Forecasting PM2.5 for Wanshouxigong District

By Group 8

Alex Funches

John Fetscher

Riya Dave

Yirui Luo

Yuxuan Zhang

Table of Contents

Introduction

As industrialization continues to expand, automobile usage continuously grows which is negatively impacting the environment than ever before. The emission of pollutants like carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2), both from automobiles and other products & industries, only exacerbates the problem. This tradeoff of economic growth and pollution is problematic especially in terms of PM2.5 particulate matter emissions because of its microscopic particles which can infiltrate the respiratory system and have associated long term effects like reduced lung function, increased rates of chronic bronchitis, and increased mortality from lung cancer and heart disease. PM2.5 is widespread and carried through the burning of fuels, but also through inside activities like cooking, burning candles, and heaters.

The objective of this consulting project is to create a model that can accurately predict the daily levels of particulate matter with a diameter less than 2.5 micrometers (PM2.5) and produce a user interface alert system that will allow its user to input values for the dependent variables and receive a numeric response for the PM2.5 prediction and indicate if it is at a safe level. The Beijing Multi-Site Air-Quality dataset was extracted from the UC Irvine Machine Learning Repository, it is a multivariate time series dataset that has taken data from different sources like its air-quality from the Beijing Municipal Environmental Monitoring Center and meteorological data from the China Meteorological Administration. The dataset we are provided has 35,064 observations and 18 variables that describe the pollutant gas levels in the Wanshouxigong station. Each observation uses the observation number, weather station, and time related variables like month, day, and year as reference points to organize and separate the recorded observations.

We are interested in building a model that can best predict the PM2.5 levels of the Wanshouxigong district in Beijing, China. We will determine the variables that most affect the PM 2.5 level on any given day. The dataset has 11 continuous variables, 6 of these variables describe the pollutant gases and their chemical makeup (these are *PM2.5, PM10, SO2, NO2, CO,* and *O3*). The 5 remaining continuous variables describe the daily weather conditions (these are the temperature (TEMP), pressure (PRES), the dew point (DEWP), rain, and wind speed (WSPM)). The 5 discrete variables are used to reference the time of occurrence and observation number (these are *No, year, month, day,* and *hour*). All of the variables were initially used to predict our response variable of PM2.5 under World Health Organization's daily guidelines that

a value of PM2.5≤75 is acceptable and a PM2.5≥150 is unacceptable. Combining the cutoffs given by our client and the World Health Organization we are using the following cutoffs to warn the public about PM 2.5 levels: PM 2.5 that is between 0-35 ug/m^3 are highly acceptable levels, 35-75 ug/m^3 are slightly acceptable levels, 75-150 ug/m^3 is slightly unacceptable levels and 150+ ug/m^3 are highly unacceptable levels of PM 2.5.

Variables	**Properties**	**Function**
No	discrete(**Identification)**	row number
Year	discrete	year of data in this row
Month	discrete	month of data in this row
Day	discrete	day of data in this row
Hour	discrete	hour of data in this row
PM2.5	continuous (**Response**)	PM2.5 concentration (ug/m^3)
PM10	continuous	PM10 concentration (ug/m^3)
SO2	continuous	SO2 concentration (ug/m^3)
NO2	continuous	NO2 concentration (ug/m^3)
CO	continuous	CO concentration (ug/m^3)
O3	continuous	O3 concentration (ug/m^3)
TEMP	continuous	temperature (degree Celsius)
PRES	continuous	pressure (hPa)
DEWP	continuous	dew point temperature (degree Celsius)
RAIN	continuous	precipitation (mm)
wd	categorical	wind direction
WSPM	continuous	wind speed (m/s)
Station	categorical	name of the air-quality monitoring site

Exploratory Data Analysis

Variable Analysis

The assignment revolves around the Beijing Wanshouxigong Air Quality dataset. The dataset includes 16 continuous variables, which are Year, Month, Day, Hour, SO2, NO2, CO, O3, PM2.5, PM10, Temperature, Pressure, Dew point temperature, precipitation, Wind Speed and Wind direction.

Figure 1[1] shows a general trend of P.M 2.5 levels across the four years of data we were provided. From this, we can conclude that PM 2.5 levels have stayed relatively consistent throughout the years. It can also be seen that every year has the same kind of parabolic trend which convinced us to explore PM 2.5 levels throughout the months of a year.

Figure 2 is a histogram that shows the amount of data points we have in every cutoff level. Since the histogram is skewed right, we can see that there are more acceptable levels than unacceptable levels in the data provided for this district. From this, we can conclude that overall, the PM 2.5 levels in the Beijing Wanshouxigong District are acceptable.

Figure 3 shows PM 2.5 values across the months of a year. While initially there seemed to be no pattern, after doing some research about the Beijing Wanshouxigong district, we split the months into different seasons. From this we can gather that PM 2.5 levels are generally lower in the warmer months (in spring and summer) and rise to higher levels as it gets cold (in fall and winter). Based on our research as to what causes PM 2.5, these trends make sense because as it gets colder, people are burning more coal and wood to generate heat in homes and other buildings which, along with other factors, increases the overall PM 2.5 levels in that time period. Similarly, as it gets hotter in the summer, the needs to generate heat lessen which also help in decreasing the overall PM 2.5 levels. In a more basic sense this cold weather remains trapped and lets pollutants accumulate unlike hot weather causing the air to rise.

Figure 4 is a histogram depicting the PM 2.5 levels throughout the days of a month. At initial glance at this visual, there seems to be no significant relationship between PM 2.5 and day.

However, if we look at PM 2.5 levels throughout the week (as seen in figure 5), we can conclude that there is a general trend of increasing levels of 2.5 from the start of the week to the

[1] All figures can be referenced in the appendix

end of the week. We can see that generally weekdays have lower PM 2.5 levels than the weekend (starting with Friday). Combining this with the research done about the sources of PM 2.5, we can see that during the weekend, it is more likely that people are at home spending time with family or outside cooking which plays a role in increasing PM 2.5 levels in the area. Since people are usually home on the weekend, there are a lot more buildings burning fuels for cooking or lights than there are when people are working during the weekday (since many people congregate in office buildings and work).

Figure 6 describes PM 2.5 levels throughout the hours in a day. From this, we can see that overall, PM 2.5 levels tend to decrease during the middle of the day and rise as the day transitions into the evening and nighttime. Combining this visual with the research done for the sources of PM 2.5, we can reason that since people are working during the middle of the day, they are congregated at work places and at nighttime, most people are in their individual homes which is an increase in buildings that need heating and energy to make food. This could be a factor that explains these trends of increases and decreases of PM 2.5 levels during the day. Additionally, this middle of the daytime is when it is usually hotter and illustrates the same trend we saw across seasons with PM2.5 being lower in hotter weather.

Methods

Data Cleaning & Validation

In starting this assignment, we first determined whether the dataset contained any errors or missing data. Upon looking through the entries of each variable, it was clear that some observations contained missing data. The following variables contained observations with missing values: PM2.5, PM10, SO2, NO2, CO, O3, TEMP, PRES, DEWP, RAIN, and WSPM. In addressing the issue, the average was taken for each variable, with the resulting average being placed in for the missing values. This enabled us to have observations with data for each variable measured in the dataset.

In addition to accounting for missing values, new variables were created to aid in developing a more accurate model. For indicating seasonal trends, a categorical variable was created for the four seasons present in the Wanshouxigong region. As shown in the PM2.5 vs Month visual (see Figure 3), PM2.5 levels tend to vary throughout seasons, with winter having the highest PM2.5 values and summer having the lowest. The seasonal variable will allow the model to change its prediction of PM2.5 depending on the season the prediction is being forecasted in. Another variable created for improving the predictive abilities of the model was a weekday variable. As seen in the PM2.5 vs Weekday visual (see Figure 5), the levels of PM2.5 change throughout the week. As mentioned above, PM2.5 levels tend to be lower on weekdays and higher on weekends. The addition of the weekday variable enables the model to account for these differences. Finally, the last variable built for improving the model was a previous day's PM2.5 variable. With wanting to predict the PM2.5 level for the next day, taking into consideration the PM 2.5 levels for the day before gives the model more indication as to what the next day's PM2.5 level will be. The inclusion of the previous level variable gives the model more accurate information as to the relative conditions of the environment in the previous day, as PM2.5 conditions usually do not drastically change day-to-day.

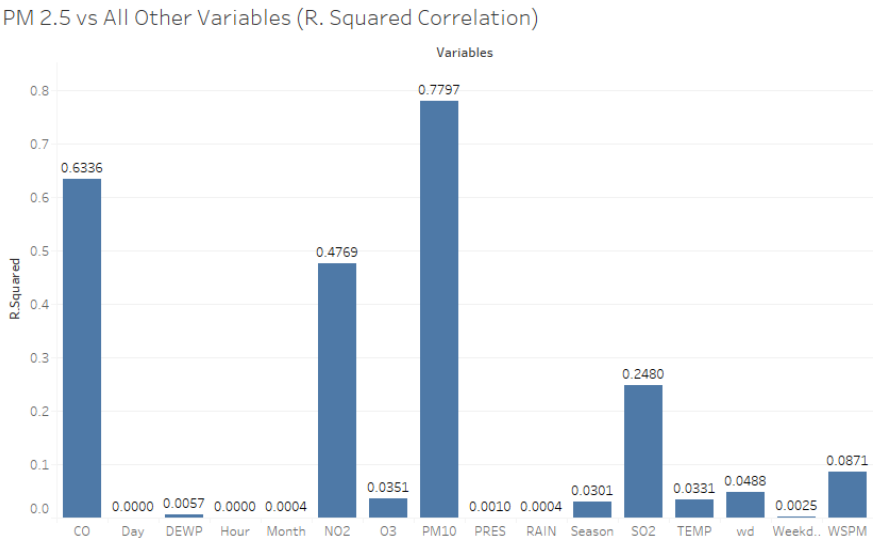
To ease the process of time series graphs as well as visualizing seasonal trends, a timestamp variable was created. Using Excel, the timestamp variable was built out through a concatenation of the values in the year, month, day, and hour variables. Another variable created to simplify data visualization processes was a PM2.5 level variable, a categorical variable indicating the air quality using WHO's PM2.5 guidelines. The categorical variable is split into 4 categories, with under 35 ug/m^3 being highly acceptable, 35 through 75 ug/m^3 being slightly

acceptable, 75 through 150 ug/m^3 being slightly unacceptable, and over 150 ug/m^3 being highly unacceptable.

In order to make our alert system easier for clients to use, we grouped by days and got the daily average dataset as the input to our model so that there is no need to input previous 24 hours weather data in order to predict next day's PM2.5. As the prediction model aims to predict the PM2.5 level for a future date, grouping information by days fully captures the data recorded for each day and lowers the standard error.

Correlation

First, we wanted to see the relationship between PM2.5 and other variables so that we could at least get some insights on which variables might have higher influence on the pollution level. As we've mentioned in the introduction, sulfates, nitrates and black carbon are main components of PM2.5, we would expect the toxic gas variables (CO, NO2, O3, SO2) to be closely related to pollution level. We performed simple linear regression of PM2.5 on each of the other 14 variables. Following figure shows the R.squared value of each simple linear regression model. The R.squared measures how closely the two variables are related. From the figure, we can observe that the PM10, carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2) and wind speed are more closely related to PM2.5 than other variables, which is identical to our expectation. Thus, in the following exploratory data analysis and model building process, we will specifically focus on these variables and see how these variables would affect the model performance.



Multicollinearity (VIF)

Variance inflation factors (VIF) measure the inflation in the variances of the parameter estimates due to collinearities that exist among the predictors. A VIF of 1 means that there is no correlation among a certain predictor and the remaining predictor variables, and hence the variance of the certain predictor is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

The following chart shows the VIF results. When diagnosing collinearity between the different variables in the data, we found that both Temperature and Dew point have high collinearity (VIF of 9.32 and 7.13 respectively). In addition, we found that pressure has a moderate collinearity (VIF of 4.82). It is important to look at collinearity because a linear regression assumes the features are independent. If we have collinearity in our data, this can result in an inaccurate linear regression model. Since one of our end goals is to come up with an alert system to tell the public about predictions of the pollution for the next day, we need to remove collinearity if we decide to use a linear regression model for our predictions.

Variables	**Tolerance**	**VIF**
year	0.7468841	1.338896
month	0.496485	2.014159
day	0.9917178	1.008351
hour	0.8474574	1.18
SO2	0.4941348	2.023739
NO2	0.3027969	3.302544
CO	0.3348848	2.986101
O3	0.3957578	2.526798
TEMP	0.1072277	9.325949
PRES	0.2074737	4.819888
DEWP	0.1402395	7.130657
RAIN	0.9720688	1.028734
WSPM	0.579531	1.725533
seasonSpring	0.3352375	2.98296
seasonSummer	0.3345464	2.989123

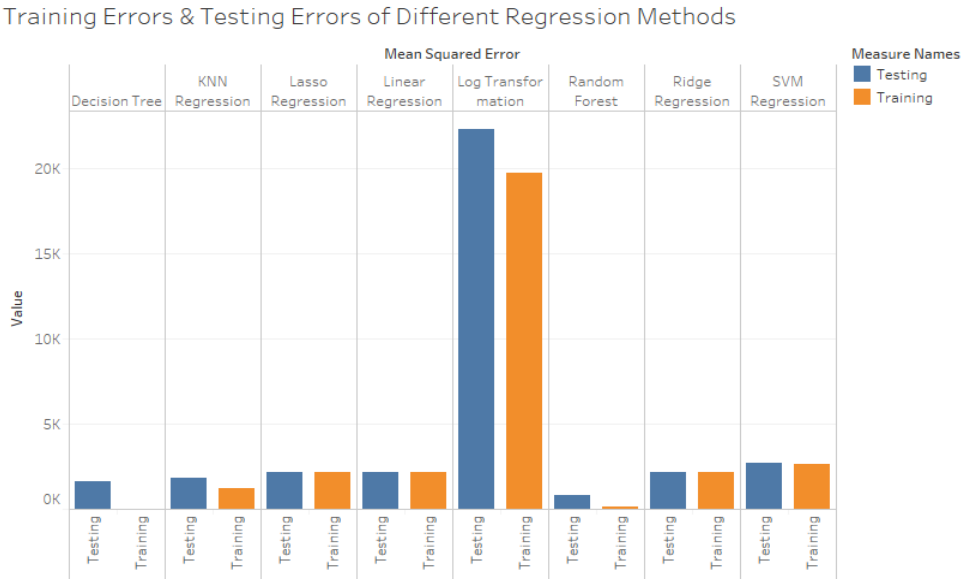
seasonWinter	0.2733753	3.657976

Models

We built seven regression models to figure out which regression methods work best for our data Decision Tree, KNN Regression, Lasso Regression, Linear Regression, Log Transformation, Random Forest, Ridge Regression, and SVM Regression. In order to prevent overfitting issues and minimize the effects of data discrepancy, we split the data to training and testing by 70/30. To measure the model performance, we use Mean Squared Errors (MSE) as a metric to estimate the training errors and testing errors in each model. After our empirical analysis, we compared the performance of each model and produced a "Model Efficacy Comparison" bar chart to visualize the model performance more intuitively.

Model Efficacy Comparison Table and Bar Chart

Models	Training Errors	Testing Errors
Linear Regression	2151.24	2137.84
Ridge Regression	2151.24	2137.83
Lasso Regression	2151.23	2137.83
Log Transformation	19772.7	22281
Decision Tree	0	1621.5
Random Forest	**104**	**779.69**
KNN Regression	1218.11	1821.07
SVM Regression	2649.41	2706.91



From the chart and figure above, random forest and decision trees performed best in all seven models. Since the testing error of decision trees is more than twice as much as random forest, we decided to use random forest as an optimal regression model. In the next section, we will present some results derived from the random forest model, including the level of importance of each variable. Although the random forest model has high prediction accuracy, it's non-parametric and might sacrifice some interpretability. We considered the interpretability of a model the same as important as accuracy, so we decided to build a linear regression model with higher interpretability next. In addition, random forest is a complex model that cannot fit in the dashboard we plan to build using Excel.

In order to build a linear regression, we need to factorize features such as month and weekday whose value does not have an actual meaning. We dummy code those variables in a way using 1 and 0 to indicate whether the observation belongs to that class. In the modeling, we remove variables, such as year, day, hour, wind direction, which are insignificant to our model. We found PM2.5 from yesterday is referential because there is high autocorrelation in PM2.5 series. After adding it into the model, R-squared is improved greatly. Finally, we need to standardize the data because some features have a large range and skewed distribution.

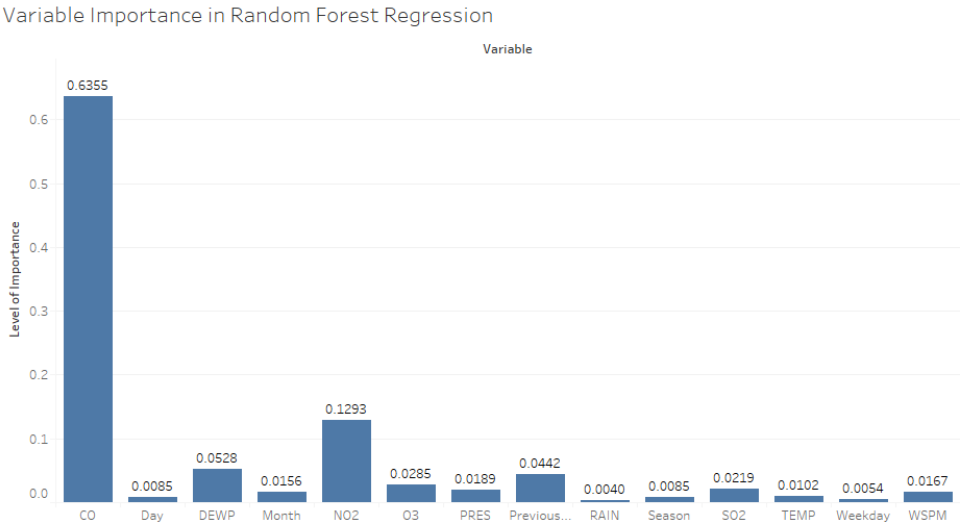
After dummy coding, the number of features we have now exceeds 30. Therefore, we employed Lasso regression which adds a penalty of coefficients to the cost function. It can guarantee accuracy not only in training but also in testing. Another advantage of Lasso regression is that it selects features automatically and will shrinkage coefficients to zero. In R, there is a high level function *cv.glmnet* can do the cross validation for us to prevent overfitting and choose the best parameter to minimize the cost function. We chose the mean squared errors as the metric to evaluate our models. After fitting a Lasso regression, the metric is lowered ten times than the linear regression model. However, the estimate from Lasso regression is no longer unbiased, so no p-value is produced and the interpretation from coefficients is limited. As Lasso shrinks the size of the coefficients the closer to 0 the coefficient of a variable is the less impact it has on affecting the PM2.5 level.

Results

In this section we will present some results and interpretations derived from different regression methods. We wanted to figure out which variables might have higher influence on the PM2.5 and pay more attention to the model performance in prediction.

Variable Importance in Random Forest Regression

We built the random forest regression model and found out some important variables that might highly influence the pollution level in Wanshouxigong district. There is a built-in function in Python to check the level of importance of each variable in the model. From the following figure, we can see the carbon monoxide (CO), nitrogen dioxide (NO2), dew point temperature (DEWP) and previous day's PM2.5 (Previous.PM2.5) have the highest influence on the pollution level, which is identical to our expectation. Except for these variables, the ozone (O3), sulfur dioxide (SO2), WSPM (wind speed), and TEMP (temperature) also have moderate influence on the pollution.



Variable Importance in Random Forest

LASSO Regression

The graph below shows the comparison of true value and predicted PM2.5 over four years. The cutoff value is also listed in the graph. We can see that there is not much difference between the two lines. The mean of squared errors is 447.1605 in the Lasso regression model. The second graph is the confusion matrix visualization for PM2.5 levels. The colors represent the level of PM2.5 according to the combination of the WHO standard and guidelines of our client

and the size represents accuracy. As the prediction transition from green to red the worse the air pollution becomes with green representing a highly acceptable level with a PM2.5 between 0-35 and red representing a highly unacceptable, hazardous level with a PM2.5 greater than 150. This model can correctly predict tomorrow's expected average PM2.5 at an accuracy above 80%. There is a low probability that we predict the correct PM2.5 level to the neighboring level. This is also true when predicting the acceptability, there is a very low probability that an unacceptable PM2.5 level is predicted as acceptable at 7.57% and an acceptable PM2.5 level being predicted as unacceptable at 19.7%. Besides, it is unlikely that a large deviation will happen in our model. Also, this higher prediction of an acceptable PM2.5 being unacceptable can lead to more caution and care when considering daily activity.

Conclusions

Based on our statistical analysis we found that PM2.5 levels increase when values for the following variables increase: Previous PM2.5, PM10, CO, NO2, O3, Pressure, and Dew Point. On the other hand, PM2.5 levels decrease when values for the following variables increase: SO2, Temperature, Rain, and Wind Speed.

Based on the data we can see that when it is colder or when more people are at home, PM 2.5 levels tend to rise. This can be explained by looking back at the causes of PM 2.5: burning fuels for cooking and heating, power plants and motor vehicles. In the colder months, more fuels are burned for heating and over the weekend people are traveling more and spending time with family which means more cooking and traveling. Due to this, we advise that the general public wear masks when out during the cooler months as well as the weekend if the reduction of outdoor activity is not possible. Also as seen by the PM2.5 reduction towards the middle of the day we advise people plan accordingly and if possible, conduct their outside activity during this time frame of the afternoon and evening. However, since our model fits the weights of the variables automatically, these trends are not seen within the coefficients for each determining variable.

User Interface

By implementing a user interface as a component of the alert system, we can give an interactive, quantifiable response for the predicted daily level of PM2.5. Using a more common and accessible program like Google Sheets to host the interface allows our client to easily make inputs and receive a numeric response based on the contained date, pollutant, and weather variables. The program itself is simplistic and appropriate for the general public and Google Sheets is something most people are familiarized and experienced with unlike R's shiny program which was an alternative we considered.

The user interface is built using our lasso regression model. Standing at a training error of 2,151.23 and a testing error of 2,137.83 the lasso model did not have the lowest error among our models. However, unlike random forest or decision tree models, this model has greater interpretability and makes a user interface where the client can input values. In building models and determining the one best suited for an interface accuracy and interpretability were both valued. The interface itself uses 13 variables: PM2.5(Today), PM10, SO2, NO2, CO, O3, Temp, Pressure, Dew Point, Rain, Wind Speed, Month, and Weekday. Utilizing two sheets to create a

dashboard the inputted values are applied accordingly to their respective coefficients, with conditional formatting indicating the range the PM2.5 level falls under. This gives the user a numeric response for the day's PM2.5 which is highlighted by color according to the safety condition of PM2.5 which they can consider and properly respond to. As stated previously in the colder months and day's when the PM2.5 is predicted as unacceptable with a highlighting of dark orange and red we recommend for people to limit their daily activity and dress appropriately by wearing masks like the N95.

Appendices

Description of Variables

Variables	**Properties**	**Function**
No	discrete(**Identification)**	row number
Year	discrete	year of data in this row
Month	discrete	month of data in this row
Day	discrete	day of data in this row
Hour	discrete	hour of data in this row
PM2.5	continuous (**Response**)	PM2.5 concentration (ug/m^3)
PM10	continuous	PM10 concentration (ug/m^3)
SO2	continuous	SO2 concentration (ug/m^3)
NO2	continuous	NO2 concentration (ug/m^3)
CO	continuous	CO concentration (ug/m^3)
O3	continuous	O3 concentration (ug/m^3)
TEMP	continuous	temperature (degree Celsius)
PRES	continuous	pressure (hPa)
DEWP	continuous	dew point temperature (degree Celsius)
RAIN	continuous	precipitation (mm)
wd	categorical	wind direction
WSPM	continuous	wind speed (m/s)
Station	categorical	name of the air-quality monitoring site

Figure 1: PM2.5 Timestamp

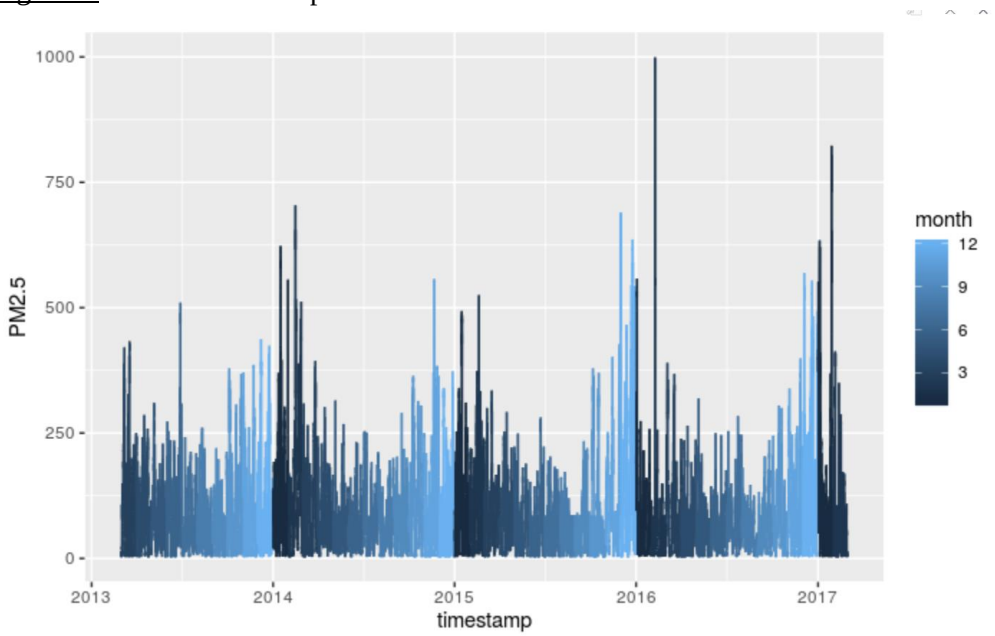


Figure 2: Histogram of PM2.5 by Cutoff Levels

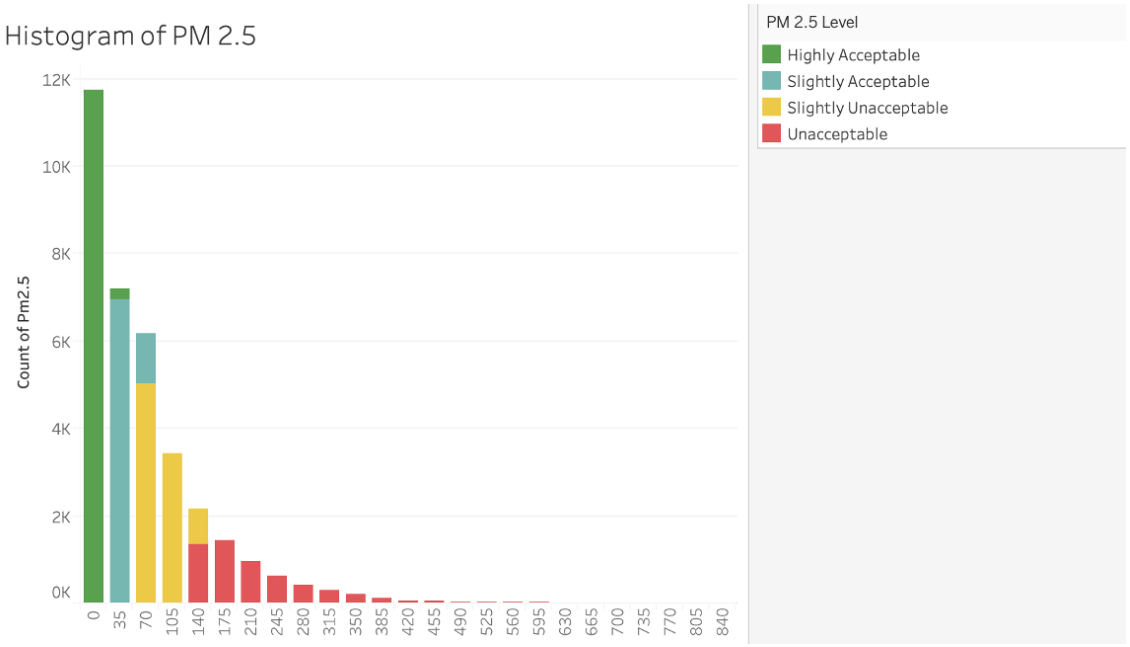


Figure 3: PM2.5 by Month

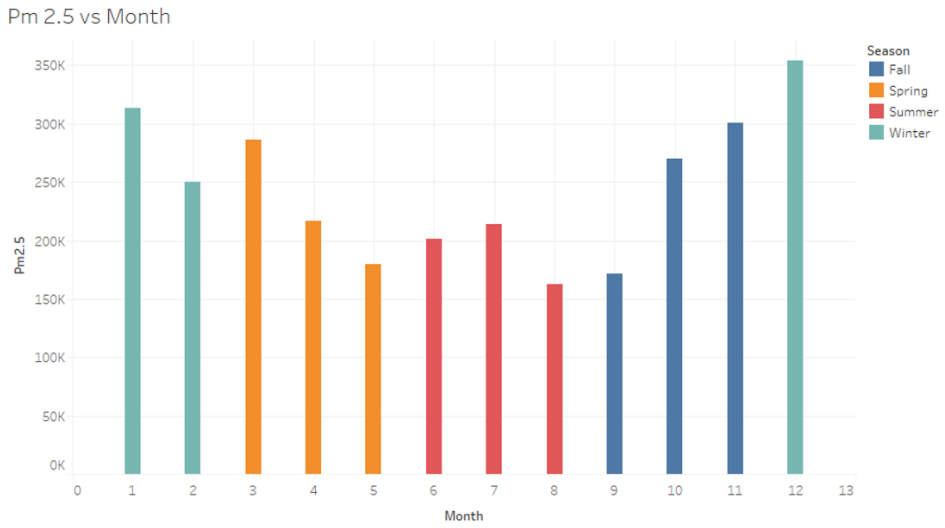


Figure 4: PM2.5 by Day

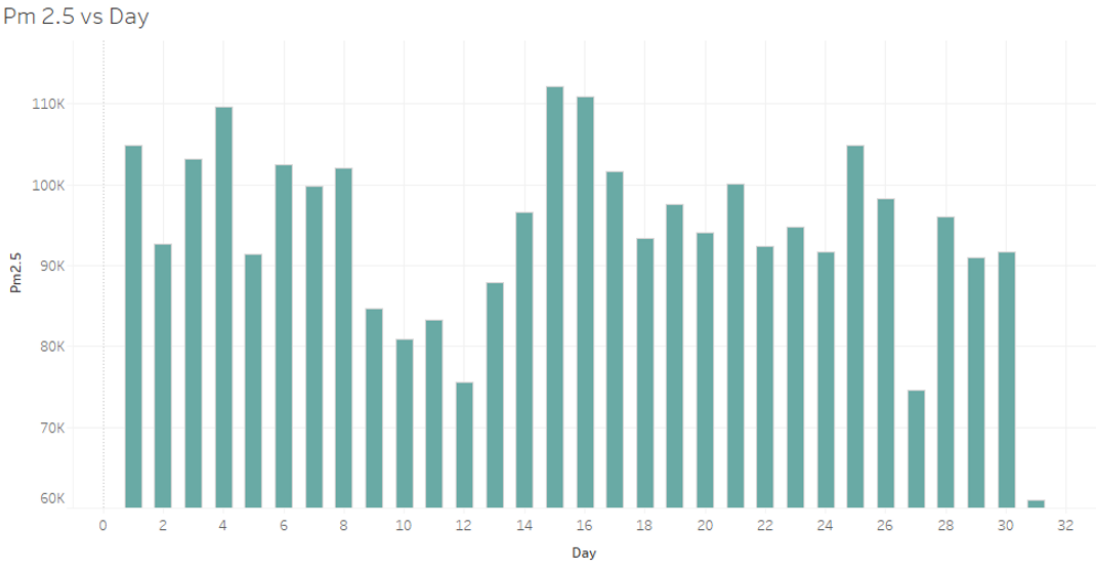


Figure 5: PM2.5 by Weekdays

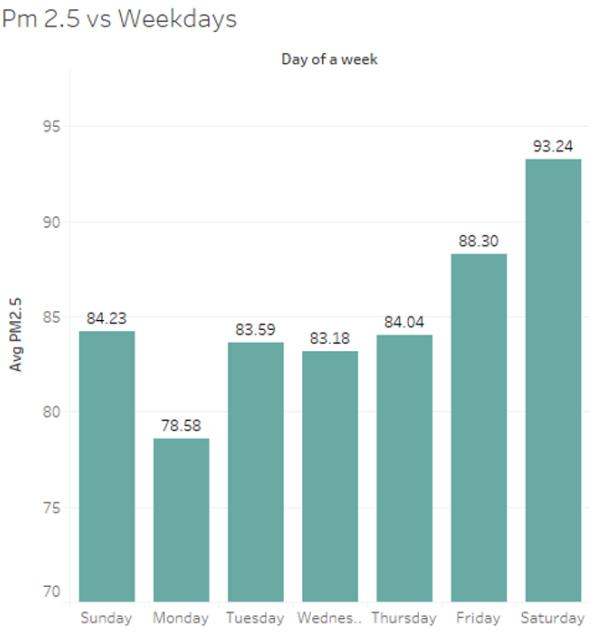


Figure 6: PM2.5 by Hour

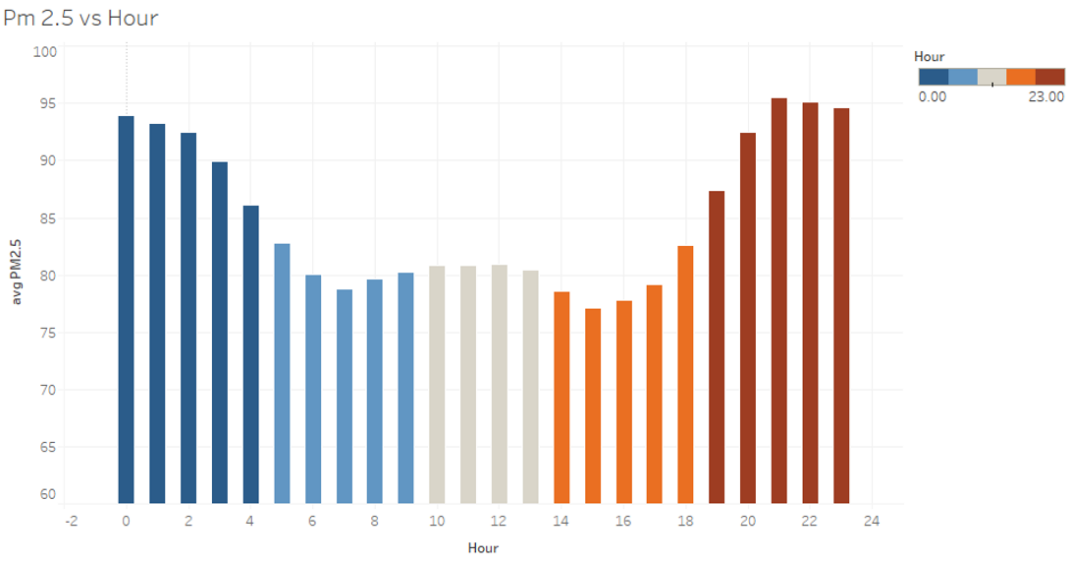


Figure7: Variable Correlation

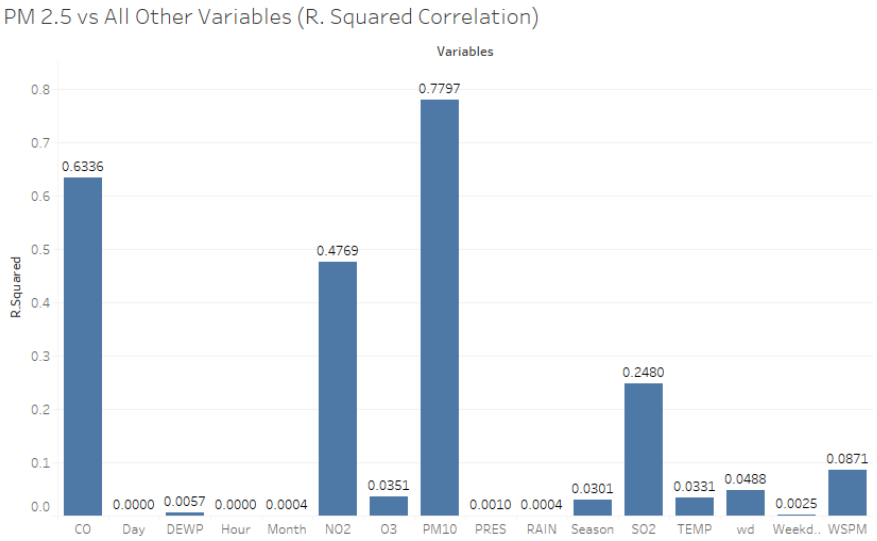


Figure8: Multicollinearity of Preliminary Model

Variables	**Tolerance**	**VIF**
year	0.7468841	1.338896
month	0.496485	2.014159
day	0.9917178	1.008351
hour	0.8474574	1.18
SO2	0.4941348	2.023739
NO2	0.3027969	3.302544
CO	0.3348848	2.986101
O3	0.3957578	2.526798
TEMP	0.1072277	9.325949
PRES	0.2074737	4.819888
DEWP	0.1402395	7.130657
RAIN	0.9720688	1.028734
WSPM	0.579531	1.725533
seasonSpring	0.3352375	2.98296
seasonSummer	0.3345464	2.989123
seasonWinter	0.2733753	3.657976

Figure9: Model Efficacy Comparison Table

Models	Training Errors	Testing Errors
Linear Regression	2151.24	2137.84
Ridge Regression	2151.24	2137.83
Lasso Regression	2151.23	2137.83
Log Transformation	19772.7	22281
Decision Tree	0	1621.5
Random Forest	**104**	**779.69**
KNN Regression	1218.11	1821.07
SVM Regression	2649.41	2706.91

Figure10: Model Mean Standard Error Comparison Table

Models	Mean Std Error
Linear Regression	445.6393
Lasso Regression	447.1605
Polynomial Model	423.0345
Polynomial Model with Interaction	284.449

Among these four models the polynomial model with interaction has the lowest mean standard error at 284.449, but it also has the least interpretability and is inappropriate for a user interface where the user can input values. This model also has more outliers and large leverage points than both the linear and polynomial model at 72 outliers and over 1,400 large leverage points.

Figure11: Model Efficacy Bar Chart

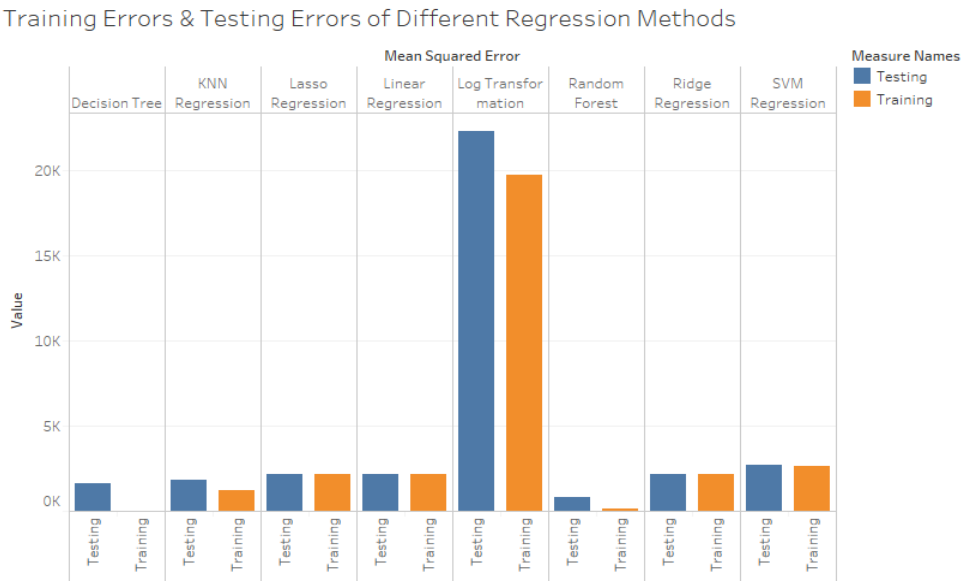


Figure12: Random Forest Variable Importance

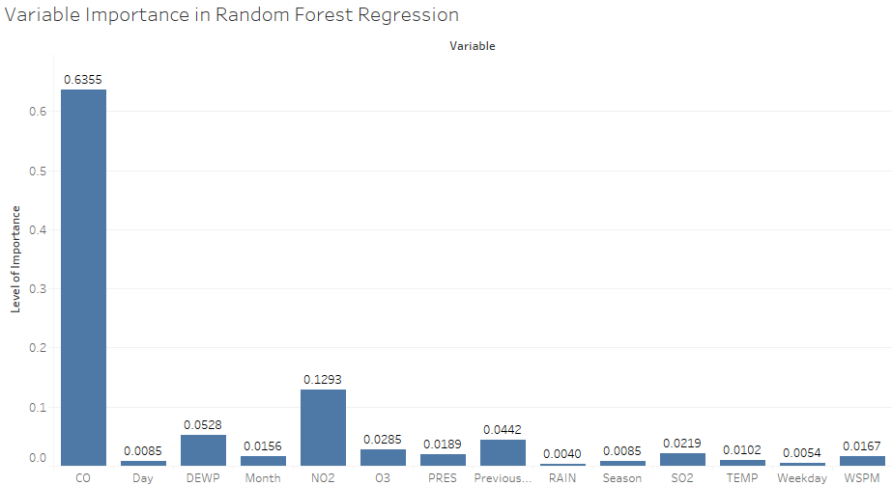


Figure13: Linear Model Parameter Estimates

Variables	**DF**	**Parameter Estimate**	**Standard Error**	**T Value**	**Pr(>|t|)**	**Variance Inflation**
Intercept	1440	-2.341e+02	1.296e+02	-1.877	0.06078	0
yesterday	1440	9.555e-02	9.509e-03	10.048	< 2e-16	1.562124
PM10	1440	5.817e-01	1.514e-02	38.416	< 2e-16	4.609223
SO2	1440	-1.188e-01	4.142e-02	-2.868	0.00419	2.374417
NO2	1440	6.678e-02	4.324e-02	1.544	0.12271	4.747168

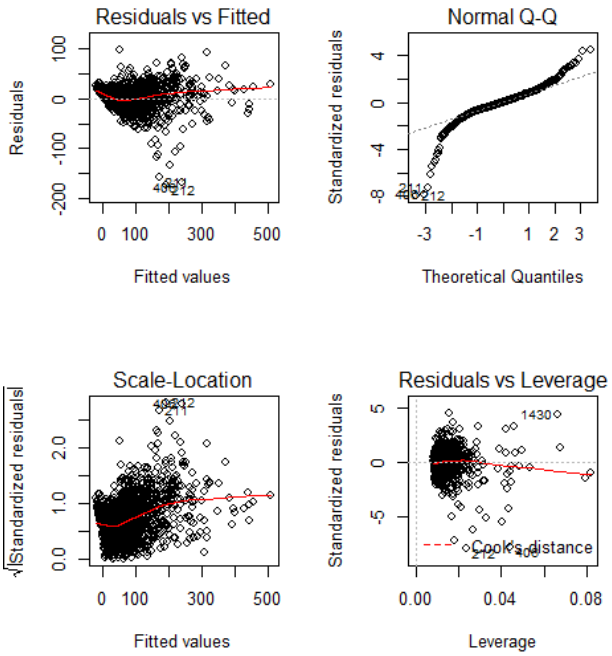
CO	1440	1.1794e-02	1.345e-03	13.346	< 2e-16	6.278567
O3	1440	1.086e-01	2.726e-02	3.984	7.11e-05	3.691925
TEMP	1440	-2.353e+00	2.491e-01	-9.445	< 2e-16	23.323716
PRES	1440	2.426e-01	1.268e-01	1.913	0.05594	5.569916
DEWP	1440	1.894e+00	1.406e-01	13.470	< 2e-16	11.494238
month2	1440	1.148e+01	2.853e+00	4.025	5.98e-05	1.870008
month3	1440	5.829e+00	3.287e+00	1.773	0.07643	2.702901
month4	1440	9.085e+00	4.246e+00	2.140	0.03256	4.377153
month5	1440	4.730e+00	5.197e+00	0.910	0.36296	6.756100
month6	1440	9.229e+00	5.978e+00	1.544	0.12284	8.675589
month7	1440	1.206e+01	6.505e+00	1.854	0.06391	10.583093
month8	1440	3.004e+00	6.434e+00	0.467	0.64063	10.355715
month9	1440	-2.631e+00	5.545e+00	-0.474	0.63526	7.463693
month10	1440	5.293e+00	4.358e+00	1.215	0.22475	4.751445
month11	1440	-2.562e+00	3.227e+00	-0.794	0.42731	2.528034
month12	1440	-5.212e+00	2.785e+00	-1.871	0.06148	1.940056

All of the variables included in the model are statistically significant at a 5% level of significance with the exception of the month dummy variables and NO2. Removing NO2 decreases the adjusted R squared of the model from .9155 to .9154. Using the model selection methods of stepwise, backward, and forward none of them result in the removal of any variables from the original linear model and lists the model's Akaike Information Criterion (AIC) as 8,958.36. As determined by selection methods this is the linear model that results in the lowest AIC and does not feature the variables of RAIN, weekday, and WSPM.

Before running predictions and checking model accuracy, we examined the diagnostics of the model to understand any potential issues and make further adjustments. Using the variance inflation function in R we conclude that there are four variables with a VIF greater than 10 indicating that there are issues of multicollinearity. In the original dataset and linear model all of the same variables are included with the exception of yesterday's PM2.5 and the highest VIF is TEMP at 7.95. After the inclusion of variables like yesterday's PM2.5 and the creation of dummy variables through the factorization of month the variables of DEWP, TEMP, month7,

and month8 become collinear. This model was a starting point and not chosen due to the issues of multicollinearity, heteroscedasticity, and non-normality. Additionally, it is not the model that results in the lowest training and testing errors(see Figure 11).

Figure14: Diagnostics for Linear Model



The Residuals vs Fitted plot appears to be positively, linearly related at higher fitted values and heteroscedastic residuals. The Normal Q-Q plot has points deviating from the line which would lead us to reject the assumption of normality. Scanning the Residuals vs Leverage plot there are no influential points. Under further scrutiny using the bptest this also leads us to reject the assumption of normality as the p value is less than 2.2e-16. There are also 60 outliers and over 1,400 large leverage points using a standard cutoff of 2*(number of columns in dataset)/(number of rows). This indicates that there are observations that diverge greatly from the overall pattern in the dataset and a lot of observations have a large effect on the regression slope and interpretation.