



# **SAFETY ANALYTICS**

## **Great Step 2019**

15.10.2019

Magnetite

Naman Paharia

Shubham Ekapure



# PREFACE

Coal has long been a reliable source of energy, but it comes with tremendous costs because it is incredibly dirty. The same chemistry that enables coal to produce energy—the breaking down of carbon molecules—also produces a number of profoundly harmful environmental impacts and pollutants that harm public health. The coal mined from coalmine results in suspension of particulate matter of different sizes in the air. These particles cause air contamination and can be dangerous to health. Air safety plays a crucial role in health, safety and security of mankind and ecology.

In the content followed we see various statistical results derived from the data obtained from CAAQMS. We observe the way in which emission vary over a month, week and a day. With such data, we can monitor air pollution on a real-time basis, and predict the weather conditions.

The extraction of coal is a regular process such that the emission data follow a rough pattern over each day. We try to predict the relationship between the parameters so that we can take effective steps to prevent pollution at the proper time and proper place.



# INDEX

## The Features

### Descriptive Statistics

- Measure of Central Tendency and Standard Deviation

- Coefficient of Variation

### Correlation between PM10 and the Data

- Correlation

- Graphical variation

### Trends of PM10 and PM2.5

### Linear Model for Particulate Matter

- Establishing models

- Statistical tools

- Observation

# The Features

**PM10 and PM2.5:** Particulate Matter are microscopic solid or liquid matter suspended in the air. The integer refers to its size in micrometre ( $\mu\text{m}$ ). The average value of its concentration is noted down. PM10 can cause short term respiratory problems whereas PM2.5 can be more dangerous and can lead to long term problems.

**To Date:** It gives us the time and date for a particular measurement, for time-series analysis of data we convert it into the index. This can be further used to find the variation for different time-periods. Statistics like mean, median and mode are mostly time-series dependent.

**Wind Speed:** The hourly measurement of the wind's speed is given by this feature. As a feature of CAAQMS, it is used to predict the direction in which the particulate matters will flow i.e. towards which city.

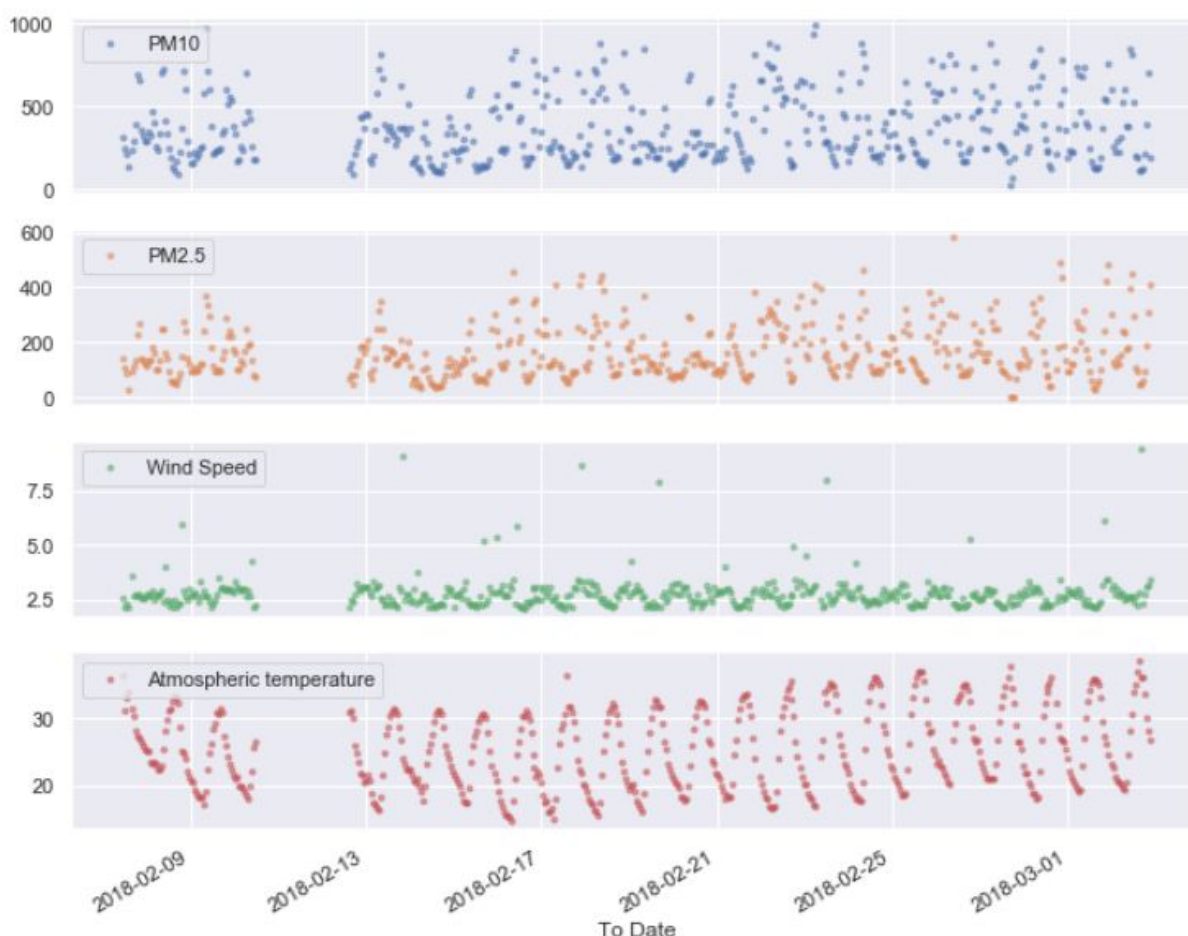
**Atmospheric Temperature:** The hourly measurement of the atmospheric temperature is given by this feature.

# Descriptive Statistics

## MEASURE OF CENTRAL TENDENCY AND STANDARD DEVIATION

Here we measure the mean, mode, median and standard deviation for the variables. But first, take a look at their behaviour over a specified time interval.

The variation over a month for all the variables represented by scatter plot.



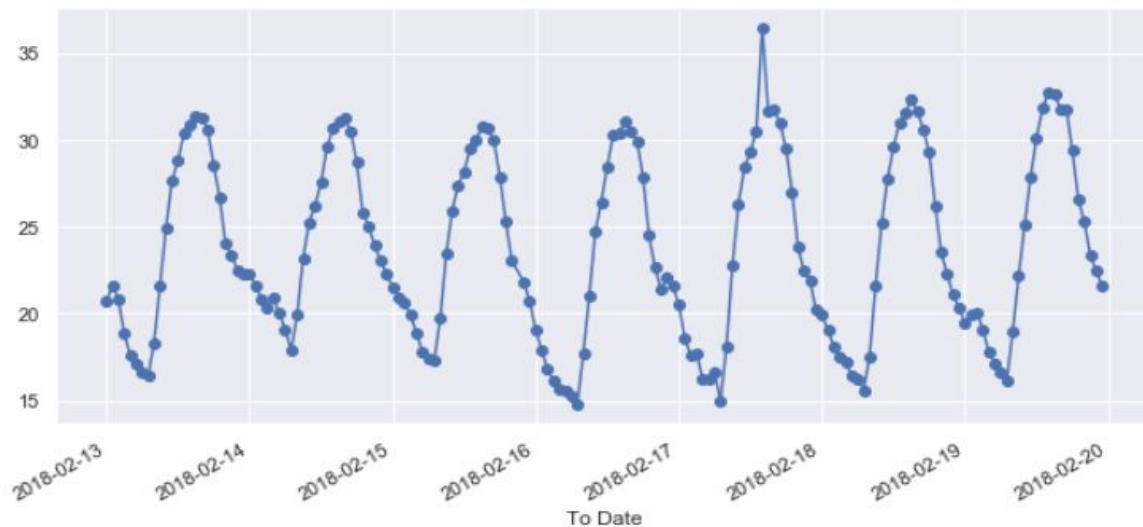
The features PM10 and PM2.5 aren't stationary i.e. their behaviour isn't exactly repeating over some time period. But wind speed and atmospheric temperature are clearly stationary and they follow a trend. Hence the mean for temperature and wind speed doesn't vary with time but PM10 and

PM2.5 has moving mean if we reduce the time window for a smaller time-period. We will see the behaviour of the mean for different time periods further. Also we can see a clear gap between the figures because the data for 11/02 and 12/02 aren't fully present.

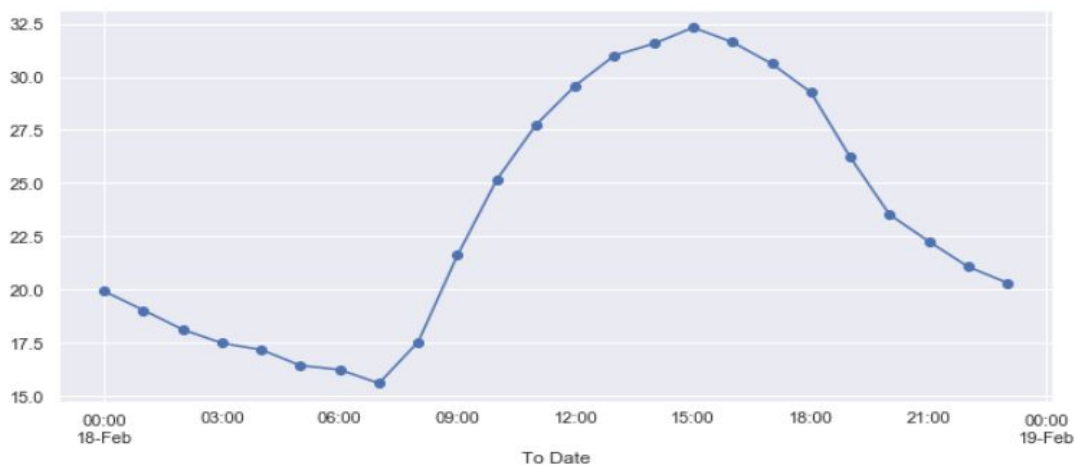
## Atmospheric Temperature:

### WEEKLY VARITAION

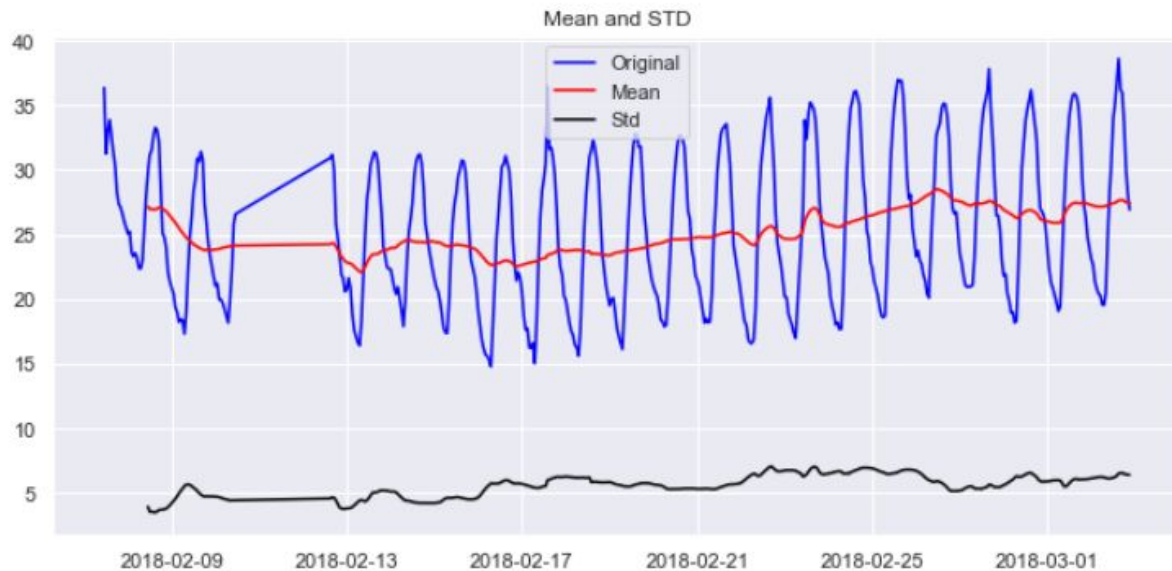
```
: ax = data.loc['2018-02-13':'2018-02-19', 'Atmospheric temperature'].plot(marker='o', linestyle='--')
```



```
0]: ax = data.loc['2018-02-18', 'Atmospheric temperature'].plot(marker='o', linestyle='--')
```



The plot above shows variation over a day.



*Clearly we have nearly constant rolling mean for **temperature** and for **wind speed** but not so for **pm10/2.5**.*

We can clearly see that the data for each day repeats itself. Hence we calculate a rolling mean for a window of 24 hours and the mean and the standard deviation is nearly stationary. As expected the temperature reaches a maximum by noon and then drops sharply. From the last plot we see that the mean lies somewhere between 25 and 26 and STD lies near 6.

By plotting a histogram and box plot for the temperature data we get the idea for the mode and median. The mode temperature is 19°C because its late winter and temperature tends to be lower for most of the times.

The measures of Central tendency are concisely given below-

MEAN = 25.34571°C

MEDIAN = 24.715°C

MODE = 19.06°C

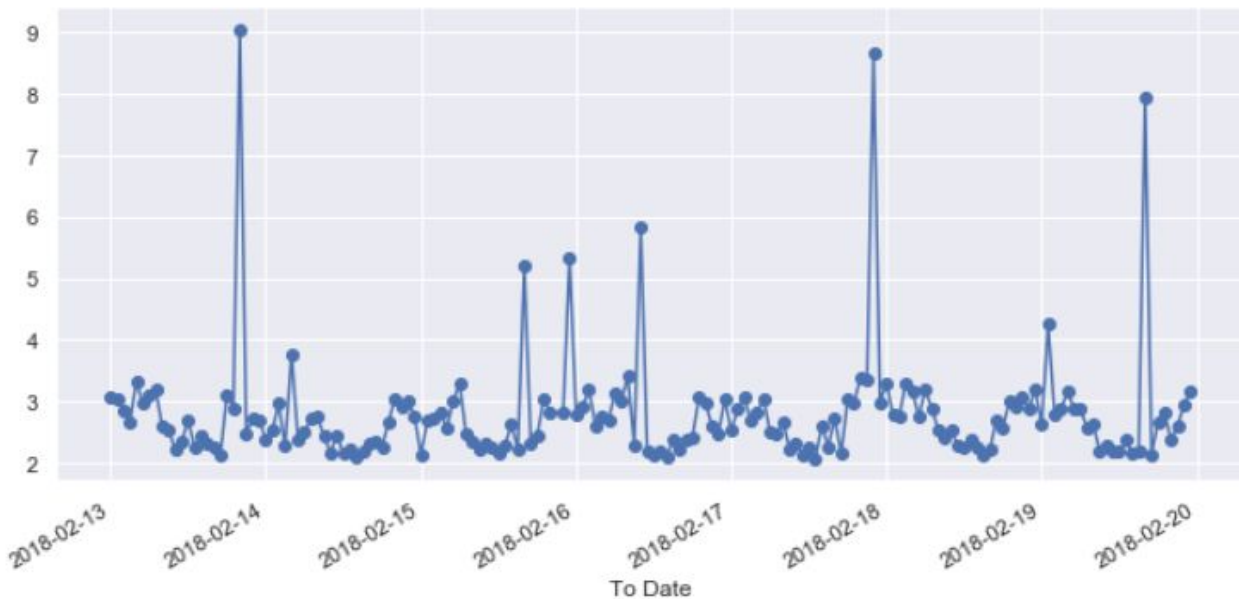
STD = 5.793065°C



## Wind Speed:

We see the variation over a week and a month followed by its rolling mean over a window of 24 hours.

### Weekly Variation

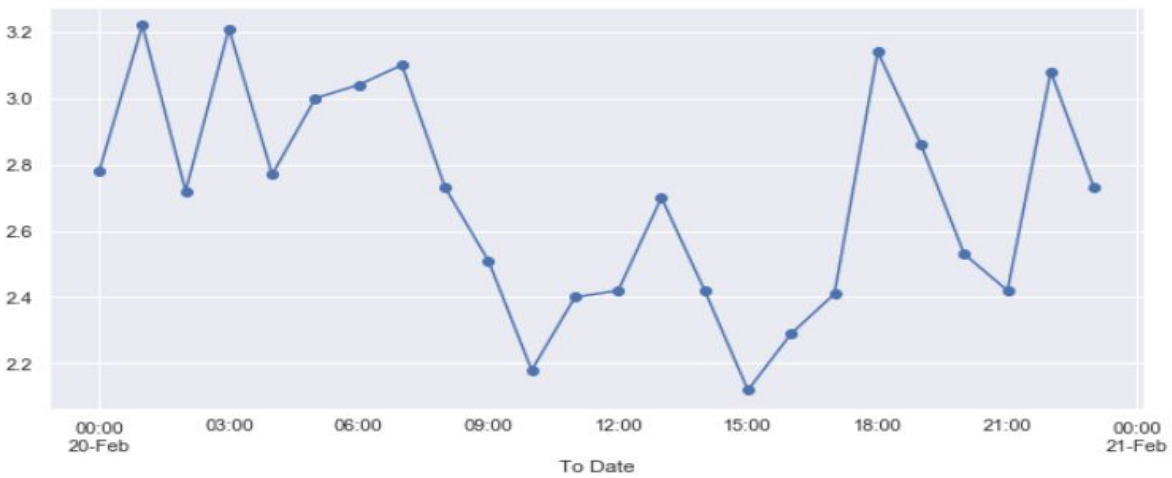
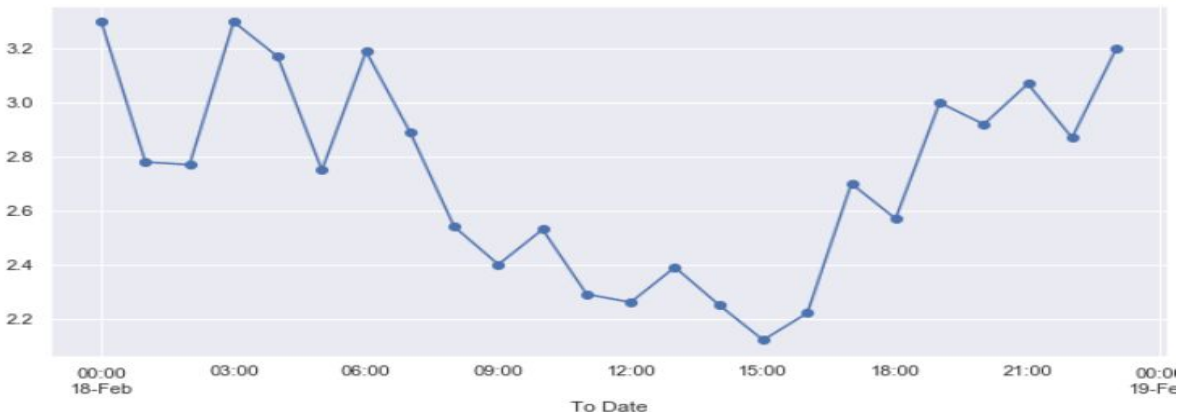


We can't see any repetition for the data for a week. Wind is a weather phenomenon as so it varies pretty randomly yet it is periodic for seasons but we have data for a single month.

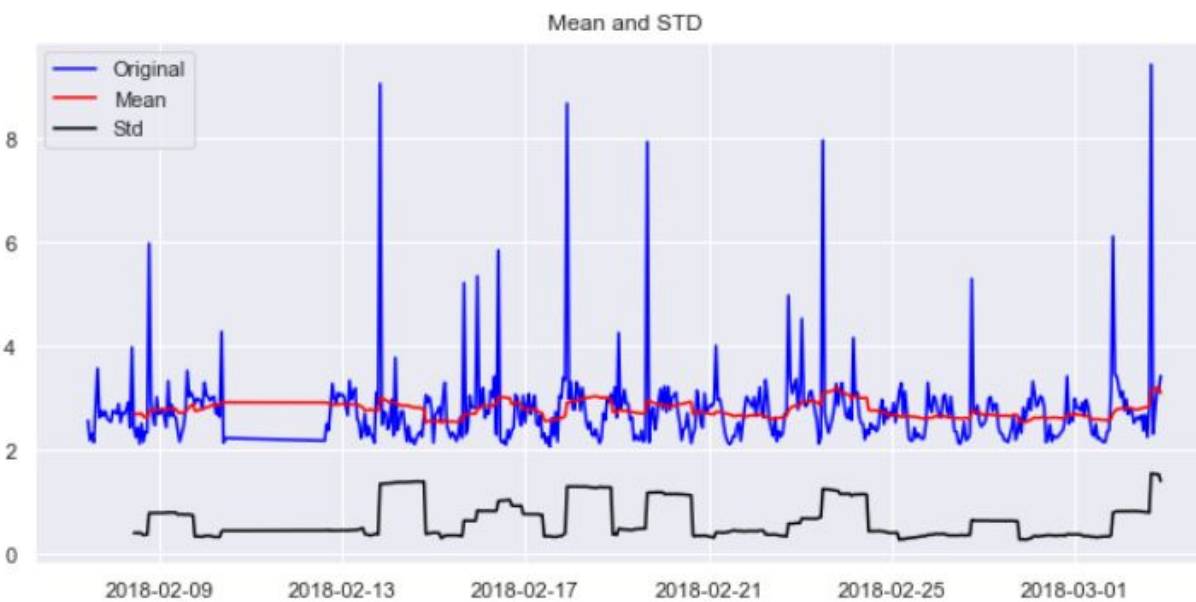
### Variation for two random days


The speed data is certainly not repeating itself on a daily basis as said above.





## Rolling mean and STD on a window of 24 hours





The mean for the data is somewhere around 3 m/s but it doesn't show increasing or decreasing trends hence the data is stationary to some extent.

The measures of Central tendency are concisely given below-

MEAN = 2.7696400000000008 m/s

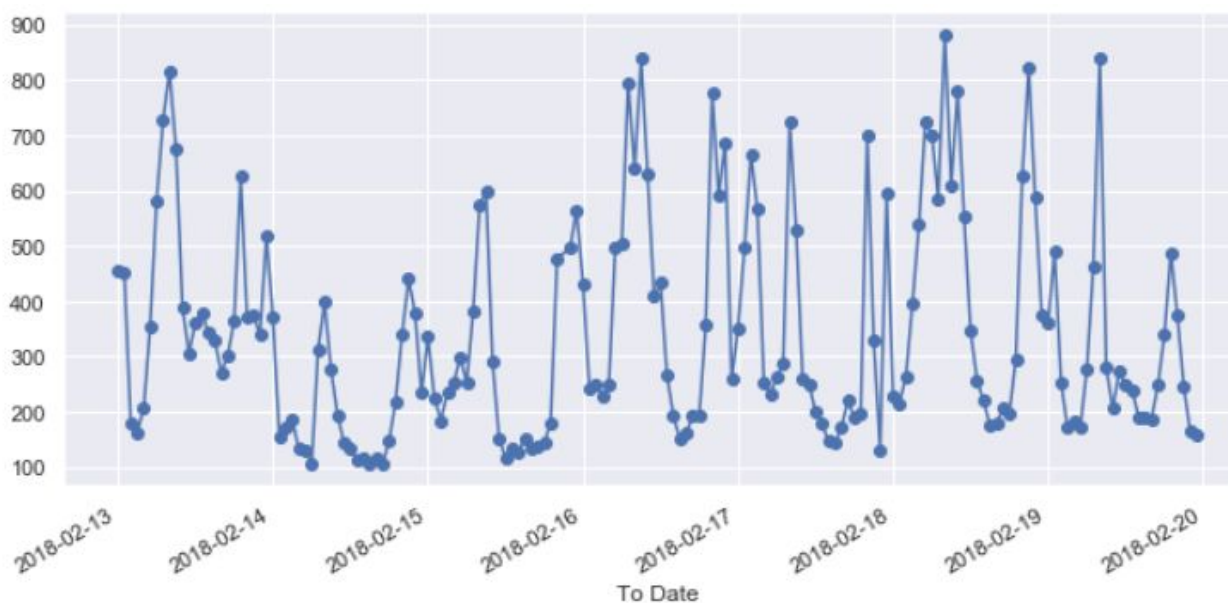
MEDIAN = 2.67 m/s

MODE = 2.18 m/s

STD = 0.7789918658459204 m/s

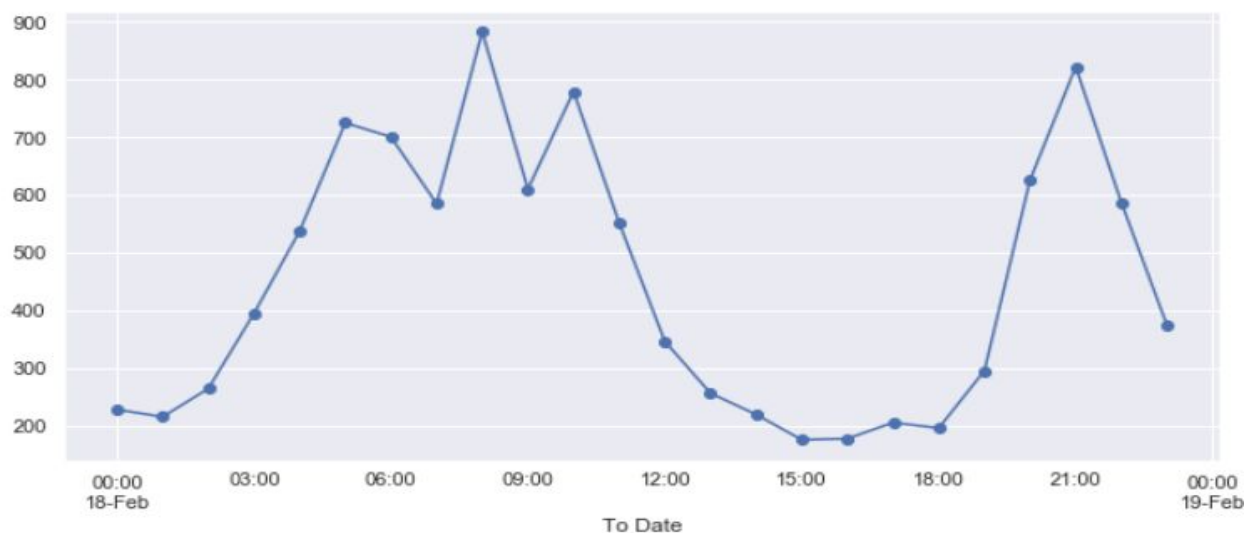
## PM10

### Weekly Variation

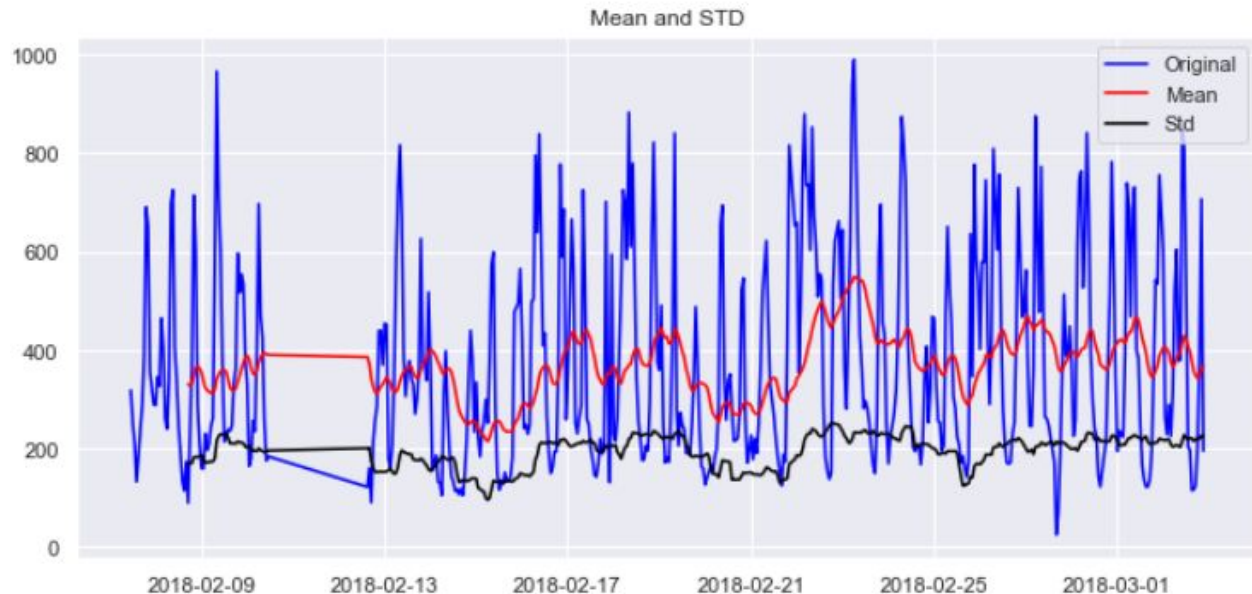


There is a trend which is being followed daily, we can see clear peak and trough each day at a fixed time. Although the magnitude may vary. These points of time are represented below.

### For a random day



The graph clearly tells us the time at which coal extraction occurs, as it corresponds to the peaks of PM10 on a daily basis.



The mean is varying on a window of 24 hours, mostly because each day a different amount of coal is mined. Also if we would have normalised the data for each day we would obtain a stationary mean and STD. The mean and STD line doesn't show increasing or decreasing trend. It's wobbling about a value.

The measure of Central tendency and STD is given below

MEAN = 367.96643999999975 unit

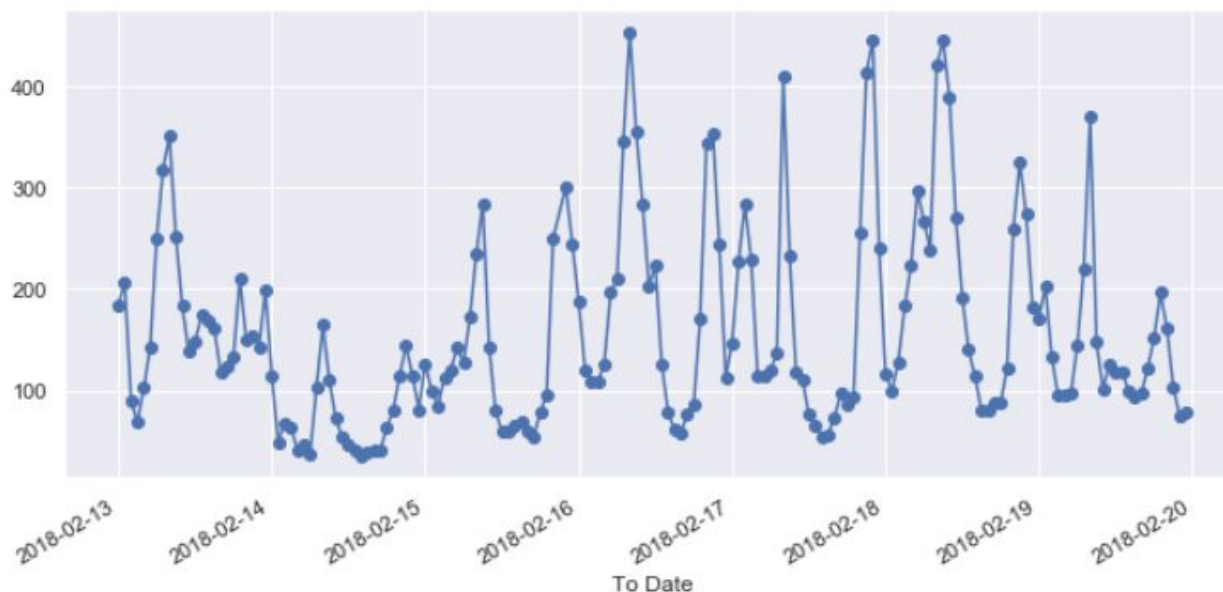
MEDIAN = 298.91999999999996 unit

MODE = 136.46 unit

STD = 203.41033986514688 unit

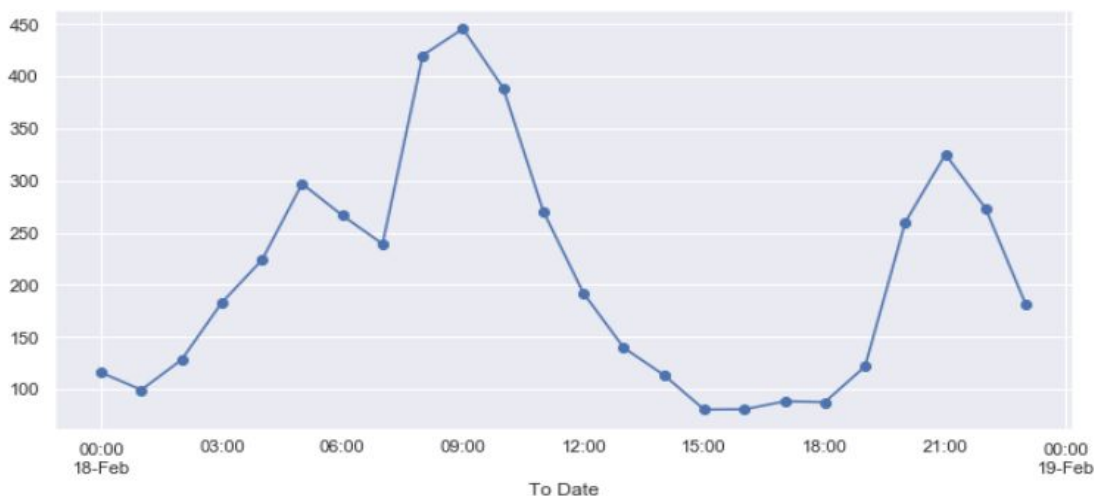
## PM2.5

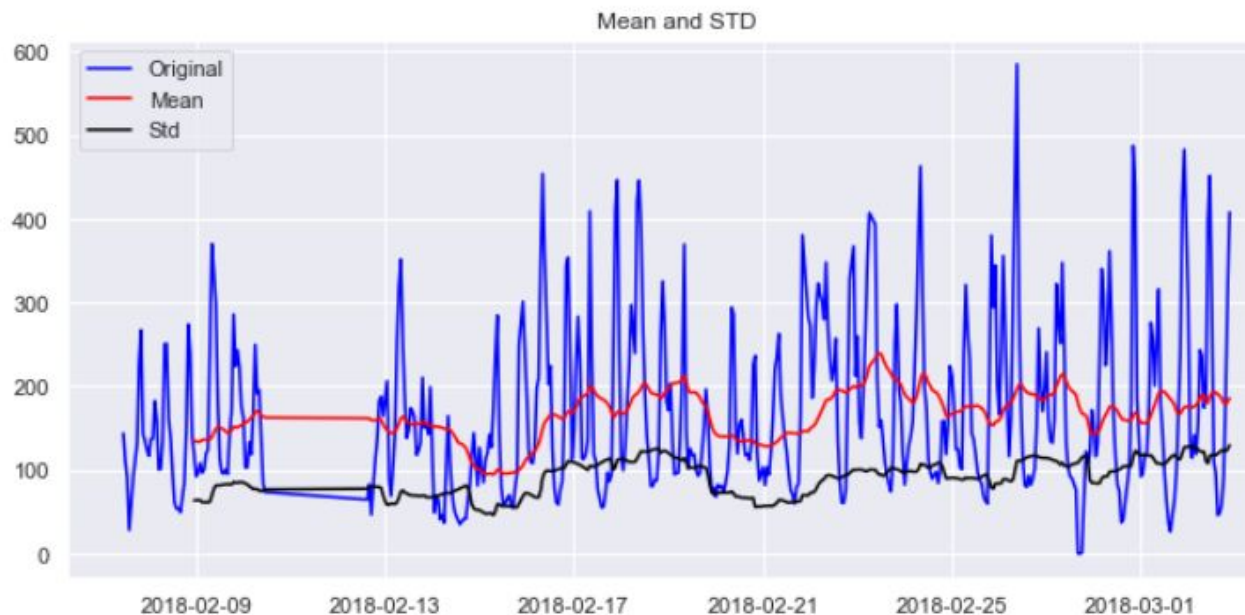
### Weekly Variation



There is a trend which is being followed daily, we can see clear peak and trough each day at a fixed time. It follows a similar trend as PM10, we will see it in further section.

### For a random day





This plot for mean and STD is similar to the one for PM10. The line for mean and STD oscillate about a fixed value because of the different magnitude of peak. The mode for

The measure of Central tendency and STD is given below

MEAN = 165.90211999999998 unit

MEDIAN = 136.685 unit

MODE = 116 units

STD = 97.24984175693604 unit

## Coefficient Of Variation

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean.

The coefficient of variation represents the ratio of the standard deviation to the mean.

$$CV = \frac{\sigma}{\mu}$$

The value of coefficient of variation for variables is given in the table below

Variable Name	Coefficient of Variation
PM10	0.553
PM2.5	0.568
Atmospheric Temperature	.228
Wind Speed	.281

The smaller value of CV for temperature tells us that the estimate is precise and it doesn't wobble a lot around its mean. On the other hand as we observed the plots of moving mean and STD for PM10 and PM2.5, there is a higher level of dispersion around the mean. We couldn't have told this by STD because it is dependent on the magnitude of the variable.



## Correlation between PM10 and the Data

**Correlation** is a statistical measure that indicates the extent to which two or more variables fluctuate together.

**Correlation coefficient** is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. We've used Pearson's correlation coefficient, which is given by

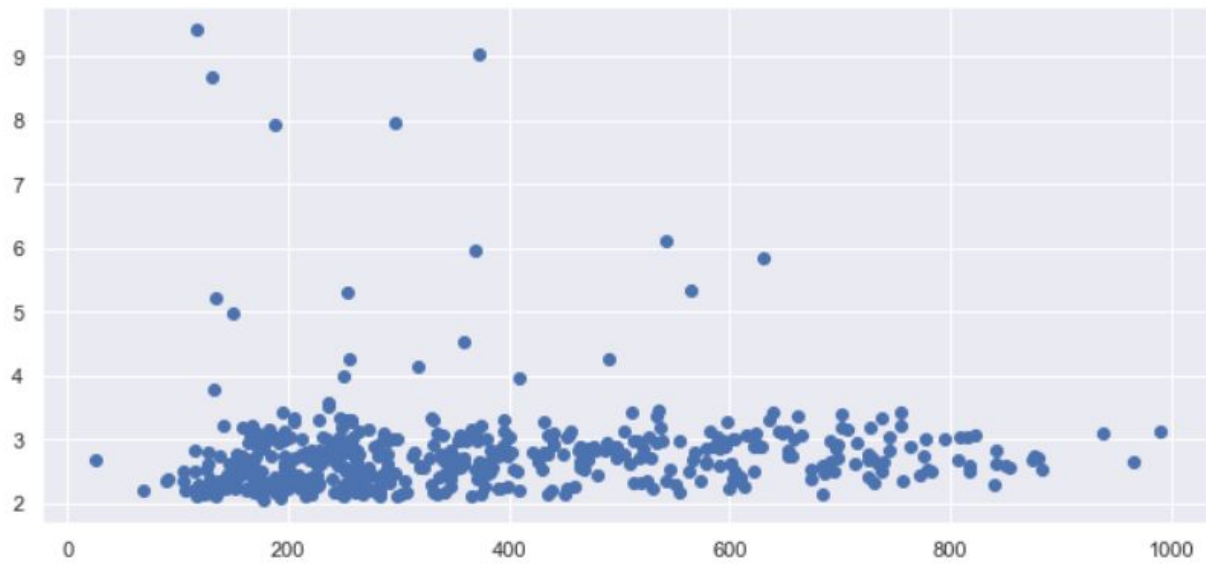
$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Where variables are  $X$  and  $Y$  with expected value  $\mu_X$  and  $\mu_Y$  and standard deviation  $\sigma_X$  and  $\sigma_Y$

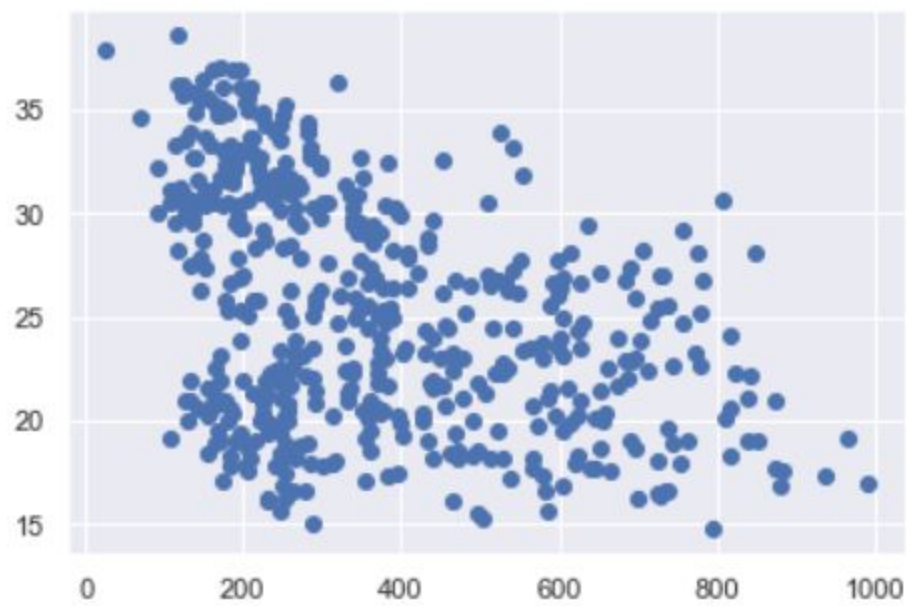
Applying the formula above we get:

**Correlation in PM10 and Atmospheric temperature** - correlation coefficient is **-0.4138**. As the coefficient is greater than 0, it shows a significant relation between PM10 and Atmospheric temperature and the negative sign shown the inverse relation between them.

**Correlation in PM10 and Wind Speed** - correlation coefficient is **0.0676**. These values are obtained at a 5% level of significance, because by default the level of significance is set to 5%.



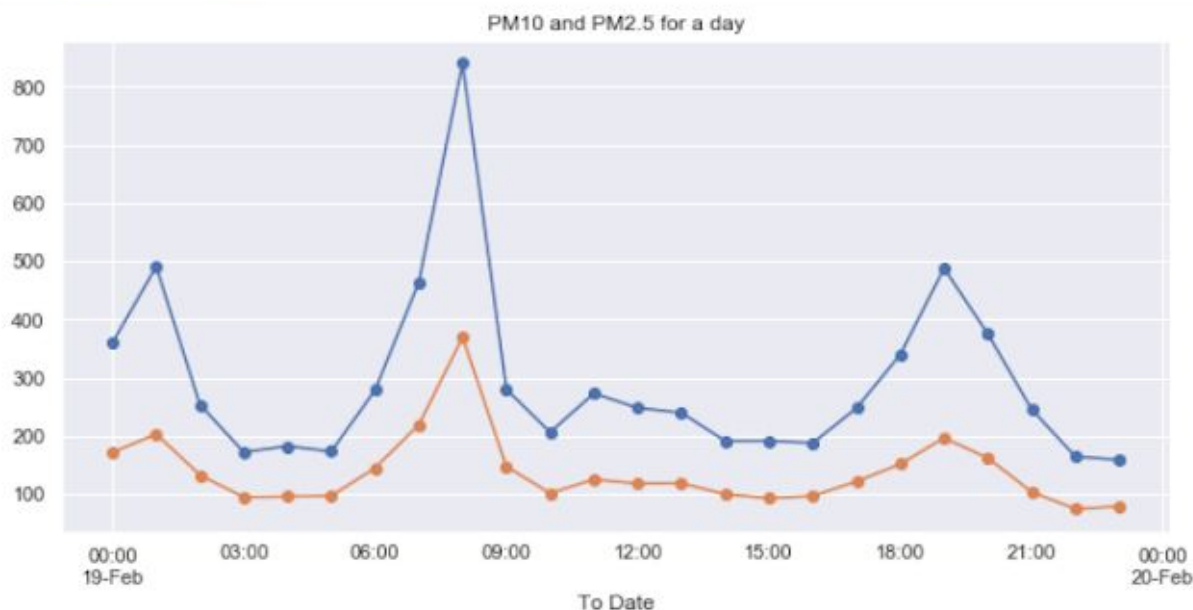
Scatter plot of PM10 vs Wind Speed



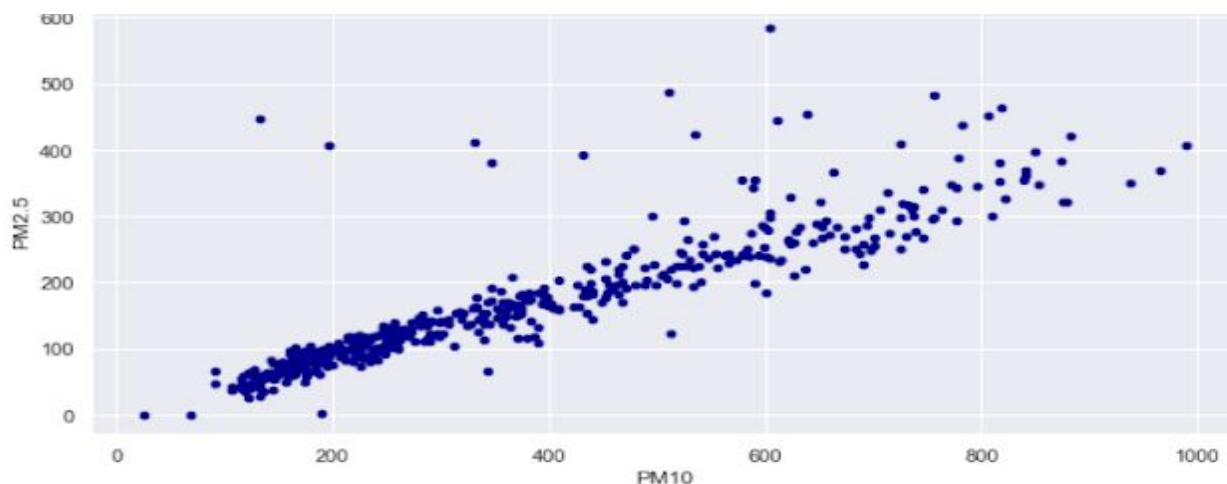
Scatter plot of PM10 vs Atmospheric Temperature


## TRENDS OF PM<sub>10</sub> AND PM<sub>2.5</sub>

Based on the given data we came up with the answer if PM<sub>10</sub> and PM<sub>2.5</sub> follow a similar trend, we used plots as well as metrics to determine the same.



This data has been extracted from the data over the entire time duration which could be easily seen in the ipynb notebook provided. Clearly we can see that the peaks and trough occur at the same instant time i.e. PM<sub>2.5</sub> follows a similar trend as PM<sub>10</sub>.





From the scatter plot we clearly see a linear relationship between PM10 and PM2.5 i.e. PM2.5 follows the same trend as PM10. This can be observed from the fact that when coal is mined both the effluents are emitted into the atmosphere.

Correlation in PM10 and PM2.5 - Correlation coefficient is **0.886**. This means that both the variables are dependent on each other through a relationship that causes PM2.5 to follow the same trend such as PM10.

# LINEAR MODEL FOR PARTICULATE MATTER

## ESTABLISHING MODELS

We have set-up linear relationship for the particulate matter with Atmospheric Temperature and Wind Speed. We have used SK Learn package from pandas to obtain relationship and take measurements for it's score. The full model for F-Test has been built using both the variables via multivariate regression.

## STATISTICAL TOOLS:

**F-Test** - **F-test in regression** compares the fits of different linear models. The smaller the value of a model the better it is.

$$F^0 = \frac{(SSR_R - SSR_{UR}) / q}{SSR_{UR} / (n - k - 1)}$$

Where, SSR is Sum Squared Error, R is reduced model, UR is full model, (n-k-1) represents the change in parameters.

**R-squared Score** - It is a statistical measure of how close the data are to the fitted regression line, it is the percentage of the response variable variation that is explained by a linear model.

$$\text{R-squared} = \text{Explained variation} / \text{Total variation}$$

**MSE Score** - Mean Squared Error basically measures the average squared error of our predictions.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

## OBSERVATION

FOR PM<sub>10</sub>

Statistical Tool	Full model	PM <sub>10</sub> & Temp	PM <sub>10</sub> & Wind
R <sup>2</sup> Score	0.18	0.18	0.004
MSE Score	35325	35190	42776
F-Test	N.A.	-0.64	35

The F-Test Score for linear model with temperature is much better than with wind speed, hence temperature is a better feature for model fitting. This is because there is a very high increase in SSE for the reduced model when Temperature is removed i.e. model with Wind Speed. On the contrary, the reduced model with Temperature has pretty much the same SSE as the full one.

FOR PM<sub>2.5</sub>

Statistical Tool	Full model	PM <sub>2.5</sub> & Temp	PM <sub>2.5</sub> & Wind
R <sup>2</sup> Score	0.16	0.158	0.012
MSE Score	7164	7181	8427
F-Test	N.A.	0.393	29.4

The F-Test Score for linear model with temperature is much better here as well. The model performs very badly with Wind Speed as there isn't any linear relationship which can be confirmed from scatter plot presented earlier.