Final Turbine Project Report
Prepared for Man Fung Leung

By Kaleb Jordan, Hyo Min Yoo, Jessica Gong and Sangjun Ko

Table of Contents

Introduction

Our client Man Fung Leung tasked our team with finding ways to reduce CO emissions in their turbine by understanding the relationship of specific turbine variables with CO. After receiving the data, three different cases were identified: Case 1 is the full data provided by the client, Case 2 is typical Total Energy Yield (TEY) range between 130 to 136 MWh and Case 3 is when TEY is greater than 160 MWh. The questions we wish to answer are:

1) Which variables have the strongest relationships with CO for each of the three cases?
2) How can we see the nature of these relationships through visualizations and modeling?

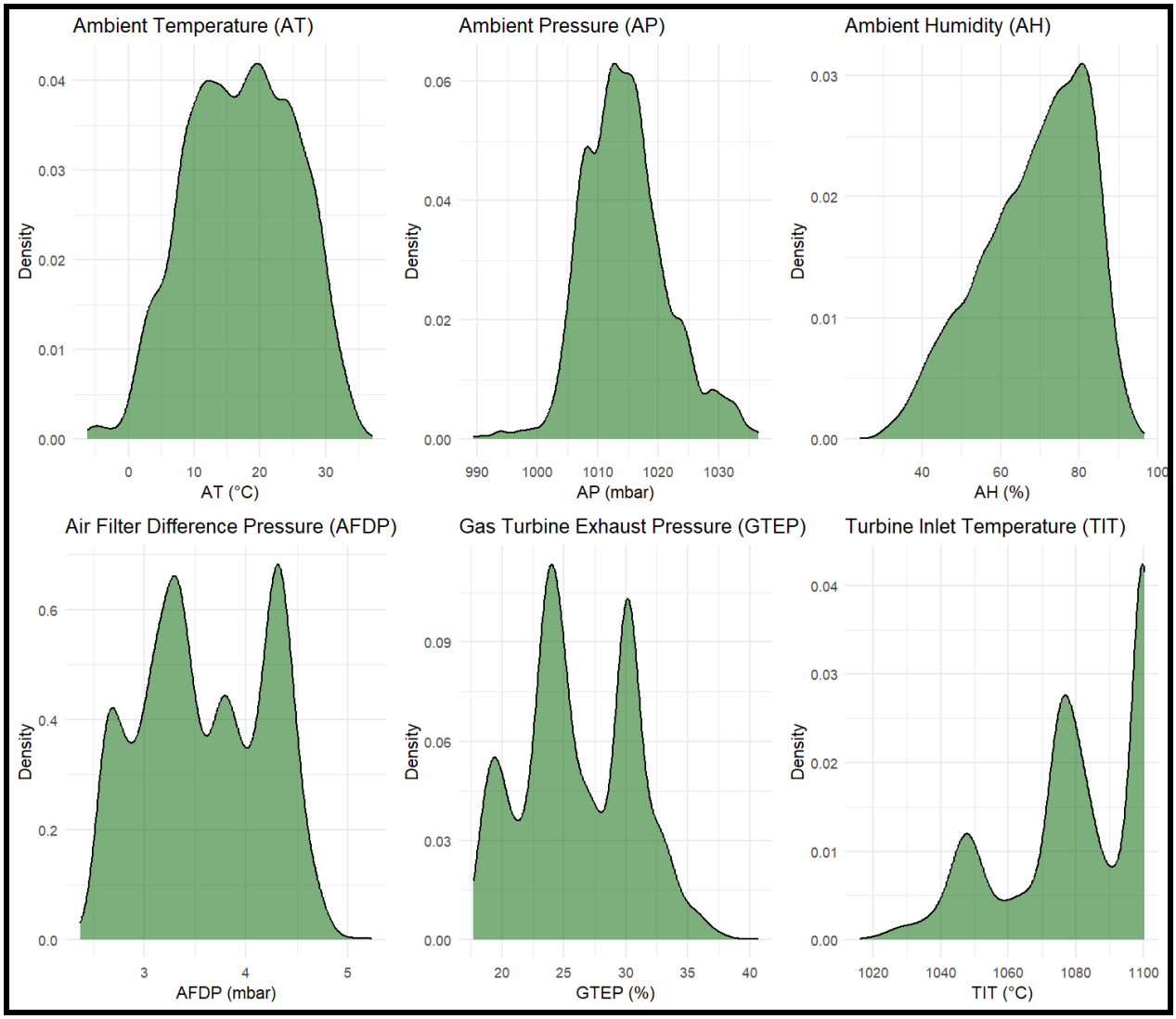
The data provided by the client included 7,384 observations with 11 continuous variables. Case 1 contains all 7,384 observations, Case 2 has 1,701 and Case 3 consists of 418. Among the 11 variables, there consisted of three ambient variables (Ambient Temperature, Ambient Humidity, and Ambient Pressure). Ambient variables cannot be adjusted but still may play an integral role in understanding CO emissions. Refer to Appendix A for more information on the variables.

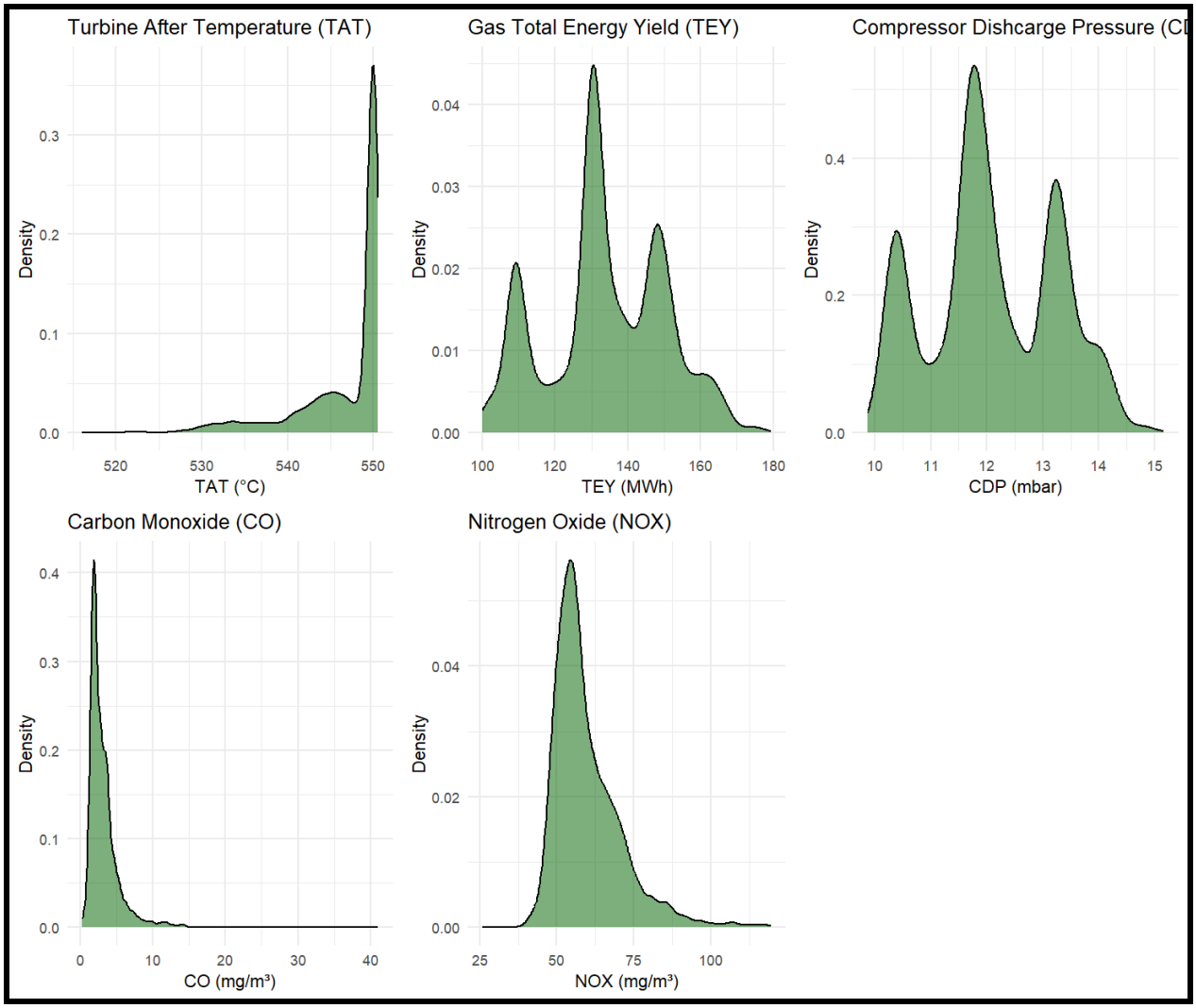
FIGURE 1: TABLE OF VARIABLES

Continuous Variables	Functions
Ambient Humidity	The amount of moisture in the air in a specific environment. It cannot be controlled.
Ambient Pressure	The pressure of the air in a specific environment. It cannot be controlled
Ambient Temperature	The temperature of a specific environment. It cannot be controlled.
Air Filter Difference Pressure	Measures the difference in pressure between the clean and dirty sides of the collector.
Gas Turbine Exhaust Pressure	The pressure of the gasses as they exit the turbine after combustion.
Turbine Inlet Temperature	The temperature of the gasses entering the turbine after being heated during combustion.
Turbine After Temperature	The temperature of gasses as they exit the

	turbine
Compressor Discharge Pressure	The pressure of air after being compressed and is ready to enter the combustion chamber.
Turbine Energy Yield	The energy output of a gas turbine over a specified time period.
Carbon Monoxide	The key pollutant to monitor, as its emissions are a direct indicator of combustion efficiency. It is our target variable.

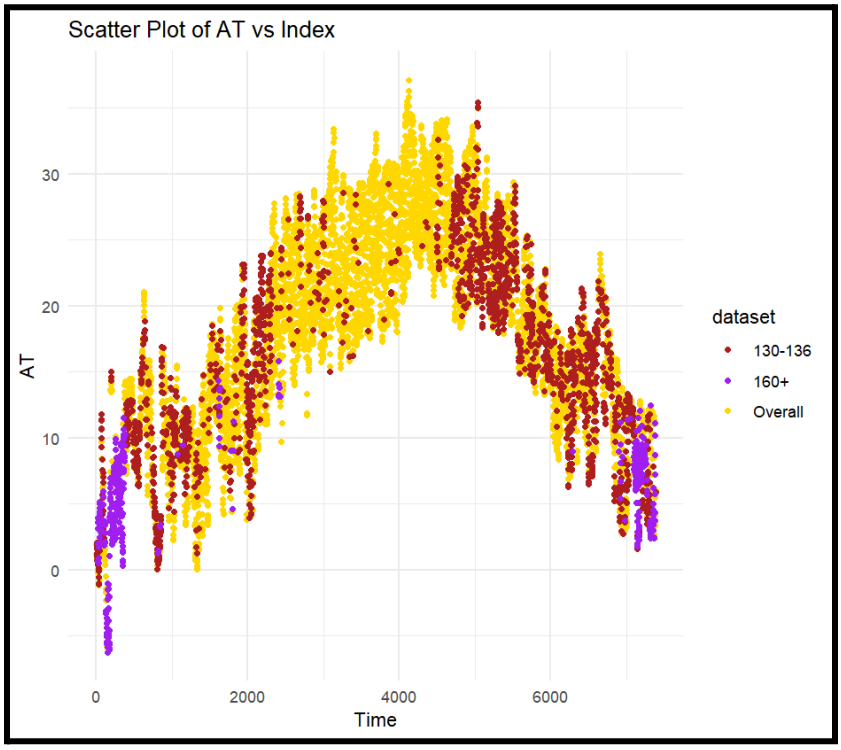
Density Plots of the Individual Variables (Overall Case)



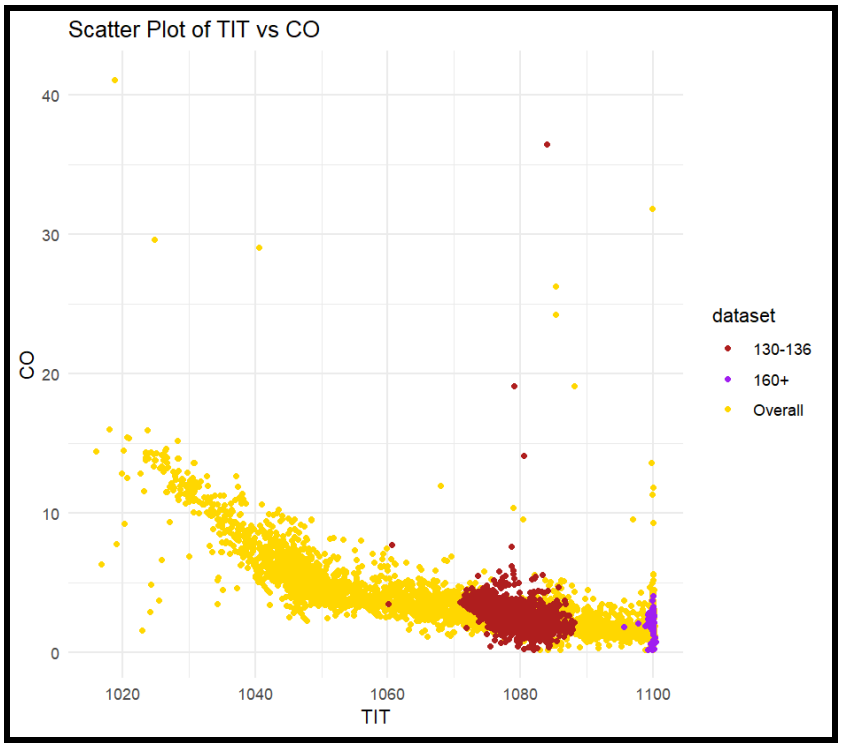


These plots identify the amount of observations for each continuous variable in the data set and use the y-axis to measure these observations in a percentage form.

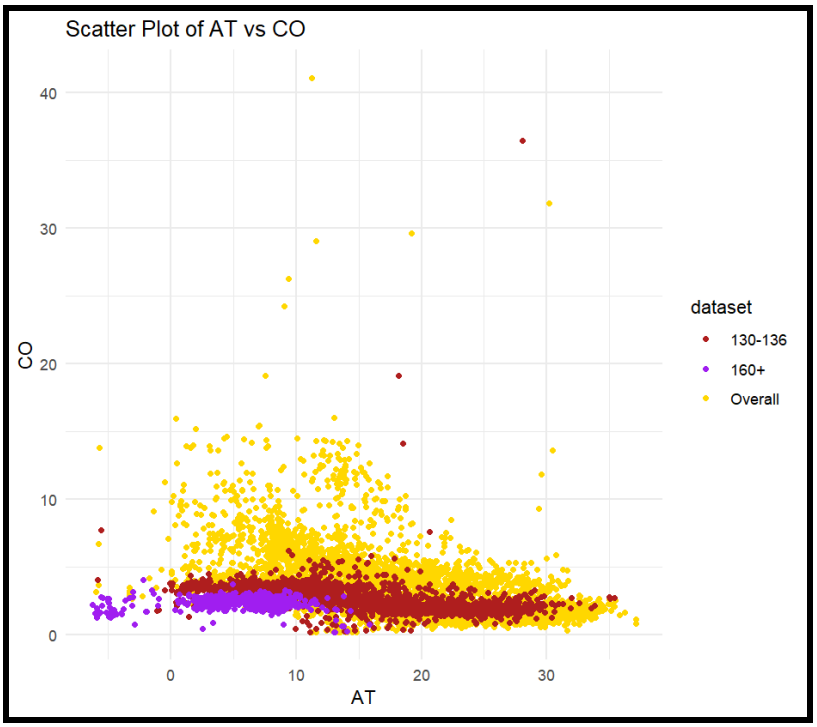
AT and Index Given TEY Case



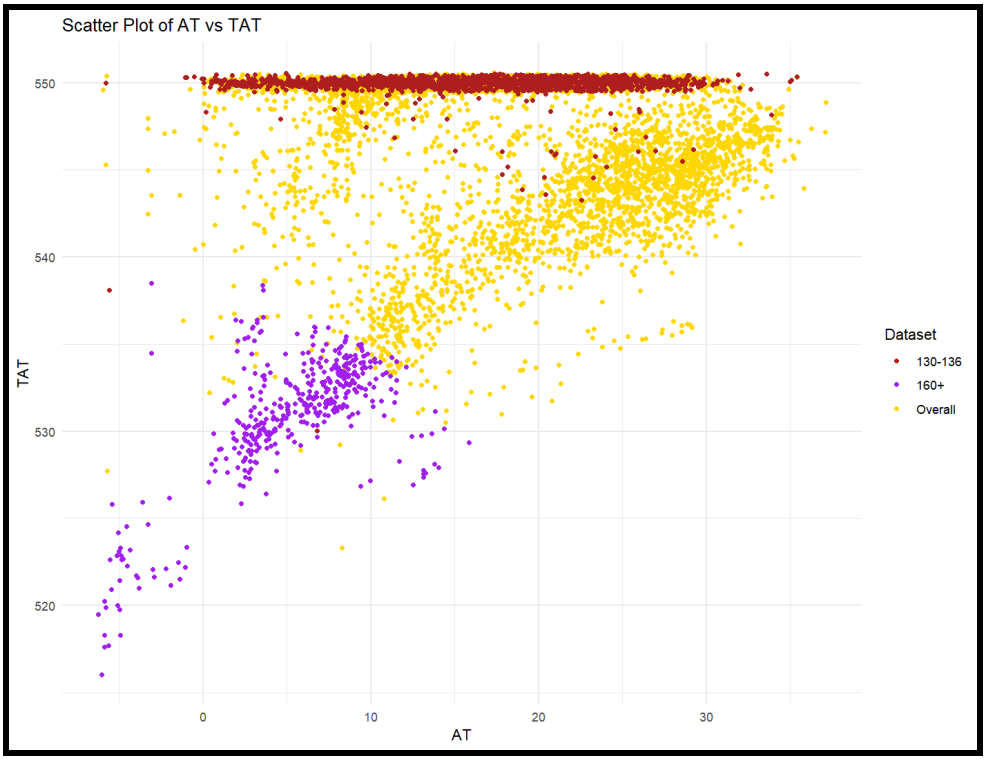
In this graph where AT is plotted against every index observation, we found that when TEY is greater than 160 MWh (Case 3), AT is limited to below approximately 10°C. In the winter months where AT is at its lowest, TEY is at its highest. This may show a negative correlation between the two variables in Case 3. Case 1 (Overall TEY Range) and Case 2 (TEY Range 130-136 MWh) does not show a similar correlation.



The scatter plot showing the relationship between TIT and CO demonstrates a decreasing curved relationship. We can visibly see each case has a cluster of points on the graph. This may signify a correlation between the variables. When TEY is between 130-136 MWh, TIT tends to be within the range of 1070-1090°C. As TIT continues to increase, CO declines with the lowest points of CO taking place at the highest TIT temperatures. The highest points of TIT take place in the 160+ MWh (Case 3) data. Therefore, CO may be at its lowest when the conditions of TIT are met within Case 3.



This graph looks at the relationship of the variables AT and CO aimed to identify a pattern. There are clear patterns that can be seen visually regarding the three cases. CO emissions are limited to below 5 mg/m³ for Case 3 (160+ MWh). In Case 2, we see a larger varying range for CO throughout the AT scale. For the other cases pertaining to the overall data set, we see much higher CO emissions which may provide useful insights to how TEY ranges, AT, and CO interact with one another.



This graph shows the relationship between AT and TAT. You can see visually that a linear relationship tends to appear in Case 1 and Case 3. As AT increases, so does TAT. In Case 2, TAT is almost always maxed out at 500°C. There appears to be a clear distinction between Case 3 and the other observations.

Methods and Results

Besides looking at the scatterplots to see individual relationships between CO and a variable, correlation matrices and models better capture these variables in context with each other.

We included a correlation matrix for each case, which measures the direction and strength of the row variable to the column variable. Darker shades indicate stronger linear relationships (i.e. how closely would the points on a scatterplot of these two variables follow a straight line). Blue indicates the direction is positive, so as the row variable is increasing in value we see the column variable increase in value as well. Red indicates the direction is negative, for example as CO emissions are increasing, we expect turbine-inlet temperature to decrease in Case 1. Typically, we run into issues when a variable we are measuring has little to no relationship with our response variable, CO, or when these variables have high correlation with each other (multicollinearity).

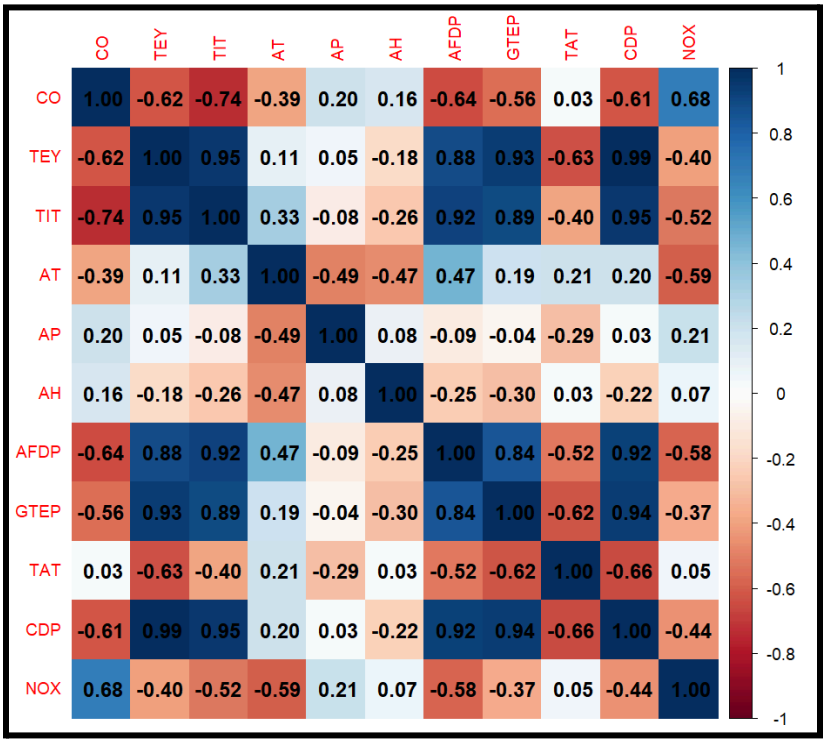
For the three separate cases, we used different types of models to explain CO emissions outside of isolating each variable individually. Models show us how variables influence CO in a larger context with each other and provide useful mathematical formulas to predict CO emissions when we change a variable.

In the overall data set, we found correlations for CO and the other variables, and saw that TIT has a high negative correlation (-0.74). Most of the other variables have moderate correlations besides TAT (0.03). TEY appeared to have strong correlations with multiple variables in the overall data set, with variables: TIT (0.95), AFDP (0.88), GTEP (0.93), and CDP (0.99).

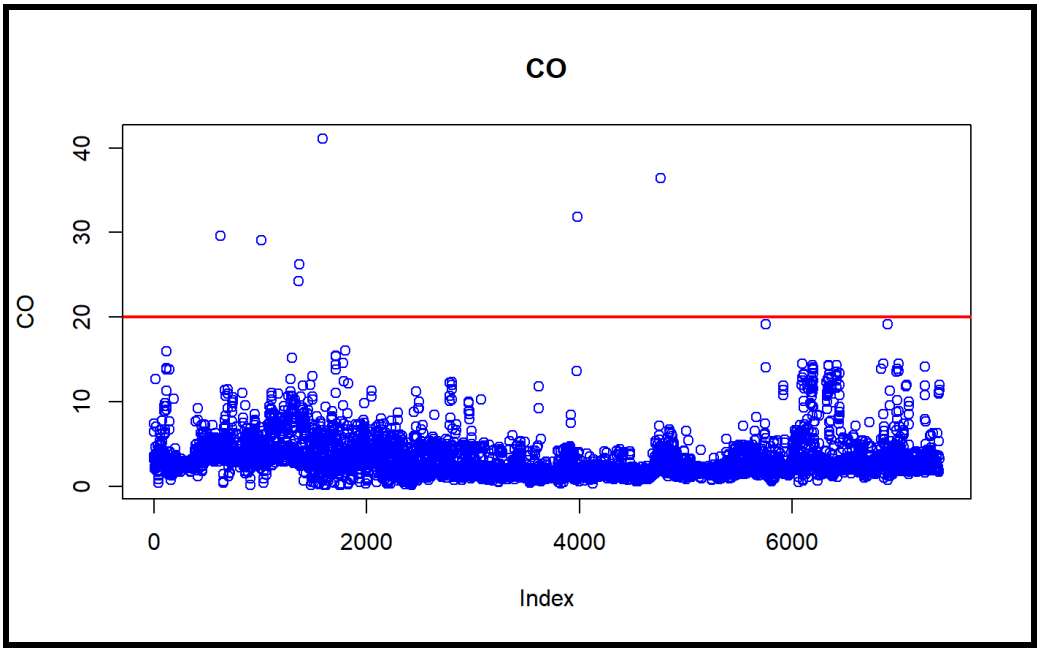
Case 2 shows different correlation values. There are no significant correlations regarding TEY and any other variables. CO also shows moderately-low correlations with some variables but nothing significant to note. AT's strong correlation with TIT (0.79) may provide some insights into their relationship.

In Case 3, there appear to be many strong negative correlations between the variables. TEY with both AT (-0.89) and TAT (-0.92). Perhaps these correlations may determine useful when building our final model.

Case 1: Overall Data Set



In case 1, CO has the strongest correlation with TIT (-0.74). AFDP (-0.64), CDP (-0.61), and GTEP (-0.56) also show large correlation with CO. All these variables had negative relationships with CO. For this case, we used a lasso model to get the most important variables. Three ambient variables (AT, AP, AH) and GTEP, TIT, TAT were chosen by lasso method (see appendix B.1). We used these 6 variables for the linear model.



According to the plot, there are only 7 data points where CO is higher than 20. Regarding that these 7 values did not show common characteristics, these were considered as outliers and were removed for the model.

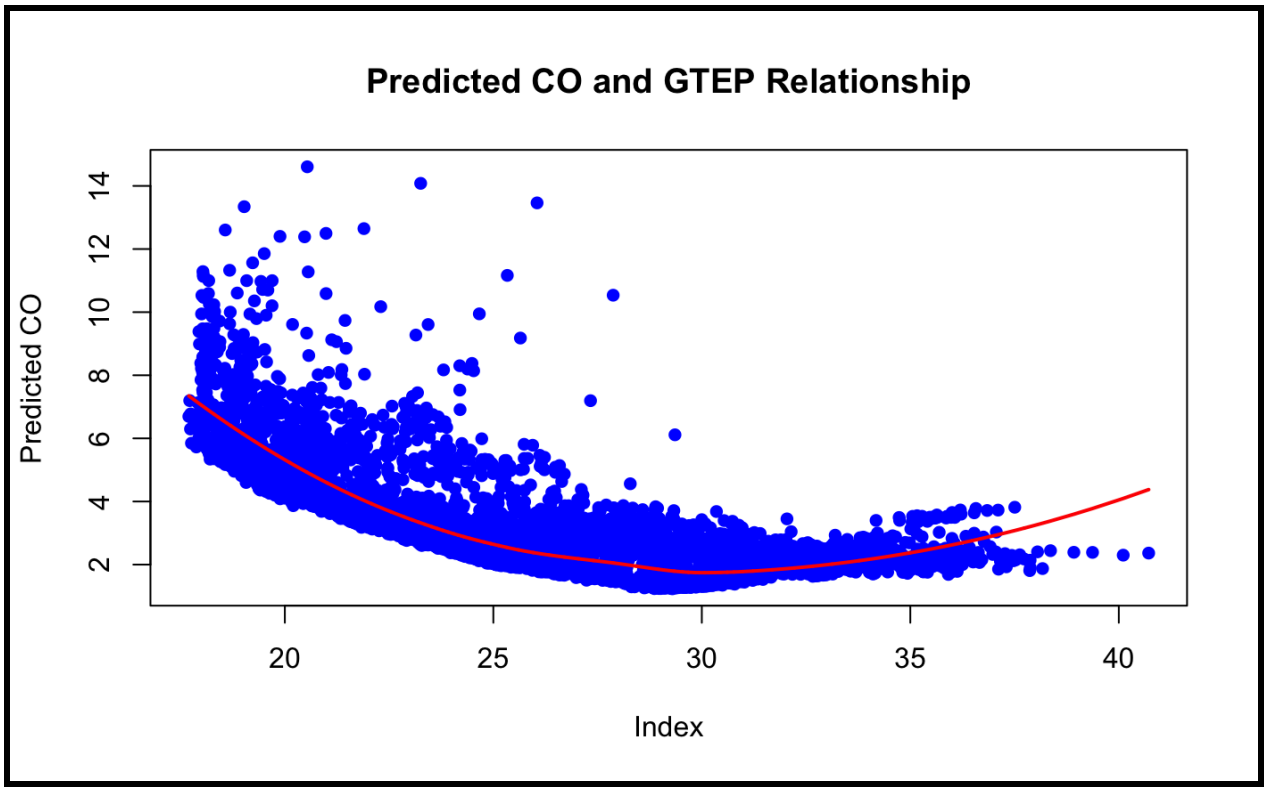
To make linear relationship between the variables and CO, we used log transformed CO using (see appendix B.2):

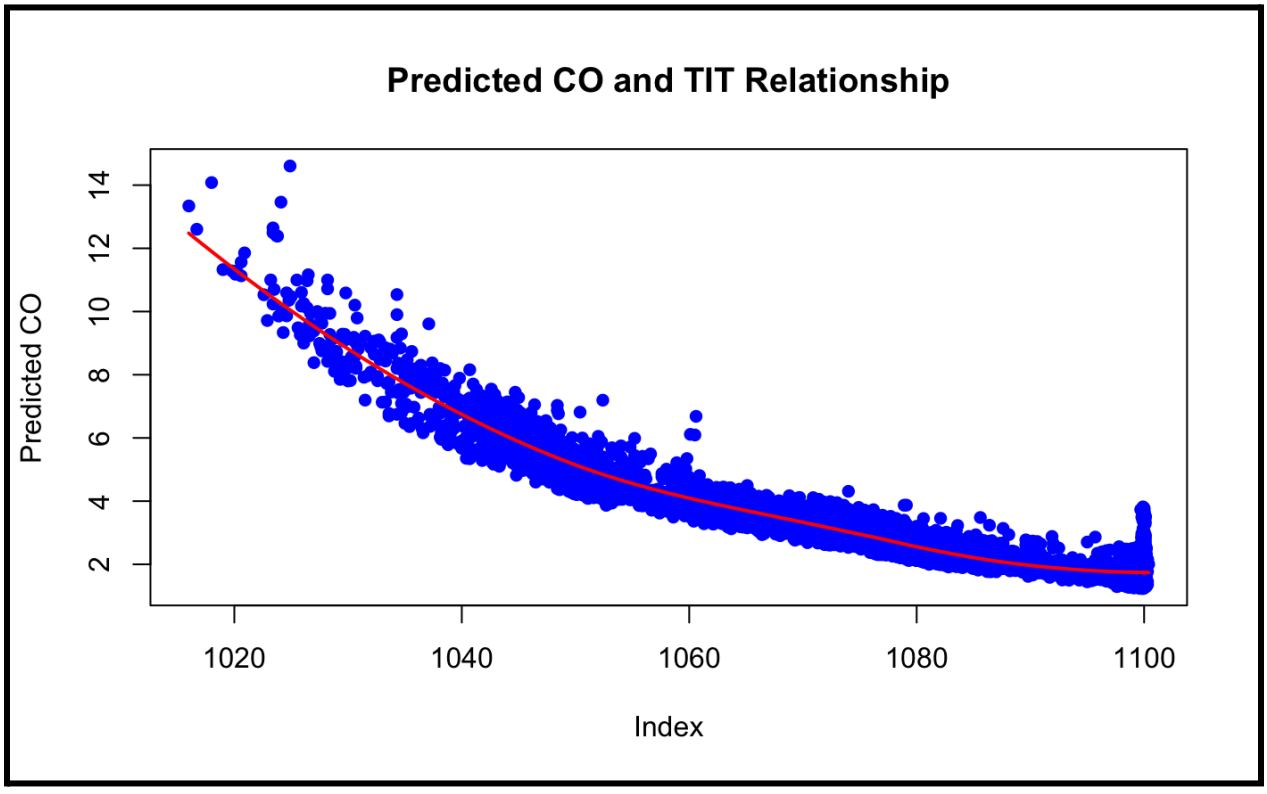
$$CO_{new} = \frac{CO^{\lambda} - 1}{\lambda}$$

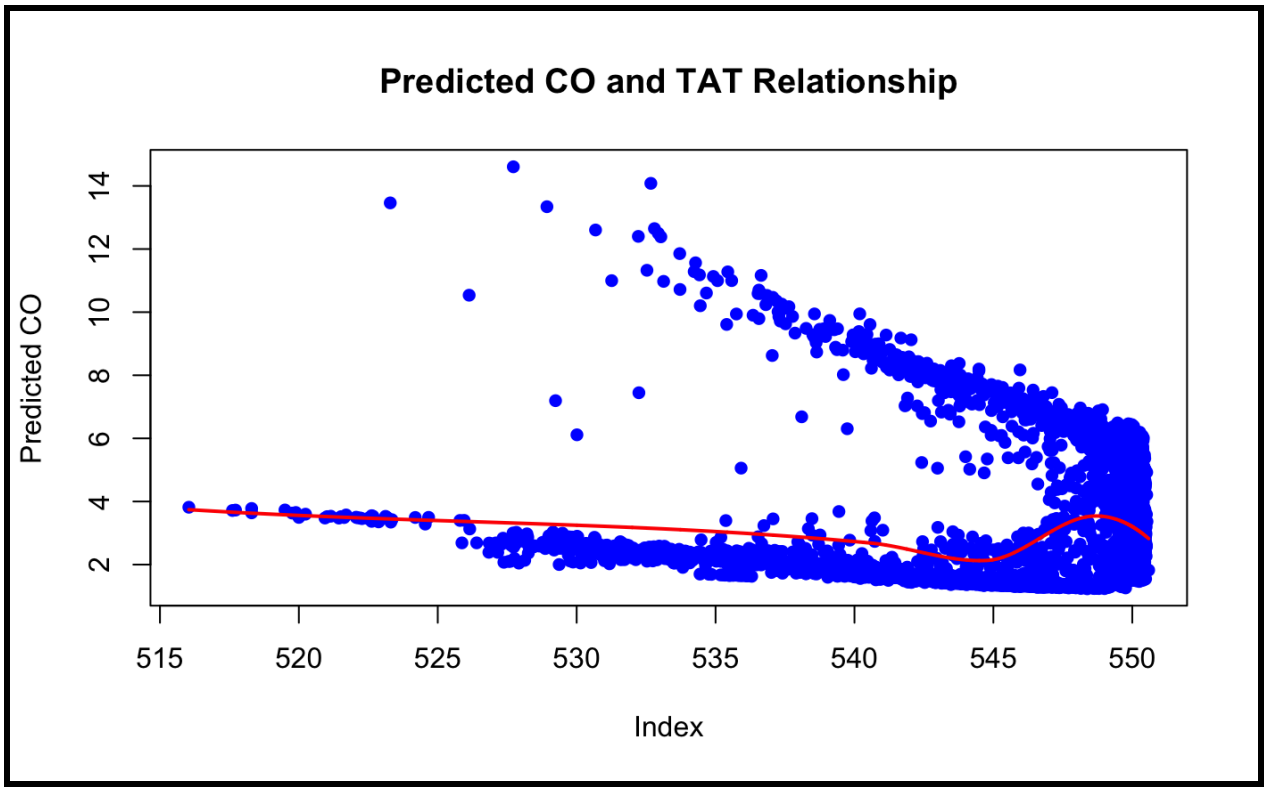
As a result, the final model was able to explain 74.1% of the variance of CO.

Generalized linear model:

CO_{new} **= 37.998 - 0.018*AT +0.004 *AP - 0.005*AH - 0.032*TIT + 0.020*GTEP - 0.012*TAT**

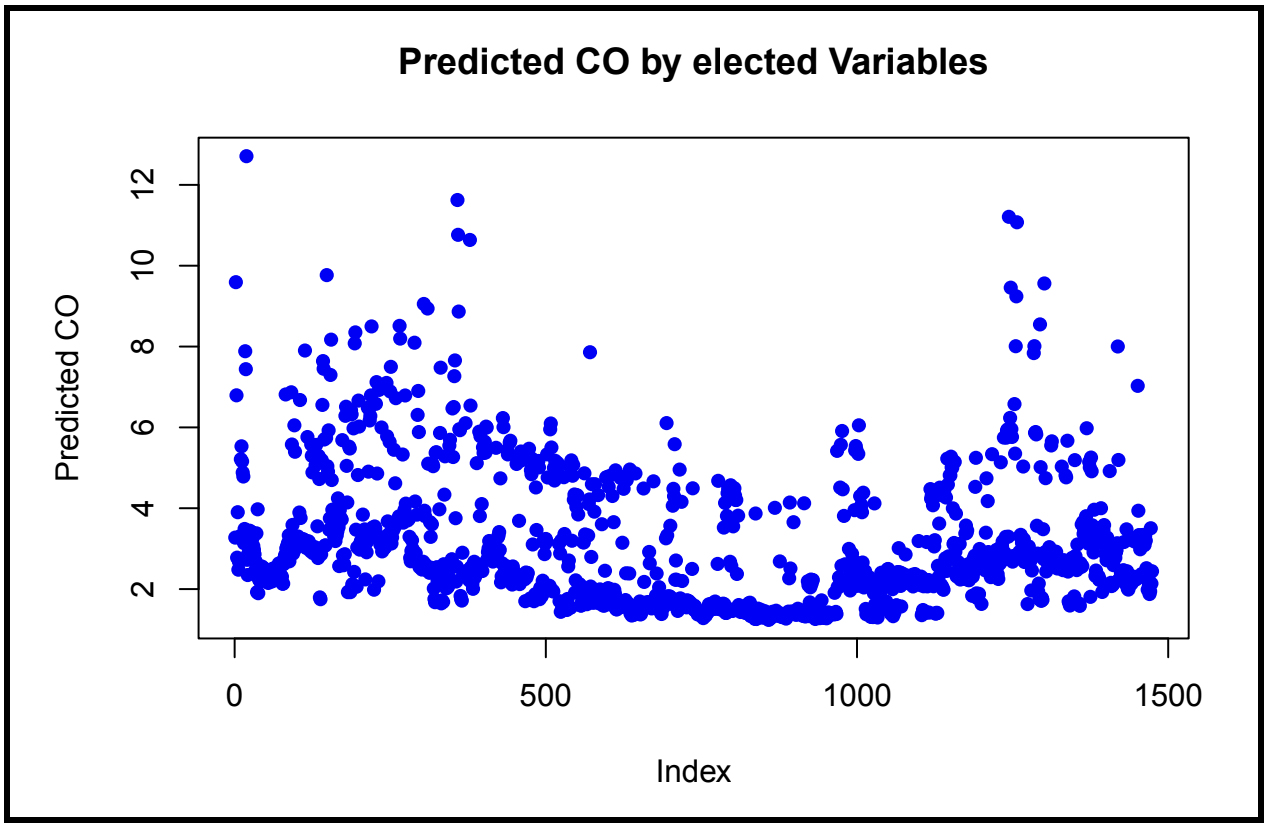

Predicted CO and GTEP Relationship
Predicted CO
2
4
6
8
10
12
14
20
25
30
35
40
Index


Predicted CO and TIT Relationship
Predicted CO
2
4
6
8
10
12
14
1020
1040
1060
1080
1100
Index

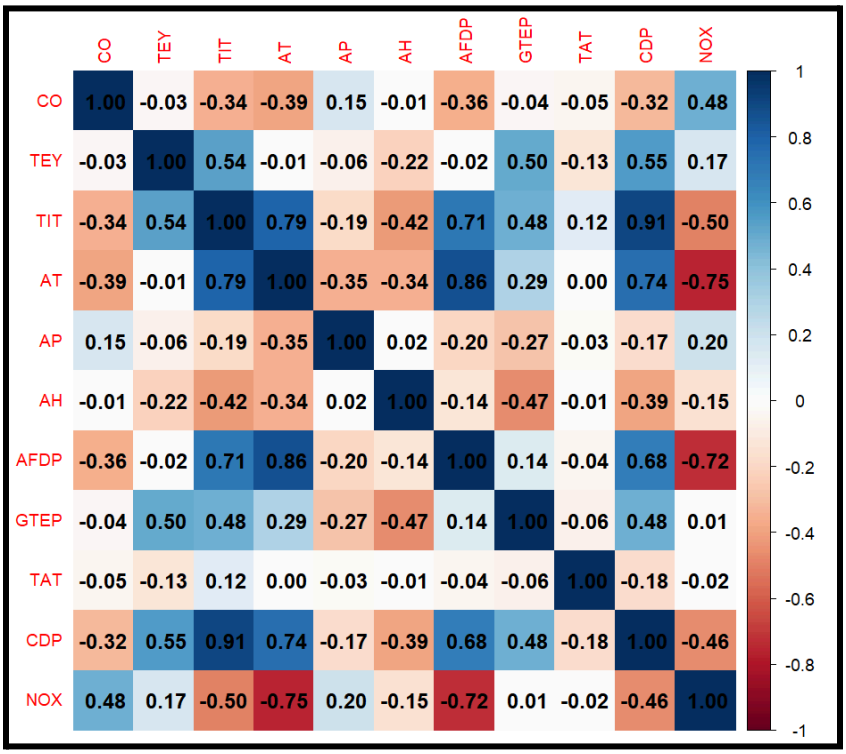


These three scatter plots show the relationship between predicted value of CO in its original scale and three variables GTEP, TIT, TAT. TAT does not have a linear relationship but the model performance is slightly higher when TAT is included and it is a significant variable picked by generalized linear model.

Using all six variables, the prediction of CO in the original scale looks like the plot below.


Predicted CO by elected Variables
Predicted CO
2
4
6
8
10
12
0
500
1000
1500
Index

Case 2: TEY Range 130-136 MWh



The model is more reliable than the correlation matrix shown above, which only captures direct linear relationships between CO and a variable. Contrary to the matrix showing no relationship between CO and AH or CO and GTEP, our model shows that when we have the four variables they chose in our model, they are all statistically significant in predicting CO emissions.

Generalized least squares model:
log(CO) = 31.184 - 0.021*AT - 0.005*AH - 0.028*TIT + 0.012*GTEP

We chose this type of model for Case 2 because it had stronger strength (lower AIC value) compared to other models we tested, and it focused on only a few turbine variables that were found to have some of the strongest relationships with CO. The baseline value of 31.184 for the logarithm of CO when all other variables are zero does not have any meaning because that is an impossible situation. The negative signs in front of AT, AH and TIT tell us that when these variable values are high, they are associated with lower CO emissions. For GTEP, the coefficient of 0.012 tells us that for each 1 mbar increase in gas turbine exhaust pressure, the logarithm of CO levels increases by 0.012. In other words, each 1-unit increase in GTEP leads to a 1.21% increase in CO emissions. More broadly, higher exhaust pressure is associated with higher CO emissions.

Table of model variables:

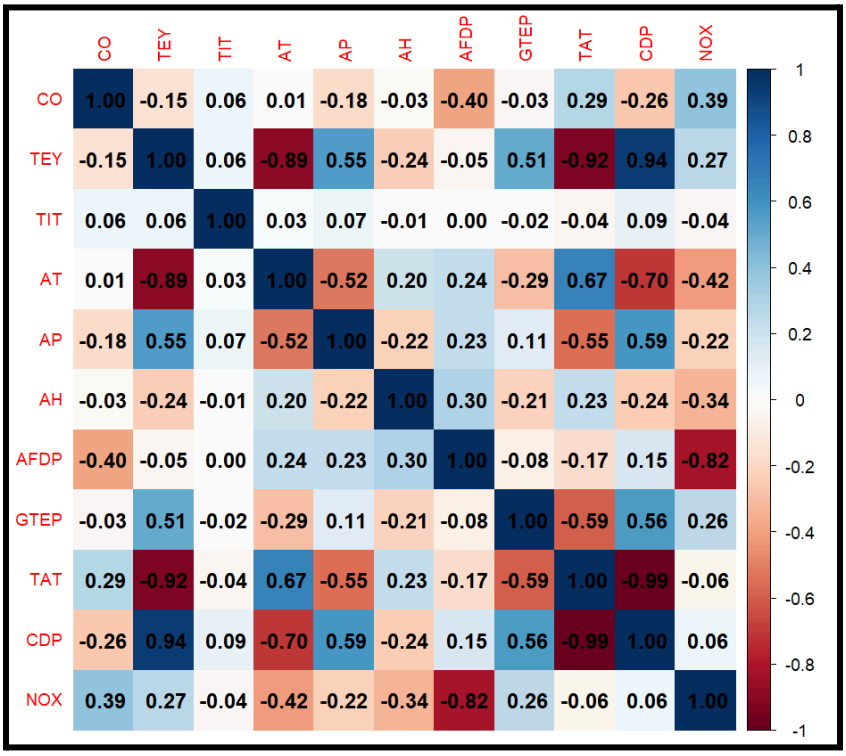
Coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	31.184448	4.255557	7.327936	0.0000
AT	-0.021124	0.002103	-10.046847	0.0000
AH	-0.004764	0.000856	-5.564955	0.0000
TIT	-0.027723	0.003999	-6.932301	0.0000
GTEP	0.012229	0.005719	2.138155	0.0326

While this may seem small, these coefficient values in the model equation are significantly significant (p-values < 0.05), with larger coefficients indicating stronger relationships. TIT has the strongest relationship with CO, with each 1-unit increase in TIT leading to a 2.75% decrease in CO emissions. The small standard errors in the model indicate our estimates are reasonably precise, and we found that on average, this model's predictions deviate from the actual CO emission values by about 0.672 mg/m^3.

We also resolved some issues along the way of deriving our final model for Case 2. Our data for Case 2 did not follow a normal distribution, so we found the best transformation for our response variable CO using box-cox, which was taking the logarithm. There were three outlier CO observations as well that were well above the other values observed, and their removal from the model substantially improved its strength (lowered the AIC value). We chose a correlation structure of AR(1) because we had time series data where our errors (residuals) from observations are moderately positively correlated (phi = 0.405) on consecutive days when data was collected due to the turbine being in similar conditions. Finally, we used the maximum likelihood method for the generalized least squares model because it is more suitable for dealing with the non-constant variance in our data. Please refer to Appendix C for more details.

CASE 3: TEY Range 160+ MWh



In Case 3, where TEY exceeds 160 MWh, we observed no strong linear relationships between CO emissions and the predictors from the correlation matrix. The highest correlation was between CO and AFDP at -0.40. TEY showed strong correlations with AT(-0.89), TAT(-0.92), and CDP(-0.94). However, these linear correlations failed to capture the complex relationships between CO emissions and turbine variables, underscoring the need for more advanced modeling techniques.

Generalized Linear Model (GLM) with Log Transformation:

$$\log(CO) = -143.5 + 0.025 \cdot AT + 0.007 \cdot AP + 0.065 \cdot GTEP + 0.195 \cdot TAT + 0.114 \cdot TEY + 0.860 \cdot CDP$$

After testing various models (see Appendix D), the Generalized Linear Model (GLM) with a log-transformed response (log(CO)) emerged as the best-performing model with an AIC of 98.66, significantly lower than the initial linear regression model's AIC of approximately 500. This model was selected after addressing key issues such as non-linearity, heteroscedasticity, and influential points. The log transformation stabilized the variance of the response variable, improving model performance and diagnostic outcomes. After stepwise selection, the final model included six predictors: AT, AP, GTEP, TAT, TEY, and CDP. These variables were selected based on their significance and contribution to the model's predictive power.

Table of model variables:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.435e+02  2.152e+01  -6.668 8.38e-11 ***
AT           2.476e-02  1.634e-02   1.515 0.130536
AP           7.234e-03  3.356e-03   2.155 0.031720 *
GTEP         6.507e-02  1.277e-02   5.096 5.31e-07 ***
TAT          1.951e-01  2.772e-02   7.038 8.22e-12 ***
TEY          1.143e-01  3.427e-02   3.336 0.000929 ***
CDP          8.603e-01  4.271e-01   2.014 0.044630 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The coefficient estimates of the final model revealed important insights. AT (Ambient Temperature) was retained in the model despite being non-significant, as it likely contributed to overall model fit and interpretability.AP (Ambient Pressure) showed a positive relationship with CO emissions, suggesting that higher atmospheric pressure slightly increases emissions. GTEP (Gas Turbine Exhaust Pressure) and TAT (Turbine After Temperature) exhibited strong positive relationships with CO emissions, highlighting their critical roles in turbine operations. TEY (Turbine Energy Yield) significantly influenced CO emissions, aligning with expectations that higher energy output leads to increased emissions. CDP (Compressor Discharge Pressure) had a moderate positive effect, indicating its additional predictive value.

The chosen GLM with log transformation provided significant improvements over both the initial linear regression and the GLS AR(1) model by addressing several issues. The log transformation effectively captured non-linear relationships between CO emissions and turbine variables. Although the Breusch-Pagan test indicated residual heteroscedasticity, the log transformation significantly stabilized the variance, and the model could be further improved with robust standard errors. Cook’s Distance analysis identified a few influential points, but the final model remained robust to their effects. While minor issues such as residual heteroscedasticity remain, the log-transformed GLM provides a robust, interpretable, and high-performing model for capturing the key relationships between turbine variables and CO emissions in this case.

Conclusion

We set out to answer two main research questions 1) Which variables have the strongest relationships with CO for each of the three cases, and 2) How can we see the nature of these relationships through visualizations and modeling?

Across all three cases, we found that decreases in CO emissions are associated with increases in TIT and decreases in GTEP. These two variables, TIT and GTEP, likely have the strongest relationship with CO. From our models, TIT consistently showed significant p-values across cases: $p<0.0001$ in Case 1, $p<0.0001$in Case 2, and $p=0.131$ in Case 3. GTEP was significant or marginally significant, with $p=0.0034$ in Case 1, $p=0.0326$ in Case 2, and $p<0.001$ in Case 3. Case 3 also introduced additional significant predictors, such as TEY ($p<0.001$) and CDP ($p=0.045$), which were unique to this scenario and reflect the distinct dynamics at high energy yields.

Comparing across cases, we see consistency between Case 1 and Case 2 in terms of the coefficients for TIT, TAT, and AFDP. Both cases demonstrate negative relationships between these variables and CO emissions, suggesting that turbine improvements targeting higher TIT values can reduce emissions. However, Case 3 presents a shift in dynamics: TAT showed a positive relationship with CO emissions, contrasting its negative effect in the other cases. At high TEY levels, additional fuel consumption and air intake appear to create diminishing returns for turbine efficiency, leading to higher TAT and increased emissions. Similarly, the emergence of TEY and CDP as significant predictors in Case 3 underscores the distinct challenges of managing emissions under elevated energy yield conditions.

Based on our analysis, we recommend monitoring TIT, GTEP, TAT and AFDP as key indicators for turbine performance. Higher TIT generally improves the efficiency of the gas turbine, allowing it to generate more power for the same amount of fuel. This reduces CO emissions as well since CO is a byproduct of incomplete combustion when there is insufficient oxygen to fully oxidize the carbon in the fuel. If the turbine inlet temperature is too low, the fuel does not burn and incomplete combustion occurs, resulting in more CO byproduct. For gas turbine exhaust pressure (GTEP), higher values mean the turbine needs to work harder to expel gases, making the combustion process less optimal and raising CO emissions. Turbine after temperature (TAT) is the temperature of these expelled exhaust gases and should theoretically have similar reasoning as TIT. However, when the total energy yield of the turbine is too high as in Case 3, higher power output involves more fuel consumption and air intake. Having higher TAT actually hinders turbine performance, so there is an upper limit to how high temperatures in the turbine should be for optimal performance. That is why we see a positive relationship in the Case 3 model for TAT compared to negative relationship in the overall data. Additionally for this case, the air filter difference pressure should be slightly increased to reduce excess air flow and create the optimal air-fuel ratio for low CO emissions.

Our preliminary results have some limitations as well. Most importantly, we generally focused on parametric models that assume a known distribution for the data, and we did not go into much depth exploring nonlinear relationships with CO. Though our model predictions had slight variability, the range of typical CO emissions is so small that it is hard to evaluate the model performance. Future attempts at modeling the turbine parameters' impacts on CO emissions may want to consider more flexible nonparametric models with less assumptions about the data at the cost of interpretability. More investigation could be done to the turbine's state when extremely large CO emission levels are detected, such as potential damage to the turbine. Additionally, we recommend creating an experimental design testing the current model and an updated model over a short period of time for a better comparison of how they perform in CO emissions.

Appendix A

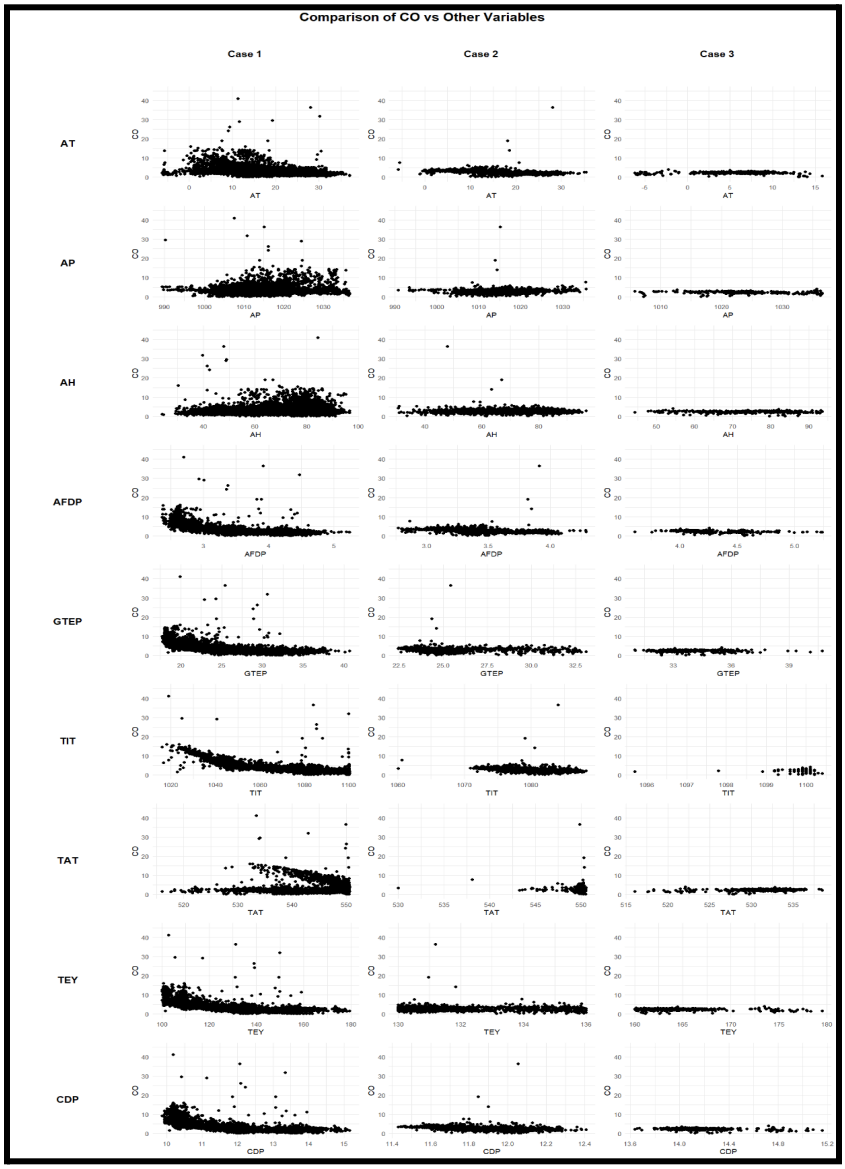
Descriptive Analysis

For more detailed information on our analysis, please refer to the code we submitted with this report that includes technical annotations of our code and additional descriptions. We felt that as a team each individual contributed equally to make this project possible.

FIGURE 1: TABLE OF VARIABLES

Variable Name	Variable Abbreviation	Variable Type	Unit	Min	Max	Mean
Ambient Temperature	AT	Continuous	°C	-6.23	37.10	17.71
Ambient Pressure	AP	Continuous	mbar	985.85	1036.56	1013.07
Ambient Humidity	AH	Continuous	%	24.08	100.20	77.87
Air Filter Difference Pressure	AFDP	Continuous	mbar	2.09	7.61	3.93
Gas Turbine Exhaust Pressure	GTEP	Continuous	mbar	17.70	40.72	25.56
Turbine Inlet Temperature	TIT	Continuous	°C	1000.85	1100.89	1081.43
Turbine After Temperature	TAT	Continuous	°C	511.04	550.61	546.16
Compressor Discharge Pressure	CDP	Continuous	mbar	9.85	15.16	12.06
Turbine Energy Yield	TEY	Continuous	MWh	100.02	179.50	133.51
Carbon Monoxide	CO	Continuous	mg/m³	0.00	44.10	2.37
Nitrogen Oxide	NOX	Continuous	mg/m³	25.90	119.91	65.29

FIGURE 2: SCATTERPLOTS OF CO AND OTHER VARIABLES BY CASE



While the CO ranges are the same for each graph, the horizontal x-axis values are not since our turbine variables have different units of measurement. Case 1 has many variables with nonlinear relationships to CO, while they appear linear in Case 2. For Case 3, CO emissions are extremely small (under 5 mg/m^3) and its relationship to other variables is harder to visually determine as the points tend to cluster within certain variable ranges. Some variables have similar relationships to CO within each case, such as AFDP, GTEP and CDP in Case 1.

Appendix B

Case 1

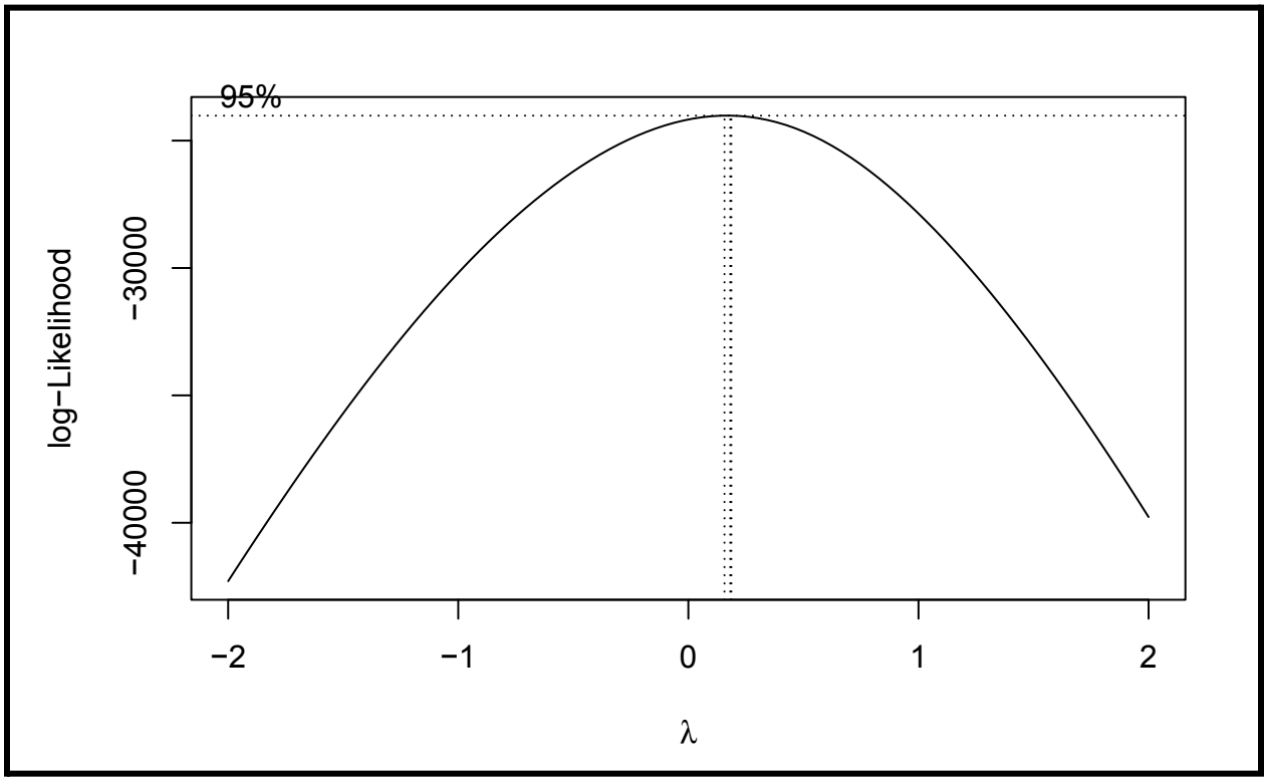
1. Variable Selection - Lasso Regularization

```
las_model <- gamlr(X, Y, family = "gaussian")
coef(las_model)
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                  seg100
## intercept 37.977218443
## AT        -0.018311989
## AP         0.002244427
## AH        -0.004510431
## AFDP       .
## GTEP       0.003410480
## TIT       -0.028296966
## TAT       -0.014804892
## TEY        .
## CDP        .
```

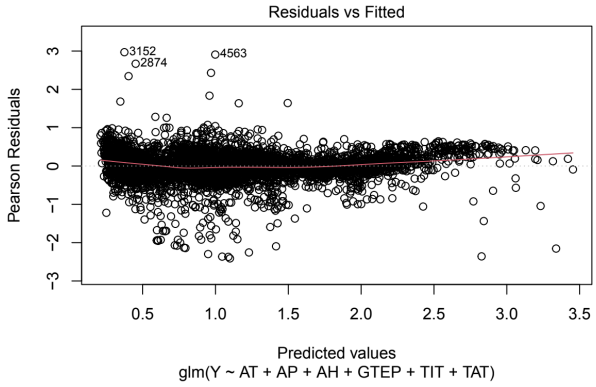
Using the lasso model with full data, three ambient variables (AT, AP, AH) and GTEP, TIT, TEY.

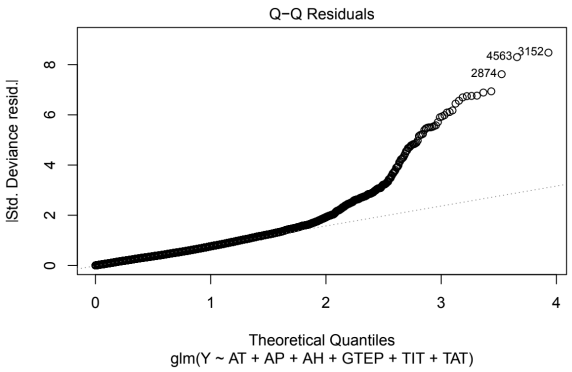
2. Box-Cox Transformation of CO

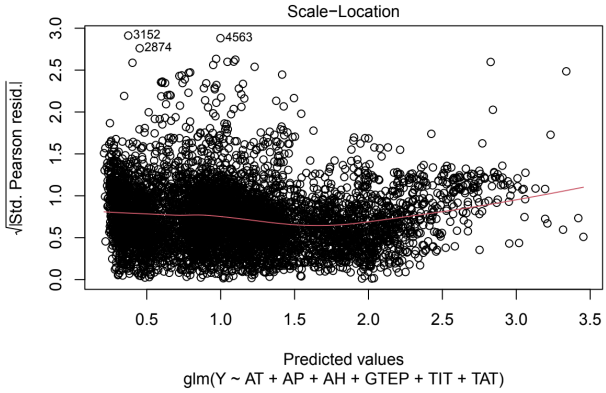


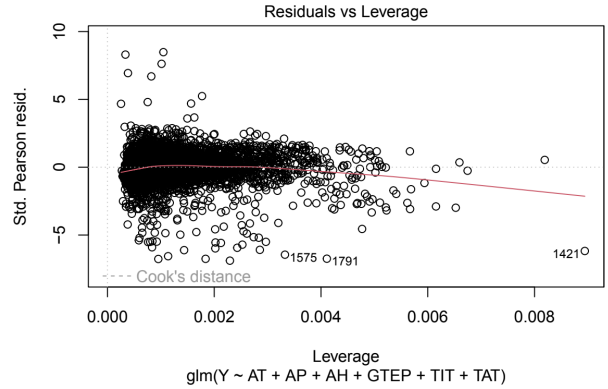
According to the Box-Cox test, CO needs log transformation for better model performance.

3. Final model residual analysis


Residuals vs Fitted
Pearson Residuals
3152
2874
4563
Predicted values
glm(Y ~ AT + AP + AH + GTEP + TIT + TAT)


Q−Q Residuals
|Std. Deviance resid.|
4563
3152
2874
Theoretical Quantiles
glm(Y ~ AT + AP + AH + GTEP + TIT + TAT)


Scale−Location
√|Std. Pearson resid.|
3152
2874
4563
Predicted values
glm(Y ~ AT + AP + AH + GTEP + TIT + TAT)


Residuals vs Leverage
Std. Pearson resid.
1575
1791
1421
Cook's distance
Leverage
glm(Y ~ AT + AP + AH + GTEP + TIT + TAT)

Appendix C

Case 2

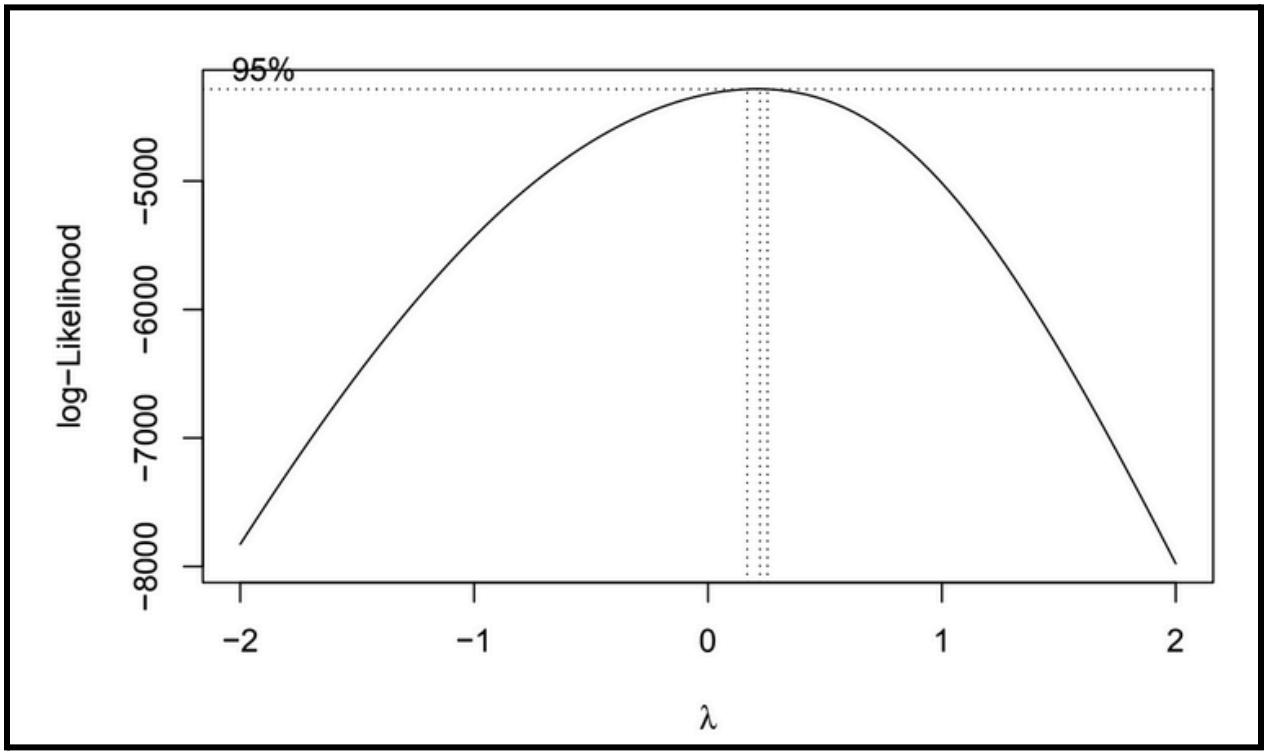
We first improved the full multiple linear regression model to the best of our ability by applying a logarithmic transformation of CO and removing three highly influential observations that are well above the third quantile for CO. That was able to reduce our AIC dramatically.

```
##             df       AIC
## lr_model    11 5382.9544
## lr_model2   11 3397.4474
## lr_model3   11  658.3472
## lr_model4   11 -503.0532
```

```
df_mid2 = filter(df_mid, CO < 10)
summary(df_mid$CO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2325  1.9786  2.6072  2.7150  3.3849 36.4540
```

Explanation of models: lr_model is the original full multiple linear regression model, lr_model2 is after removing three highly influential observations, lr_model3 tests the logarithmic transformation of CO with the removed observations and lr_model4 tests the square root transformation of CO with the removed observations.



The optimal lambda from the box-cox plot above was 0.222, so we tested both log(CO) and sqrt(CO). Moving on to other models that work better with the failed assumptions from the multiple linear regression model, we continued to use the data with removed observations, all of the predictor variables besides NOX, and a log(CO) transformation. These included different

forms of the generalized least squares model, generalized linear model, robust model and general additive model. We stuck to parametric models for easier interpretation.

	df <dbl>	AIC <dbl>
lr_model3	12.000	-7202.0374
model_gls	11.000	741.7242
model_gls_AR1	12.000	447.1554
model_gls_AR1_2	12.000	368.3355
model_glm	11.000	658.3472
model_glm2	11.000	3401.4475
model_robust	11.000	687.6560
model_gam2	42.989	483.4269

As you can see from this table, model_gls_AR1_2 had the smallest AIC value of 368. As stated previously, the AR(1) correlation structure and maximum likelihood method works well with our time series data with non-constant variance. We then moved to variable selection to balance model strength with model complexity.

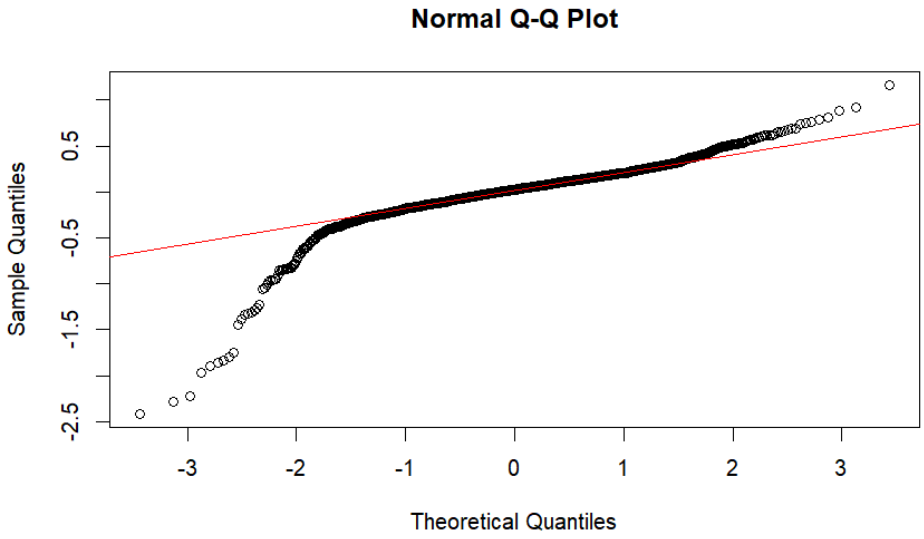
	Value <chr>	Std.Error <chr>	t-value <chr>	p-value <chr>
(Intercept)	36.50592	6.853744	5.326420	0.0000
AT	-0.01179	0.004450	-2.650682	0.0081
AP	0.00359	0.001877	1.914683	0.0557
AH	-0.00456	0.000875	-5.208133	0.0000
GTEP	0.01286	0.006217	2.068383	0.0388
TIT	-0.03699	0.007818	-4.730744	0.0000
TEY	0.02755	0.013554	2.032218	0.0423
CDP	-0.23464	0.111257	-2.108986	0.0351

While the reduced model from stepwise selection produced an AIC = 364 and all significant predictors, there are still way too many variables to focus on. We further reduced the model down to 4 predictors from the above selected variables: AT, AH, TIT and GTEP.

	Value <chr>	Std.Error <chr>	t-value <chr>	p-value <chr>
(Intercept)	31.184448	4.255557	7.327936	0.0000
AT	-0.021124	0.002103	-10.046847	0.0000
AH	-0.004764	0.000856	-5.564955	0.0000
TIT	-0.027723	0.003999	-6.932301	0.0000
GTEP	0.012229	0.005719	2.138155	0.0326

The AIC for this model, reduced_model, is 365, which is not much different from the above model. Additionally, many of the p-values are < 0.0001 which makes these predictors highly statistically significant in the context of each other. Therefore, we felt this model was better at capturing the client's wish for targeted, effective improvements to their turbine.

Using 10-fold leave-one-out cross-validation, we found that on average, our reduced_model's predictions deviates from the actual CO values by about 0.672 mg/m^3. Performing model diagnostics, we found that while there is no significant autocorrelation between the residuals, they do not follow a Normal distribution.

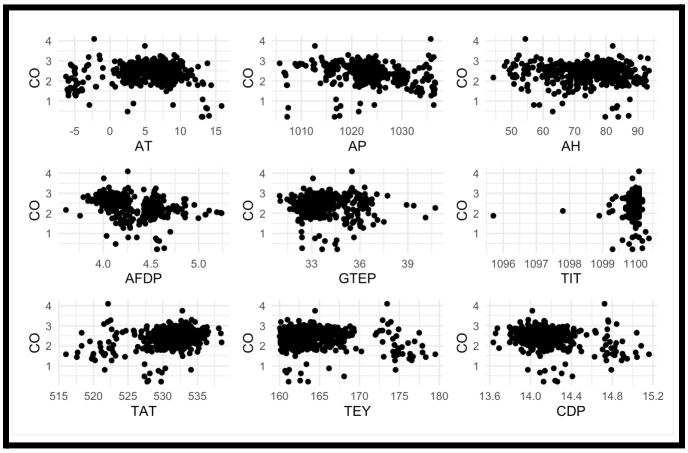


The non-normality of our residuals is the major limitation for reduced_model. Therefore, we may have some misleading conclusions from our hypothesis tests. However, when we tested the generalized linear model (GLM) that is designed to handle non-normal residuals, we saw in the earlier table that its AIC was 658, almost double our model_gls_AR1_2. Future attempts at model-fitting may want to choose other models designed to handle the type of data structure for Case 2.

Appendix D

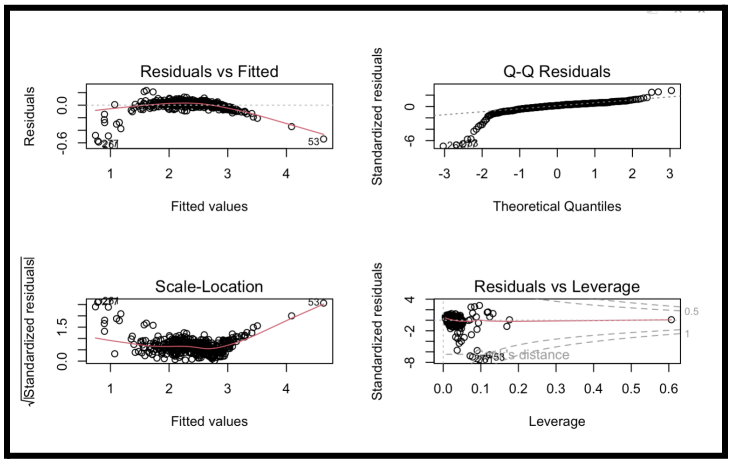
Case 3

1. EDA and Initial Modeling
    a. Variable Relationships



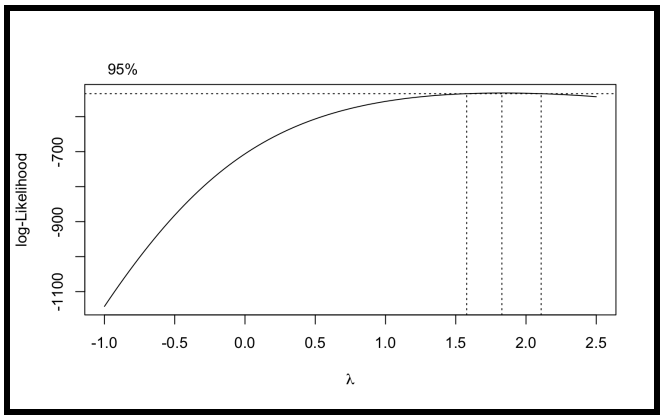
Some relationships, such as those for TIT and TEY, appear nonlinear. For example, TIT shows a cluster around 1100, which may suggest specific operational modes or a non-standard relationship with CO. Variables like AP and AH don't show strong visible trends with CO, which might indicate a weaker or less direct relationship.

    b. Initial linear model

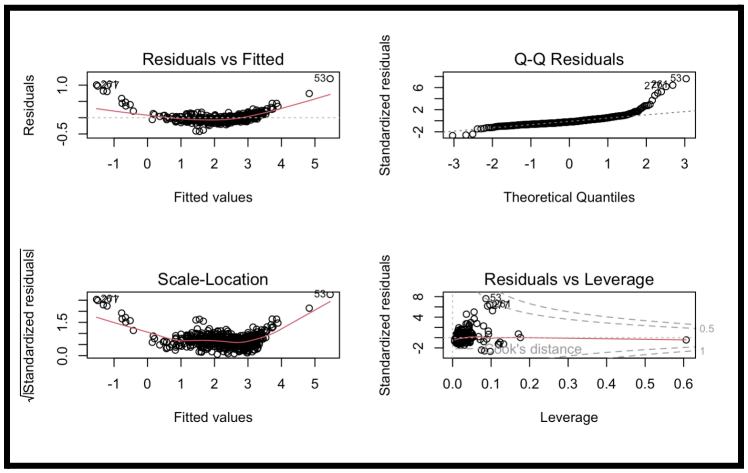


The slight curvature in residuals vs fitted suggests the relationship between the predictors and CO is not strictly linear. Residual spread is inconsistent, violating the assumption of constant variance. Outliers and deviations in the Q-Q plot tails show the residuals are not normally distributed. High-leverage points like observation 35 might disproportionately affect the model.

    c. Box Cox Transformation



d. Transformed linear model



Curvature in residuals, heavy tails in the Q-Q plot, and strong heteroscedasticity in the Scale-Location plot suggest that multiple linear regression (MLR), even with the transformed CO, is not the best approach for this case.

2. More modeling
    a. GLS, GLM, RLM

```r
# Generalized Least Squares (GLS)
model_gls_CO <- gls(CO ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP, data =
filtered_data)
model_gls_CO_transformed <- gls(log(CO) ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP,
data = filtered_data)

# Generalized Linear Models (GLM)
model_glm_CO <- glm(CO ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP, family = gaussian
, data = filtered_data)
model_glm_CO_transformed <- glm(log(CO) ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP,
family = gaussian, data = filtered_data)

model_glm_loglink_CO <- glm(CO ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP, family =
gaussian(link = "log"), data = filtered_data)

# Robust Linear Models (RLM)
model_robust_CO <- rlm(CO ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP, data =
filtered_data)
model_robust_CO_transformed <- rlm(log(CO) ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY +
CDP, data = filtered_data)
```

b. AIC scores

Model <chr>	AIC <dbl>
GLS (CO)	554.8497
GLS (CO Transfor…	167.8630
GLM (CO)	498.4980
GLM (CO Transfor…	102.0031
RLM (CO)	509.8199
RLM (CO Transfor…	154.4573

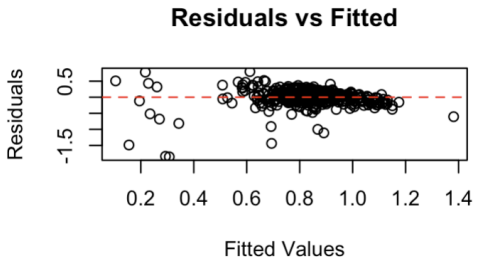
GLM log(CO) (AIC = 102.0031) is the best-performing model.
GLM (AIC = 498.4980) is the best-performing model.

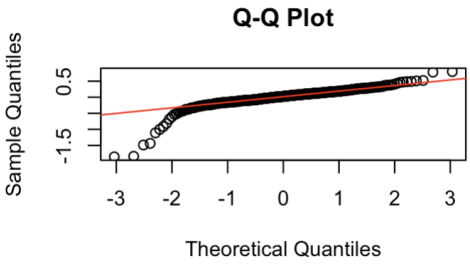
c. Variable selection

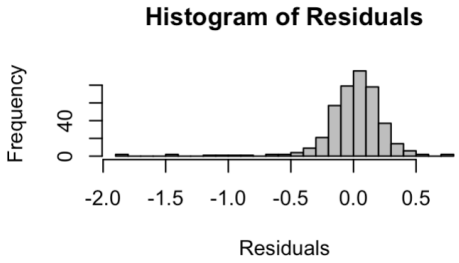
```
       Df Deviance     AIC
<none>      29.768  98.661
+ AFDP  1   29.632  98.746
- AT    1   29.935  98.989
+ AH    1   29.700  99.705
+ TIT   1   29.726 100.077
- CDP   1   30.063 100.768
- AP    1   30.105 101.359
- TEY   1   30.576 107.826
- GTEP  1   31.654 122.271
- TAT   1   33.365 144.228

Call:
glm(formula = log(CO) ~ AT + AP + GTEP + TAT + TEY + CDP, family = gaussian,
    data = filtered_data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.435e+02  2.152e+01  -6.668 8.38e-11 ***
AT           2.476e-02  1.634e-02   1.515 0.130536
AP           7.234e-03  3.356e-03   2.155 0.031720 *
GTEP         6.507e-02  1.277e-02   5.096 5.31e-07 ***
TAT          1.951e-01  2.772e-02   7.038 8.22e-12 ***
TEY          1.143e-01  3.427e-02   3.336 0.000929 ***
CDP          8.603e-01  4.271e-01   2.014 0.044630 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.07260516)

    Null deviance: 40.293  on 416  degrees of freedom
Residual deviance: 29.768  on 410  degrees of freedom
AIC: 98.661
```

The most influential predictors are GTEP, TAT, and TEY, which have the strongest statistical significance and largest coefficients. AP and CDP also contribute significantly but have smaller effects.

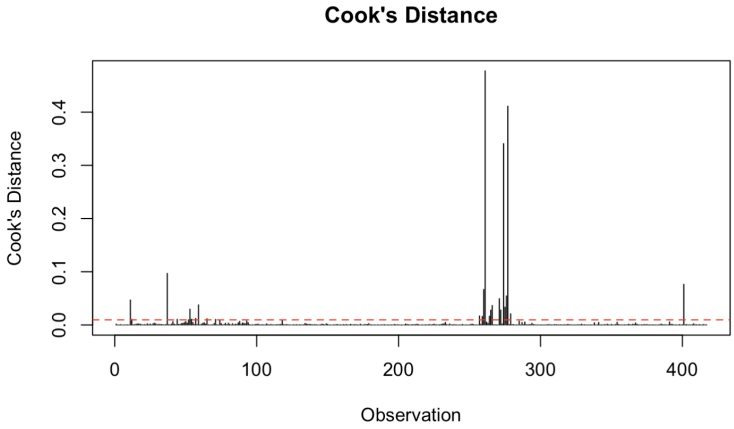
3. Model Diagnostics








```
        studentized Breusch-Pagan test

data:  stepwise_model
BP = 68.67, df = 6, p-value = 7.661e-13
```



- The residuals appear mostly random, and the Gaussian assumption for the GLM is valid based on the Q-Q plot and histogram of residuals.
- Significant heteroscedasticity (Breusch-Pagan test) indicates that the model might benefit from robust standard errors or further transformations.
- The presence of influential points (Cook’s Distance) suggests potential sensitivity to certain observations.