STAT 443 Final Consulting Project:
$PM_{2.5}$ levels in Gucheng, a region in Beijing, China

Group 4 members: Mengjia Zeng, Junseok Yang, Lavanya Upadhyaya, Wasay Siddiqui
Client: Theren Williams and Dr. Darren Glosemeyer

Project Abstract

This project focuses on pollutants such as $PM_{2.5}$ that impact the chemical composition of the air, potentially impact flora and fauna in a harmful manner. The aim is to identify specific attributes, such as precipitation and humidity, temperature, concentrations of various greenhouse gasses such as SO_2 and O_3, time of day, and seasonal changes that might influence $PM_{2.5}$ levels. To effectively relay this information, a complete linear regression analysis and variable selection techniques are completed. In addition to regression models, diagnostics are performed to evaluate model assumptions and investigate if there are any observations with a large influence on the analysis. Lastly, machine learning techniques, such as cross–validation, regularization and so on, are used to train and test the model on various iterations. The ultimate goal of this analysis is to identify variables that impact the $PM_{2.5}$ concentration in the atmosphere and predict future pollutant levels to increase caution and suggest human behavioral changes to mitigate such problems.

Table of Contents

Introduction

Climate change and fluctuations of pollution in the atmosphere has been studied for decades now, since it has a tremendous impact on human health, as well as natural rehabilitation and growth. A large part of China is experiencing substantial amounts of air pollution with impactful fine particulate matter, or PM. This study used the dataset collected in Gucheng, Beijing, with $PM_{2.5}$ and PM_{10}, where the numbers refer to fine PM with aerodynamic diameters of less than 2.5 and 10 micrometers (μm) respectively [1]. The official air quality statistics, which are taken from averaging hourly readings, are composed using $PM_{2.5}$ data from state-controlled monitoring sites.

Description of the relevant data

The goal of this project is to look at 6 main air pollutants and 6 relevant meteorological variables from 12 nationally-controlled air-quality monitoring sites in Beijing, China. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017. We used the UCI Machine Learning Repository to obtain our data, and after viewing the peer-reviewed academic journal where this data came from, we are convinced that it is accurate.

The kinds of variables that significantly impact $PM_{2.5}$ can be broken down into various "groups":

- Time variables: Year, Month, Day, and Hour
- Various pollutants and greenhouse gasses: $PM_{2.5}$, PM_{10}, NO_2, CO, O_3
- Weather variables: TEMP (temperature in ºC), PRES (pressure in hPa), RAIN (mm of rain fallen/precipitation), wd (wind direction, such as NW, N, SW, S, etc.), and WSPM (wind speed per minute). There were a few more variables such as DEWP (dew point temperature), but they were not significant to this dataset.

To determine which data affects the $PM_{2.5}$ concentration in the atmosphere, we performed visualizations determining these relevant variables. Not all the data is relevant for the results we are trying to achieve, since we want to observe the most harmful pollutants in the atmosphere, climate conditions that impact or exacerbate the spread of these pollutants, and the concentrations of these chemicals. Even though we kept the missing values in our visualizations so as to not lose too many data points, we are only factoring in necessary variables.

The variables that we need are as follows in Table 1, this data is based on which variables are statistically significant.

Variable name	Description
year	year of data in this row
month	month of data in this row
day	day of data in this row
hour	hour of data in this row
$PM_{2.5}$	$PM_{2.5}$ concentration (ug/m^3)
PM_{10}	PM_{10} concentration (ug/m^3)
NO2	NO_2 concentration (ug/m^3)
O3	O_3 concentration (ug/m^3)
CO	CO concentration (ug/m^3)
TEMP	Temperature (degree Celsius)
PRES	Pressure (hPa)
RAIN	precipitation (mm)
wd	wind direction
WSPM	wind speed (m/s)

Table 1: Description of variables that influence pollutant concentration

Methodology

We mainly applied exploratory analysis and linear regression, with some other machine learning techniques on the Gucheng data to examine significant predictors affecting the $PM_{2.5}$ data and develop a possible model to predict the $PM_{2.5}$ with information from the previous day. We used R to do our data analysis and reports, as well as the statistical methods explained below:

- Linear regression:
    - Diagnostics: Outliers, assumptions of linearity/constant variance/normality. Essentially, the diagnostics show us any points that may influence the dataset and how to modify the model to avoid those trends and behaviors.
    - Variable selection methods (Forwards and Backwards methods)

    - Multicollinearity
- Machine Learning methods
    - Cross validation
    - Regularization
    - K-nearest neighbors algorithm
    - Multinomial Regression

The number of observations with missing values was 7.3% in relation to the overall size, so we initially considered dropping these values. However, after further consideration dropping all observations with missing values could affect model performance. Due to this, we concluded that we would only drop observations with missing $PM_{2.5}$ values. This restructured dataset is what we utilized in creating visualizations in relation to the various predictors, as well as regression modeling.

In addition, a new categorical variable, PM2.5_Type, was added to the new dataset based on the critique that with the $PM_{2.5}$ concentration lower than 35 ug/m^3, the condition was reported as Low. With the $PM_{2.5}$ concentration higher than 35 ug/m^3, but lower than 75 ug/m^3, the condition was reported as Medium. With the $PM_{2.5}$ concentration higher than 75 ug/m^3, but lower than 105 ug/m^3, the condition was reported as High. Lastly, with the $PM_{2.5}$ concentration higher than 105 ug/m^3, the condition was reported as Dangerous.

Regarding our datasets for the purpose of creating visualizations, we created 4 different tibbles to represent predictor relationships with PM 2.5. To start, the visualization datasets were established using the maximum and average value of $PM_{2.5}$ and average value of all the other variables within one day to represent the different categorization of time. These different categories included year, month, weekday, and hour. For the variable RAIN, the total value was used to represent the precipitation of the corresponding time period, while the variable wd was used to represent the mode of the wind direction of the time period..

For the modeling dataset, we wrangled the dataset by grouping by the Date_ymd format. An example would be 2013-03-02 00:00:00, in which Year, Month, Day and Hour were combined. Following this, we created a column value for maximum daily PM 2.5 values to represent each entire day. An additional column created for the purpose of predicting maximum PM 2.5 values for the next day called Max_Tmrw_PM2.5. In dealing with missing values, we dropped these observations as the total proportion of them in relation to the dataset was only 2.9%, or 43 observations.

This entire modeling dataset was then split into training and testing dataframes, following a 80:20 split. Different models would be trained based on the training dataset, and evaluated using the performance of the model on the testing dataset.

We first applied the linear regression model to the training dataset and did stepwise variable selection with BIC as the criteria because many of the predictors were insignificant. After the variable selection, we were concerned about the assumptions for the linear regressions. Therefore, we used a lot of methods to check whether the model fulfilled the assumptions of the linear regression, like using VIF to check multicollinearity issues and some diagnosis plots to check normality and constant residuals. For the diagnosis plots (see Appendix B), the normality assumption was violated, which then forced us to consider using Box-Cox transformation in order to fulfill the assumptions of linear regression. From the Box-Cox plot, we chose to transform the response variable with the lambda to be 0.4. After the transformation, the diagnosis plots, especially the plot related to the normality, would fulfill the requirement. We denoted the model developed through this process as the transformed model.

In order to improve the performance of the linear regression model, we also tried Lasso and Ridge regression with cross validation technique. Following the cross validation technique, we could choose the corresponding best lambda as the one with the minimum mean cross-validated error and refit the model.

Since the concentration of $PM_{2.5}$ could be converted into a categorical variable, Dangerous, High, Medium and Low, with the criteria mentioned above, we also tried two different machine learning techniques: the Multinomial Regression Model and the k-Nearest-Neighbor (KNN) Algorithm. For the multinomial regression model, we tried different combinations of predictors and evaluated them based on the whole predicting accuracy. As for the KNN algorithm, we tried k from 1 to 30 and chose the value of k to be the one with highest predicting accuracy.

As for the comparison among models mentioned above, we first converted the concentration of $PM_{2.5}$ predicted by the transformed, Lasso and Ridge model into a categorical one and evaluated the five different models based on two different criteria. The first criterion was the total predicting accuracy, meaning the percentage of the model correctly predicting different levels of $PM_{2.5}$. The second criterion, denoted as underpredicted error, was about the numbers of underpredicted observations, specifically Dangerous or High to be predicted as Medium or Low. One assumption we had was that people should be careful and avoid going out in the category of Dangerous or High of $PM_{2.5}$, while going out with the category of $PM_{2.5}$ to be Medium or Low, even with the prediction to be Dangerous or High, should be acceptable and not harm human health. Although we wanted to have as few observations that are predicted incorrectly as possible, considering our ultimate goal was to alarm situations of Dangerous or High to people and prevent them from any outdoor activity to protect their health, predicting observations as Dangerous or High that were actually Medium or Low may be acceptable in this case. Because even though people may be alert, going out may not harm their health. But for the opposite situation, predicting observations as Medium or Low, while the actual situation was Dangerous

or High, would not be acceptable, since people may not be alerted and go out that harm their health.

Results

Visualizations

To determine what specific aspects that may influence $PM_{2.5}$ levels, a set of visualizations were produced based on the different time-series datasets.. The concentrations change based on the time of day (hourly data), particular days in the week (weekday data), time of month (monthly data), and time of year (seasonal changes). The visualizations, as well as their physical and statistical significance, are shown below.

The average and maximum visualizations were compared with one another to see which data would be more suited to explain changes in $PM_{2.5}$ concentrations. The data was examined in multiple ways to lead to the best warning system possible. In many of the trends shown below, it is clear that maximum $PM_{2.5}$ trends are better for developing more accurate early warning systems as average values undermine the severity of pollutant levels at certain time trends.

Visualizations by year

Annual trends of $PM_{2.5}$ data between March 2013 and February 2017 were taken to view any important observations that may be important for analysis.

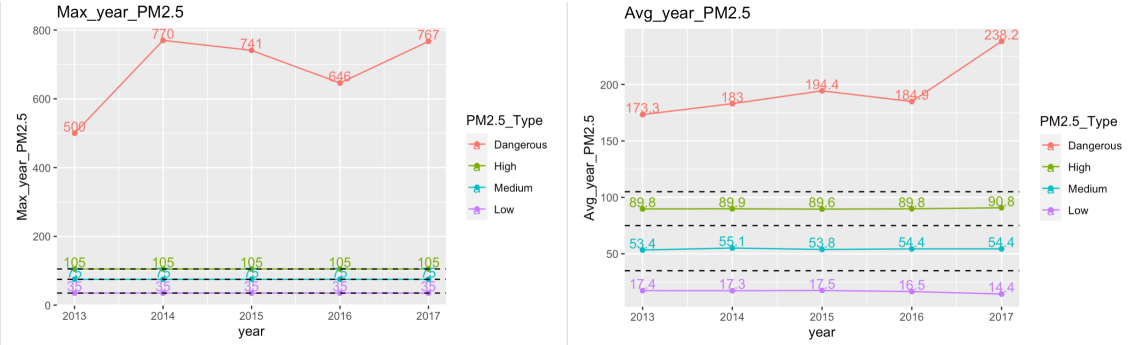

Figure 1: Visualizations of maximum (left) and average (right) PM2.5 concentration levels trend graph with respect to year

A set of visualizations were produced based on the different time-series datasets. Based on 'year', the highest maximum $PM_{2.5}$ concentration has been detected in 2014, and yet 2017 had the highest average $PM_{2.5}$ concentration. Considering that the dataset contains from March 1, 2013 to February 28, 2017, meaning that 2017 has only January and February, one possible naive

assumption would be that the winter months are more likely to have higher $PM_{2.5}$ concentration than the summer months. In other words, the $PM_{2.5}$ concentration would go up as the temperature goes down, indicating a negative correlation between them.

Since the data set was taken between the years of 2013 and 2017, it was deemed important to analyze the trends between these years. We will focus on the dangerous category of $PM_{2.5}$ concentrations since these are relatively easier to notice then the other three levels. From the yearly visualization data call mom we can see that there are different trends between the maximum yearly $PM_{2.5}$ data versus the average yearly $PM_{2.5}$ data. It seems that maximum values hit in 2014, when concentrations hit close to 800. However, the average values show a different story, since the highest concentration seems to have occurred in 2017. This distinction is due to the dangerous levels in comparison with other levels (i.e., high, medium, and low).

Visualizations by month

As for the trends of $PM_{2.5}$ levels from monthly data, dangerous levels will most likely occur in December and January with some extreme values. From the table of the average concentration of $PM_{2.5}$ for each month, it's clear that December will suffer the most from the pollutant. And in the winter, from October to March, people tend to suffer from the much worse condition of $PM_{2.5}$.

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Max. $PM_{2.5}$	767	770	458	533	337	500	375	276	311	468	546	741
Avg. $PM_{2.5}$	204.2	207.7	192.9	153.1	148.5	159.2	149.4	133.1	155.5	207.2	200.8	228.7

Table 2: Table of maximum and average $PM_{2.5}$ concentration of 'Dangerous' level with respect to month

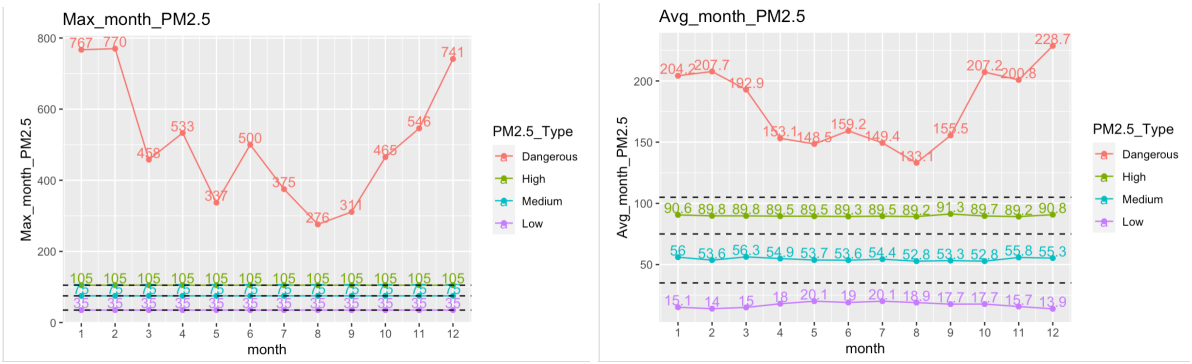

Figure 2: Visualizations of maximum (left) and average (right) $PM_{2.5}$ concentration levels trend graph with respect to month

Above are the table and visualizations of maximum and average $PM_{2.5}$ concentration based on months, and the naive intuition made in the year section was somewhat correct. Although there are some spikes going up in April and June, the overall trend line starts to decrease from March to August (Spring to Summer) and gradually increases from September to December (Fall to Winter). These increases may be due to higher usage of heating systems from residential, private, and public sectors, which may spread pollutants that travel at lower levels of the atmosphere higher. Such reasons strengthen the assumption that the $PM_{2.5}$ concentration is more likely to be higher in the winter months compared to the summer months.

To conclude the monthly visualizations, we can look at the following figure.

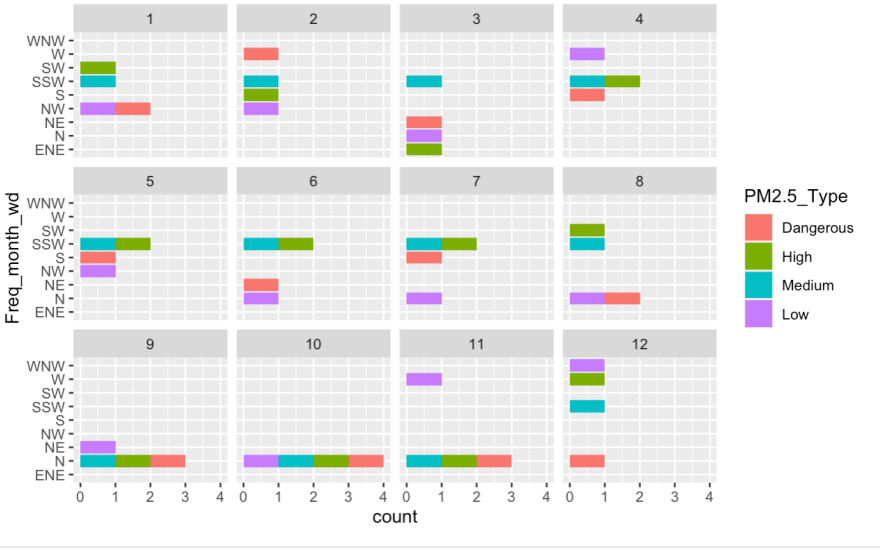

Figure 3: The frequency of wind direction based on monthly data

It seems as though wind direction is not consistent between the months of the year. Overall, there seem to be quite a few bars in the north direction. However, this is not consistent enough for us to make any conclusions based on the wind direction. Therefore, we will only consider the $PM_{2.5}$ data in our results.

Visualizations by weekday

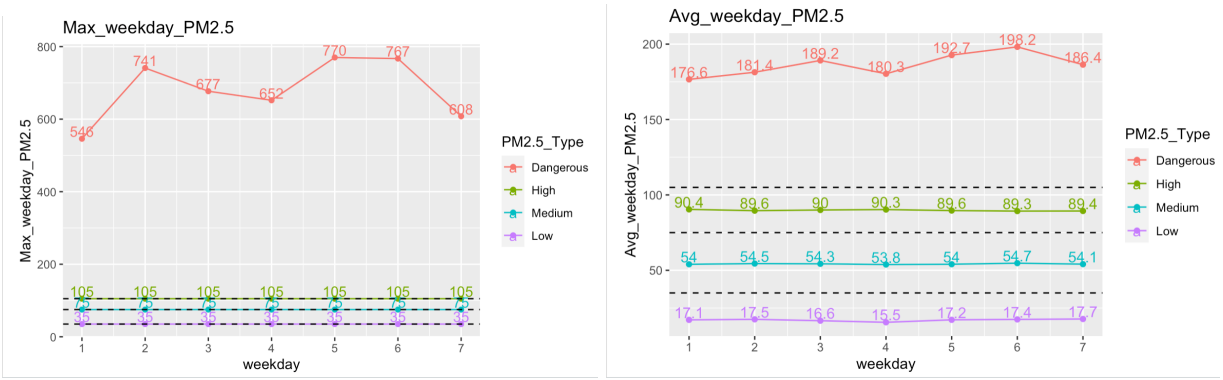

Figure 4: Visualizations of maximum (left) and average (right) PM2.5 concentration levels trend graph with respect to weekday
(1=Monday, 2=Tuesday, 3=Wednesday, 4=Thursday, 5=Friday, 6=Saturday, 7=Sunday)

For weekdays, both the maximum and average PM2.5 concentration trend graph seem to increase on Thursday and reach their highest peak on Friday and Saturday. Starting from Sunday, the graphs decrease, but soon climb slowly again and reach the second highest around Tuesday and Wednesday. From these visualizations, some possible assumptions would be that the PM2.5 concentration is higher on Friday due to rush hours and nighttime activities after working hours and more family-based outdoor activities on Saturday. Moreover, the heightened concentrations on Tuesday and Wednesday may stem from the fact that many have to travel to work. Many companies offer 3-day weekends, which may explain the decrease on Monday (for example), but an increase in traveling for work may impact $PM_{2.5}$ levels due to higher petroleum emissions. Therefore, it would be advisable to inform the public to take precaution on the weekends especially, as well as rush hours on weekdays.

To conclude the weekday change, we can look at the following figure:

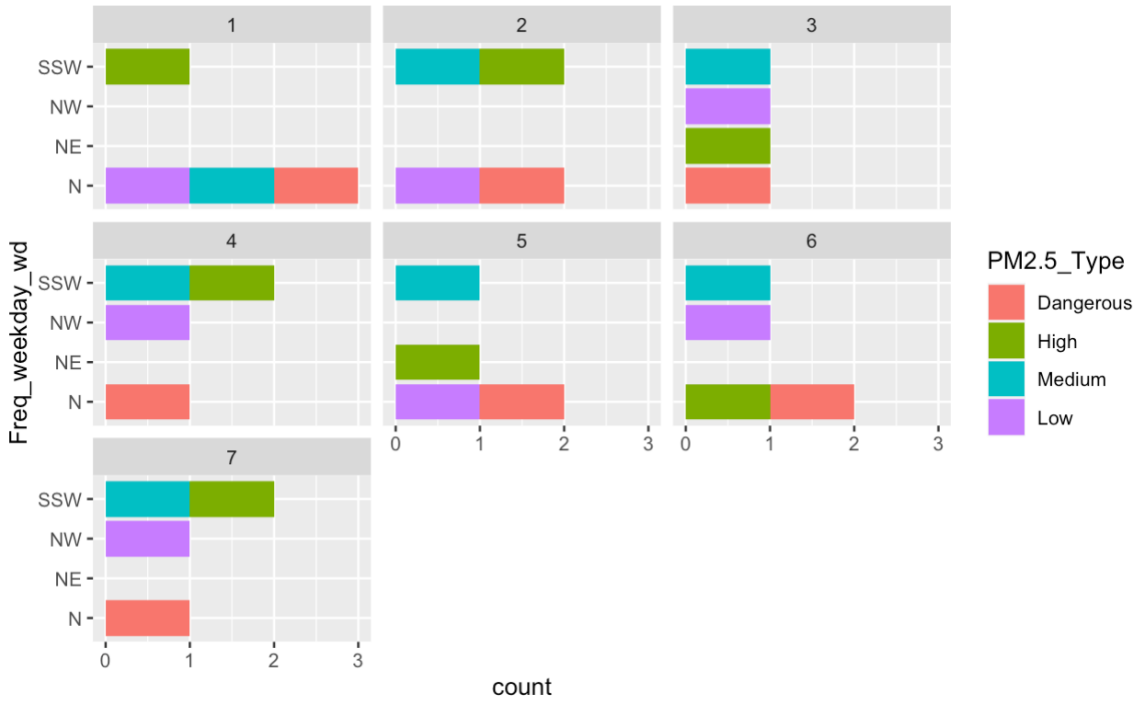

Figure 5: The frequency of wind direction based on Weekday data
(1=Monday, 2=Tuesday, 3=Wednesday, 4=Thursday, 5=Friday, 6=Saturday, 7=Sunday)

Looking at the frequency of wind direction based on the days of the week, it seems as though there are consistently dangerous levels in the north direction. Overall, there seem to be quite a few bars in the north direction. We want to be especially aware of the “dangerous” level, and the north direction has a lot of dangerous counts in particular.

Visualizations by hour

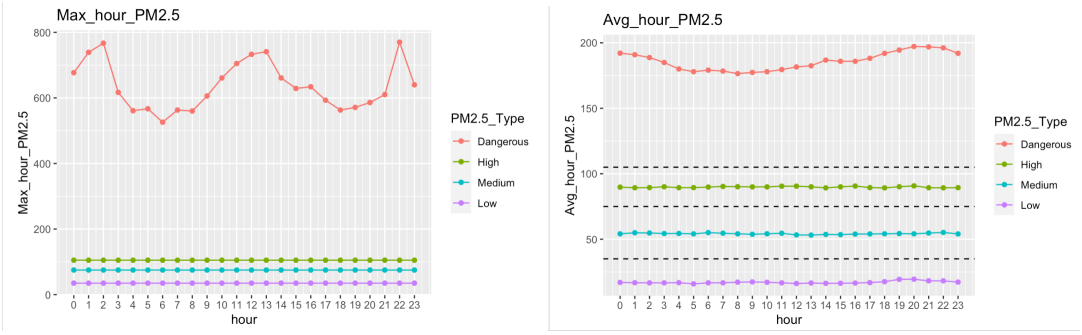

Figure 6: Visualization of maximum (left) and average (right) PM2.5 concentration levels trend graph with respect to hour

Lastly, the $PM_{2.5}$ concentration graphs reveal an interesting relationship with hour time. Considering some possible factors that contribute to the $PM_{2.5}$ level such as rush hours in the early morning (7-9am) and evening hours (5-7pm), the concentrations during these time periods show relatively lower levels than the afternoon (12-2pm) or nighttime (9pm-2am) hours. Although we do see a spike right before midnight,and around 2am, where the maximum levels reach around 800, there is a decline before the morning rush hour peaks. This may be a clue that the $PM_{2.5}$ level does not increase right away, but instead take a few hours to be concentrated and measured. Therefore, any outdoor activity around the afternoon and nighttime hours may not be recommended.

It is important to point out the significant difference in $PM_{2.5}$ behavior based on the maximum vs. average data. When we look at the average data, there does not seem to be significant fluctuations depending on the time of day; the levels seem to be consistent at around 175. From the differences between the maximum and average values, we can definitively state that using maximum data is more beneficial to developing a warning system for the public. The average data may severely undermine $PM_{2.5}$ levels at certain times of day, which can be harmful to those with pre-existing health conditions.

Visualizations by relevant air pollutants and meteorological features

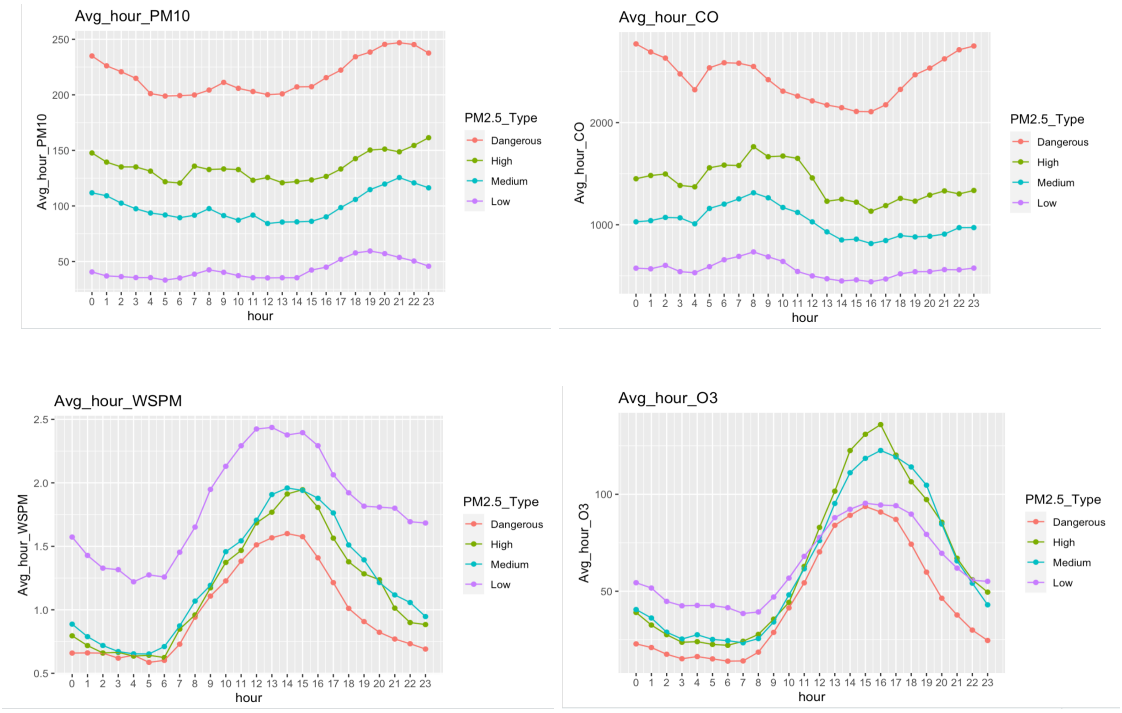

Figure 7: Visualizations of some air pollutants and meteorological trait with similar or opposite trend lines to the average $PM_{2.5}$ concentration

Beside time-series analyses, there are several other air pollutants that display almost identical trend lines to the average $PM_{2.5}$ such as PM_{10} and CO. This might indicate that these variables may have strong positive correlations with the $PM_{2.5}$ concentrations, which means that the $PM_{2.5}$ would increase as the PM_{10} and CO increase. On the other hand, WSPM (wind speed) and O3 have opposite spikes, meaning that the $PM_{2.5}$ concentrations would decrease when WSPM and O3 increase.

To sum up, some significant trends of $PM_{2.5}$ concentrations based on different time-series were detected. There were noticeable ups in the winter months and downs in the summer months. Also, Friday and Saturday were more likely to have higher concentrations, and this may be due to relatively more outdoor activities compared to other weekdays. Moreover, the nighttime and afternoon seemed to show higher spikes, therefore a suggestion of refraining outdoor activities during these time periods may be helpful in terms of protecting people's health. Lastly, some air pollutants and meteorological features such as PM_{10} and CO follow similar trends to $PM_{2.5}$ concentrations while WSPM and O3 reveal opposite trends, and these substances may be carefully monitored since this might be a signal of strong correlations with $PM_{2.5}$ and could potentially affect the modeling part later.

Correlations between different variables

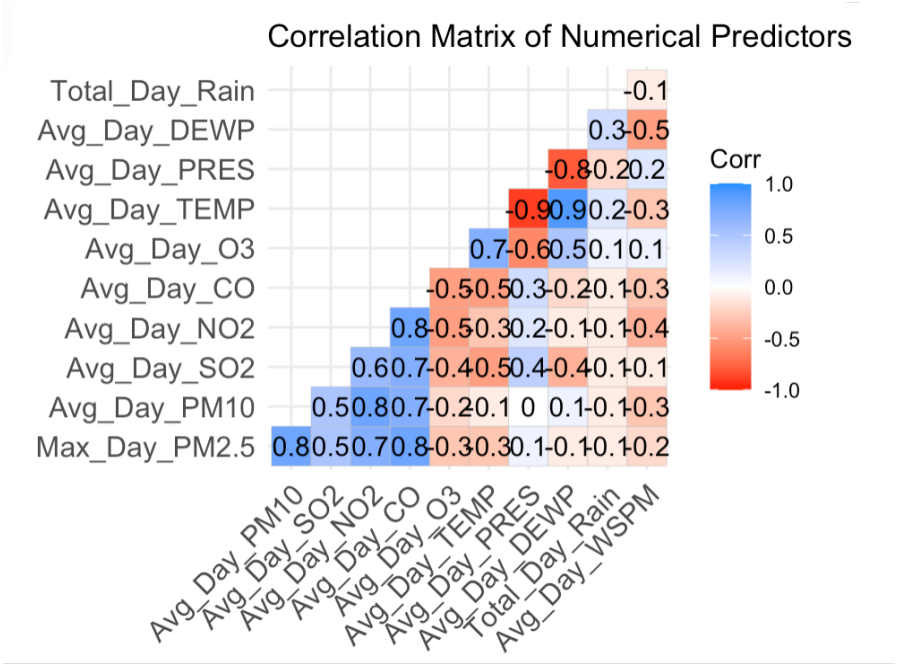

Figure 8: Correlation Coefficient Matrix of Numerical Predictors of Initial Model

From Figure 8, some variables seemed to be highly correlated with $PM_{2.5}$, such as PM_{10}, NO_2, CO. Also, there may be some variables highly correlated with each other, such as the PM_{10} and NO_2, suggesting that there may be some multicollinearity issue when fitting the linear regression model. But since we planned to use the variable selection procedures, we would still be exploring the models from the saturated one.

$PM_{2.5}$ Models: Regression and Machine Learning techniques

Model Selection Methods

We have come up with five different models with the performance as in the following table. Transformed, Lasso and Ridge Models were developed to predict the numeric outcome, which was the actual number of concentration of $PM_{2.5}$, therefore they could have evaluations like R^2 and adjusted R^2 . We would also use the threshold for $PM_{2.5}$ to divide the concentration of $PM_{2.5}$ into four different categories, Dangerous, High, Medium and Low, and further compare their performance with outcomes from the multinomial model and the KNN model with k to be 7 (see Appendix D).

Model	R^2	Adjusted R^2	Level Prediction Accuracy	Underpredicted Error
Transformed	0.285	0.269	171/283 = 0.604	17
Lasso	0.365	0.347	178/283 = 0.627	19
Ridge	0.362	0.339	177/283 = 0.623	19
Multinomial	X	X	183/283 = 0.644	27
KNN	X	X	177/283 = 0.623	27

Table 3: Comparison of Performance among Models

Based on the first criterion as mentioned above related to the whole accuracy, the best model selected should be the multinomial model, with the total accuracy to be 0.644. Lasso, Ridge and KNN would perform similarly, with the accuracy to be around 0.62. Even the transformed model would still have the accuracy to be around 0.6.

As for the second criteria which was related to the underpredicted error, as shown in Table 3, the transformed model would have the lowest number of this type of error. Table 4 and Table 5 further indicated the performance of the multinomial model and the transformed model

in predicting different levels of $PM_{2.5}$. The columns of the table would state the actual situation of $PM_{2.5}$, while the rows of the table would state the predicted situation of $PM_{2.5}$.

Predicted \ Actual	Dangerous	High	Medium	Low
Dangerous	156	29	24	13
High	0	0	0	0
Medium	10	13	26	7
Low	1	3	1	1

Table 4: Predicted vs Actual Labels Table of Multinomial Model

Predicted \ Actual	Dangerous	High	Medium	Low
Dangerous	140	19	16	10
High	19	16	20	0
Medium	8	9	15	7
Low	0	0	0	0

Table 5: Predicted vs Actual Labels Table of Transformed Model

The multinomial model would have a higher total accuracy in predicting different levels, as compared with the transformed model, except the category of High. It seemed that the multinomial model would not predict the high level of the concentration of $PM_{2.5}$. Furthermore, based on the second criteria mentioned above, the multinomial model would have 27 underpredicted errors, while the transformed model would only have 17 underpredicted errors. Under the situation with underpredicted errors, people would not be alerted for the harmful situation of high concentration of $PM_{2.5}$ and their health may be harmed due to this kind of predicting error, while for the opposite situation, the overpredicted situation, would not harm people's health, but just making them be more careful, which was acceptable.

Following the discussion above, together with another reason that the multinomial model could not predict the high level of the concentration of $PM_{2.5}$, we recommended the transformed model with the parameter of Box-Cox as 0.4 to be the final best model, with the corresponding parameters shown in Table 6.

	Coefficients	P - Value
Max_Day_$PM_{2.5}$	0.0142	<2e-16 ***
Avg_Day_SO_2	-0.0398	1.30e-05 ***
Avg_Day_NO_2	0.0650	<2e-16 ***
Avg_Day_O_3	0.0420	<2e-16 ***
Avg_Day_TEMP	-0.202	<2e-16 ***
Total_Day_Rain	-0.093	2.27e-07 ***
Avg_Day_WSPM	-1.436	4.54e-15 ***

Table 6: Estimated Coefficients and p-values in the Transformed Model

From the coefficients and the p-values of all the predictors, it was clear that all the predictors, the maximum value of $PM_{2.5}$ in the previous day, the average value of SO_2, NO_2, O_3, temperature and wind speed, together with the total amount of rain in the previous day would be significant in predicting the values of $PM_{2.5}$ in the future. Based on the coefficients of the predictors, maximum value of $PM_{2.5}$, average value of NO_2 and O_3, in the previous day had a positive correlation with $PM_{2.5}$ in the future, while the average value of SO_2, temperature and wind speed in the previous day, as well as the total amount of rain in the previous day would have a negative correlation with $PM_{2.5}$ in the future.

Conclusions and Discussion

To conclude, there were many observations and predictions that could be made from the visualizations shown and the models fitted.

First, looking at the visualizations, we can see from the monthly trends (which also indicate any seasonal changes or impacts of $PM_{2.5}$) that the $PM_{2.5}$ levels increase in the winter months for example the highest maximum $PM_{2.5}$ and average $PM_{2.5}$ values occur around the months of December through February. These results may be because of the heightened use of heating systems in homes, public facilities, and workplaces. Moreover, ridges from high pressure systems accompanied by relatively strong winds which bring in cooler air and carry pollutants with it, may cause pollutants to remain close to populated source locations. Particulate matter may be spread from vehicle exhausts, but it only rises high enough through higher wind speeds.

Therefore, we suggest that people are careful of their surroundings when they step outside during the winter months.

Moreover, we can note the increase in $PM_{2.5}$ data depending on the weekday. As explained before, $PM_{2.5}$ concentrations increase during the weekend time, possibly due to gatherings and an increase in travel, whether that be via air or personal vehicles. Therefore, we would suggest the public to be careful during the weekend times and suggest the use of masks for people especially with respiratory issues.

Lastly, the hourly data can provide insight into times of day that people should travel or be cautious of higher concentrations of $PM_{2.5}$ period. The graphs of the hourly data show an increase in $PM_{2.5}$ levels around the early morning around 6:00 AM evening around 5:00 PM and night time. Therefore, people should be cognizant of their health, especially those with pre-existing or respiratory conditions, around times of peak rush hour and nighttime (around 11pm). This may be a clue that the $PM_{2.5}$ level does not increase right away, but instead take a few hours to be concentrated and measured. Therefore, any outdoor activity around the afternoon and nighttime hours may not be recommended.

For the second part of the study, we have developed five different models, the transformed, Lasso, Ridge, Multinomial and KNN model. From the overall model, we can predict tomorrow's $PM_{2.5}$ values based on today's data, which allows for projections to ensure caution and warnings to the public. As mentioned before, we have also developed two different criteria to compare the performance of the five different models. For the first criteria related to the overall accuracy rate, the best model selected should be the multinomial, while for the second criteria of underpredicted errors, the best model should be the transformed model with much fewer underpredicted errors, as compared with the multinomial model. Together with the reason that the multinomial model could not predict the category of high concentration of $PM_{2.5}$, the best model selected and recommended should be the transformed model from the linear regression model via Box-Cox transformation with lambda to be 0.4.

For further interpretation from the transformed model, higher concentrations of $PM_{2.5}$, NO_2, O_3 in the present will result in higher concentrations of $PM_{2.5}$ in the future. Lower SO_2, temperature, rain and wind speed will also result in higher future $PM_{2.5}$ concentrations.

Overall, the information in this report provides a significant warning system of $PM_{2.5}$ levels. However, to improve the analysis, a couple of aspects can be considered. First of all, although the times of day and days of the week show the times when concentrations of pollutants are the highest, some other physical considerations are where the pollutants are coming from (i.e., facilities or certain behaviors that are the main sources of pollutants). Moreover, a dataset with less missing values will improve the overall accuracy of the model.

Appendices

Appendix A: Extended version of all variables

Variable name	Description
No.	Row number
year	year of data in this row
month	month of data in this row
day	day of data in this row
hour	hour of data in this row
PM2.5	PM2.5 concentration (ug/m^3)
PM10	PM10 concentration (ug/m^3)
SO2	SO2 concentration (ug/m^3)
NO2	NO2 concentration (ug/m^3)
CO	CO concentration (ug/m^3)
O3	O3 concentration (ug/m^3)
TEMP	Temperature (degree Celsius)
PRES	Pressure (hPa)
DEWP	dew point temperature (degree Celsius)
RAIN	precipitation (mm)
wd	wind direction
WSPM	wind speed (m/s)
station	name of the air-quality monitoring site

Table 1: Descriptions of All the Variables that are Presented in the Dataset

Appendix B: Diagnosis Plots in Exploration of Linear Models

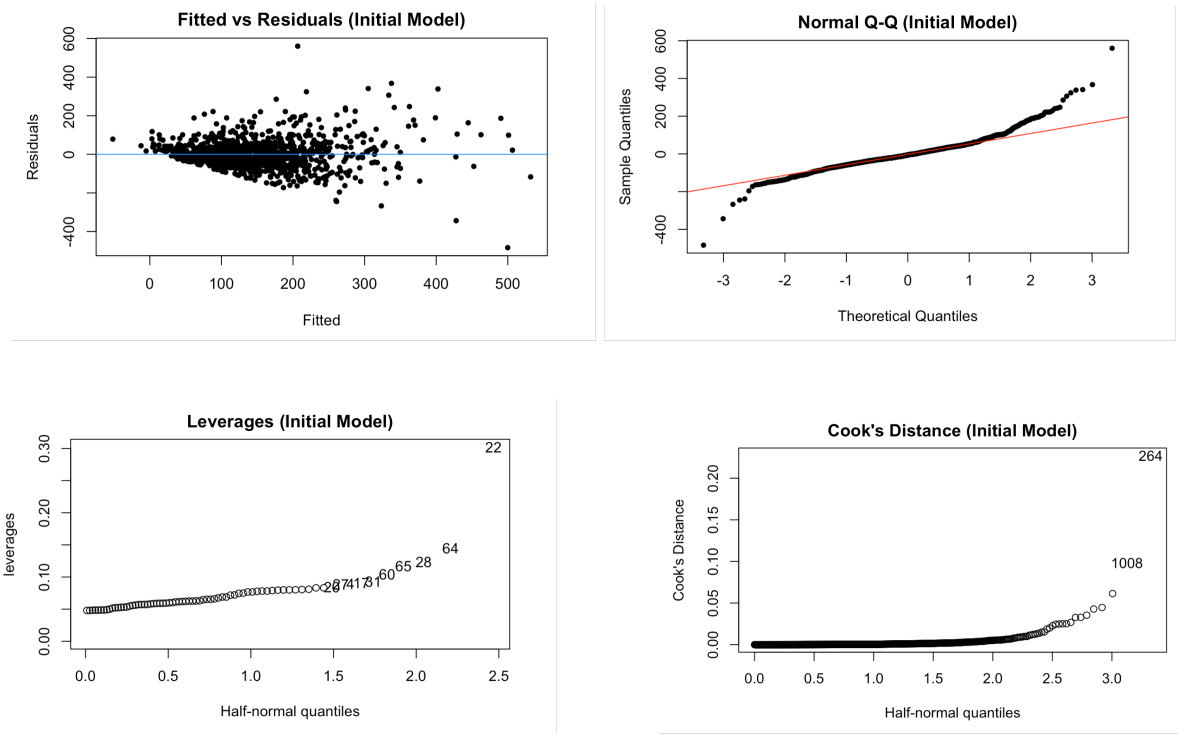

Figure 1: Diagnosis Plots of the Saturated Model

Test Type	Test Value (p-value)
Breusch-Pagan (Constant Variance)	228.19 (< 2.2e-16)
Shapiro-Wilk (Normality)	0.92 (< 2.2e-16)
Kolmogorov-Smirnov (Normality)	0.52 (< 2.2e-16)
Leverages (Maximum)	0.30
Cook's Distance (Maximum)	0.23

Table 2: Some Tests tried for the Assumptions

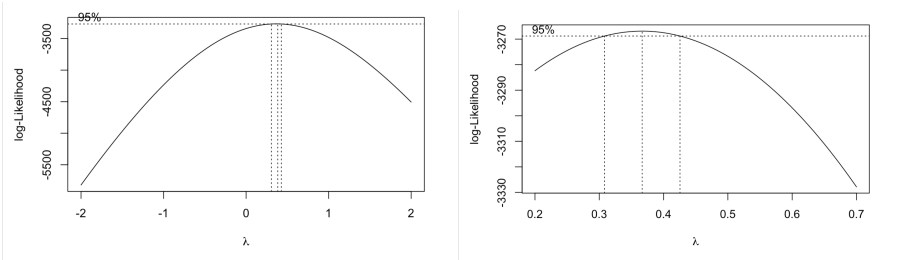

Figure 3: Plots to Select the Value of Lambda in the Box-Cox Transformation to Fulfill the Assumptions

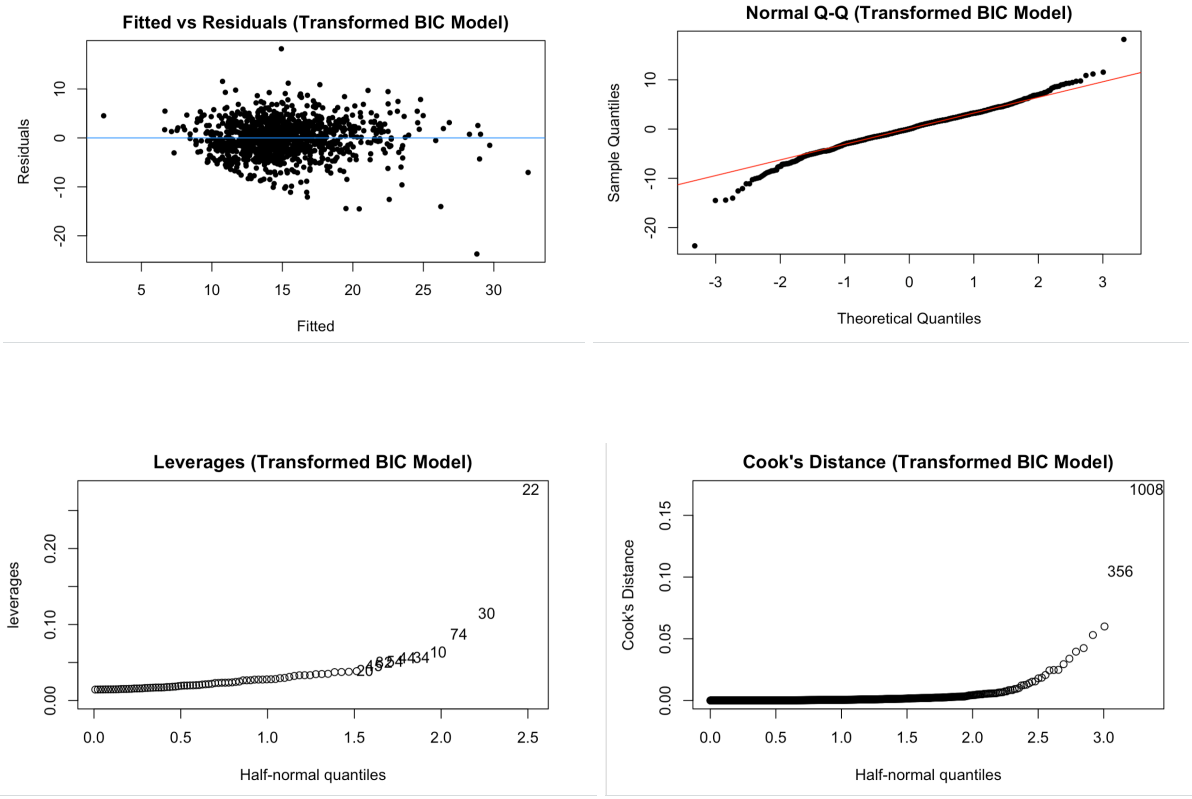

Figure 4: Diagnosis Plots of the Final Model

Test Type	Test Value (p-value)
Breusch-Pagan (Constant Variance)	103.55 (< 2.2e-16)
Shapiro-Wilk (Normality)	0.98 (< 8.1e-13)
Kolmogorov-Smirnov (Normality)	0.28 (< 2.2e-16)
Leverages (Maximum)	0.28
Cook's Distance (Maximum)	0.17

Table 3: Some Tests tried for the Assumptions (Final Model)

Appendix C: Cross Validation of Lasso and Ridge Model

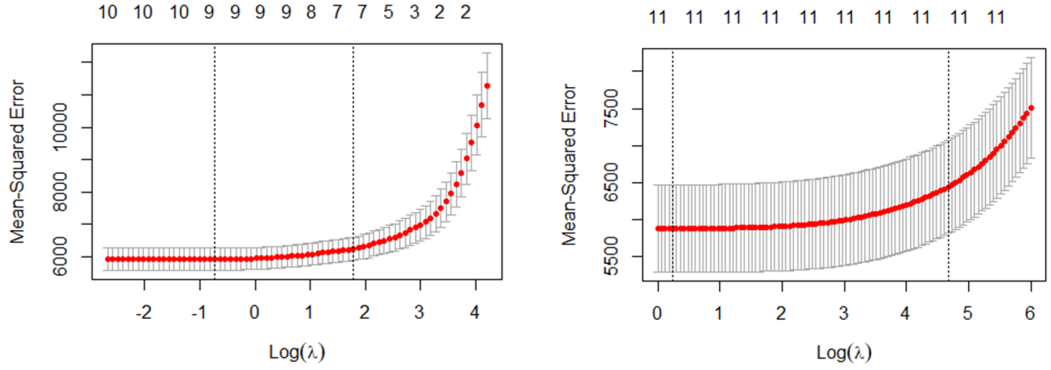

Figure 5: Corresponding Cross Validation Plots for Lasso and Ridge Regression

	Lasso	Ridge
Max_Day_PM2.5	0.3641061	3.597283e-01
Avg_Day_PM10	0.1297935	1.452975e-01
Avg_Day_SO2	-0.6011327	-6.854923e-01
Avg_Day_NO2	0.8876815	9.184284e-01
Avg_Day_CO	-	-5.341606e-04
Avg_Day_O3	0.5703099	6.065010e-01
Avg_Day_TEMP	-2.7589021	-2.798371e+00
Avg_Day_PRES	1.1100001	1.172401e+00
Avg_Day_DEWP	-	-1.073800e-01
Total_Day_Rain	-1.2873017	-1.302772e+00
Avg_Day_WSPM	-23.2653461	-2.461248e+01

Table 4 : Coefficients for Lasso and Ridge Regression

Appendix D: Accuracy in Predicting Different Levels of KNN model with K from 1 to 30

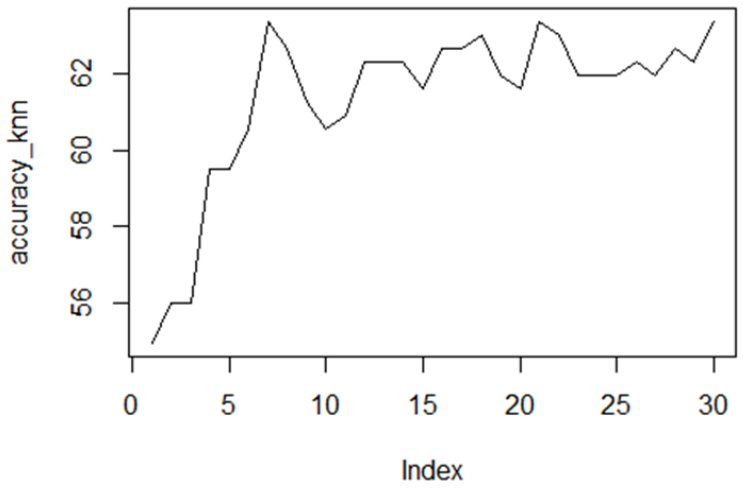

Figure 6: Accuracy of KNN model with K from 1 to 30

Code and Data

The supporting code is provided in the accompanying STAT 443 Project.rmd file.

Contributions

- Junseok Yang: Mainly focused on writing the overall baseline codes, updating the codes based on other group members' feedbacks, and providing general explanation and interpretation of both the visualizations and modeling results of the .Rmd file for other group members to articulate the report.
- Lavanya Upadhyaya: Overall → Set up initial Github file and communication. Mid-project and final presentation → introduction, visualization, and correlation analysis. Some model analysis for next steps. Report → project abstract, introduction/data description, methodology, visualization and correlation analysis, diagnostics (RMD file), administered model analysis, conclusions and next steps.
- Mengjia Zeng: Contribute to the visualization part, build the Lasso, Ridge, multinomial and KNN models in predicting, mainly focus on writing the codes in the rmd file. Contribute to the Mid Check-in slides, final presentation slides and write the methods and regression analysis parts of the final report.
- Wasay Siddiqui: Set up Git Flow on team member devices to enable proper collaboration between code versions. Responsible for different sections within both of the mid-project as well as final report. Provided input on presentation slides prior to submission for modification as well as write-ups . Wrote methodology sections across reports, as well as overviewing the final versions of both for clarity and summarization of results.

Acknowledgements

We would like to thank Professor Darren Glosemeyer and Theren Williams for their help and support throughout this process!

References

[1] Zhang, S., Guo, B., Dong, A., He, J., Xu, Z. and Chen, S.X. (2017) Cautionary Tales on Air-Quality Improvement in Beijing. Proceedings of the Royal Society A, Volume 473, No. 2205, Pages 20170457

[2] UCI Machine Learning Repository: Beijing Multi-Site air-quality data data set. (2017.). Retrieved March 27, 2022, from https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data#