Pollutants***Я*** U.S.



Final Report

Alexandra J. Rukhina

Areeb Abubaker

Jordan Bickelhaupt

Julio Delgado

Jaime Moscoso

DSC-323

November 23, 2020

Professor Gulasingam

Abstract:

Pollution has been a major issue in the world with the pollutants and gases that litter the air in our atmosphere. This also has been a major issue historically in the United States. We were interested in seeing what pollution is like in the United States and seeing how it has changed over time. We decided to examine different types of pollutants and gases in the atmosphere of the United States. This dataset spanned from 2000 to 2016, which gave us a 16-year window. We thought that this window of time would give us enough data to accurately investigate how pollution has changed in the United States during this time. The main problem we are trying to explore in our project is to analyze the data and create a model that we can predict the missing values of the pollutants and create a model for the air quality index (AQI). We plan to analyze the various levels of pollutants to create a model that maypredict the missing values of the following AQI: Nitrogen Dioxide (NO2), Ozone or Trioxygen (03), Sulfur Dioxide ( SO2), and Carbon Monoxide (CO).

We began exploring the data by calculating the correlation of the mean and the maximum value (max\_val) for each of the pollutants. We then would exclude the variables with low correlation. We would be doing this because excluding the variables with low correlation would allow us to develop an accurate model to predict the AQI. We thought it would make the most sense to exclude the variables that would exclude variables with low correlation because those variables would have little to no relationships with AQI. We created two models and then did a fivefold cross-validation test for the two models to see which one would be better. We were able to find out that the first full model was the better one because it would have better test values, and it is more important than the training data. The second model also had a variable that would have insignificant (Date\_Local), and it was insignificant because pollution does not change day by day. Pollution changes over a long period of time; we can cut that variable from the final model. From the final model, we saw that the overall AQI had gone down since 2000, which would mean that the US has made efforts to bring down the AQI. This analysis can be very positive and beneficial for countries in the future because it can be used as a precautious way to identify the AQI and react accordingly to decrease AQI levels if needed.

Intro

As stated previously in the abstract section, the main goal we would like to accomplish is to evaluate and see how the pollution in the atmosphere in the United States is and how that has changed during a time period. The main problem we want to solve in our project is to analyze the data and create a model that we can predict the missing values of the pollutants and create a model for the air quality index (AQI). Pollution has been one of the world’s biggest issues, and it has been historically for the United States. The United States has been ahead in many other issues and metrics, but pollution has not been one of those. Pollution is one of the biggest global killers and affects millions and millions of people. Pollution does not affect humans, but it also impacts animals and the environment. The United States makes up a small portion of the world’s population but the pollution that the United States has contributed a sizable portion of the world’s pollution because of the burning coal, oil, and the world’s natural gases. Therefore, we were interested in the United States’ pollution and how it has changed throughout the years.

This analysis can help see if the United States has been putting in the effort to lower the AQI and decrease its contribution to the world’s pollution. We began this process by exploring information and data from The Environmental Protection Agency.The Environmental Protection Agency has documented different levels of pollution occurring in the United States of America. We obtained the dataset called “US Pollution Data,” and we obtained the dataset from Kaggle.com. In our data analysis, we examined the different types of gases and pollutants in the atmosphere of the United States. The data set has accumulated 16 years of data from 2000 to 2016. We planned to analyze the data and create a model that can predict the missing values of the individual pollutants (NO2, O3, SO2, CO) and create a model for the air quality index. We will begin to explore the data by calculating the correlation for Mean and Maximum value (Max\_Val) of each pollutant and excluding the variables with low correlation. By excluding variables, one can increase the adjusted R-Squared value, developing an accurate model to predict the AQI. Hence, it would allow us to better evaluate the pollution level in the United States atmosphere and see how much effort the United States has put forth to lower the AQI.

Methodology:

From the dataset we used called “US Pollution Data,” which was obtained from Kaggle.com, we used methods to analyze the data and explore how the pollution in the United States is and how it has changed during this 16-year time period. The approach we took to analyze the various levels of pollutants was to create a model that will be able to predict the missing values of each AQI such as Nitrogen Oxide (NO2), Ozone or Trioxide (O3), Sulfur Dioxide (SO2), and finally Carbon Monoxide (CO). We then created an overall AQI by calculating the mean and max value (Max\_Val) for each of the pollutants and then excluding the variables that had low correlations to build a better and accurate model to predict AQI. Also, excluding the variables, one at a time with low correlations would help increase the adjusted R-squared value, which would also help build an accurate model to predict AQI. We first data prepped, and the first step was to investigate the collinearity of the data. The first thing that was immediate about the data was the number of variables, and we had to minimize that. We then were able to figure out the variables with high collinearity that needed to be taken out. We would then create an Overall AQI value to analyze the data better as a means of all AQI values.

We then used data visualization to analyze the data overall such as histograms and boxplots, to evaluate the newly created predictor value and evaluate the different trends and patterns. In the next step, to remove the skewness, we had to do a logarithmic transformation to make it more of a normal curve. After the logarithmic transformation, we then tried discovering critical relationships between the variables we must evaluate the correlation. Then we created a full model using the selection methods and using the adjusted R-Squared. After removing more variables that we discovered in the previous method that would not help us, we moved onto another method. We tried using the forward model to evaluate our model further and see if there are any more issues, such as variables to remove, to understand better how AQI works in the real world. Additionally, we will use a fitted model to see how the variables would look when using the created variable of AQI2 as the dependent. After removing the variables contributing to the issues from the full model and using the forward selection method, then get rid of all the variables that had bad p-values to make the model better. Finally, we used a 5-fold cross-validation test for two models to determine which of the two models is better.

The AQI is the most commonly accessible index of air quality for city residents in the United States and the referring factor for their decisions. Previous research shows that the forecast AQI is of high accuracy, and we regard the actual AQI as the one that people receive in forecast alerts. The data include hourly concentrations of four out of six major air pollutants, SO2, NO2, CO, and O3, monitored by local stations. Based on these concentrations, the AQI is calculated as a comprehensive indicator denoting the level of air pollution and consequential health risks. The daily AQI data for each city is calculated as

Equation (5)

where IAQIp is the individual air quality index (IAQI) of pollutant P, which is calculated by an interpolation-like process based on the concentration of P (Cp), the upper and lower bounds of the pollutant P (BPH and BPLo respectively), and the lower and upper limits of IAQI (IAQILo and IAQIHi respectively) corresponding to the range BPLo to BPH. The Cp values include daily average concentrations of SO2, NO2, O3, and CO, the maximum hourly average concentration of O3, and the maximum 8-h moving average concentration of O3, this being O3\_1st\_Max\_Hour in our data. Then, the AQI equals the maximum of the IAQIs of n (here, n = 5) types of a pollutant:

AQI = max{IAQI1, IAQI2…, IAQIn} (5)

The AQI score ranges from 0 to 500, with a higher value indicating a more serious pollution level. The AQI is divided into four classes to indicate the significance of health effects according to a series of cut-off points.

Analysis:

EDA

Our data set was very large, with a total of 80168 observations across every state in the United States. The data set originally 29 fields, those being as follows: State\_Code, County\_Code, Site\_Num, Address, State, County, City, Date\_Local, NO2\_Units, NO2\_Mean, NO2\_1st\_Max\_Value, NO2\_1st\_Max\_Hour, NO2\_AQI, SO2\_Units, SO2\_Mean, SO2\_1st\_Max\_Value, SO2\_1st\_Max\_Hour, SO2\_AQI, O3\_Units, O3\_Mean, O3\_1st\_Max\_Value, O3\_1st\_Max\_Hour, O3\_AQI, CO\_Units, CO\_Mean, CO\_1st\_Max\_Value, CO\_1st\_Max\_Hour, CO\_AQI.

After deliberating the necessity of these variables, we cut down the number to 19 by combining the pollutants' density (such as CO AQI, NO2 AQI, SO2 AQI, and O3 AQI) into one variable, overall\_AQI. This was also necessary to remove these variables' high collinearity, as we will go into later. We also did away with unnecessarily specific data such as Site\_Num and Address of the observation site, State and County\_Code, County, and City. This ended up leaving us with all numerical values except for the pollutants Units, which we felt was necessary to leave in, as it distinguished between Parts per Billion and Parts per Million.

After visualizing the quantiles, we could see that the new variable we had created, overall\_AQI, was seen to be highly skewed to the left. 95% of the data is before an AQI score of 40, which is drastic as the range was driven to 150. We proceeded with a log transformation to fix this skew, which provided us with a much more normal curve, albeit now skewed slightly to the right.

To better see any outliers, we formed the data into boxplots. One can see that there were many outliers in the data, as the minimum whisker of log\_AQI ranged from -1 to 1 maximum whisker was ranging from 2 to 5, respectively, across the states. We could boil this down to the states' size and population density, such as Wisconsin having a lower range of 1 and a higher range of 4, compared to Minnesota who ranged from -1 to 5, respectively.

Interaction

The following is speculation about our steps, as we did not have time to add this into our submitted code.

First, it would’ve been necessary to determine which variables might directly be interacting with the dependent variable, overall\_AQI. For this model, we would’ve used State and Date\_Local, to see if we could display if time and place had a significant effect on the dependent variable. For this, it would be figured that State and Date\_Local may be considered continuous variables, which requires a slightly more complicated model as follows:

overall\_AQI =

=

=

A positive value for the effect of this interaction term would imply that as time goes on (the higher the day and year), the location (State\_Code) would have a greater effect on the overall AQI, meaning a State’s population density would show in the model in regards to how much or little Date\_Local changes the AQI.

### Regression

In regression, a dataset's features are used as input of the regression model to predict the continuous-valued output. This kind of prediction is obtained by learning the relationship between the input x and the output y. We affixed the input being State\_Code, County\_Code, Site\_Num, Date\_Local, NO2\_Mean, NO2\_1st\_Max\_Hour, O3\_Mean, O3\_1st\_Max\_Hour, SO2\_Mean, SO2\_1st\_Max\_Hour, CO\_Mean, CO\_1st\_Max\_Hour, overall\_AQI, and over\_manual to the output being log\_AQI.

With predictive modeling of this data, we hoped we could then answer the question of where pollution concentrated the most, why the AQI was so poor (in terms of pollutants), how bad may it get if unresolved? Since the data isn’t qualitative, we would continue with polynomial and linear regression. By analyzing the f-stat as

F = =

we could then use the error sum of squares to measure the variance of the residuals. This helped us explain the part of the variation in log\_AQI that was not explained by the model, compared to the regression sum of squares, which explained how much variability in the observed log\_AQI values was in fact explained by our linear regression model. The final f-stat turned to be <.0001 for all variables except Site\_Num, which was left out of the final model.

Collinearity

Collinearity is a problem in linear regression analysis because independent variables correlate - it is a problem because the variables are independent; therefore, it should not correlate. It can also affect the model because this can shift the values of an independent variable in the regression analysis. We created a model of this dataset after transforming the variable of AQI into log\_AQI. After this transformation, we created a regression model using all the dataset variables having log\_AQI as the dependent variable. The reference to this model will be in Appendix M1. In the picture of this model, we can see that there is some collinearity occurring. We will take the necessary steps to get rid of collinearity to come up with a better model.

The collinearity in this model must be eliminated, so we did this by taking out the independent variables one at a time. The air quality index of the gas and pollutants are assumed to cause collinearity. The main culprits may be the max values of these gasses and pollutants. The gas and pollutants' max values must be deleted to get rid of collinearity because these independent variables have high variance inflation values. Typically, a variance inflation value of over 10 is considered a problem, and some of these variables have a variance inflation value of over 200. A high variance inflation value means that an independent variable may not actually be independent. Meaning some independent variables are affecting other independent variables, thus making them dependent variables. This is a problem because there is supposed to be only one dependent variable in linear regression analysis. The highest variance inflation value is the NO2 air quality index, so it should be removed to create a better model. The other variables that should be removed are NO2 max value, O3 max value, SO2 max value, SO2 air quality index, CO max value, and CO air quality index. Removing these variables should create a better regression model and eliminate collinearity in the model.

After playing around with the variables in this data set, we settled on a final model. The final model can be referenced in Appendix M2. The final model consists of variables with low variance inflation values. The highest variance in inflation value in this model is two. This means that all the variables minus the dependent variable in this regression model are independent. They do not affect any other independent variables, and because of this, we can continue analyzing the model with accurate results. The next thing we did in this regression analysis is to create different models using different selection methods.

Selection Method

There are different selection methods that we used to find the best predictors to include in our model. The different types of variable selection methods are forward selection, backward selection, and stepwise selection. Backward selection removes predictors based on their p values. It removes variables until there are no more predictors to remove. The final model should include the best possible predictors. Stepwise selection removes and enters predictors based on their p values until the final model has no predictors to remove or include. The final model should have the best possible predictors. Forward selection is a selection method that begins with an empty model. Therefore, it adds predictors one variable at a time until the best model is found. Additionally, the final model includes the best predictors utilizing the dependent variable - in our case, it is the log air quality index.

In our data analysis, we used two selection methods to compare models. We used a stepwise selection method, and we used a forward selection method. The reference of the stepwise selection method can be found in appendix M3. The stepwise selection method shows that the model has a .5819 adjusted r squared value. This model is identical to the one used to get rid of collinearity. This method took 5 steps to find the best model; now, we must compare it to the forward selection model to see which of these two is the best. The forward selection model can be referenced in Appendix M4. The forward selection model has an adjusted r squared value of .5849; therefore, it has different predictors than the stepwise selection model. The variable date\_local is kept in the final forward selection model. There are also some differences with some of the same predictors that the models both share. In the forward selection model, the mean of carbon monoxide has a higher negative t value of -22.46 compared to the -8.57t value in the stepwise selection model. It seems that the shared predictors have lower t values in the forward selection model. The reason for this is because the additional predictor is in the forward selection model. Therefore, we settled on using the stepwise selection-based model.

Although the stepwise selection-based model has a slightly lower adjusted r squared value, we think that the other predictors having higher t values benefit the overall accuracy of the model. We did not use the backward selection method because we thought the other two methods would determine a final model. Using these methods to compare the two models has helped us settle on the model used to do an accurate analysis of the data set. It is important to have the best model available because it prevents issues from coming up later during the analysis. After this, we will create more models using these predictors to see which one gives us the best prediction of our dependent variable.

Predicting the y variable

One of the most important things in finding the final model is having a model that can best predict the dependent variable. In our case, that would be the log air quality index variable that we transformed. We first took the air quality index variable and made it into an overall\_AQI variable by adding up the pollutants' mean variables. Then we used log transformation to change the overall\_AQI into log\_AQI. This variable will be our final variable, and it will be used in the final model. Finding the best model that best predicts the dependent variable is important because it helps us avoid biased analysis results. To find the best model, we started by creating models that used different predictors with no collinearity to have an accurate model. Eliminating models that did not do a good job predicting the log air quality index was the first step towards reaching the final model. The final model can be referenced in Appendix M2.

         The model posted in appendix M2 is the final model because of various factors. It has an adjusted r squared value of .5819. This is a good, adjusted r squared value because some variables that affect the air quality index cannot be counted as data. An example would be state laws. Some states may have laws that help fight against pollution, while others may not care about it. This can affect how much pollution comes out of a state, which explains the low adjusted r squared value. The variables chosen in this model best predict the dependent variable because they have a low p-value. They also have low variance inflation values. This means that these predictors are going to predict the best y variable. The predictor that has a great effect on the air quality index is NO2\_ means which is the mean of Nitrogen Dioxide that is in the atmosphere. The second predictor that greatly affects the y variable is SO2\_mean, which is the mean of sulfur dioxide in the atmosphere. The third predictor that greatly affects the y variable is O3\_mean, the mean of ozone in the atmosphere. The pollutant that has the least effect on the y variable is CO\_mean, which is carbon monoxide in the atmosphere. It has a negative t value, and if we had more time to analyze the data set, we could figure out why this predictor has a negative t value. We will be able to predict an accurate air quality index variable over the course of 16 years of data because of this model. The final equation of this model is 1.47811 + 0.00179(State\_Code) + 0.06588(NO2\_Mean) + 3.43851(O3\_Mean) + 0.03864(SO2\_mean) + -0.01680(CO\_Mean). This equation will yield a good prediction for the log air quality index variable.

Lastly, we came up with a model that best predicts the y-variable, and it is the last model used in the analysis. The model that was chosen does not have collinearity; therefore, all the predictors are independent. Additionally, all the predictors are essential to predict the results of the dependent variable. We also did the selection methods to find the best model. All the necessary steps were taken to create a strong linear regression model. These are some of the steps we did while doing this linear regression analysis. It will guarantee that the analysis will not have a biased result, and the findings can be presented with confidence.

Residual Plots

As far as what we are concerned, all the transformations, multicollinearity(M1) issues, and model of fit test (M2*)*are all insignificant without a residual plot analysis. Unfortunately, due to time constraints and misplacement of resources, we cannot provide the code necessary to display proper residual plots. Much of this section will be based on our previous observations and educated speculation. With that said, we will describe residual plots and any problems that would likely occur for reference as to how we approached our dataset in what could have been if the resources were properly allocated.

Our analysis cannot be used without residual plots as we cannot solely understand if our observations are truly random. If any residual plots can form a pattern of shapes varying from straight lines, u shapes, curves, or some trend, we can ultimately say there are issues with the current model. In brief, residual plots should not have the ability to predict the error for any given observation. This clear distinction helps many analysts move forward and realize any discrepancies with their model that they need to address. (However, due to the provided reasons above, any references about official residual plots can be dismissed, but evidence in other areas will be used to substitute the oversight).

Residual plots for our full model (with log transformation) may have been highly inaccurate. We included the multiple predictors that corresponded with each other such as the different chemicals with their max value and a max hour inside the model. According to our model, this caused a multicollinearity issue that we previously described with R Squares of over 0.83 (M1) and the Variance inflation values culminating to over 10 for many variables. We had to get rid of these variables to remove these discrepancies. If the access to the residual plots were there, we would reasonably see a graph with many outliers, and missing values as this original model statement excluded many of the rows with missing values for the max values and max hours as stated by the full model statement output (M2). Due to these issues, we reduce our model with our selection methods, as mentioned previously, with both backward and stepwise selection methods used to finalize the fitted model's predictors.

Our fitted model would become:

*log\_aqi = State\_Code + NO2\_Mean + O3\_Mean + SO2\_Mean - CO2\_Mean.*

With that in mind, the residual plots for our predictors would still have problems. As with analysis comes trials and tribulations to resolve as well. We didn’t realize that even after resolving multicollinearity, much of our data had missing values from this initial model and that removing the problematic predictors had the software include more of our observations, which increased our observations from 436,656 to 1,741,045observations. This was about a 40% increase in observations and sample size that wasn’t accounted for in the data’s initial reporting. However, as the resources are not available to provide accurate residual plots, we can only speculate that our data still had much cleaning to do. Many more observations make us more prone to error and influential points that may skew the data or have faulty results.

We can speculate that our residual plots with assumptions of normality, linearity, variance, and independence could be violated due to straight lines, predictable patterns, or inaccurate reporting from the initial data set because of these apparent issues. If these assumptions were to be violated, we would have to initially take care of any outliers as any transformation may still affect the residual plots and its predictors with the dependent variable. The next section on outliers will cover the process in which we would have gone in cleaning up the data with the massive amounts of outliers that were visible in the dataset.

To resolve issues from any of the assumptions, the outliers first will have to be investigated to see if any of the variables have outstanding values that are skewing other data to have inaccurate residual plots. Then next thing we would attempt like the dependant variable of

log\_aqi

is that we would have to transform the predictors to have a consistent residual plot with random reportings. This would ensure no bias, and there's a random amount of error when displaying the residual plots. Assumptions of independence and normality will need transformations if the plots are curvy or display some pattern. With log transformations, the square or inverse log might resolve the issues. Without the output required to see if these violations occurred, we can only speculate on what issues may have arisen and have been ready to tackle the issues with careful consideration and dissection.

Outliers

As we’ve mentioned, due to limited resources, the cleaning up of our data was not done accurately, and much of the errors within the dataset may still affect our final model later on. However, we can still discuss the impact that these errors will have on our model. This will help us discuss and understand what errors others will have to consider when using our model to predict any AQI values for states based on the dataset provided and inaccurate cleaning done in the dataset.

When importing the dataset, we saw significant red flags that would help us realize that our dataset needed cleaning later in the exploratory data analysis process. This started from our histogram of each chemical's AQI values in the AQI, though later reduced to a single variable of overall\_aqi. With the histogram of overall\_aqi, we can see a significant spike in AQI values in the 10-20 range, which skews our data to the right, having most of the lower end (H1). Though few instances had overall AQI values of over 60, it is still important to mention those areas as they help us realize the problematic regions of pollution and its affection for AQI. Even after transforming the overall\_aqi variables with log, it seemed not to fix all the skewness, but it was considerably better (H2). As we can see in (H2), the histogram has shifted and resembles a normally distributed graph with skewness. With the overall\_aqi variable, there was a skew with a right-leaning tail, and with the transformed dependent variable, we now have a left-leaning tail, but not as severe kurtosis as the initial variable. The only issue that would undermine the histogram transformation is the number of visible outliers when observing. If we compare the two histograms of (H1) and (H2), we can see that the mean AQI value is around 16.671 untransformed and 2.5 transformed. With the transformation, we can see the values tend to settle around the 0 to 4.5 AQI levels specified with the log. This helps us better understand that most states who have higher AQI values are somewhat towards the middle and are not on the extremes as specified in the original variable's representation.

Occurrences of outliers and influential points will affect our model statement throughout the modeling process. It is important to handle them before and after the modeling strategy to ensure no important data points are removed that are essential in computing the model. The use of the SAS software and our decision to choose a dataset with copious amounts of observations (1.7 million to approximate) took a toll on the software in calculating the individual data points and their influence on the entire model. For this reason, we had many troubles in removing or viewing the outliers in full for our dataset because of its size. With proper resources or better sampling methods, we would have the ability to remove outliers that exceeded the -3 to +3 band’s threshold, greatly improving our model and our assumptions for our residual plots, as previously mentioned. Without these removals, we have points in our data that hinder our progress in making an accurate model. As we can see in H3, the extreme values of log\_aqi values greater than 4.8 drastically impede the overall report and analysis. If we had the resources to remove the points, the issue would again arise that more points would pop up and become the new extreme outliers we would have to remove. With what we are given, it seems that it would be a never-ending issue, and for this reason, e=we kept our outliers in the final model and dataset sample. This is also important to keep in because it gives us the real-world implications of AQI and its effects on different states. This is significant because observations that outline where AQI values seem to be higher than normal can be investigated. State and federal governments can keep an eye on those states and regions to ensure they don’t exceed unhealthy amounts of certain chemicals into the atmosphere.

The same can be said about the influential points and could improve our R^2 value from 0.58 to something greater because that vast amount of outliers may tell us that our model may be worse than it may appear with those points. For discussion and having results represent the good of the world and understanding AQI levels from state to state, the influential points are kept in.

Model Improvement

Since we are insufficient in our analysis from residual plots, outliers, and certain regression models, we can assure that our model would be significantly better than what we have described throughout this report. Because of the lack of resources and missing information needed to conduct report-analysis, we had trouble cleaning up the model and fixing any issues with assumptions. For our technical report, we can say that our model is unsatisfactory. With what we have discovered initially, more cleaning and more time can be taken to improve the model by analyzing the residual plots, outliers, and more validation methods to ensure that our model would be effective in larger sample sizes. Otherwise, with the work we have currently done, many people can still use our analysis and findings to find more relationships from interaction terms and other ways to understand the AQI levels in each state.

Within the selected regression model, the relationship that each mean pollutant value had was significant. Amongst each pollutant's measurements (Max value, etc), the mean value had the closest connection in accordance with the overall AQI value that was created. Amongst the various mean values of pollutants, Nitrous oxide (NO2), a pollutant created as a byproduct of agriculture, had the most significant impact on the air quality index. As a result of this, the mean value of nitrous oxide was the strongest predictor, possibly due to the high value of standard deviation and strong impact that the pollutant has on the ecosystem.

Carbon monoxide, a pollutant created via the combustion of fossil fuels, was another pollutant that had a significant impact and correlation to the air quality index. Among the other variables used within the model, sulfur dioxide and ozone, both results of fossil fuel combustion, were also relevant and had a significant impact on the air quality index. The variables included within the model that were not pollutants and used as context references included the date and the state code.

During the 5-fold cross-validation step used to evaluate the predictive power of the model, the original data was split into training and testing sets via the ratio 0.25. For model one, 1306614 entries were used for training while 434431 entries were used for testing. In terms of the overall performance of the two models generated via the 5-fold validation, model two generated more successful ASE test results versus training results while model one did not demonstrate the same. Given that model one did not generate more successful test results, it was not a good sign for the model, given a majority was portioned for training and still did not perform better within testing. Furthermore, model two demonstrated lower F values as well as more accurate overall modeling via the R-Square and Adj R-Sq values. Where model one performed an adj R-sq value of 0.5817, model two had a slightly better performance with an adj R-sq value of 0.5851. Another factor that shows model two’s greater performance was that its variance analysis means square error was slightly lower. In terms of the 5-fold parameter estimates, model two clearly demonstrated lower t-value statistics however had very slightly larger standards of error.

The final linear model is as follows:

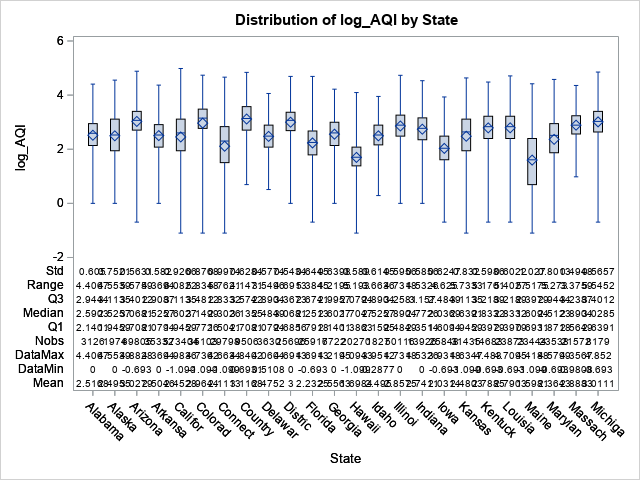
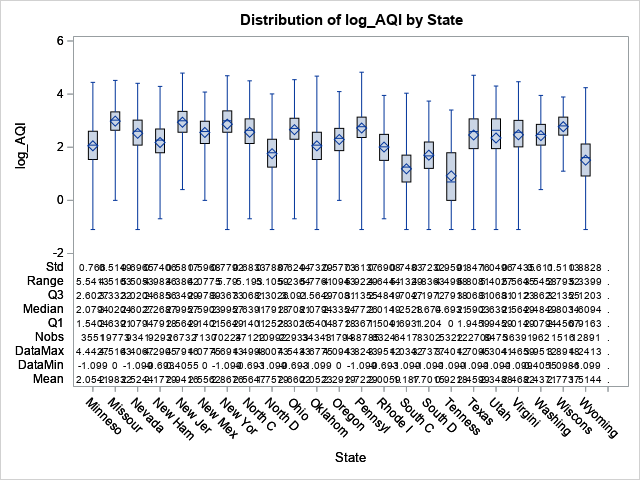
log\_AQI = 1.47811 + 0.00179\*State\_Code + 0.06588\*NO2\_Mean + 3.43851\*O3\_Mean + 0.03864\*SO2\_Mean + -0.01680\*CO\_Mean.

This is the best working model which, given the mean values, can predict the overall AQI to a valid degree of accuracy provided by the p-value. After reviewing the 5 fold, another thing that reveals the model’s potential for predicting is the low level of variance. Removing the outliers would change the resulting accuracy of the model but not enough to accept the null hypothesis. While the null hypothesis was rejected with confidence, the model could have been improved furthermore by removing the mentioned outliers within pollutant variables as well as the overall AQI variable. Further parameter estimates as well as confidence intervals may have had different results provided the log transformation was reversed to the original condition.

In conclusion, while the mean NO2 value had the strongest positive correlation to the generated overall AQI value, the mean O3 value had the most significant impact on the final model. After removing irrelevant variables as well as collinearity and defining the significant variables, the model that was developed has enough accuracy to signify the rejection of the null hypothesis.

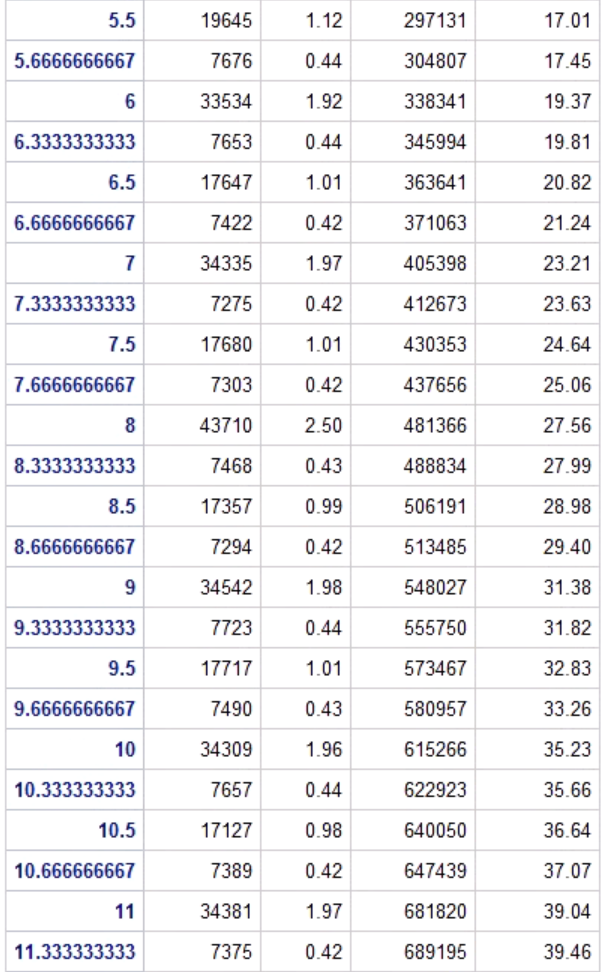
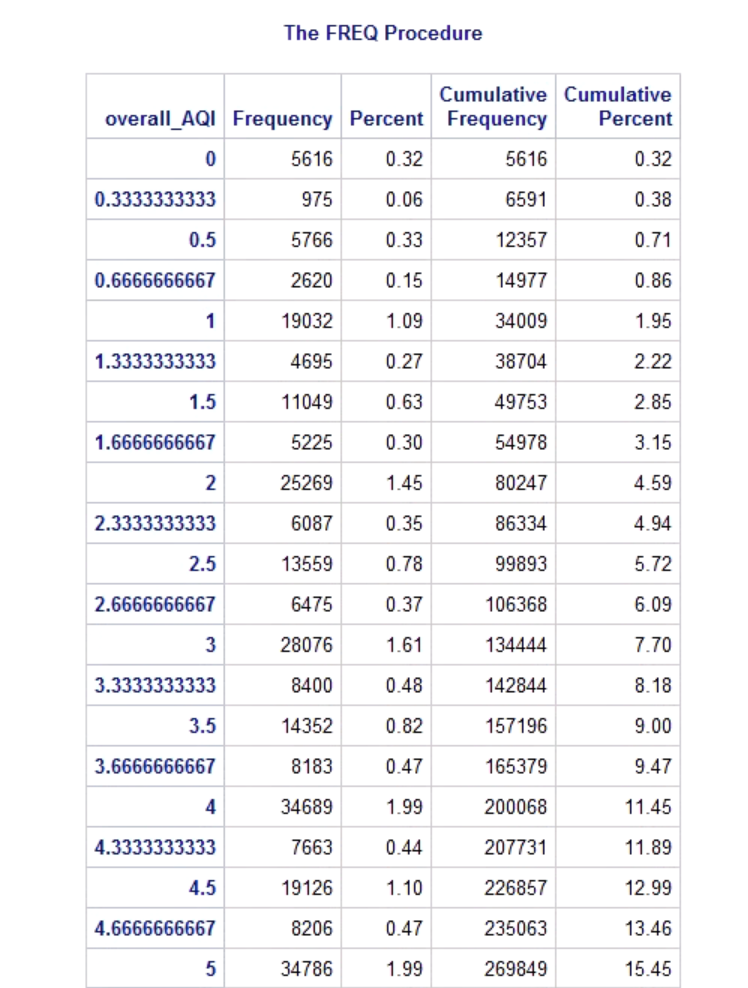
Appendix A

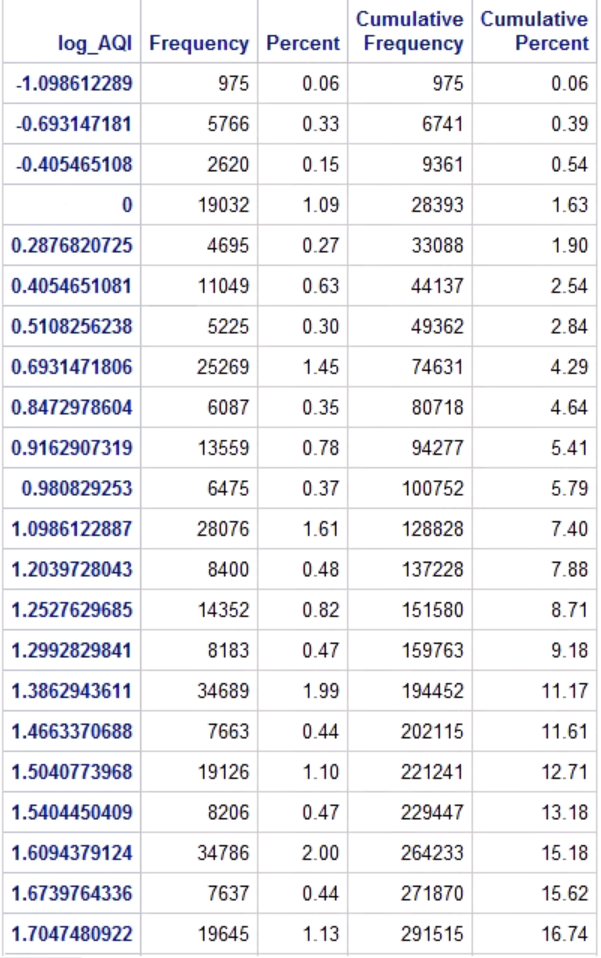
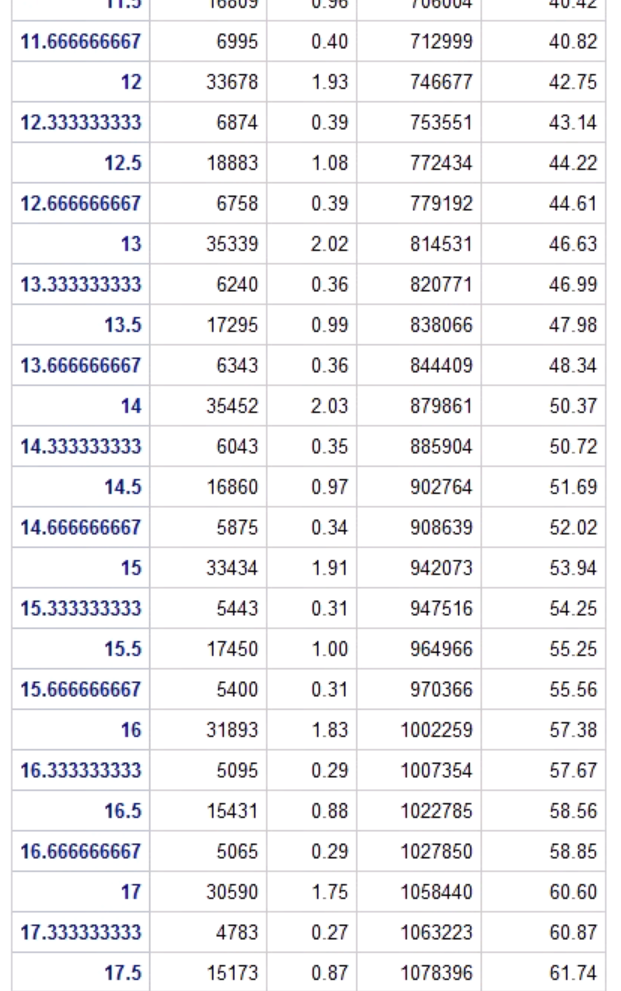
BoxPlot

B1B2

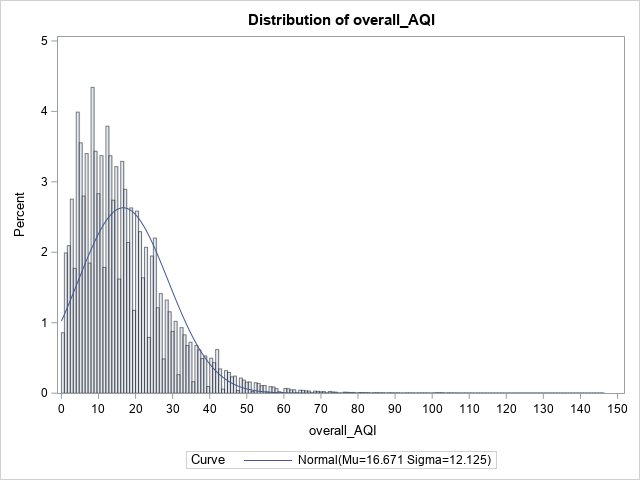
Appendix B

Frequency Table



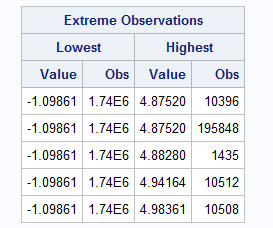


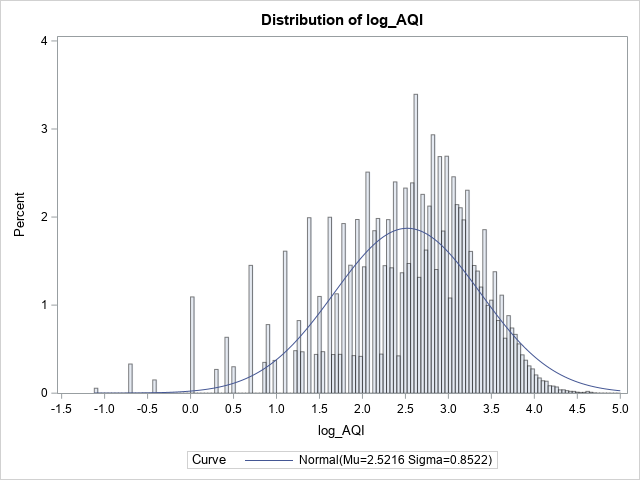
Appendix C

Histogram

H1

H3



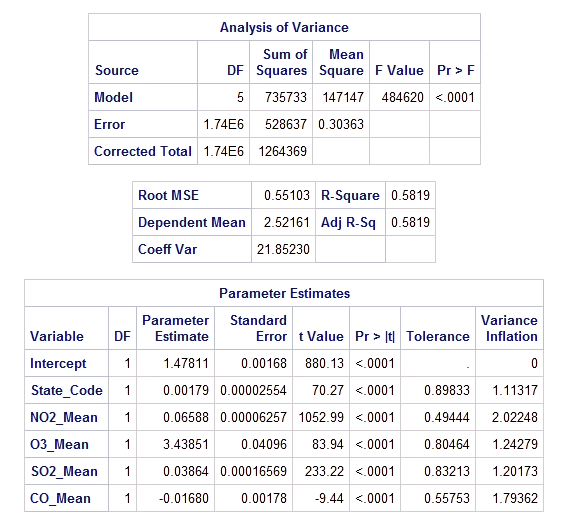


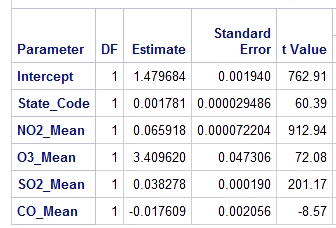
H2

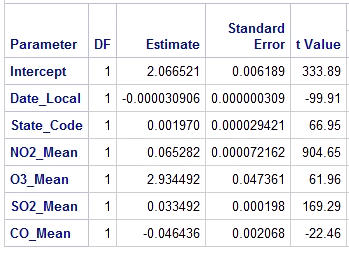
Appendix D

Models

M1 

M2. 

M3 

M4. 

Appendix E

Scatter Plots