Reducing Carbon Monoxide Emission from Gas Turbines

Consulting Group 2

Subin Cho, Chenfei Hu, Jackie Kang, Jonas Reger

INTRODUCTION

Goals

Our goal is to analyze a dataset containing ambient, process, and emission variables from a gas turbine in Turkey to discover what relationships might exist between Carbon Monoxide emission and the other variables. We want to analyze these relationships so we can advise the client on ways they might be able to make modifications to their process variables in order to reduce Carbon Monoxide emission.

Data Description

The dataset is from a study done by Turkish researchers who were interested in developing a predictive emission monitoring system to use on CO and NOx emissions. They analyzed this data to discover useful insights about these emission predictions in hopes to research what variables in the data are appropriate and ecologically valid to justify costly emission monitoring systems. We retrieved the data from the Gas Turbine CO and NOx Emission Data Set, from the UCI Machine Learning Repository. The dataset contains the following variables:

- Ambient Variables
    - Ambient Temperature (AT) °C
    - Ambient Pressure (AP) mbar
    - Ambient Humidity (AH) %
- Process Variables
    - Air Filter Difference Pressure (AFDP) mbar
    - Gas Turbine Exhaust Pressure (GTEP) mbar
    - Turbine Inlet Temperature (TIT) °C
    - Turbine After Temperature (TAT) °C
    - Compressor Discharge Pressure (CDP) mbar
    - Turbine Energy Yield (TEY) MWH
- Emissions Variables
    - Carbon Monoxide (CO) mg/m^3
    - Nitrogen Oxides (NOX) mg/m^3

The data for these variables was collected as hourly averages of sensor measurements in the turbine. There are 7,628 recorded observations spanning over 1 year. The Ambient and Process variables are input measurements, while the Emission variables are the target variables of the study that we obtained the dataset from.

METHODS

Full Data

We first created a model using the given dataset as a whole, which is not divided by the three turbine energy yield ranges. Our goal is to reduce carbon monoxide emissions from gas turbines, and since we want to predict carbon monoxide emissions with other ambient and process variables, we decided to remove the Nitrogen Oxides variable from the full data. This also meets the client's request in the previous meetings. Then, we checked if there were any missing values for data cleaning purposes, but there were no missing values. The types of data for each column are all numeric and consistent as well.

We began the modeling process by comparing a full model that contains all nine predictor variables with a null model that does not contain any of them. The analysis of variance (ANOVA) result showed us that there is a linear relationship between CO and at least some of the predictor variables. Therefore, at least some of the predictors are useful, and the full model is preferred. The R Squared value of the full model was 0.6012, and this means that about 60 % of the response variable CO is explained by the linear relationship with the nine predictor variables. However, among the nine predictor variables, a p-value of AFDP was 0.187, which is too high, and this means AFDP is statistically insignificant in this model and should be removed. After removing AFDP, the Adjusted R Squared value was rarely changed as expected, and the p-values of other variables were all close to 0 (*Figure 1.1*). However, the R Squared tends to favor large models in many cases while we prefer smaller models that do not contain many variables. To avoid this situation, we tried to consider some of the quality criteria. We used the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the stepwise search. Because these quality criteria would implicitly penalize larger models, we thought we could effectively get smaller models. Contrary to our expectations, however, all the three methods did not provide us with a different model.

Then, when we looked more closely at this model that only AFDP was removed, the biggest issue we found was a huge multicollinearity between predictor variables. For example, an approximately 0.99 correlation coefficient between GTEP and CDP can be found in the correlation matrix. This means there is a strong linear relationship between the two variables. Also, we calculated the variance inflation factors

(VIF) for each variable, and the VIF values of six variables including TIT, TEY, GTEP, CDP, TAT, and AT were over 15 (*Figure 1.2*). In practice, a VIF greater than 5 implies a huge multicollinearity issue, so we had to find a way to overcome the problem. We knew that a ridge regression model or principal component regression was a good way to solve the multicollinearity problem, but we could not apply them properly at the level we knew. Instead, we decided to compare the combinations of problematic variables and eliminate them one by one. First, we compared a model without CDP and another model without GTEP because their histograms were very similar, both are related to pressure, and their coefficient of correlation value of 0.99 was the highest. By comparing the two models using the Adjusted R Squared, we decided to remove CDP. Using the same method, we compared TIT and TAT, and eventually TAT was eliminated from the model. The new model without CDP and TAT, however, still had the multicollinearity issue, and it showed that GTEP was statistically insignificant. After removing GTEP from the model, all the problems we had eventually solved.

Another concern was 125 outliers that had a large negative effect on the model. They are called influential observations, and we could detect them using a Cook's Distance. Since they were only about 1.6 % of the full data, we could safely remove the outliers. Also, quadratic transformations were applied to AP, TIT, and TEY, and the Box-Cox transformation applied to the response variable CO (*Figure 1.3).*

Middle Range Data

We created a new dataset that is a subset of the full dataset, including only observations where Turbine Energy Yield (TEY) is between 130-136 MWH. We have two tracks of analysis that we used: 1) Build a model without NOx, and 2) Build a model with NOx. While the client indicated that the NOx variable can be ignored as it wasn't important to their interest in CO emissions, we thought it would be prudent to also build a model with NOx included to see if there might be any benefit in doing so.

Model Without NOx

First, we created a model (*Figure 2.3a*) that includes all available variables except NOx to use as a comparison with other models we develop (i.e. models modified from this full model). This full model performed poorly with an Adjusted R Squared of 0.07948. If there was a meaningful relationship to be found, then there was an issue within the data that made it difficult to build a reliable linear regression model to fit it. We investigated some possible problems that may cause this, the first being multicollinearity since some predictors are highly correlated with each other (*Figures 2.1a-b*). We also checked the model assumptions and found them to be violated (*Figure 2.4a*).

We found that GTEP, TIT, AT, and CDP were highly correlated, which can cause unstable variations in estimating the coefficients of the predictors. We addressed this issue by using the Variance Inflation Factor (VIF) criteria (i.e. VIF > 5-10) to remove the problematic variables (i.e. GTEP, AT, and TIT). The Adjusted R Squared of this new model is 0.07726, which was not an improvement. However, since the removal of those variables, the statistical significance of the remaining variables increased with the exception of AP.

Since the client was interested in simplifying the model to increase interpretability, we used BIC backward selection method to remove variables that may not be useful to the model. This criteria removed TEY and GTEP, and the new model had an Adjusted R Squared value of 0.07975, which is a small improvement from the full model. Before re-addressing the multicollinearity in this BIC model, we chose to investigate possible benefits of transformations on the predictors and the response variable.

We used Box-Cox transformation on the response variable CO and found that the model performance actually suffered from the transformation, which led us to believe that there is a likely possibility that there isn't much meaningful relationship between the predictors and CO. This is something we suspected from our initial examination of pair plots between each variable. We also tried transforming the predictors with quadratic transformation, which we reasoned might be useful since there is a very slight quadratic curve in some of the plots if there is in fact more noise present in the middle range data. This transformation (i.e. quadratic transformation on AFDP, TIT, TAT, and CDP) improved the model performance to Adjusted R Squared of 0.1607 (*Figure 2.3b*). We also found that extending the quadratic transform to all predictors improved the Adjusted R Squared value to 0.1636. There were small but insufficient improvements to satisfying the model assumptions (*Figure 2.4b*). This model includes AT, AP, AH, AFDP, TIT, TAT, and CDP.

We investigated the possibility of outliers by setting a cut-off value using Bonferroni Correction for studentized residuals of the model we built so far, and found 8 possible outliers. We used pair plots to see where these potential outliers were located and found that only 2 were clearly outliers while the others were simply part of the surrounding noise (*Figure 2.2*). We created another dataset that doesn't include these two extreme outliers. We wanted to see how the model might perform over both data sets, but decided that we would leave the decision of removing these observations up to the client and their engineers who are more familiar with the equipment and possible issues they might experience. Additionally, we found that the models that were developed so far, performed very poorly after the outliers were removed. We finally used VIF elimination to remove variables with multicollinearity issues.

Model With NOx

The methodology used for this model is very similar to what we did for the model that doesn't include NOx. We verified that both models share the same outliers so we made no changes to the dataset that doesn't include outliers. We treated the multicollinearity issue with VIF elimination and removed the same variables as before from the new full model (i.e. includes NOx). We also looked at the model diagnostics (*Figure 2.4c*) and saw that the assumptions (i.e. normality, homoscedasticity, and linearity) of linear regression were clearly violated (we looked at them for the model without NOx and found similar results). VIF elimination and removal of outliers only helped a little bit in improving the model and the data, but is not sufficient to trust any of these models.

Our VIF model over the full mid-range data performed well with an Adjusted R Squared value of 0.1164, which led us to believe that NOx is contributing largely to the explanation of variation in CO that the model provides. For this model, we found that it improved without the outliers included (Adj. R^2 = 0.1272). We used a more exhaustive both direction BIC method to find the best model, which only eliminated CDP from the full model (Adj. R^2 = 0.1687). Then we used VIF elimination on the BIC model and removed GTEP and AT (Adj. R^2 = 0.1210). This model also improved when outliers were not included (Adj. R^2 = 0.1476).

Lastly, we performed Box-Cox Transformation on CO, and quadratic transformation on all predictors since that helped improve the model performance. However, removal/inclusion of the outliers and the transformations did not sufficiently help address fundamental issues in the models developed for the mid-range data (i.e. assumptions are still violated and the models are not trustworthy). The final model containing NOX includes AP, AT, AH, AFDP, TIT, and TAT, with an Adj. R Sq. value of 0.2814 (*Figure 2.3c*). Similar to the other final model, the diagnostics didn't improve enough such that the assumptions would not be violated (*Figure 2.4d*)

All Models

We did explore the possibility of using alternative methods such as Ridge/Elastic-Net/LASSO, KNN, Decision Trees, Random Forests, etc. but we didn't have enough time or experience with them to be able to utilize them for this project. A non-parametric model would probably work best with this dataset since it doesn't rely on any assumptions and might be more interpretable in simple cases.

Finally, we wanted to see how much effect each predictor might have on CO emission levels given that all other predictors are held constant at their average values. We evaluated these effects for both final models and included the plots in the appendices section (*Figures 2.5a-b*) for each predictor. Each plot shows the relationship between the change in the target predictor and the predicted CO value. The dashed blue lines indicate the average values of the predictor and predicted CO values. The solid, vertical

orange lines indicate the predictor value that affects the greatest decreasing change in predicted CO levels, with the innermost line representing the greatest change within one standard deviation of the predictor average, and the outermost line representing the greatest change within two standard deviations.

High Range Data

For high range data (TEY > 160), given that we notice the high correlation between NOX and CO, we decided to find two models for high range Data, where one is with the NOX including and one is without the NOX.

Model without NOx

For the model without NOx, we plotted the diagnostics plot *(Figure 3.4)*, the residual VS fitted value plot suggests the distribution of error is not normal and the residual VS leverage plot suggests possible outliers. To solve the non-normality issue, we performed a Box Cox method to perform a square root transformation of the CO *(Figure 3-5)*. To address the possible outlier, we used cook's distance method to inspect the one outlier and check the model fit before and after removing the outlier(Figure 3.6). We decided to remove the outlier because the adjusted r square improved. Then, we checked the correlation between all the variables. According to Figure3-1, there is a multicollinearity issue between predictors TEY & TAT, TEY & CDP, TAT & GTEP, and TAT & CDP (correlation coefficient > 0.8). Then, we performed a t-test of partial correlation of coefficients to check the pattern of multicollinearity. After we located that AT, TIT & CDP are the predictors causing the issue. Next, we removed the problematic variables (AT, TIT & CDP) causing multicollinearity in the model *(Figure 3-1)*. However, the multicollinearity issue is not addressed because TAT still has correlation higher than 0.8 and its change can largely contribute to the change of other predictor variables (*Figure 3-2*). We consider removing TAT in the model but it will lead us to a model with very low r square that means the predictor variables cannot explain the variation of CO emission well and that is not what the client is looking for. A poor model fit cannot help much for the client to either predict the CO emission or find the relationship between CO and predictor variables. Therefore, we decided to include TAT in the model. To further justify and address containing the TAT variable that results in multicollinearity issue in the model, we conducted ridge regression to decrease the model variance and prediction error.

Since we split the data into train and test set during ridge regression, each time we would get a different value of RMSE and r square, which are important indicators to evaluate the model fit, and model coefficients, which explain the relationship between predictor variables and the response variable. To

make the result more reliable, we did a simulation for 2000 times and took the average for all key values to represent the final results (*Figure 3-7*).

Model with NOx

For the dataset having the NOX, we firstly check the relation between the CO & TEY. As the demand of our client that TEY (Turbine Energy Yield) is supposed to add in the model. Given that condition, I think it is necessary to check the relationship between CO and TEY. It turns out that they are not very correlated (the result is in the *Figure 4.1*). After knowing this, we built one full model for the high data range with NOX. The output of R reflected there are only a few predictors that are highly correlated. Under the guide of the result (*Figure 4.2*), we choose the variable AT, GTEP, TAT, NOX. Since the TEY is required by the client, I added it as well into the model, "high_after", which means after first checking. According to *Figure 4.3*, the model "high_after" performanced much better than the full model. Given that the AT's p value in the check is comparatively lower than other predictors, I created a new model (high_after_comp) to compare two models' performances. Based on the results (*Figure 4.4*), the model including AT has a better adjusted R-squared value. However, I then checked the VIF values for both models, the vif value for model high_after_comp is better, much smaller VIF values, which mean much lower multicollinearity issues (*Figure 4.5*). Hence, I decide to keep both models and keep trying to improve their performances.

Besides selecting the most appropriate model, we also found out that we could improve the overall model performance by cleaning the data set. After several functions detected (*Figure 4.6*), we found that there exist some outliers in the high range data set (with NOX). After the whole cleaning process, 4 outliers were eliminated, and by comparing the correlation maps (*Figure 4.7*), it is easy to tell that the improvements in several variables in the adjusted data set (gt_adj).

Back to the model itself, the last method we tried to improve the performance of models is introducing the transformation. For instance, I tried the existing model but with adding transformations to the TAT, NOX. (adding square root transformation to the TAT, and adding quadratic transformation to the TEY). By trying all the combinations, there are 4 new models, which are valueNOP to valueNOP 4 (means value of Nitrogen Oxide performance). In the graph *Figure 4.8*, it can tell how each model differs. In Figure *4.9*, there are the each output of four models, based one the relevant variable values, we select the valueNOP2 as our final model for the high data range with NOX.

RESULTS

Full Data

In our final model, the most effective predictors chosen are AT, AP, AH, TIT, and TEY as we removed a total of four variables including AFEP, GTEP, TAT, and CDP in the process of creating the model.

Final Model using full data is as shown below (*Figure 1.4*):

$$\frac{(\mathbf{CO}^{\lambda}-1)}{\lambda} = \beta_0 + (\beta_1\mathbf{AT}) + (\beta_2\mathbf{AP} + \beta_3\mathbf{AP^2}) + \beta_4\mathbf{AH} + (\beta_5\mathbf{TIT} + \beta_6\mathbf{TIT^2}) + (\beta_7\mathbf{TEY} + \beta_8\mathbf{TEY^2}) + \epsilon$$

λ	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
0.465	- 0.4	0.031	4.71	- 4.55	0.0077	- 108.2	7.29	22.27	- 14.54

The equation means if we want to decrease the carbon monoxide emissions, we have to decrease AT, AH, and TIT because they have positive relationships with carbon monoxide emissions. In other words, if AP and TEY values are increased, the carbon monoxide emissions will increase. Looking at the four diagnostic plots of the first model, we could notice that the assumptions of a linear model were violated, while the diagnostic plots of the final model were significantly improved (*Figure 1.5a and 1.5b).* Also, we can use the histogram of residuals plot to check whether the variance is normally distributed. Since we got a symmetric bell-shaped histogram that its mean is around zero, we are confident to say the normality assumption was not violated as well (*Figure 1.6),* and the results using our final model are valid.

Middle Range Data

We have two final models for this range of the data (i.e. one without NOx and the other with NOx). We will include performance results of the models and predicted effects of the predictors on CO emission for both models.

Model 1 (without NOx) is as shown below (*Figure 2.3b*):

$CO = \boldsymbol{b_0} + \boldsymbol{b_1}\times AT + \boldsymbol{b_2}\times AP + \boldsymbol{b_3}\times AH + \boldsymbol{b_4}\times AFDP + \boldsymbol{b_5}\times AFDP^2 + \boldsymbol{b_6}\times TIT + \boldsymbol{b_7}\times TIT^2 + \boldsymbol{b_8}\times TAT + \boldsymbol{b_9}\times TAT^2 + \boldsymbol{b_{10}}\times CDP + \boldsymbol{b_{11}}\times CDP^2 + \boldsymbol{\varepsilon}$

Model 2 (with NOx) is as shown below (*Figure 2.3c*):

$$\sqrt{CO} = \boldsymbol{b}_0 + \boldsymbol{b}_1 \times AT + \boldsymbol{b}_2 \times AT^2 + \boldsymbol{b}_3 \times AP + \boldsymbol{b}_4 \times AP^2 + \boldsymbol{b}_5 \times AH + \boldsymbol{b}_6 \times AH^2 + \boldsymbol{b}_7 \times AFDP + \boldsymbol{b}_8 \times AFDP^2 + \boldsymbol{b}_9 \times TIT + \boldsymbol{b}_{10} \times TIT^2 + \boldsymbol{b}_{11} \times TAT + \boldsymbol{b}_{12} \times TAT^2 + \boldsymbol{b}_{13} \times NOX + \boldsymbol{b}_{14} \times NOX^2 + \boldsymbol{\varepsilon}$$

We can see that of the two final models, the one including NOx has a higher Adj. R^2 value. Since NOx is an emission variable rather than an input variable, this tells us that this increase in explainability is more due to a relationship between CO and NOx as emissions, which doesn't particularly help us determine what changes to make in the input to decrease CO emissions. However, since there is some relationship between CO and NOx, we investigated how well NOx is related to the other predictors (including and excluding CO) and that those full models performed very in its initial form (i.e. we didn't do any diagnostics or transformations to improve it since it is not necessary to do so at this time).

Finally, we include two tables (shown below) showing the predicted decrease in CO when changing one predictor while keeping all others constant for both models. The plots are shown in the appendices section (*Figures 2.5a-b*). The tables show how much the predicted CO value is changed when changing a predictor by one or two standard deviations from the predictors mean (i.e. sign of direction is indicated by (±1 SE) in the left columns next to the variable names). The direction of change in the predictor values is chosen in favor of decreasing predicted CO levels. For model 1, the most effective variables are TIT, AT, AH, and AFDP. The most effective variables in model 2 are NOX, AT, AH, and TIT.

Model 1		
Predictor (±1 SE)	Within 1 SE of Mean	Within 2 SE of Mean
Ambient Temperature (-7.42)	-0.14 mg/m³	-0.29 mg/m³
Ambient Pressure (-5.79)	-0.04 mg/m³	-0.07 mg/m³
Ambient Humidity (-14.51)	-0.13 mg/m³	-0.27 mg/m³
Air Filter Difference Pressure (+0.61)	-0.11 mg/m³	-0.27 mg/m³
Turbine Inlet Temperature (+3.65)	-0.31 mg/m³	-0.99 mg/m³
Turbine After Temperature (-0.70)	-0.13 mg/m³	-0.19 mg/m³
Compressor Discharge Pressure (-0.17)	0 mg/m³	0 mg/m³

Model 2		
Predictor (±1 SE)	Within 1 SE of Mean	Within 2 SE of Mean
Ambient Temperature (-7.42)	-0.78 mg/m³	-1.3 mg/m³
Ambient Pressure (-5.79)	-0.14 mg/m³	-0.32 mg/m³
Ambient Humidity (-14.51)	-0.56 mg/m³	-1.08 mg/m³
Air Filter Difference Pressure (+0.61)	-0.21 mg/m³	-0.48 mg/m³
Turbine Inlet Temperature (+3.65)	-0.32 mg/m³	-0.75 mg/m³
Turbine After Temperature (-0.70)	-0.04 mg/m³	-0.05 mg/m³
Nitrogen Oxides (-7.28)	-0.68 mg/m³	-1.31 mg/m³

High Range Data

We come up with two models, one with and one without NOx, for the high range data. While it is not common for the firm to have NOX information available when predicting CO emission, the information can be useful if it is available.

Model 1 for high range is shown below.

$$\sqrt{CO} = 47.1870 - 0.005 * \mathbf{AP} - 0.002 * \mathbf{AH} - 0.025 * \mathbf{AFDP} - 0.112 * \mathbf{GTEP} - 0.062 * \mathbf{TAT} - 0.029 * \mathbf{TEY} + \varepsilon$$

The model has a RMSE_train of 0.2882, R square_train of 0.2033 (*Figure 3-7*). The results presented are from the average of 2000 times simulation, which can be reliable. While the standard for a good model is to be high in R square and low in RMSE, the model 1 is decent. All the predictor variables have negative coefficients in the model, which suggests that any increase of the predictor values can lead to a decrease of CO. Given the fact that the coefficients of AP and AH are small, removing them to simply the model can be considered.

To find the most effective variables impacting the change, especially the negative change, of CO, we performed the same analysis for the mid-range that holding other variables constant to examine how one variable change would affect the CO. The result (*Figure 3.8 & 3.9*) shows that predictors TAT and

GTEP have the strongest impact over the change of CO, suggesting they can be the go-to variable for the engineer to control the CO emission.

Model 2 (With NOX) for high is shown below:

$CO = 180.819 + 0.0283 * \mathbf{AT} - 7.123 * \sqrt{\mathbf{TAT}} - 0.374 * \mathbf{GTEP} - 0.0269 * \mathbf{TEY} + 4.085 * \mathbf{NOX} - 0.714 * \mathbf{NOX}^2 + \varepsilon$

With the adjusted R-squared value of 0.3963

The result is in Figure *4.8*. The table below shows how the carbon monoxide changes if there is only one variable change for (+/-) for one or two standard deviations and other predictors are considered as constants. For model 2, the most effective predictors are the TAT, NOX, the somewhat effective predictor is the GTEP, and the least effective predictor is the AT and TEY. Also, in Figure *4.10*, there are 5 specific graphs telling about the more detailed information about this.

Model 2 for high range data		
Predictor (±1 SE)	Within 1 SE of Mean	Within 2 SE of Mean
Ambient Temperature	-0.07 mg/m³	-0.15 mg/m³
Turbine After Temperature	-0.5 mg/m³	-1.01 mg/m³
Gas Turbine Exhaust Pressure	-0.36 mg/m³	-0.73 mg/m³
Turbine Energy Yield	-0.08 mg/m³	-0.16 mg/m³
Nitrogen Oxides	-0.2 mg/m³	-0.43 mg/m³

CONCLUSIONS

Full Data

Through several statistical analyses with the given data, we have succeeded in creating one final model. We first removed AFDP from the full model as it was statistically insignificant, three variables CDP, TAT, and GTEP were removed because of the multicollinearity issue, and then quadratic and a Box

Cox transformations were applied to the model. We found that AT, AP, AH, TIT, and TEY were effective predictors, and with our final model, 79 percent of the data can be explained.

We also regarded 1.6 percent of the data (125 observations) as outliers and excluded them from the modeling process. If we could figure out why and how such data were collected, we might be able to create a better consistent and reliable model that predicts carbon monoxide emissions.

In conclusion, to reduce carbon monoxide emissions, we advise the client to focus on reducing ambient temperature (AT), ambient pressure (AP), and turbine inlet temperature (TIT). Also, by increasing ambient pressure (AP) and turbine energy yield (TEY), the client will be able to see carbon monoxide emissions go down.

Middle Range Data

The middle range data was fairly difficult to fit a model to especially for linear regression, which relies a lot on assumptions that were violated in the data. Due to the poor performance of the models, continued violation of important assumptions, and lack of using non-parametric methods like Decision Tree models, we were unable to produce a good, reliable model. Therefore, we would recommend to the client to avoid using a model that was fitted to this energy yield range, but rather use a model fitted to the full data. We believe that this was a big issue in this data subset because the range was too narrow for a model to fit to any meaningful relationship (i.e. “big picture” model captures the overall relationship, while a narrow range model does not see the overall relationship but rather only sees too much noise). Since these mid-range models are likely not fitting to any meaningful relationships, it may be best to not trust predictions drawn from them. This lack of meaningful relationships between CO and the predictors is intuitively shown in the plots that were used to detect if there were any underlying clustering structures in the data (*Figures 2.6b-c*). This is also reinforced by the clustering plots that shows the clusters for the full data vs Mid-Range data with color hues based on CO levels and TEY levels (*Figure 2.6a*). We can see that the cluster for the Mid-Range TEY values correspond to the same cluster that shows very little variation in CO in the other plot. Comparing full data vs Mid-Range data we see that CO and TEY has a stronger pattern present in the full data, but nothing meaningful in the Mid-Range data.

Despite the fact that the models performed poorly over this dataset and shouldn’t be trusted, we did discover in the clustering plots and in a simple full linear regression model afterwards that NOx actually has a fairly decent relationship with the other predictors (*Figure 2.6c*). Since NOx contributed a lot to the explanation of CO’s variation in the CO models, it may be possible to develop a better model with NOx as the response variable in order to take advantage of the relationship between CO and NOx, which although isn’t very strong to begin with. Through manipulating input variables to decrease NOx, it

may be possible to also decrease CO as a result since the input variables could potentially have similar effects on all or most of the pollutants from the gas turbines.

For the Mid-Range data, we advise the client to use the model that was fitted to the full data rather than using the ones fitted here. Alternatively, the energy range could be expanded to a wider range such that a model would be able to fit more to meaningful relationships than to noise. If the client wants to pursue other options for Mid-Range analysis, our recommendation would be to develop non-parametric models that do not rely on restrictive assumptions, or to investigate whether or not if an NOx response model might help reduce CO emissions as a side effect of reducing NOx emissions.

High Range Data

The main issues when analyzing the data for this range are multicollinearity between predictor variables and the low r square - which means a poor explanation of CO by the predictor variables. To address the multicollinearity issue, we performed a series tests to locate the problematic variables,AT, TIT & CDP, and ridge regression to decrease the prediction error resulting from variable TAT. To address the low r square issue, we include NOx, which is useful in explaining the change of CO, in the model.

Overall, the dataset has a limited sample size which makes it difficult to draw useful conclusions on the modeling. However, for model 1 without NOx, all predictor variables included AP, AH, AFDP, GTEP, TAT, and TEY show negative relationship with CO. The most effective variables to decrease the CO are TAT and GTEP.

For the second model, including the NOX does increase the adjusted R-squared value for a lot. After cleaning the data and specifically variable selection, the VIF value for the final model 2 is comparatively small. To decrease the emission of carbon monoxide, we could simply decrease the Ambient Temperature (AT) and increase the Gas turbine exhaust pressure (GTEP), Total energy yield (TEY), and Nitrogen oxides (NOX). Nonetheless, given that the mode including NOX, which is also one output of the turbine, we think this model should be better as a reference, instead of as the final model representing the whole data set formula.

Final Conclusions

In conclusion, we created several models that can explain a relationship between carbon monoxide emissions and the process variables. Each model may have its own advantages and disadvantages, but when comparing the different statistical test results, we recommend using the final model using full data collected in 2012 to predict carbon monoxide emissions. With this model, to reduce carbon monoxide emissions, ambient temperature, ambient humidity, and turbine inlet temperature need

to be decreased. Also, since ambient pressure and turbine energy yield have a negative relationship with carbon monoxide, they need to be increased.

APPENDICES

Appendix - Section 1 - Full Range

Figure 1.1. Summary results of the model without AFDP.

```
Call:
lm(formula = CO ~ . - AFDP, data = gt)

Residuals:
   Min     1Q Median     3Q    Max
-5.130 -0.593 -0.066  0.394 32.403

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 163.249019   5.656467  28.861  < 2e-16 ***
AT            0.058315   0.009455   6.168 7.28e-10 ***
AP           -0.034578   0.003821  -9.049  < 2e-16 ***
AH            0.016042   0.001604  10.000  < 2e-16 ***
GTEP         -1.929487   0.088607 -21.776  < 2e-16 ***
TIT           0.316610   0.026263  12.055  < 2e-16 ***
TAT          -0.747154   0.037294 -20.034  < 2e-16 ***
TEY          -0.224337   0.024482  -9.164  < 2e-16 ***
CDP           1.284358   0.312703   4.107 4.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.564 on 7619 degrees of freedom
Multiple R-squared:  0.6011, Adjusted R-squared:  0.6006
F-statistic:  1435 on 8 and 7619 DF,  p-value: < 2.2e-16
```

Figure 1.2. Variance Inflation Factor (VIF) graph of all variables

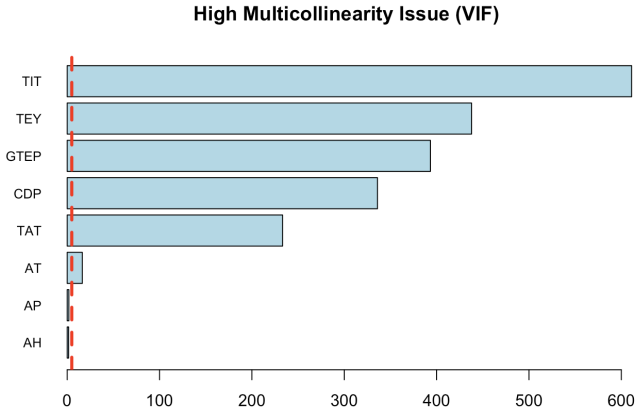


Figure 1.3. Box Cox transformation on CO

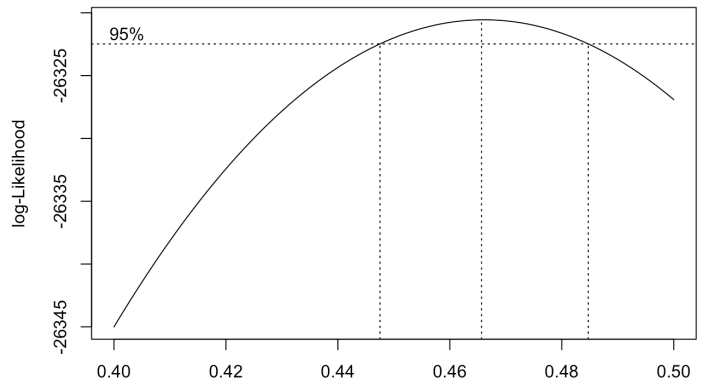


Figure 1.4. Summary results of the final model.

```
Call:
lm(formula = (((CO^lambda) - 1)/lambda) ~ AT + poly(AP, 2) +
    AH + poly(TIT, 2) + poly(TEY, 2), data = gt, subset = cd <=
    4/length(cd))

Residuals:
    Min      1Q  Median      3Q     Max
-2.4218 -0.2781  0.0045  0.2459  2.1904

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.015e-01  5.996e-02  -6.697 2.28e-11 ***
AT             3.109e-02  1.745e-03  17.817  < 2e-16 ***
poly(AP, 2)1   4.710e+00  5.270e-01   8.938  < 2e-16 ***
poly(AP, 2)2  -4.554e+00  5.227e-01  -8.713  < 2e-16 ***
AH             7.744e-03  5.017e-04  15.436  < 2e-16 ***
poly(TIT, 2)1 -1.082e+02  3.648e+00 -29.654  < 2e-16 ***
poly(TIT, 2)2  7.290e+00  8.097e-01   9.003  < 2e-16 ***
poly(TEY, 2)1  2.227e+01  3.535e+00   6.300 3.14e-10 ***
poly(TEY, 2)2 -1.454e+01  1.086e+00 -13.382  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4766 on 7494 degrees of freedom
Multiple R-squared:  0.7892, Adjusted R-squared:  0.789
F-statistic:  3507 on 8 and 7494 DF,  p-value: < 2.2e-16
```

Figure 1.5a. Diagnostic plots of the first model.

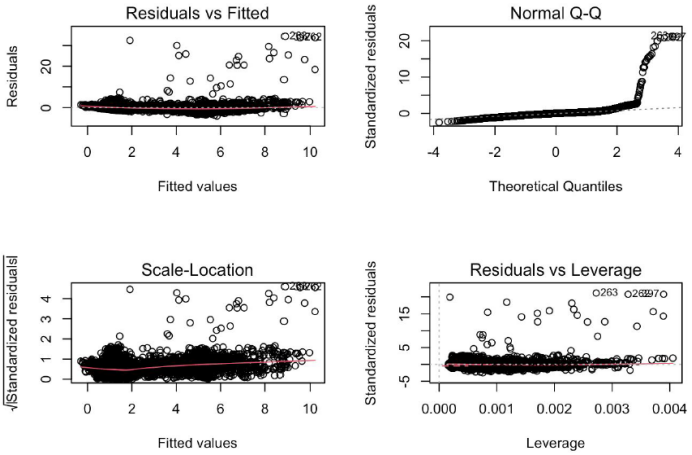


Figure 1.5b. Diagnostic plots of the final model.

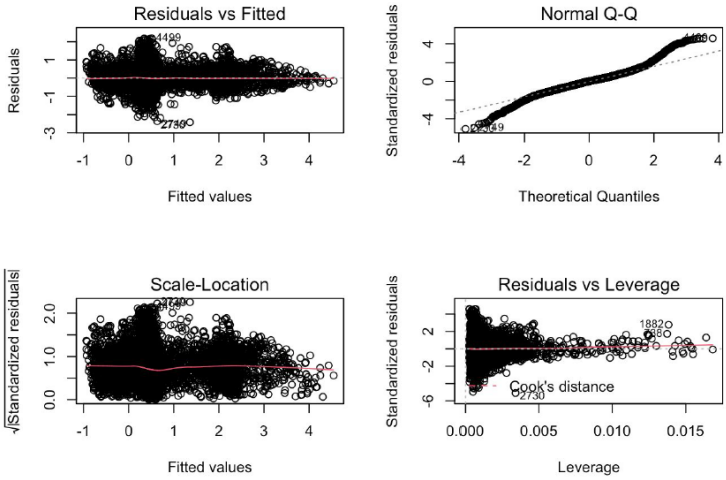
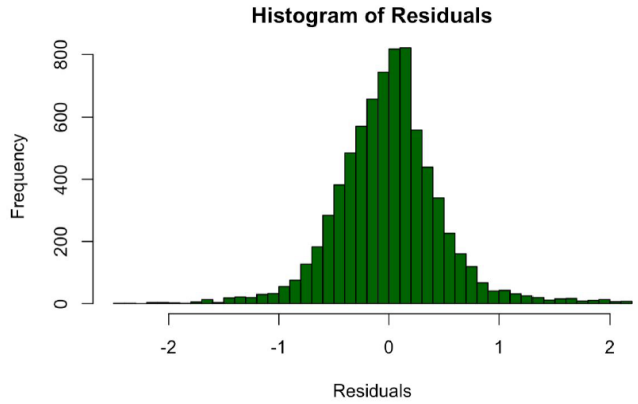


Figure 1.6. The histogram of residuals plot.



Appendix - Section 2 - Mid-Range

Figure 2.1a. Correlation graph of all variables except NOX.

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO
AT	1.00	-0.30	-0.62	0.29	0.95	0.89	0.06	-0.11	0.86	-0.02
AP	-0.30	1.00	0.04	-0.17	-0.30	-0.11	0.00	0.04	-0.13	0.01
AH	-0.62	0.04	1.00	-0.14	-0.58	-0.58	-0.01	0.08	-0.56	0.10
AFDP	0.29	-0.17	-0.14	1.00	0.37	0.29	-0.03	0.00	0.30	-0.05
GTEP	0.95	-0.30	-0.58	0.37	1.00	0.92	-0.03	0.11	0.91	-0.02
TIT	0.89	-0.11	-0.58	0.29	0.92	1.00	0.25	0.23	0.90	-0.09
TAT	0.06	0.00	-0.01	-0.03	-0.03	0.25	1.00	0.05	0.00	-0.25
TEY	-0.11	0.04	0.08	0.00	0.11	0.23	0.05	1.00	0.21	-0.03
CDP	0.86	-0.13	-0.56	0.30	0.91	0.90	0.00	0.21	1.00	-0.01
CO	-0.02	0.01	0.10	-0.05	-0.02	-0.09	-0.25	-0.03	-0.01	1.00

Figure 2.1b. Correlation plot of all variables except NOX.

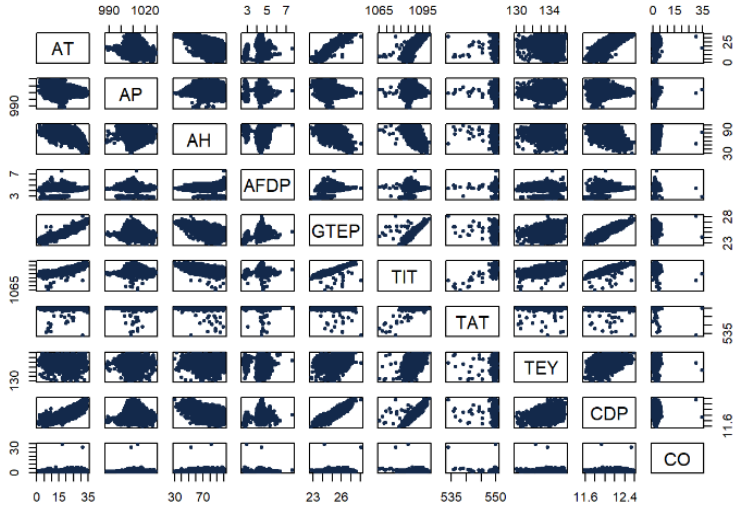


Figure 2.2. Some pair plots between CO and a few predictors to show the extreme outliers.

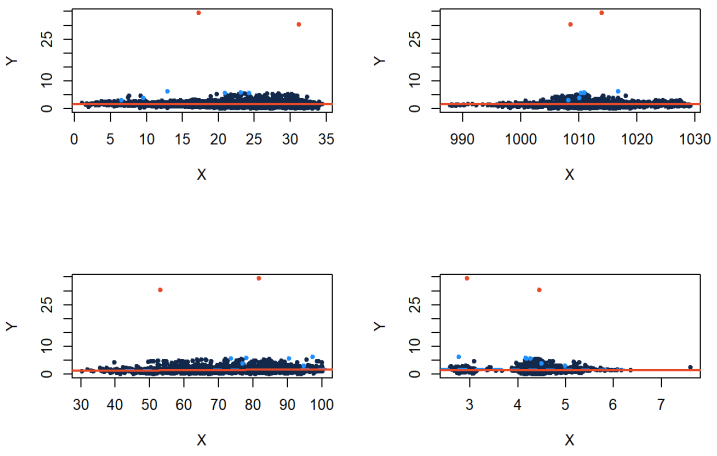


Figure 2.3a. Summary results of the Mid-Range Full Model.

```
Call:
lm(formula = CO ~ ., data = gt_mid)

Residuals:
   Min     1Q Median     3Q    Max
-4.394 -0.369 -0.054  0.224 32.600

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 206.370163  13.635908  15.134  < 2e-16 ***
AT            0.031552   0.012242   2.577  0.00999 **
AP            0.009148   0.003796   2.410  0.01600 *
AH            0.009878   0.001374   7.190 7.74e-13 ***
AFDP         -0.088187   0.029368  -3.003  0.00269 **
GTEP          0.018078   0.124452   0.145  0.88451
TIT          -0.077685   0.026377  -2.945  0.00325 **
TAT          -0.261402   0.042412  -6.163 7.84e-10 ***
TEY           0.029471   0.035248   0.836  0.40315
CDP           0.741095   0.246368   3.008  0.00265 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9409 on 3897 degrees of freedom
Multiple R-squared:  0.0816, Adjusted R-squared:  0.07948
F-statistic: 38.47 on 9 and 3897 DF,  p-value: < 2.2e-16
```

Figure 2.3b. Summary results of the final model without NOX.

```
Call:
lm(formula = CO ~ AT + AP + AH + poly(AFDP, 2) + poly(TIT, 2) +
    poly(TAT, 2) + poly(CDP, 2), data = gt_mid)

Residuals:
   Min     1Q Median     3Q    Max
-6.416 -0.351 -0.056  0.229 32.869

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6.070530   3.042977  -1.995 0.046120 *
AT               0.019646   0.005527   3.555 0.000383 ***
AP               0.006471   0.002927   2.211 0.027091 *
AH               0.009372   0.001354   6.924 5.11e-12 ***
poly(AFDP, 2)1  -3.194292   0.990177  -3.226 0.001266 **
poly(AFDP, 2)2  -2.977301   1.008120  -2.953 0.003163 **
poly(TIT, 2)1   -3.934220   3.440565  -1.143 0.252909
poly(TIT, 2)2  -18.349490   1.603611 -11.443  < 2e-16 ***
poly(TAT, 2)1  -24.983758   1.570303 -15.910  < 2e-16 ***
poly(TAT, 2)2   16.797544   0.950353  17.675  < 2e-16 ***
poly(CDP, 2)1    2.510030   2.581421   0.972 0.330940
poly(CDP, 2)2   12.503072   1.246289  10.032  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8984 on 3895 degrees of freedom
Multiple R-squared:  0.1631, Adjusted R-squared:  0.1607
F-statistic: 69.01 on 11 and 3895 DF,  p-value: < 2.2e-16
```

Figure 2.3c. Summary results of the final model with NOX.

```
Call:
lm(formula = sqrt(CO) ~ poly(AP, 2) + poly(AT, 2) + poly(AH,
    2) + poly(AFDP, 2) + poly(TAT, 2) + poly(TIT, 2) + poly(NOX,
    2), data = gt_mid_all)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2781 -0.1187 -0.0004  0.1107  4.4755

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.233828   0.003852 320.276  < 2e-16 ***
poly(AP, 2)1    2.572193   0.292404   8.797  < 2e-16 ***
poly(AP, 2)2   -1.238311   0.267165  -4.635 3.69e-06 ***
poly(AT, 2)1   21.966603   0.945212  23.240  < 2e-16 ***
poly(AT, 2)2    1.157332   0.378977   3.054  0.00227 **
poly(AH, 2)1   13.891484   0.480146  28.932  < 2e-16 ***
poly(AH, 2)2   -1.269242   0.267685  -4.742 2.20e-06 ***
poly(AFDP, 2)1 -3.095761   0.282293 -10.966  < 2e-16 ***
poly(AFDP, 2)2 -1.686331   0.278209  -6.061 1.48e-09 ***
poly(TAT, 2)1  -4.003558   0.380485 -10.522  < 2e-16 ***
poly(TAT, 2)2   2.487767   0.250173   9.944  < 2e-16 ***
poly(TIT, 2)1  -5.307118   0.690971  -7.681 1.99e-14 ***
poly(TIT, 2)2  -3.200082   0.408515  -7.833 6.08e-15 ***
poly(NOX, 2)1  14.612148   0.497860  29.350  < 2e-16 ***
poly(NOX, 2)2  -3.134226   0.288762 -10.854  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2408 on 3892 degrees of freedom
Multiple R-squared:  0.284,  Adjusted R-squared:  0.2814
F-statistic: 110.3 on 14 and 3892 DF,  p-value: < 2.2e-16
```

Figure 2.4a. Diagnostic plots of the full model without NOX.

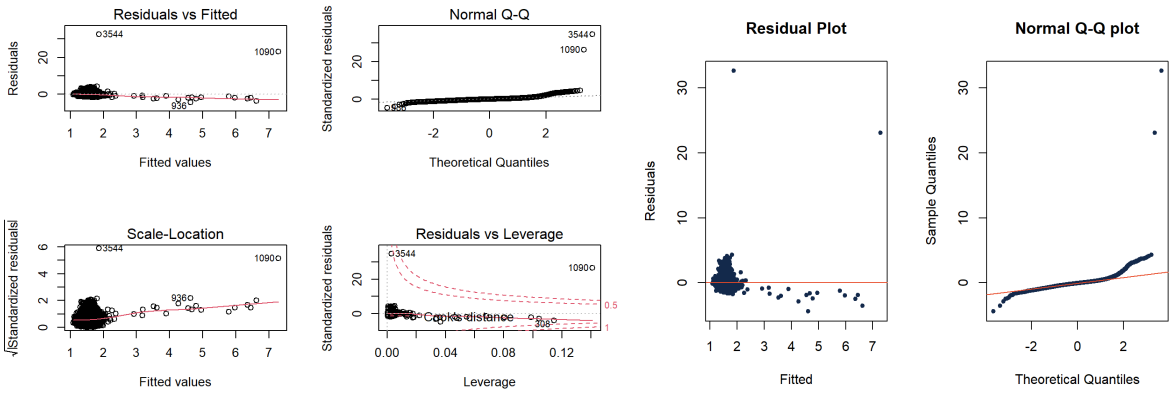


Figure 2.4b. Diagnostic plots of the final model without NOX.

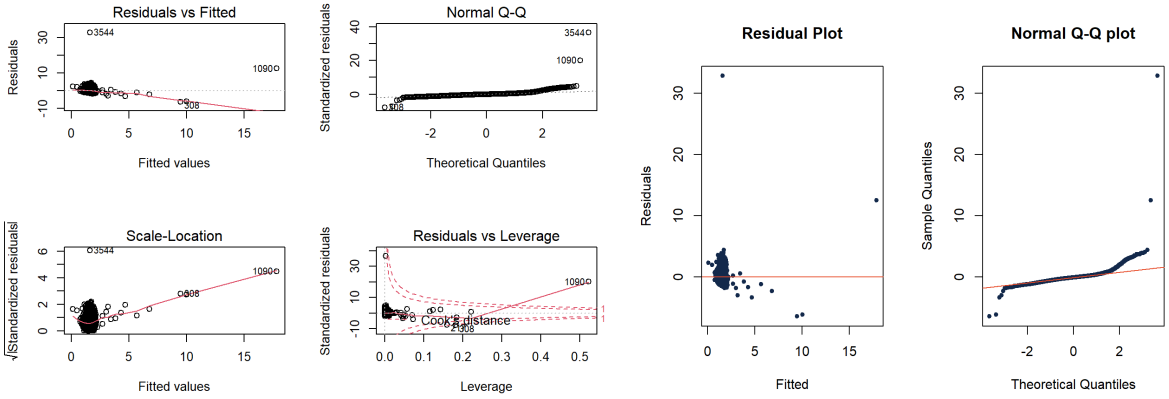


Figure 2.4c. Diagnostic plots of the full model including NOX.

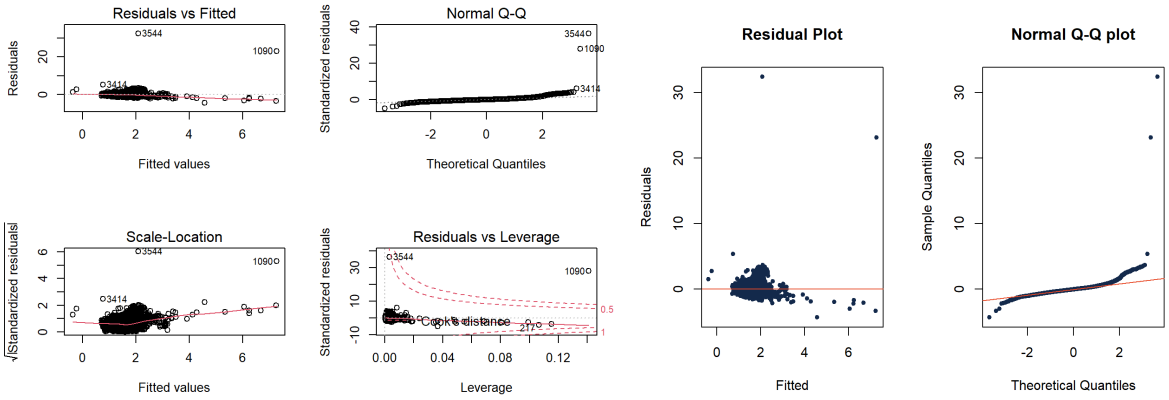


Figure 2.4d. Diagnostic plots of the final model including NOX.

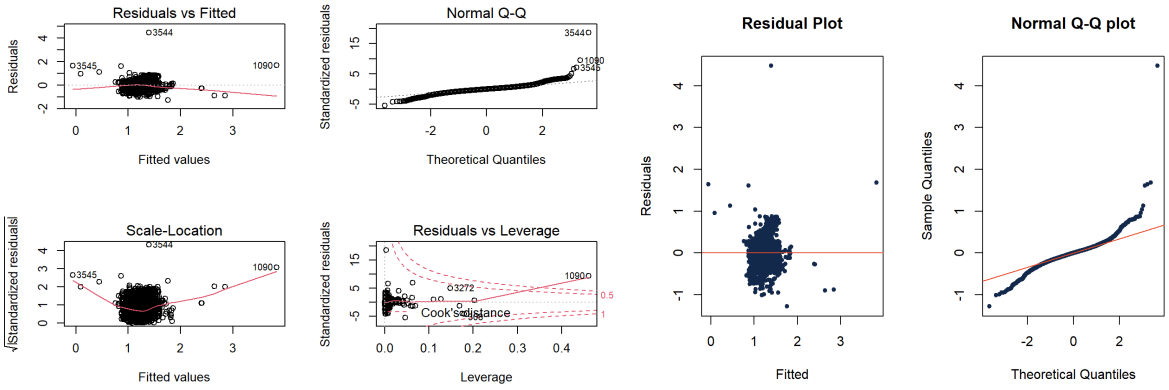


Figure 2.5a. Predicted effect in CO when changing one predictor while holding others constant. (For the final model without NOX).

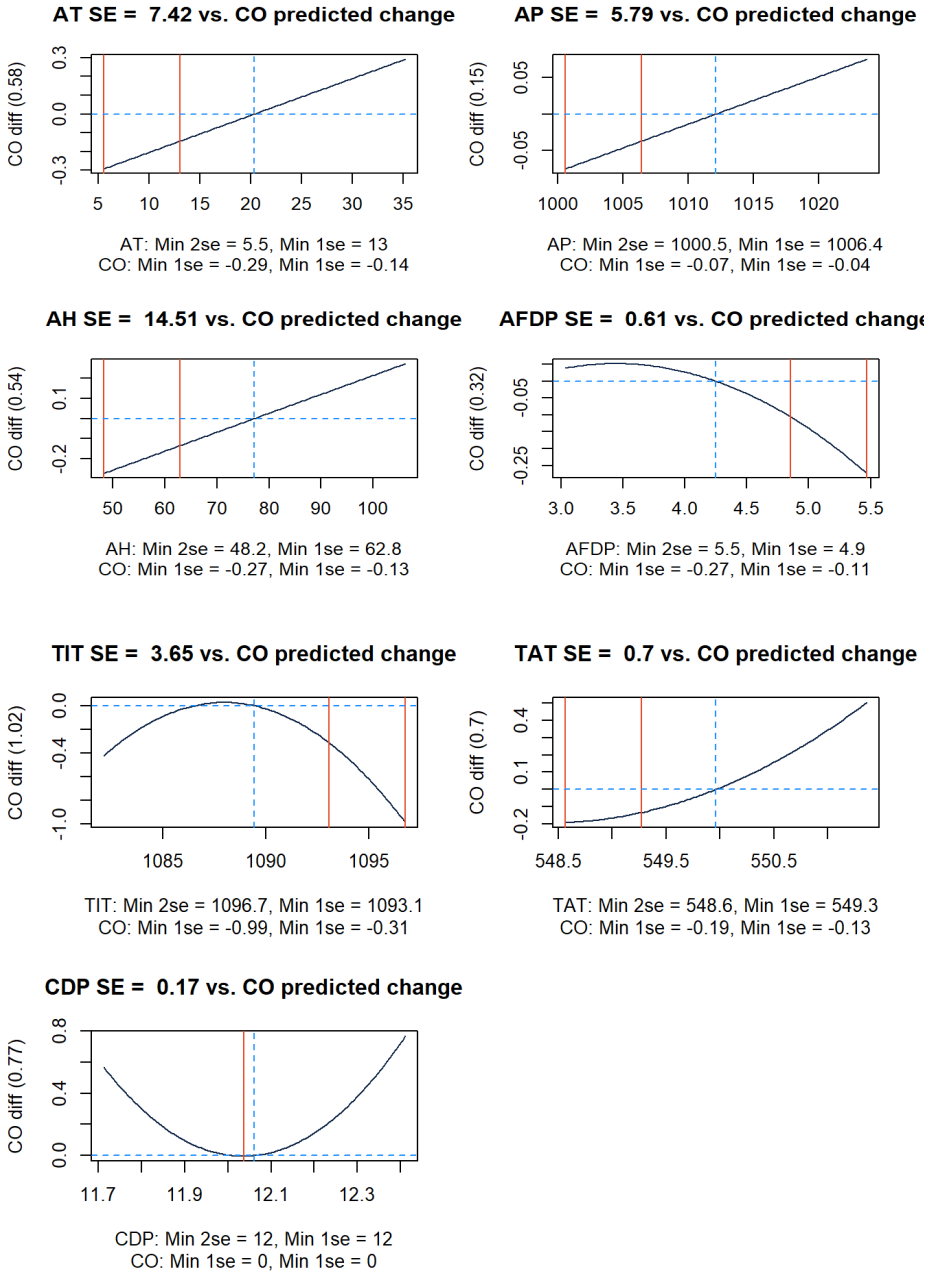


Figure 2.5b. Predicted effect in CO when changing one predictor while holding others constant. (For the final model with NOX).

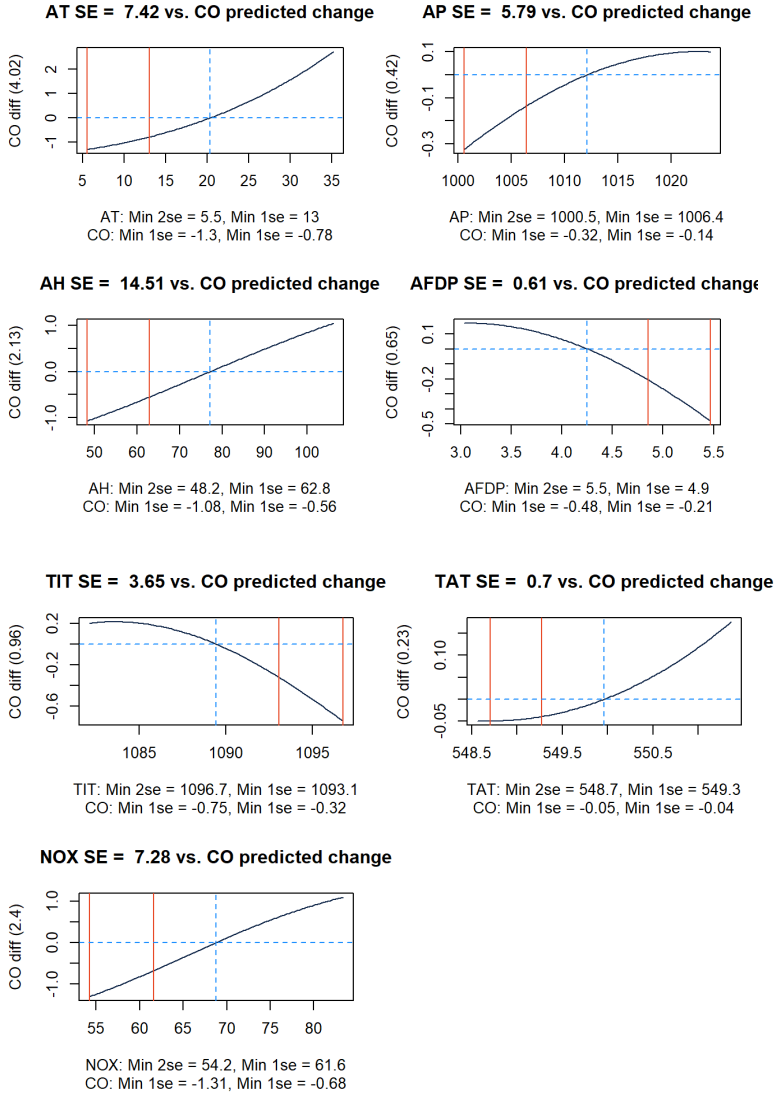


Figure 2.6a. T-SNE plots for cluster analysis. Full Data clusters with color hues based on CO (left) and TEY (right) values. The top two includes all data, while the bottom two exclude outliers.

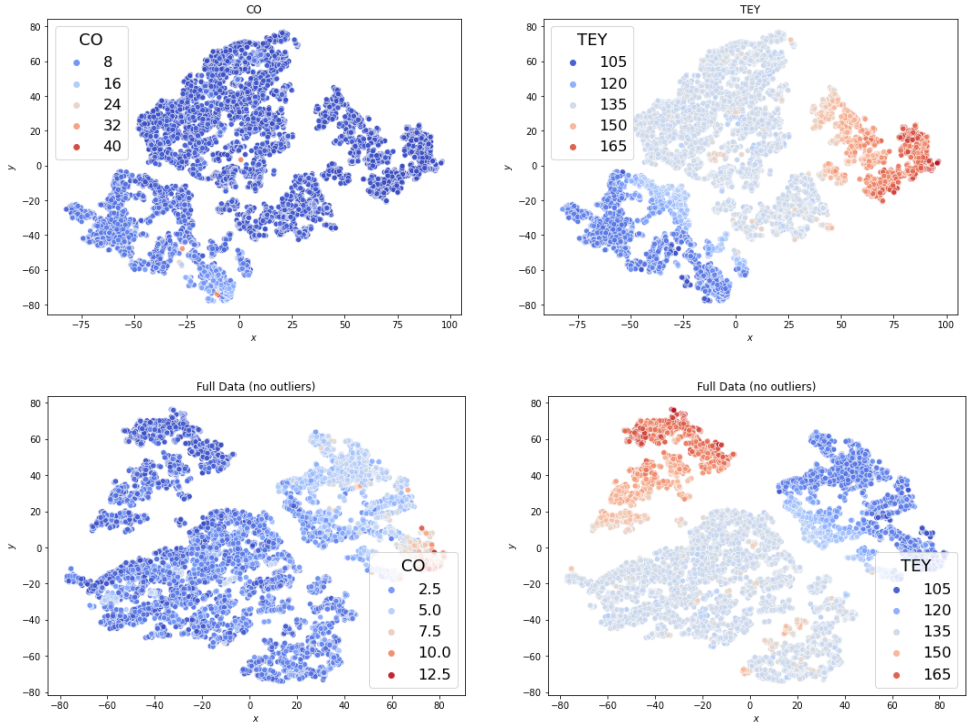


Figure 2.6b. T-SNE plots for cluster analysis. Comparing CO levels in clusters between the full data and Mid-Range data

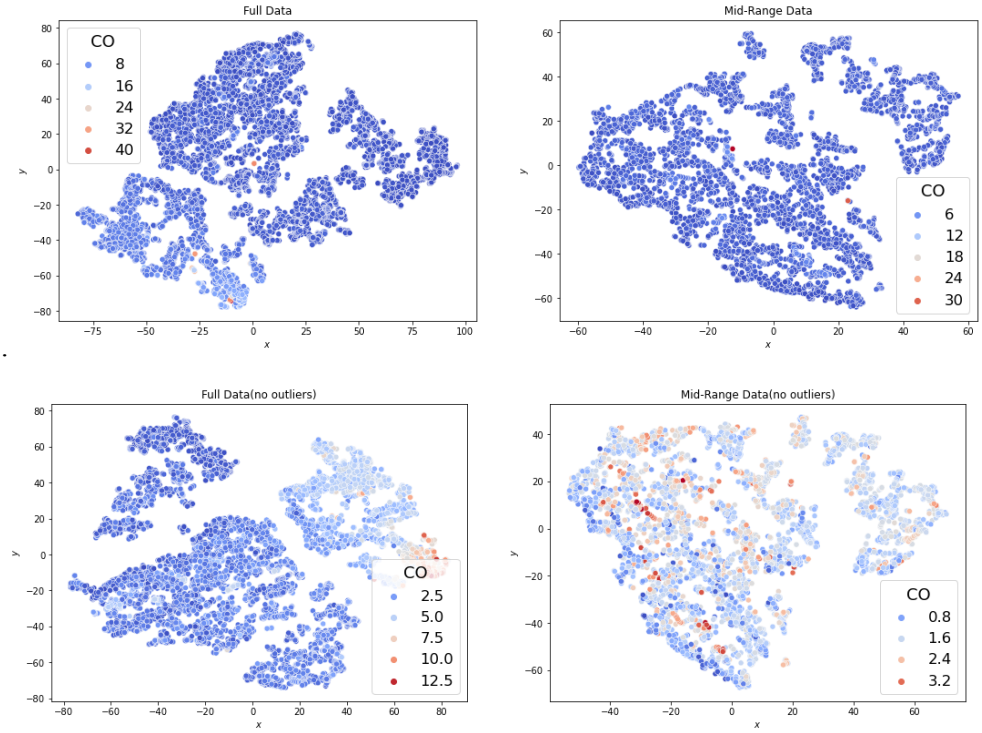
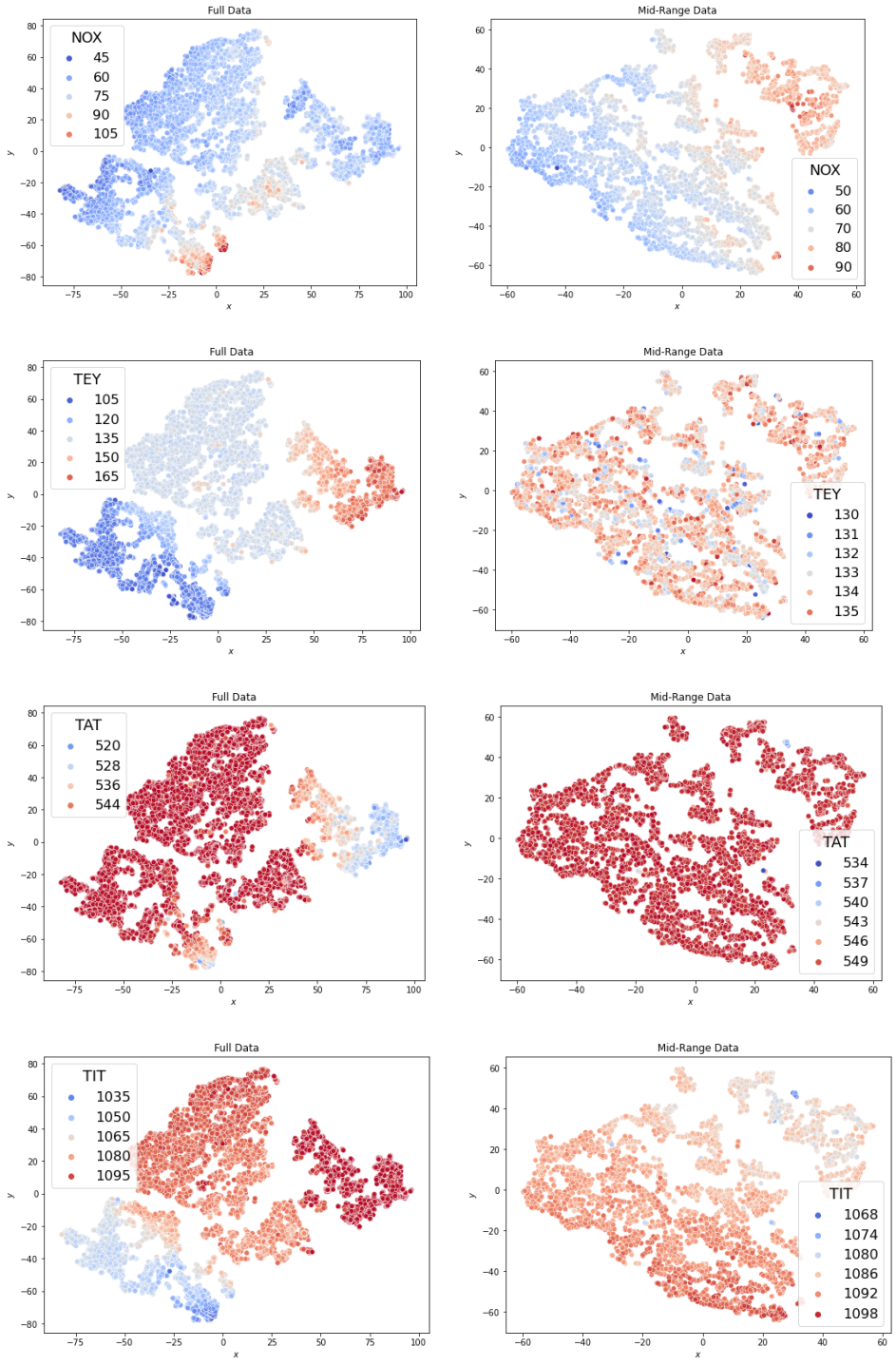
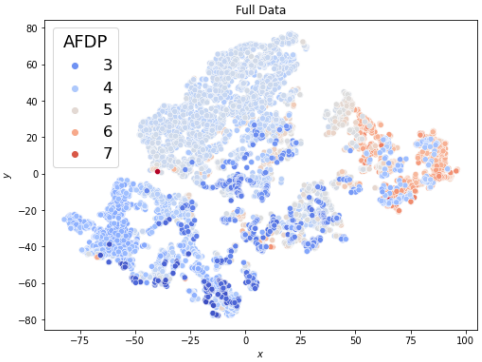
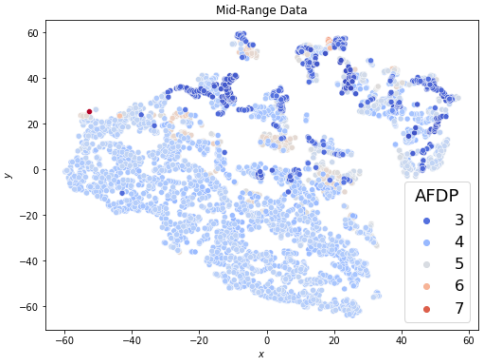
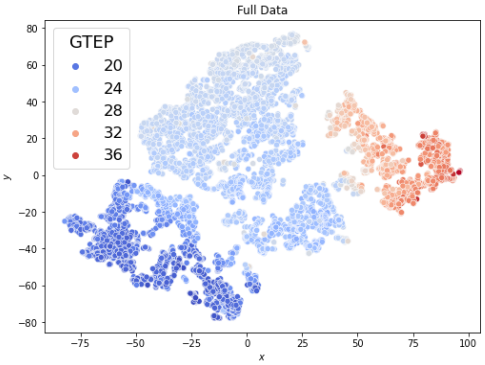


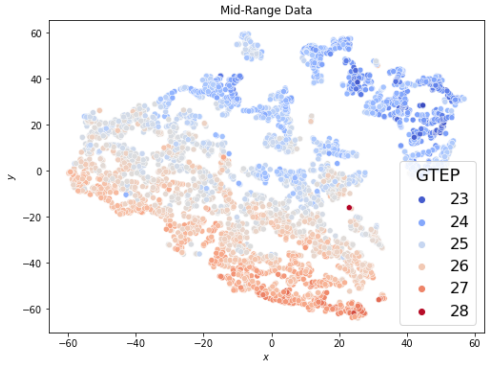
Figure 2.6c. T-SNE plots for cluster analysis. Comparing clusters between full data and Mid-Range data, looking at hues for each variable.

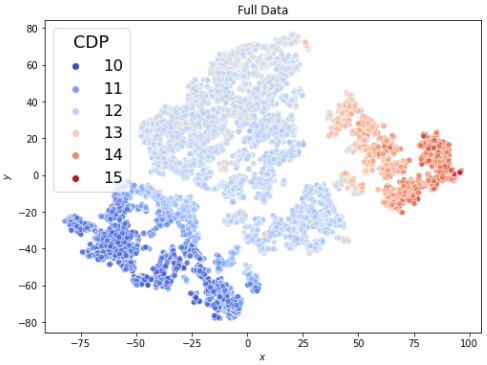


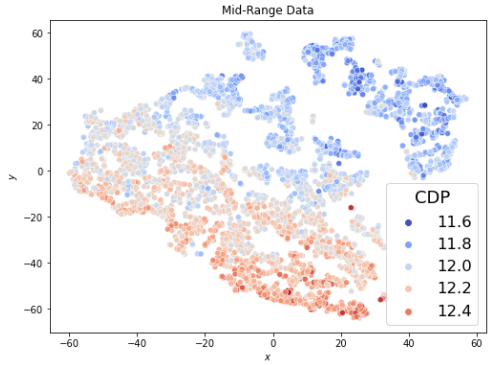

Full Data
AFDP
3
4
5
6
7
x
y

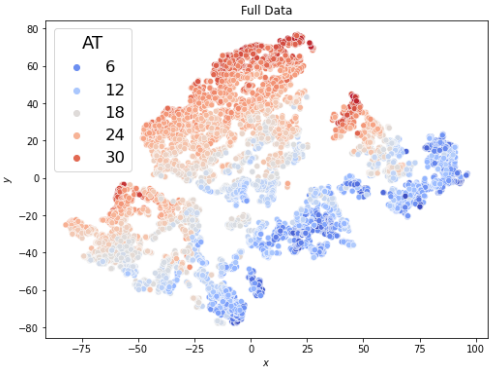

Mid-Range Data
AFDP
3
4
5
6
7
x
y

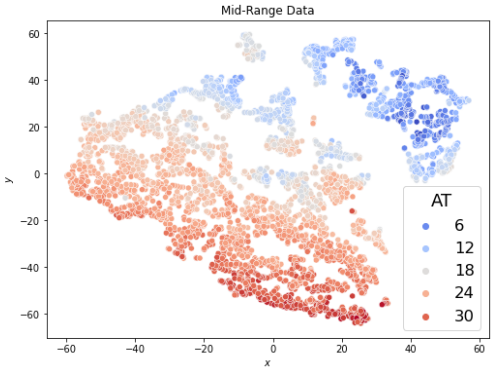

Full Data
GTEP
20
24
28
32
36
x
y

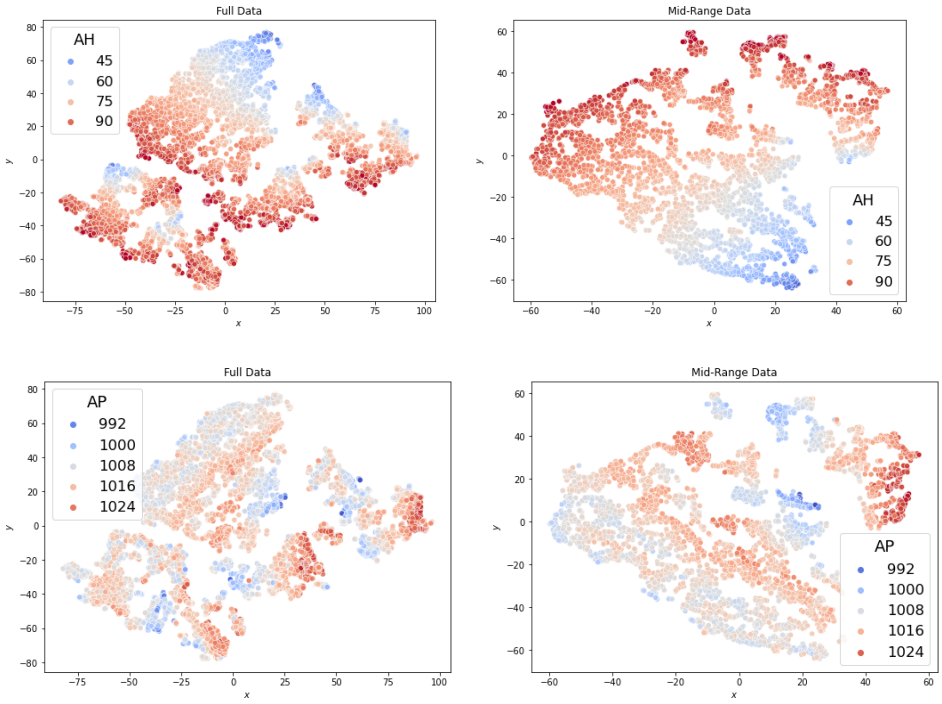

Mid-Range Data
GTEP
23
24
25
26
27
28
x
y


Full Data
CDP
10
11
12
13
14
15
x
y


Mid-Range Data
CDP
11.6
11.8
12.0
12.2
12.4
x
y


Full Data
AT
6
12
18
24
30
x
y


Mid-Range Data
AT
6
12
18
24
30
x
y



Appendix - Section 3 - High Range

Figure 3.1. Correlation Matrix Between All Variables Figure 3.2. Correlation Matrix Between Selected Variables

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO
AT	1.00	0.00	0.01	-0.04	-0.29	0.07	0.39	-0.75	-0.41	0.11
AP	0.00	1.00	-0.28	0.19	0.20	0.03	-0.32	0.09	0.31	-0.02
AH	0.01	-0.28	1.00	-0.08	-0.18	-0.02	0.20	-0.10	-0.22	-0.08
AFDP	-0.04	0.19	-0.08	1.00	0.29	0.10	-0.11	-0.03	0.13	-0.10
GTEP	-0.29	0.20	-0.18	0.29	1.00	0.07	-0.93	0.76	0.91	0.02
TIT	0.07	0.03	-0.02	0.10	0.07	1.00	-0.01	0.03	0.10	-0.08
TAT	0.39	-0.32	0.20	-0.11	-0.93	-0.01	1.00	-0.85	-0.98	-0.12
TEY	-0.75	0.09	-0.10	-0.03	0.76	0.03	-0.85	1.00	0.86	0.01
CDP	-0.41	0.31	-0.22	0.13	0.91	0.10	-0.98	0.86	1.00	0.12
CO	0.11	-0.02	-0.08	-0.10	0.02	-0.08	-0.12	0.01	0.12	1.00

	AP	AH	AFDP	GTEP	TAT	TEY	CO
AP	1.00	-0.28	0.19	0.20	-0.32	0.09	-0.02
AH	-0.28	1.00	-0.08	-0.18	0.20	-0.10	-0.07
AFDP	0.19	-0.08	1.00	0.29	-0.11	-0.04	-0.11
GTEP	0.20	-0.18	0.29	1.00	-0.93	0.76	-0.01
TAT	-0.32	0.20	-0.11	-0.93	1.00	-0.85	-0.09
TEY	0.09	-0.10	-0.04	0.76	-0.85	1.00	-0.01
CO	-0.02	-0.07	-0.11	-0.01	-0.09	-0.01	1.00

Figure 3.3. Summary for Model with all Variables

```
lm(formula = CO ~ ., data = gt_high)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72113 -0.19549 -0.04956  0.14719  1.57464

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 147.983424  48.383749   3.059  0.00233 **
AT           -0.015949   0.014733  -1.083  0.27948
AP           -0.016476   0.002281  -7.222 1.63e-12 ***
AH           -0.001693   0.001105  -1.532  0.12598
AFDP          0.014025   0.021654   0.648  0.51745
GTEP         -0.403361   0.049176  -8.202 1.54e-15 ***
TIT          -0.016053   0.050723  -0.316  0.75175
TAT          -0.170828   0.029834  -5.726 1.66e-08 ***
TEY          -0.108901   0.025412  -4.285 2.14e-05 ***
CDP           0.608197   0.295870   2.056  0.04027 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2976 on 576 degrees of freedom
Multiple R-squared:  0.2054,    Adjusted R-squared:  0.193
F-statistic: 16.54 on 9 and 576 DF,  p-value: < 2.2e-16
```

Figure 3.4. Diagnostics Plots before Transformation

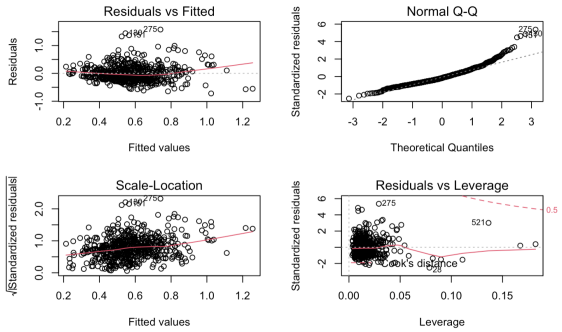
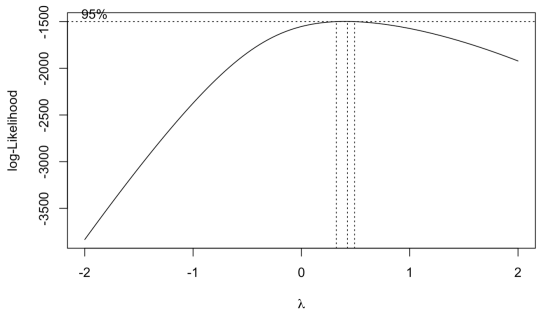


Figure 3.5. Box Cox Method



3.6 - Diagnostic Plots after Transformation

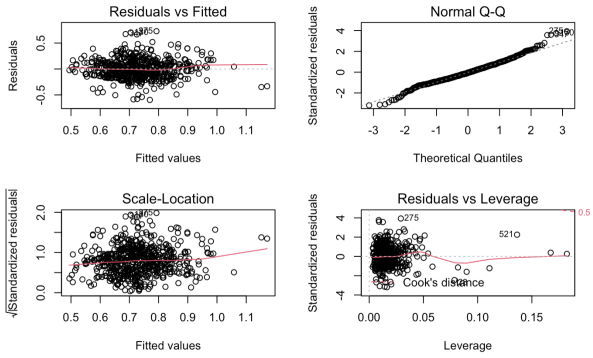


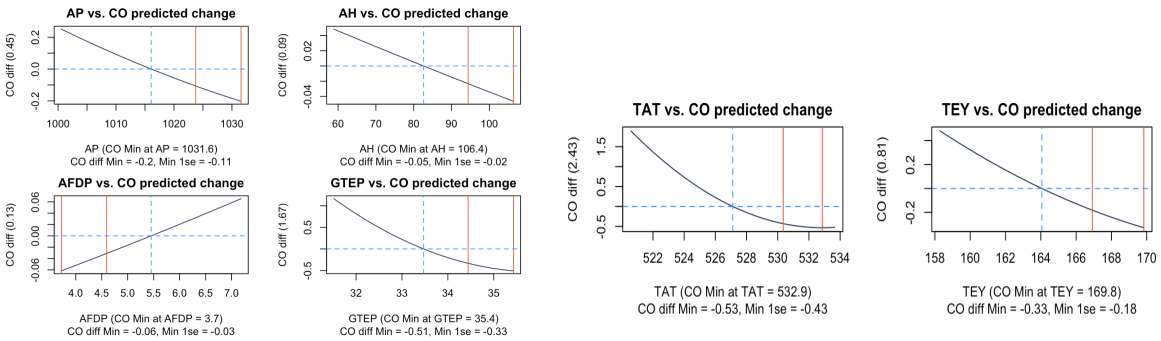
Figure 3.7. Important Values for Modeling

Average Key Values after 2000 Times Simulation			
RMSE_Train	0.2882	AP	-0.005
RMSE_Test	0.2942	AH	-0.002
R^2_Train	0.2033	AFDP	-0.025
R^2_Test	0.1713	GTEP	-0.112
Intercept	47.1870	TAT	-0.062
		TEY	-0.029

Figure 3.8. Predicted effect in CO when changing one predictor while holding others constant(For the model without NOX)

Model 1 - High Range		
Predictor(±1 SE)	Within 1 SE of Mean	Within 2 SE of Mean
Ambient Pressure (-7.51)	-0.11mg/m³	-0.2mg/m³
Ambient Humidity (-11.88)	-0.02mg/m³	-0.05mg/m³
Ambient Filter Difference Pressure(-0.87)	-0.03mg/m³	-0.06mg/m³
Gas Turbine Exhaust Pressure(-1.67)	-0.33mg/m³	-0.51mg/m³
Turbine After Temperature(-3.27)	-0.43mg/m³	-0.53mg/m³
Turbine Energy Yeild (-2.89)	-0.18mg/m³	-0.33mg/m³

Figure 3.9. Predicted effect in CO when changing one predictor while holding others constant (For the model without NOX) - Graphs



Appendix - Section 4 - High Range with NOX

Figure 4.1. The correlation check between CO & TEY

```
Call:
lm(formula = high$CO ~ high$TEY)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56348 -0.23043 -0.08066  0.18641  1.73126

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.376226   0.777435   0.484    0.629
high$TEY    0.001230   0.004738   0.260    0.795

Residual standard error: 0.3315 on 584 degrees of freedom
Multiple R-squared:  0.0001153, Adjusted R-squared:  -0.001597
F-statistic: 0.06735 on 1 and 584 DF,  p-value: 0.7953
```

Figure 4.2. The result of full model in the high range data set with NOX

```
Call:
lm(formula = high$CO ~ ., data = high)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79206 -0.15794 -0.02582  0.13515  1.64989

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.263e+02  4.299e+01   2.938  0.00344 **
AT           3.144e-02  1.362e-02   2.308  0.02133 *
AP           3.320e-04  2.433e-03   0.136  0.89149
AH           3.791e-04  9.951e-04   0.381  0.70340
AFDP         6.139e-03  1.924e-02   0.319  0.74972
GTEP        -3.688e-01  4.375e-02  -8.430 2.80e-16 ***
TIT         -4.101e-02  4.508e-02  -0.910  0.36330
TAT         -1.356e-01  2.664e-02  -5.092 4.81e-07 ***
TEY         -2.867e-02  2.346e-02  -1.222  0.22223
CDP          3.335e-01  2.636e-01   1.265  0.20629
NOX          3.869e-02  3.100e-03  12.478  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2642 on 575 degrees of freedom
Multiple R-squared:  0.3747,    Adjusted R-squared:  0.3639
F-statistic: 34.46 on 10 and 575 DF,  p-value: < 2.2e-16
```

Figure 4.3. The result of selected model in the high range data set with NOX

```
Call:
lm(formula = CO ~ AT + GTEP + TAT + TEY + NOX, data = high)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80322 -0.15740 -0.02388  0.13437  1.65499

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.006349   9.817332  10.289  < 2e-16 ***
AT            0.026667   0.009504   2.806  0.00519 **
GTEP         -0.369155   0.032128 -11.490  < 2e-16 ***
TAT          -0.161579   0.013808 -11.701  < 2e-16 ***
TEY          -0.034642   0.015139  -2.288  0.02248 *
NOX           0.038761   0.002555  15.171  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2635 on 580 degrees of freedom
Multiple R-squared:  0.3725,    Adjusted R-squared:  0.3671
F-statistic: 68.86 on 5 and 580 DF,  p-value: < 2.2e-16
```

Figure 4.4. The comparison check for having the predictor AT or not

```
Call:
lm(formula = CO ~ GTEP + TAT + TEY + NOX, data = high)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92300 -0.16244 -0.02026  0.13248  1.62652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.225419   8.231107  14.120   <2e-16 ***
GTEP         -0.362440   0.032227 -11.246   <2e-16 ***
TAT          -0.178945   0.012416 -14.412   <2e-16 ***
TEY          -0.071521   0.007556  -9.466   <2e-16 ***
NOX           0.038162   0.002561  14.901   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2651 on 581 degrees of freedom
Multiple R-squared:  0.364,     Adjusted R-squared:  0.3596
F-statistic: 83.12 on 4 and 581 DF,  p-value: < 2.2e-16
```

Figure 4.5. VIF check for model high_after, & high_after_comp

```{R}
vif(high_after)
vif(high_after_comp)
```

```
       AT      GTEP       TAT       TEY       NOX 
 5.216518  8.482570 17.390940 16.158659  1.130892 
     GTEP       TAT       TEY       NOX 
 8.435501 13.897087  3.978189  1.122984 
```

Figure 4.6. The process of cleaning the data set

```
data = read.csv('gt_2012.csv', sep = ',')
gt = as_tibble(data)


full = lm(CO ~ ., data = high)
fit1 = update(full, .~.-GTEP-AT-TIT)
critval = qt(0.05/(2*nobs(fit1)), df=df.residual(fit1)-1, lower=FALSE)
gt_adj = gt[-c(which(abs(rstudent(fit1)) > critval)[c(2, 4)]),]
```

Figure 4.7. The Correlation map comparison between the origin data (left) and the cleaned data (right)

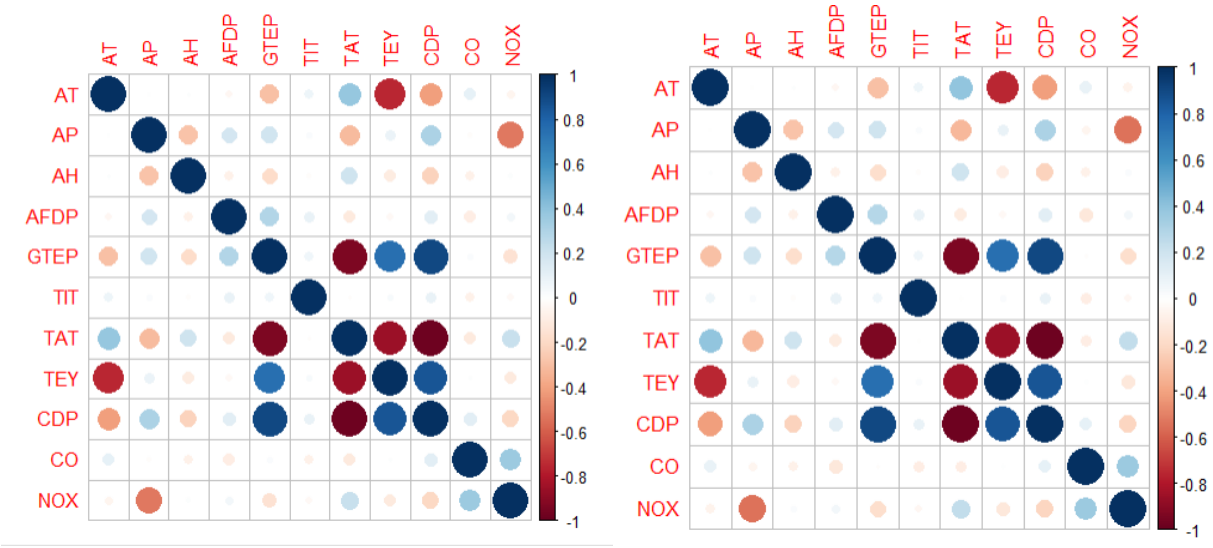


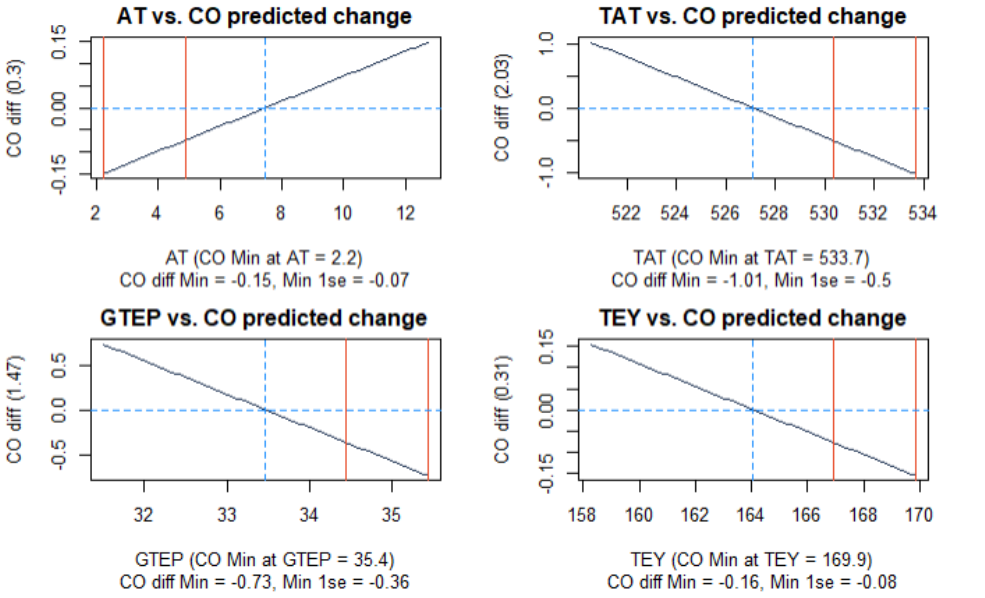
Figure 4.8. The first two model summary

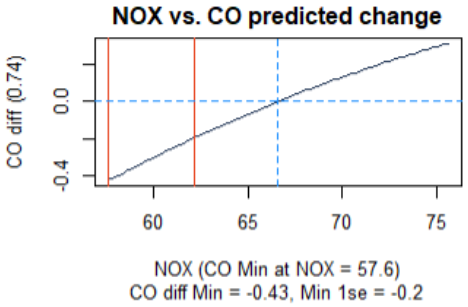
```
Call:
lm(formula = CO ~ AT + sqrt(TAT) + GTEP + TEY + poly(NOX, 2),
    data = high)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79966 -0.15933 -0.02259  0.13319  1.72630

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   190.855611  17.038691  11.201  < 2e-16 ***
AT              0.025693   0.009495   2.706  0.00701 **
sqrt(TAT)      -7.491599   0.631812 -11.857  < 2e-16 ***
GTEP           -0.369310   0.032035 -11.528  < 2e-16 ***
TEY            -0.037241   0.015163  -2.456  0.01434 *
poly(NOX, 2)1   4.249436   0.279401  15.209  < 2e-16 ***
poly(NOX, 2)2  -0.407063   0.264848  -1.537  0.12485
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2629 on 579 degrees of freedom
Multiple R-squared:  0.3764,    Adjusted R-squared:   0.37
F-statistic: 58.26 on 6 and 579 DF,  p-value: < 2.2e-16

          AT      sqrt(TAT)           GTEP            TEY poly(NOX, 2)1 poly(NOX, 2)2
    5.230194      17.416549       8.472497      16.283477       1.129255       1.014683

Call:
lm(formula = CO ~ AT + sqrt(TAT) + GTEP + TEY + poly(NOX, 2),
    data = gt_adj)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74554 -0.14848 -0.02517  0.13295  0.96354

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   180.818747  15.638960  11.562  < 2e-16 ***
AT              0.028311   0.008692   3.257  0.00119 **
sqrt(TAT)      -7.122840   0.579959 -12.282  < 2e-16 ***
GTEP           -0.373942   0.029338 -12.746  < 2e-16 ***
TEY            -0.026898   0.013911  -1.934  0.05366 .
poly(NOX, 2)1   4.084995   0.256902  15.901  < 2e-16 ***
poly(NOX, 2)2  -0.714368   0.242039  -2.951  0.00329 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2405 on 575 degrees of freedom
Multiple R-squared:  0.4026,    Adjusted R-squared:  0.3963
F-statistic: 64.58 on 6 and 575 DF,  p-value: < 2.2e-16

          AT      sqrt(TAT)           GTEP            TEY poly(NOX, 2)1 poly(NOX, 2)2
    5.222382      17.261219       8.317461      16.254813       1.141177       1.012949
```

Figure 4.9. The second two model summary

```
Call:
lm(formula = CO ~ AT + TAT + GTEP + TEY + NOX, data = gt_adj)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75169 -0.15135 -0.02359  0.13899  0.94204

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 94.468467   9.060605  10.426  < 2e-16 ***
AT           0.029690   0.008752   3.392  0.00074 ***
TAT         -0.152470   0.012747 -11.961  < 2e-16 ***
GTEP        -0.373742   0.029593 -12.630  < 2e-16 ***
TEY         -0.022887   0.013975  -1.638  0.10204
NOX          0.037717   0.002396  15.744  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2424 on 576 degrees of freedom
Multiple R-squared:  0.392,     Adjusted R-squared:  0.3867
F-statistic: 74.27 on 5 and 576 DF,  p-value: < 2.2e-16

       AT       TAT      GTEP       TEY       NOX
 5.211235 17.237638  8.329119 16.146096  1.142786

Call:
lm(formula = CO ~ sqrt(TAT) + GTEP + TEY + poly(NOX, 2), data = gt_adj)

Residuals:
     Min       1Q   Median       3Q      Max
-0.87184 -0.14954 -0.01647  0.13622  1.04845

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   206.721550  13.577802  15.225  < 2e-16 ***
sqrt(TAT)      -7.971554   0.522429 -15.259  < 2e-16 ***
GTEP           -0.366794   0.029499 -12.434  < 2e-16 ***
TEY            -0.066179   0.006992  -9.466  < 2e-16 ***
poly(NOX, 2)1   4.012350   0.258058  15.548  < 2e-16 ***
poly(NOX, 2)2  -0.748359   0.243822  -3.069  0.00225 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2425 on 576 degrees of freedom
Multiple R-squared:  0.3916,    Adjusted R-squared:  0.3863
F-statistic: 74.14 on 5 and 576 DF,  p-value: < 2.2e-16

    sqrt(TAT)          GTEP           TEY poly(NOX, 2)1 poly(NOX, 2)2
    13.776765      8.270923      4.038397      1.132576      1.011065
```

Figure 4.10. The graphs of specific explanation for the table in the results part





Contributions

All members contributed equally to the project as a whole.

Code & Data

- The dataset is in the *gt_2012.csv* file.
- The supporting code for the All-range data analysis is in the *Figure_AllRange.Rmd* file.
- The supporting code for the Mid-Range data analysis is in the *Figure_MidRange.Rmd* file and the *Figure_MidRange_Cluster.ipynb* file.
- The supporting code for the High-Range data (without NOX) analysis is in the *Figure_HighRange.Rmd* file.
- The supporting code for the High_Range data (with NOX) analysis is in the *Figure_HighRange_NOX.Rmd* file.