Analyzing the Influence of Pollutants and Meteorological Conditions on PM2.5 Concentration

[redacted]

Introduction

Fine particulate matter (PM2.5), a type of tiny particle in the air with a width of two and one-half micrometers or less, is considered to be one of the atmospheric pollutants whose effects are the greatest on public health. In 2012, air pollution in urban and outdoor areas caused 3.7 million people's death in the world (Li et al. 2017), and the World Health Organization (WHO) designated it as a Group 1 carcinogen (WHO 2017).

Beijing, a fast-growing capital city with a large urban population, has suffered from PM2.5 in the last 20 years. In 2013, the severity of the air pollution problems in Beijing peaked. According to data from Beijing municipal environmental monitoring center, the air quality approached a heavy pollution degree, and the concentrations of PM2.5 reached up to 1000 μg/m³ (Li et al. 2014). Serious air pollution not only causes health risks and economic losses but also affects the city's image in front of the world. In order to improve air quality, Beijing launched its "Five-year Clean Air Action Plan" from 2013 to 2017, aiming to accelerate the process of reducing the concentration of pollutants (Xu et al. 2021).

Since PM2.5 is a big part of the air quality concern, several research studies have focused on this specific type of air component. In particular, previous studies have shown that there is a close relationship between the level of PM2.5 and other air pollution sources. For example, a moderate to strong positive correlation was found in PM2.5 with SO_2 ($r = 0.449$) and with NO_2 ($r = 0.498$) (Xie et al. 2015). Other research also implied the influence of meteorological conditions on PM2.5 concentrations: spatial and temporal patterns of PM2.5 are closely analyzed by causality analysis models, and interaction between PM2.5 and temperature, wind speed, wind direction, humidity, precipitation, radiation, atmospheric pressure, and planetary boundary layer height. (Chen et al. 2020) This leads to a discussion of the effect of other air pollutants and meteorological conditions on PM2.5.

The purpose of this project is to focus on the level of SO_2, NO_2, CO, and O_3 and the weather conditions, including temperature, air pressure, wind direction, and precipitation to model how they correlate with, and potentially impact, the concentration of PM2.5. After understanding the potential interactions between PM2.5 and these factors, a detailed evaluation of the effectiveness of the current policies enacted in Beijing will be provided, and suggestions for future actions will be given.

Methods

The data set used in the article was obtained from the UCI Machine Learning Repository, an online platform that contains collections of databases, domain theories, and data generators for empirical analysis (Zhang et al. 2017). Song Xi Chen from the Guanghua School of Management of Peking University collected this data set for research paper *Assessing Beijing's PM2.5 Pollution: Severity, Weather Impact, APEC and Winter Heating* (Liang et al. 2015). The observations in the dataset are the hourly air pollutants data from 12 nationally controlled air-quality monitoring sites, including PM2.5, PM10, SO_2, NO_2, CO, and O_3, all of which are measured in $\mu g/m^3$, as well as temperature (°C), pressure (hPa), dew point temperature (°C), precipitation (mm), wind direction, and wind speed (m/s).

Note that the particular interest of this study is only the site Tiantan since Tiantan is one of the most famous scenic spots and historical sites in Beijing: its large exposure to citizens and tourists makes it a good representation of the overall air quality in Beijing. The Tiantan air quality data used is part of the "Beijing Multi-Site Air-Quality Data Data Set" obtained by the Beijing Municipal Environmental Monitoring Center. The full dataset that was downloaded includes information from 12 nationally controlled air quality monitoring sights around Beijing. Additionally, data from the nearest weather station from the Chinese Meteorological Administration was used to supplement this data. The data were collected between March 1st, 2013, and February 28th, 2017.

Result

Multiple linear regression models were used when analyzing the data. According to the dataset, the response variable PM2.5 is continuous and quantitative; so are the levels of PM10, SO_2, NO_2, CO, O_3, temperature, pressure, and rain. The purpose of conducting a multiple linear regression is to see the power

of influence of the eight independent variables on the level of PM2.5. This could also be used to predict the future level of PM2.5 and provide insight into which factors should be controlled the most.

To achieve the optimal model with the least possible number of variables, an AIC backward selection was employed in the original regression model. According to the result, pressure (PRES) should be removed from the model. Having PM2.5 as the response variable, PM10, SO_2, NO_2, CO, O_3, TEMP, and RAIN will generate the lowest AIC (224741.5), indicating that the best fit model for the PM2.5 is:

$$\text{PM2.5} = \beta_0 + \beta_1\text{PM10} + \beta_2\text{SO}_2 + \beta_3\text{NO}_2 + \beta_4\text{CO} + \beta_5\text{O}_3 + \beta_6\text{TEMP} + \beta_7\text{RAIN} + \varepsilon$$

Before continuing with the analysis, the normality and constant variance assumptions for regression were checked. To begin with, for a linear regression to be appropriate to use, the residuals should be normally distributed. A histogram of the residuals generated by the AIC-selected model was plotted (Figure 1). According to the plot, the residuals are centered at 0 and are roughly bell-shaped, generally implying a normal distribution. However, different conclusions were drawn when a quantile-quantile plot (QQ-plot) was graphed (Figure 2). Since the data points do not lie on the straight line, the distribution of the residuals is not normal, indicating that the normality assumption is not met. This might be due to the time-series characteristic or the long tail, symmetric distribution of the dataset.

Moving on, the common variance of residuals was verified by graphing the diagnostic plots. The fitted values were plotted on the x-axis, and the residuals for the model were graphed on the y-axis (Figure 3). The plot shows that there is a lower bound cutoff on the residual values. The plot also presents a "conning effect" of the residuals, which could indicate the heteroscedasticity of variances. Aiming to resolve the non-constant variance, logarithm (Figure 4) and square root (Figure 5) transformations were employed for the response variable. However, heteroscedasticity of variances still existed, which might be probably due to a variable in the dataset (either a predictor or the response) that has a natural boundary on it that is cutting off the data on the lower side.

Finally, outliers and influential data points were examined. Here, outliers were defined as data points that are outside three standard deviations from the mean ($z < -3$ or $z > 3$); influential data points were defined as observations with Cook's Distance greater than the 50^{th} percentile of the F distribution with $v_1 = 7$ and $v_2 = 32835$ degrees of freedom. 462 outliers were found, and no influential data points were observed. According to the scatter plot of the level of PM2.5 across the time points (Figure 6), the red points represent the outliers. These outliers would not be ignored from the dataset because most of them are on the higher end of the distribution of PM2.5 with an approximate cyclical trend across the time. This might be due to the assumptions not being met for linear regression and suggests that other models should be employed in the future to see if they provide a better fit for the data. Note that the data were measured and collected on a timely basis, meaning that the independence assumption might be violated. This potential issue could possibly be solved by performing a time series analysis. However, since the knowledge learned didn't support such modeling and research, multiple linear regression models were used for the following analysis.

After testing the conditions, the results from the AIC-selected model were investigated. The equation below was the expression of the concentration of PM2.5 in terms of the level of PM10, SO_2, NO_2, CO, O_3, temperature, and precipitation.

$$\text{PM2.5} = -22.069 + 0.566\text{PM10} + 0.019\text{SO}_2 + 0.185\text{NO}_2 + 0.0216\text{CO} + 0.038\text{O}_3 + 0.276\text{TEMP} + 0.913\text{RAIN} + \varepsilon$$

According to this, the level of PM10, SO_2, NO_2, CO, O_3, temperature, and precipitation are all positively correlated with the level of PM2.5. Among these variables, precipitation, temperature, and PM10 have the largest coefficients, indicating that they impose a greater influence on the level of PM2.5. What's more, a high adjusted R^2 of 0.8532 suggests that around 85% of the variation in PM2.5 can be explained by the linear relationships with the explanatory variables in this model. This is relatively high, indicating that the model has a high goodness of fit.

Discussion

Based on the results, except for pressure, all other variables originally hypothesized to affect the response were statistically significant in predicting PM2.5. This is reasonable because pressure might have a competing predictive power with precipitation since low pressure often causes more rain (Yu et al.

2018). Practically, the positive correlation between temperature and PM2.5 concentration can be empirically justified: because of the temperate continental climate in Beijing, the wind is lower during summer, which makes the pollutants suspended in the still air for a longer time, whereas the howling wind during winter promotes the diffusion of the pollutants and thus decrease the level of PM2.5.

As mentioned above, there are serious limitations related to the use of the linear model. When checking the assumptions for linear regression, three violations were observed: the data points in the dataset are not independent of each other, the residuals are not normally distributed, and there is no common variance for the residuals. Because of this, a linear regression model should be used with caution even if it provides a high adjusted R^2, especially when there is huge cyclical fluctuation for the response variable across the time.

Broadly speaking, these models are still not sufficient to decide which variables provide the greatest influence on the level of PM2.5, since a dominant factor was not found here. For future research, factor analysis can be conducted, and structural equation models can be built to group observed variables into factors to better understand the existence of the latent variables. Moreover, for data collection, more information should be collected for each month, such as the level of human activities, to help researchers get a sense of how the level of PM2.5 is related to people's daily life. This can also help the citizens be aware of the air condition issues and make policies or change their lifestyle based on the results.

Conclusion

The meteorological conditions are hard to control, but the level of other pollutants such as PM10, SO_2, NO_2, CO, and O_3 can be manually adjusted or manipulated through policies. Since the level of these pollutants usually changes accordingly with PM2.5 concentration, it's reasonable to assume that these pollutants might come from the same source or might have interaction effects on each other. Thus, if other pollutants are controlled properly, the level of PM2.5 may also be restricted effectively.

In the National Air Quality Action Plan (2013), the Chinese government implemented several policies regarding coal combustion limitation, vehicle emission standards, and transparent air quality reporting. Results show that coal combustion contributes to 40% of the total PM2.5 concentration on the national average (Ma, et al. 2017). Since 2013, new coal-fired plants were prohibited in target regions, and existing coal plants were required to reduce emissions or be replaced with natural gas. In 2017, China's largest coal producer Shanxi Province closed 27 coal mines. By early 2018, Beijing had shut down its last coal-fired power plant and canceled its future building plan (Air Quality Life Index). These actions reduced coal combustion pollutants such as SO_2, CO, and NO_2, which might indirectly control the PM2.5 concentration.

Moreover, in large cities such as Beijing, Shanghai, and Guangzhou, the number of cars on the road on any given day and the number of new license plates issued each year were restricted to control vehicle emissions. The emissions standards were strictly enforced as well. "In late 2017, China suspended the production of 553 car models that do not meet fuel economy standards, including the ones made by foreign and state-run companies." (Air Quality Life Index) Controlling vehicle emissions and implementing stricter standards for fuel standards, the Chinese government successfully controlled the emission of PM2.5, PM10, SO_2, NO_2, and CO.

Last but not least, China has built a nationwide network of air pollution monitors and makes the data available to the public. Over 5000 monitoring stations were built in China by March 2017 (Air Quality Life Index). The increasing transparency in government reporting of air quality statistics increased public awareness and engagement in controlling air quality and facilitated the process of reducing PM2.5 concentrations.

Other than directly controlling PM2.5 concentration in the air, reducing emissions of other polluted particles such as PM10, SO_2, NO_2, and CO is also an effective approach, and the Chinese government has already put great effort into improving the air quality. The policies above aligned with the conclusion of this study, and they indeed effectively reduced PM2.5 concentration and improved air quality in Beijing. The next step would be to continue the implementation of these policies and adjust based on the future predictions and analysis of PM2.5 concentration.

Appendix

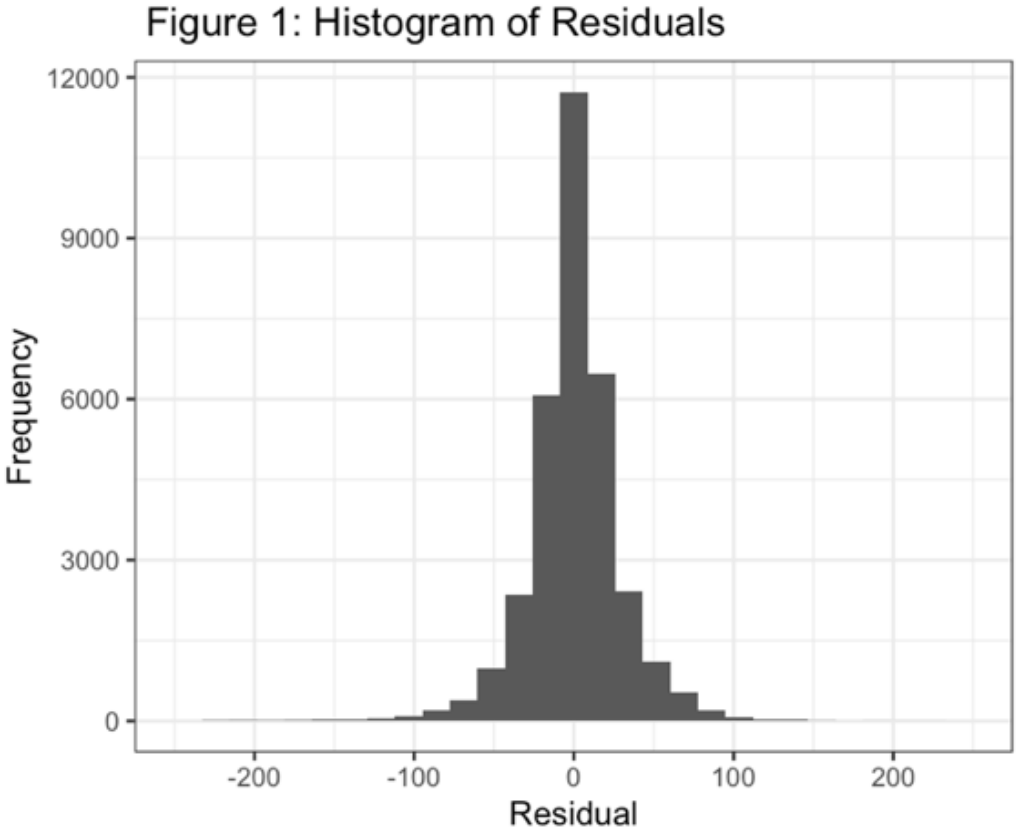


Figure 1: Histogram of Residuals
12000
9000
6000
3000
0
Frequency
-200
-100
0
100
200
Residual

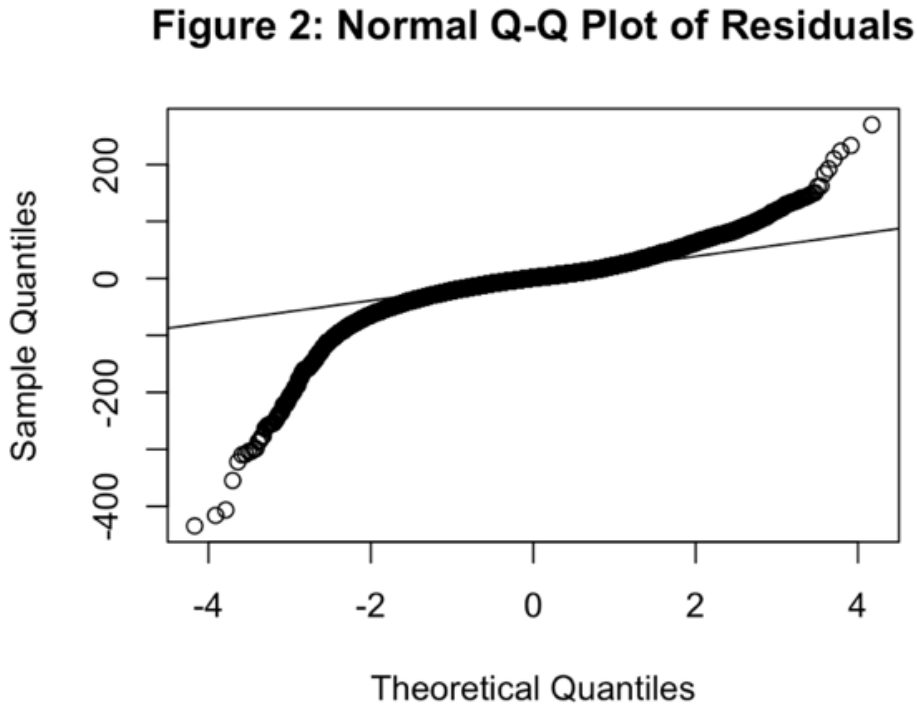


Figure 2: Normal Q-Q Plot of Residuals
Sample Quantiles
200
0
-200
-400
-4
-2
0
2
4
Theoretical Quantiles

Figure 3: Diagnostic Plot for PM2.5

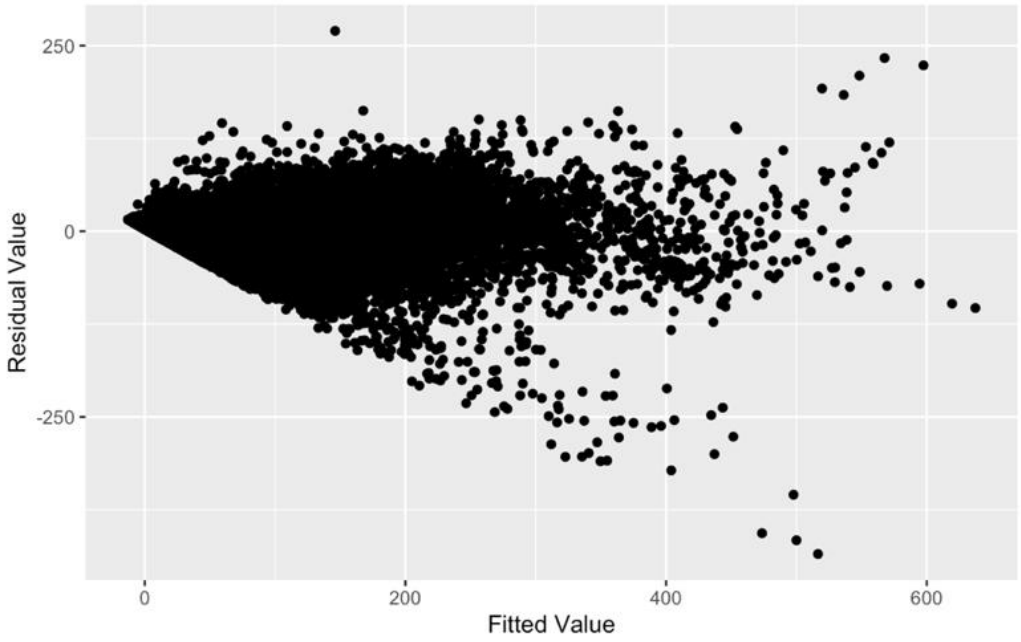


Figure 4: Diagnostic Plot for ln(PM2.5)

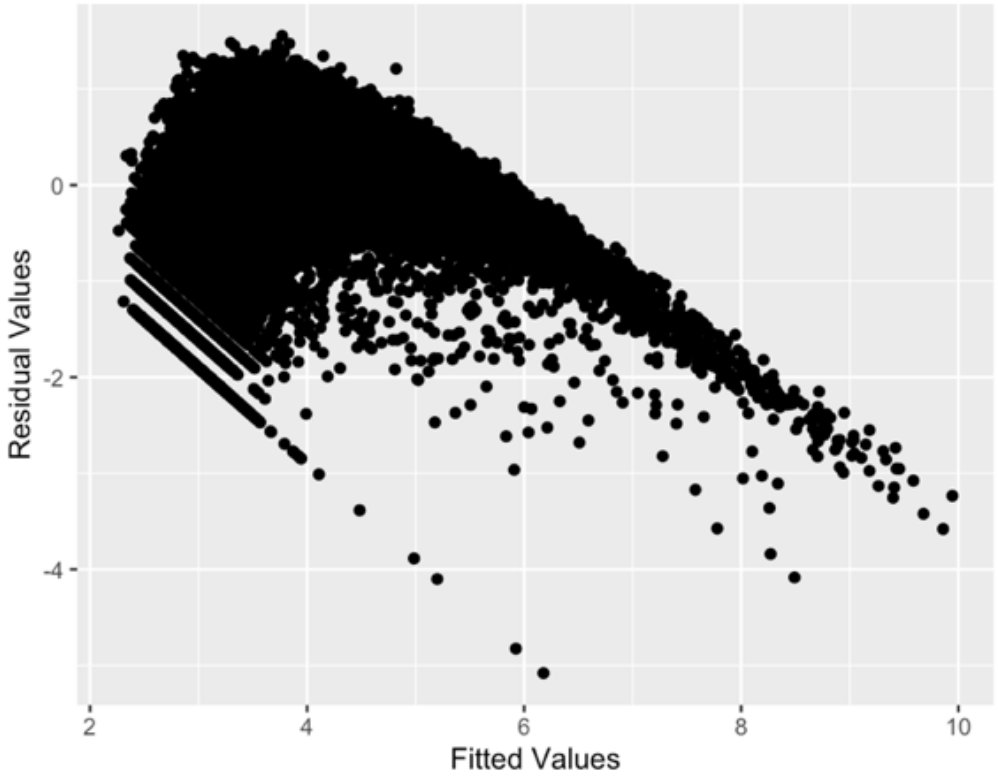


Figure 5: Diagnostic Plot for Square Root of PM2.5

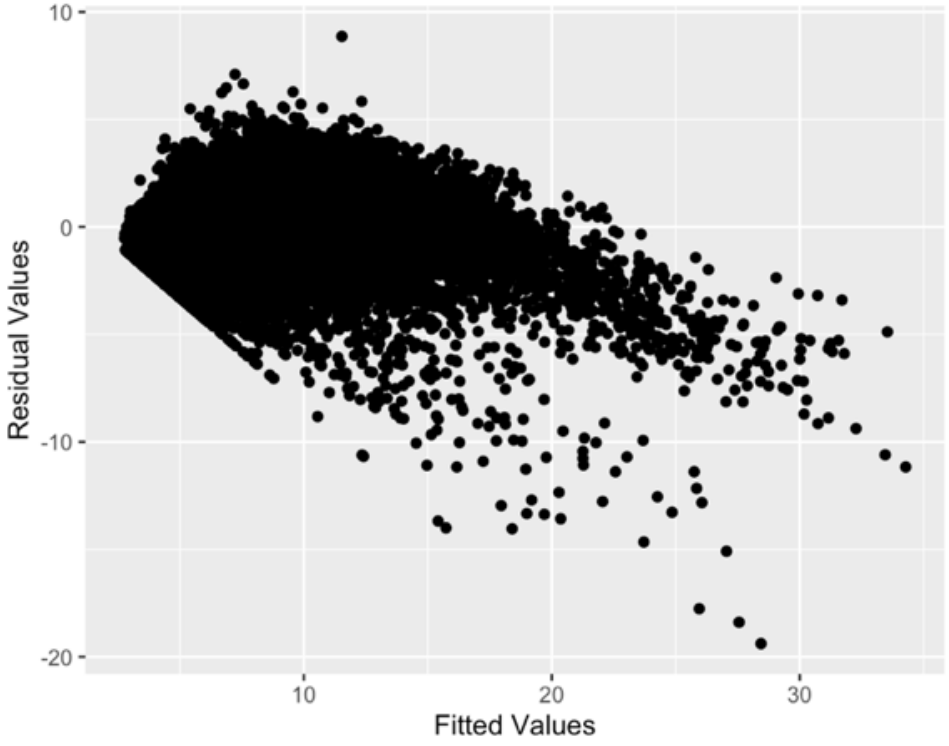
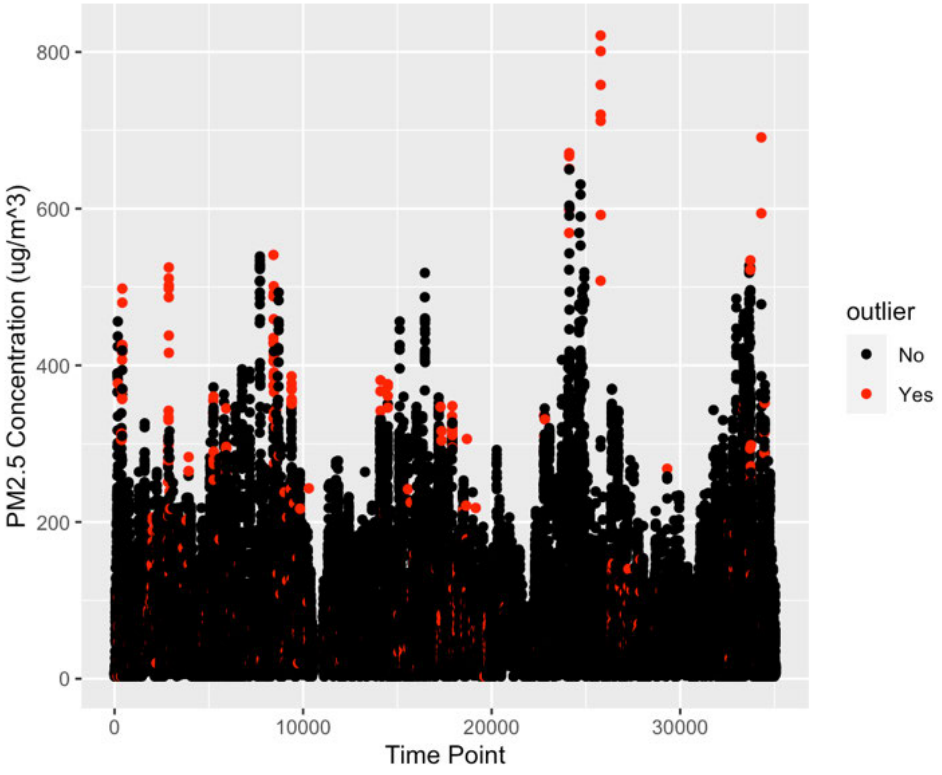


Figure 6: Scatter Plot of the Level of PM2.5 Across Time Points



Bibliography

Dataset:

Zhang, S., et al. "Cautionary Tales on Air-Quality Improvement in Beijing." *Proceedings of the Royal Society A*, Volume 473, 2017 No. 2205, Pages 20170457. https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data.

Source:

Air Quality Life Index. "China: National Air Quality Action Plan (2013)." *Air Quality Life Index*. https://aqli.epic.uchicago.edu/policy-impacts/china-national-air-quality-action-plan-2014/.

Chen, Ziyue, et al. "Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism." *Environment International*, Volume 139, 2020, 105558, ISSN 0160-4120, https://doi.org/10.1016/j.envint.2020.105558.

Li, Li, et al. "The health economic loss of fine particulate matter (PM2.5) in Beijing." *Journal of Cleaner Production*, Volume 161, 2017, Pages 1153-1161, ISSN 0959-6526. https://doi.org/10.1016/j.jclepro.2017.05.029.

Liang, Xuan, et al. "Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating." *Proceedings of the Royal Society*. A, Mathematical, Physical, and Engineering Sciences, 471(2182), 20150257–20150257. 2015. https://doi.org/10.1098/rspa.2015.0257.

Ma, Li & Zhao, Qingzhu. "The first 'annual report' on PM2.5: Beijing experienced severe pollution one day per week on average in 2013." *The Beijing News*. January 3, 2014. http://scitech.people.com.cn/n/2014/0103/c1007-24012560.html.

World Health Organization. Regional Office for Europe. "Evolution of WHO air quality guidelines: past, present and future." *World Health Organization*. Regional Office for Europe. 2017. https://apps.who.int/iris/handle/10665/341912.

Xie, Yangyang, et al. "Spatiotemporal variations of PM2.5 and PM10 concentrations between 31 Chinese cities and their relationships with SO_2, NO_2, CO and O_3." *Particuology*, Volume 20, 2015, Pages 141-149, ISSN 1674-2001, https://doi.org/10.1016/j.partic.2015.01.003.

Xu, Xianmang, et al. "Health risk and external costs assessment of PM2.5 in Beijing during the 'Five-year Clean Air Action Plan'." *Atmospheric Pollution Research*, Volume 12, Issue 6, 2021, 101089, ISSN 1309-1042, https://doi.org/10.1016/j.apr.2021.101089.

Yu, Ziwen, et al. "The bridge between precipitation and temperature – Pressure Change Events: Modeling future non-stationary precipitation." *Journal of Hydrology*, Volume 562, 2018, Pages 346-357, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2018.05.014.