The Effects of reduced mobility on Chicago Air Quality Index

Zachary Lowell
Hiwot Gebreyohannes
Jordan Gilliland
Prachiben Chilkesh Patel
Alexander Rosenblum

7.15.2020

Table of Contents

Table of Contents 2

Abstract 3

Introduction 5

Methodology 8

Analysis, Results and Findings 11

Future Work 29

References 31

Appendix 34

Abstract

Starting in November of 2019, a virus by the name of SARS-CoV-2 began its initial transmission from patient zero in Wuhan, China, to people across the planet located within the United States. Experts say SARS-CoV-2 originated in bats. That's also how the coronaviruses behind Middle East respiratory syndrome (MERS) and severe acute respiratory syndrome (SARS) got started.This virus, famous at this point, is well known for its negative symptoms, high transmission rate, and relatively high mortality rate - making it one of the deadliest pandemics modern society has had to deal with.

Within the United States, the spread was handled by initial tracing of those with the disease, to restaurant shutdowns, to the extreme of stay at home orders, forcing citizens to stay in the comfort of their own home to stop transmission rates from spiking. The common phrase that was spoken across the country - "Flatten the Curve" - was ubiquitous with SAR-CoV-2 (Covid-19 from this point onwards) - meaning that by staying indoors and cutting off the transmission between people, the rate of infection would gradually decrease - allowing for an overwhelmed hospital to be able to handle the influx of patients.

Through these stay at home orders, the U.S. economy took massive hits, from small businesses going under, to large companies declaring bankruptcy. This led to an overall unemployment rate going from what was 3.8% to 20.0% in the span of three months from February 2020 to May 2020. Through this, we saw a decrease in travel to offices through cars, a decrease in plane activity due to travel, and a decrease in pollutant emission due to shutdown of non-essential companies.

Due to this decreased stimulation of emissions from the stay at home order forced from Covid-19, we tended to see a decreased air quality index (AQI from this point onwards), which is a quantitative air quality measure, across the United States. Being that the team writing this paper is in Chicago, it made the most sense to analyze how these factors contributed to the overall AQI within this city. By using the variables focused on AQI that were affected by the stay at home order through Covid-19 (emissions, weather patterns, and travel), we are able to create a regression analysis model that can predict the overall AQI if the stay at home order were to continue or if it were to be lifted allowing travel restrictions and work in office and factory to resume.

This research papers aims to show the correlation and effects of COVID-19 on the city of Chicago, particularly on how contributions from Air travel, Car travel, pollutant emissions, Temperature, and Precipitation, through the last two and a half years have affected the AQI and the sudden impact of the decrease of the variables due to the saty at home order has on AQI.

After conducting our analysis using the data described above, we were able to determine that there was an effect on the overall Chicago AQI due to the stay at home order. We inevitably had to transform the data initially as the AQI has a right skewed trend within the data, so we determined that we would do a logarithmic transformation on the data. One the transformation was done, we proceeded to conduct model diagnostics and selection to ensure the independent variables contributed to the dependent variable and that whatever outliers and influential points within the data that occurred, were taken out. Once this was done, we produced a final model and conducted validation to ensure it was producing the best outcome. Although our model produces a lower than expected adjusted R^2 value, other methods of validation ensure us that the stay at home order affected the overall AQI in the city. Our final model equation can be seen below:

$$ln(Chicago\ AQI) = 3.86043 - Avg\ Wind\ Speed * 0.01951 + Avg\ Temperature * 0.00641 + NO2 * 0.00930 - Flight\ Count * 0.000136 - SAH * 0.10646$$

This shows that of the entire dataset sourced, our final model for prediction of the AQI within Chicago encompasses wind speed, temperature, NO_2 levels, flight count, and stay at home order.

Introduction

Scope

It is well documented that automobiles and airplanes powered by fossil fuels emit exhaust with high concentrations of substances that are harmful when inhaled[1] . Areas with high population densities are particularly affected by this reality due to the high volume of daily traffic, and major cities which act as employment hubs attracting commuters from surrounding suburbs will see an average air quality considerably lower than rural areas with similar emissions regulations[2].

Recently, the COVID-19 pandemic and the associated lockdown that many states and cities put into place caused a significant drop in daily traffic in areas around the United States that are typically very heavy with commuter activity. Additionally, air travel has reduced significantly as travel bans and concerns for personal safety have kept most would-be vacationers and business trippers from boarding planes. This unique situation gives the opportunity to assess just to what extent daily commuter traffic and air travel have on local air quality, and how much harm to local populations, wildlife, and global climate can be prevented if automobile and air traffic are reduced or made significantly more fuel efficient. Due to technical and logistical limitations, this study will be limited to the Chicagoland area, specifically the area covered by the Cook County, IL geographic area defined by the EPA Air Quality Index Daily Values Report[17].

Overview of AQI

Air Quality Index is a daily measure of air pollution from common harmful sources, ranging from 0 to 500 where 0 is the cleanest the air can be and 500 indicates extreme hazard. It was first officially used by US government agencies to report air health in 1977 as an amendment to the Clean Air Act (originally under the name Pollution Standards Index with the official name being changed to Air Quality Index (AQI) by the EPA in a 1999 revision) to measure ambient levels in populated areas of ground-level ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), fine particulate matter ($PM_{2.5}$) and coarse particulate matter (PM_{10})[3].

These pollutants are measured as running averages of concentration of each pollutant over time by over a thousand monitors across the United States[4] , with the AQI for each pollutant being calculated individually based on the relative hazard associated with the concentration measured. Overall AQI is simply reported as the highest AQI value of the individual pollutants measured. For example, if the AQI of O_3 at a given time is 127 and all other pollutants have AQI values between 50 and 100, the overall AQI will be reported as 127 since that is the highest of the individual values[5] .

As previously stated, AQI is measured on a scale from 0 to 500. Typically, any reported value under 100 indicates that the air is healthy to breathe. You may be familiar through the weather app on your smartphone with the phrase “Unhealthy Air Quality for Sensitive Groups” – this alert means that AQI for that day falls between 101 and 150. The specific groups affected change slightly depending on the main offending pollutant, but generally sensitive groups include young children, elderly adults, people with heart or lung disease, and people who spend much time outdoors (construction workers, for example)[4]. Reported AQI values above 150 are considered unhealthy to everyone, with values between 300 and the maximum 500 being considered emergency conditions.

According to the AQI Daily Values Report published by the EPA, the most common main pollutants reported since 2018 for the Cook County, IL geographic area are $PM_{2.5}$ and Ozone.[17] To appreciate the scope of the effect these pollutants have, it is important to understand what they are, where they come from, and through what mechanisms they affect life daily and in the long-term.

AQI Value	**Actions to Protect Your Health from Ozone**
Good (0 - 50)	None
Moderate (51 - 100)*	Unusually sensitive people should consider reducing prolonged or heavy outdoor exertion
Unhealthy for Sensitive Groups (101 - 150)	The following groups should reduce prolonged or heavy outdoor exertion: - People with lung disease, such as asthma - Children and older adults - People who are active outdoors
Unhealthy (151 - 200)	The following groups should avoid prolonged or heavy outdoor exertion: - People with lung disease, such as asthma - Children and older adults - People who are active outdoors Everyone else should limit prolonged outdoor exertion.
Very Unhealthy (201 - 300)	The following groups should avoid all outdoor exertion: - People with lung disease, such as asthma - Children and older adults, - People who are active outdoors. Everyone else should limit outdoor exertion

Table summarizing AQI values and vulnerable groups associated with each AQI level classification[4]

Fine Particulate Matter

Fine particulate matter, or $PM_{2.5}$, is a broad category of pollutants defined by being solid or liquid inhalable particles smaller than 2.5 micrometers (μm) in diameter. For reference, the average diameter of a human hair is 50-70 µm. $PM_{2.5}$ can come from many sources, but they are primarily formed from atmospheric chemical reactions of pollutants from power plants, industry, and automobile traffic[6]. $PM_{2.5}$ can also be emitted directly from sources like construction sites and smokestacks.

One of the more commonly experienced detrimental effects of $PM_{2.5}$ is haze[8] . Though visibility reduction can be caused by a wide array of airborne substances, $PM_{2.5}$ tend to be the worst contributors as they are particularly effective at reducing visibility as its small size allows it to both travel great distances and scatter light with high efficiency[7]. This effect of $PM_{2.5}$ mostly affects places where visibility is typically high, like rural areas or national parks, though some large cities like Los Angeles also commonly experience reduced visibility from particulates.

Perhaps of greater import to city dwellers, however, are the risks to personal health associated with breathing air that is high in $PM_{2.5}$. EPA.gov cites numerous scientific studies that link exposure to particulate matter with nonfatal heart attacks, irregular heartbeat, aggravated asthma, decreased lung function, increased coughing, difficulty breathing, and even premature death in people with heart or lung disease[8] . Kheirbek et al. estimates that $PM_{2.5}$ from on-road sources alone in New York City contribute to 320 deaths and 870 hospitalizations and ER visits per year, and Wylie et al. found that an increase in exposure to $PM_{2.5}$ among pregnant women in Dar es Salaam, Tanzania is linked with an increased risk of low birth weight, stillbirth, and prematurity[9,10].

Ground-Level Ozone

It may be confusing to see ozone (O_3) described as a pollutant. After all, isn't one of the reasons that it is important to reduce pollution to conserve the ozone layer that protects the earth from the sun's more harmful UV radiation? Well, yes, but the natural ozone layer is high in the upper atmosphere and does not pose any risk of being breathed in. When at ground-level, however, ozone is harmful to many organisms, including people in many of the same sensitive groups outlined for $PM\text{-}_{2.5}$. Ozone at ground-level is formed in the same way as most $PM_{2.5}$ in that it is not emitted directly from polluting sources but rather is formed in the environment from pollutants emitted from automobiles and industry[11].

Ground level ozone is hazardous in many of the same ways that $PM\text{-}_{2.5}$ is. Short term ozone exposure has been found to cause adverse respiratory effects that are more pronounced in those with asthma, emphysema, and chronic bronchitis, and long-term exposure to ozone is likely to increase risk of asthma development and even cause premature death. Ozone exposure can be particularly damaging to children, as their lungs are still developing and would receive a higher dose of ozone per body mass[12].

Ozone is also damaging to many plant and tree species native to the United States, leading to reduced photosynthesis and growth, and increased sensitivity to disease and harm from severe weather. As such, an overabundance of ground-level ozone can have long term adverse effects on an ecosystem, causing loss of species diversity and habitat quality[13]. To make matters worse, one of the effects of climate change is that it leads to more favorable conditions for ozone to form at ground level. The ozone then, being a greenhouse gas, contributes in turn to the changing climate – creating a feedback loop of poisonous air that is constantly worsening[12].

Why now?

We suspect that the lower rates of travel and emissions resulting from the COVID-19 lockdown may have led to an increase in air quality in Chicago. Zambrano-Monserrate et al. claim in their paper titled “Indirect effects of COVID-19 on the environment“ that there are both positive and negative associations from orders to stay at home, the positive of which includes improvement in air quality, clean beaches, and environmental noise reduction. Though they also cite that there are negative secondary aspects such as “reduction in recycling and the increase in waste, further endangering the contamination of physical spaces (water and land), in addition to air”, these effects are outside the scope of our study[14].

Given the clear effects that climate change is already having on the earth, it is important to understand through what mechanisms the planet’s health is deteriorating. While inexplicably tragic, the COVID-19 pandemic may allow us to learn a little bit more about humankind’s impact on the environment, and that opportunity should not be wasted. Poor air quality is something that, if allowed to worsen too much, could affect everyone, and is already impacting the lives of many. It is for these reasons that we set out to model the changing air quality at this unique point in time.

Methodology

For this study - we began using four different datasets, each containing their own information. Dataset one contained information relative to air quality and pollutant density in the air, dataset two contained information relative to flight data, dataset three contained information relative to bus congestion, and dataset four contained data relative to weather information.

Our dependent variable in this study is the daily Air Quality Index (or AQI from here on out) within the city of Chicago. Our predictors in our case were the pollutants daily value contributing to Chicago AQI, the number of flights going through Chicago (daily), the number of buses driving through congestion in Chicago (daily), and the stereotypical weather components/telemetry data contributing to daily weather within Chicago. Furthermore, we decided to add in another variable that is the data in which the stay at home order due to Covid-19 began.

- Air Quality Index within Chicago: Measures the average amount of pollutant within the air and is a numeric data type that could be classified as an integer. This has a range from 0 to 500, 0 representing great quality air, 500 representing extremely polluted air (Source: https://www.epa.gov/outdoor-air-quality-data/air-quality-index-daily-values-report) [15]

- Daily Weather Telemetry in Chicago: Measures the average amount of inches of precipitation occurring in Chicago, wind speed, and temperature, daily, and can be classified as a floats (Source: https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094846/detail)[16]

- Daily emissions of CO, SO_2, NO_2, Ozone, $PM_{2.5}$, PM_{10}: Measures the daily pollutant emissions value within Chicago, and is classified as an integers (Source: https://www.epa.gov/outdoor-air-quality-data/air-quality-index-daily-values-report) [17]

- Daily Flight Count: Measures the amount of flights that are arriving and departing from both O'Hare and Midway airports and is an integer (Source: https://www.transtats.bts.gov/ONTIME/Departures.aspx)[18]
- Daily Traffic/Bus Congestion: Measures the daily amount of traffic congestion within Chicago and is an integer (Source: https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/ef4k-dci7)[19]

- Stay-at-home Order: Is a binary yes or no field that is 'yes' for dates when the stay at home order was in action and 'no' when it was not.

Of this data provided above, we used a two and a half year lookback period, which consisted of the date range from January 1st 2018 to April 30th 2020. This gives us roughly ~900 records with eight columns of data.

Our first goal after collecting all four individual datasets for a specific time frame and merging them into a singular dataframe. The next step involved data cleaning, ensuring that all the dates lined up, there were minimal null values within the dataset, and the data types were all aligned and set properly.

Once this was done, we then imported the data into SAS and performed pre-processing steps including creating dummy variables and any necessary new variables from the date (standardization and normalization if necessary). We then do data exploration using scatter plots, descriptive statistics and visualizations. Transformations required for the response variable and predictor variables will also be determined at this step.

We then determined a final model that is adequate for fitting the data and understanding the relationship between the variables. Furthermore collinearity tests will be performed.

Data analysis within SAS includes the following:

- Fitting the full model
- Checking for assumptions and diagnostics/fixing issues
- Splitting data into training and testing sets
- Running model selection using a training set - We will fit multiple models and compare - same with variable selection techniques (stepwise, cp, etc.)
- Checking assumptions and diagnostics on the final model including analyzing residual plots and checking for outlier and influential points
- Examining the relationship and associations among the variables using the final model
- Checking the test performances using the test set
- Compare the performances
- Compute two predictions including the prediction intervals using the regression model
- Apply validation techniques to evaluate the predictive power of the final model

Finally, we will capture this information and cite impressions of the model and areas of improvement within the model.

Analysis, Results and Findings

Data Cleaning

With the variety of factors we wanted to explore regarding air pollution, it was clear that we were not going to find a single dataset that included all the variables. Furthermore, we had to make a decision regarding what we were going to use in order to measure our dependent variable (air pollution). The most useful data set we found was the EPA's site that keeps track of pollutant data[15]. This dataset has multiple benefits. First, it provides both recent and historical data. This is necessary because we not only wanted to be able to view the current data on air quality, but also needed to compare the recent data with past data. Second, the site breaks the data down into daily values. This provides us the opportunity to look at a single day's air quality index and determine the factors that influence this value, rather than, for instance, looking at an average air quality value for a month. Third, the site provides an overall AQI index along with other major pollutants.This allows us to include these other pollutants in our model. One of the issues with this data was that while the overall AQI values were up-to-date, CO, SO_2, PM_{10}, and NO_2 only had data values up until April 30th of this year. At the time, we decided to keep the date range of our analysis from January 1, 2018 up to June 30th, as we wanted to have our data as recent as possible. We wanted this recent data in order to have more days where the Stay-At-Home order was in effect so that we could better measure the effect of the order on air quality. So, we delayed the decision to modify our data range until more data came into play.

Once we decided upon this dataset, it provided the primary structure for our data. We created a spreadsheet in which each observation (row) would serve as a single day. The columns would be our dependent variable (the particular day's AQI index) and our independent variables.

While it was beneficial to discover this daily data, it also provided more structural challenges for the rest of our datasets. Not only did we need to find datasets where our desired variables were included, but we also needed these sets to produce a daily value with which we could fit into our primary dataset. If a dataset provided simply a summary number for a year or month, it was not beneficial, as we needed a dynamic daily value in order to measure how the variable affected the air quality index for the day.

Our next search for a variable with a daily dataset was weather. While sites that provide current weather data are abundant, it was a little more difficult to find historical weather data that was broken down as a daily value. Luckily, the NOAA keeps track of historical weather data from a variety of weather stations, many of which have the data broken down by day[16]. The most reliable (i.e. few missing values) weather station in the Chicago region was O'Hare airport, so we decided to use it as our weather station. The NOAA allows a user to choose which measures to include in a dataset and download data over a specified date range. We chose to include precipitation, average daily temperature, and average wind speed for our variables. Once we had that data in a set, it was simply a matter of copying and pasting the values in order to line up the dates in this new dataset with our primary set. After this was completed, we had a dataset that included (for each day) an overall AQI index, pollutant levels for CO, Ozone, SO_2, PM_{10}, $PM_{2.5}$, and NO2, and also Average Wind Speed, Precipitation, and Average Temperature.

This left us with our variables that we had decided to use in order to measure the mobility of people within the Chicago region. The first of these variables was air traffic. One might think that it wouldn't be too difficult to find a dataset that simply had the daily cumulative flights in and out of Chicago, but, alas, it was not that simple. For instance, at FlyChicago.com (the Chicago Department of Aviation's official website) the narrowest range for cumulative flight data is monthly. Our best dataset ended up being on the Bureau of Transportation Statistics site, which has daily flight data ranging from April of this year back to 1987[18]. Once again, we encounter a dataset with its most recent data being April 30th, 2020. Since we were already losing many of our pollutant variables after April 30th, with the addition of the Flight Count variable being lost, we decided to modify the range of our data observations. Instead of January 1st, 2018 until June 30th, 2020, our observations now end on April 30th, 2020.

The available flight data includes departure times, delays, taxi time, reason for delay, among others. Using the "Actual Departure Time" and "Actual Arrival Time" data for the set, we could produce one row of data for each flight in and out of the Chicago airports. Unfortunately, you cannot simply produce a list that gives these departure and arrival times. The data sets are separated by airline and airport. So, for each airport, we went through by airline and downloaded the arrival and departure flight data for each day, back to January 1, 2018. This produced 48 separate data files. To produce a daily flight count, we created a Python program (provided in the appendix) that iterated through the files and counted each time a particular date appeared in the data (as this would indicate either a departing or arriving flight). We created a dictionary to keep track of the counts. After the files were processed, the dictionary was written into our primary dataset. This gave us a column for Total Flight Data to add to our primary data set.

One more mobility measure that we wanted to add to our analysis was traffic congestion. An initial dataset that we thought would be useful was the City of Chicago's "Average Daily Car Count". It turned out that this data was not actually a daily count of Chicago traffic. Rather, it was simply an estimate based on a "car census" taken in 2006. Instead of relying on actual traffic counts, we found a dataset that estimated Chicago traffic congestion[19]. In order to estimate traffic congestion, the city uses its extensive bus network. Each time a bus measures congestion, it sends data to the city. As you can imagine, this creates a massive data set. We found ourselves with a spreadsheet containing millions of observations. In order to produce the data in workable form for our primary dataset, we conducted a similar process to the one that formatted our daily plane data. Each time a bus registered a traffic congestion issue, it added to our "congestion count" for the day (see R Code in appendix). This way, we had a variable that gave an estimate for the traffic congestion during each day in our dataset. In order to join this data with our existing dataset, we could not do a simple copy and paste. This traffic congestion data was missing a large number of observations (more on that below), so we had to use a formula to join the data by date. After this join, we had a dataset with nearly all the variables we wanted to add into our analysis.

The final variable to include with our dataset was a variable to indicate whether or not the day had a Stay-At-Home order in place. The governor of Illinois put a stay-at-home order in place on March 20th, 2020. In order to create a variable for this order, we simply created a column that had a "Yes" if the order was in place and a "No" if the order was not in effect. The result was a data column that simply had a series of "No" values from January 2018 until March 19th, 2020. After that date, the column became a "Yes" value for the rest of the observations in the dataset.

Our full dataset then consisted of the following:
Each day (from January 1, 2018 until April 30th, 2020) acting as an observation.

The following variables:

- Overall AQI Value
- The following daily pollutant values: CO, Ozone, SO_2, PM_{10},$PM_{2.5}$, and NO_2.
- Average Wind Speed (mph)
- Precipitation (in.)
- Average Temperature (°F)
- Total Flight Count
- Bus Count (Traffic Congestion)
- Stay-At-Home Order

Pre-processing

With our complete dataset, we were prepared to input the data into SAS for our analysis. We named our initial dataset "AQI" and used an infile method to input our dataset. All but two of our variables were numerical, making our input rather straightforward. Both the Date variable and stay at home variable are text values, rather than numerical. We don't need to worry about the date variable, as this is simply used to provide the organizing structure for our observations. The Stay-at-Home variable, on the other hand, is a variable that we want to include in our analysis so we needed to create a dummy variable for measurement. This variable is about as basic of a variable as one can find, as it is simply a "Yes" or "No" value. Since this data is already in binary form, for our SAS analysis it was simply a matter of creating a single dummy variable. We decided to have a "1" indicate the Stay at Home order was in effect on that day and "0" indicate that the order was not in effect.

After the data input, we did a "proc print" in order to check that our dataset had uploaded correctly. After performing some basic analysis on our AQI variable (histogram, boxplot, etc.), we ran a full regression model with all our variables. At this point, it became more clear that we had a significant number of missing values in our bus data. This is disconcerting for a couple reasons. The first reason this missing data is troubling is more from a statistical analysis standpoint. Without these observations (approximately 200) our dataset becomes too small for our analysis. We would have to exclude each of these observations, reducing our total number of observations to below 600. Furthermore, once we split our data into training and testing sets, the training set would be even more reduced. This significant reduction in the number of observations is only the first item that causes concern for our analysis.

The second concern comes from outside the statistical standpoint and stems from the purpose of conducting our analysis in the first place. The desire for this statistical analysis is to see the effects (if any) of reduced mobility on air quality during the COVID-19 pandemic. One of the primary air pollutants is vehicular traffic[24]. If we are able to see that the reduction of traffic congestion is correlated to the reduction of air quality, it strengthens the argument that the COVID-19 pandemic has created an environment in which a reduction in commuting has impacted with the air quality in the Chicago region. If we are unable to use the traffic congestion data, we have lost one of our three indicators for human mobility (the other two being Flight Counts and the Stay-At-Home Order).

With this in mind, we considered a few options that we could pursue to keep our bus data. Option number one was to move back our date range so that we could include more total observations. If we moved our start date to January 2017, for example, it would simply require us to expand our other data variables to include these dates and then put them in our dataset. This would hopefully allow us to have more bus data and the total number of observations would increase. The issue with this option is that the bus data gets even more sparse as dates move further into the past. 2017 is missing even more data than

2018, so the reliability of this bus data as an indicator for traffic congestion does not improve if we try to expand our date range.

A second option would be to try to calculate a value for the missing traffic data. While this is tempting, it would create a bit of a flaw in our data. Since we are relying on variables that are dynamic and change daily, if we were to simply calculate or estimate a value for traffic congestion on a particular day, it would not truly reflect the potential impact on the day’s air quality. For instance, it is possible that we use an estimated traffic count on a day that happened to have extremely bad traffic. If this same day had poor air quality and our estimate did not approximate the day’s actual traffic congestion, one might assume that the poor air quality was unrelated to the traffic data. If we do this estimated calculation with over 200 observations, our analysis would produce skewed results, as we would have to rely on data values that did not reflect the realities of each day in Chicago.

Since we were not satisfied with these options, we decided that it would be best to simply drop our bus variable from our analysis. Having over 200 observations that had either artificial values or were simply not calculated in our analysis were not options we wanted to move forward with in our analysis. While it is true that this traffic congestion variable is an important indicator of mobility in the Chicago area, we still have the Stay-at-Home Order variable and the Flight Count variable to use as measures of reduced mobility in the pandemic.

Once we decided to remove the bus count variable from our dataset, we went back in and looked for any more missing values that could impact our data. Over the entire dataset we now had only 70 missing values. These missing values were dispersed across observations and variables, with no single variable accounting for a large portion of the missing values. At this point, we were comfortable with moving forward in our data analysis.

Data Exploration

Descriptive analysis can be described as the summary of quantitative measures used to describe the variables being used within a model. This can be inclusive things such as:

- Mean and Standard Deviation
- 5 Point Summary (Min, Max, 25th Percentile, Median, 75th Percentile)
- Histogram Distributions
- Box Plots
- Correlation values

Of the 851 observations used, we initially began by analyzing the dependent variable, in our case, Chicago AQI. This returned:

- Average/Mean: 56
- Standard Deviation: 19.62
- 5 Point Summary:
    - Min - 23
    - 25th Percentile - 43
    - Median - 53
    - 75th Percentile - 64
    - Max - 177

Furthermore, after initial numeric analysis, we then proceeded to plot both the histogram of the Chicago AQI variable as well as the box plot of the Chicago AQI vs. that of the stay at home order variable. From the initial histogram of Chicago AQI, we saw that the distribution was an extremely right skewed and unimodal distribution *(figure 1a).* The box plot (*figure 1d*) shows that at the time the stay at home order began, there were much more condensed variations of AQI compared to the overall non-stay at home Chicago AQI. Distribution from the box plot for dates prior to the stay at home order showed that the overall min in comparison was much lower, and the overall max in comparison was much higher. The median, 25th and 75th percentiles, however, seemed to be more aligned with each other.

From this, we determined that we needed to transform the data in order to have a proper distribution. We decided to run a logarithmic transformation on the Chicago AQI variable to account for this skewness within the data. As can be seen by figure 2a within the appendix, by conducting a log transformation, our data distribution follows that of almost a perfect unimodal normal distribution which is exactly what we wanted.

We proceeded to analyze the Pearson correlation distribution between the independent variable and the newly transformed dependent variable. The table below is a representation of the values:

Correlation Between	***Correlation Value***
Chicago AQI - Avg Wind Speed	-0.39162
Chicago AQI - Precipitation	-0.06137
Chicago AQI - Avg Temperature	0.36354
Chicago AQI - CO	0.15096
Chicago AQI - Ozone	0.72518
Chicago AQI - SO2	0.25414
Chicago AQI - PM_{10}	0.35871
Chicago AQI - $PM_{2.5}$	0.67336
Chicago AQI - NO_2	0.34168
Chicago AQI - Total Flight Count	0.10426
Chicago AQI - Total Bus Count	-0.08391
Chicago AQI - Stay At Home	-0.05251

Full model

Based on observations made in the data cleaning and exploration stage, we fitted a full model excluding the bus count variable and applying logarithmic transformation on the response variable.

$$\begin{aligned}\boldsymbol{ln}(\boldsymbol{Chicago\ AQI}) &= \mathbf{2.95720} - \boldsymbol{Avg\ Wind\ Speed} * \mathbf{0.00135} - \boldsymbol{Precipitation} \\ &\quad * \mathbf{0.00232} \\ &- \boldsymbol{Avg\ Temp} * \mathbf{0.00176} - \boldsymbol{CO} * \mathbf{0.00479} + \boldsymbol{O3} * \mathbf{0.00771} + \boldsymbol{S02} * \mathbf{0.00068811} \\ &+ PM10 * \mathbf{0.00073330} + PM2.5 * \mathbf{0.01409} + \boldsymbol{NO2} * \mathbf{0.00215} \\ &- \boldsymbol{Total\ Flight\ Count} * \mathbf{0.00001760} - \boldsymbol{Stay} - \boldsymbol{at} - \boldsymbol{Home} * \mathbf{0.02537}\end{aligned}$$

The standardized estimates for the full model indicate that the top three most important predictors are $PM_{2.5}$, O_3 and Average Temperature, in order of importance. All three predictors are significant at $\alpha = 0.05$ with associated p-value of <0.0001. NO_2 is also significant with p-value of 0.0012. *(Figure 3a)*

Both total flight count and stay-at-home variables have a negative impact on the response variable. For the dummy variable stay-at-home, there isn’t a significant difference in air quality between before and after the stay-at-home order was put into place on March 21st as indicated by the parameter estimate of -0.02537 and standardized estimate of -0.01841*(Figure 3a)*.

There are few assumptions made when using linear regression to model the relationship between a response variable and it’s predictors. These assumptions make it possible to develop measures of reliability for the least squares line.

- Assumption of linearity assumes that the relationship between the response variable and a predictor is linear.
- Assumption of constant variance assumes that the variance of the probability distribution of the residual is constant for all settings of the predictors.
- Assumption of independence assumes that the error associated with one value of the response variable has no effect on the errors associated with other response values.
- Assumption of normality assumes the probability distribution of the residual is normal.

The analysis of the residuals is useful in detecting possible violations of these assumptions. If model assumptions hold, points should be randomly scattered inside a band centered around the mean of the residuals. We performed residual and normality analyses of residuals in SAS for analyzing these assumptions.

For the full model, after logarithmic transformation on the response variable, assumption of linearity is violated as indicated by the residual versus predictor plots and the scatterplot for precipitation and total flight count. *(Figure 3c - 3l and Figure 2e)*

The assumption of constant variance, however, does not seem to be violated since the residual against the predicted values plot on the full model does not show a pattern and the points within ± 3 standardized/studentized residual bands are randomly scattered *(Figure 3m)*. Independence and constant variance assumptions are generally related. If one is violated, then the other one is violated as well. We can also note few observations that are outside the ± 3 standardized/studentized residual band as possible outliers.

The assumption of normality for the logarithmic transformed model also seems to be satisfied as indicated by the normal probability plot of the residuals *(Figure 4n)* which shows a near straight line. This re-affirms the decision made to transform the response variable at the data exploration step.

Various diagnostic techniques exist for checking the validity of these assumptions, and these diagnostics suggest remedies to be applied when the assumptions appear to be invalid. Consequently, we applied these diagnostic tools in every regression analysis.

a. Multicollinearity: This is defined when two or more of the predictors in a regression model are moderately or strongly correlated with each other (not with the dependent variable). Another way to think of this is that one predictor can be used to calculate another predictor. There are two types of multicollinearity, one being perfect multicollinearity which is when two or more variables are perfectly correlated, and near perfect multicollinearity which exists when two variables are highly correlated (much more common in this instance). This can be analyzed by looking at the Pearson correlation between specific dependent variables or the VIF associated with the model and the predictor. Causes of multicollinearity include insufficient data, dummy variables being incorrectly used, including two identical variables, and using a variable that is actually a combination of other variables. The best ways to fix this issue are by either adding more data of dropping on or more of the variables.

b. Outliers and influential points: These two types of points go hand in hand. An outlier can be defined as an observation within the data that is different from the majority of cases and cannot be explained by the model itself. An influential point can be defined as an observation that has excessive influence on the fit of the model. These observations are often valid in natural variation, however, they can also arise from errors inclusive of coding errors, entry errors, and misspecification model errors. Once influential points and outliers have been found, the analyst must verify the data and determine if there is such a drastic effect on the model from these points that they need to be removed.

c. Goodness-of-fit: This metric measures the goodness of fit of the data within the model produced. This is done through a specific type analysis by looking at the "Null Hypothesis", which states that there is no relation between the dependent variable and any of the independent variables. If the overall F statistic is high and the associated P value is <.0001, we are able to reject the null hypothesis and state that the independent variables have significance and contribute to the dependent variable.

Using SAS, we were able to analyze all three of these metrics within our model and dataset, and using our combined knowledge and experience, determined what to remove and keep within the data set.

After conducting an analysis on the studentized residuals and the cook’s distance of the data, we were able to remove the following points as they were both outliers and influential points that were drastic enough to affect the overall model.

- Points: 72, 73, 82, 108, 155, 183, 188, 226, 289, 290, 291, 404, 563

Once the influential and outlier-based observations were removed, we proceeded to analyze the multicollinearity within the model and if there was any occurrence. In order to analyze this, we produced two different metrics to analyze, the Pearson correlation graph, and a plot wise scatter. A Pearson correlation analysis produces a table that contains the correlation value between every variable. This number produced is statistically significant in the sense that the closer to zero this calculation is, the less the variables relate. If this correlation calculation is equal to positive one then we can say that there is a strong positive correlation (as one goes up the other goes up), and if it's closer to negative one, we can say there is a strong negative correlation (as one goes up the other goes down). The scatter plots are used to visualize the correlation, two variables that have strong correlation will have a linear angled plotted result, whereas those that do not have strong correlation will be randomly dispersed, graphically.

After conducting these two metrics, we discovered interesting results. We found that there was much higher correlation between the AQI and the Ozone and $PM_{2.5}$/PM_{10} variables. This is ok as these are between the dependent variable and the predictors, however we then looked at the relationship between these two variables. To our surprise there was little collinearity between these two variables, however, as referenced within the introduction, the AQI is calculated by taking the max of all the individual pollutant gasses contributing to it, meaning that Ozone and $PM_{2.5}$/PM_{10} were producing the maxes of almost every occurrence. Furthermore, we analyzed the VIF of the variables contributing to the complete model and discovered that there were no values above ten which is the threshold that states there is multicollinearity occurring for that variable. We proceeded to remove these three variables as we believed this would have ignored the other variables and overfit the model. We referred to this as structural multicollinearity where our experience and knowledge override the statistical insignificance between the two independent variables.

Finally, when using the full model, we analyzed the P score and F value associated with the model and were able to conclude that we can reject the null hypothesis that none of the variables contribute to the dependent variable. This can be seen below:

$$\beta_0 = 0$$

$$: None\ of\ the\ predictors\ included\ in\ the\ model\ have\ an\ association\ to\ the\ y\ variable.$$

$$\beta_0 \neq 0$$

$$: At\ least\ one\ of\ the\ predictors\ included\ in\ the\ model\ has\ an\ association\ to\ the\ y\ variable.$$

F-value: 68.11, P-score : < 0.0001

Conclusion: Since the P-score is less than 0.05 (meaning the model has significant contributors), we can reject the null hypothesis and conclude that there is at least one predictor that has a significant effect on the dependent variable.

Overall, when doing this analysis, we ultimately decided to remove all the influential and outlier points occurring within the model as they contributed and swayed the final model more than an insignificant amount. Furthermore, we removed pollutant values as there was too high a correlation between certain values that would have overfit the model as they are a direct relation between each other. This included removing Ozone, $PM_{2.5}$ and the PM_{10} variables, leaving the only pollutants contributing being CO and NO_2, otherwise known as the two main human contributing pollutants.

Reduced Model

From the adjusted R^2 of the full transformed model we can conclude that 87.89% of the variation in AQI is explained by all predictors in the full model which seemed plausible since AQI is already a function of the pollutant variables. However, since the purpose of our study is to analyze the impact of reduced mobility on air quality, it might not be relevant to include the pollutant variables that would not explain the direct impact of reduced mobility. Based on the study of outdoor air pollutants such as Nitrogen Dioxide, Sulfur Dioxide, and Carbon Monoxide, these are the main pollutants that not only are emitted by humans, but also cause the most harm to our health[

Refitting the full model with variables that are relevant for our study while also removing observations flagged as outliers and influential points, we get the reduced model below.

$$\begin{aligned}\boldsymbol{ln(Chicago\ AQI)} &= \boldsymbol{3.85030 - Avg\ Wind\ Speed * 0.01976 - Precipitation} \\ &\quad \boldsymbol{* 0.02982} \\ &\boldsymbol{- Avg\ Temp * 0.00650 - CO * 0.00958 + NO2 * 0.00852} \\ &\boldsymbol{- Total\ Flight\ Count * 0.00014003 - Stay-at-Home * 0.13329}\end{aligned}$$

At α=0.05, total flight count and stay-at-home predictors are now significant with precipitation being the only insignificant variable. Removing the pollutants that are not directly caused by human mobility will isolate the variation caused by the variables left in the model. Consequently, the new adjusted R^2 was reduced to 34.97%. This implies that the variables we identified as being directly linked to human mobility do not do well in describing the variation in AQI for Chicago.

In the reduced model we still observe linearity assumption violation in the regression model. *(Figure 4h - Figure 4j*). All parameters do not have VIF above 10 and hence collinearity still does not seem to be a problem. We flagged four new observations as outliers and influential points which were not removed. *(Figure 4f - Figure 4g)*. The reduced model has F-value of 62.92 with associated p-value of <0.0001. At α=0.05, we can reject the null hypothesis and conclude there is at least one predictor in the reduced model that is still significantly associated with AQI.

To test how well the final model will perform, we randomly split 75% of the data set for training the model and 25% for testing the model. This will enable us to assess the model's performance using unseen data. We used SAS to randomly split our data with 630 observations for training and 210 observations for testing, which is sufficient for making inference about how well the model would perform. *(Figure 5)*

The goal of model selection is to collect the variables that best contribute to the model itself. The variables used are known as explanatory variables, independent variables, and predictors, which are used to calculate the dependent/outcome variable. We used four different methods of analysis for model selection:

- Stepwise
- Backward Elimination
- Adjusted R^2 Selection
- Mallows' CP Selection

Each of which has its own unique method of analyzing the variables contributing the model and determining the best fit data. Luckily, SAS has the capabilities to autonomously cover each of these methods. We were able to run all four selection options on our dataset relative to AQI in Chicago; a summary of what each method does can be seen below as well as what the output conveyed relative to our model.

STEPWISE: This model selection method focuses an iteration of forward selection by adding a significant predictor variable to a model if the P score associated with the variable is above a threshold (in our case, α = 0.05) as well as if there is a high correlation coefficient. Once this variable is added, this method looks at the variables within the model and deletes any that is not statistically significant - once this check is done, another variable is added into the model and the process terminates when none of the variables outside of the model are considered significant

Of the seven independent variables used within the model to determine Chicago AQI (those being average wind speed, precipitation average, average temperature, CO emission, NO_2 emission, total flight count, and stay at home order), by using the stepwise function, the model determined that average wind speed, average temperature, NO_2 emission, and daily flight count were significant relative to the dependent variable AQI. This was completed in four steps and produced the following regression equation:

$$\begin{aligned}\boldsymbol{ln(Chicago\ AQI)} &\\ &= \mathbf{3.75 - \mathit{Avg\ Wind\ Speed} * .0195 - \mathit{Avg\ Temp} * .006 + \mathit{NO2} * .0095} \\ &\qquad \mathbf{- \mathit{Flight\ Count} * .000087}\end{aligned}$$

With R^2 value of .3467.

BACKWARDS: This model selection begins with all independent variables within the model and moves backwards, removing one variable at a time dependent on significance. Each iteration, the selection method removes a variable from the model if it is not statistically significant (in our case, a = 0.05). At each step, the variable with the smallest contribution to the dependent variable is deleted, stopping only when there is no further improvement that can be made to the model.

Of the seven independent variables used within the model to determine Chicago AQI (those being average wind speed, precipitation average, average temperature, CO emission, NO_2 emission, total flight count, and stay at home order), by using the backwards selection function, the model determined that average wind speed, average temperature, NO_2 emission, and daily flight count were significant relative to the dependent variable AQI. This was completed in three steps and produced the following regression equation:

$$\begin{aligned}\boldsymbol{ln(Chicago\ AQI)} &\\ &= \mathbf{3.75 - \mathit{Avg\ Wind\ Speed} * .0195 - \mathit{Avg\ Temp} * .006 + \mathit{NO2} * .0095} \\ &\qquad \mathbf{- \mathit{Flight\ Count} * .000087}\end{aligned}$$

With R^2 value of .3467.

MALLOWS' CP: This model selection looks at CP statistics for selecting the most ideal model. This selection runs every combination of independent variables and produces the most ideal combination of significant variables along with an associated value. The best combination of variables is associated with a CP value that roughly measures the value of P, where P is equal to the number of variables plus one.

Of the seven independent variables used within the model to determine Chicago AQI (those being average wind speed, precipitation average, average temperature, CO emission, NO_2 emission, total flight count, and stay at home order), by using the mallows' CP selection function, the model determined that all seven of these variables are to be used as it has a CP value of 8.0 and the P variable is equal to seven variables plus one, meaning CP = P, validating this variable and model selection. This produced the following regression equation:

$$\boldsymbol{ln(Chicago\ AQI) = 3.85 - Avg\ Wind\ Speed * .0197 - Precipitation * -.0298 + Avg\ Temp * .006 + CO * .0095 + NO2 * .0095 - Flight\ Count * .000087 - SAH * .133}$$

With R^2 value of .3554

ADJUSTED R^2: Similar to MALLOWS' CP, this model selection creates iterations of every combination of the model and produces statistics associated with it. The main difference between these model selection methods is that instead of calculating the CP value, this method calculated the adjusted R2 and R2 value associated with each combination. The best combination of variables is associated with the highest adjusted R^2 and R^2 values.

Of the seven independent variables used within the model to determine Chicago AQI (those being average wind speed, precipitation average, average temperature, CO emission, NO_2 emission, total flight count, and stay at home order), by using the adjusted R^2 selection function, the model determined that all seven of these variables are to be used as it has the highest R^2 and adjusted R^2 values. This produced the following regression equation:

$$\boldsymbol{ln(Chicago\ AQI) = 3.85 - Avg\ Wind\ Speed * .0197 - Precipitation * -.0298 + Avg\ Temp * .006 + CO * .0095 + NO2 * .0095 - Flight\ Count * .000087 - SAH * .133}$$

With R^2 value of .3556 and an Adjusted R^2 value of .3492.

After running various variable selection methods, we were left with the choice between several possible candidates for the final model. The first, which was output by the Adjusted R^2 and Mallow's CP selection methods, included all 7 predictors. The second candidate for the final model, which included the 4 predictors average wind speed, average temperature, NO_2 emission, and daily flight count. Since the difference in Adjusted R^2 is negligible when the 3 additional variables are removed from the first model, we decided to remove the first model from consideration, leaving us with the 4-predictor model. However, the impetus for this paper was to analyze the effect of the COVID-19 lockdown on Chicago's AQI, and while this model does take flight counts into consideration, it doesn't account for ground traffic in any way. So we proposed a third model, which includes the same predictors as the 4 predictor model, but also included the stay at home order logistical predictor. These are the two models we set out to validate side-by-side in order to determine the final model. Because the model that includes the stay at home order is our primary interest, this model was named "Model 1" and the model with only 4 predictors is "Model 2". The training dataset generated in a previous step and containing 630 observations was used to train and generate performance statistics for both models. See below for the equations for both possible final models, as well as a table comparing both models' diagnostics for training and testing data.

Model 1:

$$ln(Chicago\ AQI)\ =\ 3.86043\ -\ Avg\ Wind\ Speed\ *\ 0.01951\ +\ Avg\ Temperature\ *\ 0.00641\ +NO2\ *\ 0.00930\ -\ Flight\ Count\ *\ 0.000136\ -\ SAH\ *\ 0.10646$$

Model 2:

$$\boldsymbol{ln(Chicago\ AQI)\ =\ 3.77284\ -\ Avg\ Wind\ Speed\ *.01963-\ Avg\ Temp\ *.00618\ +\ NO2\ *.00926\ -Flight\ Count\ *.00009307}$$

Train	Model 1	Model 2
RMSE	**0.25083**	0.25137
R^2	**0.3534**	0.3496
Adj-R^2	**0.3482**	0.3454
GOF	OK	OK
Residuals	OK	OK
Test		
RMSE	**0.25149**	0.25352
MAE	**0.20215**	0.20535
R^2	0.3494	**0.3388**
Adj-R^2	**0.3335**	0.3259

CV-R^2	**0.004**	0.0108

Training and testing diagnostics for Model 1 vs. Model 2

Adj-R^2 for the test data was calculated manually using the formula,

$$Adj.R^2 = \frac{(1-R^2)(n-1)}{n-p-1}$$

where *n* = 210 observations for both models and *p* = 5 for Model 1 and *p* = 4 for Model 2. It can be seen from the above table that Model 1 clearly outperforms Model 2 both in terms of fit statistics for the training data and ability to accurately predict the AQI value when tested with unseen data. This is good, as now we have validation past simple "business needs" for including the stay at home order predictor as we know that it improves the model when included.

Final Model

$$ln(Chicago\ AQI)\ =\ 3.86043\ -\ Avg\ Wind\ Speed\ *\ 0.01951\ +\ Avg\ Temperature\ *\ 0.00641\ +NO2\ *\ 0.00930\ -\ Flight\ Count\ *\ 0.000136\ -\ SAH\ *\ 0.10646$$

The stay-at-home variable has a p-value of 0.055, which is insignificant at $\alpha=0.05$. We, however, decided to keep this variable in our final model since it would be significant at $\alpha=0.1$ and performance metrics above indicate that the model that includes the stay-at-home variable performs better than the model that does not include this variable *(Figure 36-37)*. The standard estimates for the final model show that weather variables and NO_2 are still the most important predictors in the final model.

The final model has adjusted R^2 of 34.82%. This implies that there is no improvement in describing the variation in AQI in the final model versus the reduced model where the variables we identified as being directly linked to human mobility. We still see that human mobility does not do well in describing the variation in AQI for Chicago.

The final model satisfies all regression model assumptions of linearity, constant variance, independence and normality. The residual plots show points that are randomly scattered inside a band centered around the mean of the residuals and the normal probability plot still shows a linear line. *(Figure 9b - Figure 9g)*. Collinearity still does not seem to be a problem as described by the VIF statistics for each predictor in the final model *(Figure 9a).* We flagged two observations as outliers and influential points which, however, were not removed from the final model. *(Figure 9h)*. The final model has F-value of 68.11 with associated p-value of <0.0001. At $\alpha=0.05$, we can reject the null hypothesis and conclude there is at least one predictor in the final model that is still significantly associated with AQI.

The effect of each individual predictor was calculated using the following equation,

$$\% \, Change\ in\ Overall\ AQI = (e^{\beta_x} - 1) * 100$$

where $\beta_x$equals the parameter estimate for the predictor in question. According to our final model, the specific effect of each predictor (with all others held constant) on the overall AQI are as follows. An increase in average wind speed by 1 mph will reduce the overall AQI by 1.97%. An increase in temperature by 1℉ will increase the overall AQI by 0.64%. An increase in measured NO_2 by 1 will increase the overall AQI by 0.93%. An increase in total flight count by 1 will increase the overall AQI by 0.01%. Finally, and most importantly, having a stay at home order in place can be expected to reduce the overall AQI by 11.23%.

In Summary, the variables in our full model do well in describing the variation in AQI. When we reduce our model to variables that have direct human influence, our model does not do as well in describing the variation in Chicago's AQI though still showing to be significant predictors. There are possible factors that could explain this. Our data set was limited to Chicago. As discussed in the Chicago Tribune newspaper titled "Many cities around the globe saw cleaner air after being shut down for COVID-19. But not Chicago.", Chicago's situation is unique. During the stay-at-home orders, the article suggests that diesel emissions have not slowed down in Chicago which are likely the biggest pollution contributors to the city's air quality. Further work will need to be done to better understand the impacts of reduced mobility on air quality in Chicago and across the world.

Computing Predictions

The final model includes the 5 predictors of average wind speed, average temperature, NO_2 emission, daily flight count, and stay at home order. To test our model, we created two hypothetical observations for which our model could predict the day's overall AQI.

Hypothetical Observation #1 is a 40℉ day with average wind speeds of 9 mph. NO_2 levels are at 27, daily flight count is 1900 and stay at home order is in place. Our model predicts that ln(AQI) for this hypothetical day is 3.8277 with a 95% Confidence Interval of (3.7308, 3.9245) and a 95% Prediction Interval of (3.356, 4.3297). Because the dependent variable was log transformed, these values will need to be re-transformed in order to make sense. Taking the antilog of the above values and rounding to the nearest integer gives a predicted overall AQI of 46, with a 95% Confidence Interval of (42, 51) and a 95% Prediction Interval of (28, 76).

For Hypothetical Observation #2, there is no stay-at-home order in place. Average temperature is 19°F and average wind speed is 13 mph. NO_2 levels are at 47, and flight count is 2200. For this hypothetical, the predicted value for ln(AQI) is 3.8666 with 95% Confidence Interval of (3.8168, 3.9165) and 95% Prediction Interval of (3.3715, 4.3617). Taking the antilog of these values and rounding to the nearest integer gives us a predicted overall AQI of 48, with a 95% Confidence Interval of (45, 50) and a 95% Prediction Interval of (29, 78).

Future Work

There are few additional avenues worth exploring based on what we've discovered in our analysis, which are inclusive of the effect of covid-19 on the environment and in turn the effect on one's health. Furthermore, there are references of the impact of Covid-19 on solar power and the effect it had on the efficiency of them due to lower pollutant emission, as well as references on how Covid-19 has actually increased the overall amount of pollutants being emitted.

We were only able to gather data up to April 30th, 2020 even though the stay-at-home order was not lifted until late May of 2020. More observations in the stay-at-home category should provide a more balanced split and can help in more accurately measuring the impact of mobility on air quality for further analysis.

We can also make the case to include data past May as although the stay-at-home order has been lifted, people are not moving around like normal. If people's lifestyles change, there could be long term effects on air quality and ultimately climate change. From the article "Coronavirus Covid-19 Remote Work from Home Office Re-opening" [22] estimates that when the pandemic is over, 30 percent of the entire workforce will work from home at least a couple times a week. Before the pandemic, that number was in the low single digits. Further study should be done to analyze what the impact of increased work-from-home opportunities would be long term beyond the stay-at-home orders.

Another additional avenue worth exploring is collecting a more accurate indicator of mobility, such as google maps data. This will enable us to identify true variations in mobility that are caused by stay-at-home orders, increased work-from-home or any other future impacts on mobility.

This study should be expanded to include other major cities across the world. Many of these cities have seen air quality improvements during the COVID-19 crisis, except Chicago. According to a Chicago Tribune's analysis of state monitoring data, average soot concentrations in Chicago have not seen a significant change recently. Likely culprits being diesel emission in the city that have been a chronic problem for Chicago. More can be learned by expanding data bounds[21]. According to the article 'Will Covid-19 have a lasting impact on the environment?'[24]. 'the industries, transport networks and businesses have closed down, it has brought a sudden drop in carbon emissions. Compared with this time last year, levels of pollution of one of the big cities of the USA that is New York pollution is reduced by nearly 50% because of certain measures to contain the virus. Kimberly Nicholas, a sustainability science researcher at Lund University in Sweden, is the different reasons that emissions have dropped. Take transport, for example, which makes up 23% of global carbon emissions. These emissions have fallen in the short term in countries where public health measures, such as keeping people in their homes, have cut unnecessary travel. Driving and aviation are key contributors to emissions from transport, contributing 72% and 11% of the transport sector's greenhouse gas emissions respectively.'

High levels of air pollution according to the EPA **Air Quality Index** directly affect people with asthma and other types of lung or heart disease. The size of particles is directly linked to their potential for causing health problems. Small particles (known as PM2.5 or fine particulate matter) pose the greatest problems because they bypass the body's natural defenses and can get deep into your lungs and potentially your bloodstream. Exposure to such particles can affect both your lungs and your heart. Another future research topic that would be interesting to look at is analyzing health using AQI as a predictor variable and indirectly measuring the impact of mobility on overall health.

References

1) Cars, Trucks, Buses and Air Pollution. (2008). Retrieved July 11, 2020, from https://www.ucsusa.org/resources/cars-trucks-buses-and-air-pollution

2) Strosnider, H., Kennedy, C., Monti, M., &amp; Yip, F. (2017). Rural and Urban Differences in Air Quality, 2008–2012, and Community Drinking Water Quality, 2010–2015 — United States. MMWR. Surveillance Summaries, 66(13), 1-10. doi:10.15585/mmwr.ss6613a1

3) Air Quality Index Reporting, Volume 64, Number 149 40 CFR 58 § I. Background (1999).

4) Wayland, M. (2014). Air quality index: A guide to air quality and your health (p. 3) (United States, Environmental Protection Agency, Office of Air Quality Planning and Standards). Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Outreach and Information Division.

5) About Air Data Reports. (2018, September 21). Retrieved July 11, 2020, from https://www.epa.gov/outdoor-air-quality-data/about-air-data-reports

6) Particulate Matter (PM) Basics. (2018, November 14). Retrieved July 11, 2020, from https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

7) Malm, W. C. (2000). Transport and Transformation of Atmospheric Particulates and Gases Affecting Visibility. In *Introduction to visibility* (p. 24). Fort Collins, CO: Cooperative Institute for Research in the Atmosphere, NPS Visibility Program, Colorado State University.

8) Health and Environmental Effects of Particulate Matter (PM). (2020, April 13). Retrieved July 14, 2020, from https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm

9) Kheirbek, I., Haney, J., Douglas, S., Ito, K., &amp; Matte, T. (2016). The contribution of motor vehicle emissions to ambient fine particulate matter public health impacts in New York City: A health burden assessment. Environmental Health, 15(1). doi:10.1186/s12940-016-0172-6

10) Wylie, B. J., Matechi, E., Kishashu, Y., Fawzi, W., Premji, Z., Coull, B. A., . . . Roberts, D. J. (2017). Placental Pathology Associated with Household Air Pollution in a Cohort of Pregnant Women from Dar es Salaam, Tanzania. Environmental Health Perspectives, 125(1), 134-140. doi:10.1289/ehp256

11) Ground-level Ozone Basics. (2018, October 31). Retrieved July 13, 2020, from https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics

12) Zhang, J. (., Wei, Y., &amp; Fang, Z. (2019). Ozone Pollution: A Major Health Hazard Worldwide. Frontiers in Immunology, 10. doi:10.3389/fimmu.2019.02518

13) Ecosystem Effects of Ozone Pollution. (2017, February 27). Retrieved July 11, 2020, from https://www.epa.gov/ground-level-ozone-pollution/ecosystem-effects-ozone-pollution

14) Zambrano-Monserrate, M. A., Ruano, M. A., &amp; Sanchez-Alcalde, L. (2020). Indirect effects of COVID-19 on the environment. Science of The Total Environment, 728, 138813. doi:10.1016/j.scitotenv.2020.138813

15) Air Quality Index Daily Values Report. (2018, November 26). Retrieved July, 2020, from https://www.epa.gov/outdoor-air-quality-data/air-quality-index-daily-values-report

16) National Centers for Environmental Information (NCEI). (n.d.). Daily Summaries Station Details. Retrieved July, 2020, from https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094846/detail

17) Air Quality Index Daily Values Report. (2018, November 26). Retrieved July, 2020, from https://www.epa.gov/outdoor-air-quality-data/air-quality-index-daily-values-report

18) Detailed Statistics Departures. (n.d.). Retrieved July, 2020, from https://www.transtats.bts.gov/ONTIME/Departures.aspx

19) Chicago Traffic Tracker - Historical Congestion Estimates by Segment - 2018-Current - Dashboard: City of Chicago: Data Portal. (n.d.). Retrieved July 14, 2020, from https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/ef4k-dci7

20) Illinois News. (n.d.). Retrieved July, 2020, from https://www2.illinois.gov/Pages/news-item.aspx?ReleaseID=21288

21) Hawthorne, M. (2020, May 14). Many cities around the globe saw cleaner air after being shut down for COVID-19. But not Chicago. Retrieved July, 2020, from https://www.chicagotribune.com/news/environment/ct-met-covid-chicago-air-quality-20200514-rqam273qqfbmnfsyn3vzm4f45e-story.html

22) Molla, R. (2020, May 21). Office work will never be the same. Retrieved July, 2020, from https://www.vox.com/recode/2020/5/21/21234242/coronavirus-covid-19-remote-work-from-home-office-reopening

23) U.S. Air Quality Continues to Improve. (2019, July 31). Retrieved July, 2020, from https://www.instituteforenergyresearch.org/regulation/u-s-air-quality-continues-to-improve/

24) Will Covid-19 have a lasting impact on the environment? (2020, March 27). Retrieved July, 2020, from https://www.bbc.com/future/article/20200326-covid-19-the-impact-of-coronavirus-on-the-environment

25) Outdoor Air Pollution: Nitrogen Dioxide, Sulfur Dioxide, and Carbon Monoxide Health Effects. (2007, April 01). Retrieved July, 2020, from https://www.sciencedirect.com/science/article/abs/pii/S0002962915325933

Appendix

Descriptive Statistics before transformations

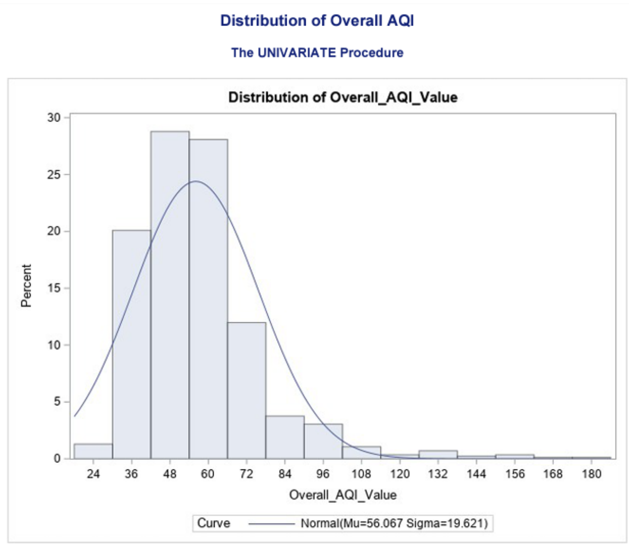


Figure 1a

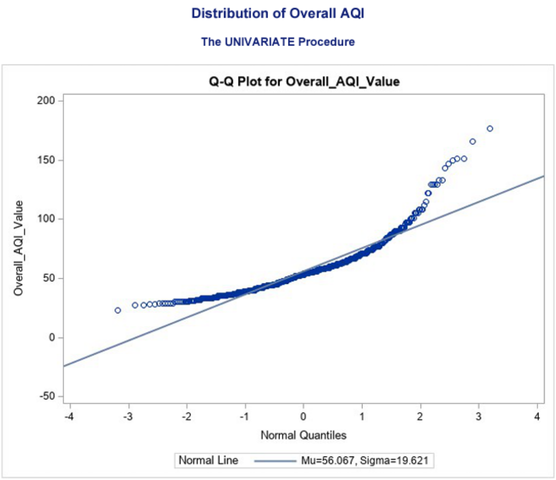


Figure 1b

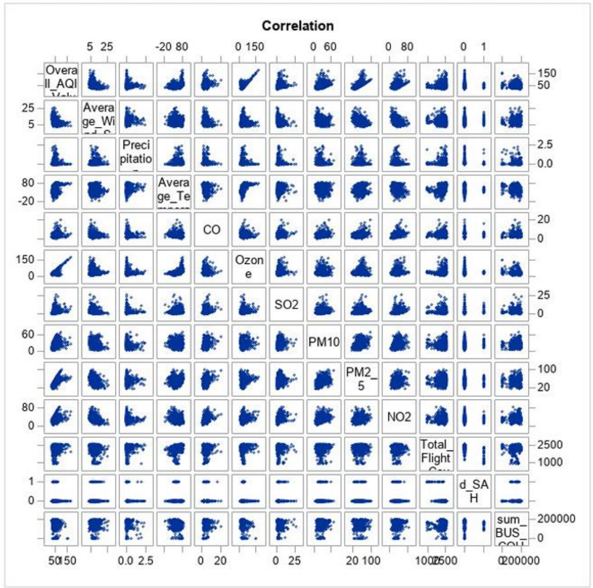


Figure 1c

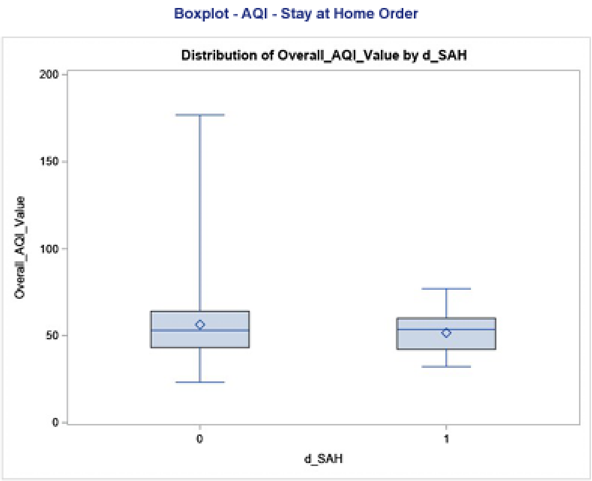
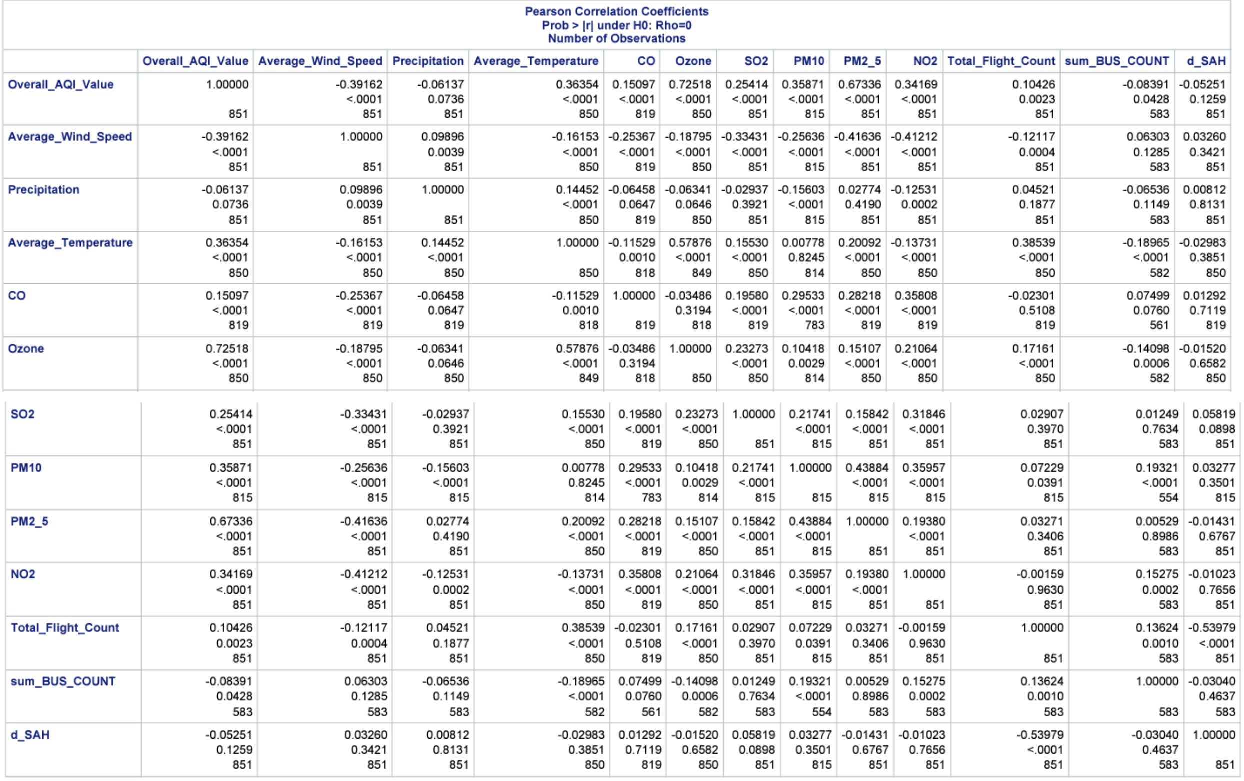


Figure 1d



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

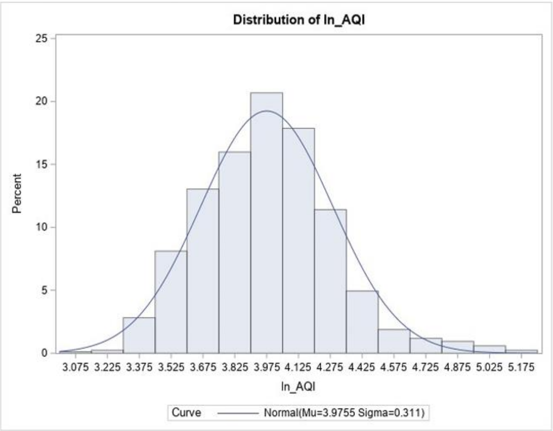
	Overall_AQI_Value	Average_Wind_Speed	Precipitation	Average_Temperature	CO	Ozone	SO2	PM10	PM2_5	NO2	Total_Flight_Count	sum_BUS_COUNT	d_SAH
Overall_AQI_Value	1.00000 851	-0.39162 <.0001 851	-0.06137 0.0736 851	0.36354 <.0001 850	0.15097 <.0001 819	0.72518 <.0001 850	0.25414 <.0001 851	0.35871 <.0001 815	0.67336 <.0001 851	0.34169 <.0001 851	0.10426 0.0023 851	-0.08391 0.0428 583	-0.05251 0.1259 851
Average_Wind_Speed	-0.39162 <.0001 851	1.00000 851	0.09896 0.0039 851	-0.16153 <.0001 850	-0.25367 <.0001 819	-0.18795 <.0001 850	-0.33431 <.0001 851	-0.25636 <.0001 815	-0.41636 <.0001 851	-0.41212 <.0001 851	-0.12117 0.0004 851	0.06303 0.1285 583	0.03260 0.3421 851
Precipitation	-0.06137 0.0736 851	0.09896 0.0039 851	1.00000 851	0.14452 <.0001 850	-0.06458 0.0647 819	-0.06341 0.0646 850	-0.02937 0.3921 851	-0.15603 <.0001 815	0.02774 0.4190 851	-0.12531 0.0002 851	0.04521 0.1877 851	-0.06536 0.1149 583	0.00812 0.8131 851
Average_Temperature	0.36354 <.0001 850	-0.16153 <.0001 850	0.14452 <.0001 850	1.00000 850	-0.11529 0.0010 818	0.57876 <.0001 849	0.15530 <.0001 850	0.00778 0.8245 814	0.20092 <.0001 850	-0.13731 <.0001 850	0.38539 <.0001 850	-0.18965 <.0001 582	-0.02983 0.3851 850
CO	0.15097 <.0001 819	-0.25367 <.0001 819	-0.06458 0.0647 819	-0.11529 0.0010 818	1.00000 819	-0.03486 0.3194 818	0.19580 <.0001 819	0.29533 <.0001 783	0.28218 <.0001 819	0.35808 <.0001 819	-0.02301 0.5108 819	0.07499 0.0760 561	0.01292 0.7119 819
Ozone	0.72518 <.0001 850	-0.18795 <.0001 850	-0.06341 0.0646 850	0.57876 <.0001 849	-0.03486 0.3194 818	1.00000 850	0.23273 <.0001 850	0.10418 0.0029 814	0.15107 <.0001 850	0.21064 <.0001 850	0.17161 <.0001 850	-0.14098 0.0006 582	-0.01520 0.6582 850
SO2	0.25414 <.0001 851	-0.33431 <.0001 851	-0.02937 0.3921 851	0.15530 <.0001 850	0.19580 <.0001 819	0.23273 <.0001 850	1.00000 851	0.21741 <.0001 815	0.15842 <.0001 851	0.31846 <.0001 851	0.02907 0.3970 851	0.01249 0.7634 583	0.05819 0.0898 851
PM10	0.35871 <.0001 815	-0.25636 <.0001 815	-0.15603 <.0001 815	0.00778 0.8245 814	0.29533 <.0001 783	0.10418 0.0029 814	0.21741 <.0001 815	1.00000 815	0.43884 <.0001 815	0.35957 <.0001 815	0.07229 0.0391 815	0.19321 <.0001 554	0.03277 0.3501 815
PM2_5	0.67336 <.0001 851	-0.41636 <.0001 851	0.02774 0.4190 851	0.20092 <.0001 850	0.28218 <.0001 819	0.15107 <.0001 850	0.15842 <.0001 851	0.43884 <.0001 815	1.00000 851	0.19380 <.0001 851	0.03271 0.3406 851	0.00529 0.8986 583	-0.01431 0.6767 851
NO2	0.34169 <.0001 851	-0.41212 <.0001 851	-0.12531 0.0002 851	-0.13731 <.0001 850	0.35808 <.0001 819	0.21064 <.0001 850	0.31846 <.0001 851	0.35957 <.0001 815	0.19380 <.0001 851	1.00000 851	-0.00159 0.9630 851	0.15275 0.0002 583	-0.01023 0.7656 851
Total_Flight_Count	0.10426 0.0023 851	-0.12117 0.0004 851	0.04521 0.1877 851	0.38539 <.0001 850	-0.02301 0.5108 819	0.17161 <.0001 850	0.02907 0.3970 851	0.07229 0.0391 815	0.03271 0.3406 851	-0.00159 0.9630 851	1.00000 851	0.13624 0.0010 583	-0.53979 <.0001 851
sum_BUS_COUNT	-0.08391 0.0428 583	0.06303 0.1285 583	-0.06536 0.1149 583	-0.18965 <.0001 582	0.07499 0.0760 561	-0.14098 0.0006 582	0.01249 0.7634 583	0.19321 <.0001 554	0.00529 0.8986 583	0.15275 0.0002 583	0.13624 0.0010 583	1.00000 583	-0.03040 0.4637 583
d_SAH	-0.05251 0.1259 851	0.03260 0.3421 851	0.00812 0.8131 851	-0.02983 0.3851 850	0.01292 0.7119 819	-0.01520 0.6582 850	0.05819 0.0898 851	0.03277 0.3501 815	-0.01431 0.6767 851	-0.01023 0.7656 851	-0.53979 <.0001 851	-0.03040 0.4637 583	1.00000 851

Figure 1e

Descriptive Statistics after logarithmic transformation

Distribution of AQI (log)

The UNIVARIATE Procedure



Distribution of AQI (log)

The UNIVARIATE Procedure

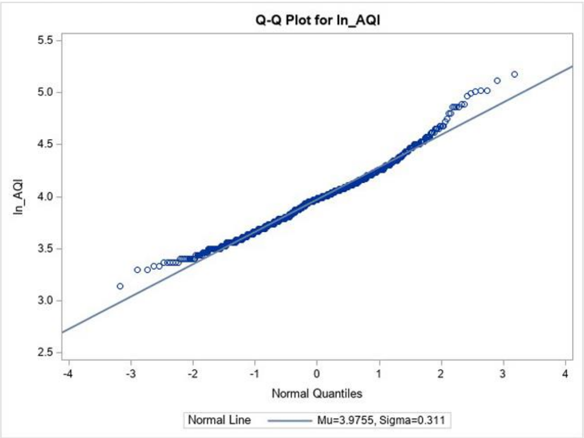


Figure 2a **Figure 2b**

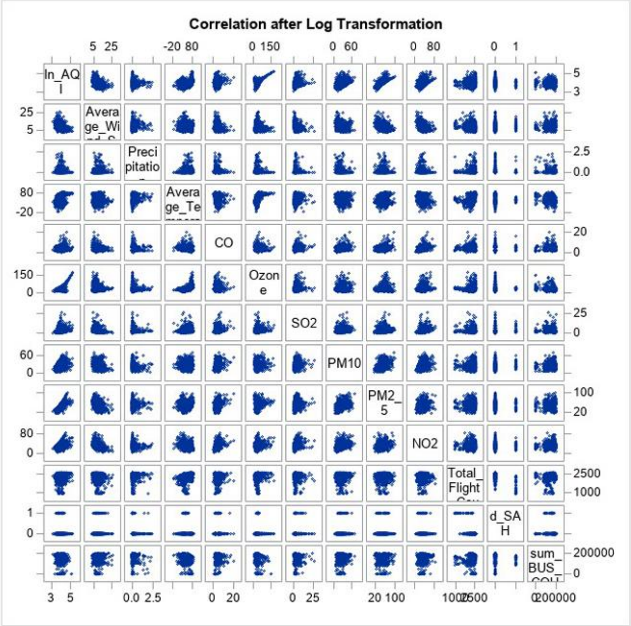


Figure 2c

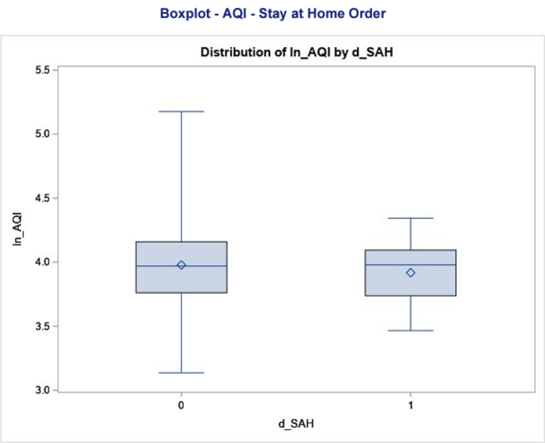
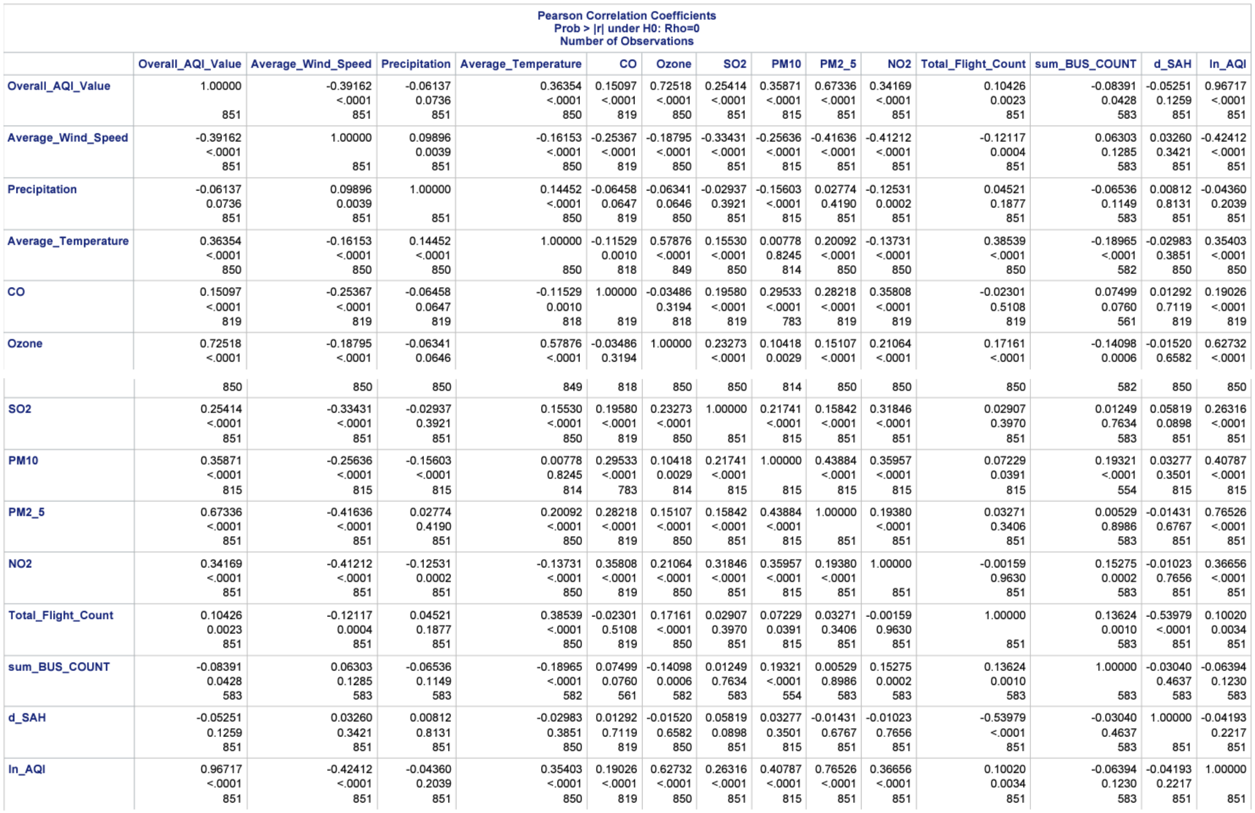


Figure 2d

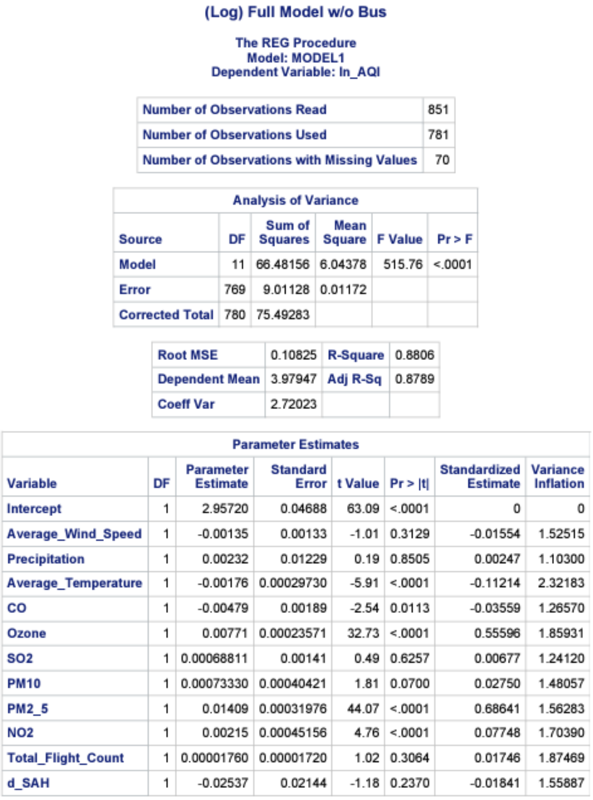


Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Overall_AQI_Value	Average_Wind_Speed	Precipitation	Average_Temperature	CO	Ozone	SO2	PM10	PM2_5	NO2	Total_Flight_Count	sum_BUS_COUNT	d_SAH	ln_AQI
Overall_AQI_Value	1.00000 851	-0.39162 <.0001 851	-0.06137 0.0736 851	0.36354 <.0001 850	0.15097 <.0001 819	0.72518 <.0001 850	0.25414 <.0001 851	0.35871 <.0001 815	0.67336 <.0001 851	0.34169 <.0001 851	0.10426 0.0023 851	-0.08391 0.0428 583	-0.05251 0.1259 851	0.96717 <.0001 851
Average_Wind_Speed	-0.39162 <.0001 851	1.00000 851	0.09896 0.0039 851	-0.16153 <.0001 850	-0.25367 <.0001 819	-0.18795 <.0001 850	-0.33431 <.0001 851	-0.25636 <.0001 815	-0.41636 <.0001 851	-0.41212 <.0001 851	-0.12117 0.0004 851	0.06303 0.1285 583	0.03260 0.3421 851	-0.42412 <.0001 851
Precipitation	-0.06137 0.0736 851	0.09896 0.0039 851	1.00000 851	0.14452 <.0001 850	-0.06458 0.0647 819	-0.06341 0.0646 850	-0.02937 0.3921 851	-0.15603 <.0001 815	0.02774 0.4190 851	-0.12531 0.0002 851	0.04521 0.1877 851	-0.06536 0.1149 583	0.00812 0.8131 851	-0.04360 0.2039 851
Average_Temperature	0.36354 <.0001 850	-0.16153 <.0001 850	0.14452 <.0001 850	1.00000 850	-0.11529 0.0010 818	0.57876 <.0001 849	0.15530 <.0001 850	0.00778 0.8245 814	0.20092 <.0001 850	-0.13731 <.0001 850	0.38539 <.0001 850	-0.18965 <.0001 582	-0.02983 0.3851 850	0.35403 <.0001 850
CO	0.15097 <.0001 819	-0.25367 <.0001 819	-0.06458 0.0647 819	-0.11529 0.0010 818	1.00000 819	-0.03486 0.3194 818	0.19580 <.0001 819	0.29533 <.0001 783	0.28218 <.0001 819	0.35808 <.0001 819	-0.02301 0.5108 819	0.07499 0.0760 561	0.01292 0.7119 819	0.19026 <.0001 819
Ozone	0.72518 <.0001 850	-0.18795 <.0001 850	-0.06341 0.0646 850	0.57876 <.0001 849	-0.03486 0.3194 818	1.00000 850	0.23273 <.0001 850	0.10418 0.0029 814	0.15107 <.0001 850	0.21064 <.0001 850	0.17161 <.0001 850	-0.14098 0.0006 582	-0.01520 0.6582 850	0.62732 <.0001 850
SO2	0.25414 <.0001 851	-0.33431 <.0001 851	-0.02937 0.3921 851	0.15530 <.0001 850	0.19580 <.0001 819	0.23273 <.0001 850	1.00000 851	0.21741 <.0001 815	0.15842 <.0001 851	0.31846 <.0001 851	0.02907 0.3970 851	0.01249 0.7634 583	0.05819 0.0898 851	0.26316 <.0001 851
PM10	0.35871 <.0001 815	-0.25636 <.0001 815	-0.15603 <.0001 815	0.00778 0.8245 814	0.29533 <.0001 783	0.10418 0.0029 814	0.21741 <.0001 815	1.00000 815	0.43884 <.0001 815	0.35957 <.0001 815	0.07229 0.0391 815	0.19321 <.0001 554	0.03277 0.3501 815	0.40787 <.0001 815
PM2_5	0.67336 <.0001 851	-0.41636 <.0001 851	0.02774 0.4190 851	0.20092 <.0001 850	0.28218 <.0001 819	0.15107 <.0001 850	0.15842 <.0001 851	0.43884 <.0001 815	1.00000 851	0.19380 <.0001 851	0.03271 0.3406 851	0.00529 0.8986 583	-0.01431 0.6767 851	0.76526 <.0001 851
NO2	0.34169 <.0001 851	-0.41212 <.0001 851	-0.12531 0.0002 851	-0.13731 <.0001 850	0.35808 <.0001 819	0.21064 <.0001 850	0.31846 <.0001 851	0.35957 <.0001 815	0.19380 <.0001 851	1.00000 851	-0.00159 0.9630 851	0.15275 0.0002 583	-0.01023 0.7656 851	0.36656 <.0001 851
Total_Flight_Count	0.10426 0.0023 851	-0.12117 0.0004 851	0.04521 0.1877 851	0.38539 <.0001 850	-0.02301 0.5108 819	0.17161 <.0001 850	0.02907 0.3970 851	0.07229 0.0391 815	0.03271 0.3406 851	-0.00159 0.9630 851	1.00000 851	0.13624 0.0010 583	-0.53979 <.0001 851	0.10020 0.0034 851
sum_BUS_COUNT	-0.08391 0.0428 583	0.06303 0.1285 583	-0.06536 0.1149 583	-0.18965 <.0001 582	0.07499 0.0760 561	-0.14098 0.0006 582	0.01249 0.7634 583	0.19321 <.0001 554	0.00529 0.8986 583	0.15275 0.0002 583	0.13624 0.0010 583	1.00000 583	-0.03040 0.4637 583	-0.06394 0.1230 583
d_SAH	-0.05251 0.1259 851	0.03260 0.3421 851	0.00812 0.8131 851	-0.02983 0.3851 850	0.01292 0.7119 819	-0.01520 0.6582 850	0.05819 0.0898 851	0.03277 0.3501 815	-0.01431 0.6767 851	-0.01023 0.7656 851	-0.53979 <.0001 851	-0.03040 0.4637 583	1.00000 851	-0.04193 0.2217 851
ln_AQI	0.96717 <.0001 851	-0.42412 <.0001 851	-0.04360 0.2039 851	0.35403 <.0001 850	0.19026 <.0001 819	0.62732 <.0001 850	0.26316 <.0001 851	0.40787 <.0001 815	0.76526 <.0001 851	0.36656 <.0001 851	0.10020 0.0034 851	-0.06394 0.1230 583	-0.04193 0.2217 851	1.00000 851

Figure 2e

Full Model transformed



(Log) Full Model w/o Bus

The REG Procedure
Model: MODEL1
Dependent Variable: ln_AQI

Number of Observations Read	851
Number of Observations Used	781
Number of Observations with Missing Values	70

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	66.48156	6.04378	515.76	<.0001
Error	769	9.01128	0.01172		
Corrected Total	780	75.49283			

Root MSE	0.10825	R-Square	0.8806
Dependent Mean	3.97947	Adj R-Sq	0.8789
Coeff Var	2.72023		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	2.95720	0.04688	63.09	<.0001	0	0
Average_Wind_Speed	1	-0.00135	0.00133	-1.01	0.3129	-0.01554	1.52515
Precipitation	1	0.00232	0.01229	0.19	0.8505	0.00247	1.10300
Average_Temperature	1	-0.00176	0.00029730	-5.91	<.0001	-0.11214	2.32183
CO	1	-0.00479	0.00189	-2.54	0.0113	-0.03559	1.26570
Ozone	1	0.00771	0.00023571	32.73	<.0001	0.55596	1.85931
SO2	1	0.00068811	0.00141	0.49	0.6257	0.00677	1.24120
PM10	1	0.00073330	0.00040421	1.81	0.0700	0.02750	1.48057
PM2_5	1	0.01409	0.00031976	44.07	<.0001	0.68641	1.56283
NO2	1	0.00215	0.00045156	4.76	<.0001	0.07748	1.70390
Total_Flight_Count	1	0.00001760	0.00001720	1.02	0.3064	0.01746	1.87469
d_SAH	1	-0.02537	0.02144	-1.18	0.2370	-0.01841	1.55887

Figure 3a

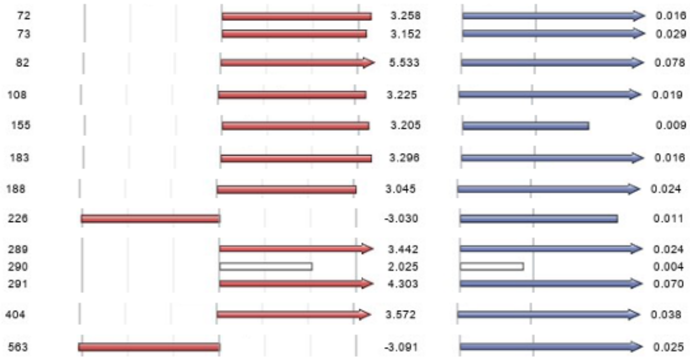


Figure 3b

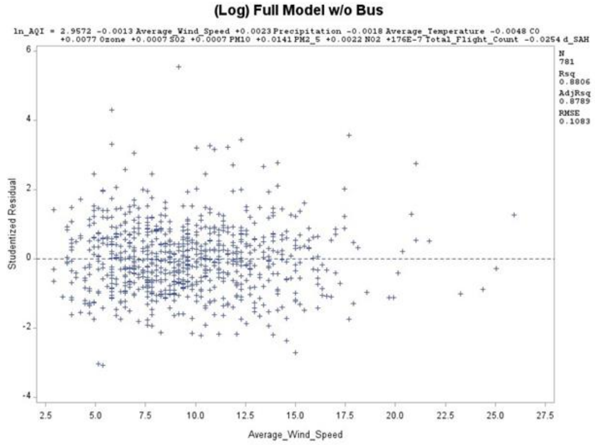


Figure 3c

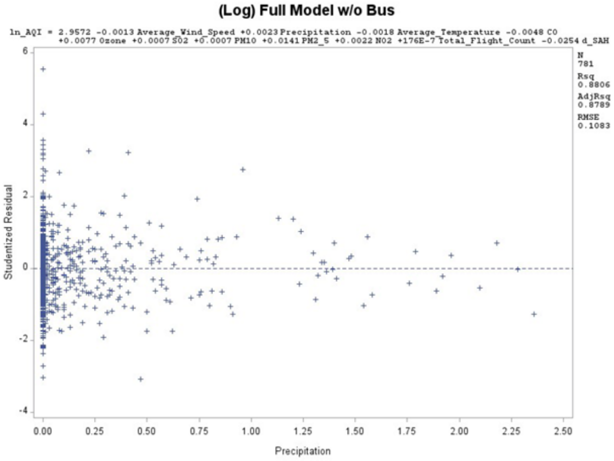


Figure 3d

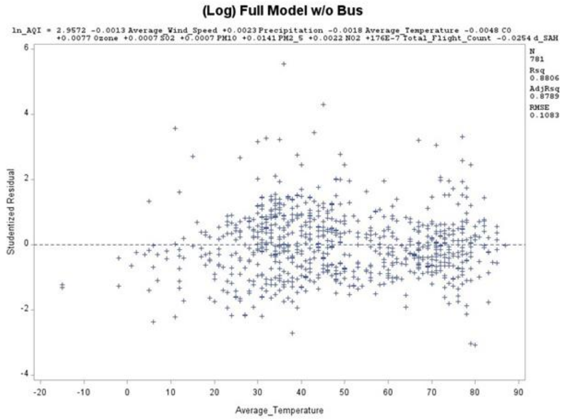


Figure 3e

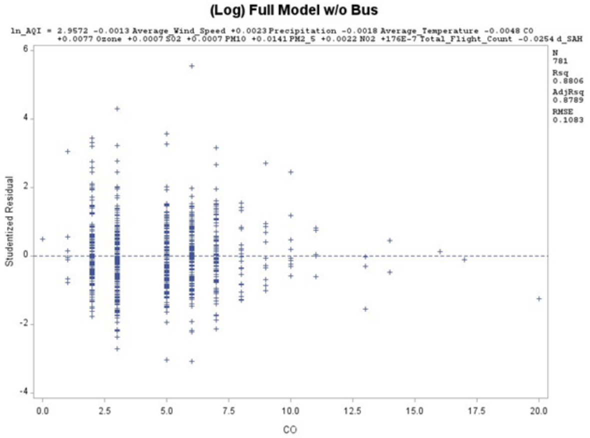


Figure 3f

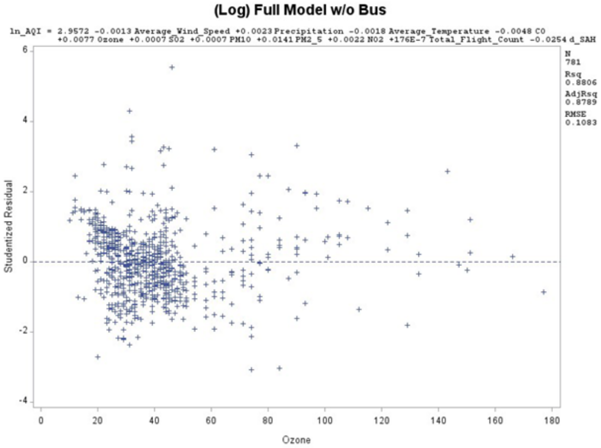


Figure 3g

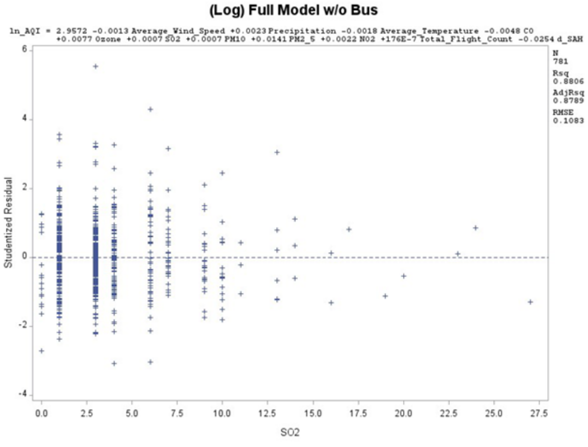


Figure 3h

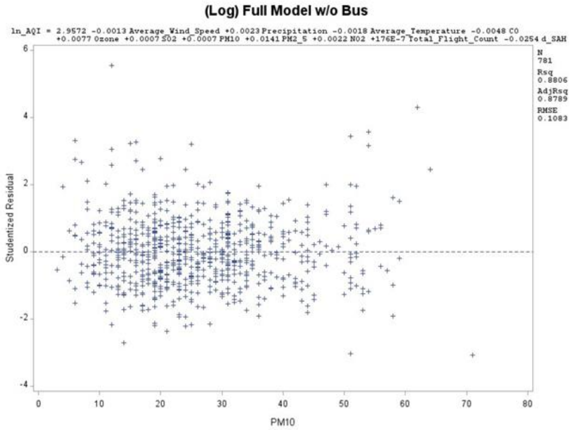


Figure 3i

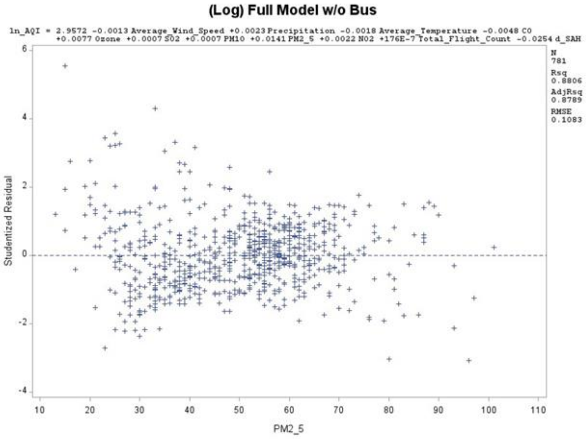


Figure 3j

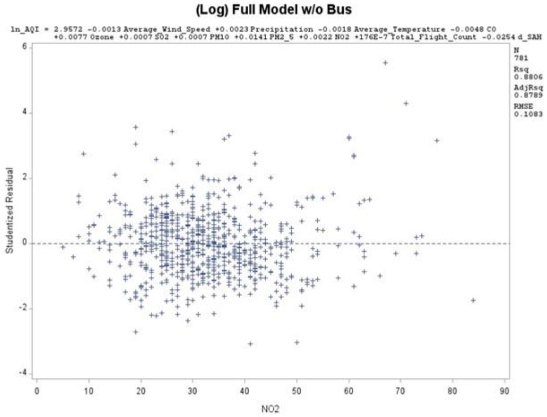


Figure 3k

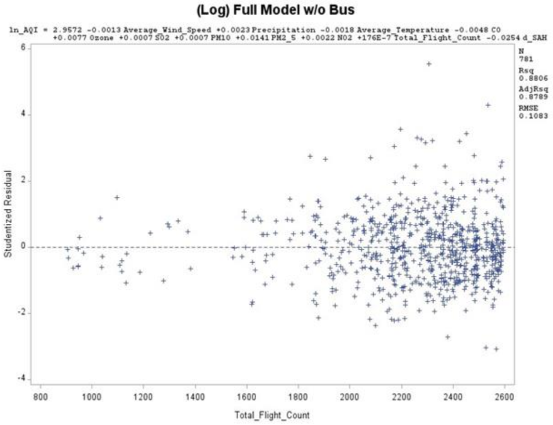


Figure 3l

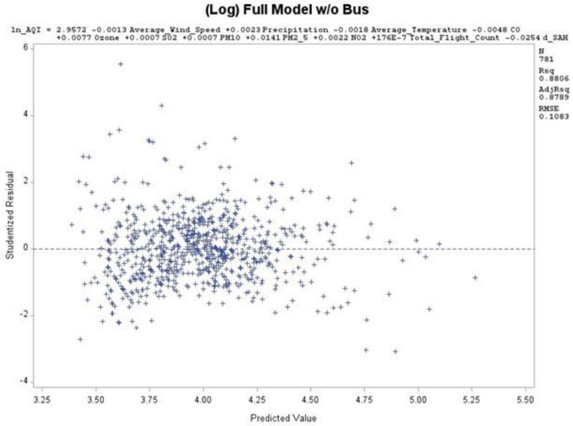


Figure 3m

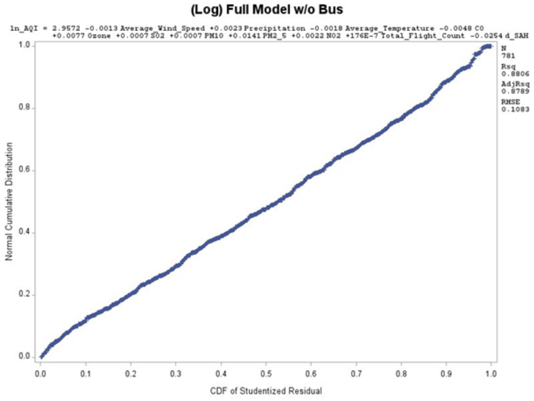
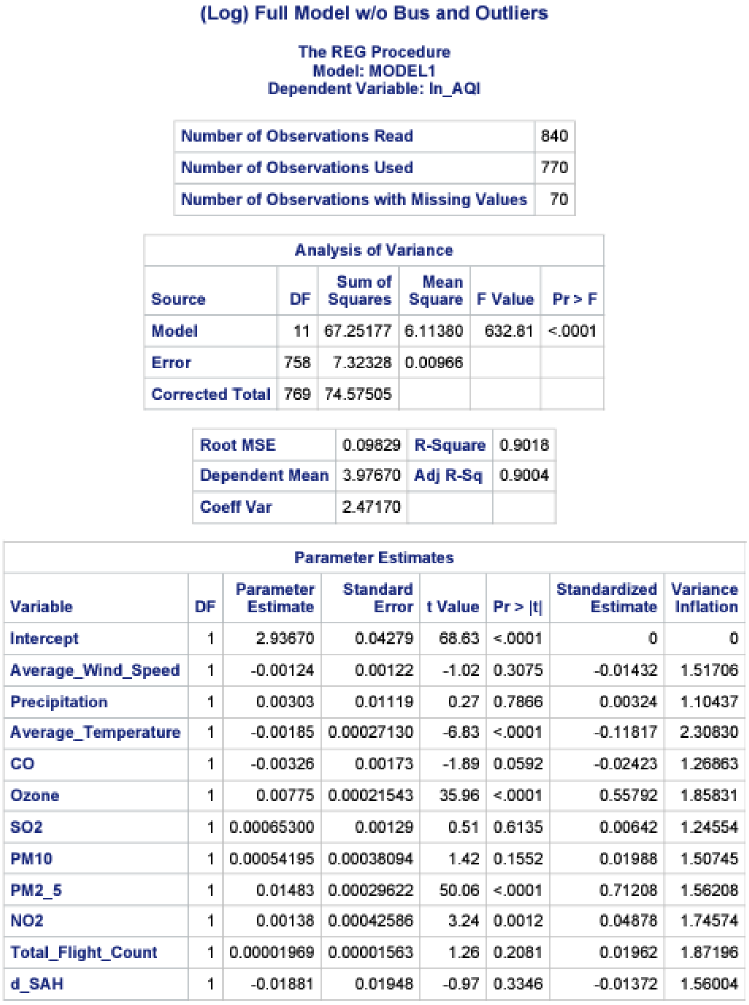


Figure 3n

Full model excluding observations flagged as outliers and influential points



(Log) Full Model w/o Bus and Outliers

The REG Procedure
Model: MODEL1
Dependent Variable: ln_AQI

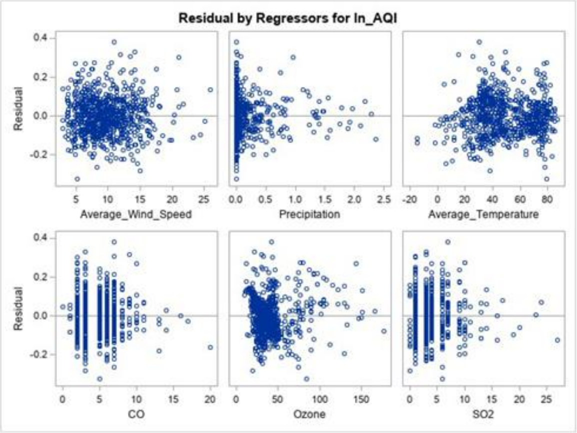
Number of Observations Read	840
Number of Observations Used	770
Number of Observations with Missing Values	70

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	67.25177	6.11380	632.81	<.0001
Error	758	7.32328	0.00966		
Corrected Total	769	74.57505			

Root MSE	0.09829	R-Square	0.9018
Dependent Mean	3.97670	Adj R-Sq	0.9004
Coeff Var	2.47170		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	2.93670	0.04279	68.63	<.0001	0	0
Average_Wind_Speed	1	-0.00124	0.00122	-1.02	0.3075	-0.01432	1.51706
Precipitation	1	0.00303	0.01119	0.27	0.7866	0.00324	1.10437
Average_Temperature	1	-0.00185	0.00027130	-6.83	<.0001	-0.11817	2.30830
CO	1	-0.00326	0.00173	-1.89	0.0592	-0.02423	1.26863
Ozone	1	0.00775	0.00021543	35.96	<.0001	0.55792	1.85831
SO2	1	0.00065300	0.00129	0.51	0.6135	0.00642	1.24554
PM10	1	0.00054195	0.00038094	1.42	0.1552	0.01988	1.50745
PM2_5	1	0.01483	0.00029622	50.06	<.0001	0.71208	1.56208
NO2	1	0.00138	0.00042586	3.24	0.0012	0.04878	1.74574
Total_Flight_Count	1	0.00001969	0.00001563	1.26	0.2081	0.01962	1.87196
d_SAH	1	-0.01881	0.01948	-0.97	0.3346	-0.01372	1.56004

Figure 4a



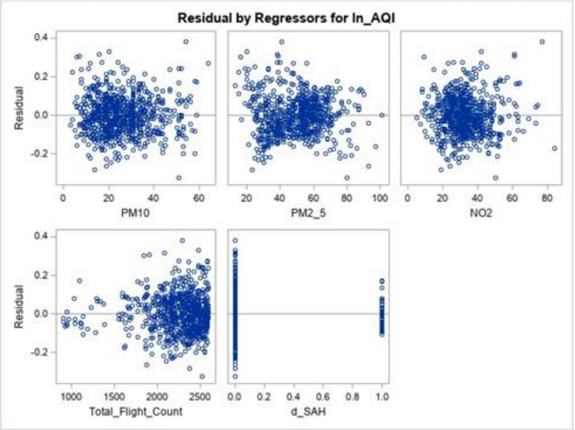
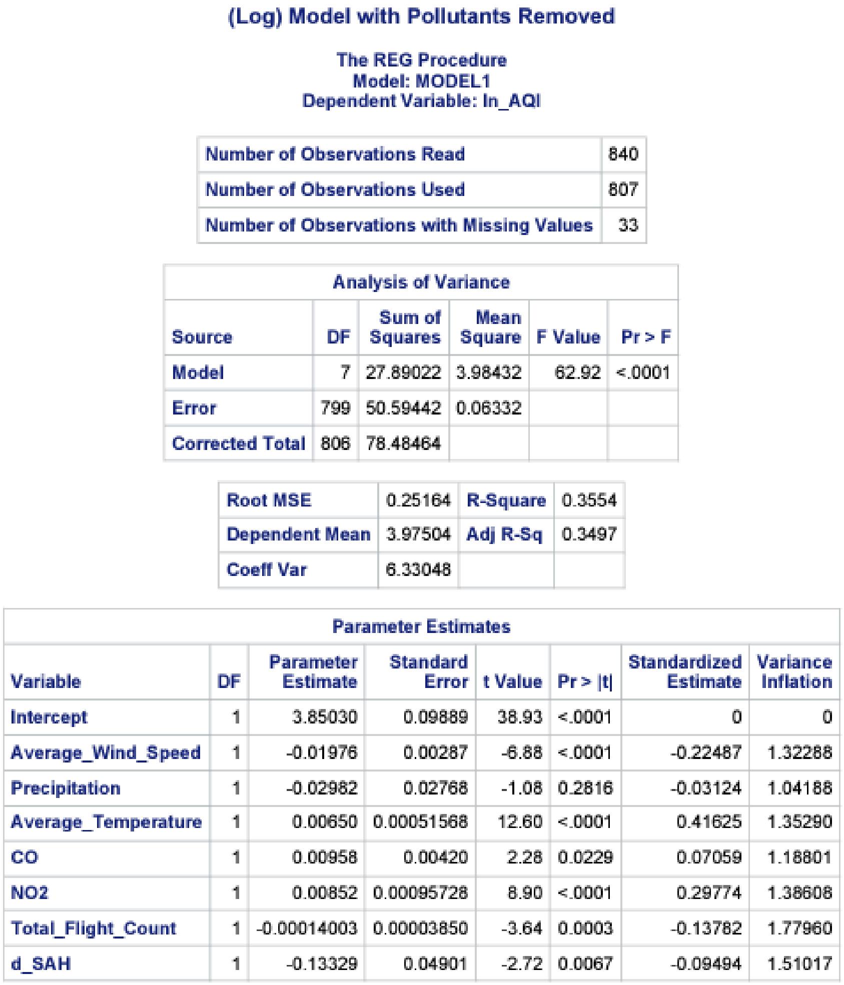


Figure 4b **Figure 4c**

Reduced model



(Log) Model with Pollutants Removed

The REG Procedure
Model: MODEL1
Dependent Variable: ln_AQI

Number of Observations Read	840
Number of Observations Used	807
Number of Observations with Missing Values	33

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	27.89022	3.98432	62.92	<.0001
Error	799	50.59442	0.06332		
Corrected Total	806	78.48464			

Root MSE	0.25164	R-Square	0.3554
Dependent Mean	3.97504	Adj R-Sq	0.3497
Coeff Var	6.33048		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	3.85030	0.09889	38.93	<.0001	0	0
Average_Wind_Speed	1	-0.01976	0.00287	-6.88	<.0001	-0.22487	1.32288
Precipitation	1	-0.02982	0.02768	-1.08	0.2816	-0.03124	1.04188
Average_Temperature	1	0.00650	0.00051568	12.60	<.0001	0.41625	1.35290
CO	1	0.00958	0.00420	2.28	0.0229	0.07059	1.18801
NO2	1	0.00852	0.00095728	8.90	<.0001	0.29774	1.38608
Total_Flight_Count	1	-0.00014003	0.00003850	-3.64	0.0003	-0.13782	1.77960
d_SAH	1	-0.13329	0.04901	-2.72	0.0067	-0.09494	1.51017

Figure 4f

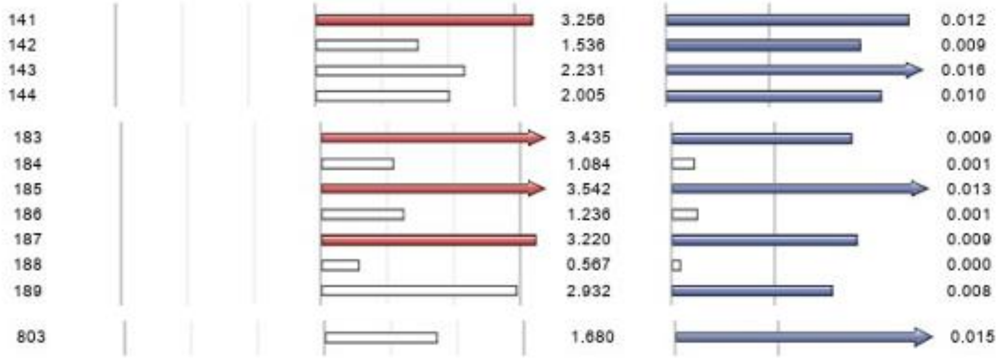
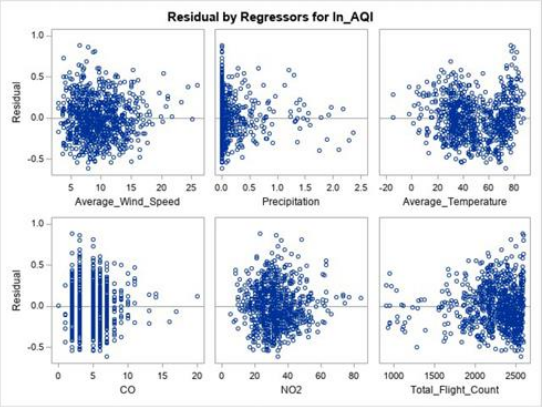


Figure 4g



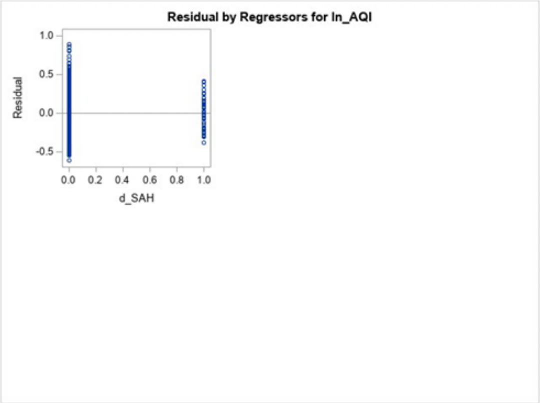


Figure 4h **Figure 4i**

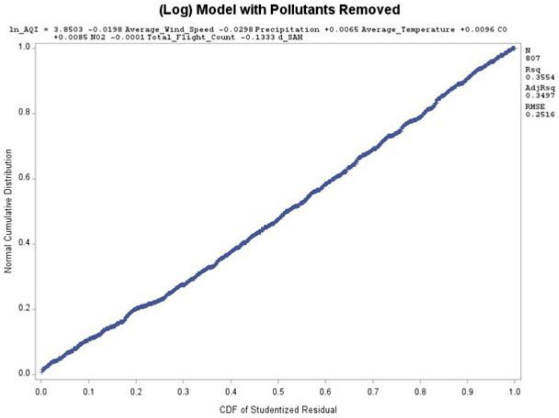
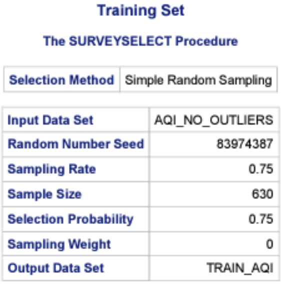


Figure 4j

Split data for training and testing set



Training Set

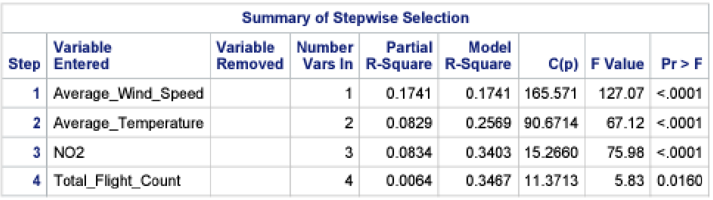
The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	AQI_NO_OUTLIERS
Random Number Seed	83974387
Sampling Rate	0.75
Sample Size	630
Selection Probability	0.75
Sampling Weight	0
Output Data Set	TRAIN_AQI

Figure 5

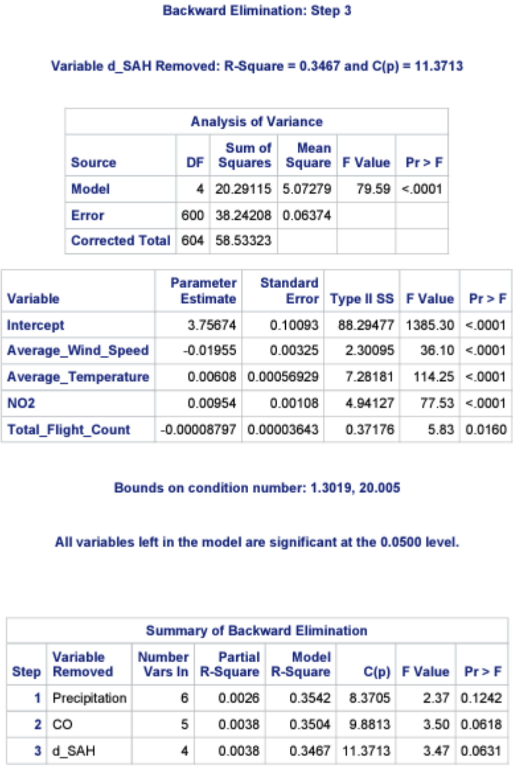
Model Selection: Stepwise



Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Average_Wind_Speed		1	0.1741	0.1741	165.571	127.07	<.0001
2	Average_Temperature		2	0.0829	0.2569	90.6714	67.12	<.0001
3	NO2		3	0.0834	0.3403	15.2660	75.98	<.0001
4	Total_Flight_Count		4	0.0064	0.3467	11.3713	5.83	0.0160

Figure 6a

Model Selection: Backwards



Backward Elimination: Step 3

Variable d_SAH Removed: R-Square = 0.3467 and C(p) = 11.3713

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	20.29115	5.07279	79.59	<.0001
Error	600	38.24208	0.06374		
Corrected Total	604	58.53323			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.75674	0.10093	88.29477	1385.30	<.0001
Average_Wind_Speed	-0.01955	0.00325	2.30095	36.10	<.0001
Average_Temperature	0.00608	0.00056929	7.28181	114.25	<.0001
NO2	0.00954	0.00108	4.94127	77.53	<.0001
Total_Flight_Count	-0.00008797	0.00003643	0.37176	5.83	0.0160

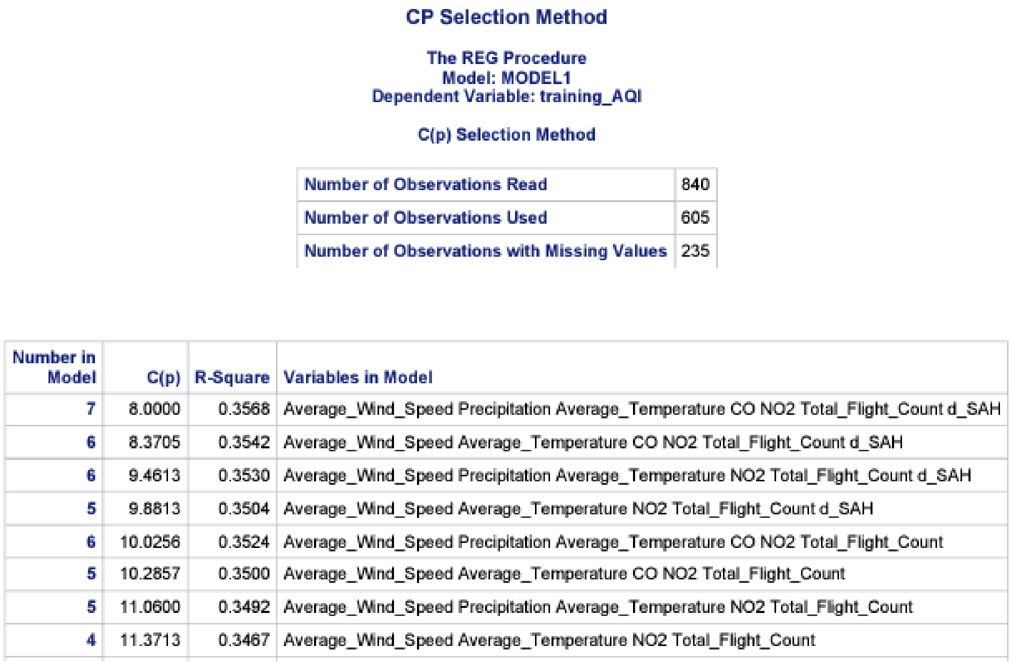
Bounds on condition number: 1.3019, 20.005

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Precipitation	6	0.0026	0.3542	8.3705	2.37	0.1242
2	CO	5	0.0038	0.3504	9.8813	3.50	0.0618
3	d_SAH	4	0.0038	0.3467	11.3713	3.47	0.0631

Figure 6b

Model Selection: CP



CP Selection Method

The REG Procedure
Model: MODEL1
Dependent Variable: training_AQI

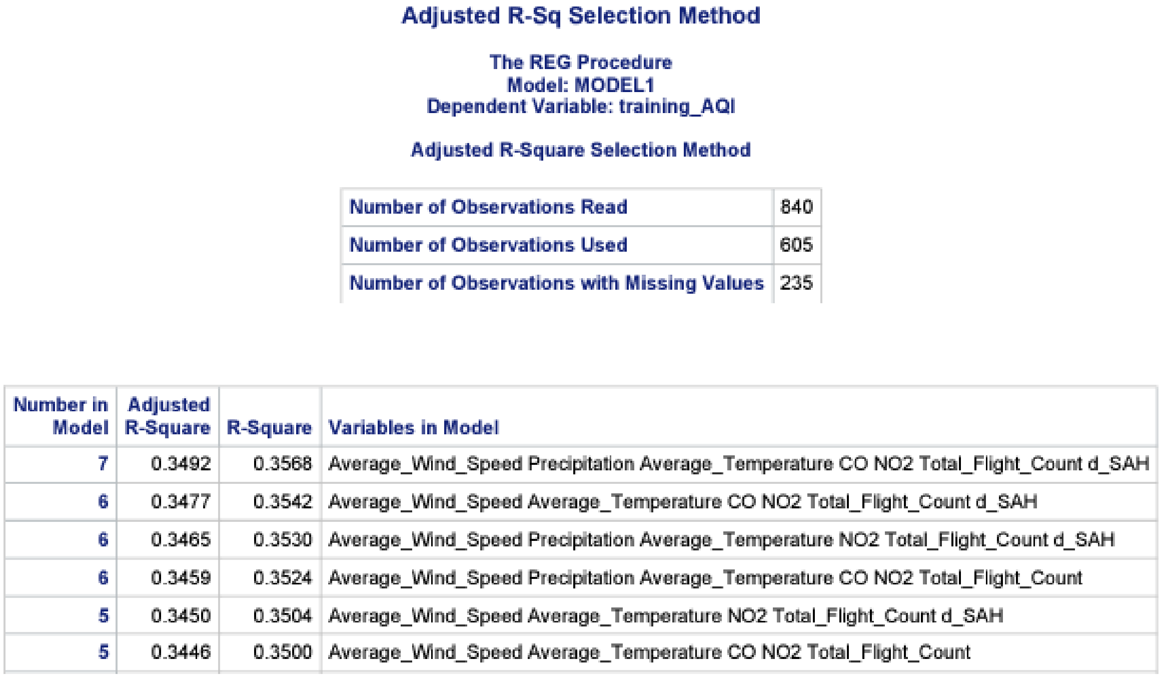
C(p) Selection Method

Number of Observations Read	840
Number of Observations Used	605
Number of Observations with Missing Values	235

Number in Model	C(p)	R-Square	Variables in Model
7	8.0000	0.3568	Average_Wind_Speed Precipitation Average_Temperature CO NO2 Total_Flight_Count d_SAH
6	8.3705	0.3542	Average_Wind_Speed Average_Temperature CO NO2 Total_Flight_Count d_SAH
6	9.4613	0.3530	Average_Wind_Speed Precipitation Average_Temperature NO2 Total_Flight_Count d_SAH
5	9.8813	0.3504	Average_Wind_Speed Average_Temperature NO2 Total_Flight_Count d_SAH
6	10.0256	0.3524	Average_Wind_Speed Precipitation Average_Temperature CO NO2 Total_Flight_Count
5	10.2857	0.3500	Average_Wind_Speed Average_Temperature CO NO2 Total_Flight_Count
5	11.0600	0.3492	Average_Wind_Speed Precipitation Average_Temperature NO2 Total_Flight_Count
4	11.3713	0.3467	Average_Wind_Speed Average_Temperature NO2 Total_Flight_Count

Figure 6c

Model Selection: Adjusted-R2



Adjusted R-Sq Selection Method

The REG Procedure
Model: MODEL1
Dependent Variable: training_AQI

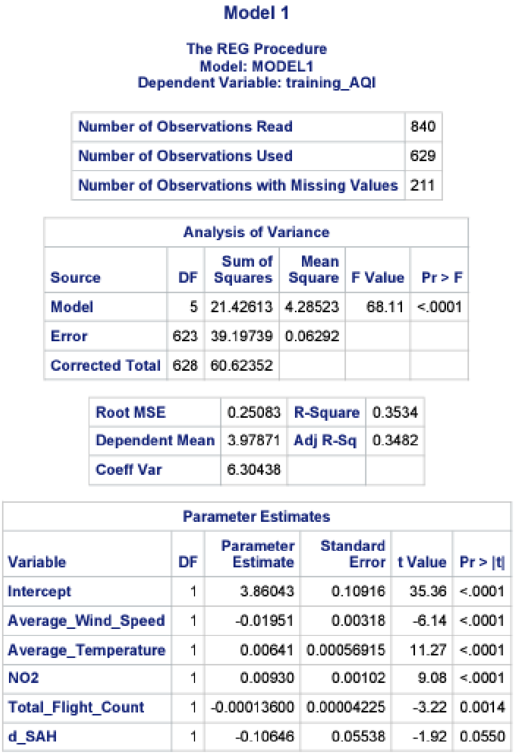
Adjusted R-Square Selection Method

Number of Observations Read	840
Number of Observations Used	605
Number of Observations with Missing Values	235

Number in Model	Adjusted R-Square	R-Square	Variables in Model
7	0.3492	0.3568	Average_Wind_Speed Precipitation Average_Temperature CO NO2 Total_Flight_Count d_SAH
6	0.3477	0.3542	Average_Wind_Speed Average_Temperature CO NO2 Total_Flight_Count d_SAH
6	0.3465	0.3530	Average_Wind_Speed Precipitation Average_Temperature NO2 Total_Flight_Count d_SAH
6	0.3459	0.3524	Average_Wind_Speed Precipitation Average_Temperature CO NO2 Total_Flight_Count
5	0.3450	0.3504	Average_Wind_Speed Average_Temperature NO2 Total_Flight_Count d_SAH
5	0.3446	0.3500	Average_Wind_Speed Average_Temperature CO NO2 Total_Flight_Count

Figure 6d

Model comparison for two selected models



Model 1

The REG Procedure
Model: MODEL1
Dependent Variable: training_AQI

Number of Observations Read	840
Number of Observations Used	629
Number of Observations with Missing Values	211

Analysis of Variance

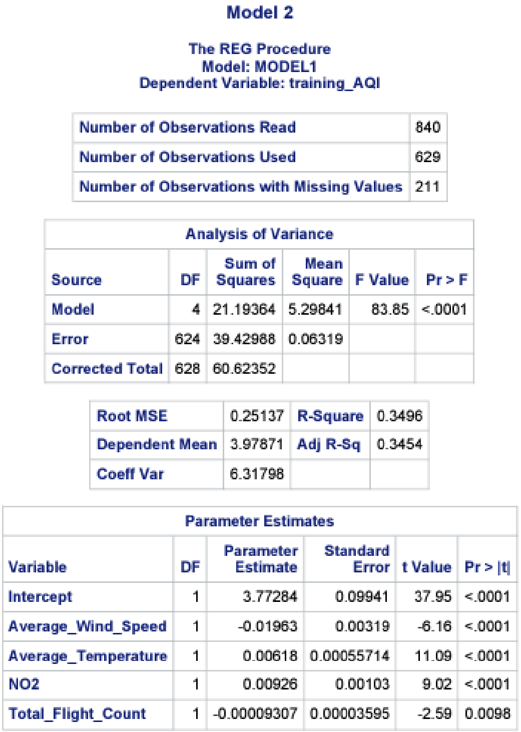
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	21.42613	4.28523	68.11	<.0001
Error	623	39.19739	0.06292		
Corrected Total	628	60.62352			

Root MSE	0.25083	R-Square	0.3534
Dependent Mean	3.97871	Adj R-Sq	0.3482
Coeff Var	6.30438		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.86043	0.10916	35.36	<.0001
Average_Wind_Speed	1	-0.01951	0.00318	-6.14	<.0001
Average_Temperature	1	0.00641	0.00056915	11.27	<.0001
NO2	1	0.00930	0.00102	9.08	<.0001
Total_Flight_Count	1	-0.00013600	0.00004225	-3.22	0.0014
d_SAH	1	-0.10646	0.05538	-1.92	0.0550

Figure 7a



Model 2

The REG Procedure
Model: MODEL1
Dependent Variable: training_AQI

Number of Observations Read	840
Number of Observations Used	629
Number of Observations with Missing Values	211

Analysis of Variance

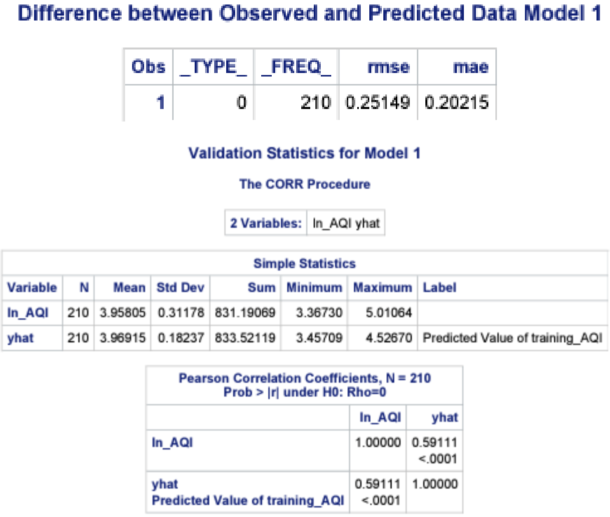
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	21.19364	5.29841	83.85	<.0001
Error	624	39.42988	0.06319		
Corrected Total	628	60.62352			

Root MSE	0.25137	R-Square	0.3496
Dependent Mean	3.97871	Adj R-Sq	0.3454
Coeff Var	6.31798		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.77284	0.09941	37.95	<.0001
Average_Wind_Speed	1	-0.01963	0.00319	-6.16	<.0001
Average_Temperature	1	0.00618	0.00055714	11.09	<.0001
NO2	1	0.00926	0.00103	9.02	<.0001
Total_Flight_Count	1	-0.00009307	0.00003595	-2.59	0.0098

Figure 8b



Difference between Observed and Predicted Data Model 1

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	210	0.25149	0.20215

Validation Statistics for Model 1

The CORR Procedure

2 Variables:	In_AQI yhat

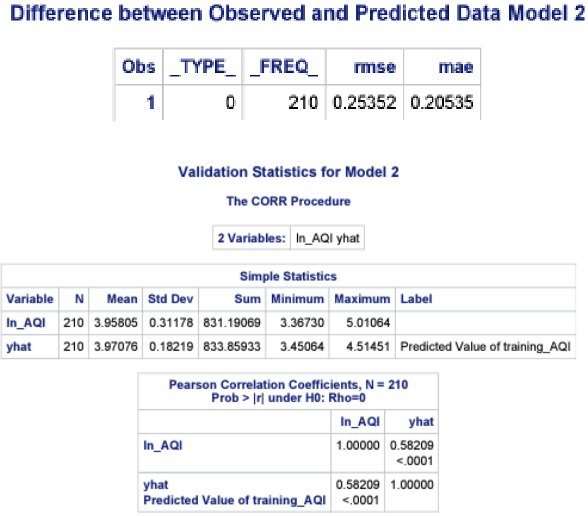
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
In_AQI	210	3.95805	0.31178	831.19069	3.36730	5.01064	
yhat	210	3.96915	0.18237	833.52119	3.45709	4.52670	Predicted Value of training_AQI

Pearson Correlation Coefficients, N = 210
Prob > \|r\| under H0: Rho=0

	In_AQI	yhat
In_AQI	1.00000	0.59111 <.0001
yhat Predicted Value of training_AQI	0.59111 <.0001	1.00000

Figure 7b



Difference between Observed and Predicted Data Model 2

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	210	0.25352	0.20535

Validation Statistics for Model 2

The CORR Procedure

2 Variables:	In_AQI yhat

Simple Statistics

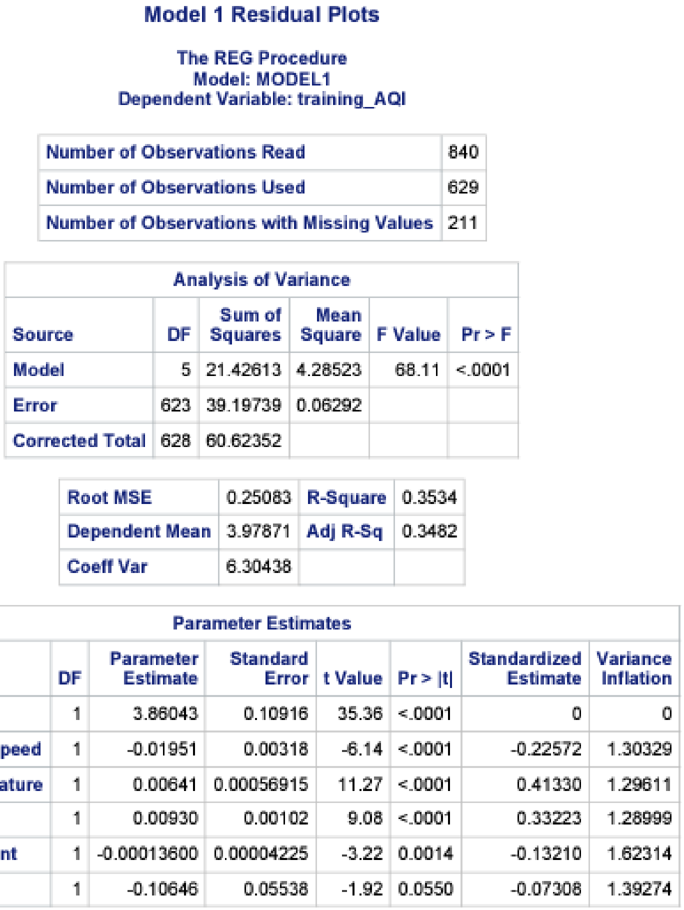
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
In_AQI	210	3.95805	0.31178	831.19069	3.36730	5.01064	
yhat	210	3.97076	0.18219	833.85933	3.45064	4.51451	Predicted Value of training_AQI

Pearson Correlation Coefficients, N = 210
Prob > \|r\| under H0: Rho=0

	In_AQI	yhat
In_AQI	1.00000	0.58209 <.0001
yhat Predicted Value of training_AQI	0.58209 <.0001	1.00000

Figure 8b

Final Model



Model 1 Residual Plots

The REG Procedure
Model: MODEL1
Dependent Variable: training_AQI

Number of Observations Read	840
Number of Observations Used	629
Number of Observations with Missing Values	211

Analysis of Variance

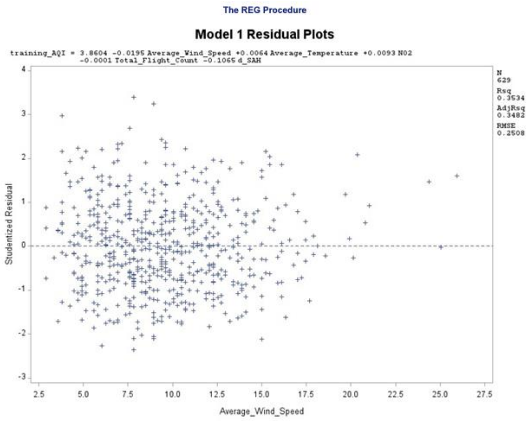
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	21.42613	4.28523	68.11	<.0001
Error	623	39.19739	0.06292		
Corrected Total	628	60.62352			

Root MSE	0.25083	R-Square	0.3534
Dependent Mean	3.97871	Adj R-Sq	0.3482
Coeff Var	6.30438		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	3.86043	0.10916	35.36	<.0001	0	0
Average_Wind_Speed	1	-0.01951	0.00318	-6.14	<.0001	-0.22572	1.30329
Average_Temperature	1	0.00641	0.00056915	11.27	<.0001	0.41330	1.29611
NO2	1	0.00930	0.00102	9.08	<.0001	0.33223	1.28999
Total_Flight_Count	1	-0.00013600	0.00004225	-3.22	0.0014	-0.13210	1.62314
d_SAH	1	-0.10646	0.05538	-1.92	0.0550	-0.07308	1.39274

Figure 9a



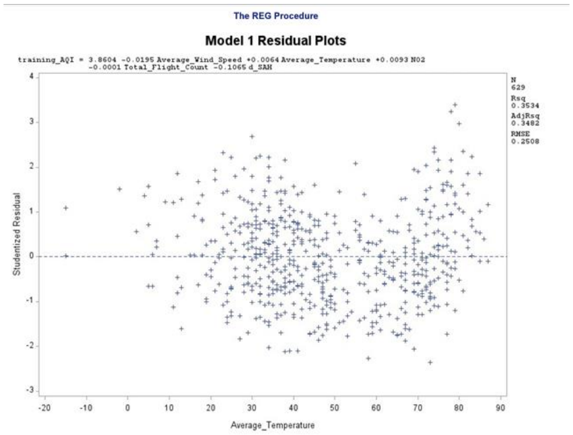


Figure 9b

Figure 9c

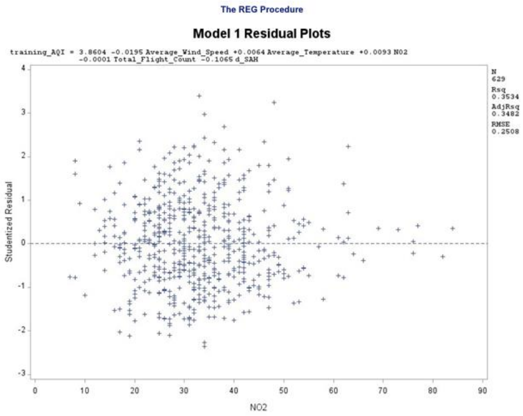


Figure 9d

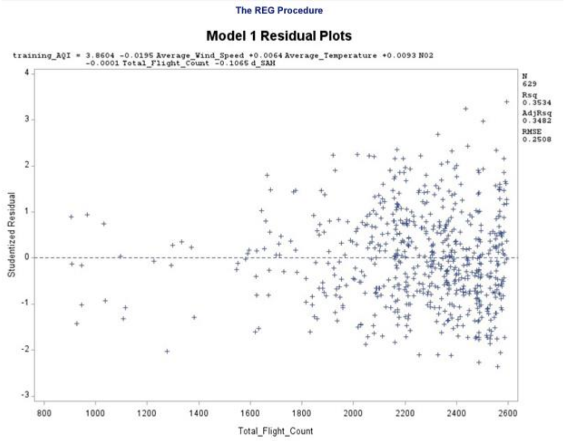


Figure 9e

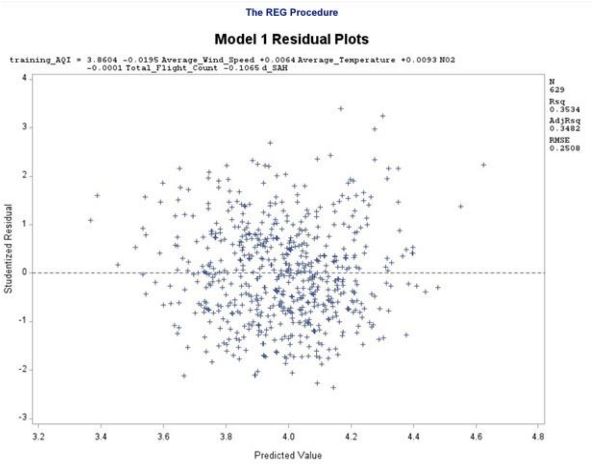


Figure 9f

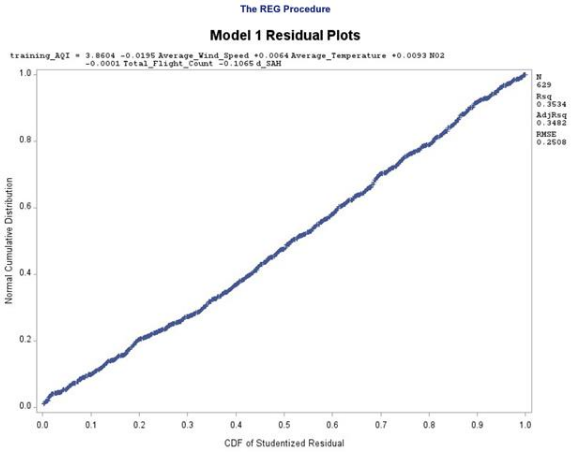
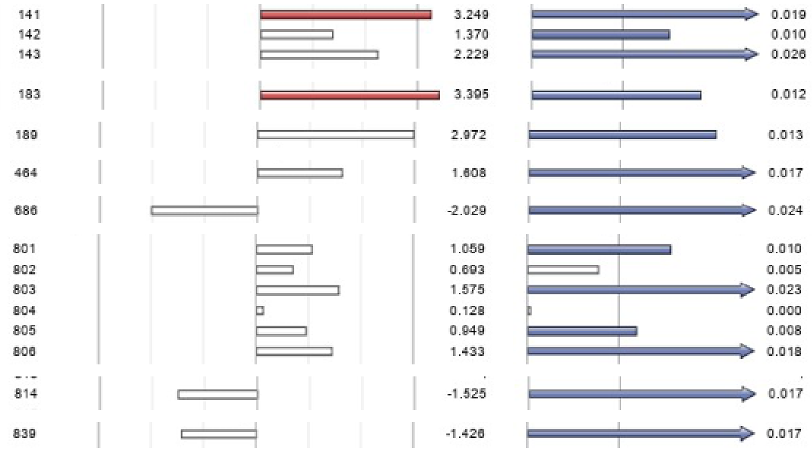


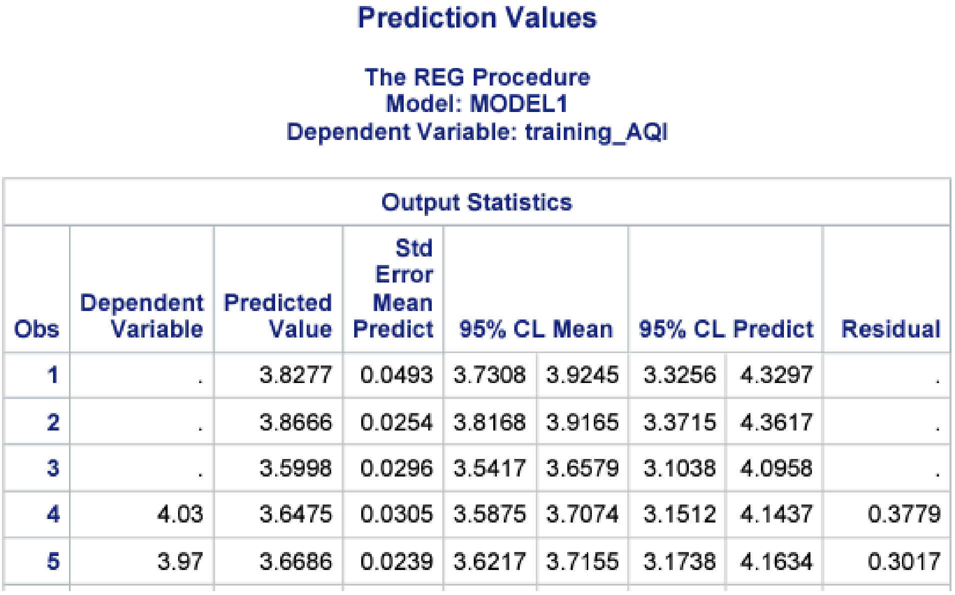
Figure 9g



141	3.249	0.019
142	1.370	0.010
143	2.229	0.026
183	3.395	0.012
189	2.972	0.013
464	1.608	0.017
686	-2.029	0.024
801	1.059	0.010
802	0.693	0.005
803	1.575	0.023
804	0.128	0.000
805	0.949	0.008
806	1.433	0.018
814	-1.525	0.017
839	-1.426	0.017

Figure 9h

Predicted Values



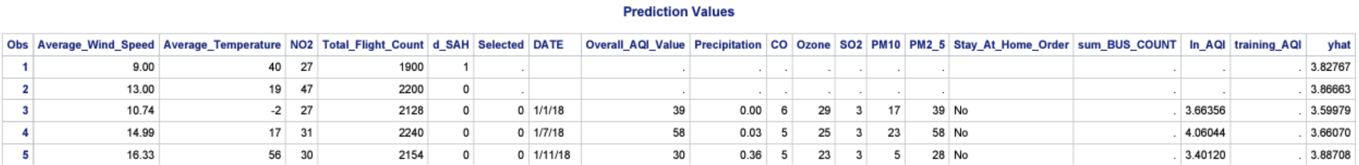
Prediction Values

The REG Procedure
Model: MODEL1
Dependent Variable: training_AQI

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	3.8277	0.0493	3.7308	3.9245	3.3256	4.3297	.
2	.	3.8666	0.0254	3.8168	3.9165	3.3715	4.3617	.
3	.	3.5998	0.0296	3.5417	3.6579	3.1038	4.0958	.
4	4.03	3.6475	0.0305	3.5875	3.7074	3.1512	4.1437	0.3779
5	3.97	3.6686	0.0239	3.6217	3.7155	3.1738	4.1634	0.3017

Figure 10a



Prediction Values

Obs	Average_Wind_Speed	Average_Temperature	NO2	Total_Flight_Count	d_SAH	Selected	DATE	Overall_AQI_Value	Precipitation	CO	Ozone	SO2	PM10	PM2_5	Stay_At_Home_Order	sum_BUS_COUNT	ln_AQI	training_AQI	yhat
1	9.00	40	27	1900	1	.		.	.	.	.	.	.	.		.	.	.	3.82767
2	13.00	19	47	2200	0	.		.	.	.	.	.	.	.		.	.	.	3.86663
3	10.74	-2	27	2128	0	0	1/1/18	39	0.00	6	29	3	17	39	No	.	3.66356	.	3.59979
4	14.99	17	31	2240	0	0	1/7/18	58	0.03	5	25	3	23	58	No	.	4.06044	.	3.66070
5	16.33	56	30	2154	0	0	1/11/18	30	0.36	5	23	3	5	28	No	.	3.40120	.	3.88708

Figure 10b

R Code for Gathering Bus Data

```
library(tidyverse)

library("RSocrata")

Chicago_Traffic_Tracker <- read.socrata(

  "https://data.cityofchicago.org/resource/sxs8-h27x.json",

  app_token = "ZN6fIBGFYfwBqHt6WuV2oRiXM",

 )

Chicago_Traffic_Tracker_byDay <- Chicago_Traffic_Tracker %>%

        group_by(substring(TIME,1,10)) %>%


        mutate(date_time=substring(TIME,1,10),count_date_time=count(date_time),count_distinct_SEGMENT_ID=count_d(SEGMENTID),sum_BUS_COUNT=sum('BUS COUNT')) %>%

        ungroup()
```

Python Code for Compiling Flight Data

```
import csv
import os

daily_counts = dict()
total_count = 0

with open('FlightCountsTotal.csv', 'w') as file:
    for f_name in os.listdir():
        if f_name.endswith('.csv'):
            with open(f_name) as csv_file:
                csv_reader = csv.reader(csv_file, delimiter = ',')
                line_count = 0
                for row in csv_reader:
                    if line_count == 0:
                        print(f'Column names are {", ".join(row)}')
                        line_count +=1
                    else:
                        for date in row:
                            if date in daily_counts:
                                daily_counts[date] += 1
                            else:
                                daily_counts[date] = 1
                line_count += 1

    for key in daily_counts.keys():
        file.write("%s,%s\n"%(key, daily_counts[key]))
```