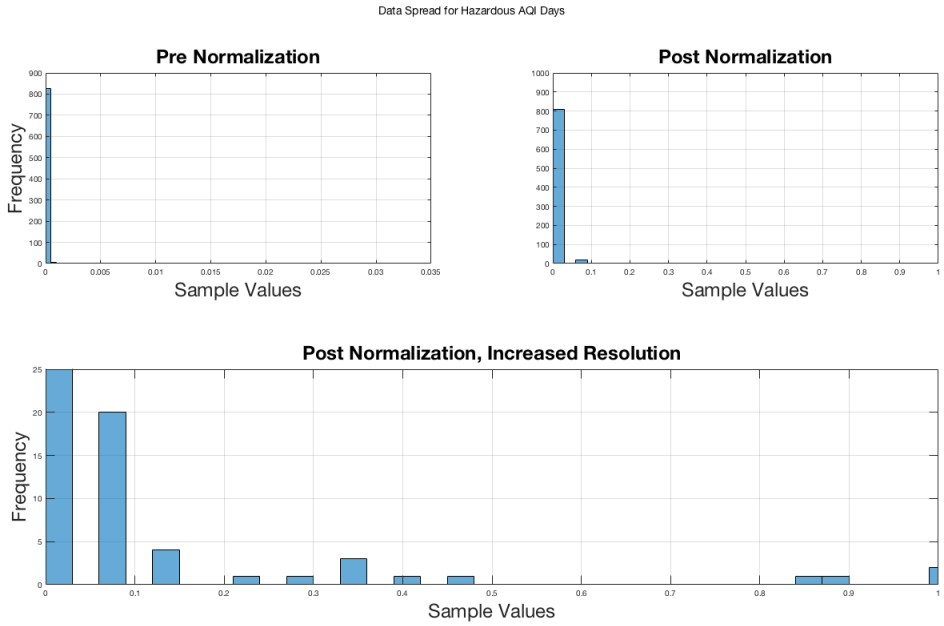

Figure 14 - Appendix A, Data Spread for Hazardous AQI Days

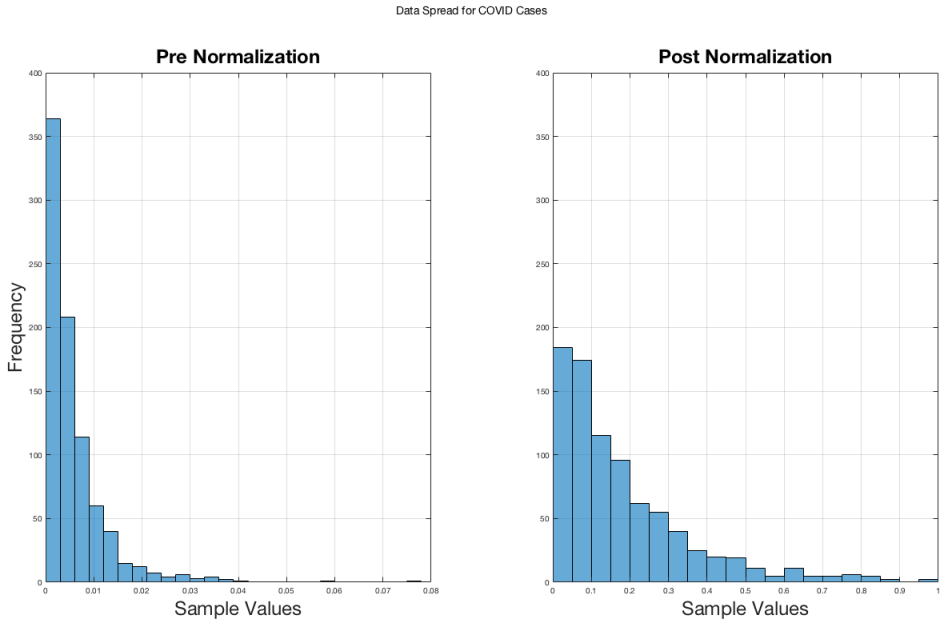

Figure 15 - Appendix A, Data Spread for COVID-19 Cases

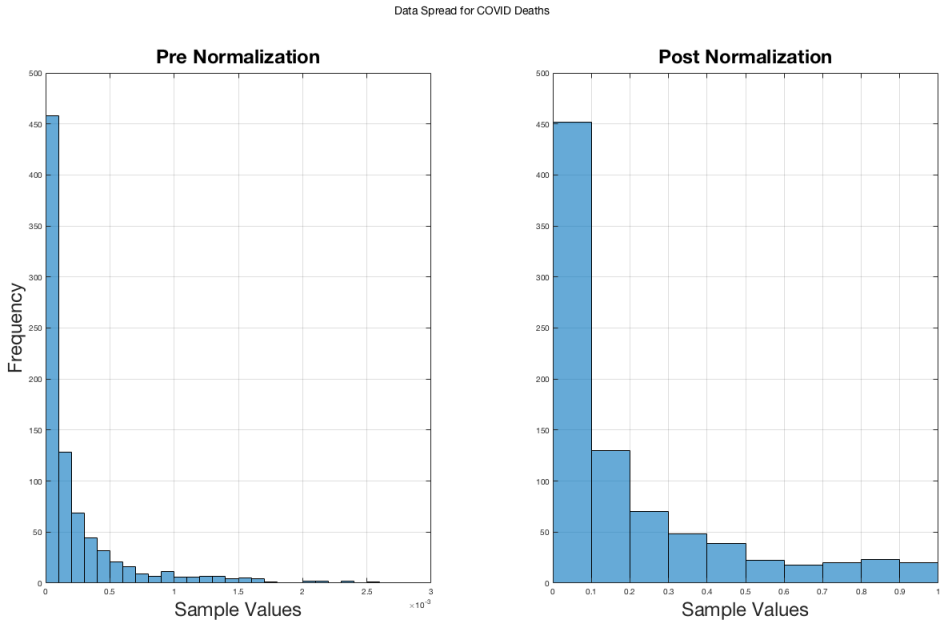

Figure 16 - Appendix A, Data Spread for COVID-19 Deaths