ANALYSIS OF AIR POLLUTION LEVELS AND ITS CONTRIBUTING FACTORS

1. ABSTRACT

High concentrations of particulate matter ($PM_{2.5}$) and frequent air pollution episodes in Beijing have attracted widespread attention. This paper utilizes data from the UCI Machine learning repository which provides the hourly data set which contains the PM2.5 data of US Embassy in Beijing and also the meteorological data from Beijing Capital International Airport. By learning the PM2.5 readings and meteorological records from 2010–2015, the severity of PM2.5 pollution in Beijing is quantified with a set of statistical measures. This paper shows the step by step analysis of the influence of various meteorological factors on the PM2.5 concentration, which can be used to monitor PM2.5 pollution in a location. Just like how weather forecast helps to know if it is going to rain today and if one has to take an umbrella while going out, same way a pollution forecast can help a person to decide if he/she needs to go out with a mask or not. Through some assumptions made based on domain literature research, in this paper, we predict the pm2.5 values using various machine learning methods and are testing the validity of the assumptions.

2. INTRODUCTION

Due to rapid economic growth, industrialization and urbanization, China has experienced severe air pollution problems, and Beijing, the capital, political and cultural center of China, is among the most polluted cities in the country. Air pollution has been recognized as a major concern in China. Concentrations of particulate matter with an aerodynamic diameter of 2.5 μm or smaller (PM2.5) in Beijing have attracted global attention due to the high levels, as well as its associations with adverse effects on human and ecological health. Due to high levels of pollution, orange haze alert was declared in 2014. An orange alert is issued when the average air quality index is forecast to exceed 200 for three consecutive days and one of those day is forecast to be over 300.

Our analysis uses hourly PM2.5 readings taken at the US Embassy in Beijing located at (116.47 E, 39.95 N), in conjunction with hourly meteorological measurements at Beijing Capital International Airport (BCIA), obtained from UCI Machine Learning Repository. Both data series run from 1 January 2010 to 31 December 2014. Although the embassy and the airport are 17km apart, they experience almost the same weather. The US Embassy started to announce hourly PM2.5 readings from April 2008 at a different location. We did not use the data in 2008 and 2009 due to a large number of missing values and the embassy moving to its current location in 2009.

3. THE DATA

The data used is from the UCI Machine learning repository. It is the hourly meteorological data of Beijing for the period of 2010 – 2014. The data is a multivariate time series data with 43824 instances and 13 different attributes.

PM2.5 is taken as the target variable and the features considered are Timestamp (year, month, day, hour), Dew Point, Temperature, Pressure, Combined wind direction, Cumulated wind speed, Cumulated hours of snow, Cumulated hours of rain.

A common way of judging air pollution is to look at the concentration of fine particulate matter (PM2.5) Particulates less than 2.5 micrometers in diameter (PM 2.5) are believed to pose the largest health risks. Compared with the coarser particles, PM2.5 is smaller in size, larger in surface area, and more easily transported, which implies more toxicity and harmful substances that can penetrate deep into the human body. PM2.5 can stay in the atmosphere for a long time and travel for a long distance. Therefore, it has a greater impact on human health and the quality of the atmospheric environment.

According to the US (EPA) standard, 35μg/m^3 (the European Union uses 25μg/m^3) is the highest PM2.5 level for acceptable air quality, while 150μg/m^3 is widely viewed as very unhealthy and even hazardous.

AQI Category	Index Values	Revised Breakpoints (μg/m^3, 24-hour average)
Good	0 - 50	0.0 – **12.0**
Moderate	51 - 100	12.1 – **35.4**
Unhealthy for Sensitive Groups	101 – 150	35.5 – **55.4**
Unhealthy	151 – 200	55.5 – 150.4
Very Unhealthy	201 – 300	150.5 – 250.4
Hazardous	301 – 400	250.5 – 350.4
Hazardous	401 – 500	350.5 – 500

4. METEROLOGICAL FACTORS

Various Meteorological factors can significantly affect PM2.5 mass concentration, which can help to reduce or aggravate the air pollution. The meteorological factors that are mainly related to PM2.5 concentrations in different cities also vary due to the differences in emission intensity and diffusion conditions of pollutants. It was found that the meteorological factors related to PM2.5 concentration during winter in Beijing are relative humidity, average daily temperature, average wind speed, wind direction and minimum temperature.

All the meteorological variables, except the wind direction, are continuous, the upper bound is infinity. The meteorological variables are mutually correlated.

It has been observed that wind tends to alleviate the air pollution, and hence stronger wind are always welcomed during the worst of the high PM2.5 episodes. Lack of wind has been frequently blamed for high $PM_{2.5}$ in Beijing, far more often than anthropogenic activities that contribute to the pollution. The weather data had 16 wind directions. A study shows that the directions can be grouped into five broad categories: northwest (NW), which includes W, WNW, NW, NNW and N; northeast(NE),for NNE, NE and ENE; southeast(SE), covering E, ESE, SE, SSE and S; southwest (SW), having SSW, SW and WSW; and calm and variable (CV). The decision to allocate E to SE and W to NW was based on the locations of major polluting industries around Beijing. It is said that a northerly wind in all seasons helps to significantly reduce the PM2.5 levels. In contrast, a southerly wind does not reduce pollution; rather it generally increases it, particularly in the summer.

Situated at the northwest corner of the NCP, Beijing is hemmed in by Taihang Mountain to the west and Yan Mountain to the north. The benefit of northerly wind is due to a lack of heavily polluting industry in the region north of Beijing. However, the mountains cause accumulation of the polluted air under a southerly wind. The south and the east of Beijing on the NCP are dense with heavy industries, which consume enormous amounts of coal and other fossil fuels. The annual coal consumption in the NCP was more than 1 billion tones in 2012, constituting 25% of China's and 15% of the world's consumption, in a densely populated region that accounts for only 5.6% of China's land area. Additionally, there are more than 5 million cars in Beijing, which also contribute to its air pollution.

There are other meteorological factors which also contribute in increasing or decreasing the PM2.5 levels. A decrease in the dew point & an increase in the pressure is usually accompanied by the arrival of the northerly wind, which brings in drier and fresher air. PM2.5 concentrations show a remarkable seasonal variability with the highest during the winter and the lowest during the summer. The effect of rainfall on the removal of particulate matter seems to be positive. The average PM2.5 concentration decreased by 56.3% following the rainfall, and PM2.5 mass concentration was less than 60 μg/m3 within 72 h after the rainfall. Within 1 h after the rainfall, the PM2.5 concentration level stayed almost unchanged, and it kept declining within the next 12 hours.

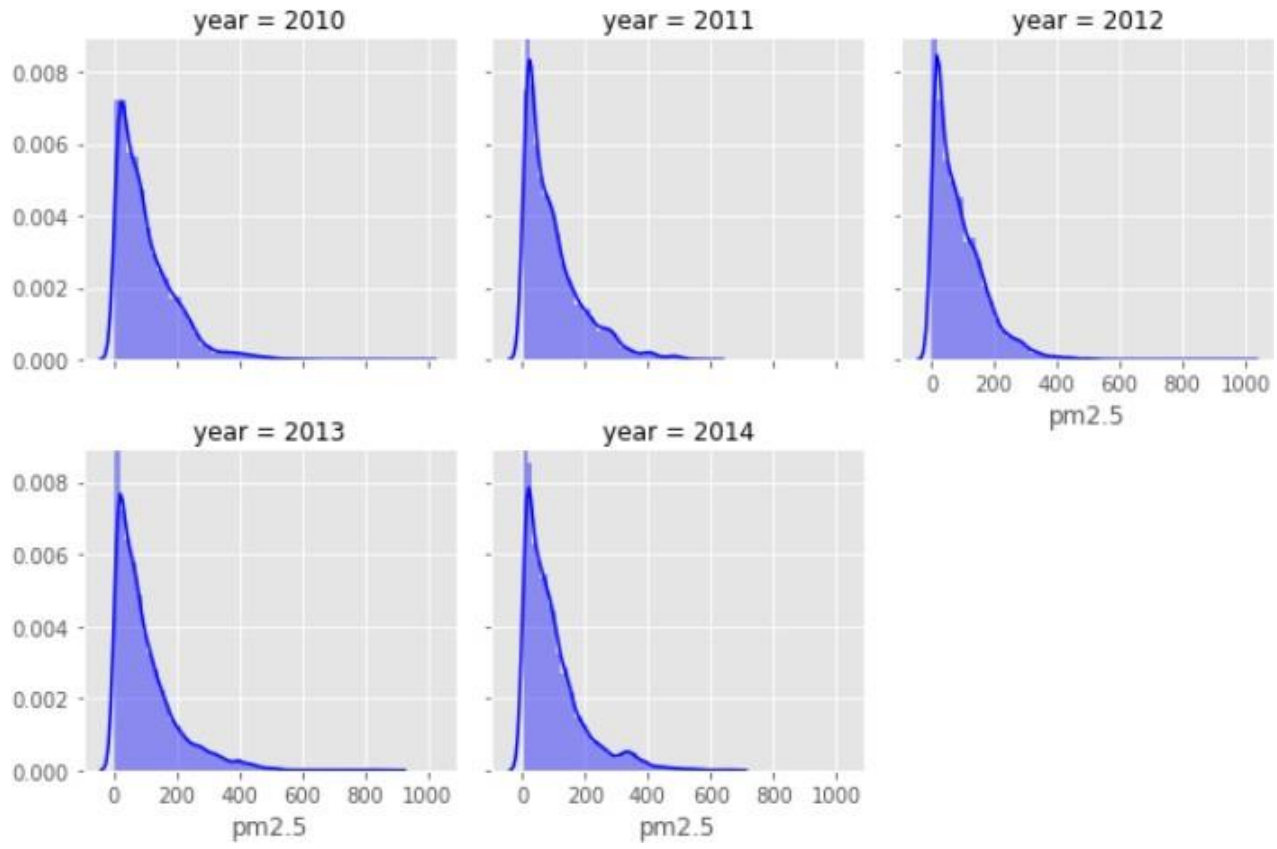
5. RESEARCH QUESTION

Air pollution has been the most notorious factor for varied health conditions all over the world. With the pollution levels sky rocketing, it's helpful to know what the air pollution level is right now. Just like how weather forecast tells you if it is going to rain today and if you have to take an umbrella with you, same way it would be nice to have a pollution forecast which helps a person to decide if he/she needs to go out with a mask or not.

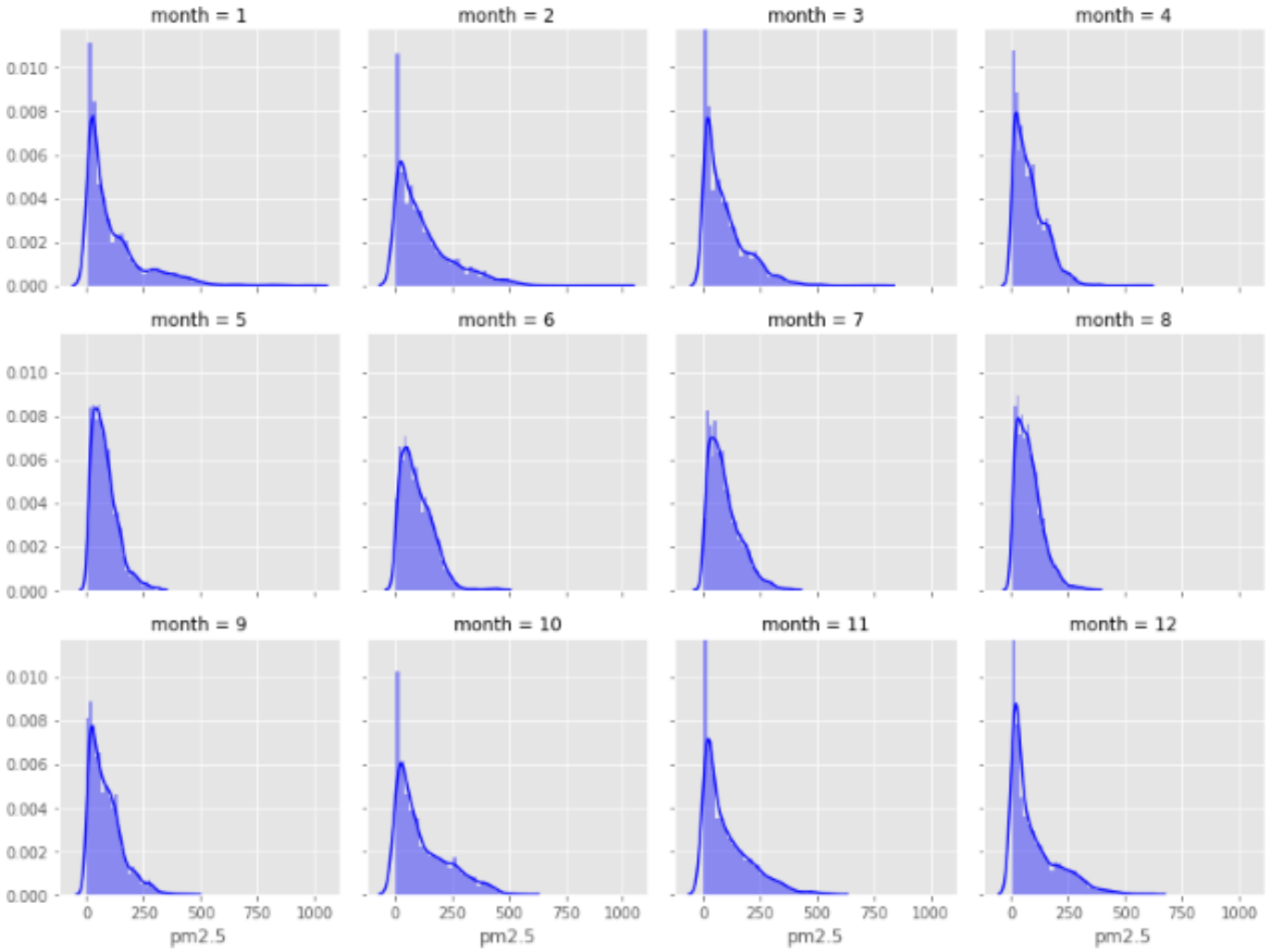
From the background research we have made major assumptions that Wind speed and direction, Dew point, temperature, month and hour of the day are the major predictors of pollution. We aim to validate the assumptions made by predicting the PM2.5 values using Machine Learning Models.

6. TRENDS OBSERVED FROM THE DATA

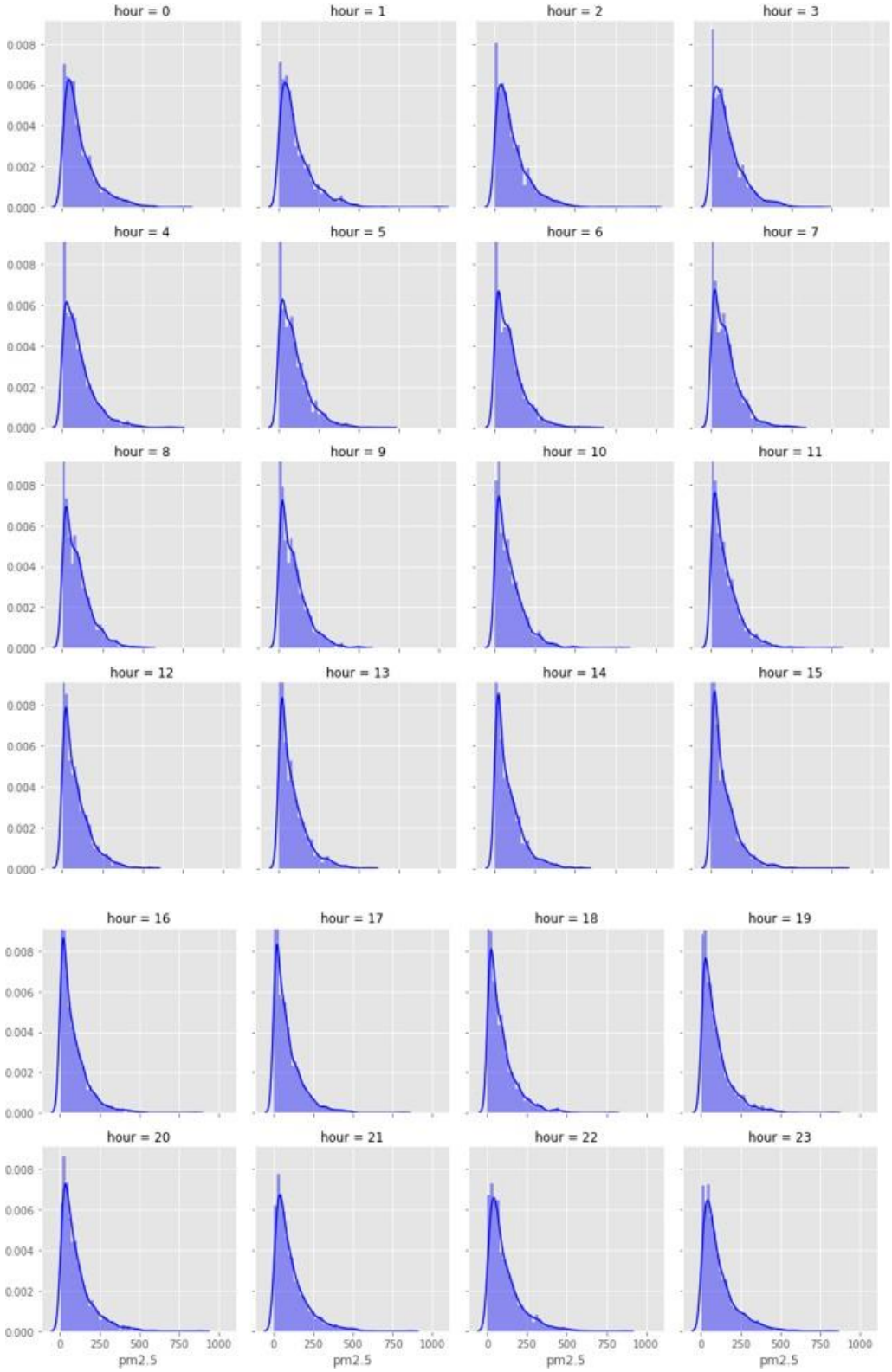
PM2.5 level distribution for every year for the period of 2010 – 2014



PM2.5 level distribution for every month for the period of 2010 – 2014



PM2.5 level distribution for every day for the period of 2010 – 2014



7. DATA PRE-PROCESSING

1. **FEATURE SHIFTING/LAG FEATURES**
2. **HANDLING MISSING VALUES**
3. **HANDLING OUTLIERS**

FEATURE SHIFTING / LAG FEATURES

In our first runs through the model the performance was not ideal. We theorized this is because our current data ignores all previous data. As one can imagine air pollution rises and falls gradually. However, in our dataset this information does not exist. As a result, we investigated the concept of feature shifting/ lag features.
Lag features are the classical way that time series forecasting problems are transformed into supervised learning problems. The simplest approach is to predict the value at the next time (t+1) given the value at the previous time (t-1). The supervised learning problem with shifted values looks as follows:

Value(t-1), Value(t+1)
Value(t-1), Value(t+1)
Value(t-1), Value(t+1)

The addition of lag features is called the sliding window method. A difficulty with the sliding window approach is how large to make the window for your problem.

We figured out good starting point to be to perform a sensitivity analysis and try a suite of different window widths to in turn create a suite of different “views” of our dataset and see which results in better performing models. We hoped to find a point of diminishing returns.

In order to find the appropriate window width, we shifted features to create previous hour features – going back – 3 hours, 5 hours, 7 hours. Then compared the results to find a point of diminishing returns.

No	year	month	day	hour	DEWP	TEMP	PRES	cbwd	Iws	...	pm2.5_2	TEMP_2	Iws_2	DEWP_2	cbwd_2	pm2.5_3	TEMP_3	Iws_3	DEWP_3	cbwd_3
1	2010	1	1	0	-21	-11.0	1021.0	NW	1.79	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2010	1	1	1	-21	-12.0	1020.0	NW	4.92	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	2010	1	1	2	-21	-11.0	1019.0	NW	6.71	...	NaN	-11.0	1.79	-21.0	NW	NaN	NaN	NaN	NaN	NaN
4	2010	1	1	3	-21	-14.0	1019.0	NW	9.84	...	NaN	-12.0	4.92	-21.0	NW	NaN	-11.0	1.79	-21.0	NW
5	2010	1	1	4	-20	-12.0	1018.0	NW	12.97	...	NaN	-11.0	6.71	-21.0	NW	NaN	-12.0	4.92	-21.0	NW
6	2010	1	1	5	-19	-10.0	1017.0	NW	16.10	...	NaN	-14.0	9.84	-21.0	NW	NaN	-11.0	6.71	-21.0	NW
7	2010	1	1	6	-19	-9.0	1017.0	NW	19.23	...	NaN	-12.0	12.97	-20.0	NW	NaN	-14.0	9.84	-21.0	NW
8	2010	1	1	7	-19	-9.0	1017.0	NW	21.02	...	NaN	-10.0	16.10	-19.0	NW	NaN	-12.0	12.97	-20.0	NW
9	2010	1	1	8	-19	-9.0	1017.0	NW	24.15	...	NaN	-9.0	19.23	-19.0	NW	NaN	-10.0	16.10	-19.0	NW
10	2010	1	1	9	-20	-8.0	1017.0	NW	27.28	...	NaN	-9.0	21.02	-19.0	NW	NaN	-9.0	19.23	-19.0	NW

HANDLING MISSING VALUES

Of the 43000 plus rows, around 2000 values were missing, which is less than 5% of the data. We devised a few approaches to handle the missing values:

1. Median – Replacing the values with the median values of the data from same time from previous years. Tried to compare the values and assigned the median.
2. Mean – Replacing the values with the mean value of the data an hour before and after the missing value.

Each of these approaches produced RMSE (Root Mean Squared Error) values around 40 -50 which was not a very favorable result.

As a final option, we opted to drop the null values as they constituted of less than 5% of the entire data. Dropping the null values proved to be advantageous. This produced RMSE values of around 20-21.

```
#Total number of rows
df.shape
```

```
(43824, 13)
```

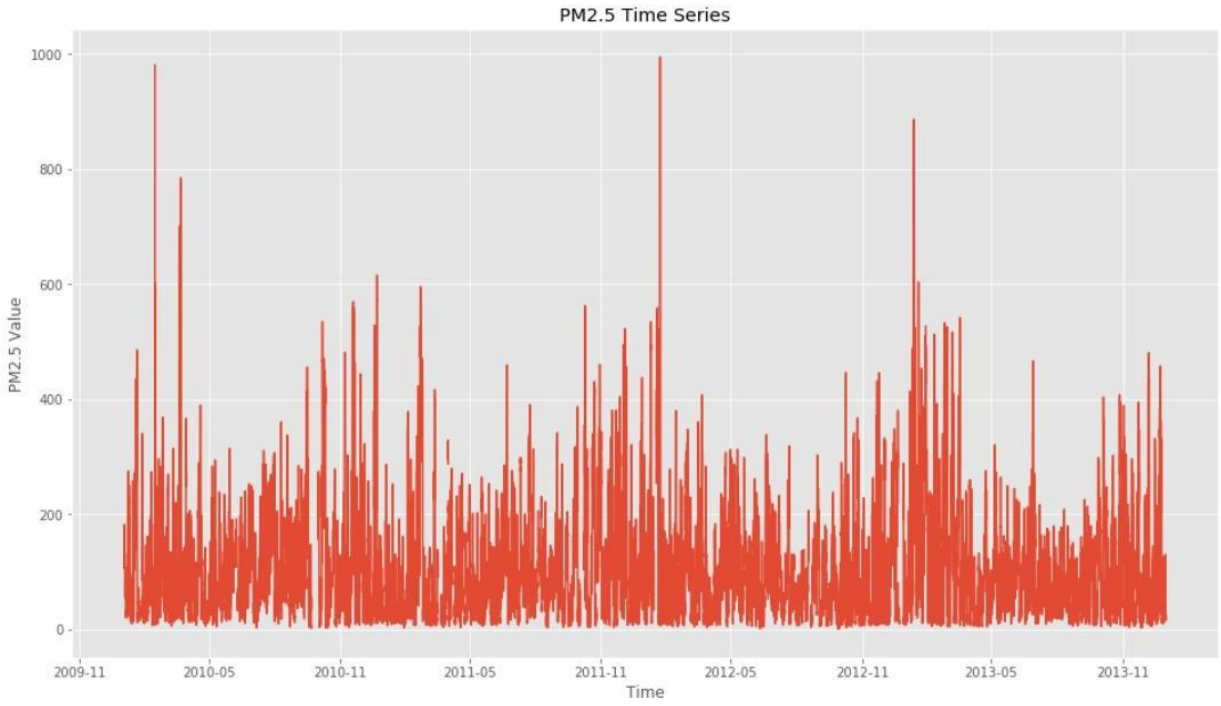
```
#Total number of null values in each column
df.isnull().sum(axis = 0)
```

```
No          0
year        0
month       0
day         0
hour        0
DEWP        0
TEMP        0
PRES        0
cbwd        0
Iws         0
Is          0
Ir          0
pm2.5    2067
dtype: int64
```

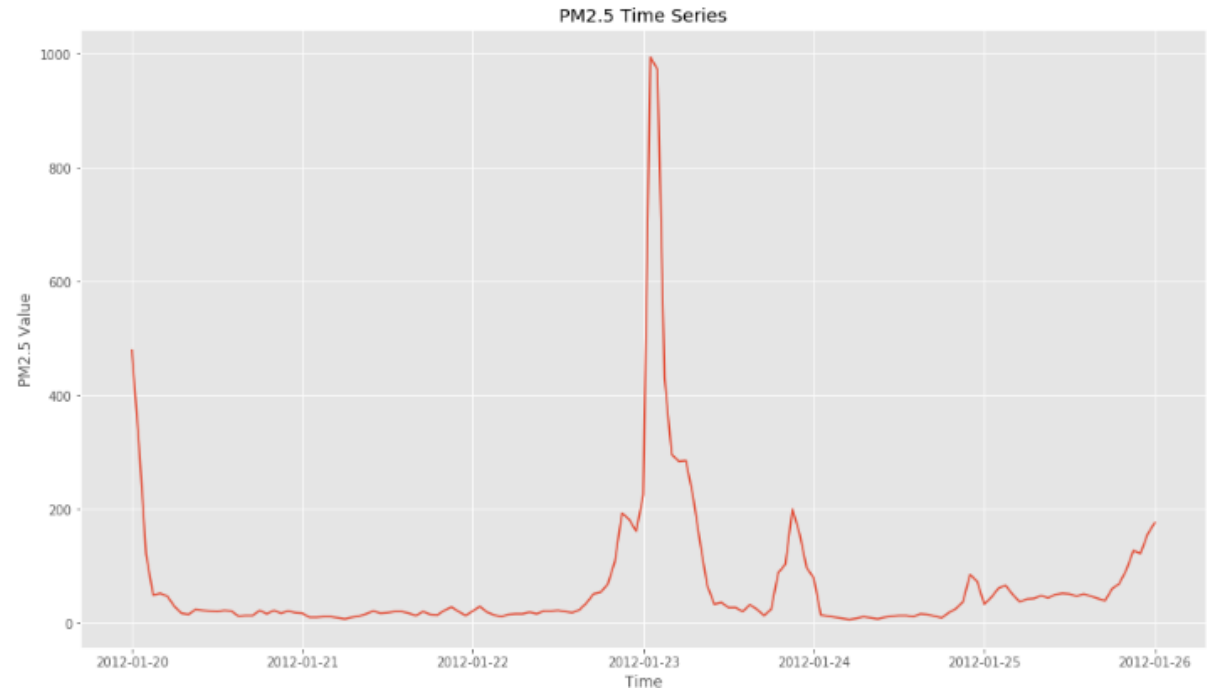
HANDLING THE OUTLIERS

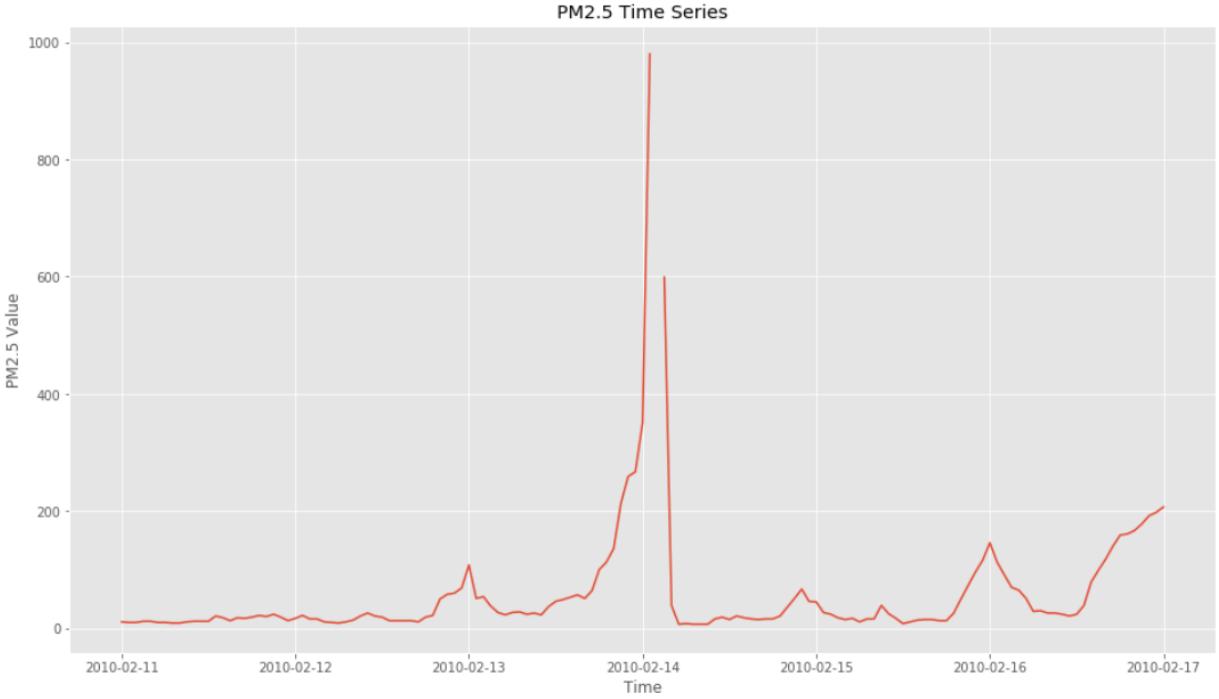
An outlier is any data point that is distinctly different from the rest of your data points. In general, outliers belong to one of two categories: a mistake in the data or a true outlier. The first type, a mistake in the data, could be as simple as typing 10000 rather than 100.00 – resulting in a big shift as we're analyzing the data later. The second type, a true outlier, would be like finding something significant in your dataset. The significance of this data might be the reason for the skew in the results. It's important to distinguish these types. It's subjective. It's up to the analyst to determine which data points are outliers in any given dataset.

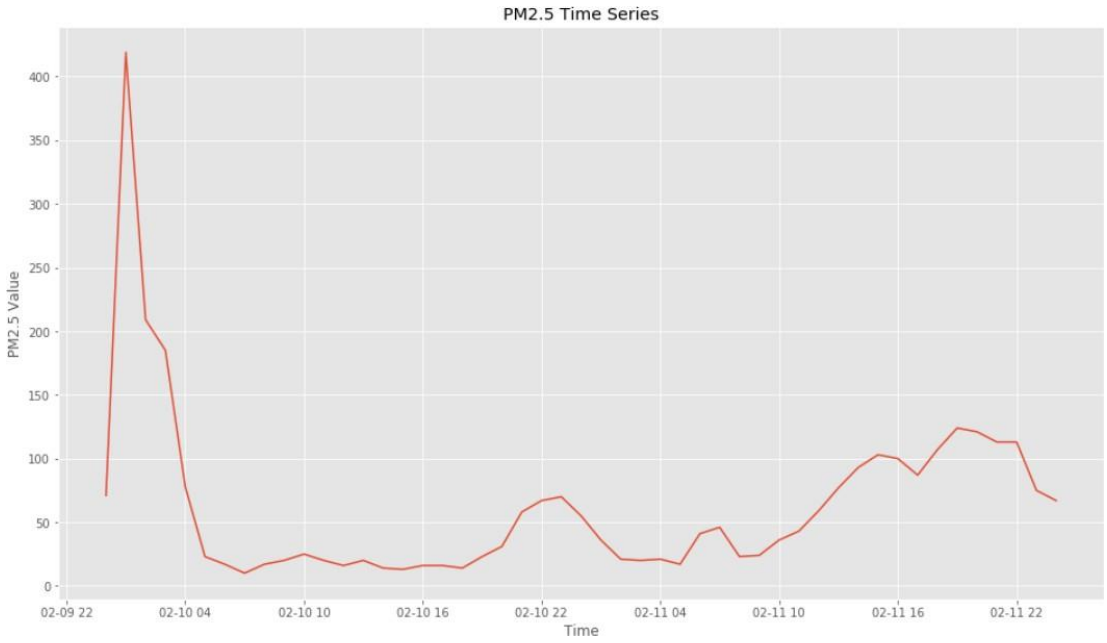
We wanted to determine if this data contained outliers and if it contained outliers, which type of outliers were these - was the question we tried to address.



From the time-series graph It can observed that values above 600 seem to be short-lived. Considering the time of these points and how quickly they come and go these are clear outliers. But we wanted to find out which type outliers these were. So, taking a closer look,


PM2.5 Time Series
PM2.5 Value
1000
800
600
400
200
0
2012-01-20
2012-01-21
2012-01-22
2012-01-23
2012-01-24
2012-01-25
2012-01-26
Time

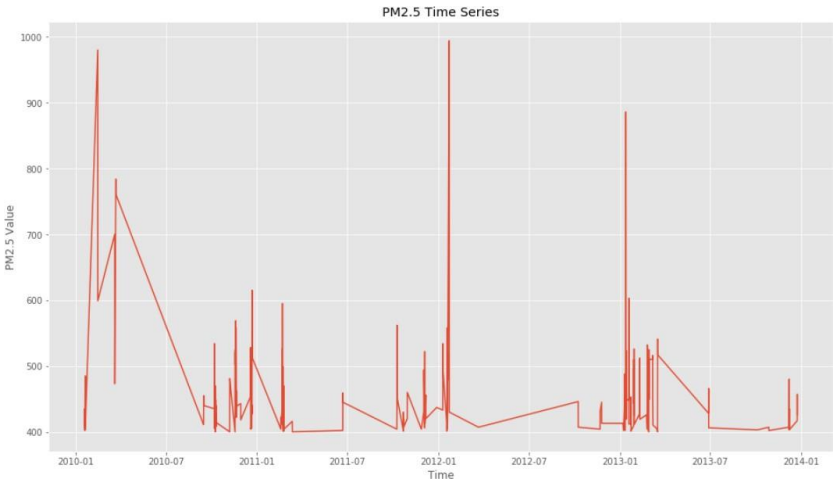

PM2.5 Time Series
PM2.5 Value
1000
800
600
400
200
0
2010-02-11
2010-02-12
2010-02-13
2010-02-14
2010-02-15
2010-02-16
2010-02-17
Time



From taking a closer look at the outliers, it can be observed that the major spikes were observed during the January and February months of every year. Digging deep revealed that these days when there were major spikes was the Chinese New Year.

Year	Chinese Year	Dates
2010	4708	14th February
2011	4709	3rd February
2012	4710	23rd January
2013	4711	10th February
2014	4712	31st January

High values of pm2.5 was observed on these dates. This is the look at the dates and values above 500.



After finding out that the outliers were observed on the Chinese New Year, we tried to take a closer look at the amount of spike.

	datetime	pm2.5
8	2010-02-14 01:00:00	980
9	2010-02-14 03:00:00	599

	datetime	pm2.5
9553	2011-02-03 01:00:00	348.0
9573	2011-02-03 21:00:00	250.0
9574	2011-02-03 22:00:00	288.0
9575	2011-02-03 23:00:00	289.0

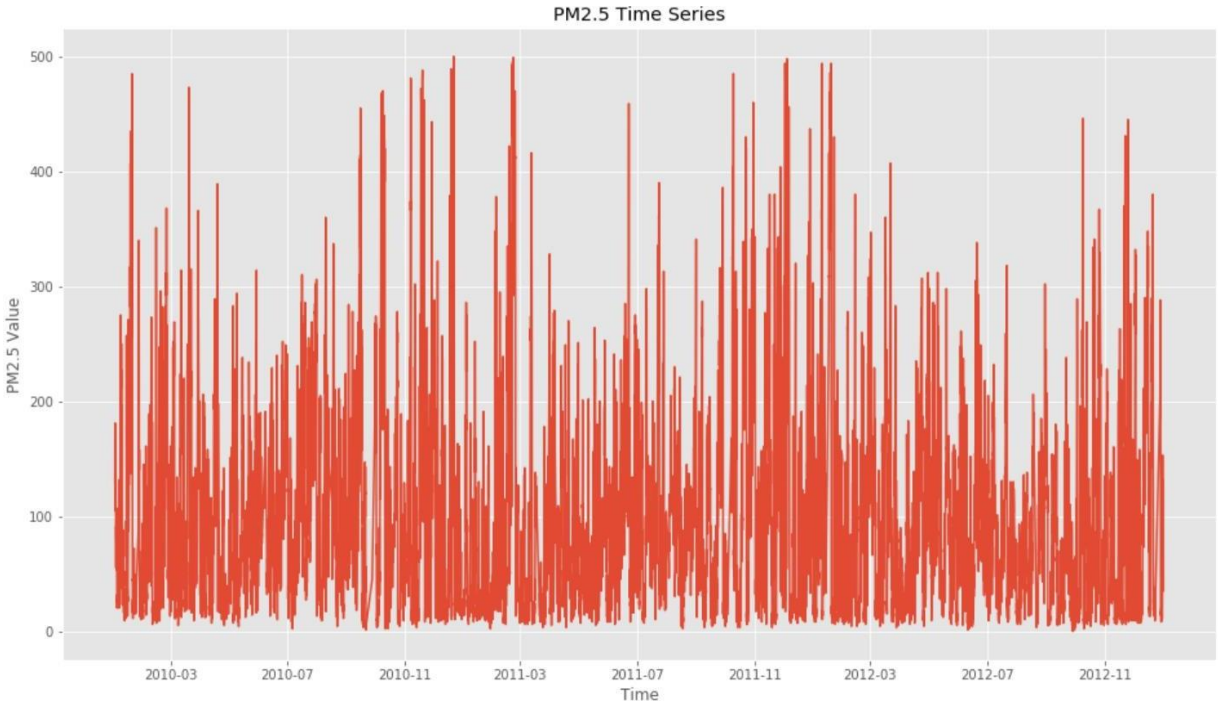
	datetime	pm2.5
276	2012-01-23 01:00:00	994
277	2012-01-23 02:00:00	972
278	2012-01-23 03:00:00	430

	datetime	pm2.5
27265	2013-02-10 01:00:00	419.0
27266	2013-02-10 02:00:00	209.0
27337	2013-02-13 01:00:00	229.0
27338	2013-02-13 02:00:00	246.0
27339	2013-02-13 03:00:00	263.0

	datetime	pm2.5
466	2014-01-31 01:00:00	469

From this data it can be observed that every year around mid- night before new year or early morning of the new year PM2.5 values are very high. It is worth to be noted that 2011 was a special case because this year the city experienced benign weather conditions and government controls on coal burning & vehicle exhausts cut pollution to more than half the normal levels. This was attributed as the reason why 2011 has low pm2.5 around the new year. In the year 2012, even after imposing rules against fireworks, there were cases of rule violation and this lead to a dangerous increase in the pm2.5 levels.

Finally, after analyzing all of this information related to the outliers, we decided to drop the data with pm2.5 values greater than 500. After dropping the values, this how the cleaned up time series looked like-



8. FEATURE SELECTION

PM2.5 is set to be the target variable. Predictor features selected are Dew point, Temperature, Month, Hour, Wind direction, Wind speed. Further on before proceeding with model fitting, the data was split into train and test set. This was done using the SciKit learn's built in libraries. We did a 70:30 split. That means, 70% of the data was assigned for training and 30% of the data was assigned for testing. Higher the percentage of training data available, better trained is the model and hence can make better predictions. Following the data split, we proceeded with model fitting.

9. MODEL FITTING

Model Selection

Predictive models are extremely useful for forecasting future outcomes and estimating metrics that are impractical to measure. For example, data scientists could use predictive models to forecast crop yields based on rainfall and temperature, or to determine whether patients with certain traits are more likely to react badly to a new medication. Here we are using the same concept to predict the pm2.5 values in Beijing based on various different predictive factors.

1) Linear Regression

Linear regression is one of the simplest and most common supervised machine learning algorithms that data scientists use for predictive modelling. Linear regression describes the relationship between a response variable (or dependent variable) of interest and one or more predictor (or independent) variables. It helps us to separate the signal (what we can learn about the response variable from the predictor variable) from the noise (what we can't learn about the response variable from the predictor variable).

Linear regression is a basic yet super powerful machine learning algorithm. Linear regression is widely used in different supervised machine learning problems, it focuses on regression problem (the value we wish to predict is continuous). It deals with dataset of a single feature per data point.

2) Multi-Layer Perceptron Regressor

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function $f(\cdot):Rm \rightarrow Ro$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. The advantages of Multi-layer Perceptron are:

- Capability to learn non-linear models.
- Capability to learn models in real-time.

Class MLP Regressor implements a multi-layer perceptron (MLP) that trains using backpropagation with no activation function in the output layer, which can also be seen as using the identity function as activation function. Therefore, it uses the square error as the loss function, and the output is a set of continuous values. MLP Regressor also supports multi-output regression, in which a sample can have more than one target.

3) Extreme Gradient Boost Regressor

XGBoost is an algorithm that has recently been dominating in applications of machine for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. Generally, XGBoost is fast. Really fast when compared to

other implementations of gradient boosting. XGBoost dominates structured or tabular datasets on classification and regression predictive modelling problems. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. The two reasons to use XGBoost are also the two goals of the project:

- Execution Speed.
- Model Performance.

4) Ridge Regression

Ridge regression is an extension for linear regression. It's basically a regularized linear regression model. The λ parameter is a scalar that should be learned as well. A super important fact we need to notice about ridge regression is that it enforces the β coefficients to be lower, but it does not enforce them to be zero. That is, it will not get rid of irrelevant features but rather minimize their impact on the trained model.
We tried fitting multiple other models too, like -

1. Lasso Regression
2. Elastic Net Regression
3. Decision Forest
4. Extra Trees
5. Decision Tree
6. Boosted Decision Tree

10. PREDICTION

We considered the Root Mean squared Error to determine the best fit. The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data ie, how close the observed data points are to the model's predicted values. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction. Hence, RMSE is our best fit prediction tool.

One other prediction parameter considered is R-squared, it has the useful property that its scale is intuitive: it ranges from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model, and one indicating perfect prediction. Improvement in the regression model results in proportional increases in R-squared.

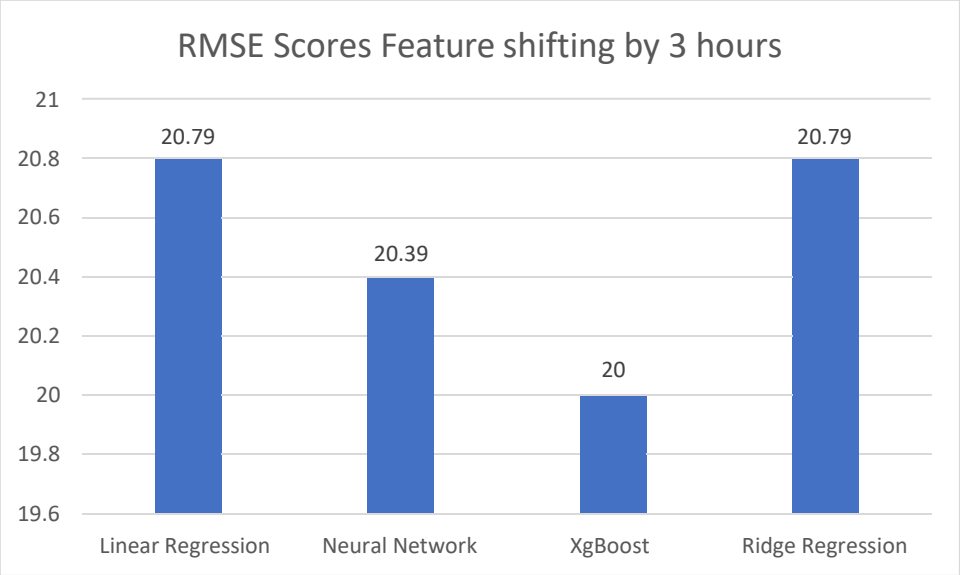
11. RESULTS AND DISCUSSION

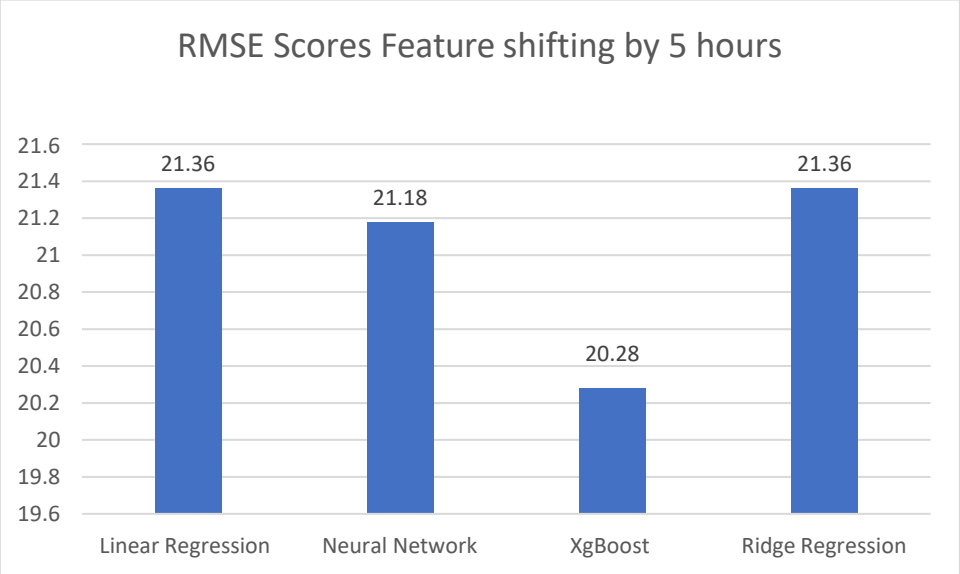
As mentioned earlier, to find the appropriate window width, we shifted features to create previous hour features – going back – 3 hours, 5 hours, 7 hours. Then compared the results to find a point of diminishing returns. Here are the results –

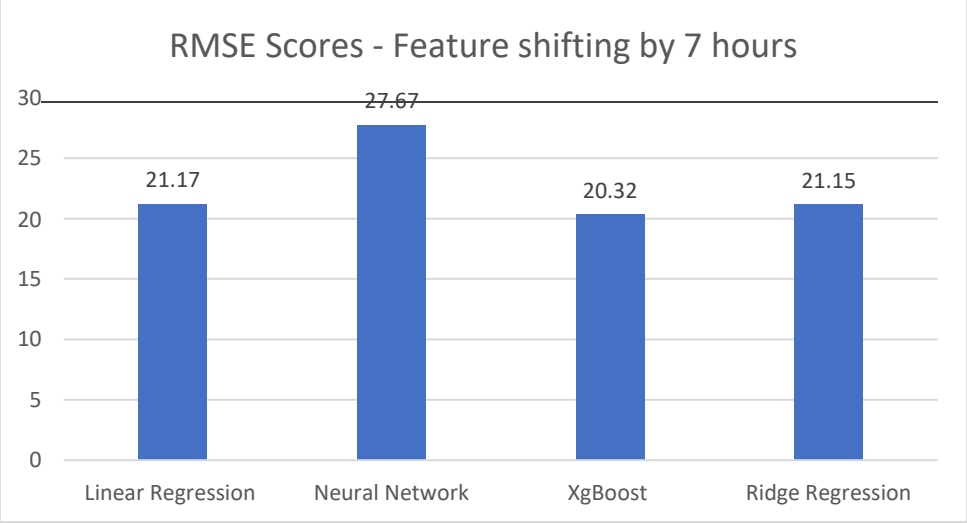
```
COMPARISON SCORES - Feature shifting by 3 hours


Linear regression score:  0.9434184588566525
Lasso regression score:  0.943381330281167
ElasticNet regression score:  0.9433675148970504
Decision forest score:  0.9442163189985464
Neural network regression score:  0.9455794745730408
Extra Trees score:  0.9467953808265207
Boosted decision tree score:  0.9425613651857501
XGBoost score: 0.9476438624381426
Ridge Regression score: 0.943418463241306
Naive Bayes score: 0.021525465031272846


RMSE:
Linear regression RMSE: 20.79
Lasso RMSE: 20.80
ElasticNet RMSE: 20.80
Decision forest RMSE: 20.64
Neural network RMSE: 20.39
Extra Trees RMSE: 20.16
Boosted decision tree RMSE: 20.95
XGBoost RMSE: 20.00
Ridge Regression RMSE: 20.79
Naive Bayes RMSE: 20.79
```

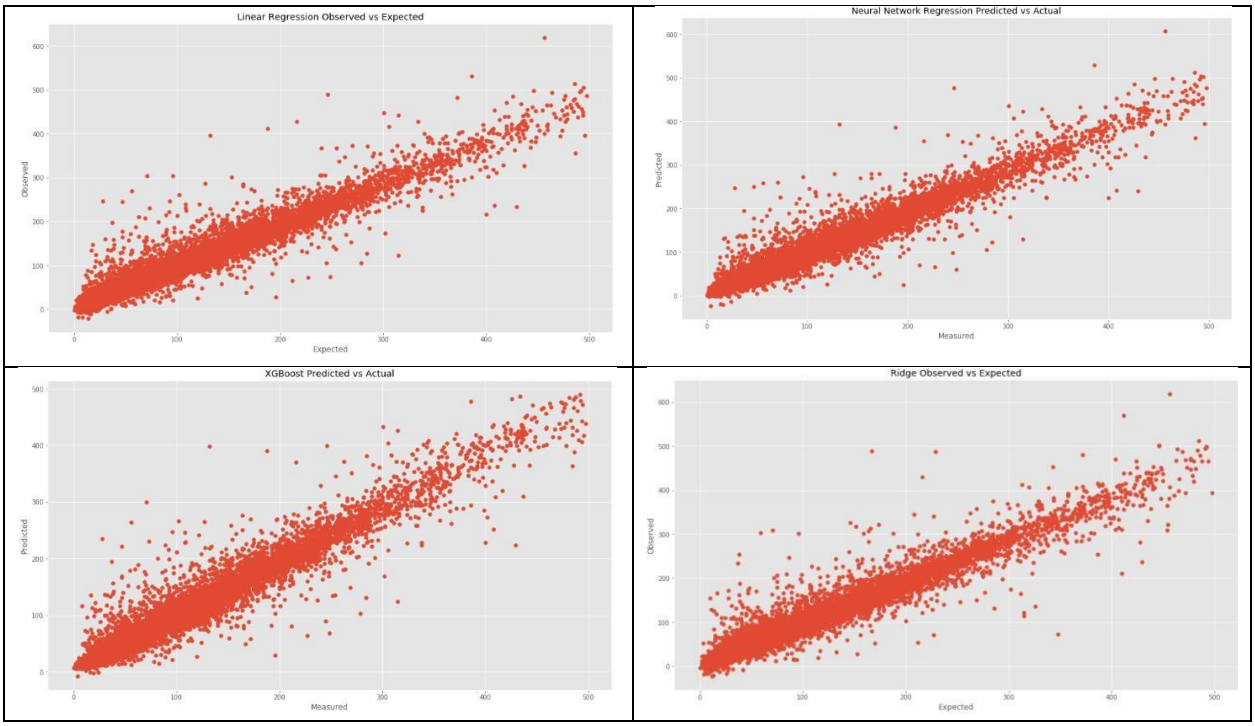
```
COMPARISON SCORES - Feature shifting by 5 hours


Linear regression score:  0.9414371826453103
Lasso regression score:  0.9414362939298039
ElasticNet regression score:  0.9414350178079701
Decision forest score:  0.9430850789458225
Neural network regression score:  0.9424041277232377
Extra Trees score:  0.9444098674305059
Boosted decision tree score:  0.9415269063502714
XGBoost score: 0.947211767766174
Ridge Regression score: 0.9414371941266567
Naive Bayes score: 0.016476760390195917


RMSE:
Linear regression RMSE: 21.36
Lasso RMSE: 21.36
ElasticNet RMSE: 21.36
Decision forest RMSE: 21.06
Neural network RMSE: 21.18
Extra Trees RMSE: 20.81
Boosted decision tree RMSE: 21.34
XGBoost RMSE: 20.28
Ridge Regression RMSE: 21.36
Naive Bayes RMSE: 21.36
```

```
COMPARISON SCORES - Feature shifting by 7 hours


Linear regression score:  0.941437113280058
Lasso regression score:  0.9413149088806759
ElasticNet regression score:  0.9413644246607168
Decision forest score:  0.9418515999785358
Neural network regression score:  0.8997868551433231
Extra Trees score:  0.944349836890589
Boosted decision tree score:  0.9406877015851911
XGBoost score: 0.9459407299653343
Ridge Regression score: 0.9414427110963889
Naive Bayes score: 0.016127698287982797


RMSE:
Linear regression RMSE: 21.15
Lasso RMSE: 21.17
ElasticNet RMSE: 21.16
Decision forest RMSE: 21.08
Neural network RMSE: 27.67
Extra Trees RMSE: 20.62
Boosted decision tree RMSE: 21.29
XGBoost RMSE: 20.32
Ridge Regression RMSE: 21.15
Naive Bayes RMSE: 21.15
```


RMSE Scores Feature shifting by 3 hours
21
20.8
20.6
20.4
20.2
20
19.8
19.6
20.79
20.39
20
20.79
Linear Regression
Neural Network
XgBoost
Ridge Regression


RMSE Scores Feature shifting by 5 hours
21.6
21.4
21.2
21
20.8
20.6
20.4
20.2
20
19.8
19.6
21.36
21.18
20.28
21.36
Linear Regression
Neural Network
XgBoost
Ridge Regression


RMSE Scores - Feature shifting by 7 hours
30
25
20
15
10
5
0
27.67
21.17
20.32
21.15
Linear Regression
Neural Network
XgBoost
Ridge Regression



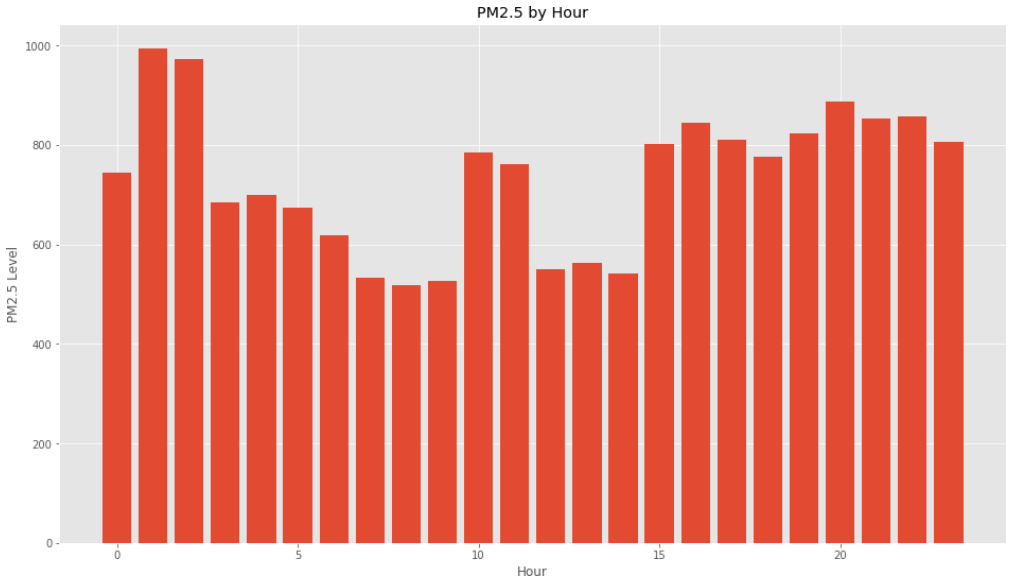
The window width of 3 proved to be most favorable and Of all the models used, Extreme Gradient Boost performed best with an RMSE score of 20.00

12. INSIGHTS

Meteorological factors can significantly affect PM2.5 mass concentration, which can help to reduce or aggravate the urban air pollution. Following are our observations -

1) Hour of the day

Generally, the daily pattern of PM2.5 shows minimal concentrations at 7-8 a.m., followed by an increase until noon due to the morning rush hour. The PM concentrations maintain stable concentrations until 3 p.m., when a second peak occurs and raises the PM concentration to maximum levels around 6 p.m., i.e., during the evening rush hour. When people are on the move the pollution builds. Road traffic is one of the more influential PM sources. The concentration decreases during the night time, which suggests the limited effect of domestic heating emissions on PM concentrations. The daily pattern also shows that the mixing layer dynamics have a limited effect on PM concentrations.



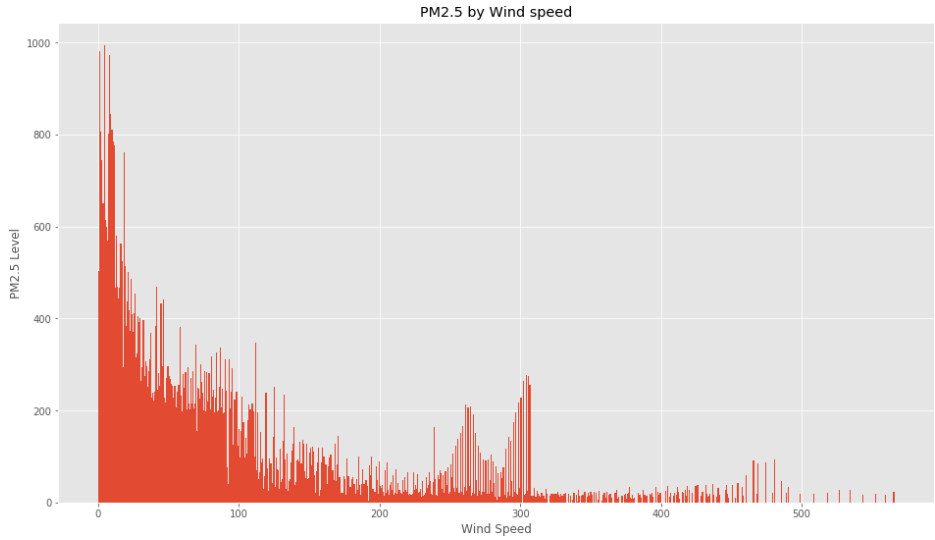
2) MONTH OF THE YEAR

The monthly average PM2.5 concentrations are lowest in summer (June, July, August), which are followed by spring (March, April, May) and autumn (September, October, November), and are highest in winter (January, February, December). We found that during high temperature weather in summer, although PM2.5 mass concentration was 2 to 3 times higher than that of low temperature period, the high temperature weather was still helpful in the diffusion of pollutants. Generally, $PM_{2.5}$ concentrations show a remarkable seasonal variability with the highest during the winter and the lowest during the summer. The winter maximum $PM_{2.5}$ level is due to the increasing anthropogenic activities such as fossil-fuel and biomass burning for heating in the cold season. Furthermore, more frequent occurrences of stagnant weather and temperature

inversion during the cold period may facilitate the accumulation of air pollutants. The $PM_{2.5}$ keeps in high abundance at the evening hours in the cold seasons because of increasing emissions for heating and stagnant atmospheric conditions. Winter months have the worst pollution levels, but some summer months have average levels in the unhealthy range.

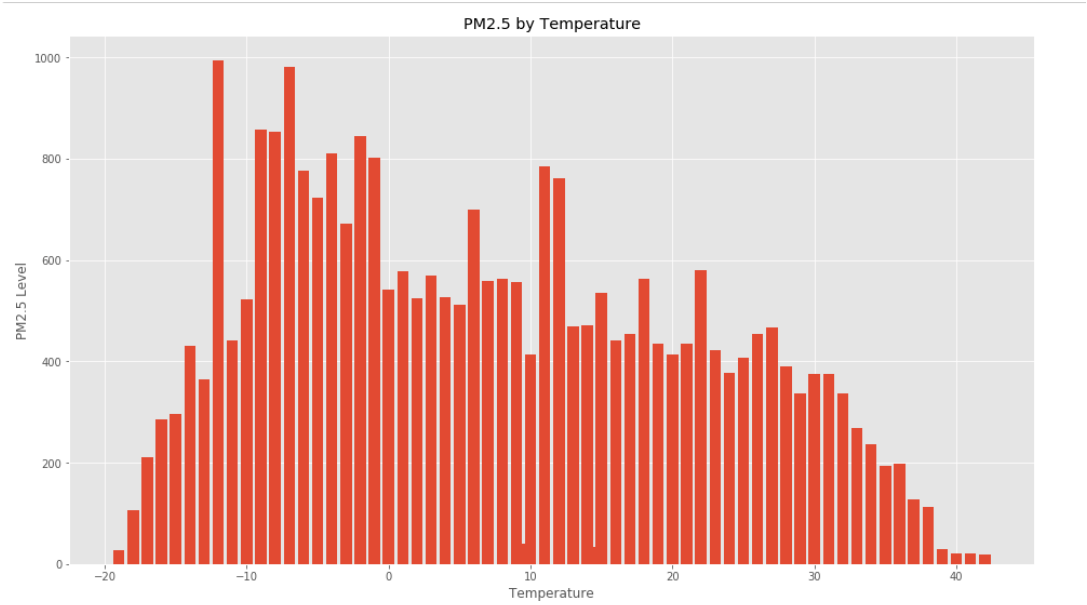
3) PM2.5 by Wind Speed

It is found that wind speed has an inverse effect on the pm2.5 concentration. With high speed winds, the concentration of the pm2.5 particles decrease. As an important meteorological factor, wind influences the horizontal and vertical transport of air pollutants. It also affects the concentration and diffusion of pollutants directly. The PM2.5 concentration reduces significantly with increase in wind speed. On the contrary, low wind speed inhibits the diffusion of PM2.5 and makes the PM2.5 accumulate on the surface. Hence it is true that wind tends to alleviate the air pollution, and stronger wind during the worst of the high $PM_{2.5}$ episodes is helpful.



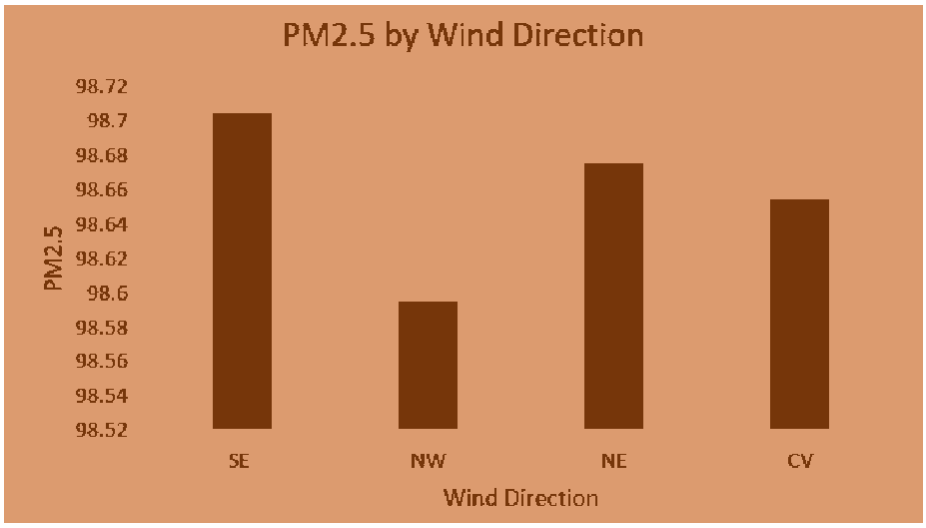
4) PM2.5 by Temperature

When the temperature is higher, the air convection at lower surface is stronger, which benefits the upward transport of particulate matter. In high temperature weather, PM2.5 mass concentration was 2 to 3 times higher than that of low temperature period. While temperature is important, it isn't as simple as cold = pollution.



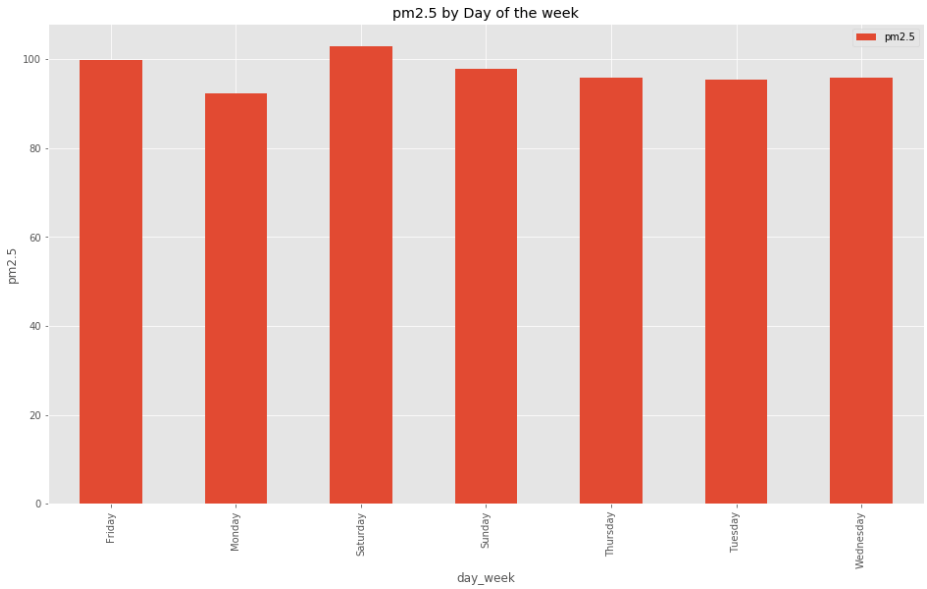
5) PM2.5 by Wind direction

Besides wind speed, wind direction influences the transport direction of Particulate matter. With respect to the wind direction, the baseline wind distribution in Beijing is dominated by NW and SE, with winters being dominated by NW and summers by SE. The benefit of northerly wind is due to a lack of heavily polluting industry in the region north of Beijing. However, the mountains cause accumulation of the polluted air under a southerly wind.



6) PM2.5 by day of the week

The day of the week has some small bearing, It can be seen that on Saturdays and Fridays the pm2.5 values are slightly higher as compared to other days. Domain research shows that people mostly use their wood stoves on the weekends for recreational purposes. It can also be understood that majority of the population might be on the move on weekends which might also be the reason for weekends having higher levels of pollution.



Among all meteorological factors, only pressure has the similar monthly tendency with PM2.5 concentrations while temperature, rainfall, wind speed and wind direction have opposite tendency with PM2.5.

13. GOING FORWARD

- Going forward, we would like to refresh the entire study with most recent data to understand the present pollution picture.
- Scrap the web to collect our own data to perform the same kind of analysis.
- Since we are dealing with time series data, we know that time series data tends to be correlated in time, and exhibits a significant autocorrelation.
- We would like to try to use Long Short-Term Memory(LSTM) and Auto Regressive Integrated Moving Average(ARIMA) models, which have a good reputation of dealing time series data sensibly.

Long Short-Term Memory

Time series prediction problems are a difficult type of predictive modeling problems. Unlike regression predictive modeling, time series also adds the complexity of a sequence dependence among the input variables. A powerful type of neural network designed to handle sequence dependence is called recurrent Neural networks. The Long Short-Term Memory network or LSTM network is a type of recurrent neural network used in deep learning because very large architectures can be successfully trained. Long Short-Term Memory models are extremely powerful time-series models. They can predict an arbitrary number of steps into the future.

ARIMA (Auto Regressive Integrated Moving Average)

A popular and widely used statistical method for time series forecasting is the ARIMA model. ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data. An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts. ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average. It is a generalization of the simpler Auto Regressive Moving Average and adds the notion of integration. Adopting an ARIMA model for a time series assumes that the underlying process that generated the observations is an ARIMA process. This may seem obvious but helps to motivate the need to confirm the assumptions of the model in the raw observations and in the residual errors of forecasts from the model.

14. OTHER RESEARCH WORKS

- Analysis of hourly PM2.5 readings for Beijing: the US Embassy and the Chinese Government.
- Daily PM10 analysis for Shanghai, Beijing, Tianjin and Suzhou (12.5 year period).
- Demonstrating the significant effect of synoptic meteorological conditions on PM2.5 pollution in the Beijing-Tianjin-Hebei area.
- Prediction of any day's PM2.5 at 8am based on data available until 8am of the previous day

Research is done by comparing varying data sources, different cities, different areas, Same time on different days.

15. CONCLUSION

The frequent air pollution episodes and poor air quality in Beijing in recent years warrant attention and analysis. From the results obtained, it can be reassured that the assumption made with regards to the factors responsible for the increase/decrease of pm2.5 values is indeed true. All our analyses point to a conclusion that, up-to the end of 2014, the PM2.5 pollution in Beijing had not improved a lot over the 2012 levels, and the air quality had in fact got worse. The analyses also indicate that a fundamental shift from mainly coal-based energy consumption to much greener alternatives in Beijing and the surrounding North China Plain is the key to solving the PM2.5 problem in Beijing.

16. PEER REVIEW

As part of the peer review, we had the students contribute a lot of productive and useful suggestions. Here is how we handled the suggestions.

1. We had suggestions on using Ridge Regression in place of Linear Regression. – We fit a ridge model to the data and observed that both Linear regression and ridge regression had same values of RMSE and performed the same.
2. About using Auto-correlation models – This is a task on our cards for the future
3. About using LabelEncoder to encode the wind direction values – We tried encoding the values and tried fitting the models. We observed that the performance was comparable.
4. About scraping the web to collect data for analysis – This is on our cards for future work. We would be very much interested to collect data for New Delhi, India to conduct similar kind of analysis.

17. REFERENCES

https://www.rapidinsightinc.com/handle-outliers/
https://royalsocietypublishing.org/doi/full/10.1098/rspa.2015.0257
https://machinelearningmastery.com/basic-feature-engineering-time-series-data-python/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5650241/
https://towardsdatascience.com/how-not-to-use-machine-learning-for-time-series-forecasting-avoiding-the-pitfalls-19f9d7adf424?fbclid=IwAR3Ahv0dAp5y9BztqTzunHacJhgmNT3iYS8eLWwX4sRU0yESThMwrF_xg-E