UNIVERSITY OF CALGARY

Air Under the Microscope: Modelling Pollutant Effects on Air Quality

Team members:

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

DATA 603: Statistical Modelling With Data

Agenda

01/ Rationale/Aim

02/ Research Questions

03/ Dataset

04/ Approach & Main Results

05/ Conclusion

Rationale:

- Clean air is essential for **healthy living**, and air quality directly influences both health and **quality of life**;
- Air pollution significantly impacts health, causing an estimated **4.2 million premature deaths in 2019** due to respiratory diseases, cardiovascular conditions, and cancers;
- **Environmental damage** from pollutants affects ecosystems, depleting soil nutrients and harming biodiversity; and
- Air quality can vary significantly, even within hours, having a **reliable monitoring** and **forecasting** system is crucial.

Aim:

- Deliver comprehensive **insights into Calgary's air quality**; and
- Build a **model that can measure and forecast air quality**, with the ultimate goal of **saving** and **improving lives**.

Air Quality

Moderate Health Risk

Scale: Canada (AQHI)

Air quality index is 5, which is worse than yesterday at about this time.

Health Information

No need to modify your usual outdoor activities unless you experience symptoms such as coughing and throat irritation. At-risk populations should consider reducing or rescheduling strenuous activities outdoors if experiencing symptoms.

Primary Pollutant

Nitrogen Oxides (NO_2)

Nitrogen oxides typically come from traffic, fires and power stations.

Research Questions

1. **Which months have the best and worst air quality?**
2. **Which communities have the best and worst levels of air quality?**
3. **What factors contribute most to poor air quality in Calgary or specific communities?**
4. **Can the model the predict future Air Quality?**

Dataset

Dependent Variable

Independent Variables

Air Quality Index (AQI) = **Seasons** + **Regions** + **Nine Pollutants**

Air Quality Index (AQI)

- Quantitative
- Standardized scale from 1 to 10

Seasons

- Categorical
- Summer
- Winter
- Fall
- Spring

Regions

- Categorical
- North West
- South East
- North East

Nine Pollutants

- Quantitative
- Parts per billion, et cetera.
- PM2.5 Mass
- Carbon Monoxide
- Nitrogen Dioxide
- Nitric Oxide
- Ozone
- Methane
- Total Hydrocarbons
- Total Oxides of Nitrogen
- Non-methane Hydrocarbons

Source: City of Calgary

Additional Information:

- 2015–2018: Most consistent pollutant data (3 years).
- ~15% missing data. Imputed with median across regions and seasons.

Approach & Main Results

Approach (Model Building)

1. **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.
4. **Addressing Assumption Violations (Normality and Heteroscedasticity):** Transform model and remove outliers.
5. **Transformation and Model Updates:** For each transformation or addition of a new feature, re-check assumptions to ensure they still hold.

Model Selection

Full Model

Station Name + Season + Methane + Nitric Oxide + Nitrogen Dioxide + Ozone + Non-methane Hydrocarbons + PM2.5 Mass + Total Hydrocarbons + Total Oxides of Nitrogen

All-Possible-Regressions Selection Procedure *(Adjusted R-squared)*

Stepwise Selection Procedure *(0.05 - 0.1)*

All

Without Seasons

ANOVA: *(0.05)*

```
Pr(>F)
0.1024
```

Reduced Model:

Station Name + ~~Season~~ + Methane + Nitric Oxide + Nitrogen Dioxide + Ozone + Non-methane Hydrocarbons + PM2.5 Mass + Total Hydrocarbons + Total Oxides of Nitrogen

Multicollinearity

	VIF	detection
Station.NameCalgary Northwest	1.6912	0
Station.NameCalgary Southeast	2.4372	0
Carbon.Monoxide	4.2789	0
Methane	30886.5359	1
Nitric.Oxide	159.2768	1
Nitrogen.Dioxide	71.9613	1
Ozone	1.7358	0
Non.methane.Hydrocarbons	2470.4887	1
PM2.5.Mass	1.7077	0
Total.Hydrocarbons	40544.7424	1
Total.Oxides.Of.Nitrogen	395.8184	1

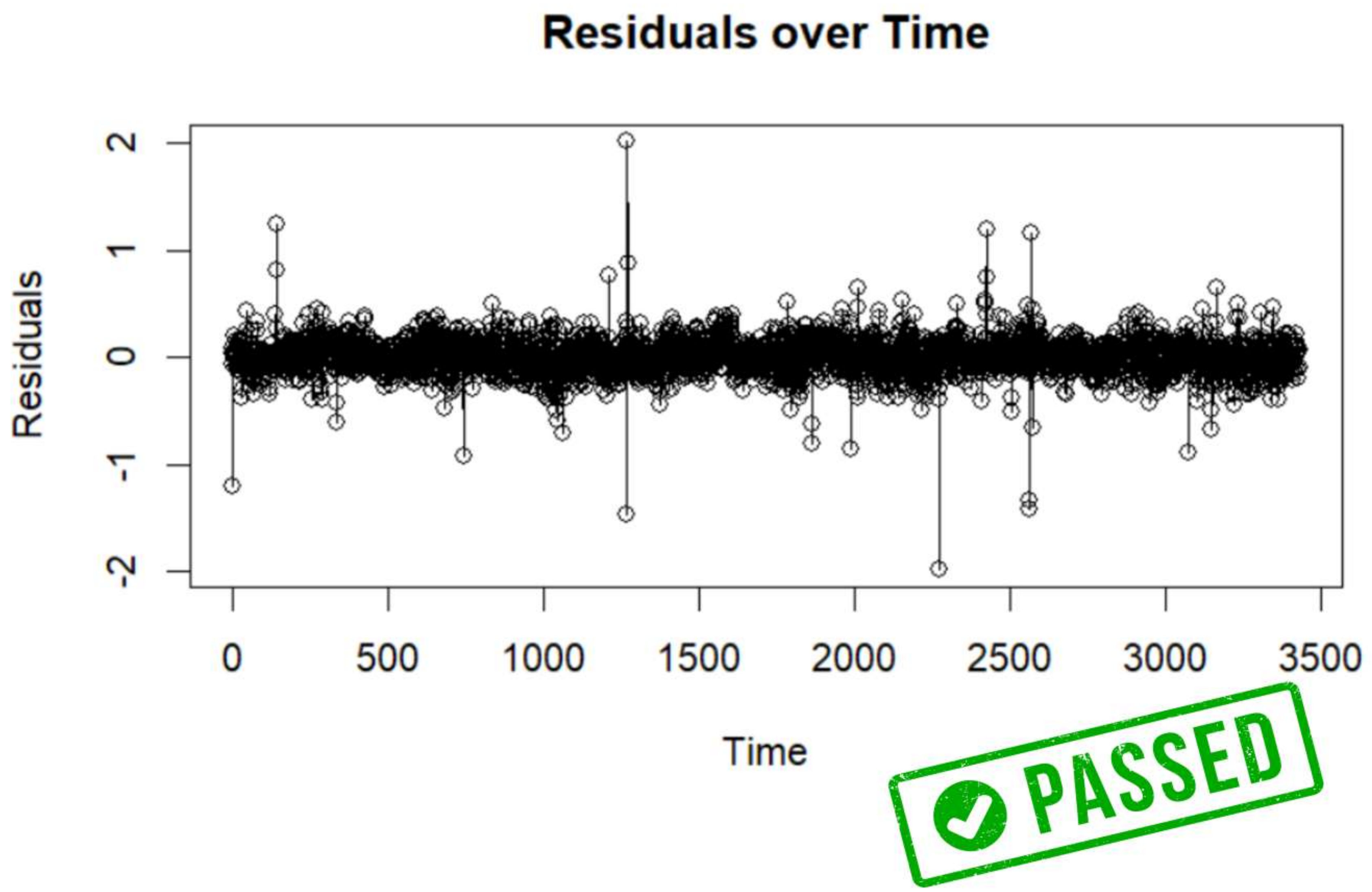
Total Oxides of Nitrogen includes **Nitric Oxide** and **Nitrogen Dioxide**

Total Hydrocarbon includes **Methane** and **Non-methane Hydrocarbons**

New Best Additive Model: **Station Name + ~~Season~~ + Methane + Nitric Oxide + Nitrogen Dioxide + Ozone + Non-methane Hydrocarbons + PM2.5 Mass + ~~Total Hydrocarbons~~ + ~~Total Oxides of Nitrogen~~**

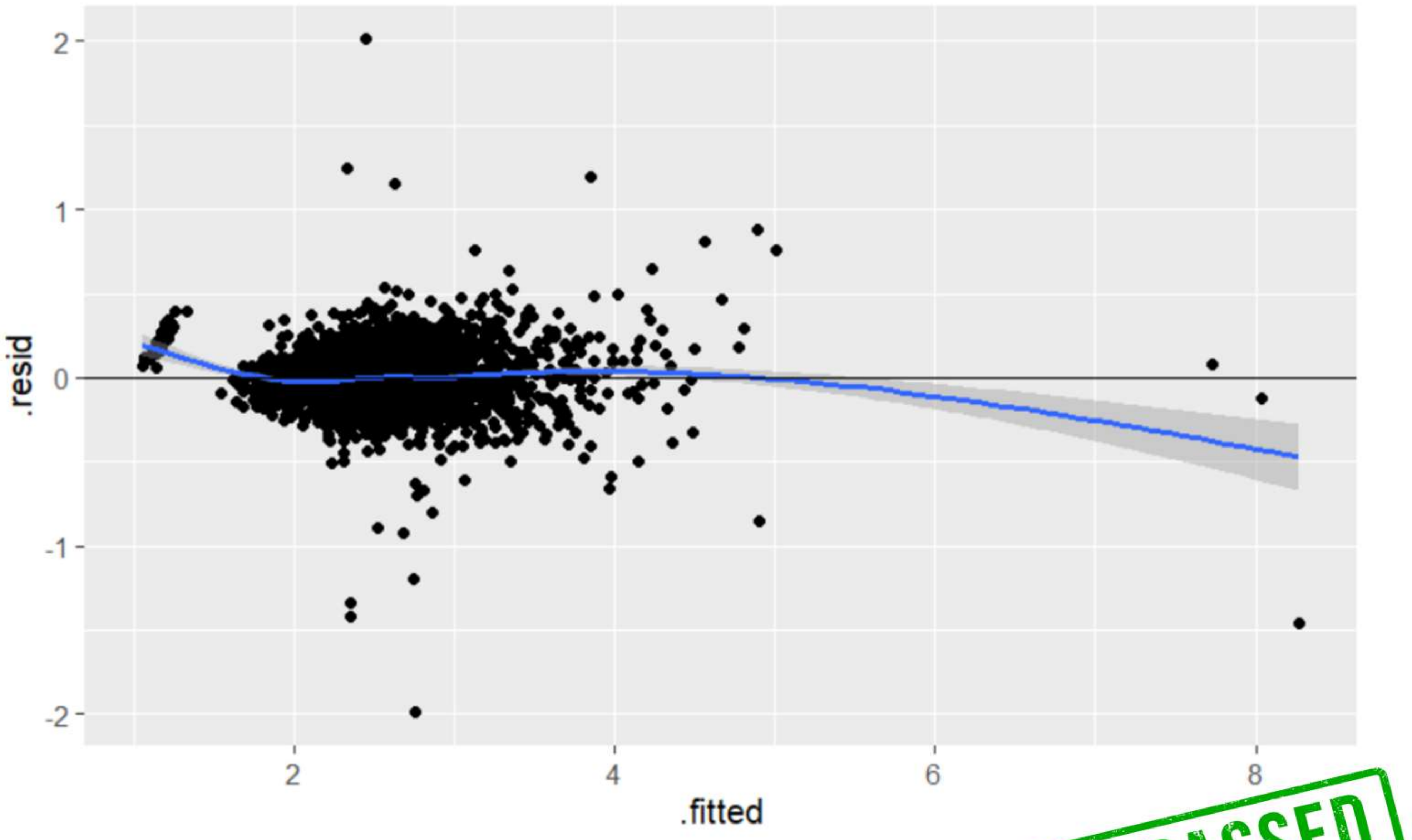
Checking Assumptions

Independence



Residuals are randomly scattered around zero with no pattern

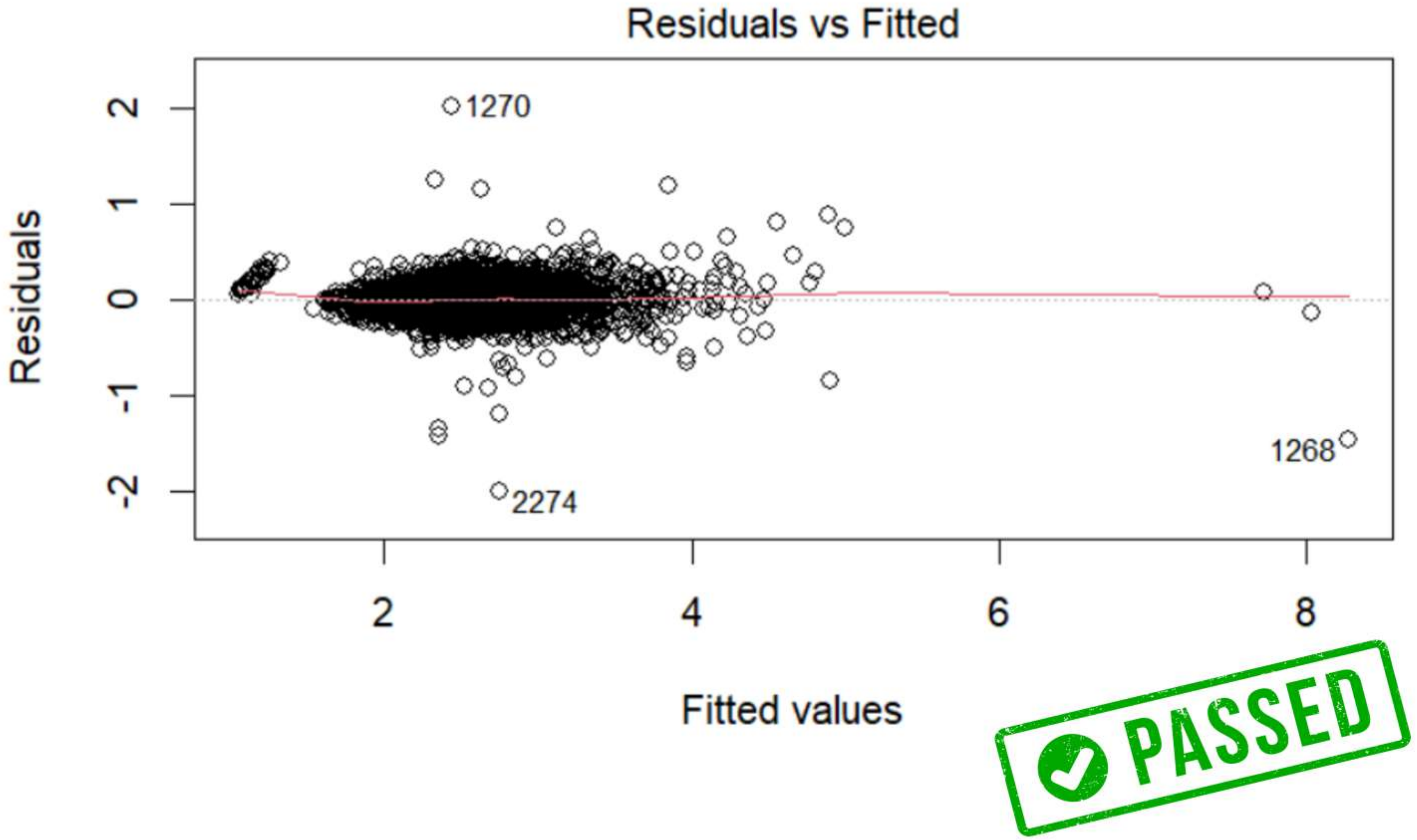
Linearity



No discernable pattern, slight curvature.

Checking Assumptions

Heteroscedasticity

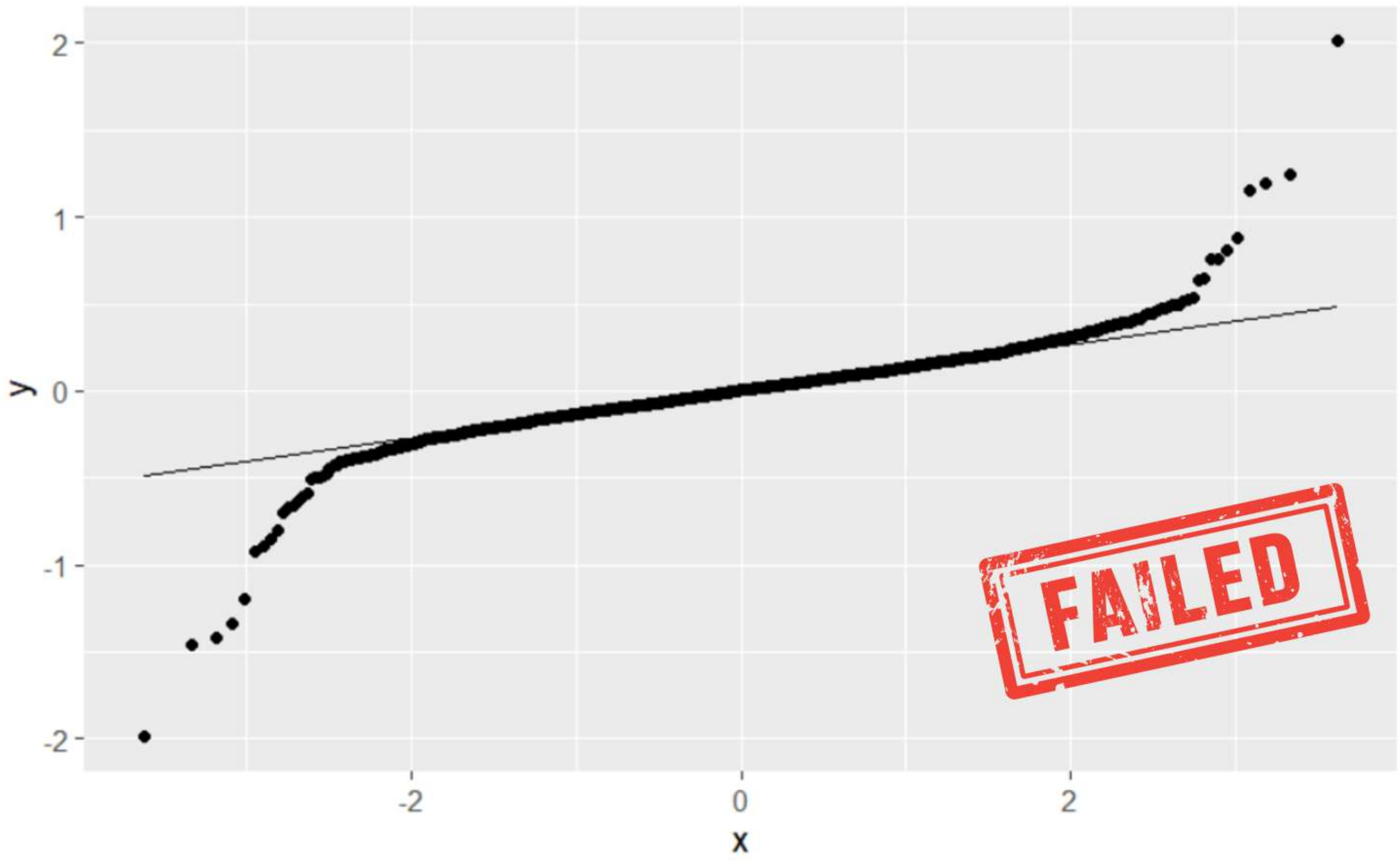


Quite horizontal, No funneling

Breush-Pagen Test:

`p-value < 2.2e-16`

Normality



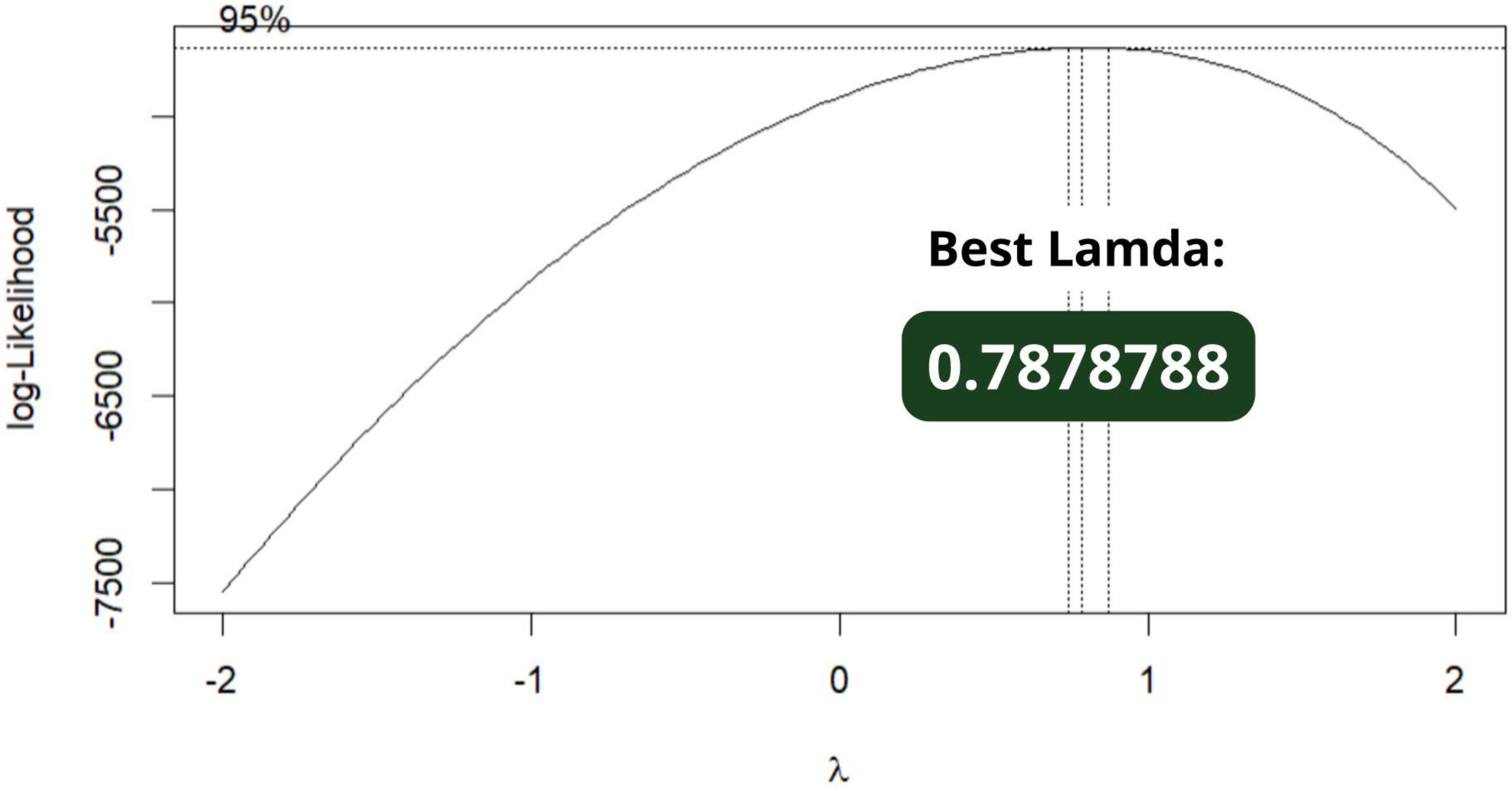
Does not follow the line completely.

Sharpio-Wilk Test:

`p-value < 2.2e-16`

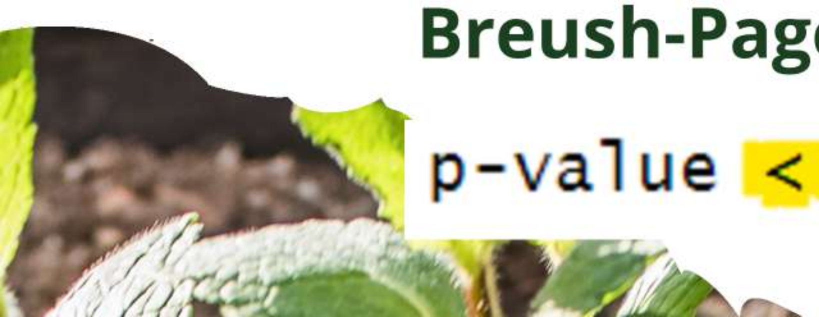
Addressing Assumption Violations:

BOXCOX TRANSFORMATION:



Best Lamda:

0.7878788



Breush-Pagen Test:

`p-value < 2.2e-16`

FAILED

Sharpio-Wilk Test:

`p-value < 2.2e-16`

FAILED

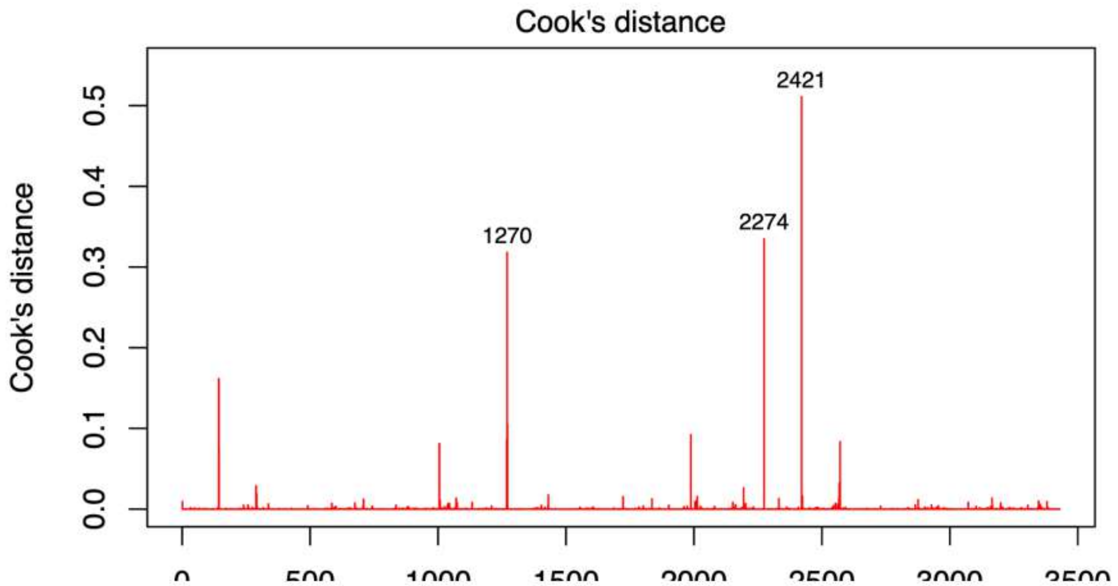
Addressing Assumption Violations and Model Updates

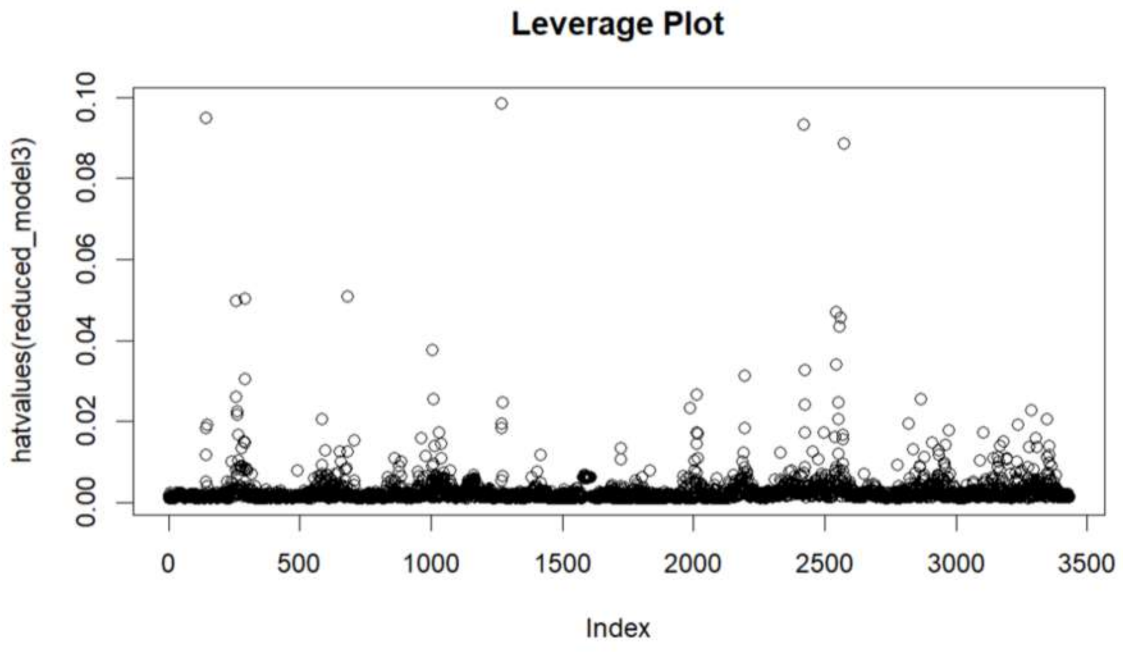
Influential Outliers:

Created a Cook's Distance Plot and Leverage Plot to identify the influential outliers and removed them, then we re-ran our assumption tests.

Cook's Distance Plot

Leverage Plot





Interaction Terms:

Made a model starting with all possible interaction terms. Removed insignificant interaction terms, re-ran our assumption tests

Higher Order Terms:

Made a model with higher order terms based on the gg-pairs plots and re-ran our assumption tests.

ALL FAILED NORMALITY AND HETEROSEDASTICITY

Solution: Robust Regression Model

What is it?

- Use when assumptions are **violated—outliers or heteroscedasticity or Normality.**
- Provide **reliable estimates** in the presence of assumption violations
- Flexible approach that remains **effective when traditional regression methods would fail**.

What it does?

- **Minimizes the influence of outliers and violations** of assumptions to **provide reliable model estimates** in less-than-ideal data conditions.
- It achieves this by using **weighting schemes that reduce the impact of extreme values** during the model fitting process.

Robust Sub-Models (Final)

Submodel for Station.NameCalgary Northwest:

$$
\begin{gathered}
\hat{Y}_{Air.Quality.Index} = 1.1939 - 0.6793X_{Carbon.Monoxide} - 0.1380X_{Methane} + 0.07381X_{I(Methane^2)} \\
+7.583X_{Nitric.Oxide} + 34.16X_{Nitrogen.Dioxide} + 31.123X_{Ozone} + 315.9X_{I(Ozone^2)}+ \\
1.5185X_{Non.methane.Hydrocarbons} + 0.0069X_{PM2.5.Mass} - 27.94X_{Carbon.Monoxide}X_{Nitric.Oxide}+ \\
31.91X_{Carbon.Monoxide}X_{Nitrogen.Dioxide}+0.07115X_{Carbon.Monoxide}X_{PM2.5.Mass}-7.715X_{Methane}X_{Nitrogen.Dioxide}+ \\
417.0X_{Nitric.Oxide}X_{Ozone} - 508.1X_{Nitrogen.Dioxide}X_{Ozone}
\end{gathered}
$$

Submodel for Station.NameCalgary Northeast:

$$
\begin{gathered}
\hat{Y}_{Air.Quality.Index} = 1.067 + 0.5202X_{Carbon.Monoxide} - 0.1380X_{Methane} + 0.07381X_{I(Methane^2)} \\
+7.671X_{Nitric.Oxide} + 44.78X_{Nitrogen.Dioxide} + 33.51X_{Ozone} + 315.9X_{I(Ozone^2)} \\
+0.02846X_{Non.methane.Hydrocarbons} + 0.01117X_{PM2.5.Mass} - 27.94X_{Carbon.Monoxide}X_{Nitric.Oxide}+ \\
31.91X_{Carbon.Monoxide}X_{Nitrogen.Dioxide}+0.07115X_{Carbon.Monoxide}X_{PM2.5.Mass}-7.715X_{Methane}X_{Nitrogen.Dioxide}+ \\
417.0X_{Nitric.Oxide}X_{Ozone} - 508.1X_{Nitrogen.Dioxide}X_{Ozone}
\end{gathered}
$$

Submodel for Station.NameCalgary Southeast

$$
\begin{gathered}
\hat{Y}_{Air.Quality.Index} = 1.1250 - 0.0674X_{Carbon.Monoxide} - 0.1380X_{Methane}+ \\
0.07381X_{I(Methane^2)} + 11.093X_{Nitric.Oxide} + 33.73X_{Nitrogen.Dioxide} + 30.206X_{Ozone}+ \\
315.9X_{I(Ozone^2)} + 0.21416X_{Non.methane.Hydrocarbons} + 0.01105X_{PM2.5.Mass}- \\
27.94X_{Carbon.Monoxide}X_{Nitric.Oxide}+31.91X_{Carbon.Monoxide}X_{Nitrogen.Dioxide}+0.07115X_{Carbon.Monoxide}X_{PM2.5.Mass}- \\
7.715X_{Methane}X_{Nitrogen.Dioxide} + 417.0X_{Nitric.Oxide}X_{Ozone} - 508.1X_{Nitrogen.Dioxide}X_{Ozone}
\end{gathered}
$$

Key Findings/Conclusion

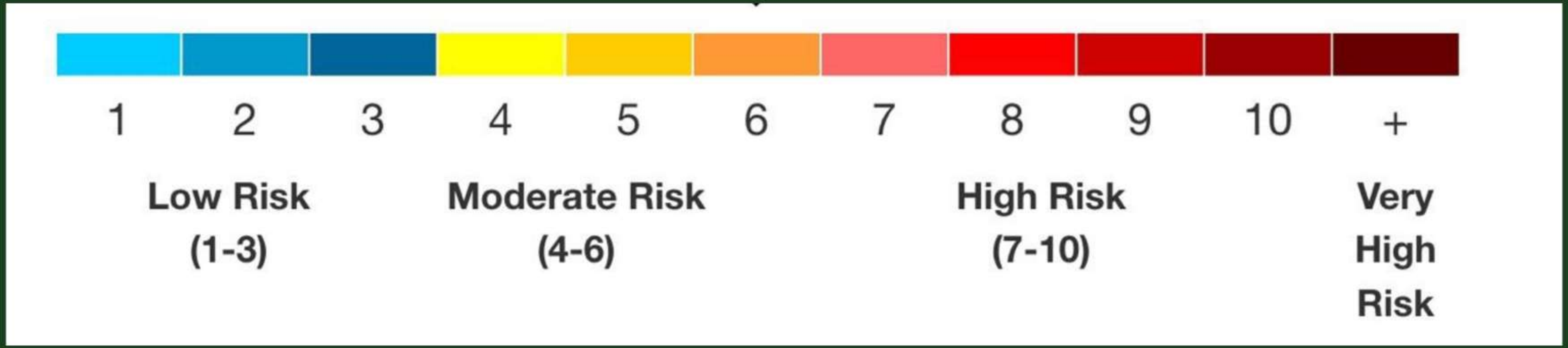
1.Which months have the best and worst air quality?
Seasonality did not appear significant in predicting air quality, as it was excluded from our best additive model.

2.Which communities have the best and worst levels of air quality?
Challenging to determine which communities have the best and worst levels of air quality in general - due to the **complexity of the model**. Generally, Northwest Calgary had lower coefficients, and Northeast Calgary had higher coefficients.

3.What factors contribute most to poor air quality in Calgary or specific communities?
Carbon Monoxide, Methane, Nitric Oxide, Nitrogen Dioxide, Ozone, Non-methane Hydrocarbons, and PM2.5 Mass.

4.Can the model predict future Air Quality?
The **robust model is a strong choice for forecasting (Adjusted R-squared ~94%),** especially if outliers are a concern or if similar irregularities are expected in future data.

Additional insights




```r
# Print the confusion matrix
print(confusion_matrix)
```

```
##                    Actual
## Predicted          Low Risk Moderate Risk High Risk
##   Low Risk             2847            78         0
##   Moderate Risk          74           428         0
##   High Risk               0             0         3
```

```r
# Print the confusion matrix
print(confusion_matrix)
```

```
##                    Actual
## Predicted          Low Risk Moderate Risk High Risk
##   Low Risk             2784            75         0
##   Moderate Risk          62           339         0
##   High Risk               0             0         0
```

Thank You!

Any Questions?

Let's clear the air, one insight at a time, for healthier lives and a greener tomorrow!