University of Connecticut School of Business

OPIM 5671 – Data Mining and Business Intelligence

**BEIJING AIR POLLUTION TIME SERIES FORECASTING**

**Instructor**

Prof. Jing Peng

jing.peng@uconn.edu

**Team - A4**

Amit Anand Kumar- Aak20008

Akash Deep Konda- Kad21008

Chaitanya Bandaru- Chb21009

Lohith Krishna Pogala- Lkp21003

Sathwik Pendyala- Sap21007

Venkata Karteek Thota- Vet21001

**Table of contents**

[**Introduction**](#_heading=h.gnoj9tcixlwx) **3**

[**Literature**](#_heading=h.nd97462lnnyu) **3**

[**Data Exploration**](#_heading=h.ot7u92l0crvd) **5**

[**Data Description**](#_heading=h.x0owg5fie4t7) **6**

[**Basic Statistics**](#_heading=h.h3gfy91nwxz) **7**

[**Data Pre-Processing**](#_heading=h.7f8916hx0swh) **7**

[**Data Analysis**](#_heading=h.uqsgt6o5dfyv) **8**

[**Time Series tests**](#_heading=h.p6ghkfga5nh9) **9**

[**Data Visualization**](#_heading=h.peaw3yx5ehix) **11**

[4. Relation between regressors](#_heading=h.frnnjub42fmu) 12

[5. Relation between the output variable and regressors:](#_heading=h.deu55kfad3zu) 13

[**Predictive Modeling**](#_heading=h.fxt6e65i8bme) **15**

[a. Teams Approach](#_heading=h.hv7e6qa3r2th) 15

[b. Models Used](#_heading=h.7dtf1660fo8b) 16

[c. Models Comparison](#_heading=h.mj69krn63x8y) 18

[**Finding**](#_heading=h.5rf2quvxgn5g)**s 19**

[**Responsibilities of Each Team Member**](#_heading=h.e771he4mhy9f) **19**

[**References**](#_heading=h.fd7hbc2hz8qa) **20**

# Introduction

Air pollution is one of the greatest scourges of our day, not only because of its impact on climate change, but also because of its impact on public and individual health due to increased sickness and death. Air pollution has a variety of negative health consequences. Even on days when air pollution is low, vulnerable, and sensitive people's health can be harmed. Cough, shortness of breath, wheezing, asthma, respiratory disease, and high hospitalization rates are all linked to short-term exposure to air pollution (a measurement of morbidity).

Apparently, air pollution is an issue that affects all regardless of how developed a nation is. This is due to the fact that its causes are universal: automobiles, factories, homes, agriculture, landfills, and even natural phenomena, and Beijing, China's capital, is no exception to this global significant issue. Beijing is the world’s most populous city with an estimated population of over 21 million people.

Beijing has extreme climate conditions, the temperature can increase upto 42°C and can drop down till -27°C The main reason for Beijing pollution has been caused by increased vehicle emissions caused by increased population. The pollutant that affects people the most is particulate matter, usually abbreviated as PM and used as a measure of air pollution.

Although particles with a diameter of 10 microns or less (≤PM10) can penetrate and embed deep in the lungs, the ones that are more harmful to health are those with a diameter of 2.5 microns or less (≤PM2.5). Smog levels have risen in Beijing in the last decade resulting in degradation of air quality and prompting international concerns. PM2.5 (particles with a diameter of less than 2.5 micrometers) can penetrate deep into the lungs, irritate and erode the alveolar wall, and impair lung function as a result.

As a result, it's critical to look at the effects of PM2.5 on the respiratory system and help Beijing, China in combating its current air pollution problems. This raises the need for accurate forecasting of pollution levels, in order to take precautions before things get disastrous. The forecasting model would help in Disaster mitigation, Creating an early warning system. The primary goal of this project is to come up with a model given the weather conditions and pollution for prior days, we can forecast the pollution for the upcoming days.

# Literature

There is a very extensive literature on the design and test of models of environmental pollution, especially in the atmosphere. Recent models are more concerned with the results of pure forecasting models, but they do not go into great detail on the causes and their temporal correlations. They anticipated the results for the next three days based on the examination of pollution level estimates, but couldn't truly foresee beyond that. So these are some of the voids that our work has attempted to fill. In addition to developing the forecasting model, we gained valuable insights into how temperature, humidity, windspeed, and dew point affect PM 2.5 levels. We also forecasted PM 2.5 concentrations for the next two months based on an analysis of feature significance criteria and pollution level estimates.

We're looking at five years' worth of data, including hourly PM 2.5 concentrations in the air, Dew Point, Temperature, and Cumulative Wind and Snow Hours. Wind speed and direction over time. To assess different pollutant forecasting models and their properties, we offer a technique based on exactness and robustness criteria. We realized the necessity of parameter adjustment with Autoregressive, First Difference, and Mean Averages. In the process we leveraged Python, SAS and Tableau to come up with the optimal visualization and insights.

Furthermore, we improved the performance of the vanilla ARIMA model by incorporating regressors and event interventions that detected the seasonality patterns in the data. Our aim was to investigate alternative hyper parameter tuning options that would allow us to obtain the best performance model with the least amount of error and high accuracy. As a result, our best models provide a 2 month ahead, very reliable prediction of pollutant concentrations in the air in the studied area, which can be utilized to plan and implement various interventions and steps to mitigate the population's effects.

Unlike autoregressive approaches, our method changes the dataset to remove the temporal order of individual instances by imputing the time dependency through additional input variables known as delayed variables. Lagged variables allow us to see if there's a link between the past and present values of attributes in a time series.

Our general overview of the approach undertaken for this project started with imputing the missing value on the raw data to prepare it for further processing. Time-series analysis is then performed on the imputed data to understand and extract the underlying patterns of the data. The data are then modeled using various forecasting methods as mentioned. The models created thus are then used to train on the entire dataset to produce the next days of forecasting, which is then made the basis for the subsequent discussion presented in the latter part of this paper.

# 

# Data Exploration

Our data Source is a UCI machine learning repository and this data is used to forecast the Air pollution level in Beijing, China and we have collected the data from Jan 1st, 2010 to Dec 31st, 2014.

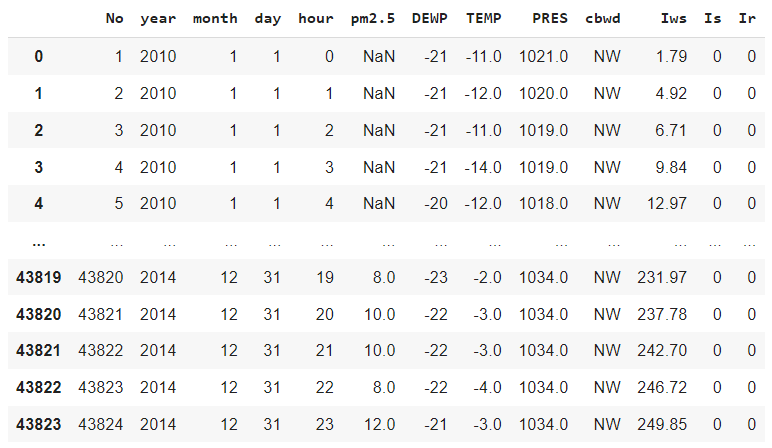
****

Figure: Beijing air pollution dataset from Jan 1st, 2010 to Dec 31st, 2014

**Data Source link: https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data#**

# 

# 

# 

# Data Description

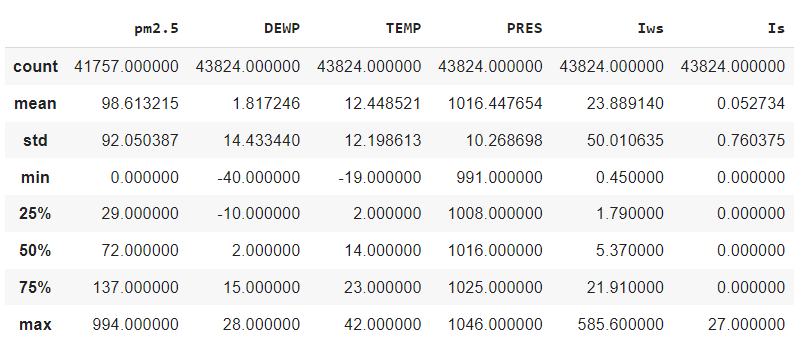
Our Data Set is collected from the source link given below. The main aim of the project is to forecast the pm 2.5 level in Beijing, China. The data which we collected is on an hourly basis i.e., the data of all the parameters were given on an hourly basis.

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| **No** | **int** | **Row Number** |
| **year** | **int** | **Year of the data in this row** |
| **month** | **int** | **Month of the data in this row** |
| **day** | **int** | **Day of the data in this row** |
| **hour** | **int** | **Hour of the data in this row** |
| **pm2.5** | **float** | **PM 2.5 Concentration (ug/m^3)** |
| **DEWP** | **int** | **Dew point** |
| **TEMP** | **float** | **Temperature at beijing at various time** |
| **PRES** | **float** | **Pressure** |
| **cbwd** | **object** | **Combined wind direction** |
| **Iws** | **float** | **Cumulative wind speed (m/s)** |
| **Is** | **int** | **Cumulated hours of snow** |
| **Ir** | **int** | **Cumulated hours of rain** |

Figure 2: Data Description

# Basic Statistics

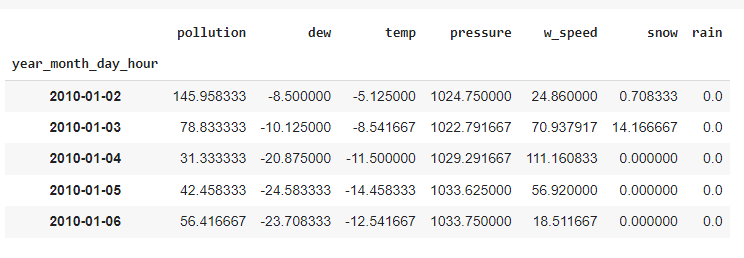
The below figure shows the basic summary statistics of our dataset. If we look at the mean pm 2.5 value which is 98.61, which is almost on the verge of entering the poor category as per the IMECA(Índice Metropolitano de la Calidad del Aire) Scale. The standard deviation of the pm2.5 column is also very high, which tells us that the other parameters are significantly contributing to the increase and decrease of the pollution levels. We can also see that the temperature of Beijing varies from 19 degree celsius to 42 degree celsius, which shows that the climatic variations are significant in beijing, which in turn affect the pollution levels.

****

# Data Pre-Processing

Data Preprocessing is the most essential step in any data related projects. Here in our project as well the dataset we used needed pre-processing steps for it to be made model ready. So, we performed various data pre processing steps.

The data we received was on an hourly basis, i.e., the parameters were logged hourly. Since the data is huge we decreased the granularity and converted it to a daily basis. Further we have considered null values to be zero as they only have 1826 records which is considerably small.We choose to convert the null values to zero instead of interpolation because converting into zero gave the best and accurate results when compared to the original data.

****

# Data Analysis

So after doing the data preprocessing steps the next thing which we were curious about and the thing that needs to be checked is whether our target variable, PM 2.5 concentration time series plot is stationary or not. So for that we need to check whether the time series plot has trend, seasonality or not.

Here in the graph below we could not see any trend as there is no continuous increase or decrease slope in the graph but we could not conclude whether it has seasonality or not. So, for that reason we have conducted white noise, unit root and seasonality tests in SAS and pasted the outcomes of them in the picture below.

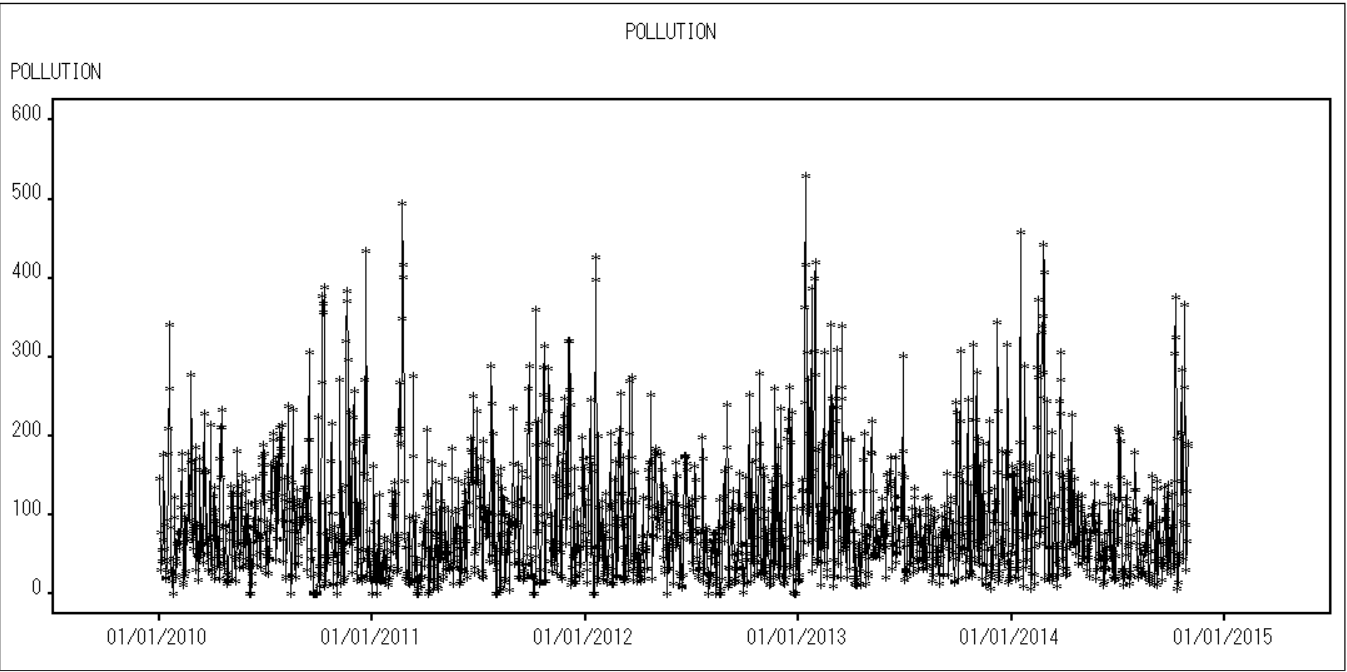
****

Fig: This figure showing if there is any trend or not in the time series plot

# Time Series tests

For a time series forecast we need to check for trend, seasonality and stationary before we forecast data.

A time series is stationary if its mean, variance and covariance are constant over time. A time series will not have trend and seasonality if it is stationary.

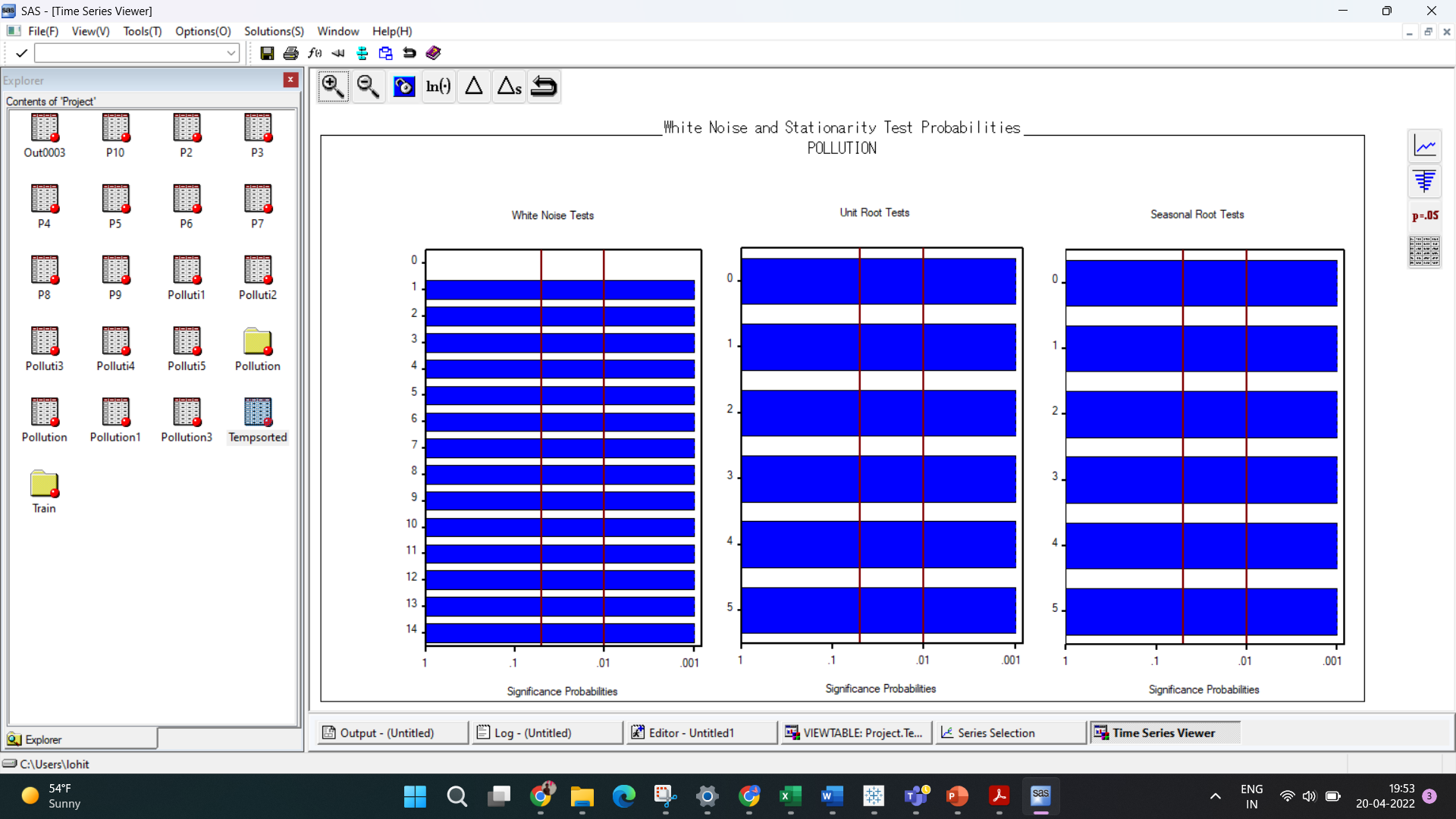
****

Fig: Prediction error tests

From the above plots we can observe white noise tests,Unit root tests and second root tests. From white noise test plot we can clearly say that the time series does not contain any white noise, From Unit root test plot we can say that the time series is stationary over time. From the seasonal root test plot for our time series we say that there is no seasonality.

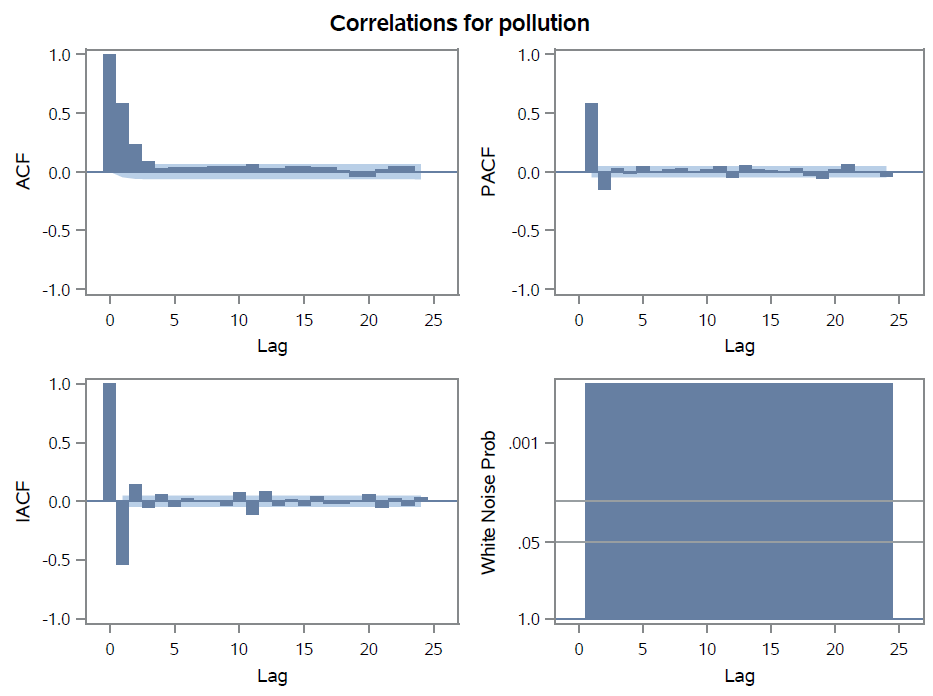
****

Fig: Autocorrelation plots

From the ACF plot we observe that there is an exponentially drop to zero relatively; this tells us that the time series is stationary. The PACF plot shows the partial correlation coefficients between the series and lag1 , lag2.

From the correlation plot we can say that temperature and dew are highly positively correlated and temperature and pressure are highly negatively correlated. And also dew and pressure are highly negatively correlated. Apart from those features, rest all do not show any kind of correlations among themselves

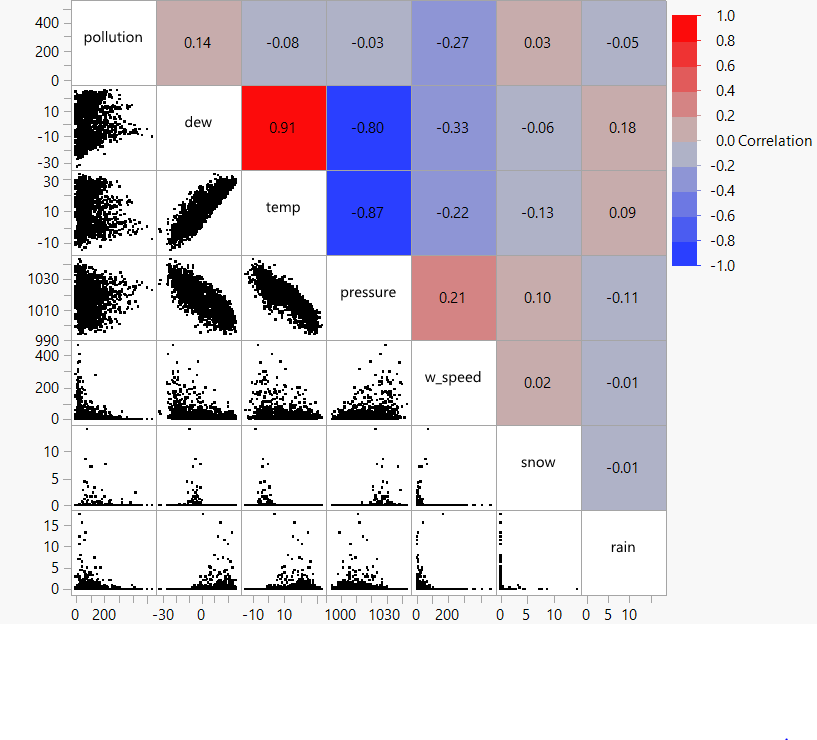
****

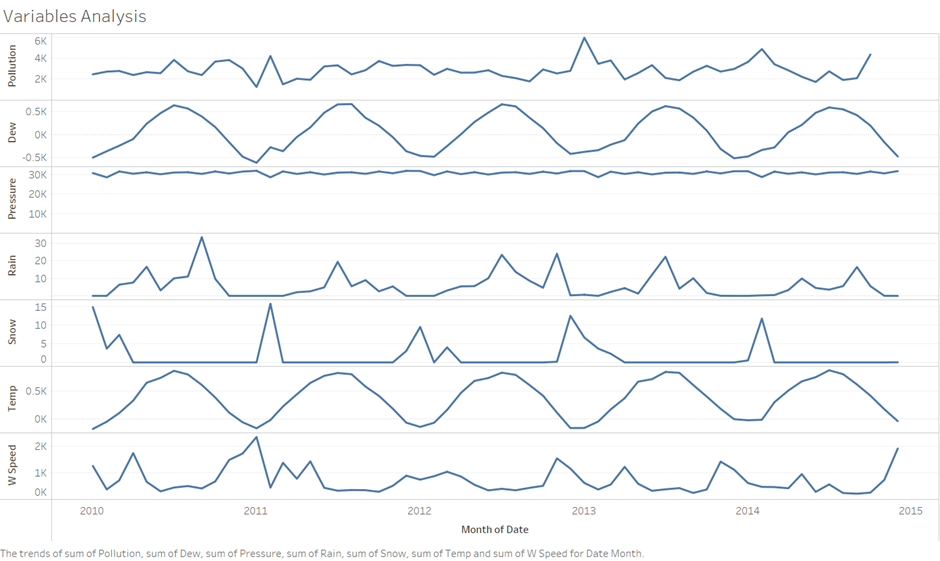
Fig: Correlation heatmap

# 

# Data Visualization

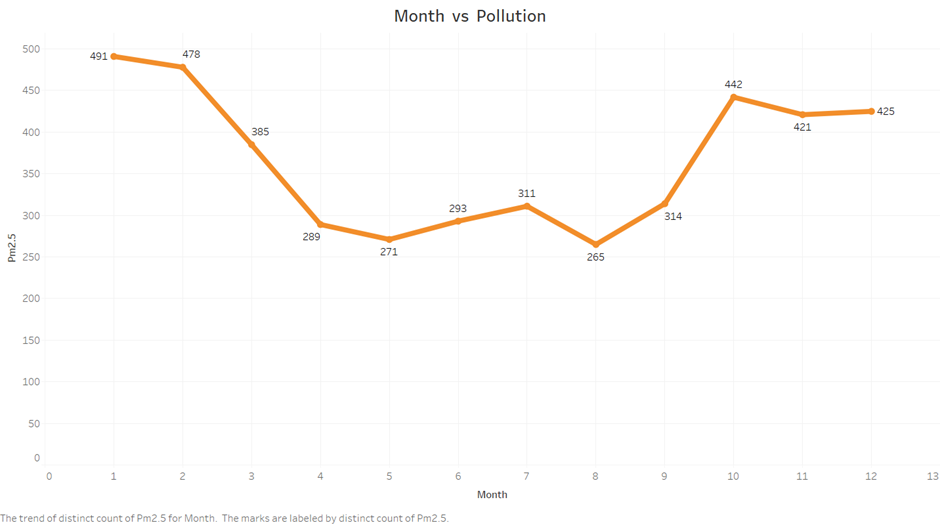
To get more information from the dataset and to build the best models we have created some visualizations in Tableau and found the following

**1. Monthly distribution of Target Variable and Regressors**



We want to know the distribution of each and every variable in this graph, as well as which variables are related to the target variable. We learned from this visualization that snow has an impact on pollution because the ice particles that makeup snow store various gaseous and particulate substances. We also learned that wind has a significant influence on pollution, which will be covered in forthcoming visualizations.

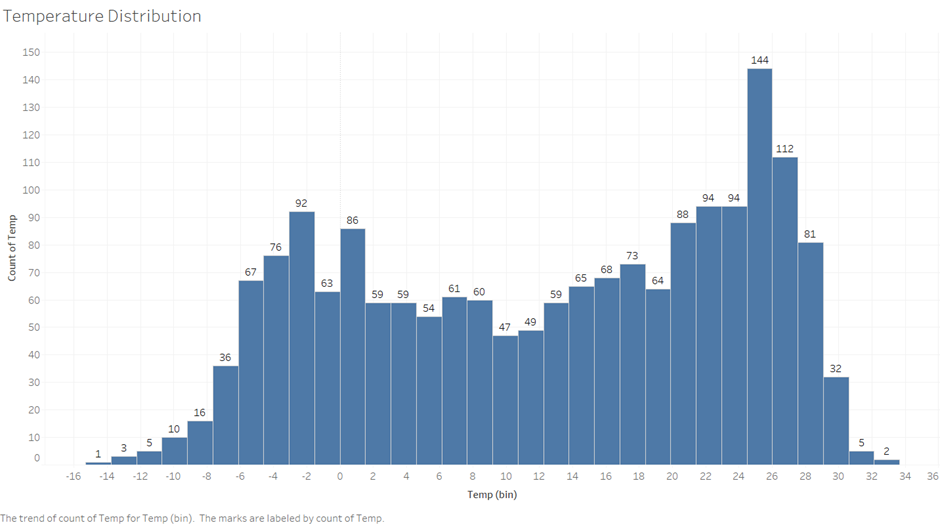
**2. Monthly pollution levels**



We've seen that pollution levels are higher at the beginning and end of each year, i.e., during the winter season, due to a convergence of natural occurrences that exacerbate air pollution. In cloudy conditions, pollution tends to be confined beneath the cloud. In this sort of temperature inversion, the warm air acts as a lid, concealing air impurities.

What impact does a temperature inversion have on pollution? Warm air acts as a cap over colder air, inhibiting vertical mixing and keeping cooler air closer to the surface. The inversion traps pollutants released into the air by vehicles, fireplaces, and industries near the ground, resulting in poor air quality.

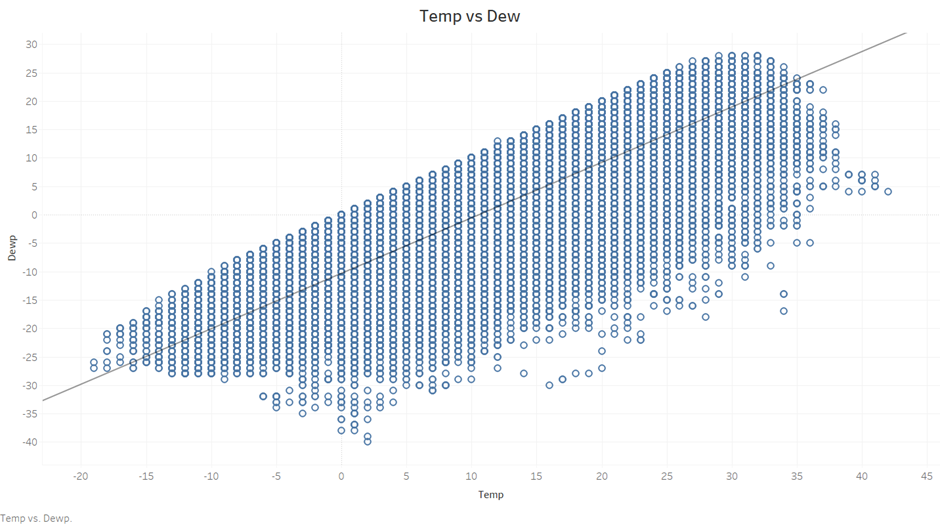
**3. Temperature Distribution**



The temperature in Beijing ranges from -16 to +36 degrees, and we can see the temperature ranges on the X-axis and the number of days with those temperatures on the Y-axis in the graph above. Regardless, the majority of temperatures range from -6 to +30 degrees, which is one of the causes of the high concentration of PM 2.5 in the environment.

## 4. Relation between regressors

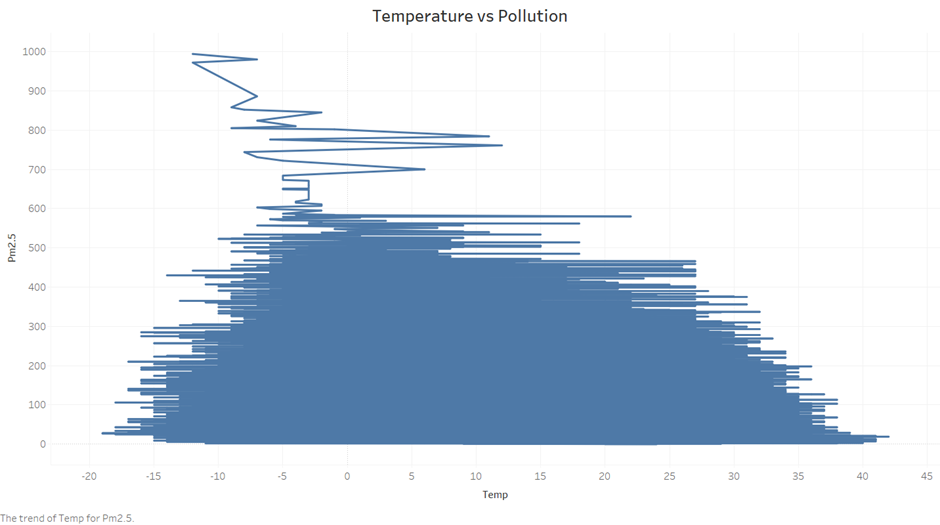
As we observed from the correlation plot that there is a correlation between temperature and dew, we wanted to investigate further, so we created a visualization between these two regressors, as shown below



Even though the dew point temperature is determined by pressure rather than the air temperature, the above graph shows that when the temperature is low, the dew point temperature is also low, and as the temperature rises, the dew point temperature rises as well. As a result, we can draw that these two factors have a positive relationship.

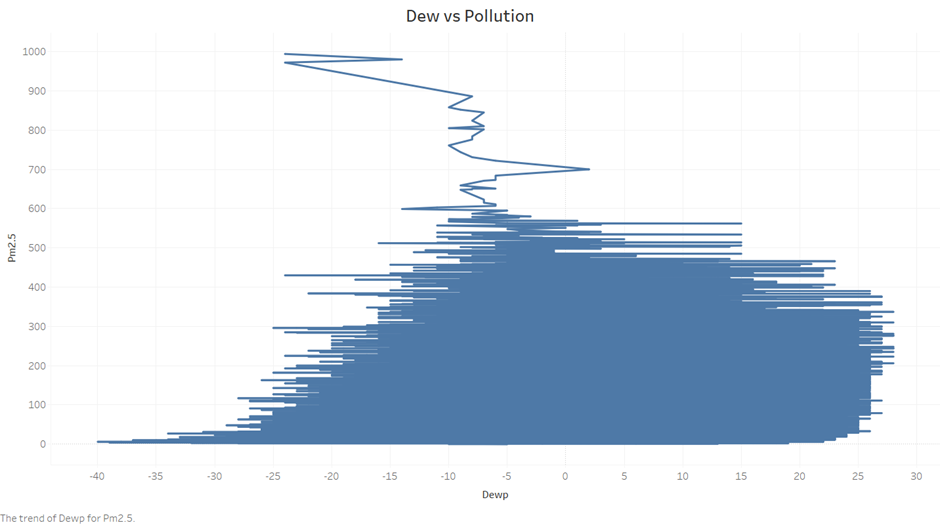
## 5. Relation between the output variable and regressors

**a. Temperature v/s Pollution**



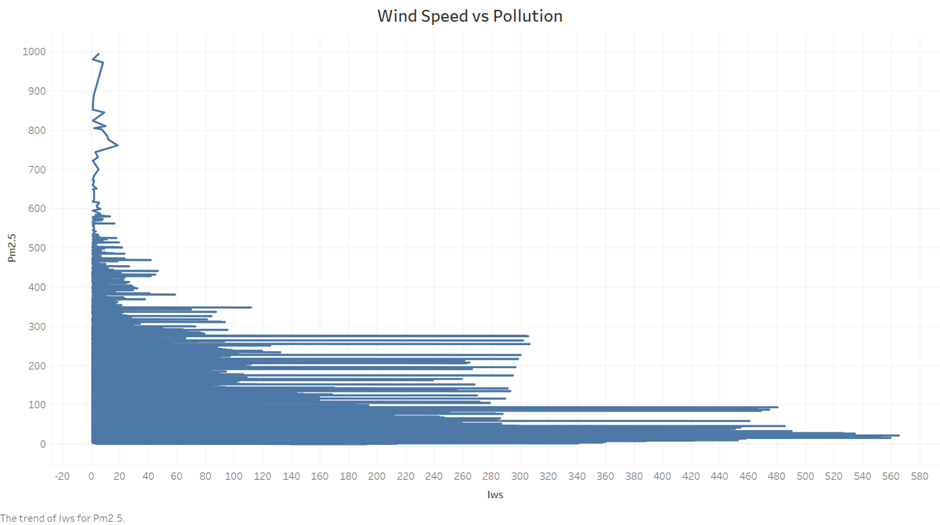
As we can see, when temperatures dip below zero, there are many pollution particles in the atmosphere, however when temperatures rise, the pollutant particles decrease because pollutants become lighter and scatter in the atmosphere as temperature rises. As a result, the concentration of contaminants in the air is reduced.

**b. Dew v/s Pollution**



We can surmise from the graph that pollution levels are low when the dew point temperature is low and there are many pollutant particles in the atmosphere; however, when the dew point temperature is moderate (-25 to -5 degrees) and the dew point temperature rises, pollutant particles in the atmosphere decay slowly.

**c. Wind speed v/s Pollution**



When the wind speed is high, pollution levels are low, and when the wind speed is low, there are more pollutant particles. Higher wind speeds cause more air pollutants to be dispersed, resulting in lower air pollution concentrations in locations with higher winds. The air becomes more turbulent as the earth heats up during the day, causing air contaminants to scatter in the air.

# Predictive Modeling

## a. Teams Approach

As the team started to come up with ideas for modeling types and approaches, There are many methods for time series forecasting that we can consider, and there is no clear winner. The model you use should always be based on how your data looks and what you're trying to achieve. Some models may be more robust against outliers, but they perform worse than the more sensible models. However, depending on the use case, they may still be the best option.

To begin, we profiled our dataset with the SAS Time Series Forecasting System to determine whether there was any pattern or seasonality. Our dataset displayed no trend or seasonality, according to the findings of these analyses. This can be easily confirmed by looking at the time series.

Our dataset has 1826 records in it. We opted to exclude the last two months' data (Nov-2014, Dec-2014) from the dataset because we have data from January 2010 to December 2014. We will forecast pollution levels for these two months and compare them to actual pollution levels to evaluate how close our model is to the outcomes. Hold-out sample is considered for 30 time periods.

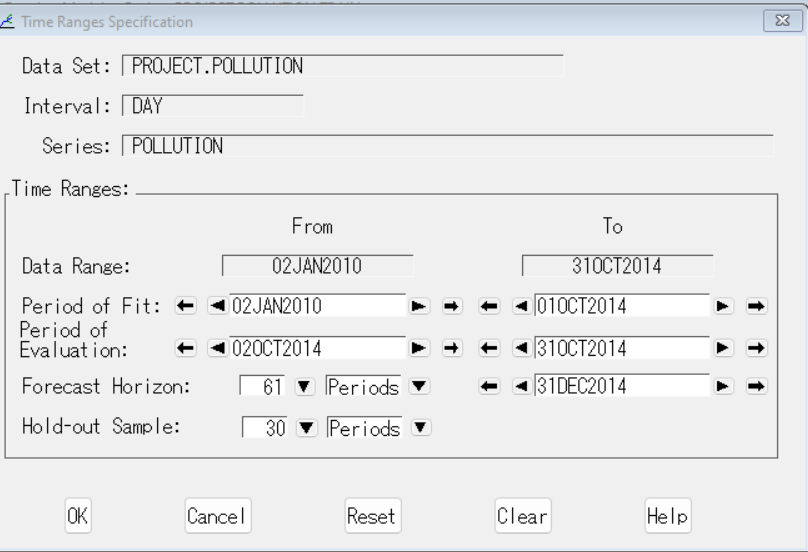


Fig: Time Ranges for the model training

## b. Models Used

We have performed an Exponential smoothing model. The Simple Exponential Smoothing (SES) approach considers the next time step to be an exponentially weighted linear function of previous time steps. In order to work properly, this approach requires that our time series be non-stationary (no trend or seasonality).

We have also performed Damped Trend, Linear, Winters Additive Exponential smoothing models for which we were getting high Mean Absolute Percent Error. We have moved on to apply ARIMA Models. Seasonality, trend, and noise are the three factors utilized in an ARIMA model to model the major elements of a time series. These values are denoted by the letters p, d, and q.

p is the parameter related to the auto-regressive element of the model, which integrates past values i.e. lags of the dependent variable. If p is 5, for example, the predictors for x(t) will be x(t-1)....x (t-5).

Number of Differences (d): d is a parameter in the integrated part of the model that controls how much differencing should be applied to a time series.

The number of MA (Moving Average) terms (q) represents the size of the model's moving average lagging forecast errors in the prediction equation. If q is 5, for example, the predictors for x(t) are e(t-1)....e(t-5), where e(i) represents the difference between the moving average at ith instant and the actual value. In ARIMA, non-stationarity series will require a level of differencing (d) >0. Using PACF, ACF plots of ARIMA, select the lag values for the Autoregression (AR) and Moving Average (MA) parameters, p and q, respectively. Also, we have tried including events but it did not show any effect on MAPE.

We have duplicated the model and then tried different q values. Then, we have attempted to fit a linear trend to these models adding dew, pressure, snow, rain, wind speed as regressors and temperature as dynamic regressor with order 1 in the numerator. We found the lowest MAPE value with this model, hence we consider it to be the best.

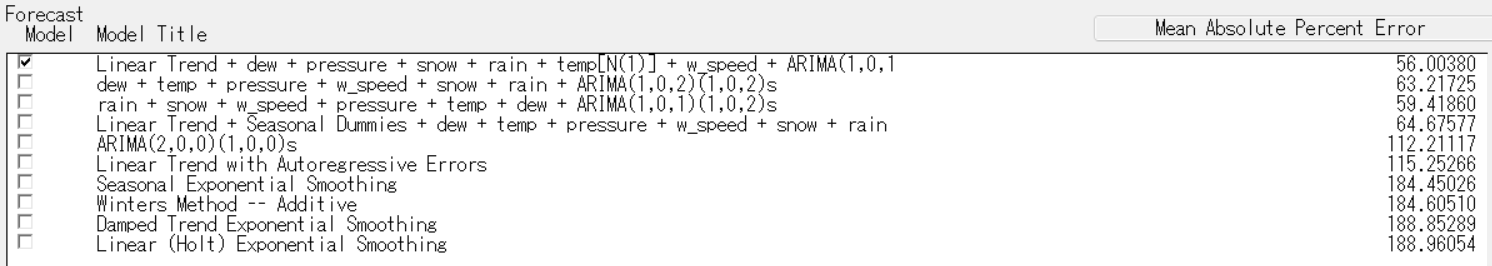


Fig: Models used for the forecasting

The below figure shows the forecast levels of pollution for the last two months.

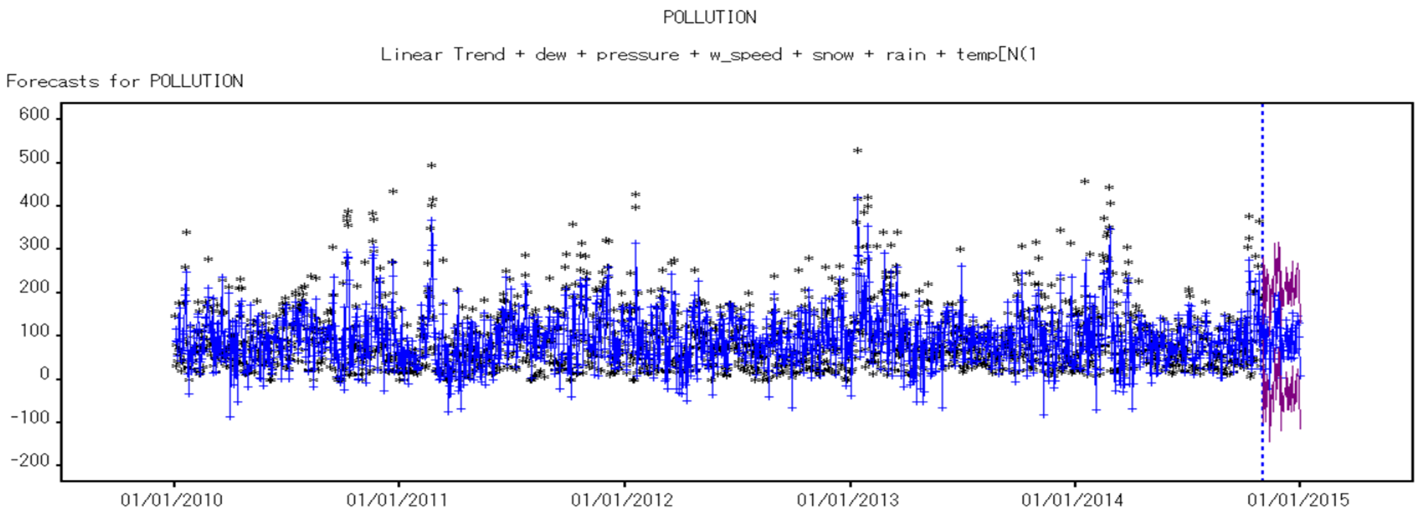
****

Fig: Forecast of best models

We have exported the data of the top-4 models into a csv file and then used tableau for getting the visualizations of the metrics for these models.

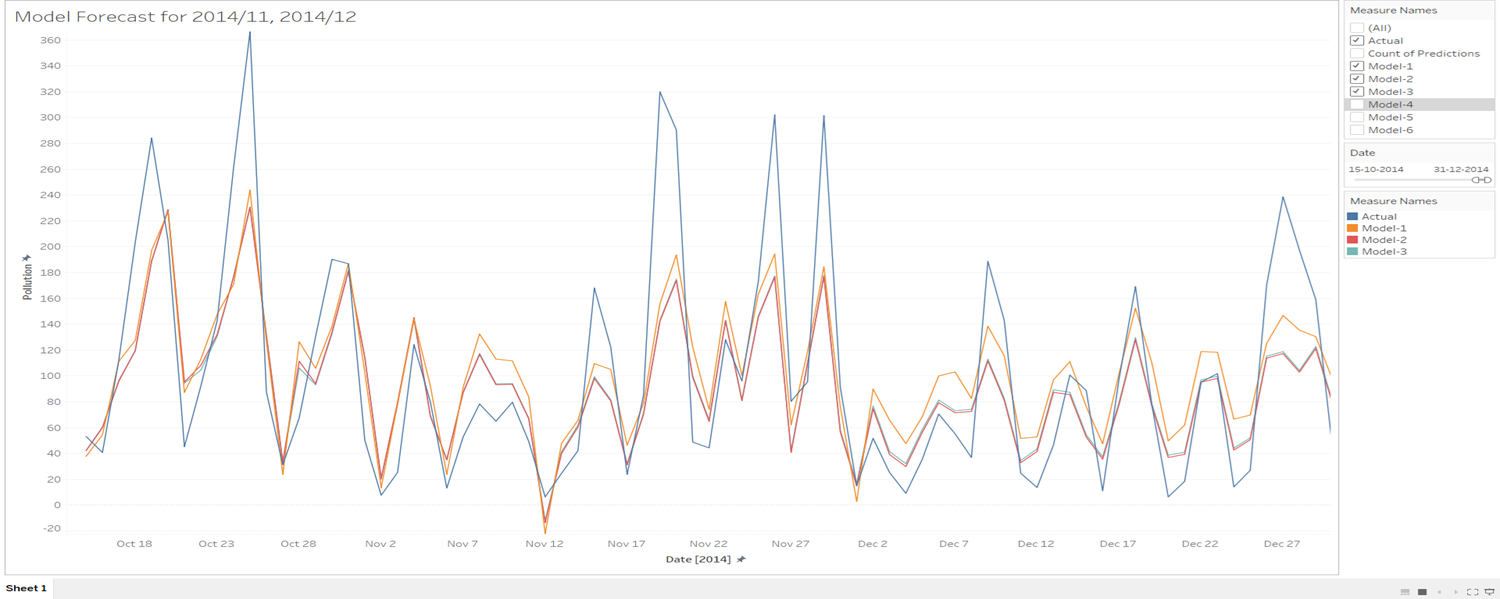
****

Fig: Model forecast for the two months

The blue line in the figure above represents actual pollution levels, while the orange line indicates our best model's expected pollution levels. We can see that there will be high pollution levels on November 18 and low pollution levels on November 12, as predicted by our model.

## c. Models Comparison

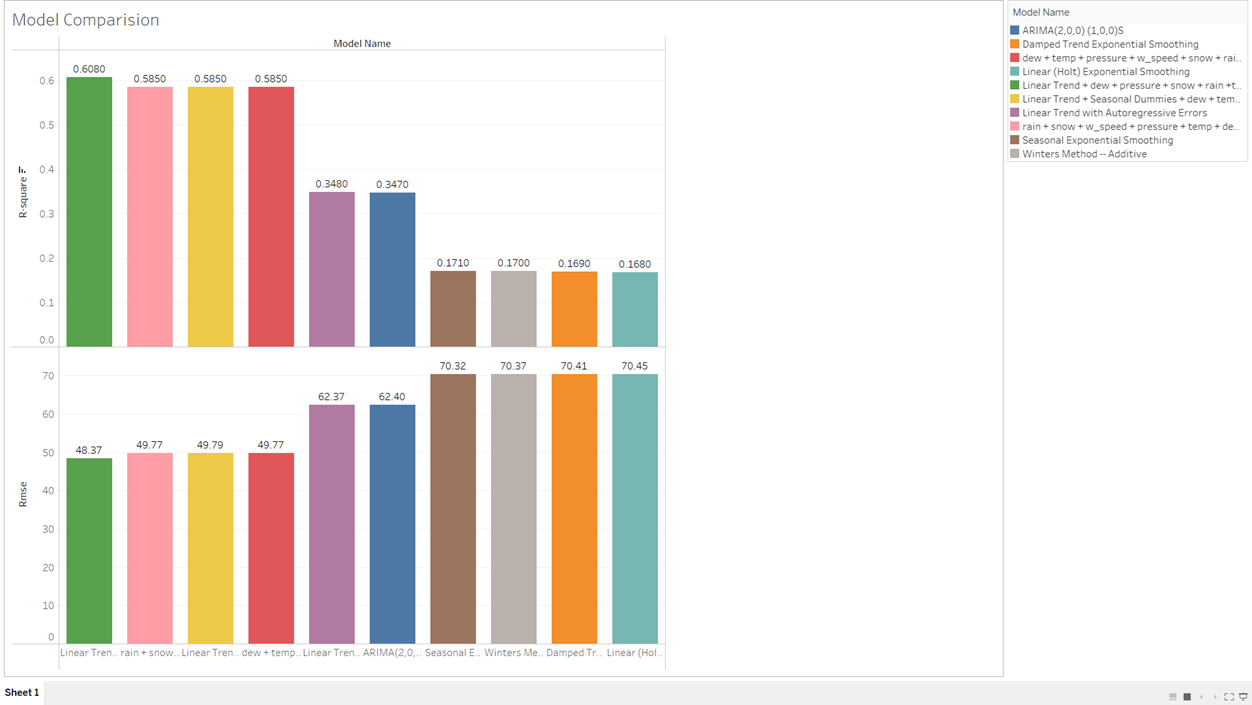
****

Fig: Model comparison

Model comparison is performed on all the models and we can see that our best model has the highest R-square value of 0.6080 and RMSE of 48.37%.

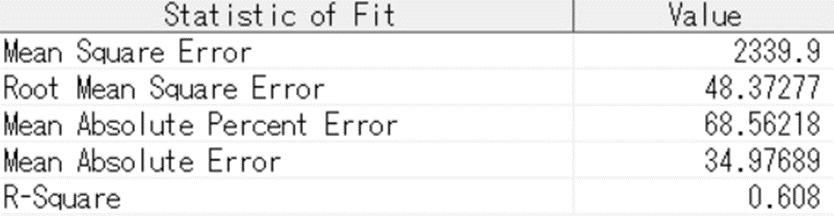
****

Fig: Statistics of fit

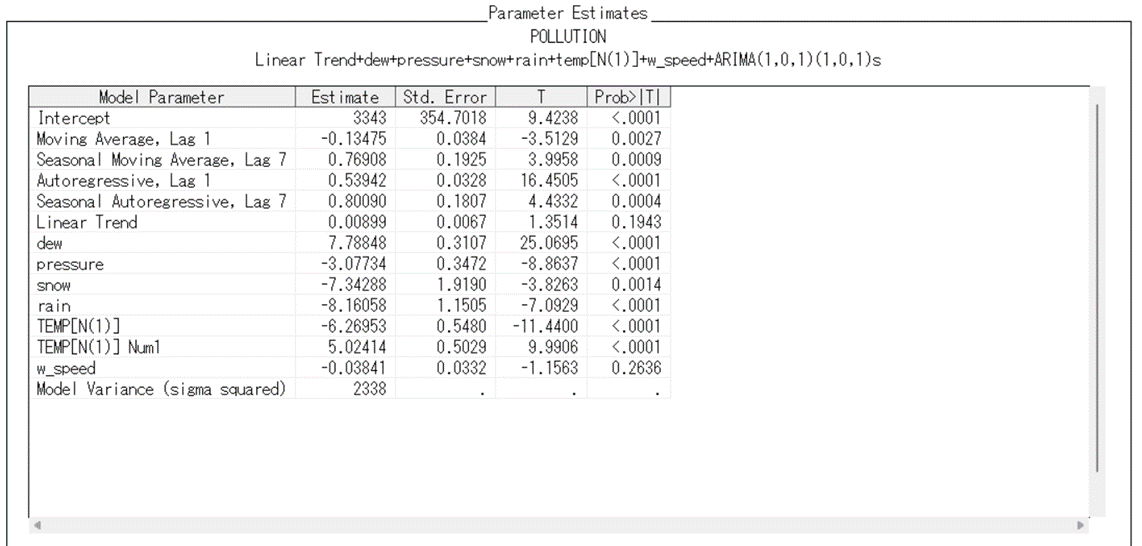
****

Fig: Parameter estimates

We can see that the Parameter estimates are high for dew and temp with 7.78 and 5.02 respectively in our best model which shows they play a major role in estimating the pollution levels. These are least for rain with -8.16.

# Findings

* Based on past data, we were able to Forecast the PM2.5 levels of next month that could be used to take necessary precautions for that coming period.
* Dew and Temperature has higher parameter estimate in our best model, It is evident that the dew and temperature plays major role in estimating pollution
* From this, we can observe that the pollution levels are high during Jan, Feb, Oct, Nov, Dec which is winter, as there will be much moisture content in the atmosphere. (increased after 2013)
* Highest Pollution rate is observed in January every year
* We also observed that when the wind speed is low, the pollution is noted to be high, which can be because pollution causing particles are not being spread out of the city
* Reducing factory emissions during this period to balance the pollution levels during this period helps safeguard the disasters caused by pollution
* During winters, the government should come up with innovative ideas like odd-even rules in order to restrict the transportation of vehicles, which in turn helps in reducing the pollution levels.

# Responsibilities of Each Team Member

| Data preprocessing using Python | Amit Anand  Akash deep Konda |
| --- | --- |
| Exploratory data analysis | Chaitanya Bandaru  Venkata karteek Thota |
| Modeling | Lohith krishna Pogala  Sathwik Pendyala |
| Presentation & Report | Amit Anand  Akash deep Konda  Chaitanya Bandaru  Venkata karteek Thota  Lohith krishna Pogala  Sathwik Pendyala |

# References

1. Data source (<https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>)
2. Air quality index and PM2.5 air pollution in Beijing (<https://www.iqair.com/us/china/beijing>)
3. Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating (<https://royalsocietypublishing.org/doi/10.1098/rspa.2015.0257>)
4. Multivariate Time Series Forecasting with LSTMs in Keras by machine learning mastery (<https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/>)
5. Analysis and Forecast of Beijing’s Air Quality Index Based on ARIMA Model and Neural Network Model (<https://www.mdpi.com/2073-4433/13/4/512/html>)