STAT443-Final Report

Keyu Chen, Ilia Lomasov, Brianna Suits, Nicholas Choi

Before starting the report, we would like to point out that we tried to put as much technical stuff into appendix as we could, so that the report is as easy to read as possible. However, some of the terminology should be present in the main part for demonstration and explanation purposes. We will be more than glad to clarify what is not very understandable, so please feel free to write to us anytime. Our team worked organically and the contribution of each member was important.

Context

By request of our client, a we conducted a study that tried to empirically search for parameters, varying which would reduce the emissions of nitrogen oxides, and not reduce the energy yield. The objectives were, however, not only to search for relationships of different parameters with energy yields and emissions, but also to develop a model that would balance the three main aspects:

- Simplicity of the model (for engineers to implement)
- Accuracy of influence of features on emissions
- Explanatory and prediction power of the whole model

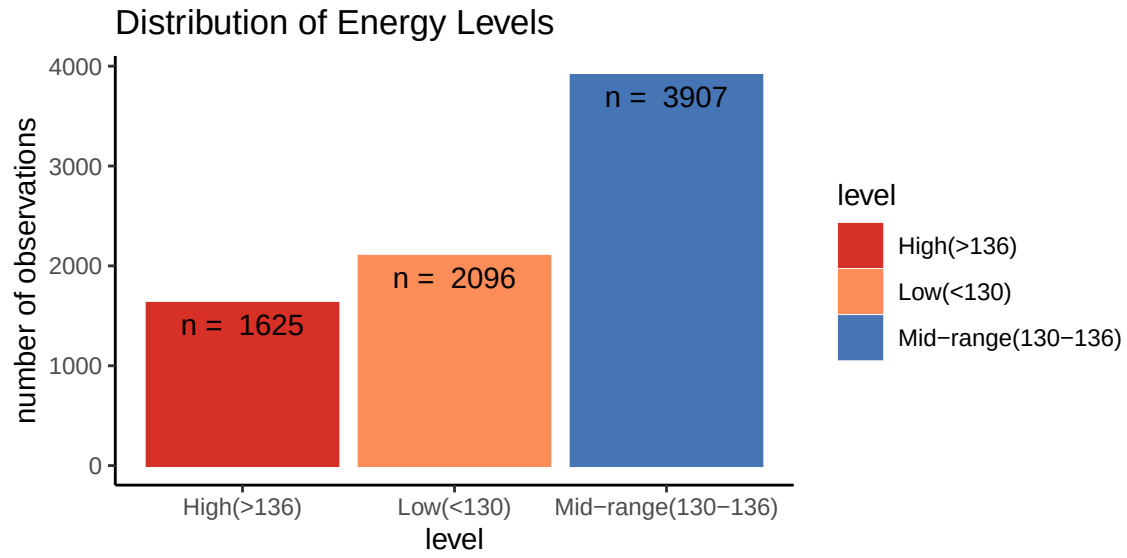
The data that was given to us contains 36733 instances of 11 sensor measures aggregated over one hour (by means of average or sum). Full description of the variables, their units of measurement and some of the descriptive statistics are presented below:

Variable	Abbr.	Unit	Min	Max	Mean
Ambient temperature	AT	°C	–6.23	37.10	17.71
Ambient pressure	AP	mbar	985.85	1036.56	1013.07
Ambient humidity	AH	(%)	24.08	100.20	77.87
Air filter difference pressure	AFDP	mbar	2.09	7.61	3.93
Gas turbine exhaust pressure	GTEP	mbar	17.70	40.72	25.56
Turbine inlet temperature	TIT	°C	1000.85	1100.89	1081.43
Turbine after temperature	TAT	°C	511.04	550.61	546.16
Compressor discharge pressure	CDP	mbar	9.85	15.16	12.06
Turbine energy yield	TEY	MWH	100.02	179.50	133.51
Carbon monoxide	CO	mg/m^3	0.00	44.10	2.37
Nitrogen oxides	NO_x	mg/m^3	25.90	119.91	65.29

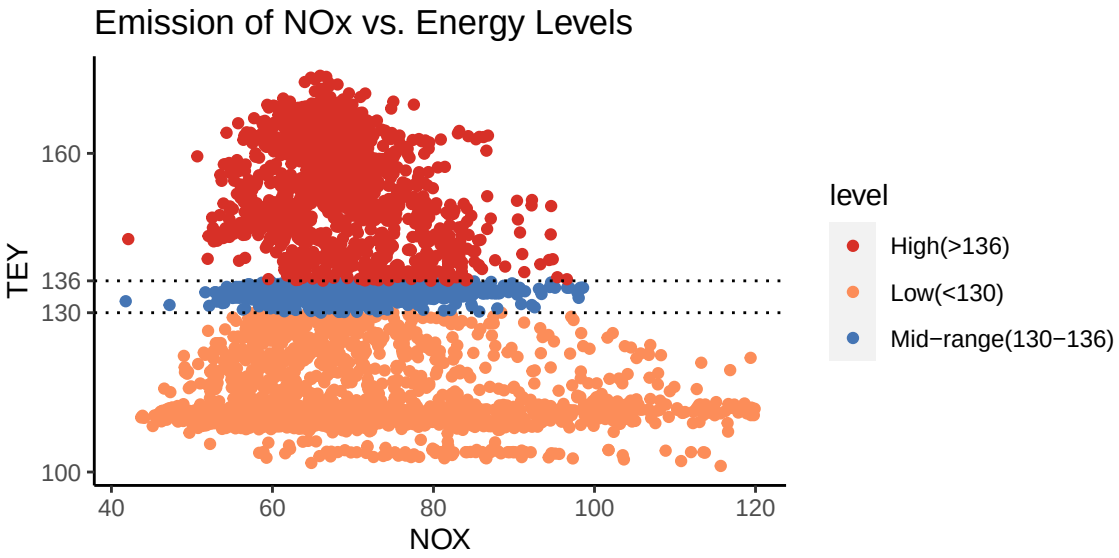
Figure 1: Data

The data was split into 4 parts on client's request: low ($TEY < 130$ [MWH]), mid-range or typical ($130 \leq TEY \leq 136$), upper mid-range ($136 < TEY < 160$) and high ($TEY \geq 160$), and a model was developed for each level separately. Running ahead, we did not form any kind of model that would describe the whole dataset, as on any range of **TEY** it would be less accurate than the best models found for those specific ranges.Furthermore, we examined the structure of the data for the upper mid-range and high range

and found them very similar, which allowed us to combine them into a single category which we called high ($TEY \geq 136$). Below is the plot that shows how many observations are in each group - significant percentage of data is within a small interval of $130 \leq TEY \leq 136$.



And below is a plot that shows how different the structure of **NOX** in relation to **TEY** is for different levels of **TEY**.



1. Mid-range data

This part is focused on the mid-range levels of Turbine energy yield. The algorithms for low, mid-range and high energy yields are very similar, however, as we already demonstrated, the structure of data are different in those four categories of **TEY**.

We start with the mid-range data, which was the main focus of the project since the very beginning.

1.1. Choice of response variable for our analysis

The regression model is used to describe individual effects of a chosen set of explanatory variables on a single response variable. [Appendix p.1]

In the model that is constructed, the interpretation of any coefficient [notated b_j] of an explanatory variable is as follows. Keeping all other explanatory variables fixed, an increase/decrease in an explanatory variable by 1 unit of its measurement results on average in an increase/decrease of response variable by b_j units of its measurement. For example,

$$NOX = 1 + 2 \cdot AT - 3 \cdot TEY$$

means that if we increase **AT** by 1 mbar and keep **TEY** constant, the corresponding increase in **NOX** will be 2 mg/m3. While if we increase **TEY** by 1 MWH and keep **AT** constant, the corresponding decrease in **NOX** will be 3 mg/m3. The former is exactly the scenario that we are looking for.

Therefore, in the setup that we have, the choice of **NOX** (Nitrogen Oxides) as a response variable was obvious - by analogy, we try to see how other features might be tweaked in order to decrease emissions (based on their coefficients) while keeping **TEY** constant.

1.2. Collinearity detection

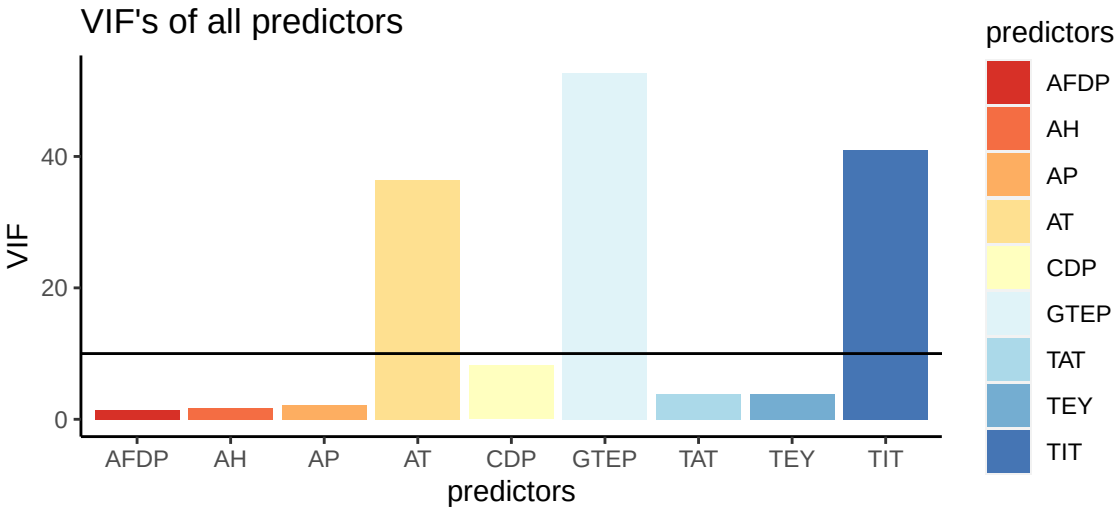
Collinearity is high linear interdependence of two or more explanatory variables. It is important to detect it since in its presence the coefficients become inaccurate, and we cannot make any conclusions based on them. Besides, collinearity is very closely related to correlation between those variables. And if two variables are highly correlated then by definition it means that it is very difficult (if not impossible) in practice to vary one of them while keeping the other one fixed.

To formally detect collinearity among the whole set of explanatory variables, we used VIF (Variance Inflation Factor) as a measure and 10 as a benchmark. [Appendix p.2]

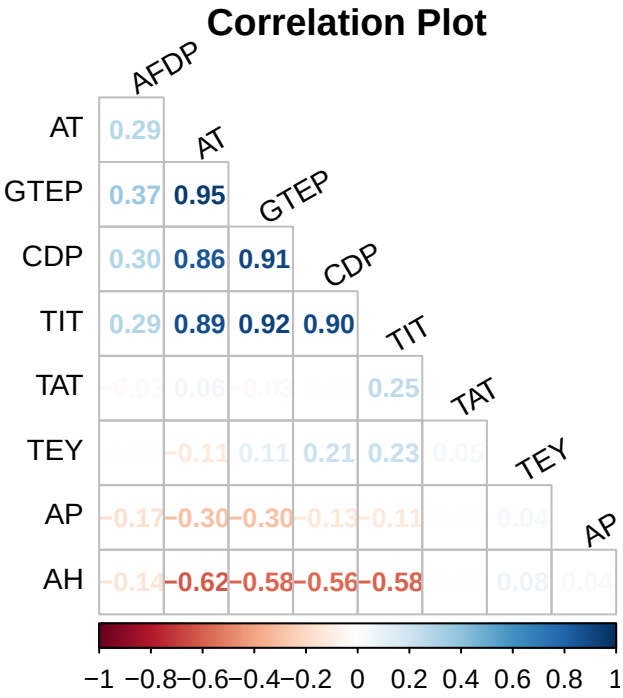
Below are the variables which do not pass the check and their VIF values:

AT	GTEP	TIT
36.44344	52.73392	40.94007

And below is the plot of VIF of all variables (with a horizontal line being the benchmark):



As we can see, our analysis detected that Ambient temperature (**AT**), Gas turbine exhaust pressure (**GTEP**) and Turbine inlet temperature (**TIT**) are very closely correlated with one another. A plot below presents the numeric values of those correlations:



Absolute value of correlation of two variables is the fraction of times they move together. Its sign is the direction in which they move together. If it is positive, they move together in the same direction. If it is negative - the opposite. Diagonal entries are always equal to one since variable moves together with itself in the same direction 100% of times. As we can see from the table, **AT** and **GTEP** move together in the same direction **95**% of times, **AT** and **TIT** move together in the same direction **89**% of times, **GTEP** and **TIT** move together in the same direction **92**% of times. In the context of our study, it means that it is nearly impossible to tweak any of those variables keeping other two fixed, and it only makes sense to change them together, looking at their coefficients in the final model.

1.3. Linear variable selection, outliers & checking assumptions

The larger the number of variables, the lower the chances that some of the features were not captured by the model. However, when there are just too many variables, overfitting problem arises. Overfitting is making the coefficients too much dependent on the existing sample. It means that if we are given a new sample (for example, a sample that the turbine sensors will collect after implementation of one some of our proposals), the errors in the analysis of this out-of-sample data will be large, because the model is not trained enough to deal with data that is not the same as it was "fed" with. Hence, we need to use a criterion that would penalize the choice of 'too many' features for analysis. For those needs, a statistic called Mallow's C_p was used. [Appendix p.3].

And the variables in model that was chosen using Mallow's C_p is presented below:

[All the variables to the right of "∼" sign are predictors. If the variable is present first with a plus and later with a minus sign, it means that it has been removed from the model.]

```
## NOX ~ (AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP + CO) -
##     CO - CDP
```

Then, detection of outliers should be done. In presence of outliers, the coefficients of the model become inaccurate. To give a simple example, if a member of your 10-people family wins $1mln in a lottery, while other 9 win nothing, it means that if you estimate average winnings, you get a result of $1mln/10=$10000. However, considering 1mln people participate, the lottery winner is just an outlier, and in reality average winnings are just $1 per person. To formally detect outliers, Cook's distance is used. Here, nothing was detected. [Appendix p.4.1]

Variable selection and estimation of the coefficients is only one part of the modeling process. Some of the coefficients might be insignificant, which means that the corresponding feature does not actually have a big enough impact on the response variable. This in turn means that the previously defined interpretation of those coefficients is invalid. And to accurately test both individual significance of a variable and joint significance of a group of variables, *homoskedasticity* and *normality* assumptions are needed [Appendix p.4.2 and p.4.3 respectively]. In other words, we need it to be 100% sure that we keep all the significant features and get rid of all the insignificant ones. (It does not affect the values of those coefficients, only their significance).

In order to test *homoskedasticity*, Breusch-Pagan and White tests were used. They both presented us with extremely strong statistical evidence that *heteroskedasticity* (opposite of *homoskedasticity*) is present in the linear model.

In order to test *normality*, Shapiro-Wilk test was used. It presented us with extremely strong statistical evidence that the assumption of *normality* is in fact violated by the linear model [Appendix p.4.4]. This test, however, cannot detect *normality* without assuming *homoskedasticity*, which is impossible by obvious reasons. Therefore, we can only use its results for comparison of different models with each other.

1.4. Nonlinear variable selection, outliers & checking assumptions

In order to see if power (or log) transformation of the response variable is appropriate, Box-Cox transformation is used [Appendix p.5.1]. And for this data, Box-Cox procedure prescribes to use ln(`NOX`) as the new response variable. This is favorable for us too since logarithms are easy to interpret. [formal interpretation is in Appendix p.5.2]

Non-formally, the interpretation is very similar to the one given for linear model (in part 1 of this report) but now the change in the response variable is not unit but percentage change.

After the transformation, the same procedure as in the linear case is done with this new model. There is still extremely strong evidence that *heteroskedasticity* is present, and that *normality* assumption is violated. However, as we said earlier, this model performs better in this regards than the previous one. [Appendix p.4.5]. Below is the model at this point: **NOXnl** (or log of **NOX**) as predictor.

```
## NOXnl ~ (AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP +
##     CO + NOX) - CO - CDP - NOX
```

1.5. Reducing the model

The model chosen at the previous step is the best one in terms of describing the given dataset, as well as predicting the levels of emissions given a new sample of turbine sensor measurements. [Appendix p.5.3]. High correlation and multicollinearity, however, causes the coefficients of the model to be inaccurate. And they might still be significant due to the sample size being large and as a result, standard errors being large. Hence it is reasonable to look how much 'worse' the model becomes if 1 or 2 of those 3 variables are removed, using information criteria. Akaike Information Criterion was used for this purpose because it is more consistent with the original criterion used (Mallow's C_p). [Appendix p.6.1].

As it stands, removing **AT** results in a very large (about 11.05%) model worsening. Removing **GTEP** results in a modest (about 1.16%) model worsening. Removing **TIT** results in a negligible (about 0.01%) model worsening. So, at the first step, **TIT** can be removed with very limited consequences. If after that **GTEP** is removed too, the cumulative worsening becomes about 2.10%, which is also acceptable, especially if simplicity of the model is highly valued. Moreover, if **GTEP** is not removed, collinearity in the model remains present, which means that individual coefficients of the model might still be inaccurate (at least more so than if **GTEP** is removed). Also, prediction power of the model as a whole is not a concern of this study, and accuracy of the coefficients is key. So, both **TIT** and **GTEP** are removed from the model at this stage. Model at this point includes the following set of variables:

```
## NOXnl ~ (AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP +
##     CO + NOX) - CO - CDP - NOX - TIT - GTEP
```

Moreover, AP and AH, despite not being correlated with the rest of the features, cannot be tweaked. So, we also checked if any/both of those variables can be removed from the model. And Removing AH results in a very large (about 25.01%) model worsening. Removing AP results in a modest (about 1.40%) model worsening. So, AP was also removed from the model.

Model at this point:

```
## NOXnl ~ (AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP +
##     CO + NOX) - CO - CDP - NOX - TIT - GTEP - AP
```

Now, in this model, TAT is not only very small by absolute value but also becomes insignificant, and the model actually improves by about 0.01% (based on the same criteria) if it is removed.

Marginal effect of each variable left in the final model on **NOXnl** is presented below:

AT	AH	AFDP	TEY
-1.5447	-0.487	2.9995	0.4883

So, the average percentage decrease in **NOX** that is associated with 1 mbar decrease of **AFDP** is approximately **3.00**%:

AFDP
2.99946

Which is quite different from **1.63**% that the non-reduced model yielded.

The same procedure (to determine prediction power) was done to this reduced model, and the results suggested that the model has only worsened by $\approx 3.70\%$ in terms of prediction power. Evidence suggests that on average predicted **NOXnl** differ from the actual ones by only $\approx 4.00\%$, which is a very slight deviation and signalizes to us that the model is still extremely good for predictions. The model was checked for robustness, and it is about as robust as the non-reduced one. [Appendix p.6.3]

2. High levels of TEY

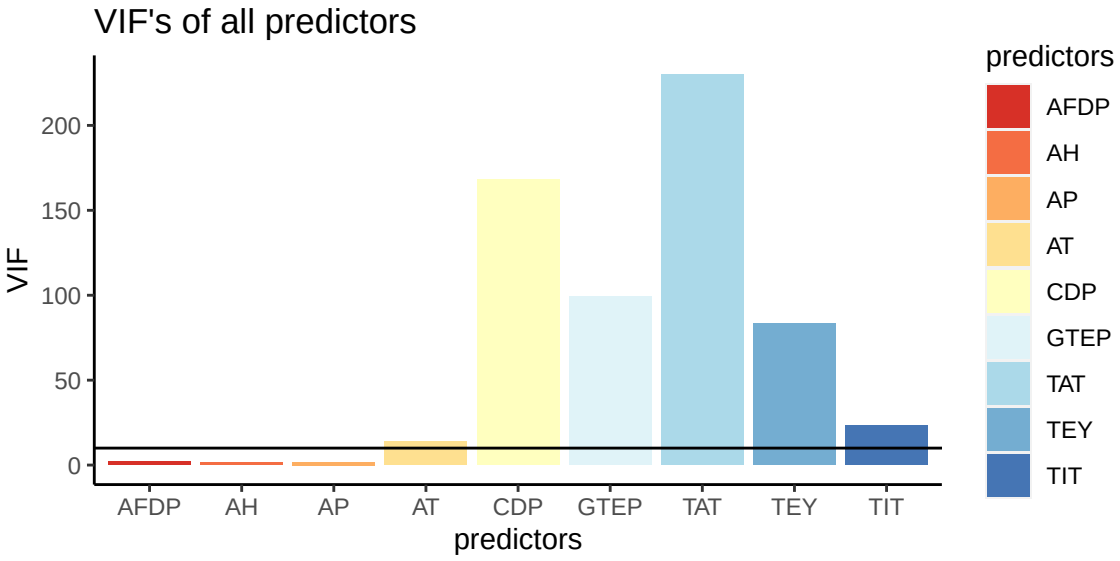
This section runs the same analysis that was done for mid-range **TEY**, hence some of the explanations are redundant. Please refer to section 1.

2.1. Collinearity detection

Below are the variables whose VIF goes beyond the benchmark of 10, and their corresponding VIF's

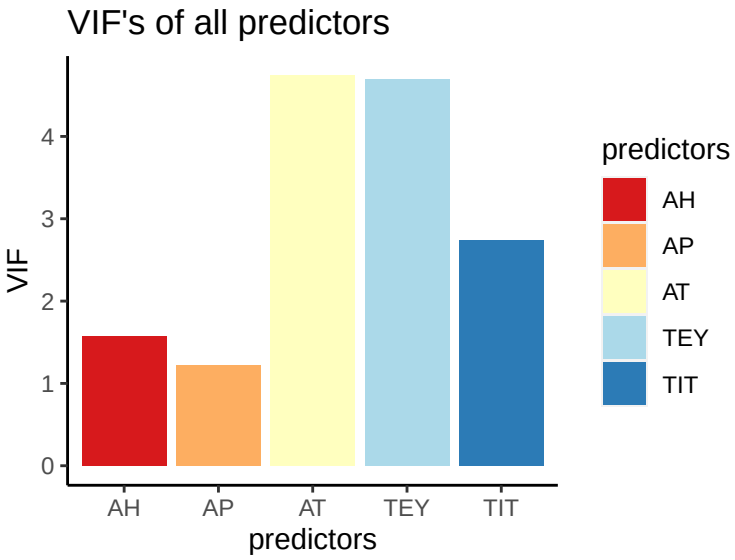
AT	GTEP	TIT	TAT	TEY	CDP
13.94014	99.17859	23.15502	229.7289	83.13647	168.1653

Evidence suggests that **AT**, **GTEP**, **TIT**, **TAT**, **TEY**, and **CDP** are collinear. Below is the plot of their VIF's, with horizontal line at $VIF = 10$ being the benchmark:



Here, the steps in finding the best model that are used before are irrelevant since one of the collinear variables is **TEY**. And it is crucial for us to find the model where **TEY** is not correlated with any other variables. So, it is reasonable to start removing other variables from the full model one by one. At each step the variable, removing which hurts the model the least (based on the same criterion as for mid-range), is removed.

At the 1st step, removing **CDP** improves the model by $\approx 0.012\%$, so, **CDP** is removed. At the 2nd step, there are no more variables, removal of which improves the model. Eliminating **ADFP** results in the least worsening though, namely 0.295%. At the 3rd step, eliminating **GTEP** results in the least worsening, 0.098%, hence, it is removed. At the 4th step, eliminating **TAT** results in the least worsening, 0.028%, hence, it is removed. At the 5th step, **TIT** is the only variable left that could be tweaked in order to reduce **NOX**. Among others, the least "important" variable is **AH**, and removing it results in a pretty significant worsening of 2.869%. Moreover, at this point collinearity is gone:



Hence, the model that is chosen for the high-range **TEY** is:

```
## NOX ~ (AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP + CO) -
##     CO - CDP - AFDP - GTEP - TAT
```

And marginal effects of each variable on **NOX** is:

AT	AP	AH	TIT
-1.0953	-0.3373	-0.1654	0.3761

Which means that keeping **TEY** fixed, a decrease of **TIT** by 1ºC is associated with an average reduction in **NOX** emissions by **0.3761** mg/m3. **AT**, **AP** and **AH**, again, can be taken into account: the model prescribes that producing energy on a day with higher ambient temperature, pressure or humidity on average means lower **NOX** emissions. There are no outliers in this model, full robustness check was done in the appendix p.7. Cross validation (also in appendix p.7) suggested that on average the model misses the predictions by approximately **13.4**% which is decent enough for this range of data.

3. Low levels of TEY

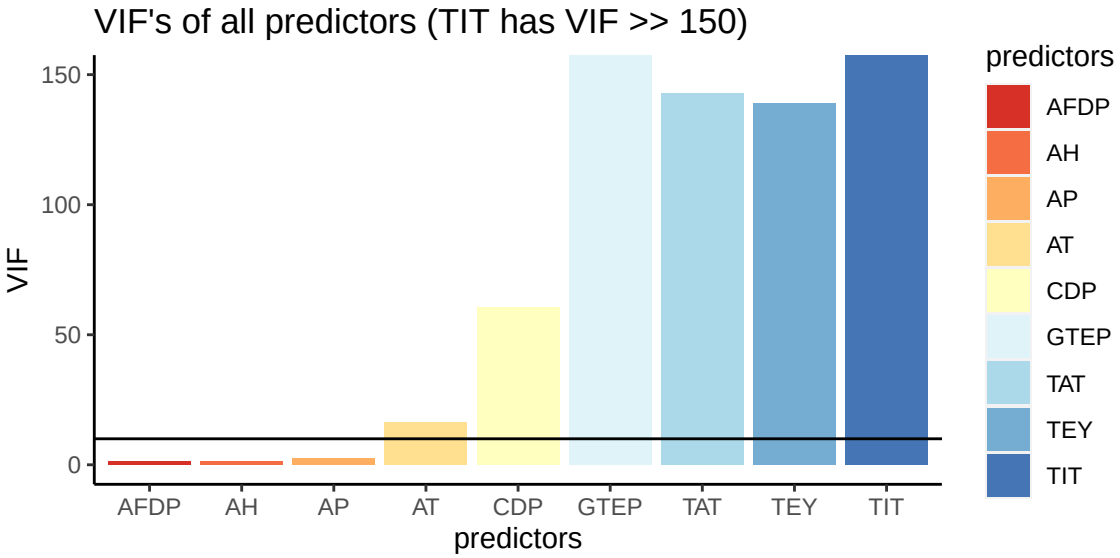
This section runs the same analysis that was done for high and mid-range **TEY**, hence some of the explanations are redundant. Please refer to section 1.

3.1. Collinearity detection

Below are the variables whose VIF goes beyond the benchmark of 10, and their corresponding VIF's

AT	GTEP	TIT	TAT	TEY	CDP
16.438	168.1873	450.0382	142.6999	138.8083	60.51543

Evidence suggests that **AT**, **GTEP**, **TIT**, **TAT**, **TEY**, and **CDP** are collinear. Below is the plot of their VIF's, with horizontal line at $VIF = 10$ being the benchmark:



Here, the steps in finding the best model that are used before are irrelevant since one of the collinear variables is **TEY**. And it is crucial for us to find the model where **TEY** is not correlated with any other variables. So, it is reasonable to start removing other variables from the full model one by one. At each step the variable, removing which hurts the model the least (based on the same criterion as for mid-range), is removed.

At the 1st step, removing **TAT** worsens the model by 0.028%, so, **TAT** is removed. At the 2nd step, **AFDP** is the one that should be romoved but it is the only tweakable variable left that is NOT collinear with **TEY**, so it cannot be removed. Eliminating **AP** results in the least worsening among the rest, namely 0.019%. At the 3rd step, eliminating **TIT** results in the least worsening, 0.214%, hence, it is removed. At the 4th step,

eliminating **CDP** results in the least worsening, 0.319%, hence, it is removed. At the 5th step, we have to remove **GTEP** (otherwise due to collinearity we would not have any results), and removing it results in a pretty significant worsening of 3.127%. Total model worsening is 3.726%.

Hence, the model that is chosen for the high-range **TEY** is:

```
## NOX ~ (AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP + CO) -
##     CO - TAT - AP - TIT - CDP - GTEP
```

And marginal effects of each variable on **NOX** is:

AT	AH	AFDP
-2.4768	-0.3318	1.1158

Which means that keeping **TEY** fixed, a decrease of **AFDP** by 1 mbar is associated with an average reduction in **NOX** emissions by **1.12** mg/m3. **AT** and **AH**, again, can be taken into account: the model prescribes that producing energy on a day with higher ambient temperature, pressure or humidity on average means lower **NOX** emissions. There are no outliers in this model, full robustness check was done in the appendix p.8. The underlying issue of this final model (for low **TEY**) is that now **TEY** is not significantly different from 0. What's more is that the model was better off if it was removed, as well as **AFDP**, during steps 3 and 4 of the elimination. Which means that although the results exists, they are not very suggestive. In absence of **TEY** in the model, the. Cross validation (also in appendix p.8) suggested that on average the model misses the predictions by approximately **8.9**% which is quite decent for this range of data, and better than for the high **TEY** data. We have to admit to a misinterpretation of given values during preparation of the presentation, where we claimed prediction error to be very high, which it apparently is not.

4. Summary of the results

In short:

1. The suggestions of the model for mid-range data about how to reduce emissions while keeping **TEY** constant are trustworthy, and the model itself is good for predicting emissions given new hypothetical data.
2. The suggestions of the model for high-range data about how to reduce emissions while keeping **TEY** constant are fairly trustworthy, and the model itself is okay for predicting emissions given new hypothetical data.
3. The suggestions of the model for high-range data about how to reduce emissions while keeping **TEY** constant are not very trustworthy, but the model itself is quite decent for predicting emissions given new hypothetical data.

More broadly:

4.1 Mid-range data

The final model for mid-range data suggests the following.

1. Suggestions on how to reduce emissions while keeping TEY fixed:

Measure	Direction of change (by 1 UoM)	Interpretation	Units of measurement (UoM)	Associated decrease in NOX
AT	up	produce on a warmer day	1ºC	1.545 %
AH	up	produce on a more humid day	1%	0.487 %

Measure	Direction of change (by 1 UoM)	Interpretation	Units of measurement (UoM)	Associated decrease in NOX
AFDP	down	decrease manually	1mbar	2.999 %

2. If used for predicting emissions using new hypothetical data, this final model on average gives the predictions that are inaccurate by about **4**% of the actual value, which is extremely good
3. In case of the non-reduced model, it cannot be used to make any suggestions but it is even better at predictions than the reduced one: its average prediction error is about **3**.**85**%.

4.2 High levels of TEY

The final model for high levels of **TEY** suggests the following.

1. Suggestions on how to reduce emissions while keeping TEY fixed:

Measure	Direction of change (by 1 UoM)	Interpretation	Units of measurement (UoM)	Associated decrease in NOX
AT	up	produce on a warmer day	1ºC	1.095 mg/m3
AP	up	produce on a day with higher atmospheric pressure	1mbar	0.337 mg/m3
AH	up	produce on a more humid day	1%	0.165 mg/m3
TIT	down	decrease manually	1ºC	0.376 mg/m3

2. In **95%** of the most common values of emissions at this level of **TEY**, if used for predicting emissions using new hypothetical data, this model on average gives the predictions that are inaccurate by about **13**.**4**% of the actual value, which is not great but acceptable for this type of model.

4.3 Low levels of TEY

The final model for low levels of **TEY** suggests the following.

1. Suggestions on how to reduce emissions while keeping TEY fixed:

Measure	Direction of change (by 1 UoM)	Interpretation	Units of measurement (UoM)	Associated decrease in NOX
AT	up	produce on a warmer day	1ºC	3.05 mg/m3
AH	up	produce on a more humid day	1%	0.282 mg/m3
AFDP	down	decrease manually	1mbar	0.142 mg/m3
GTEP	down	decrease manually	1mbar	10.496 mg/m3

2. The suggestions are not very trustworthy, and should be tested very carefully in practice, as their theoretical justification is not very strong.
3. In **95%** of the most common values of emissions at this level of **TEY**, if used for predicting emissions using new hypothetical data, this model on average gives the predictions that are inaccurate by about **8**.**9**% of the actual value, which is decent enough for this type of model.

Appendix

1. Linear regression

Assume our data is coming from the model $Y_i = \alpha + \beta_1 X_i^{(1)} + ... + \beta_p X_i^{(p)} + \epsilon_i \ , \quad 1 \leq i \leq n$, where n is the number of observations, and for i-th observation, Y_i and ϵ_i are the realizations of response variable and irreducible error respectively, $X_i^{(1)} ... X_i^{(p)}$ are the values of p different explanatory variables (or simply features). Then for each $j \in (1, p)$, $\hat{\beta}_j$ is the j-th coordinate of the vector $(\alpha \ \ \hat{\beta}_1 \ \ ... \ \ \hat{\beta}_p)^T$ that minimizes the residual sum of squares. And $\hat{\beta}_j$ is interpreted as follows:

Keeping all other variables fixed, an increase/decrease in $X_i^{(j)}$ by 1 unit of its measurement results on average in an increase/decrease of Y_i by $\hat{\beta}_j$ units of its measurement.

2. Variance inflation factor.

$VIF_j = \frac{1}{1-R_j^2}$, where R_j^2 is the percentage of variation in feature j (or $X^{(j)}$), explained by all other features. The closer R_j^2 is to 1, the larger the VIF_j, so, benchmark $VIF_j \geq 10$ detects the ones with "too high" correlation (either positive or negative) with other variables.

3. Mallow's C_p

$C_p = RSS_p + 2(p+1) - n = \sum_{i=1}^{n} (Y_i - \hat{Y_i}^{(p)})^2 + 2(p+1) - n$, where $\hat{Y_i}^{(p)}$ - fitted response variable when p features are used. C_p needs to be minimized with respect to p. The algorithm that is implemented in R tells which set of features minimizes this objective function.

4. Normality and Homoskedasticity

4.1. Outliers detection

Cook's Distance is the measure used for detecting outliers. If it is smaller than 0.5 (rule-of-thumb benchmark) for all observations, there are no problems with outliers. Here, maximal Cook's distance among all observations is:

```
## [1] 0.1344056
```

Which means there are no outliers in the linear model.

4.2. Homoskedasticity.

It is assumed that in the initial model, (from p.1)

$$Var(\epsilon) = \sigma^2 I_n \quad \text{, where } \sigma^2 > 0, \ I_n - n \times n \ \text{ identity matrix}$$

In words, it means that the errors are uncorrelated and have the same variance. This is known as constant variance or homoskedasticity. When this assumption is violated, the problem is known as heteroskedasticity. The problem with heteroskedasticity is that without constant variance assumption, neither t-tests for individual significance nor F-tests for joint significance of any group of variables are valid and can be relied on. It happens due to the fact that by their construction, test statistics for those tests are distributed the way they are only under constant variance assumption.

We use 2 diferent tests to detect Heteroskedasticity: White and Breusch-Pagan tests. They both test the same hypothesis (null - errors are uncorrelated and have constant variance, alternative - errors are correlated or have different variance). Below you can see p-values of each:

```
##         White test Breusch-Pagan test
##       1.048616e-17       1.266981e-12
```

[`Ae-B` $:= A * 10^{-B}$] Which means that both tests yield a p-value smaller than 0.001, which presents us with extremely strong statistical evidence that heteroskedasticity is present.

As a remedy for heteroskedasticity, robust (or heteroskedasticity consistent or White) standard errors [s.e.] are used instead of regular standard errors for individual significance t-tests, and Wald test (which takes into account robust s.e.) is used instead of regular F-test for joint significance.

The p-values (rounded to the nearest 5th digit in its decimal representation) of t-tests for individual significance are presented below:

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY
0	0	0	0	0	0.0357	0.0095	0

And p-value of F-test for joint significance of the model is presented below:

```
## [1] 0
```

From the two types of tests it can be seen that all of the feature coefficients (i.e., excluding intercept) are significant at at least 5% level, presenting evidence that even if constant variance assumption is violated, the chosen model is extremely significant as a whole, with each of the chosen features strongly significant individually.

4.3. Normality assumption.

Formally speaking, it is assumed that in the initial model, (from p.1)

$$\epsilon = (\epsilon_1 \ \ ... \ \ \epsilon_n)^T \sim N(0, \sigma^2 I_n) \quad \text{, where } \sigma^2 > 0, \ I_n - n \times n \ \text{ identity matrix}$$

4.4. Shapiro-Wilk test:

null hypothesis is that the distribution of errors is in fact normal. An alternative hypothesis that the distribution of errors is not normal. P-value of the test statistic for the linear model is shown below, and it is smaller than 0.001, which presents some extremely strong statistical evidence that errors are in fact not distributed normally, and that some transformations of the response variable should be done. Below is the exact p-value:

```
## [1] 3.816131e-34
```

4.5. P-value of the test statistic for the model with NOXnl

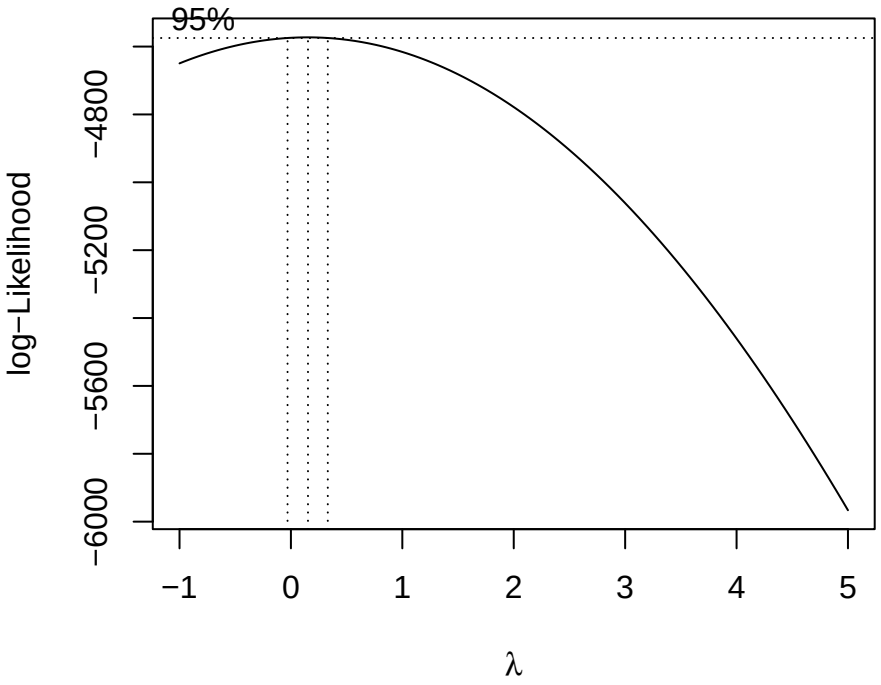
- is shown below, and it is smaller than 0.001, which presents some extremely strong statistical evidence that errors are in fact not distributed normally. However, now this evidence is weaker than it was for the fully linear model, and the normality assumption is more likely to be satisfied.

```
## [1] 0
```

5. Finding a transformation of the response variable

5.1. Box-Cox transformation.

It takes the response variable $\{Y_i\}_{i=1}^n$ and raises it to the power λ element-wise, which means that the new response is $\{Y_i^\lambda\}_{i=1}^n$. After that, it maximizes the likelihood of the data, and gives the range of possible maximizers.



In our case the interval includes $\lambda = 0$, in which case Box-Cox procedure prescribes to use $\ln(Y)$ as the new response variable.

5.2. Interpretation of the logarithm of the response variable.

Given the model $\ln(Y) = a + b \cdot X$, by our usual interpretation, $\Delta X = 1$ is associated with $\Delta\ln(Y) = b$.

$$b = \Delta\ln(Y) = \ln(Y + \Delta Y) - \ln(Y) = \ln(\frac{Y + \Delta Y}{Y}) \Rightarrow \frac{Y + \Delta Y}{Y} = e^b$$

So, interpretation of coefficients in the model $\ln(Y_i) = \alpha + \beta_1 X_i^{(1)} + ... + \beta_p X_i^{(p)} + \epsilon_i$ is as follows:

Keeping all other variables fixed, an increase/decrease in $X_i^{(j)}$ by 1 unit of its measurement results on average in an increase/decrease of Y_i by $\exp[\hat{\beta}_j] \times 100\%$ percents.

5.3. How well does the model perform as a whole?

In order to assess how well the model predicts emissions given the data that was not in the initial sample, *cross validation* is used. There are different types of cross validation, but the one that was used for this model is 10-fold cross validation. Formal steps are presented below:

1. Divide the whole set of observation into 10 samples of equal size, take the 1st one of those samples out
2. Fit the model using all other samples: $\hat{Y}_{(1)} = X_{(1)}\hat{\beta}_{(1)}$
3. Predict the value of **NOX** (Y in this case) of each observation in the left-out sample using the estimated model: $\hat{Y}_i = x_i^T \hat{\beta}_{(1)}, \quad i \in (1, n_1)$
4. Obtain prediction errors for each of those observations: $PE_i = \hat{Y}_i - Y_i, \quad i \in (1, n_1)$

5. Repeat steps 1-4 for all other samples
6. Obtain performance statistics, e.g., $MAE = \frac{1}{n}\Sigma_{i=1}^{n}|PE_i|$, $RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}PE_i^2}$
7. Assess them depending on the units of measurement of the response variable.

Now, for the model in question these statistics are presented below:

RMSE	MAE
0.0516258	0.0378091

Interpretation of RMSE and MAE is similar but MAE is easier to explain. MAE is 0.0378, which means that on average log of predicted NOx emissions is different from the actual one by 0.0378. Meaning that predicted NOx emissions on average differ from the actual ones by $(e^{0.0378} - 1) \cdot 100\% \approx 3.85\%$, which is a very slight deviation.

6. Reducing the model

6.1. AIC, BIC and Mallow's C_p

$$C_p = RSS_p + 2(p+1) - nAIC = RSS_p + 2(p+1) = C_p - nBIC = RSS_p + (p+1) \cdot \ln(n)$$

6.2. Cross Validation for reduced model

RMSE	MAE
0.0544764	0.039178

Again, by the same interpretation, predicted NOx emissions on average differ from the actual ones by $(e^{0.0392} - 1) \cdot 100\% \approx 4.00\%$, which is still very small deviation, albeit the model now has worsened by $\approx 3.70\%$.

```
round(abs(100*(0.0378-0.0392)/0.0378),2)
```

```
## [1] 3.7
```

6.3. Robustness checks for reduced model

Maximal Cook's Distance (outlier detection) is $0.036 < 0.5 \Rightarrow$ no outliers are detected:

```
## [1] 0.03636511
```

P-values of White and Breusch-Pagan tests for heteroskedasticity detection:

```
##         White test Breusch-Pagan test
##       8.999881e-26       8.318928e-18
```

Compared with the p-values for the non-reduced model:

```
##         White test Breusch-Pagan test
##       1.048616e-17       1.266981e-12
```

P-value of Shapiro-Wilk test for the reduced model:

```
## [1] 1.845426e-33
```

Compared with the p-value for the non-reduced model:

```
## [1] 3.816131e-34
```

Hence, overall, the model is less homoskedastic but errors are more likely normally distributed.

The p-values (rounded to the nearest 5th digit in its decimal representation) of t-tests for individual significance using are presented below:

AT	AH	AFDP	TEY
0	0	0	2e-05

And p-value of F-test for joint significance of the model is presented below:

```
## [1] 0
```

Which presents us with very strong statistical evidence that the chosen features are significant individually, and that the model is significant as a whole.

7. Robustness checks and cross validation for high level model

7.1. Outliers

Maximal cook's distance is less than 0.5 which means there are no outliers in the model:

```
## [1] 0.2200152
```

7.2 Normality and heteroskedasticity

P-values of White and Breusch-Pagan tests for heteroskedasticity detection:

```
##         White test Breusch-Pagan test
##       8.788597e-17       1.576328e-06
```

Compared with the p-values for the reduced mid-range model:

```
##         White test Breusch-Pagan test
##       8.999881e-26       8.318928e-18
```

Means that there is a weaker case for heteroskedasticity for high-TEY data

P-value of Shapiro-Wilk test for the reduced model:

```
## [1] 7.557905e-27
```

Compared with the p-value for the reduced mid-range model:

```
## [1] 1.845426e-33
```

Hence, overall, the model is more homoskedastic but errors are less likely normally distributed.

The p-values (rounded to the nearest 5th digit in its decimal representation) of t-tests for individual significance using are presented below:

AT	AP	AH	TIT	TEY
-1.09526	-0.33728	-0.16536	0.37609	-0.77973

And p-value of F-test for joint significance of the model is presented below:

```
## [1] 8e-306
```

Which presents us with very strong statistical evidence that the chosen features are significant individually, and that the model is significant as a whole.

7.3 Cross validation

RMSE	MAE
4.403065	3.198855

$MAE \approx 3.2$, which means that $|\hat{NOX} - NOX| \approx 3.2$

Since 95% of the values of NOX lie within 57.5 and 81.4, which is a range of length 23.9, that average deviation is $3.2/23.9 \approx 0.134$ or 13.4%. Which is not very small but not very large either.

8. Robustness checks and cross validation for high level model

8.1. Outliers

In the model that has not been reduced yet, Cook's distance is higher than 0.5 for 1 observation (#1258), which means it is likely an outlier and should be removed:

```
## [1] 1258
```

So, no outliers are left in the reduced model.

8.2 Normality and heteroskedasticity

P-values of White and Breusch-Pagan tests for heteroskedasticity detection:

```
##         White test Breusch-Pagan test
##       3.591767e-84       1.744170e-24
```

Compared with the p-values for the reduced mid-range model:

```
##         White test Breusch-Pagan test
##       8.999881e-26       8.318928e-18
```

Means that there is a weaker case for heteroskedasticity for low-TEY data

P-value of Shapiro-Wilk test for the low-TEY model:

```
## [1] 6.400777e-30
```

Compared with the p-value for the reduced mid-range model:

```
## [1] 1.845426e-33
```

Hence, overall, the model is less homoskedastic but errors are more likely normally distributed.

The p-values (rounded to the nearest 5th digit in its decimal representation) of t-tests for individual significance using are presented below:

AT	AP	AH	TIT	TEY
-1.09526	-0.33728	-0.16536	0.37609	-0.77973

And p-value of F-test for joint significance of the model is presented below:

```
## [1] 8e-306
```

Which presents us with very strong statistical evidence that the chosen features are significant individually, and that the model is significant as a whole.

8.3. Cross validation

RMSE	MAE
6.341536	4.601719

$MAE \approx 4.6$ which means that $|\hat{NOX} - NOX| \approx 4.6$

Since 95% of the values of NOX lie within 50.32 and 101.86, which is a range of length 51.54, that average deviation is 4.59/51.54. ≈ 0.089 or 8.9%. Which is quite small.