UNIVERSITY OF CALGARY

December 1, 2024

Air Under the Microscope:

Modelling Pollutant Effects on Air Quality

Group Members:

1. Harleen Kaur
2. Mackenzie Kreutzer
3. Shahzaib Rahat
4. Michael-Lee Wilson

Prepared for:

DATA 603 (L02): Statistical Modelling with Data

Air Under the Microscope: ***Modelling Pollutant Effects on Air Quality***

1. INTRODUCTION

1.1. MOTIVATION

1.1.1. Context

Clean air is essential for healthy living, and air quality directly influences both health and quality of life (Tennessee Department of Health, n.d.). Poor air quality is a major environmental risk to health and has been linked to cardiovascular and respiratory issues (NASA, n.d.). Globally, air pollution in urban and rural areas contributed to an estimated 4.2 million premature deaths in 2019, mainly due to fine particulate matter exposure, which is associated with respiratory diseases, cardiovascular conditions, and cancers (World Health Organization, n.d.).

Beyond human health, air pollution also harms the environment, causing acidification and eutrophication of water bodies, which leads to soil nutrient depletion, plant death, and harm to animal populations (NASA, n.d.). Air quality can vary significantly, even within hours, having a reliable monitoring and forecasting system is crucial. Monitoring air quality allows the public and authorities to make informed decisions, such as avoiding strenuous outdoor activities during high pollution periods or implementing policy measures to limit pollutant-producing activities (Environment and Climate Change Canada, n.d.).

In Calgary, air pollution arises from diverse sources, including vehicle emissions, industrial activities, and natural events like wildfires. Addressing these challenges requires a reliable monitoring and forecasting system. Our project aims to address these challenges by modelling pollutant effects on Calgary's air quality to provide data-driven insights for improving health outcomes and environmental sustainability.

1.1.2. Problem

The central issue is the inconsistent and incomplete data on air pollutants in Calgary. When reviewing the city's open data, it becomes apparent that some pollutants are monitored more frequently than others, creating gaps in the dataset. This raises an important question: are all the necessary pollutants being tracked to accurately measure air quality?

Additionally, there is a lack of awareness regarding how various factors, such as seasonal changes, location, and specific pollutants, impact air quality. Many individuals and policymakers may not fully understand how these elements interact. Furthermore, they may not be aware of the necessary actions to take in order to mitigate the effects of pollution, such as reducing sources like vehicle emissions, and the precautionary measures required to protect themselves, such as knowing when to avoid strenuous outdoor activities.

To address these challenges, our project aims to build a model that investigates whether pollutants with more consistent data can still provide reliable insights into air quality. We will also explore which factors most influence air quality and offer actionable recommendations, such as when precautionary measures should be taken, and which pollution sources should be reduced to improve air quality. Ultimately, our goal is to enhance air quality monitoring, forecasting, and decision-making through a data-driven approach.

1.1.3. Challenges

Challenges faced by researchers in addressing this problem include:

1. **Data Inconsistencies**: The original dataset exhibited multiple inconsistencies, such as variations in the number of pollutants reported and differences in reporting stations. Identifying a period with consistent data required a

thorough inspection and cleaning process.

2. **Time-Series Complexity**: As this is time-series data, it may violate statistical assumptions like independence due to autocorrelation. Researchers must test for these issues and apply adjustments to ensure valid modelling. Additionally, the randomness in AQI values—caused by external factors like weather—can make it difficult to build a predictive model that performs well across different conditions.

3. **Sensor Variability**: Air quality sensors may have varying levels of accuracy and calibration, introducing potential errors into the dataset. While these discrepancies are beyond the researchers' control, efforts will be made to address them to the best extent possible, such as using techniques like outlier detection and imputation.

1.2. OBJECTIVES

1.2.1. Overview

The intent of this project is to deliver comprehensive insights into Calgary's air quality. Specifically, it aims to identify which pollutants are most effective for reliably predicting air quality and to equip individuals and policymakers with actionable information to make informed decisions. By improving air quality monitoring, forecasting, and decision-making, this project aspires to foster healthier communities and support sustainable environmental practices.

1.2.2. Goals & Research Questions

Goals:

1. **Develop a Predictive Model**: Create a model that leverages consistent pollutant data to accurately assess and forecast air quality.
2. **Identify Key Influences**: Pinpoint the pollutants, seasonal patterns, and geographical factors that most significantly affect air quality in Calgary.
3. **Recommend Actionable Solutions**: Provide evidence-based guidance to reduce pollution sources and suggest precautionary measures to mitigate health risks.
4. **Increase Awareness**: Educate individuals and policymakers about the health implications of poor air quality and promote preventive, informed behaviours.

Research Questions:

1. **Which months have the best and worst air quality?**
   Aim: To identify seasonal trends in air quality, which can inform recommendations for scheduling outdoor activities during times of lower pollution.

2. **Which communities have the best and worst levels of air quality?**
   Aim: To highlight areas that may require additional air quality management and help individuals with respiratory conditions make informed decisions about where to live or engage in outdoor activities.

3. **What factors contribute most to poor air quality in Calgary or specific communities?**
   Aim: To provide policymakers with insights on which pollutants or activities have the most impact on air quality, enabling targeted interventions to reduce pollution.

4. **Can the model predict future Air Quality?**
   Aim: To explore if historical AQI and pollutant data can help forecast future air quality levels, potentially leading to a useful tool for anticipating high-risk pollution days.

2. METHODOLOGY

2.1 Data

The dataset we are using is in CSV format and contains daily time-series records from 1993 to 2019 (City of Calgary, n.d.-a). It provides information on air quality across various regions in Calgary, Alberta, Canada. Key variables include the Air Quality Index (AQI) and levels of more than ten pollutants, such as particulate matter (PM), nitrogen oxides (NOx), ozone (O_3), sulfur oxides (SOx), and carbon monoxide (CO).

For our analysis, we focused on cleaning and preparing the data for meaningful insights. To ensure consistency, we selected the most recent period (2015 to 2018) with the least amount of missing data—approximately 15% overall. This narrowed the dataset to three years and nine pollutants across three regions: Northwest, Southeast, and Central Calgary. To account for seasonal impacts, we added a column indicating the season for each record.

Since most of the data was skewed, as observed during chart analysis, we decided to impute missing values using the median. This was done by grouping the data by season and station name (region) to ensure that imputation captured local and seasonal variability effectively.

Variable Types and Measurements:

1. **Air Quality Index (AQI):** This is a quantitative variable representing the air quality on a standardized scale from 1 to 10, where 1 indicates the best air quality and 10 indicates the worst. AQI is calculated based on the concentration of pollutants and serves as our response variable.
2. **Pollutant Levels:** The pollutant data are also quantitative variables and are recorded in various units depending on the pollutant type. The nine pollutants and their corresponding measurement units are presented in **Table 1** below:

Table 1: *Pollutants and Their Measurement Units*

Pollutant	Measurement Unit
PM2.5 Mass	Micrograms per cubic meter (μg/m³)
Carbon Monoxide	Parts per million (ppm)
Nitrogen Dioxide	Milligrams per litre (mg/L)
Nitric Oxide	Parts per billion (ppb)
Ozone	Parts per billion (ppb)
Methane	Parts per million (ppm)
Total Hydrocarbons	Parts per million (ppm)
Total Oxides of Nitrogen	Parts per million (ppm)
Non-methane Hydrocarbons	Parts per million (ppm)

3. **Regions in Calgary:** his is a qualitative variable that identifies different areas within Calgary: Northwest, Southeast, and Central Calgary. This variable allows us to examine regional variations in air quality.
4. **Seasons:** Qualitative, includes all four Winter, Summer, Fall and Spring, helping to analyze seasonal variations in air quality, as pollutant levels differ across these periods.

In our analysis, **pollutants**, **seasons**, and **regions** will serve as the predictor variables, while the **AQI** will be the response variable we aim to model and forecast. By examining pollutant levels along with seasonal and regional data, we aim to

identify key patterns and factors that drive variations in air quality.

2.2 Approach

The dataset was initially cleaned using Python, as it is an effective tool for handling and cleaning large datasets. The data was narrowed down to a more consistent timeframe, focusing on the period with the least amount of missing data. Missing values were then imputed to ensure a complete and reliable dataset for analysis.

Our group will use R for the data and visual analytics solution, as it provides the necessary statistical functions for analyzing the relationships in our dataset. Specifically, we will apply a multiple linear regression model, using relevant predictors (e.g., pollutant levels, seasonal data, and regional variations) to examine their impact on the Air Quality Index (AQI), our response variable. We will use stepwise regression and an All-Possible-Regressions Selection Procedure to determine our initial model, as we have a large number of predictors, and this is an efficient way to create the best additive model.

We will test for multicollinearity to ensure that we are not overfitting our model. We will also include interaction terms to capture joint effects between predictors and higher-order terms to model non-linear relationships between predictors and the outcome. This can help the model explain more of the variation in the response variable. Testing these assumptions is a critical part of ensuring that the model provides reliable and accurate predictions. Additionally, we will test the assumptions as we fit the different models to ensure they do not violate the conditions required for using the statistical functions. Doing the following steps ensures that the model provides reliable and accurate predictions.

This approach will work well because if a predictor has a significant coefficient, it indicates a meaningful relationship with AQI, and the magnitude of the coefficient will show the strength of this impact. Conversely, if a predictor is removed from the model and the results do not significantly change, it suggests that the predictor has little or no influence on air quality. This model will allow us to identify the key factors affecting air quality and provide actionable insights for our research questions.

2.3 Workflow

What steps (workflow task list) are required? Which of these steps is particularly hard? What to do if the hard steps don't work out

1. **Model Selection:**
    - **Best Additive Model:**
        - Use **All-Possible-Regressions Selection Procedure** and **Stepwise Selection Procedure** to find the optimal model.
2. **Multicollinearity Check:**
    - **Test for Multicollinearity** to identify and remove redundant predictors.
3. **Assumptions Check:**
    - **Linearity, Independence, Normality, and Heteroscedasticity:**
        - Check these assumptions, applying transformations as needed, such as $\mathbf{X^2}$ and **log(X)**.
4. **Addressing Assumption Violations:**
    - **Normality and Heteroscedasticity:**
        - Investigate **influential outliers** to improve the model.
        - Consider adding **interaction terms** to account for additional variations in the response variable.
5. **Transformation and Model Updates:**
    - For each transformation or addition of a new feature:
        - **Re-check assumptions** to ensure they still hold.

Real-life datasets often don’t meet the assumptions required for traditional regression, such as linearity, normality, or constant variance of errors. These violations can pose challenges and may require ongoing adjustments throughout the

analysis process. When assumptions are violated—such as the presence of outliers or heteroscedasticity—**Robust Regression** can be used as an alternative. This method is designed to provide reliable estimates even in the presence of assumption violations, offering a more flexible approach that remains effective when traditional regression methods would fail due to issues like non-normality or non-constant variance of errors (UCLA, n.d.).

2.4 Contributions

For an effective project, clear collaboration and well-defined responsibilities are essential. The responsibilities for each team member are as follows:

Harleen Kaur & Shahzaib Rahat
Responsibilities: Both Harleen and Shahzaib worked together on the model-building process, including developing the regression model, assessing its performance, and fine-tuning it to ensure accurate predictions.

Mackenzie Kreutzer & Michael-Lee Wilson
Responsibilities: Mackenzie and Michael-Lee collaborated on interpreting the results and discussing how they addressed the research questions. Michael-Lee also managed data collection and preprocessing tasks, such as cleaning the data, handling missing values and outliers, and performing feature engineering.

3 MAIN RESULTS OF THE ANALYSIS

3.1 Results

To identify the best regression model that can be used to answer the research questions identified in Section 1, various steps needed to be completed, starting with model selection and ensuring that the model satisfies various assumptions.

3.1.1 **Model Selection**

To start our model selection, we identified the full additive model. In this model, we had our dependent predictor variable as the Air Quality Index score, with the other variables as our independent variables. This model can be written as:

Air Quality Index = 0.789 - 0.041X(Calgary Northwest) - 0.065X(Calgary Southeast) + 0.001X(Spring) + 0.017X(Summer) + 0.012X(Winter) + 0.520X(Carbon Monoxide) + 8.468X(Methane) - 8.954X(Nitric Oxide) + 7.040X(Nitrogen Dioxide) + 40.166X(Ozone) + 8.681X(Non-methane Hydrocarbons) + 0.036X(PM2.5 Mass) - 8.371X(Total Hydrocarbons) + 12.807X(Total Oxides of Nitrogen)

Using this full additive model we will use the following hypothesis to identify any insignificant variables.

$$H0 : \beta1 = \beta2 = \beta3 = \ldots = \beta p = 0$$
$$H1 : \text{at least one of } \beta i \text{ is not zero}$$

The standard alpha used to conduct this t-test is $\alpha = 0.05$. Using this alpha we identified that all of the season variables had a p-value greater than our alpha. This indicates that we should fail to reject H0 for the season variable as we do not have enough evidence to conclude that Season contributes significantly to predicting the Air Quality Index.

3.1.1.1 **Best Additive Model: Stepwise Selection Procedure**

To confirm that Season should be removed from the model, a stepwise selection procedure was preformed with the alpha to enter the model set at $\alpha = 0.05$, and the alpha to leave the model set at $\alpha = 0.1$. Our hypothesis for this test was as

follows:

$$H0 : \beta1 = \beta2 = \beta3 = \ldots = \beta p = 0$$
$$H1 : \text{at least one of } \beta i \text{ is not zero}$$

Through this process, there were eleven variables identified, and ten steps were performed. Within these ten steps, the final model that was determined included all variables, except for Season, which is the same as what was seen with the t-test done above. This means that for the variable of season, we should fail to reject H0 as we do not have enough evidence to conclude that Season contributes significantly to predicting the Air Quality Index.

The reduced model with season removed based on both the t-test and the stepwise procedure is as follows:

> **Air Quality Index** = 0.799 - 0.041X(Calgary Northwest) - 0.066X(Calgary Southeast) + 0.524X(Carbon Monoxide) + 8.537X(Methane) - 8.291X(Nitric Oxide) + 7.573X(Nitrogen Dioxide) + 40.150X(Ozone) + 8.770X(Non-methane Hydrocarbons) + 0.036X(PM2.5 Mass) - 8.440X(Total Hydrocarbons) + 12.144X(Total Oxides of Nitrogen)

3.1.1.2 **Best Additive Model: All-Possible-Regressions Selection Procedure**

Next, All-Possible-Regressions Selection Procedure was conducted on the full model to confirm that the same variables should be removed. Through the matrix that was produced (Appendix A), the general trend is that the adjusted R squared value increased as predictors are added until predictor 12, where the increase was minimal. This means that the model explains a higher proportion of variance as predictors are added to the model.

The lowest BIC value is present at row 8 (Appendix A), sitting at -7363.624, suggesting that the model with 8 predictors is the most parsimonious. The Cp value approaches the number of predictors plus one, thirteen, at predictor 12 where the Cp sits at 12.814 indicating the model is balanced (Appendix A).

Based on the above as we are wanting less variance in our model, we want to go with the higher adjusted R squared value. As the only difference between predictor 12, 13, and 14 is the inclusion of SeasonSpring, SeasonSummer and SeasonWinter, we should keep the variable of Season in our model.

With the variable of Season included back in our model using this method, our model can be written out again as:

> **Air Quality Index** = 0.789 - 0.041X(Calgary Northwest) - 0.065X(Calgary Southeast) + 0.001X(Spring) + 0.017X(Summer) + 0.012X(Winter) + 0.520X(Carbon Monoxide) + 8.468X(Methane) - 8.954X(Nitric Oxide) + 7.040X(Nitrogen Dioxide) + 40.166X(Ozone) + 8.681X(Non-methane Hydrocarbons) + 0.036X(PM2.5 Mass) - 8.371X(Total Hydrocarbons) + 12.807X(Total Oxides of Nitrogen)

3.1.1.3 **Best Additive Model: ANOVA Table**

The two above methods gave two separate models that we could use moving forward. In order to narrow it down to one model we created an ANOVA table to evaluate (Appendix B). Model 1 is the model without the variable of Season, and model 2 is the full model that includes the variable of Season.

Using the ANOVA table we can test:

$$H0 : \beta1 = \beta2 = \beta3 = \ldots = \beta p = 0$$
$$H1 : \text{at least one of } \beta i \text{ is not zero}$$

The p-value for the added variable of Season is 0.1024 which is greater then our standard alpha of $\alpha = 0.05$ (Appendix B). This means we should reject the null hypothesis as at least one predictor is not zero and contributes to the model. We will

be using the model with Season removed as we preform our assumption tests.

Our final additive model is as follows:

> **Air Quality Index** = 0.799 - 0.041X(Calgary Northwest) - 0.066X(Calgary Southeast) + 0.524X(Carbon Monoxide) + 8.537X(Methane) - 8.291X(Nitric Oxide) + 7.573X(Nitrogen Dioxide) + 40.150X(Ozone) + 8.770X(Non-methane Hydrocarbons) + 0.036X(PM2.5 Mass) - 8.440X(Total Hydrocarbons) + 12.144X(Total Oxides of Nitrogen)

3.1.2 **Multicollinearity**

Our first test is to test for multicollinearity which checks to see if we have any redundant predictors. The Multicollinearity test provided us with an output of graphs that compared variables in our model (Appendix C). From this graphical output we determined that the variables Methane, Nitric.Oxide, Nitrogen.Dioxide, Non.methane.Hydrocarbons, Total.Hydrocarbons and Total.Oxides.Of.Nitrogen show high collinearity (Appendix C).

Due to this, we reduced the model to:

> **Air Quality Index** = 0.880 - 0.037X(Calgary Northwest) - 0.049X(Calgary Southeast) + 0.547X(Carbon Monoxide) + 39.032X(Ozone) + 0.036X(PM2.5 Mass) + 0.104X(Total Hydrocarbons) + 9.970X(Total Oxides of Nitrogen)

In doing this reduction, we noticed that there was a small change in our Adjusted R squared value, it decreased from 0.8859 in our additive model to 0.875 in our reduced model.

We then created another reduced model that is as follows:

> **Air Quality Index** = 0.826 - 0.044X(Calgary Northwest) - 0.064X(Calgary Southeast) + 0.539X(Carbon Monoxide) + 0.087X(Methane) + 4.059X(Nitric Oxide) + 18.880X(Nitrogen Dioxide) + 40.004X(Ozone) + 0.452X(Non-methane Hydrocarbons) + 0.036X(PM2.5 Mass)

This model has an Adjusted R squared of 0.8851 which is higher then the first reduced model, and not significantly different from the final additive model's Adjusted R squared of 0.8859. Due to this, this is the model we will be continuing with.

We then tested it against VIF for both of the reduced models (Appendix D and E) and confirmed that reduced model 2 should be the model we use going forward. Our final model at this point is as follows:

> **Air Quality Index** = 0.826 - 0.044X(Calgary Northwest) - 0.064X(Calgary Southeast) + 0.539X(Carbon Monoxide) + 0.087X(Methane) + 4.059X(Nitric Oxide) + 18.880X(Nitrogen Dioxide) + 40.004X(Ozone) + 0.452X(Non-methane Hydrocarbons) + 0.036X(PM2.5 Mass)

3.1.3 **Assumptions Check**

The next step is to check assumptions on our new reduced model. We need to check assumptions for linearity, indpendence, normality and heteroscedasticity to ensure that our model reflects an accurate relationship.

3.1.3.1 **Independence**

To test independence, we examed a plot to see if there is a visible trend, cyclic pattern or clustering which would indicate that the residuals are not independent (Appendix F). From the graph output, there is no discernable patter so we can conclude that the assumption of linearity is met.

3.1.3.2 **Linearity**

We can use the same graph (Appendix F) to test linearity. As there is no discernable pattern, we can conclude that the assumption of linearity is met.

We also conducted a GGpairs Plot (Appendix G). We can see there is a very small curvature in the plot, which seems to occur near the outliers. This confirms that the assumption of linearity is met as there is not a large curve.

3.1.3.3 **Heteroscedasticity**

Testing Heteroscedasticity we use the same chart as Independence (Appendix F) as well as another chart output (Appendix H). Both of these outputs show that the scale-location plot is quite horizontal and there is no funneling in the residual plot. This confirms that the assumption of Heteroscedasticity is met.

We also performed the Breush-Pagen Test to formally test whether the variance of residuals will depend on the values of one or more predictors. Our hypothesis for this test is as follows:

H0 : The variance of the residuals is constant (homoscedasticity)
H1 : The variance of the residuals is not constant (heteroscedasticity is present)

Running the Breush-Pagen test on our model gives us a p-value of 2.2e-16 meaning that we can reject the null hypothesis with an alpha of $\alpha = 0.05$. We can conclude that we do have heteroscedasticity, and can try an higher order model to remove this issue.

3.1.3.4 **Normality**

A histogram was created to test this assumption (Appendix I). Based on the histogram, it appears that the data is normal and the assumption is met. However, we also created a QQ plot to test this assumption (Appendix J). The QQ plot shows a curve at the beginning and the end, indicating that the data may not lay fully normal.

We then conducted a Sharpio-Wilk normality test to see if the data is normal. Our hypothesis for this test is as follows:

H0 : The data follows a normal distribution
H1 : The data does not follow a normal distribution

Running the Sharpio-Wilk test on our model gives us a p-value of 2.2e-16 against an alpha of $\alpha = 0.05$. This means we fail to reject null hypothesis and conclude that the data is not normal.

3.1.4 **Addressing Assumption Violations**

To address the heteroscedasticity and normality assumptions we tried to correct both with a boxcox model. Once this was completed, we performed both the Breush-Pagen Test and the Sharpio-Wilk test again.

For the Breush-Pagen test the hypothesis is as follows:

H0 : The variance of the residuals is constant (homoscedasticity)
H1 : The variance of the residuals is not constant (heteroscedasticity is present)

And for the Sharpio-Wilk test the hypothesis is as follows:

H0 : The data follows a normal distribution
H1 : The data does not follow a normal distribution

After running both tests again after the Boxcox model, both tests gave p-values of 2.2e-16 against an alpha of $\alpha = 0.05$ which means we reject the null hypothesis for the Breush-Pagen test and conclude that the data still has heteroscedasticty. We also conclude that the data does not follow a normal distribution and we can reject the null hypothesis.

3.1.4.1 **Removing Influential Outliers**

We then tried transforming the predictors to identify which variables had heteroscedasticty. Using an alpha of $\alpha = 0.05$, we concluded that the variables carbon.monoxide, nitric.oxide, nitrogen.dioxide, ozone, and non.methane.hydrocarbons which all had a p-value greater then 0.05.

We attempted to solve this issue and failed, so our next step was to investigate the outliers to try to improve this model. To do this we created a Cook's Distance Plot (Appendix K) as well as a leverage plot (Appendix L). We could see the outliers present in these two plots. Taking this, we calculated leverage and Cook's distance to identify the influential points, and then removed them. Once they were removed, we refit the model with the outliers removed. The new model equation is as follows:

> **Air Quality Index** = 0.76 - 0.032X(Calgary Northwest) - 0.050X(Calgary Southeast) + 0.335X(Carbon Monoxide) + 0.0931X(Methane) + 7.499X(Nitric Oxide) + 19.866X(Nitrogen Dioxide) + 42.695X(Ozone) + 0.384X(Non-methane Hydrocarbons) + 0.033X(PM2.5 Mass)

We did a t-test on this model to ensure that all the variables remained significant with an alpha of $\alpha = 0.05$. We can confirm that the p-values were all below this alpha and they remained significant. We then reran the Cook's Distance and leverage plots to see if there were any other influential outliers, and there were none present.

After we removed the influential outliers, we reran our assumption tests to confirm that independence, linearity, heteroscedasticity and normality assumptions are all met. Independence and linearity both passed, however heteroscedasticity and normality both failed again.

When we ran the Breush-Pagen test on the model with the influential terms removed, we used the following hypothesis:

H0 : The variance of the residuals is constant (homoscedasticity)
H1 : The variance of the residuals is not constant (heteroscedasticity is present)

The p-value was 2.2e-16 still which against an alpha of $\alpha = 0.05$ means that we reject the null hypothesis and conclude that heteroscedasticity is still present.

The Sharpio-Wilk normality test was also run again against an alpha of $\alpha = 0.05$, and our hypothesis was as follows:

H0 : The data follows a normal distribution
H1 : The data does not follow a normal distribution

The p-value on this test is 0.00398 meaning that we reject the null hypothesis and conclude that the data does not follow a normal distribution.

We attempted to run a boxcox on the model with the influential terms removed, and then ran both the Breush-Pagen and Sharpio-Wilk test again, but received the same result.

3.1.4.2 **Interaction Terms**

As removed influential outliers did not change the results on our assumptions, we started adding interaction terms. Our first interaction model included all of the possible interaction terms. We conducted a t-test based on an alpha of $\alpha = 0.05$ and our hypothesis for this test is as follows:

$$H0 : \beta1 = \beta2 = \beta3 = \ldots = \beta p = 0$$
$$H1 : \text{at least one of } \beta i \text{ is not zero}$$

After we conducted this test we could reject the null as at least one of the betas were not zero. We then took the significant interactions which had a p-value below the alpha and created a new model.

Air Quality Index = −0.150 + 0.180X(Calgary Northwest) + 0.055XCalgary Southeast − 0.164X(Carbon Monoxide) + 0.493X(Methane) + 12.41X(Nitric Oxide) + 53.50X(Nitrogen Dioxide) + 73.15X(Ozone) − 0.865X(Non-methane Hydrocarbons) + 0.022X(PM2.5 Mass) + 0.178X(Calgary Northwest * Carbon Monoxide) + 0.313X(Calgary Southeast * Carbon Monoxide) − 11.46X(Calgary Northwest * Nitrogen Dioxide) − 8.71X(Calgary Southeast * Nitrogen Dioxide) − 0.564X(Calgary Northwest * Nitric Oxide) + 2.793X(Calgary Southeast * Nitric Oxide) − 3.912X(Calgary Northwest * Ozone) − 2.454X(Calgary Southeast * Ozone) + 1.948X(Calgary Northwest * Non-methane Hydrocarbons) + 0.910X(Calgary Southeast * Non-methane Hydrocarbons) − 0.00385X(Calgary Northwest * PM2.5 Mass) − 0.00175X(Calgary Southeast * PM2.5 Mass) − 18.83X(Carbon Monoxide * Nitric Oxide) + 27.69X(Carbon Monoxide * Nitrogen Dioxide) − 8.646X(Carbon Monoxide * Ozone) + 0.0448X(Carbon Monoxide *PM2.5 Mass) − 1.355X(Methane * Nitric Oxide) − 11.78X(Methane * Nitrogen Dioxide) − 11.30X(Methane * Ozone) + 173.2X(Nitric Oxide * Ozone) - 518.7X(Nitrogen Dioxide * Ozone) + 0.0186X(Non-methane Hydrocarbons * PM2.5 Mass)

We then compared the reduced model to the model with the interaction terms in an ANOVA Table (Appendix M). Based on this ANOVA table we can conclude that with an alpha of $\alpha = 0.05$, the p-value of 2.2e-16 indicates that the interactive model significantly improves the model fit over the first additive model.

After this interactive model was conducted, we then tested for the assumptions of independence, linearity, heteroscedasticity and normality.

Independence passed with a residuals over time plot (Appendix N). We could see that the linearity assumption did not pass with a ggpairs plot as there was some curvature (Appendix O). As there was some curvature, we went to evaluate which variables could be causing us to not meet the linearity assumption (Appendix P). Based on these graphs we rose some of the terms to the higher order and tested their significance against an alpha of $\alpha = 0.05$. Our new model with significant higher order terms included is as follows:

Air Quality Index = 1.078 + 0.113X(Calgary Northwest) + 0.070XCalgary Southeast − 0.276X(Carbon Monoxide) − 0.188X(Methane) + 0.084X(Methane^2) + 4.063X(Nitric Oxide) + 47.48X(Nitrogen Dioxide) + 33.87X(Ozone) + 294.6X(Ozone^2) − 0.510X(Non−methane Hydrocarbons) + 0.0205X(PM2.5 Mass) + 0.126X(Calgary Northwest * Carbon Monoxide) + 0.294X(Calgary Southeast * Carbon Monoxide) − 8.513X(Calgary Northwest * Nitrogen Dioxide) − 8.813X(Calgary Southeast * Nitrogen Dioxide) − 0.393X(Calgary Northwest * Nitric Oxide) + 2.295X(Calgary Southeast * Nitric Oxide) − 2.2412X(Calgary Northwest * Ozone) − 3.324X(Calgary Southeast * Ozone) + 1.858X(Calgary Northwest * Non−methane Hydrocarbons) + 0.722X(Calgary Southeast * Non−methane Hydrocarbons) − 0.00429X(Calgary Northwest * PM2.5 Mass) − 0.00074X(Calgary Southeast * PM2.5 Mass) −

13.91X(Carbon Monoxide * Nitric Oxide) + 16.59X(Carbon Monoxide * Nitrogen Dioxide) + 0.0503X(Carbon Monoxide *PM2.5 Mass) − 7.841X(Methane * Nitrogen Dioxide) + 348.2X(Nitric Oxide * Ozone) − 513.0X(Nitrogen Dioxide * Ozone)

Using this model with higher order terms we created a ANOVA table that compared this to the reduced interaction model (Appendix Q). Using an alpha of $\alpha = 0.05$ the p-value for the model with the higher order terms is 0.0001675 meaning we should go with the interaction and higher order term model.

After raising terms to the higher order we re-ran the linearity test and fixed our curvature issue (Appendix R). We can now fail to reject the null hypothesis and conclude that the data is linear. After running the heteroscedasticity test again, we reject the null hypothesis still and conclude that there is still heteroscedasticity as the p-value on the Breush-Pagan test is still 2.2e-16.

In conclusion, even after completing the boxcox on our new higher order model, we were still unable to conclude that the data was both normal and did not have heteroscedasticity. Our assumption is that this is due to the number of zeros that are present in the data set.

Due to the failure of assumptions we used a robust regression model to provide reliable estimates and offering a more flexible approach (UCLA, n.d.). After using robust regression, our final model is as follows:

Air Quality Index = 1.067 + 0.1269X(Calgary Northwest) + 0.05803XCalgary Southeast − 0.5202X(Carbon Monoxide) − 0.1380X(Methane) + 0.07381X(Methane^2) + 7.671X(Nitric Oxide) + 44.78X(Nitrogen Dioxide) + 33.51X(Ozone) + 315.9X(Ozone^2) + 0.02846X(Non−methane Hydrocarbons) + 0.01117X(PM2.5 Mass) + 0.1591X(Calgary Northwest * Carbon Monoxide) + 0.4528X(Calgary Southeast * Carbon Monoxide) − 10.62X(Calgary Northwest * Nitrogen Dioxide) + 11.05X(Calgary Southeast * Nitrogen Dioxide) − 0.01702X(Calgary Northwest * Nitric Oxide) + 3.422X(Calgary Southeast * Nitric Oxide) − 2.387X(Calgary Northwest * Ozone) − 3.303X(Calgary Southeast * Ozone) + 1.490X(Calgary Northwest * Non−methane Hydrocarbons) + 0.1857X(Calgary Southeast * Non−methane Hydrocarbons) − 0.004302X(Calgary Northwest * PM2.5 Mass) − 0.0001203X(Calgary Southeast * PM2.5 Mass) − 27.94X(Carbon Monoxide * Nitric Oxide) + 31.91X(Carbon Monoxide * Nitrogen Dioxide) + 0.07115X(Carbon Monoxide *PM2.5 Mass) − 7.715X(Methane * Nitrogen Dioxide) + 417.0X(Nitric Oxide * Ozone) − 508.1X(Nitrogen Dioxide * Ozone)

The submodel for Station.Name = Calgary Northwest is:

Air Quality Index = 1.1939 − 0.6793X(Carbon Monoxide) − 0.1380X(Methane) + 0.07381X(Methane^2) + 7.6583X(Nitric Oxide) + 34.16X(Nitrogen Dioxide) + 31.123X(Ozone) + 315.9X(Ozone^2) + 1.5185X(Non−methane Hydrocarbons) + 0.0069X(PM2.5 Mass) − 27.94X(Carbon Monoxide * Nitric Oxide) + 31.91X(Carbon Monoxide * Nitrogen Dioxide) + 0.07115X(Carbon Monoxide *PM2.5 Mass) − 7.715X(Methane * Nitrogen Dioxide) + 417.0X(Nitric Oxide * Ozone) − 508.1X(Nitrogen Dioxide * Ozone)

The submodel for Station.Name = Calgary Southwest is:

Air Quality Index = 1.1250 − 0.0674X(Carbon Monoxide) − 0.1380X(Methane) +

0.07381X(Methane^2) + 11.093X(Nitric Oxide) + 33.73X(Nitrogen Dioxide) + 30.206X(Ozone) + 315.9X(Ozone^2) + 0.02141 6X(Non−methane Hydrocarbons) + 0.01105X(PM2.5 Mass) − 27.94X(Carbon Monoxide * Nitric Oxide) + 31.91X(Carbon Monoxide * Nitrogen Dioxide) + 0.07115X(Carbon Monoxide *PM2.5 Mass) − 7.715X(Methane * Nitrogen Dioxide) + 417.0X(Nitric Oxide * Ozone) − 508.1X(Nitrogen Dioxide * Ozone)

The submodel for Station.Name = Calgary Northeast is:

Air Quality Index = 1.067 + 0.5202X(Carbon Monoxide) − 0.1380X(Methane) + 0.07381X(Methane^2) + 7.671X(Nitric Oxide) + 44.78X(Nitrogen Dioxide) + 33.51X(Ozone) + 315.9X(Ozone^2) + 0.02846X(Non−methane Hydrocarbons) + 0.01117X(PM2.5 Mass) − 27.94X(Carbon Monoxide * Nitric Oxide) + 31.91X(Carbon Monoxide * Nitrogen Dioxide) + 0.07115X(Carbon Monoxide *PM2.5 Mass) − 7.715X(Methane * Nitrogen Dioxide) + 417.0X(Nitric Oxide * Ozone) − 508.1X(Nitrogen Dioxide * Ozone)

4 CONCLUSION AND DISCUSSION

4.1 Approach

Overall, is the approach we took promising? Please elaborate. What different approach or variant of this approach is better?

This project set out to unravel the story of Calgary's air quality, aiming to pinpoint the key pollutants that most reliably predict air conditions. The ultimate goal is to provide individuals and policymakers with actionable insights to support informed decisions for a healthier environment and improved quality of life.

The final regression model we developed is designed to do exactly that. We will leverage this model to address these insights, beginning with using it to address our research questions:

1. **Which months have the best and worst air quality?**

Aim: To identify seasonal trends in air quality, which can inform recommendations for scheduling outdoor activities during times of lower pollution.

Our analysis of the air pollutants dataset revealed that seasonality did not appear significant in predicting air quality, as it was excluded from our best additive model. This result is unexpected, given the common understanding that air quality tends to worsen during specific months, such as winter, when atmospheric conditions trap pollutants closer to the ground. We find this discrepancy intriguing and believe further investigation is needed to understand why seasonality was not significant. It may be the case that air quality is more closely linked to specific weather statistics, which are better at capturing seasonal effects. This question remains open for future exploration by our group.

2. Which communities have the best and worst levels of air quality?

Aim: To highlight areas that may require additional air quality management and help individuals with respiratory conditions make informed decisions about where to live or engage in outdoor activities.

It may be challenging to determine which communities have the best and worst levels of air quality in general, as the levels

of pollutants vary across different regions. However, specific pollutants can be identified, and their levels can be compared. This allows government officials to target specific pollutants for reduction in their air quality management efforts. For example, the Northeast region has the highest Carbon Monoxide level (0.5202), compared to the Southeast (-0.0674) and Northwest (-0.6793). The significant difference in these levels is an important topic worth exploring further.

In general, it was noted that Calgary Northwest shows lower coefficients for many pollutant interactions compared to other locations. For example, there was a lower influence in Ozone where the Northwest is sitting at 31.23 units compared to 33.51 untis in the Northeast. This can suggest that Calgary Northwest can experience better air quality as pollutants generally have a smaller impact on the AQI, however further investigation into this topic should be done.

It was also noted that Calgary Northeast in general had the highest coefficients for pollutants compared to the other locations. An example is Nitrogen Dioxide is sitting at 44.78 units which is significantly higher then the other locations. This can indicate that Calgary Northeast has the worst air quality in the city, due to the pollutants contributing more heavily to increasing AQI, but this should be investigated further.

3. What factors contribute most to poor air quality in Calgary or specific communities?

Aim: To provide policymakers with insights on which pollutants or activities have the most impact on air quality, enabling targeted interventions to reduce pollution

The factors that contribute most to air quality are the pollutants included in the model, as they are significant in affecting air quality. These pollutants include:

- **Carbon Monoxide**: Emitted primarily by vehicle exhaust, industrial processes, and heating systems.
- **Methane**: Released during the production and transport of coal, oil, and natural gas, as well as from agricultural sources, especially livestock.
- **Nitric Oxide**: A byproduct of combustion processes, particularly from vehicles and power plants.
- **Nitrogen Dioxide**: Produced from the burning of fossil fuels, especially in vehicles, industrial processes, and power generation.
- **Ozone**: Formed when sunlight reacts with pollutants like nitrogen oxides and volatile organic compounds.
- **Non-methane Hydrocarbons**: Emitted from vehicles, industrial processes, and the evaporation of fuels and solvents.
- **PM2.5 Mass**: Fine particulate matter that comes from sources such as vehicle exhaust, industrial emissions, and wildfires.

Given the complexities of the model, such as, interactions among the pollutants, it is challenging to determine which contribute most to air quality. Additionally, it seems that regional factors play a significant role in determining air quality. It would be valuable to explore why these regions differ in their contribution to air quality. One area the group could investigate is the type of activities that take place in these regions, which may influence pollutant levels and their impact on air quality.

Two interaction terms also showed high predictive values. Carbon Monoxide and Nitrogen Dioxide together had a coefficient of 31.91units indicating that the combination of Carbon Monoxide and Nitrogen Dioxide contribute to a higher AQI or worse air quality. Another interaction term of Nitric Oxide and Ozone also had a high positive coefficient of 417.0 units indicating that higher levels of both Nitric Oxide and Ozone will significantly increase the AQI and contribute to worse air quality.

4. Can the model predict future Air Quality?

Aim: To explore if historical AQI and pollutant data can help forecast future air quality levels, potentially leading to a useful tool for anticipating high-risk pollution days.

Achieving a reliable model was challenging due to issues with normality and heteroscedasticity. To address this, we employed a robust regression model, which provides reliable estimates by minimizing the influence of outliers and influential data points during the estimation process.

While we can use the robust model for forecasting, there are a few important considerations. Robust regression models are designed to be less sensitive to outliers and violations of assumptions such as normality and homoscedasticity, making them particularly useful when these issues are present in the data. This makes the robust model a strong choice for forecasting, especially if outliers are a concern or if similar irregularities are expected in future data.

The estimated Adjusted R-squared of 94.84% indicates that the model explains a substantial portion of the variation in the Air Quality Index, which suggests that this model is a valuable tool for anticipating high-risk pollution days.

4.2 Future Work

Further exploration of our model can include investigating even higher order terms. In our model, we only went to the second degree for the higher order terms, but we could explore if raising variables to the 3rd, 4th or 5th power would increase our model. We also could explore using different regression models such as non-linear modeling to see if that would give us different results. Further exploration into seasonality of the data including why Seasons was excluded and if there is another data point we could use to explore seasonality can be a future research question for the team.

Follow up work for this project can include expanding the dataset to include data from more specific neighbourhoods. Having data from specific neighbourhoods instead of generalizing quadrants of the city may highlight and help us answer our research question two better. Breaking the location down into smaller sections such as neighbourhoods will also allow us to answer more specific questions surrounding urban vs industrial areas. We could also expand the dataset to include additional predictors such as sulfur dioxide to give a more comprehensive picture of air quality. Finally we could expand the timeframe of this dataset to gather more datapoints across various timeframes to see if there is a significant impact of the model.

In summary, our best model used additive, interactive and higher order terms. The best model was unable to meet the heteroscedasticity and normality assumptions despite our best efforts, however using a robust regression model we were able to create a model that didn't need to hit all of these assumptions. We were able to use our model to help answer our four research questions identifying that in general Carbon Monoxide, Methane, Nitric Oxide, Nitrogen Dioxide and Ozone if increased in the air, will worsen air quality. Finally we have identified areas where we could continue to expand our research in this area to better support Calgarians knowledge on air quality across the city.

5 REFERENCES

1. City of Calgary. (n.d.-a). *Historical air quality by parameter*. City of Calgary Open Data Portal. https://data.calgary.ca/Environment/Historical-air-quality-by-parameter/7g8h-ukcq
2. City of Calgary. (n.d.-b). *Calgary air quality story*. City of Calgary Open Data Portal. https://data.calgary.ca/stories/s/u45n-7awa
3. Environment and Climate Change Canada. (n.d.). *Understanding the Air Quality Health Index messages*. Government of Canada. https://www.canada.ca/en/environment-climate-change/services/air-quality-health-

index/understanding-messages.html
4. NASA. (n.d.). *What is air quality?* NASA. https://www.nasa.gov/general/what-is-air-quality/
5. Tennessee Department of Health. (n.d.). *Air quality and health*. Tennessee Department of Health. https://www.tn.gov/health/cedep/environmental/healthy-places/healthy-places/environmental-quality/eq/air.html
6. UCLA. (n.d.). *Robust Regression | R Data Analysis Examples.* https://stats.oarc.ucla.edu/r/dae/robust-regression/
7. World Health Organization. (n.d.). *Ambient (outdoor) air quality and health*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

APPENDIX

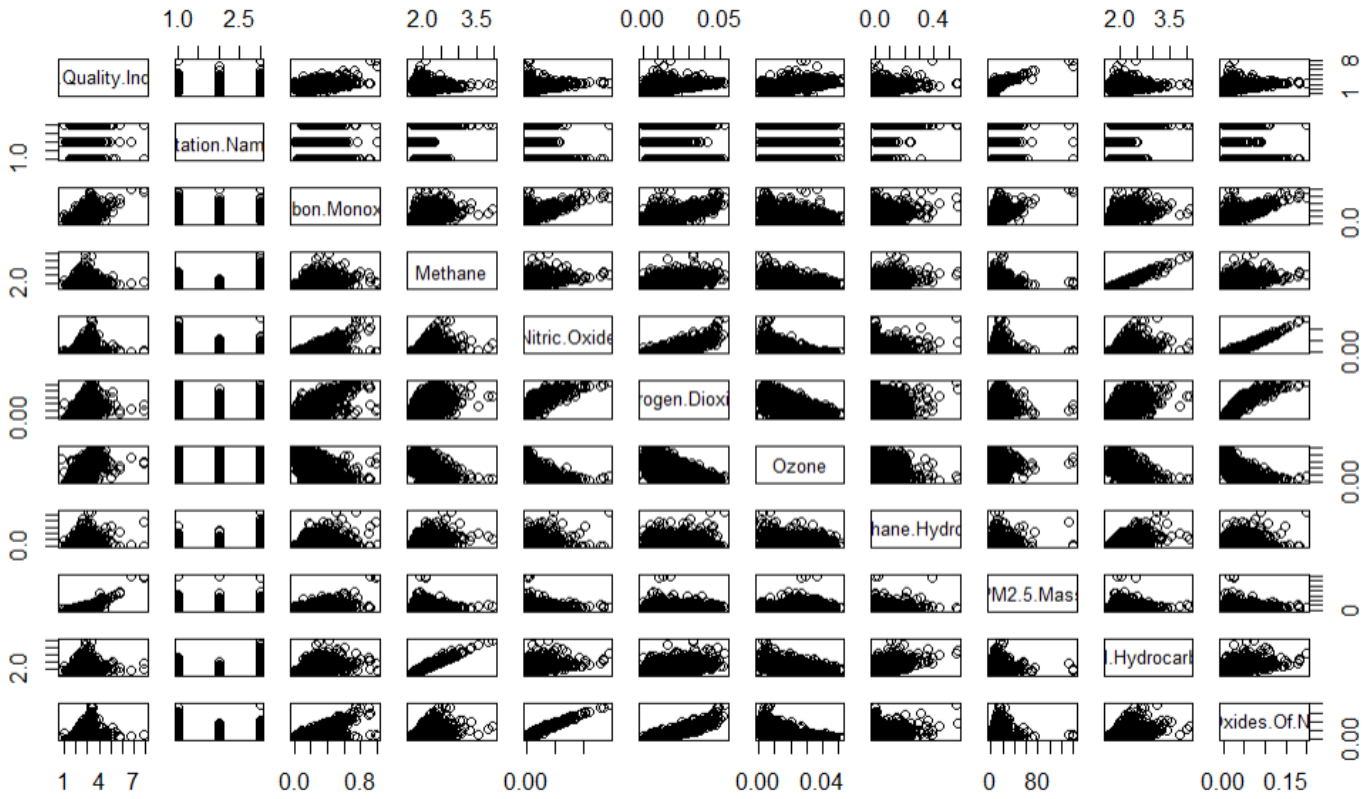
Appendix A - All-Possible-Regressions Selection Procedure Summary Table

```
        rsquare          cp       BIC       RSS AdjustedR
 [1,] 0.4599114 12823.52743 -2096.675 457.13025 0.4597539
 [2,] 0.7152136  5144.30464 -4283.714 241.04278 0.7150474
 [3,] 0.8711398   454.98890 -6995.602 109.06711 0.8710270
 [4,] 0.8812273   153.48841 -7267.063 100.52906 0.8810886
 [5,] 0.8832921    93.36438 -7319.077  98.78139 0.8831217
 [6,] 0.8840034    73.96549 -7331.903  98.17940 0.8838000
 [7,] 0.8848376    50.86528 -7348.521  97.47328 0.8846020
 [8,] 0.8856154    29.46498 -7363.624  96.81499 0.8853479
 [9,] 0.8857984    25.95767 -7360.977  96.66005 0.8854979
[10,] 0.8860363    20.80228 -7359.987  96.45876 0.8857029
[11,] 0.8862889    15.20104 -7359.459  96.24492 0.8859229
[12,] 0.8864347    12.81434 -7355.720  96.12152 0.8860359
[13,] 0.8864948    13.00520 -7349.396  96.07062 0.8860629
[14,] 0.8864950    15.00000 -7341.261  96.07048 0.8860297
```

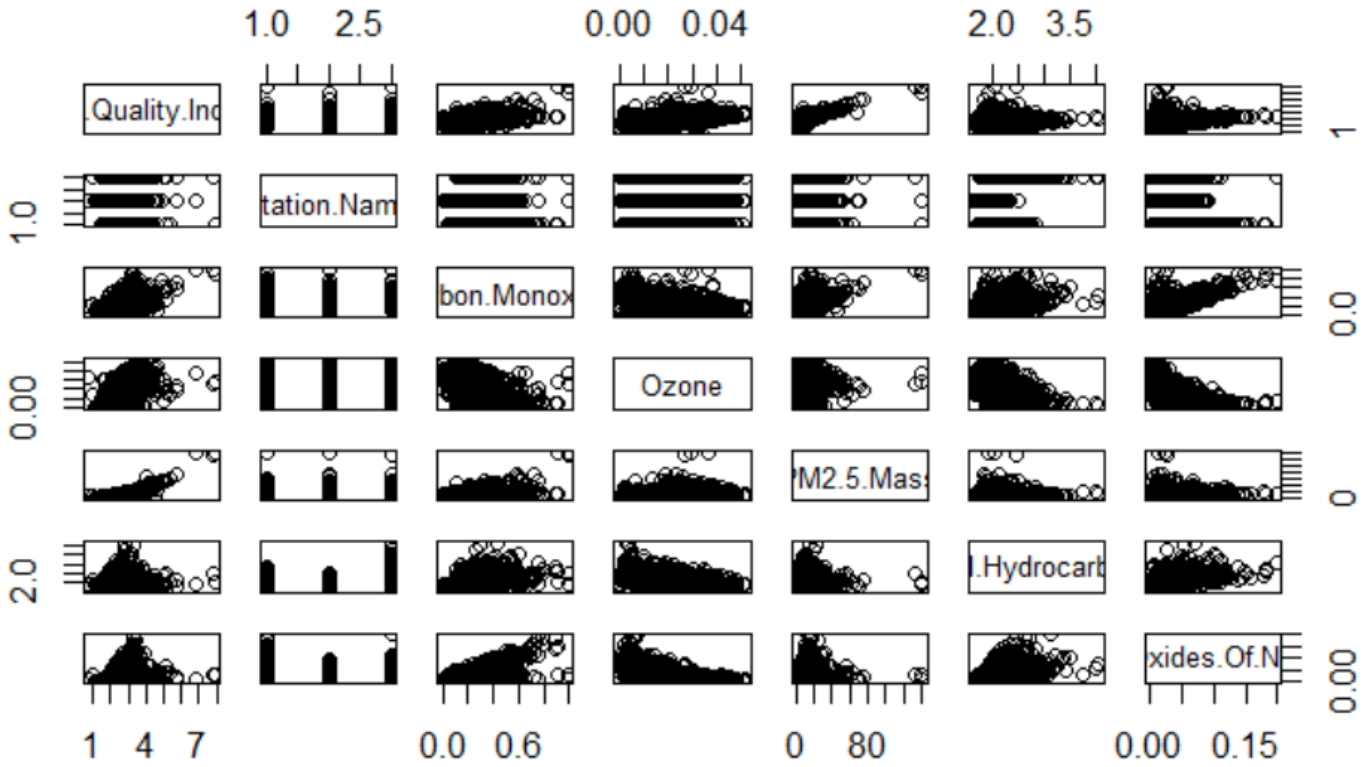
Appendix B - ANOVA Table for Additive Models

Model	Residual DF	RSS	DF	Sum of Squares	F-Statistic	P-value
Model 1	3418	96.245				
Model 2	3415	96.070	3	0.17445	2.067	0.1024

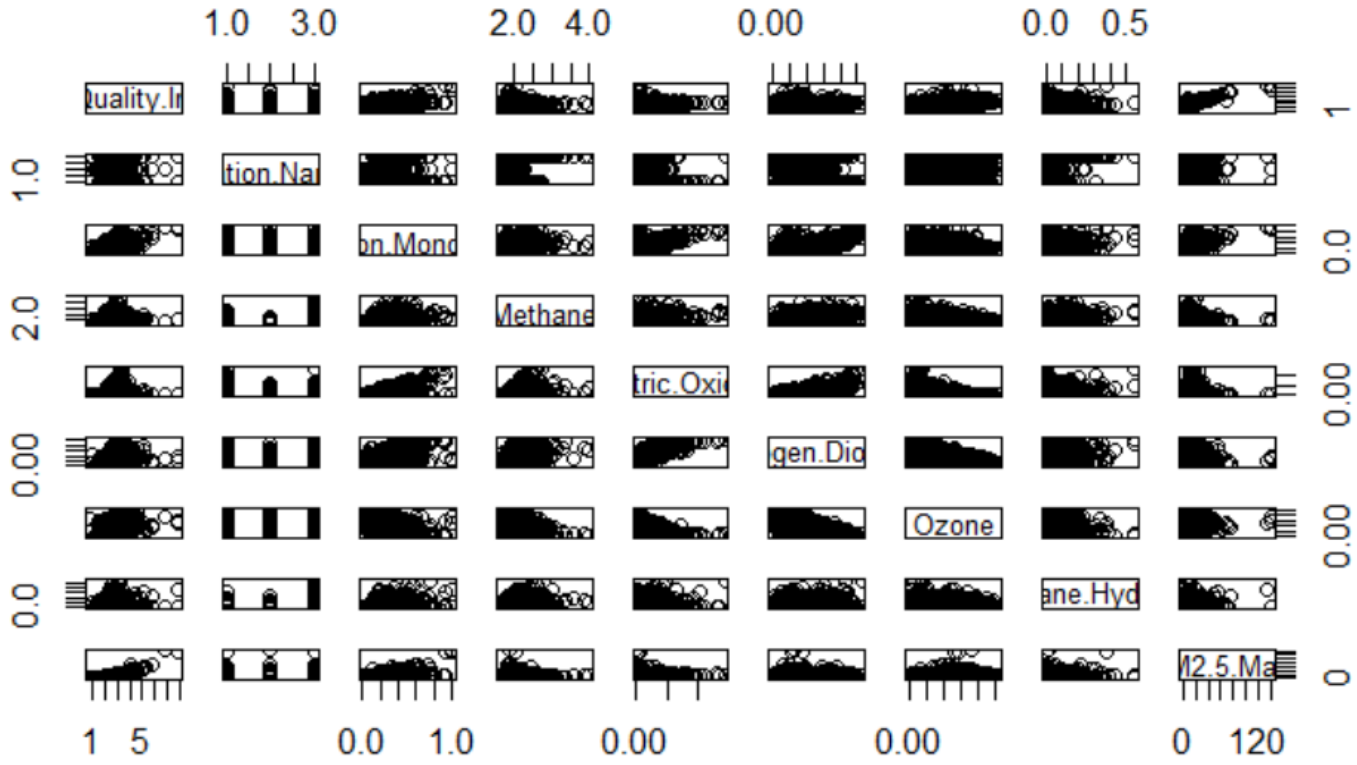
Appendix C - Multicollinearity Graph Output



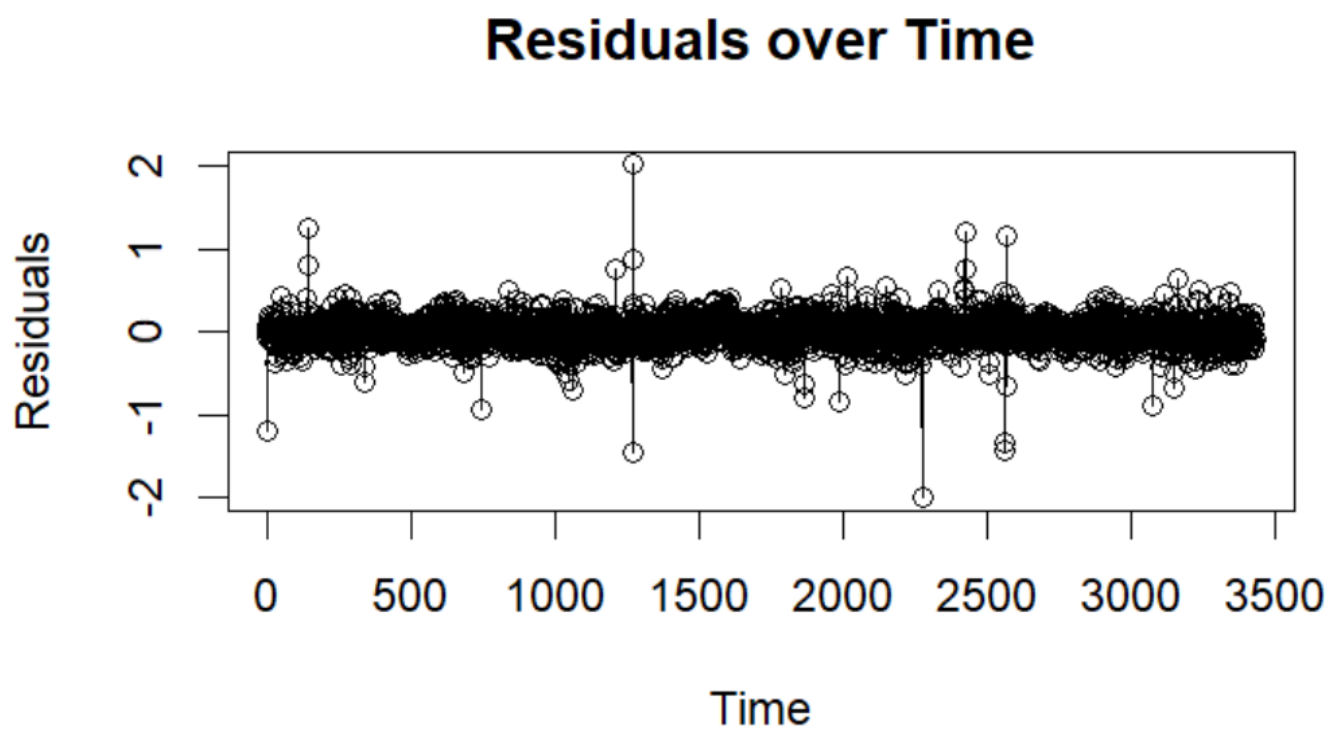
Appendix D - VIF Output - Reduced Model 1



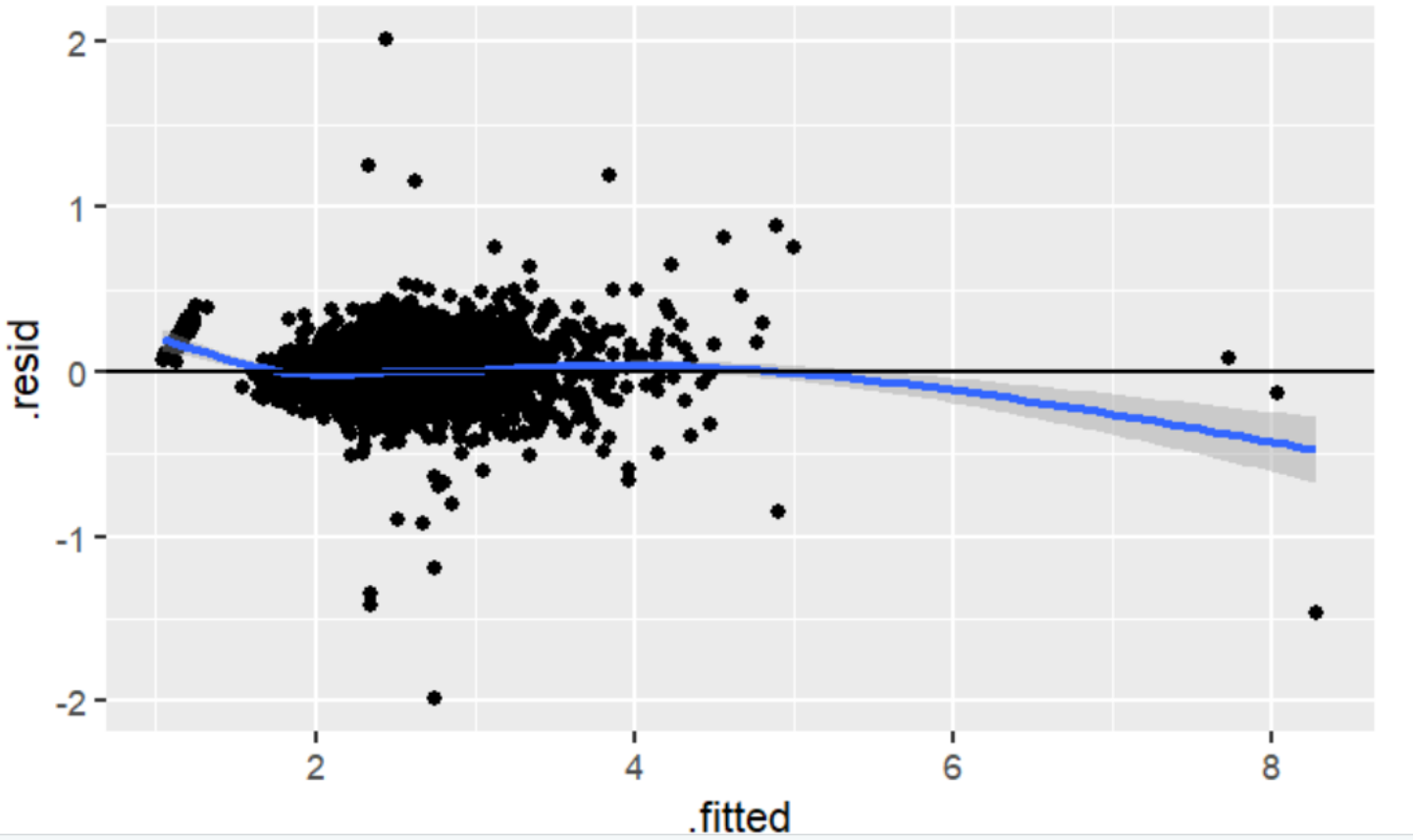
Appendix E - VIF Output - Reduced Model 2



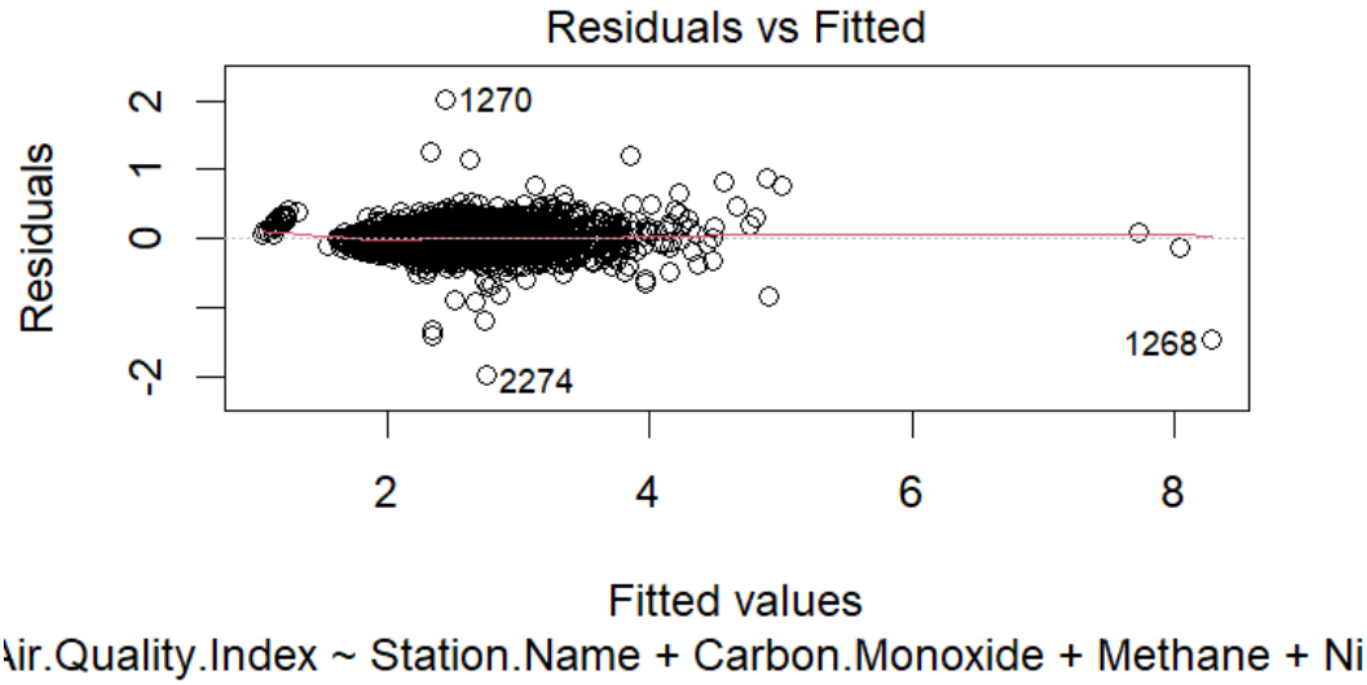
Appendix F - Linearity Test



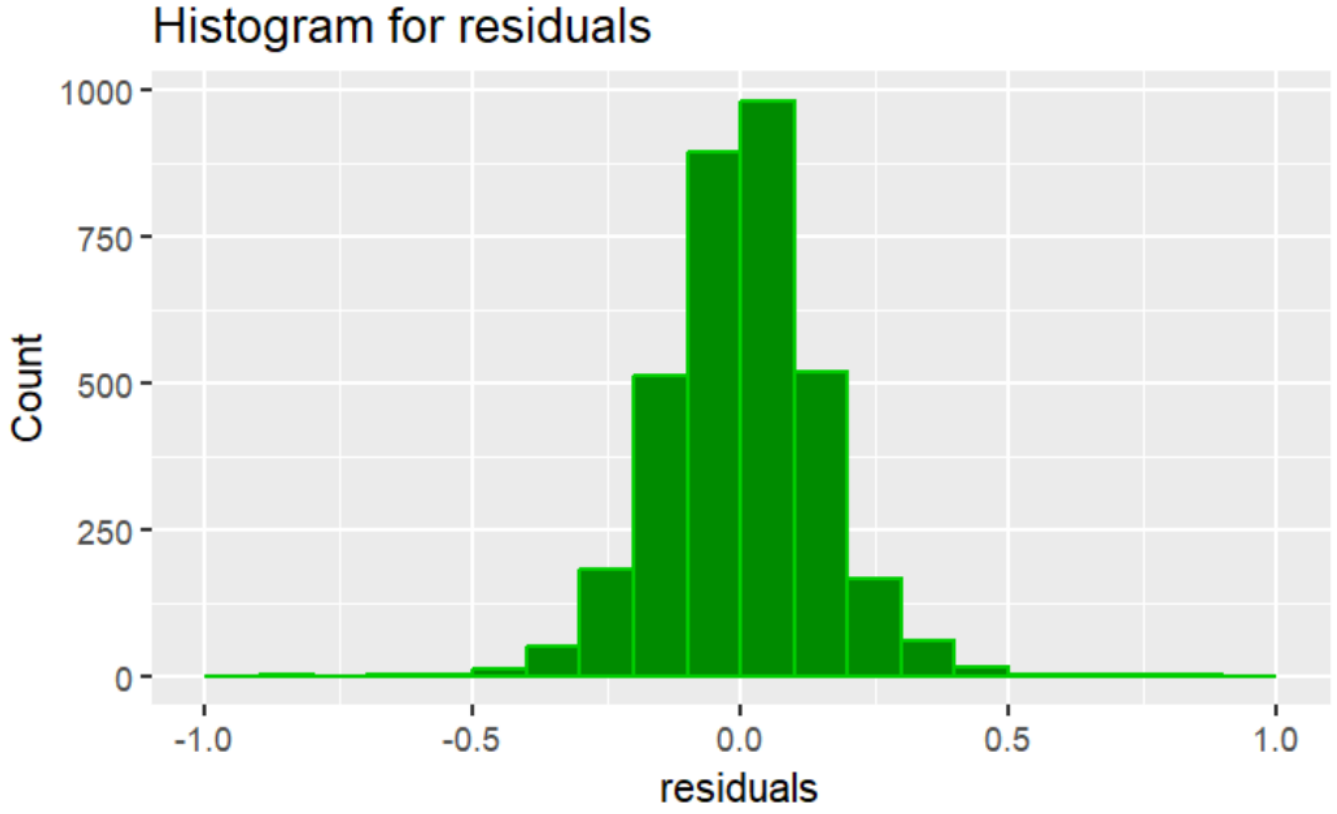
Appendix G - GGpairs plot for Linearity



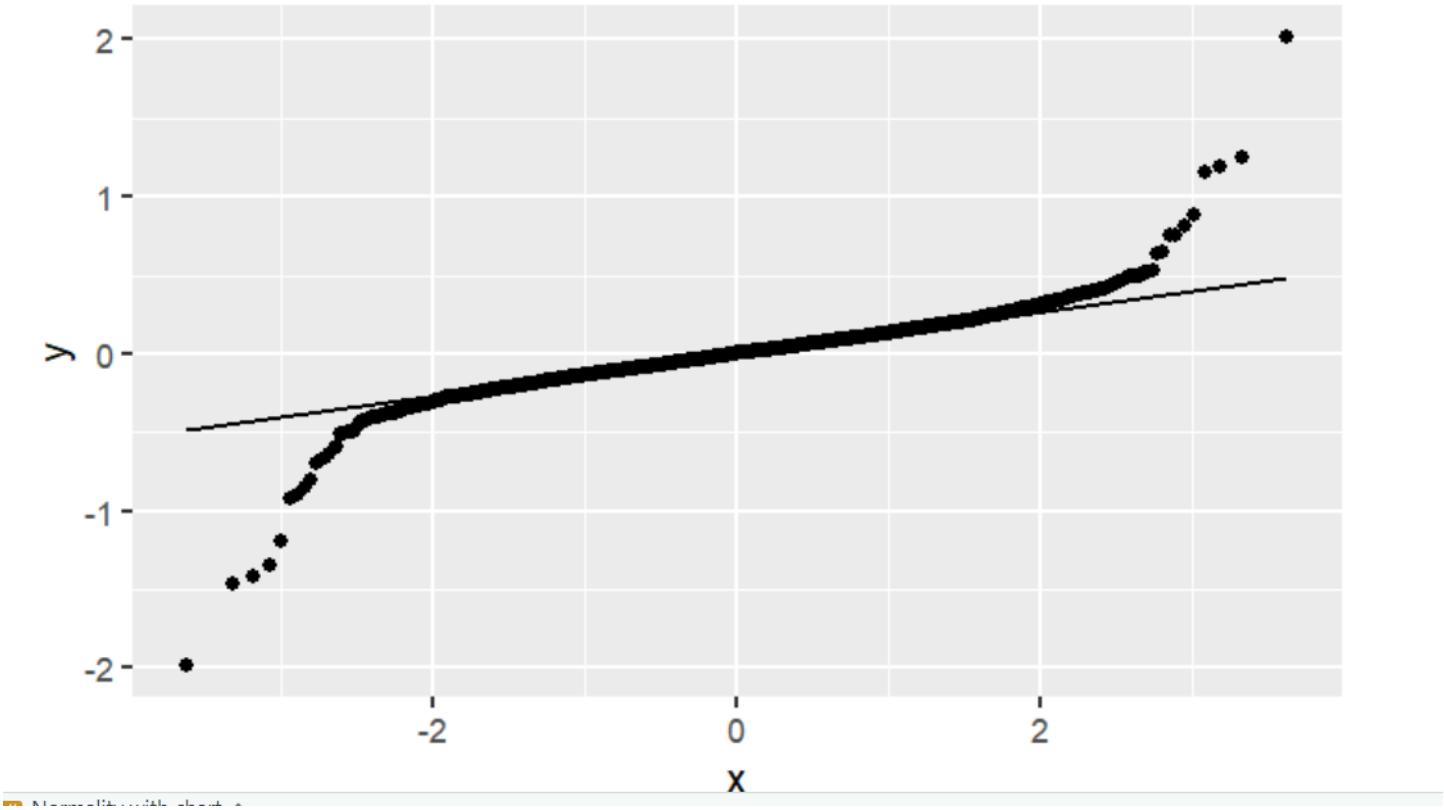
Appendix H - Heteroscedasticity Graph



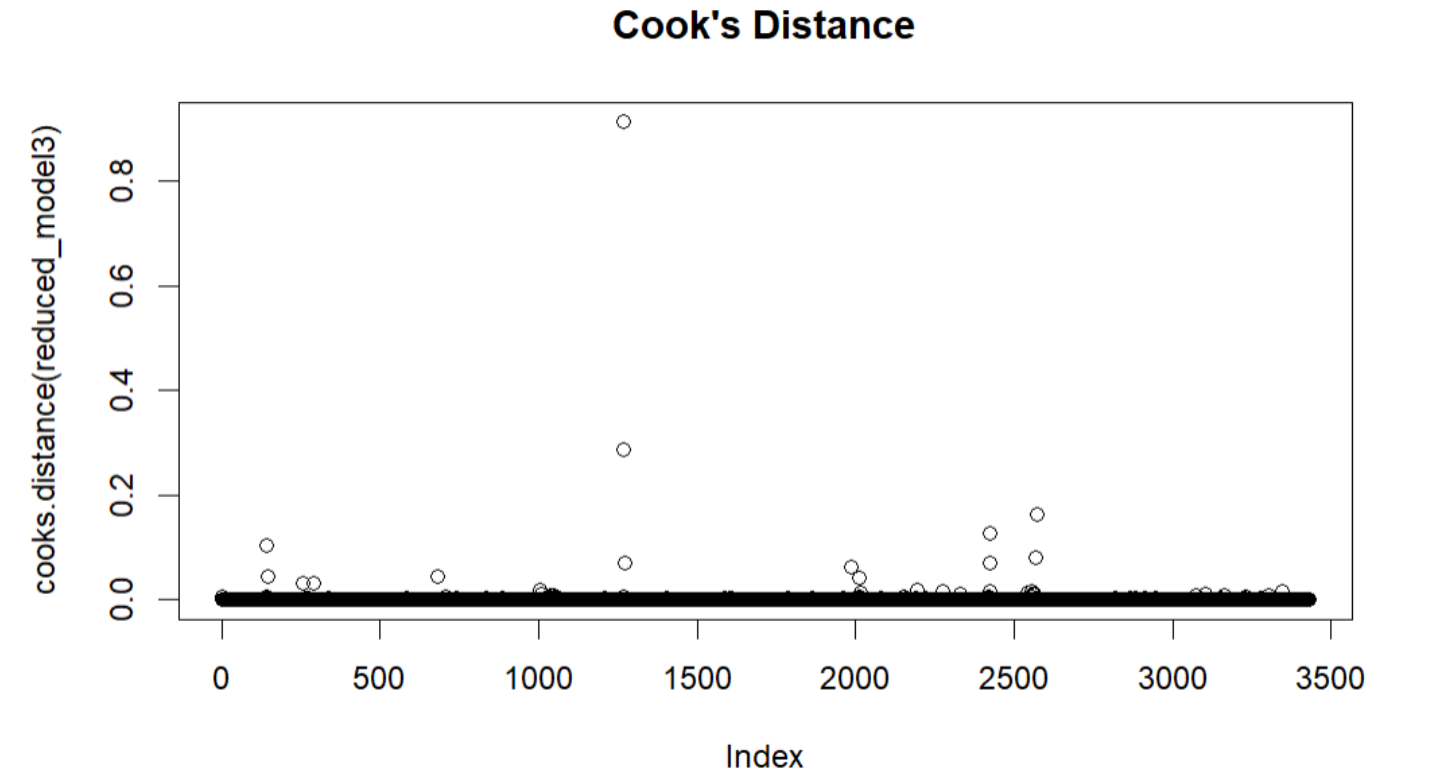
Appendix I - Normality Histogram



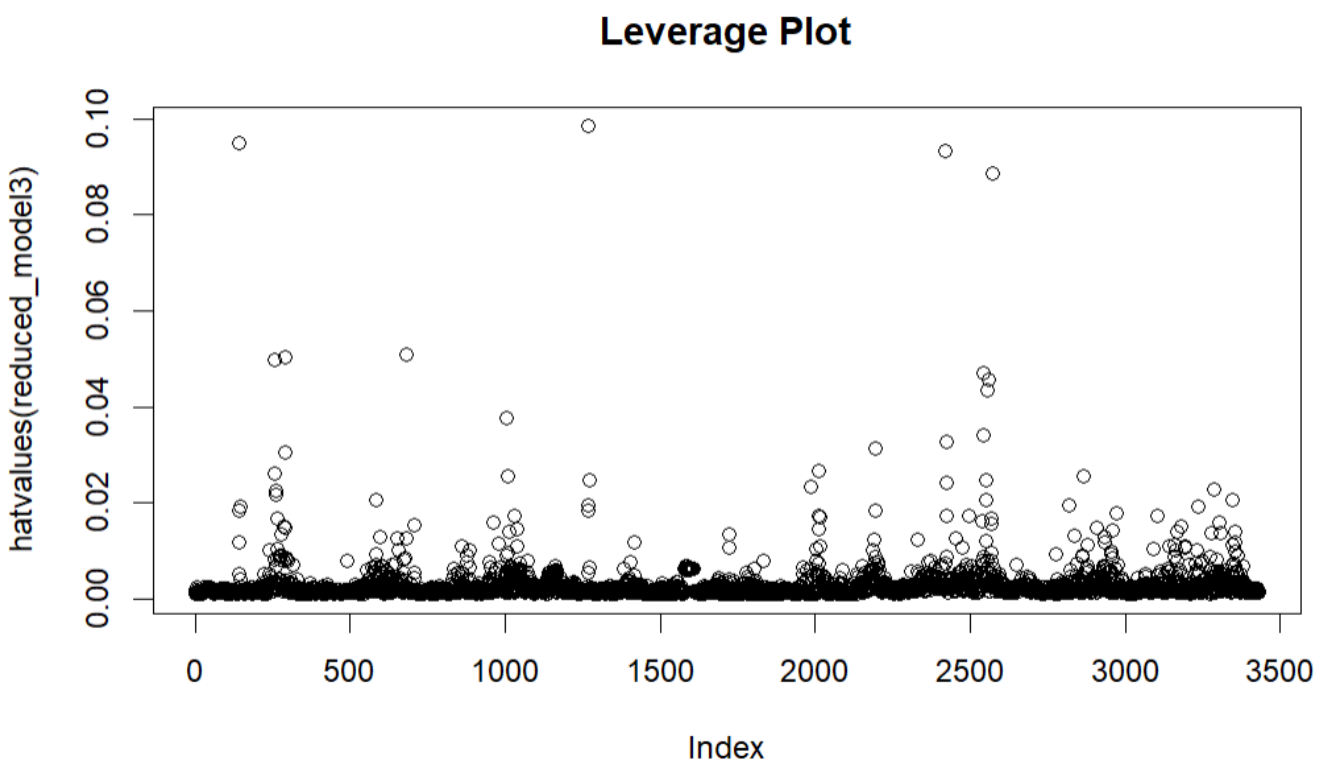
Appendix J - QQ Plot to test Normality



Appendix K - Cook's Distance Plot



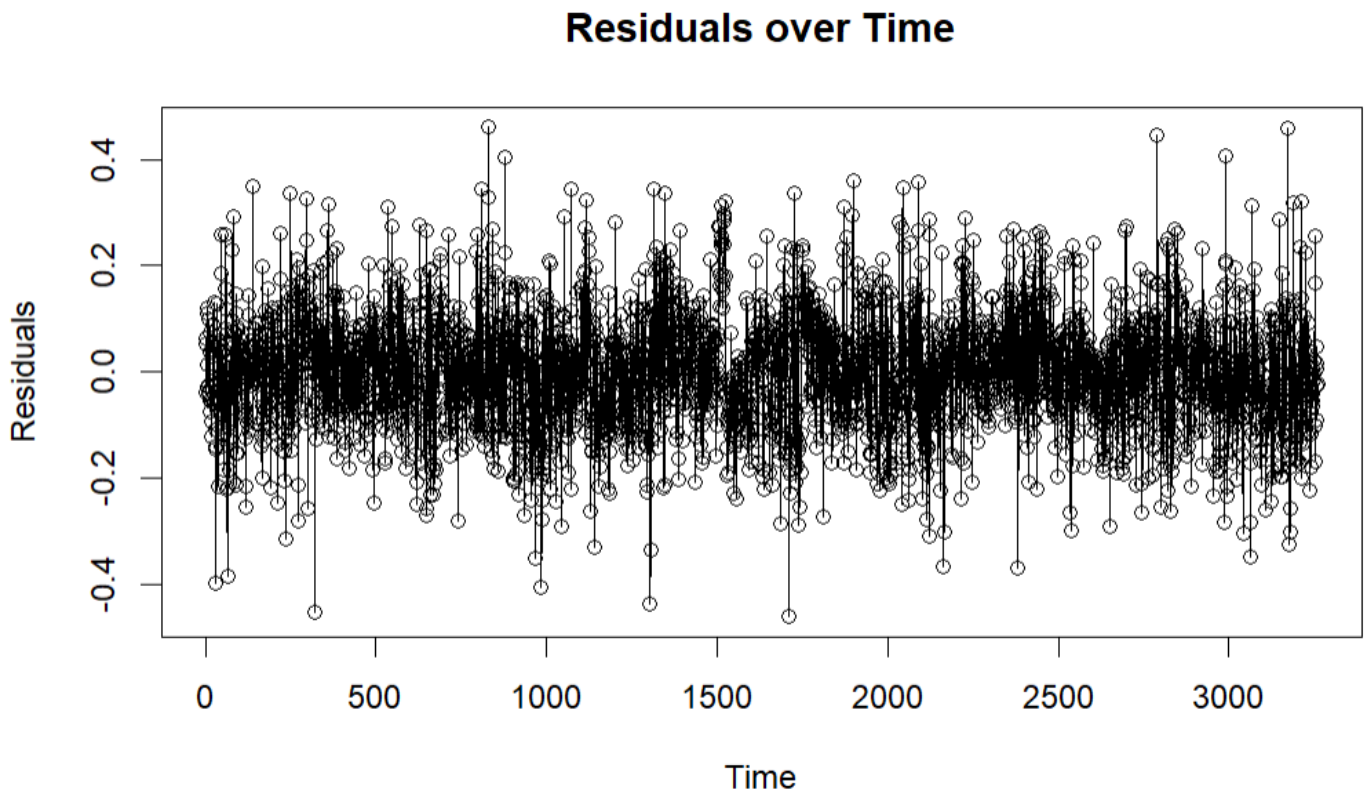
Appendix L - Leverage Plot



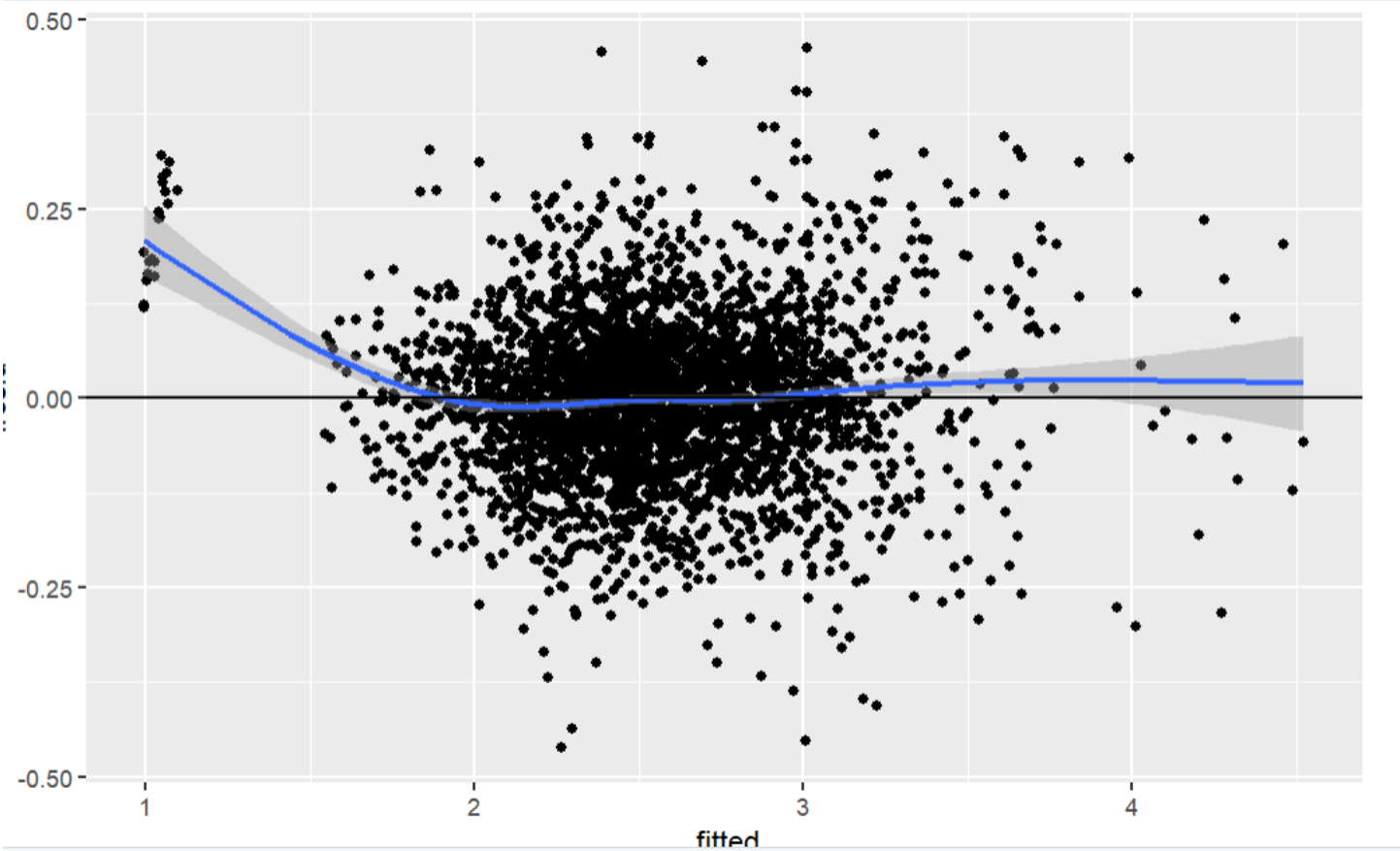
Appendix M - ANOVA Table for Interactive Model

Model	Residual DF	RSS	DF	Sum of Squares	F-Statistic	P-value
Model 1	3227	39.155				
Model 2	3250	48.243	-23	-9.0882	32.566	2.2e-16

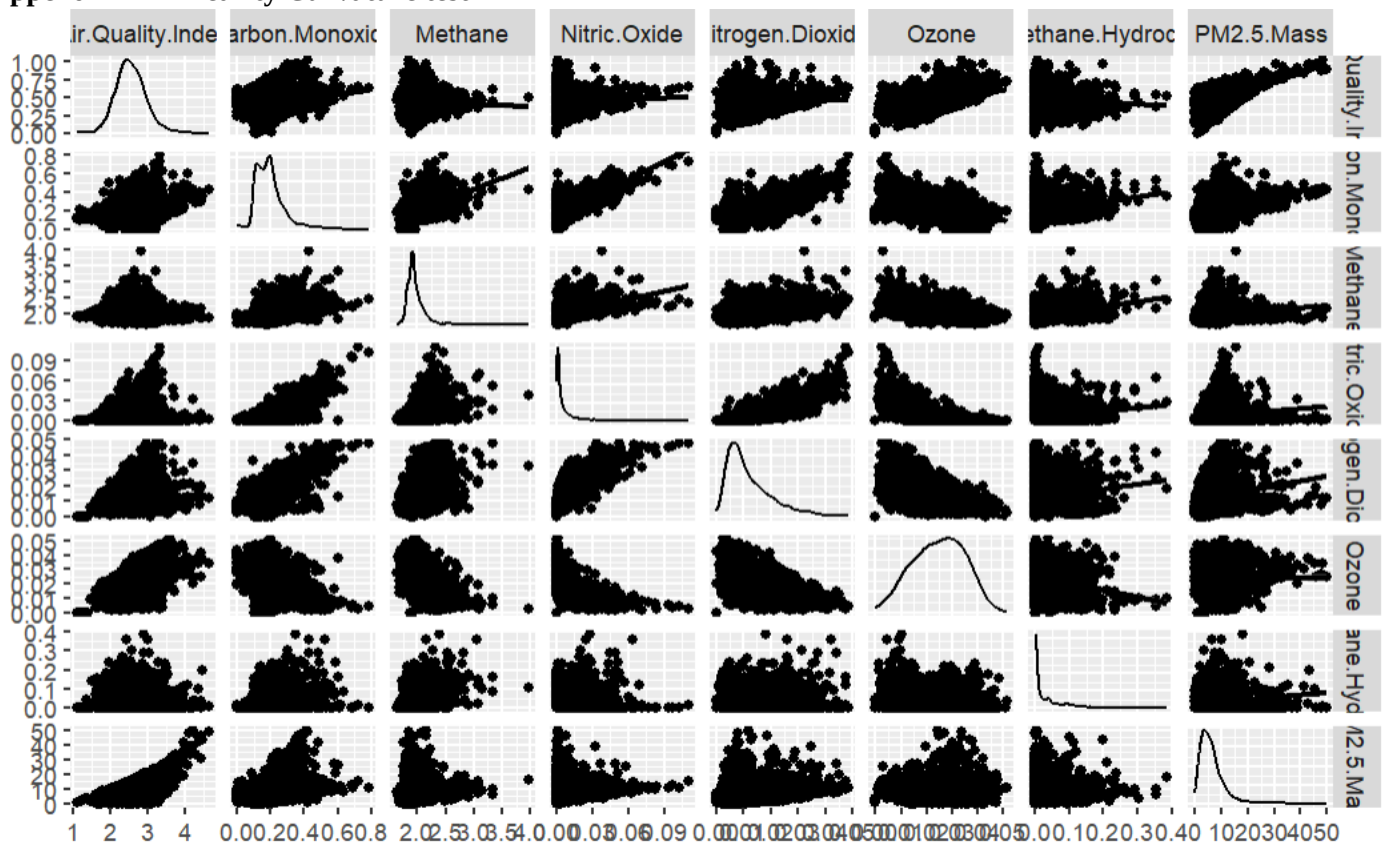
Appendix N - Independence Plot for Interaction Terms



Appendix O - GGPairs plot with interaction terms



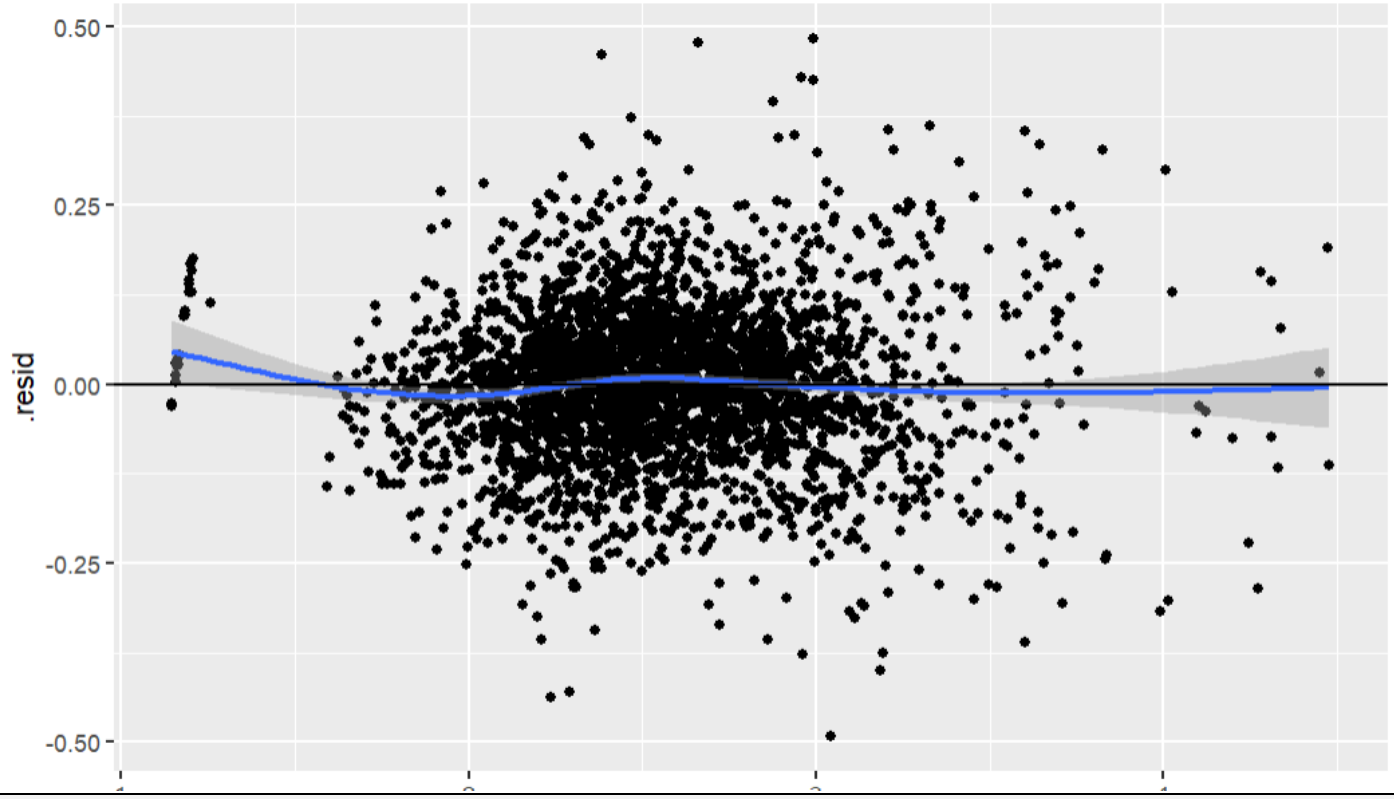
Appendix P - Linearity Curvature test



Appendix Q - ANOVA comparing Interaction to Higher Order

Model	Residual DF	RSS	DF	Sum of Squares	F-Statistic	P-value
Model 1	3230	37.515				
Model 2	3225	37.230	5	0.28478	4.9337	0.0001675

Appendix R - Linearity test on Higher Order Model



End of Project Report