Mapping the Polluted Skies

Analyzing Air Quality Metrics

Zitong LUO

Abstract

In recent years, while numerous studies have explored the impact of various factors on air quality indicators, little attention has been given to the relationship between different categorical indices of Air Quality Index. This study aims to conduct a preliminary analysis of the relationship between AQI categorical indices, identifying potential patterns, and examining regional variations. The research investigates available data on air quality. Exploratory Data Analysis provides insights into the dataset, highlighting average AQI values worldwide and identifying highly polluted regions in Asia and Africa. Factors affecting AQI, such as Ozone, PM2.5, CO, and NO2, are examined through statistical tests and correlation analysis. Model building and diagnostics involve K-Means Clustering and linear regression to understand data grouping and estimate PM2.5 using CO, Ozone, and NO2, where we encountered challenges. A case study on India reveals regional variations in AQI patterns, where the AQI gradually increases from the south and east towards the northwest, with one possible explanation of monsoon season in India.

Contents

1 Introduction

1.1 Research Context

Recent years have witnessed the proliferation of evaluating the possible influencing factors of air quality, such as economy, geography and policy. For example, a study investigated the impact of quarantine measures implemented during the COVID-19 pandemic on air quality, with focusing on the Air Quality Index (AQI), PM2.5, and tropospheric NO2 levels (Benchrif et al. 2021).

However, research on the relationship between AQI indices seems to have not been addressed yet. In this study, based on available data, I conducted a preliminary analysis of the relationship between AQI categorical indices, sought possible patterns, and separately analyzed patterns for individual regions.

The paper proceeds as follows. Section 1 provides a brief introduction to the research background and relevant data. Section 2 conducts Explorative Data Analysis. Section 3 focuses on model building and diagnostics. Section 4 presents a case study on India. And section 5 concludes and discusses our findings.

1.2 Data and Variables

The data provided from Kaggle contains 12 unique features (here we only use 8 of them) and 23463 records of city air quality. The dataset uses integer values to describe items of "AQI Value", while it uses verbal descriptions "Good", "Moderate", "Unhealthy for Sensitive Groups", "Unhealthy", "Very Unhealthy" and "Hazardous" to describe items of "AQI Category".

Table 1: Definitions and Summary Statistics of Variables

Variable	Type	Definition	Mean	SD	Min	Max
Country	Factor	Name of the country				
City	Factor	Name of the city				
AQI	Continuous	Overall AQI value of the city	72.01	56.06	6	500
AQI.Cat	Factor	Overall AQI category of the city				
CO	Continuous	AQI value of Carbon Monoxide of the city	1.37	1.83	0	133
Ozone	Continuous	AQI value of Ozone of the city	35.19	28.1	0	235
NO2	Continuous	AQI value of Nitrogen Dioxide of the city	3.06	5.25	0	91
PM2.5	Continuous	AQI value of Particulate Matter with a diameter of 2.5 micrometers or less	68.52	54.8	0	500

2 Explorative Data Analysis

In this section we focuses on gaining a deeper understanding of the data before building models. We aims to provide insights into the dataset, identify important variables, and understand their relationships.

2.1 AQI Overview

The Average AQI Value World Map, by providing a comprehensive overview of average AQI values, highlights areas that are highly polluted and those with relatively cleaner air. Notably, the map demonstrates that Asia and Africa, in particular, exhibit a concentration of highly polluted regions.

One striking observation is that India emerges as the 9th most polluted country. However, what sets India apart from other highly polluted countries is the largest standard deviation of its AQI, implying an unbalanced air quality conditions within India. Since Korea has only one observation in this data, its highest AQI value is not representative.

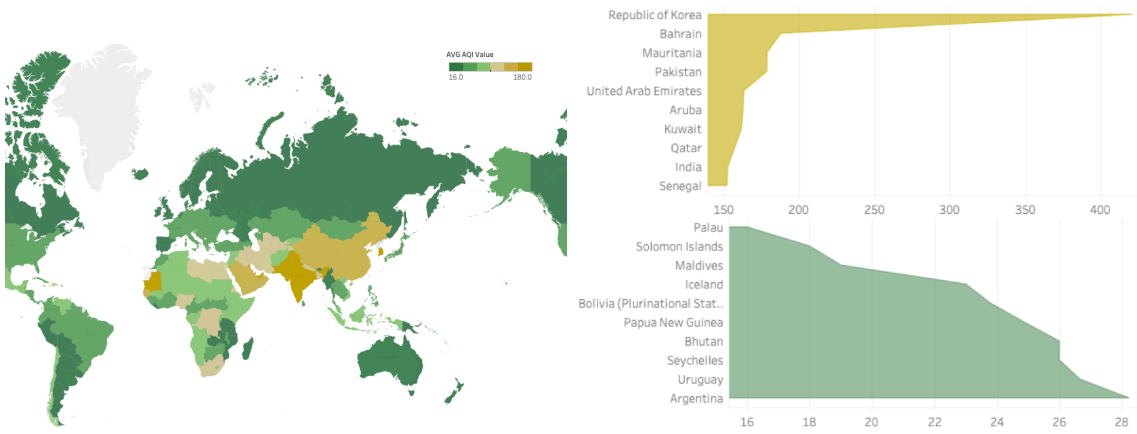

Figure 1: Average AQI Value World Map and Rankings

2.2 Factor Insights

Labeling the AQI categories ('Good' to 'Hazardous') from 1 to 6, the Ridgeline Charts of the pollutant AQI values have shown a relatively significant difference in Ozone and PM2.5 AQI values, while the other two need a further test.

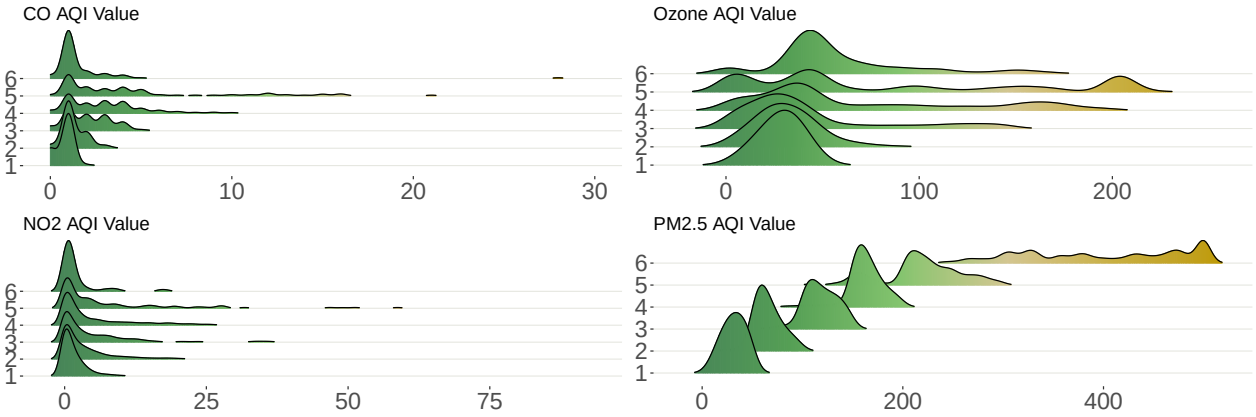

Figure 2: Pollutant AQI Values across Overall AQI Categories

To test the differences in means of the other two pollutants, CO and NO2, we opt to use the Tukey's HSD test for conducting the family-wise comparison. The results in Figure 3 indicate that, with a 95% family-wise confidence interval, the AQI values within CO are significantly different. However, for two pairs (2-6, 3-4) in NO2, there is not much difference.

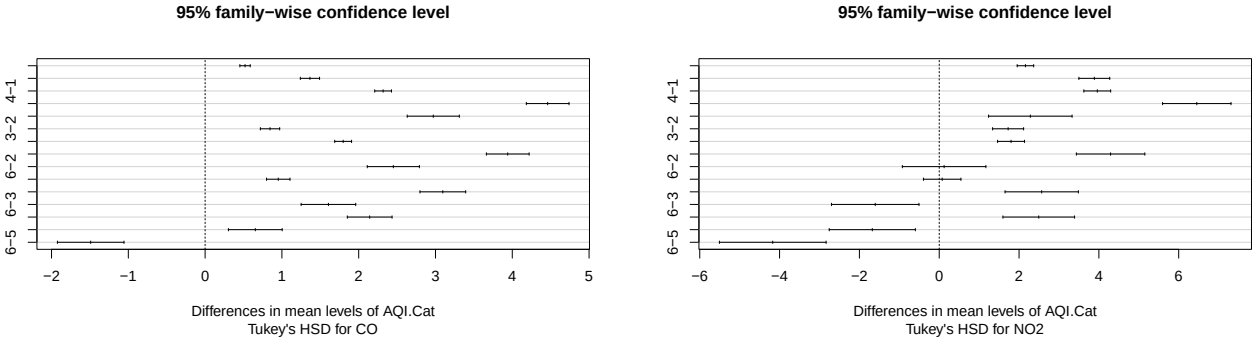

Figure 3: Family-wise CI by Tukey's HSD

We can further look into the correlations among the AQI values. The Correlogram shows an overall positive correlations with one exception between Ozone and NO2. It is noticeable that the correlation between overall AQI values and PM2.5 AQI values has reached 0.98. Based on the calculation formula of AQI that

$$OverallAQI = Max\{CategoricalAQI\} \tag{1}$$

where PM10 and SO2 are also pollutants but not included in the dataset, the correlation implies that PM2.5 is supposed to be the main contributor.

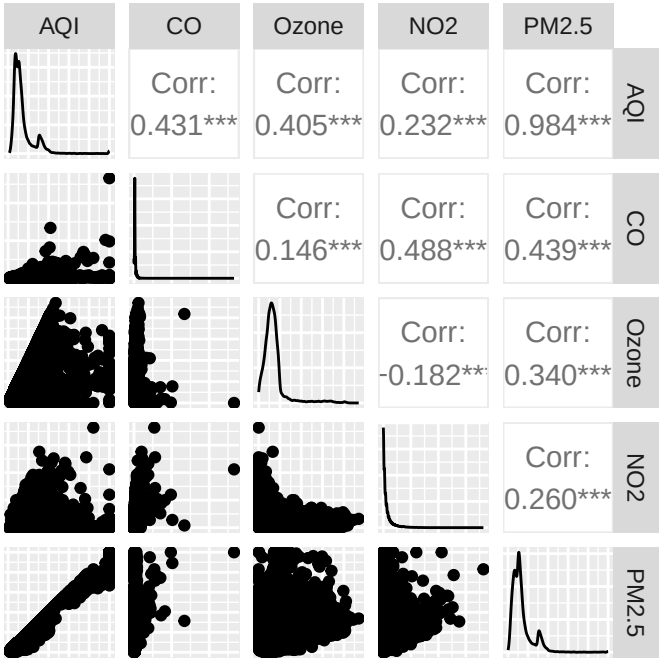

Figure 4: Correlogram of AQI Values

3 Model Building & Diagnostics

In this section, we delve deeper through models. First, we examine whether the observed values can be grouped well by clustering, and then we consider the fitting relationship between variables by linear regression analysis.

3.1 K-Means Clustering

Firstly, we apply K-Means Clustering to see if the data can be well grouped based on 4 categorical AQIs. In order to eliminate the possible scaling effects of variables, we have standardized the variables here.

By plotting Within cluster Sum of Squares against the number of clusters, it seems that there is no significant elbow point (Left panel of Figure 5). Taking k = 3 as an example, the Silhouette Plot in the upper-right panel shows an average silhouette width of 0.63. But the Silhouette Score Plot in the lower-right panel has shown that the data may not be able to well clustered with k > 2. The above information indicates that the data may not be able to well grouped based on 4 categorical AQIs.

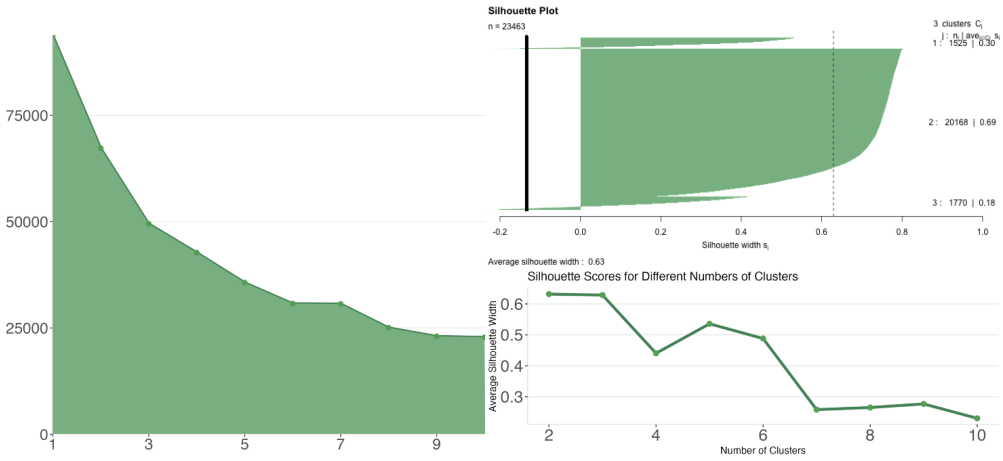

Figure 5: Elbow Plot, Silhouette Plot and Silhouette Score Plot

3.2 Linear Regression Model

As the calculation formula of AQI has fixed (shown in Equation (1)), it is meaningless to study the linear relationship between overall AQI and categorical AQI, so we consider the relationship between categorical AQI. Considering that PM2.5 is the primary contributor to the AQI values, we propose building a Linear Regression Model to estimate PM2.5 based on CO, Ozone, and NO2. As PM2.5 is highly right-skewed, we derive a Box-cox Plot to determine the appropriate transformation (As PM2.5 contains value of 0, here we add 1 to all the values). As shown in Figure 6, within the 95% confidence interval, we choose the lambda of 0.1.

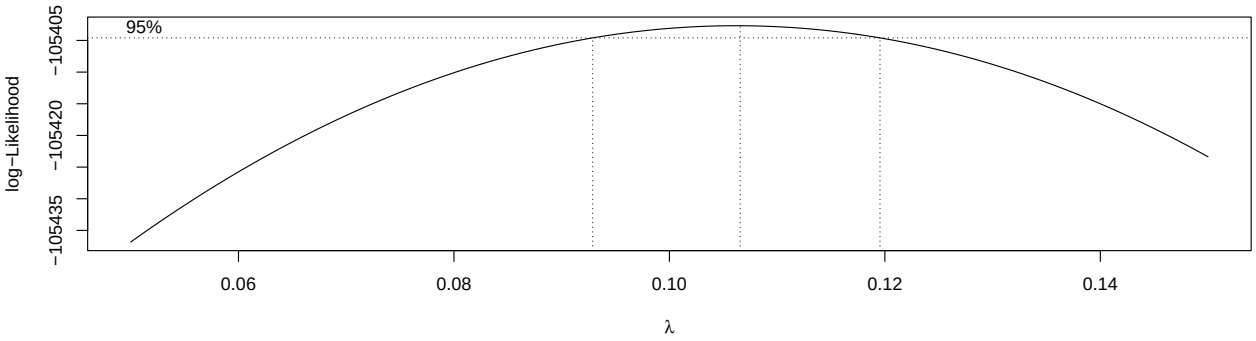

Figure 6: Box-cox Plot of Dependent Variable

After fitting the model

$$(PM2.5+1)^{0.1} = \beta_0 + \beta_1 Ozone + \beta_2 NO_2 + \beta_3 CO + e \tag{2}$$

we get the estimates all significantly differing from 0 (more precisely, greater than 0), indicating that they are all necessary variables. By plotting the diagnostic plots, we observe significant heteroskedasticity, nonlinear relationship and possible outliers.

Table 2: OLS Regression Results

Variable	Estimate	Std.Err.	t-stat	p-value
(Intercept)	1.42092	0.00103	1381.79809	0
Ozone	0.00117	0.00002	54.91086	0
NO2	0.00476	0.00013	36.74853	0
CO	0.01414	0.00037	38.27783	0

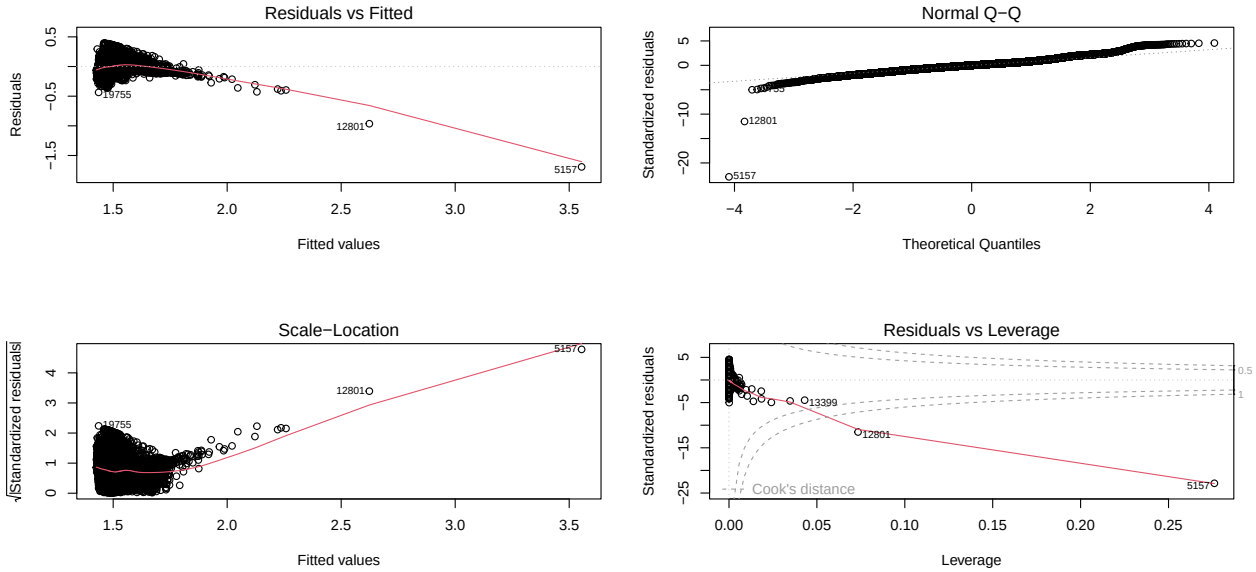

Figure 7: Diagnostic Plots of OLS Regression

To address the issue of unequal variance, we attempt to use a Weighted Least Squares (WLS) model. However, the results are not satisfactory (shown in Appendix Figure 11).

Additionally, we examine possible outliers using the Cook's Distance Plot (Appendix Figure 12), revealing that the 5157th and 12801st observations are considered influential due to their exceptionally high CO and Ozone AQI values, respectively.

In an effort to capture any non-linear relationships, we consider including quadratic & interaction terms in the model. We employ Mallow's Cp as the criteria. However, the result indicates that all nine parameters should be included (Appendix Table A3), with a Cp value of 10. Despite fitting the model with these terms, the problem remains unresolved, and the interpretability of the model is significantly reduced.

The above diagnosis reveals unsatisfactory results, which could be attributed to the limitations of the existing data fitting and significant variations in air quality patterns (economy, geography, policy etc.) across different regions worldwide. Therefore, it may be more valuable to study local and national data than to study global AQI data.

4 Case Study on India

Therefore, we proceeded with an analysis focused on a single country, as mentioned in the beginning, India, which exhibits interesting phenomena in its air quality patterns. The AQI gradually increases from the south and east towards the northwest.

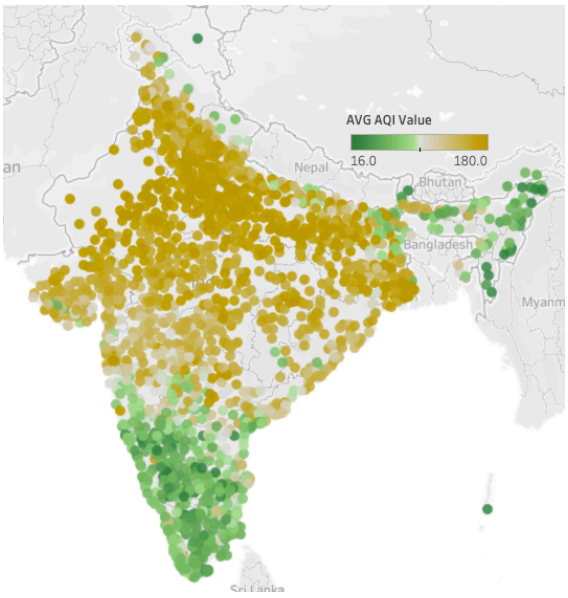

Figure 8: AQI Values for Different Cities in India

This phenomenon is well reflected in regression fitting as well. Here we obtained additional latitude and longitude data for Indian cities and used them as independent variables.

As shown in 3, the regression for

$$OverallAQI = \beta_0 + \beta_1 Latitude + \beta_2 Longitude + e \tag{3}$$

is shown in the left side. And the regression for

$$PM2.5 = \beta_0 + \beta_1 Ozone + \beta_2 NO_2 + \beta_3 CO + e \tag{4}$$

is shown in the right side. From the regression results of this single country, we can see that the AQI index gradually increases as the latitude goes up, and gradually increases as the longitude goes down (moving westward) - which is consistent with the intuitive conclusion we drew from the Figure 8 earlier.

Table 3: Two OLS Regressions Results

Variable	Estimate	Std.Err.	t-stat	p-value	Variable	Estimate	Std.Err.	t-stat	p-value
(Intercept)	113.50865	16.78303	6.7633	0	(Intercept)	97.18935	2.03364	47.79077	0.00000
Latitude	6.60498	0.18517	35.66985	0	Ozone	0.57584	0.04102	14.03974	0.00000
Longitude	-1.48852	0.21499	-6.92349	0	NO2	3.45227	0.58987	5.85257	0.00000
					CO	-0.17803	1.65433	-0.10761	0.91432

In the diagnostic plots of the two regressions, compared with the original OLS regression using the global dataset, there is relatively better satisfaction in terms of linearity, homoscedasticity, normality, and other assumptions.

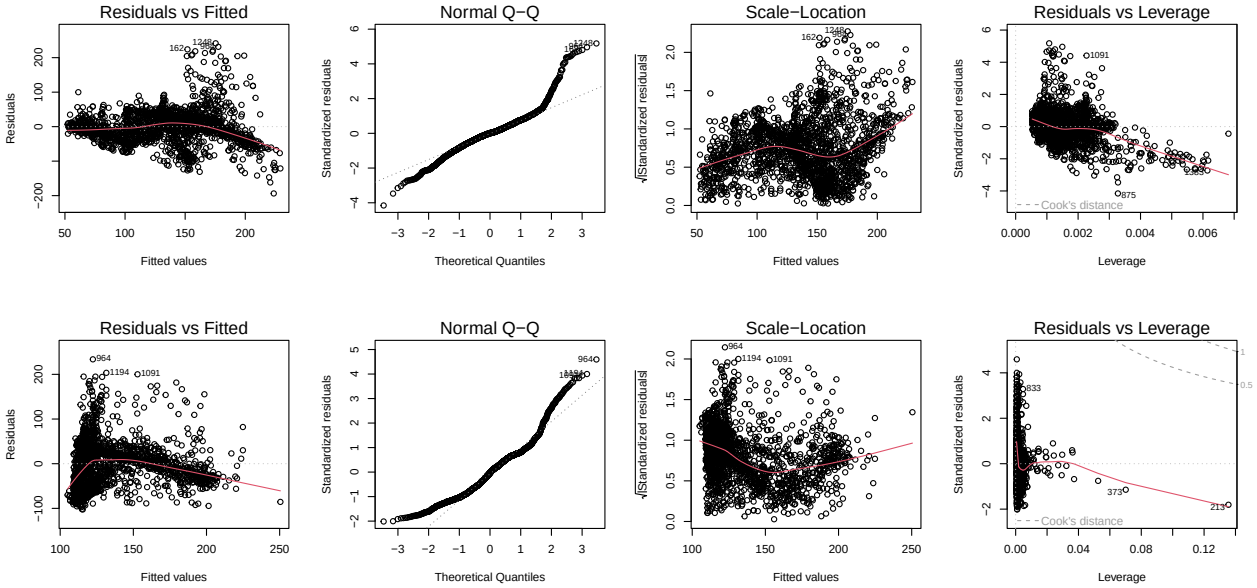

Figure 9: Diagnostic Plots of Two OLS Regressions

One possible explanation for this pattern is the influence of the monsoon season in India. As the Figure 10 shows, the direction of the increase in AQI in India is highly consistent

with the direction of the monsoon. Studies have also shown a high correlation between the Indian monsoon and air pollution (Lau et al. 2009).

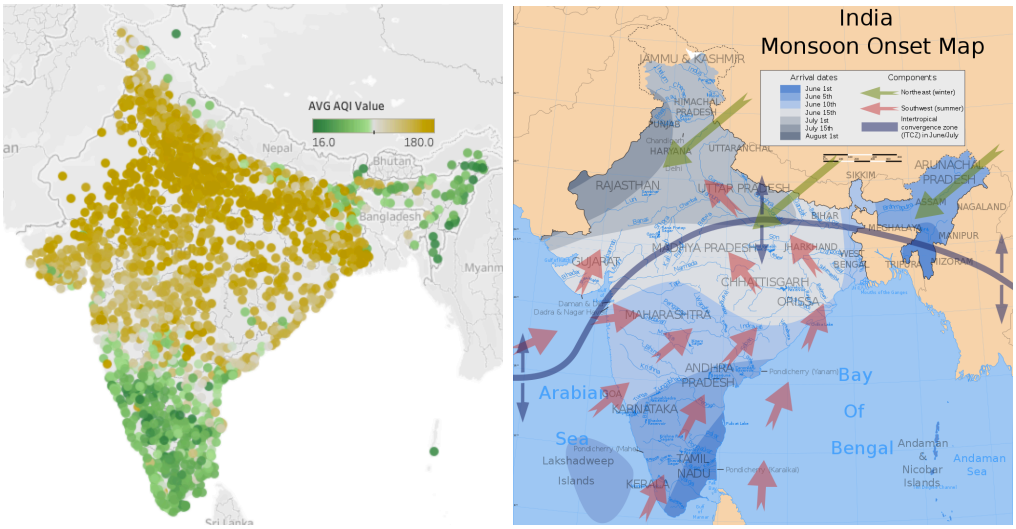

Figure 10: AQI Values and Monsoon Trends Comparison*

5 Conclusion and Discussion

In the analysis of AQI data, we explored various aspects related to pollution. We overview the data in a geographic aspect, finding that Asia and Africa exhibit a concentration of highly polluted regions. We examined different pollutants, with a particular focus on PM2.5 as the primary contributor to AQI values. We attempted to cluster and build regression models to understand the relationships between pollutants, where we encountered challenges which posed difficulties in achieving optimal results. And this could be attributed to the limitations of the existing data fitting and significant variations in air quality patterns across different regions worldwide.

However, with focusing on a single country, the problems seem to be better solved. In our example of India, the AQI values are highly possibly correlated with the monsoon season in India.

The importance of transitioning from a reactive approach to a proactive stance in safeguarding air quality has always been paramount. Through further analysis using more detailed data, for example, predictive analysis with time-series data and causal analysis with more diverse data such as economics, policy, geography we can enhance our understanding of the issue and pave the way for comprehensive protection measures to be implemented.

6 References

[1] Benchrif, A., Wheida, A., Tahri, M., Shubbar, R. M., & Biswas, B. (2021). Air quality during three covid-19 lockdown phases: AQI, PM2. 5 and NO2 assessment in cities with more than 1 million inhabitants. Sustainable Cities and Society, 74, 103170.
[2] Lau, W. K., Kim, K. M., Hsu, C. N., & Holben, B. N. (2009). Possible influences of air pollution, dust-and sandstorms on the Indian monsoon. World Meteorological Organization (WMO) Bulletin, 58(1), 22.
[3] Monsoon. (2023, June 11). In Wikipedia. https://en.wikipedia.org/wiki/Monsoon

A Appendix

WLS results and related Diagnostic Plots are shown below:

Table A1: WLS Regression Results

Variable	Estimate	Std.Err.	t-stat	p-value
(Intercept)	1.41545	0.00105	1346.23504	0
Ozone	0.00110	0.00002	47.32261	0
NO2	0.00384	0.00014	27.04861	0
CO	0.02228	0.00048	46.18811	0

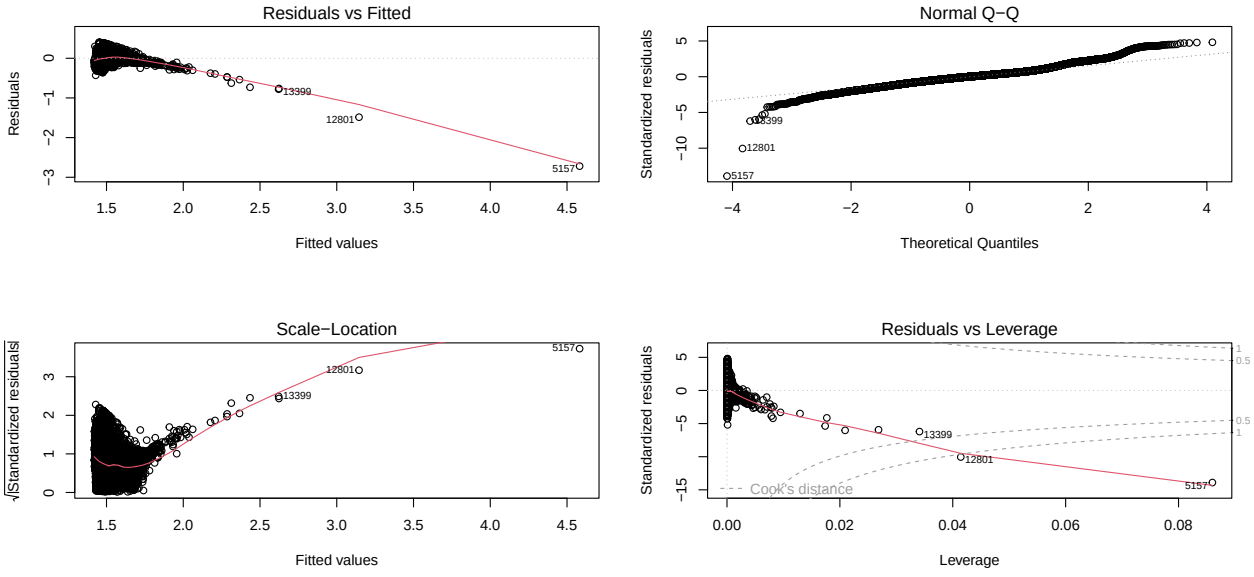

Figure 11: Diagnostic Plots of WLS Regression

Cook's Distance Plot and detailed information of outliers are shown below:

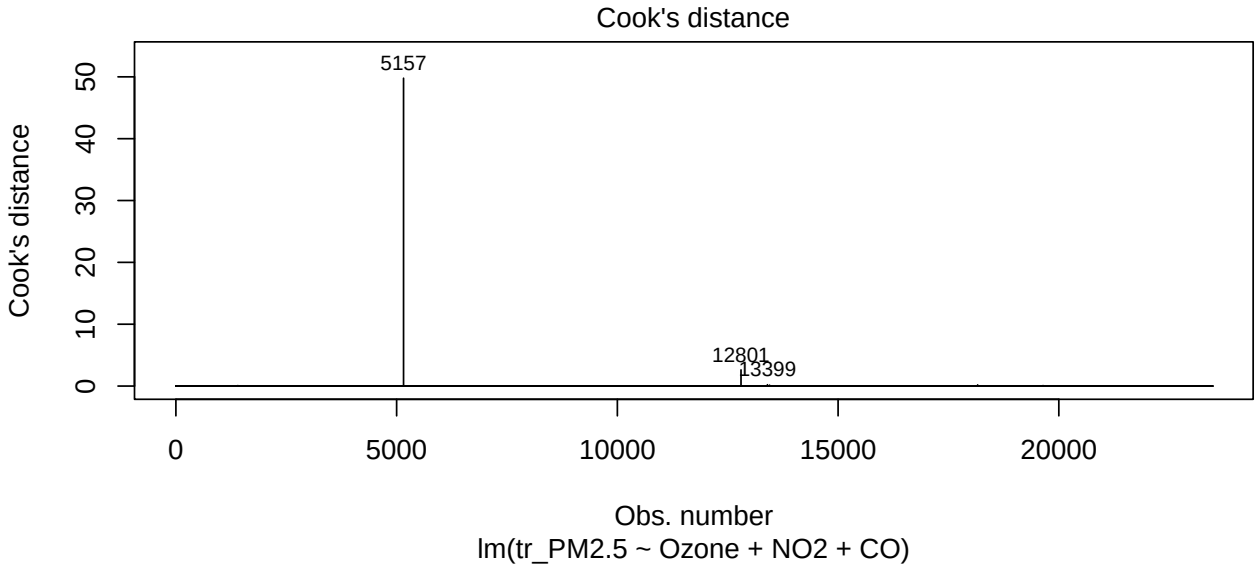

Figure 12: Cook's Distance Plot

Mallow's Cp model selection results are shown below:

Table A2: Determined Outliers

	Country	City	AQI	CO	Ozone	NO2	PM2.5
5157	United States of America	Durango	500	133	0	53	500
12801	Malaysia	Miri	209	67	209	2	157

Table A3: Mallow's Cp Model Selection Results

	CO	NO2	Ozone	CO*CO	NO2*NO2	Ozone*Ozone	CO*NO2	CO*Ozone	NO2*Ozone
9	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
8	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
8	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE