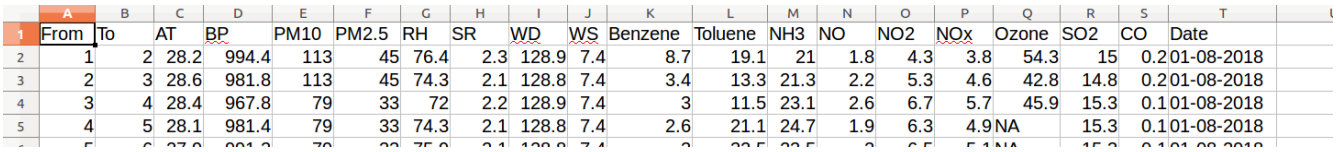
DATA DESCRIPTION

The dataset was obtained from the Central Pollution Control Board (CPCB) website. The dataset used in this project consists of Air Pollution data observed in the area Ashok Vihar in New Delhi from 1st August, 2018 to 3rd November, 2018 in time intervals of 15 minutes. The dataset consists of the following attributes:

Variable Abbr.	**Variable**	**Variable Type**	**Unit of Measurement**	**Data Type**
PM10	Suspended Particulate Matter	Pollutant	µg/m3	Continuous
PM2.5	Dust 2.5	Pollutant	µg/m3	Continuous
NO	Nitric Oxide	Pollutant	µg/m3	Continuous
NO2	Nitrogen Dioxide	Pollutant	µg/m3	Continuous
NOx	Oxides of Nitrogen	Pollutant	µg/m3	Continuous
SO2	Sulphur Dioxide	Pollutant	µg/m3	Continuous
CO	Carbon Monoxide	Pollutant	mg/m3	Continuous
NH3	Ammonia	Pollutant	µg/m3	Continuous
Benzene	Benzene	Pollutant	µg/m3	Continuous
Toluene	Toluene	Pollutant	µg/m3	Continuous
Ozone	Ozone	Pollutant	µg/m3	Continuous
AT	Atmospheric Temperature	Meteorological	Celcius	Continuous
BP	Bar Pressure	Meteorological	mmHg	Continuous
RH	Relative Humidity	Meteorological	%	Continuous
SR	Solar Radiation	Meteorological	W/m2	Continuous
WD	Wind Direction	Meteorological		Continuous
WS	Wind Speed	Meteorological	m/s	Continuous



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	From	To	AT	BP	PM10	PM2.5	RH	SR	WD	WS	Benzene	Toluene	NH3	NO	NO2	NOx	Ozone	SO2	CO	Date
2	1	2	28.2	994.4	113	45	76.4	2.3	128.9	7.4	8.7	19.1	21	1.8	4.3	3.8	54.3	15	0.2	01-08-2018
3	2	3	28.6	981.8	113	45	74.3	2.1	128.8	7.4	3.4	13.3	21.3	2.2	5.3	4.6	42.8	14.8	0.2	01-08-2018
4	3	4	28.4	967.8	79	33	72	2.2	128.9	7.4	3	11.5	23.1	2.6	6.7	5.7	45.9	15.3	0.1	01-08-2018
5	4	5	28.1	981.4	79	33	74.3	2.1	128.8	7.4	2.6	21.1	24.7	1.9	6.3	4.9	NA	15.3	0.1	01-08-2018

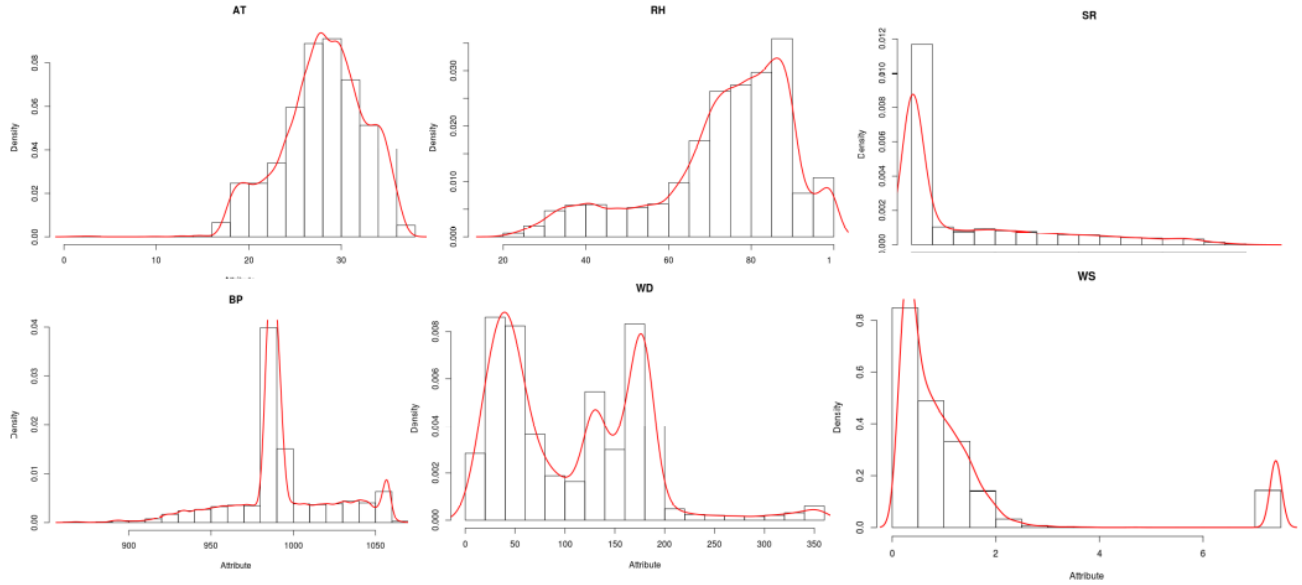
1. Data for the attributes Temperature and Vertical Wind Speed(VWS) were NULL throughout the dataset, and hence, the attributes Temperature and VWS were removed.
2. Data for those days that had no data for the key attributes like PM10, PM2.5, NO2 were removed.

DATA EXPLORATION

Summary statistics:

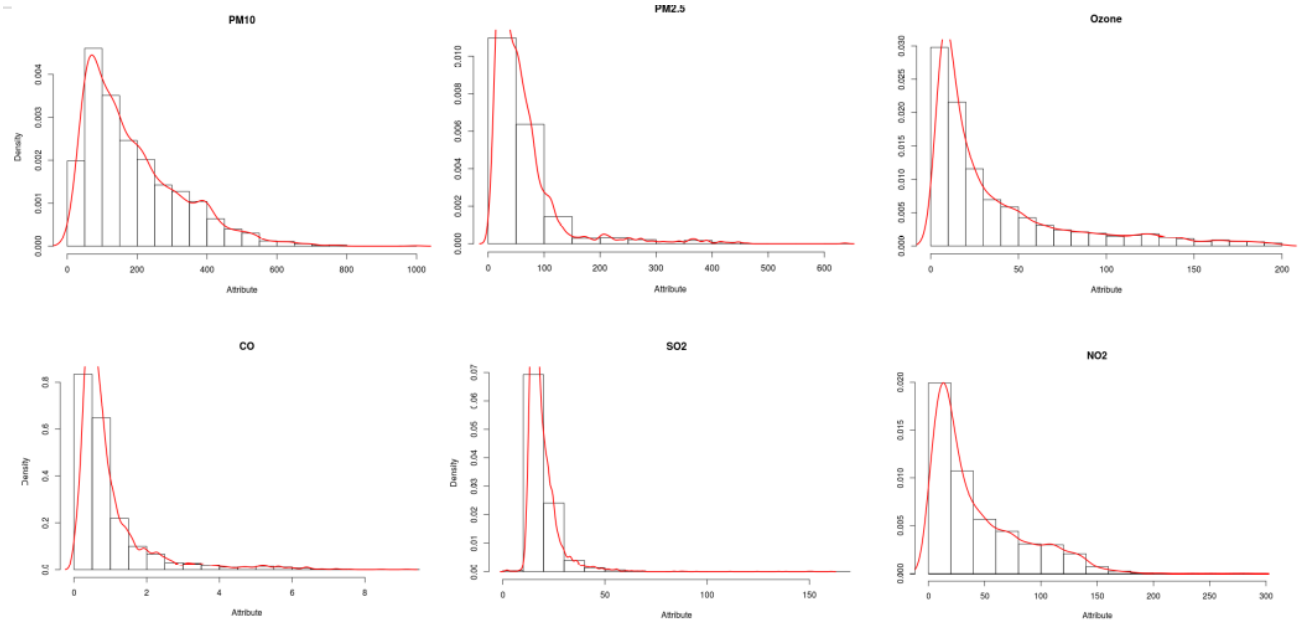
```
Console ~/Desktop/DM/
> summary(rawdata)
      From               To                AT              BP              PM10            PM2.5
 Min.   : 1.0   Min.   : 1.0   Min.   : 1.0   Min.   : 862.2   Min.   :  2.0   Min.   :  1.00
 1st Qu.:24.0   1st Qu.:24.0   1st Qu.:25.4   1st Qu.: 983.8   1st Qu.: 81.0   1st Qu.: 25.00
 Median :48.0   Median :48.0   Median :28.3   Median : 988.3   Median :149.0   Median : 46.00
 Mean   :48.4   Mean   :48.4   Mean   :28.0   Mean   : 992.5   Mean   :188.5   Mean   : 62.16
 3rd Qu.:72.0   3rd Qu.:72.0   3rd Qu.:31.2   3rd Qu.:1003.3   3rd Qu.:263.0   3rd Qu.: 74.00
 Max.   :96.0   Max.   :96.0   Max.   :37.4   Max.   :1067.5   Max.   :998.0   Max.   :637.00
                               NA's   :19     NA's   :13       NA's   :125     NA's   :92
       RH              SR             Temp               WD              WS          VWS
 Min.   :19.70   Min.   :  0.3   Mode:logical   Min.   :  0.40   Min.   :0.100   Mode:logical
 1st Qu.:66.88   1st Qu.:  1.2   NA's:8949      1st Qu.: 42.15   1st Qu.:0.300   NA's:8949
 Median :77.30   Median :  4.8                  Median : 97.10   Median :0.700
 Mean   :73.69   Mean   :135.9                  Mean   :106.26   Mean   :1.249
 3rd Qu.:85.70   3rd Qu.:234.3                  3rd Qu.:169.40   3rd Qu.:1.300
 Max.   :99.40   Max.   :839.0                  Max.   :360.00   Max.   :7.500
 NA's   :5       NA's   :5                      NA's   :194      NA's   :211
    Benzene         Toluene            NH3               NO              NO2              NOx
 Min.   : 0.1   Min.   :  2.8   Min.   : 8.00   Min.   :  0.10   Min.   :  0.10   Min.   :  0.30
 1st Qu.: 1.8   1st Qu.: 17.4   1st Qu.:19.00   1st Qu.:  2.80   1st Qu.: 12.50   1st Qu.: 10.80
 Median : 3.7   Median : 31.0   Median :23.60   Median :  5.50   Median : 27.40   Median : 19.20
 Mean   : 6.6   Mean   : 58.9   Mean   :27.58   Mean   : 26.23   Mean   : 43.07   Mean   : 43.19
 3rd Qu.: 7.9   3rd Qu.: 61.7   3rd Qu.:32.80   3rd Qu.: 16.70   3rd Qu.: 64.97   3rd Qu.: 46.83
 Max.   :47.9   Max.   :497.6   Max.   :89.20   Max.   :417.90   Max.   :285.10   Max.   :399.60
 NA's   :77     NA's   :84      NA's   :447     NA's   :941      NA's   :639      NA's   :621
     Ozone             SO2               CO                Date
 Min.   :  0.20   Min.   :  1.00   Min.   :0.0000   01-08-2018 :  96
 1st Qu.:  8.70   1st Qu.: 14.80   1st Qu.:0.4000   01-09-2018 :  96
 Median : 19.20   Median : 17.00   Median :0.6000   01-10-2018 :  96
 Mean   : 37.32   Mean   : 19.23   Mean   :0.9766   01-11-2018 :  96
 3rd Qu.: 48.60   3rd Qu.: 21.30   3rd Qu.:1.1000   02-08-2018 :  96
 Max.   :199.80   Max.   :160.70   Max.   :9.2000   02-09-2018 :  96
 NA's   :349      NA's   :755      NA's   :1303     (Other)    :8373
> View(rawdata)
>
```

Characteristics of the Meteorologic Attributes:



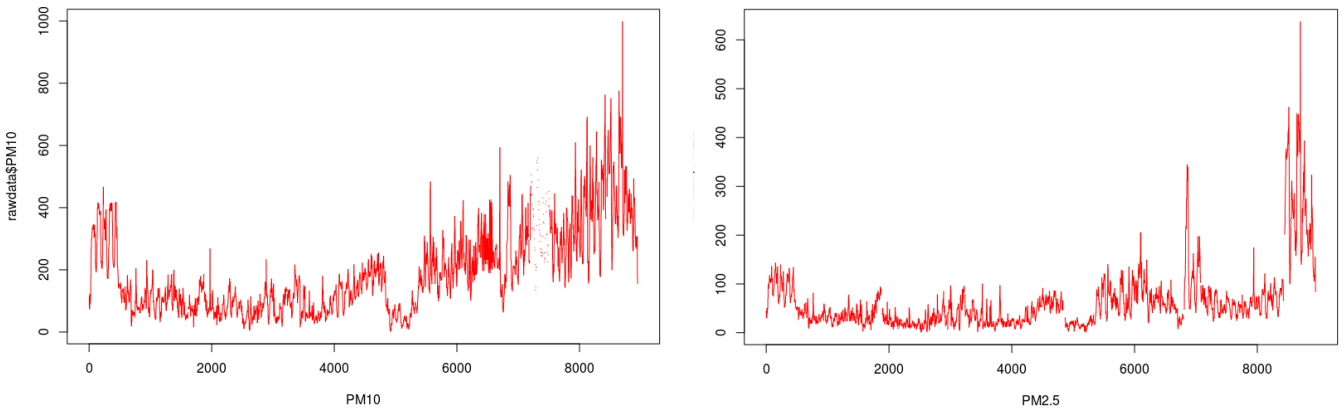
1. Most of the distributions are assymetric- Left or right skewed.
2. The attribute WD has a bimodal distribution.
3. The attributes AT, BP, and WS have outliers present.

Characteristics of the Pollutant Attributes:

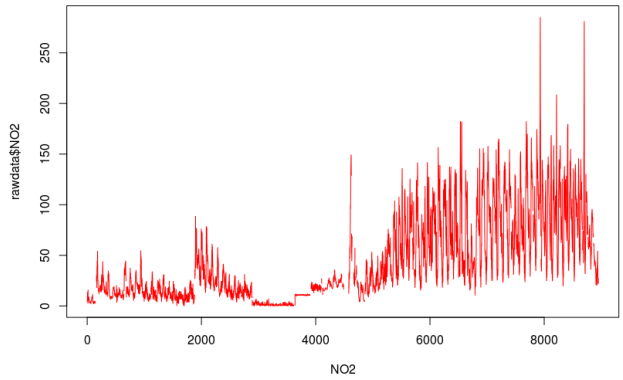


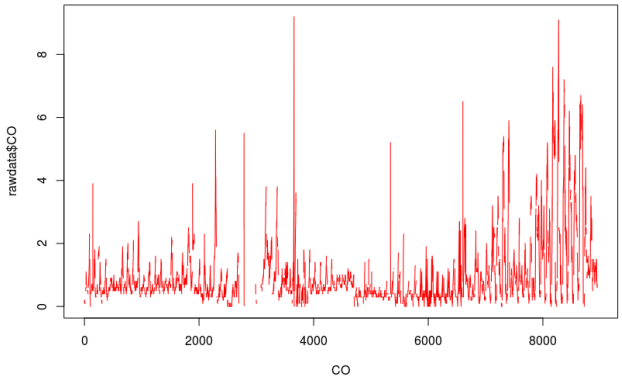
All the prime pollutants are skewed right.

Distribution of Features and Labels:

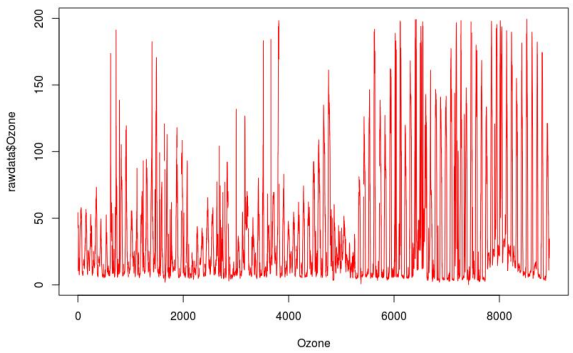


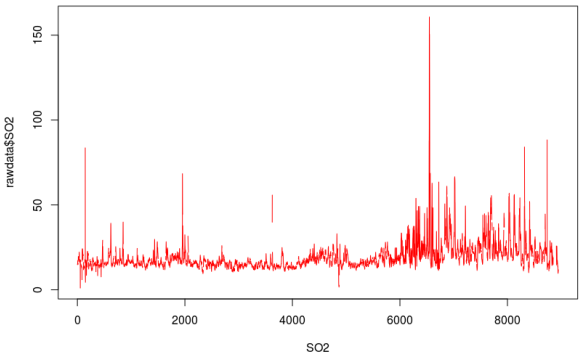
- PM10 level has stayed within the permissible level from August to the end of September, post which there is a dip during the end of September for a while, and then it seems to have increased at a constant rate.
- PM2.5 level has stayed more or less within the permissible range till the end of October, and during the end of October. It has gone considerably beyond the permissible range in the beginning of October, and has peaked up in the month November.





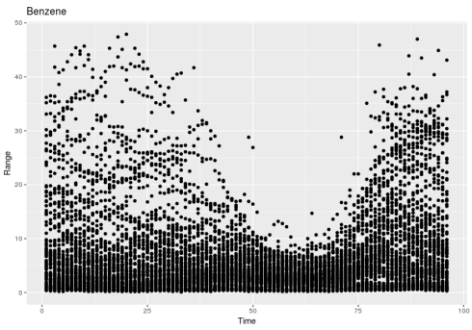
- NO2 level has also increased drastically in the middle of September.
- CO level has been sparking up from time to time for a very short while. However, CO level has frequently been high from the end of October.

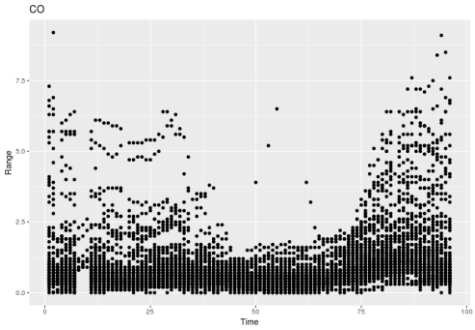


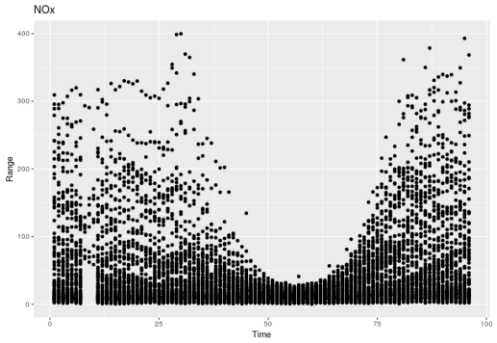


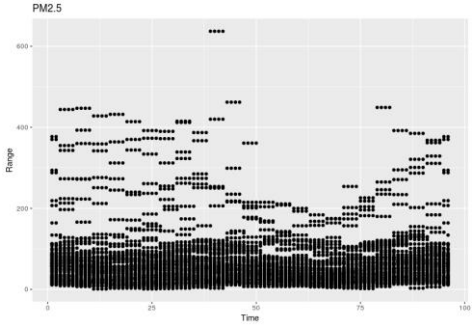
- Ozone level has been constantly high during the months October and November.
- SO2 level has been the same more or less, except during the month of October where it has peaked to a very high value, and then it falls down during the month of November.

We can infer that there is a general trend in most (if not all) pollutants to increase moderately from the end of September or the beginning of October. (Autumn). In most cases, it only increases after that. We can say that they increase at a very high rate during the month of November and after (Winter).





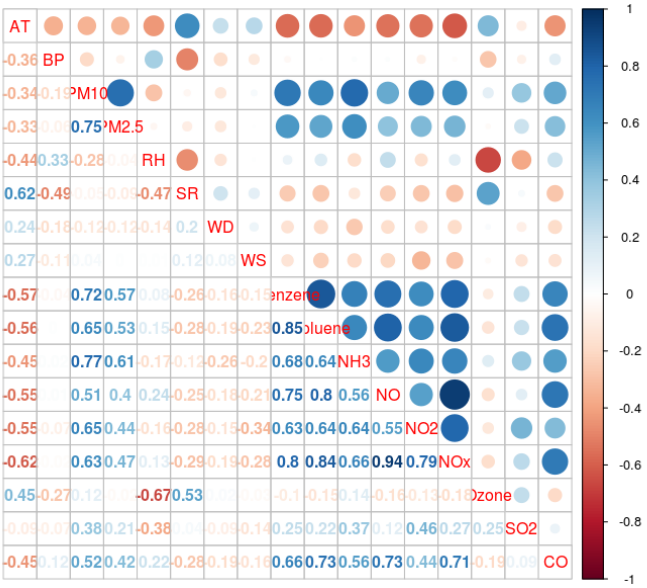


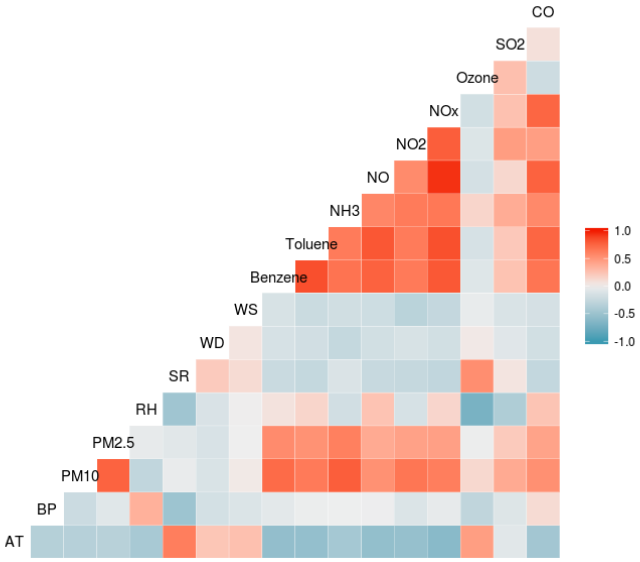


Most of the pollutant levels tend to fall during the afternoon i.e. from noon to 3-4 PM.

Since, the pollutants increase drastically during winters (average temperature is LOW), and decrease during the afternoon (average temperature is HIGH), the pollutants seem to have an indirect correlation with the temperature.

Relationships between attributes:



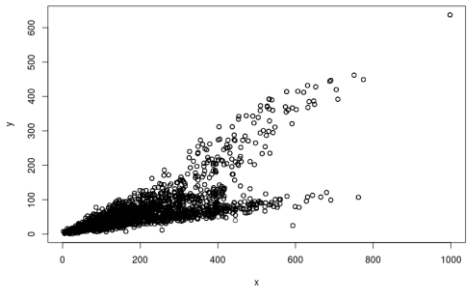


The following inferences have been made by observing these correlation plots:
- The pollutants are strongly related among themselves, for example PM10-PM2.5, Benzene-Toluene, Benzene-NOx, Nox-CO, NH3-PM10 etc.
- AT i.e. Air Temperature is the meteorological attribute that is the most correlated to the pollutants.
- Ozone is affected by multiple meteorological attributes like AT(Air Temperature), RH(Relative Humidity), and SR(Solar Radiation).
There are two types of relations between the attributes:

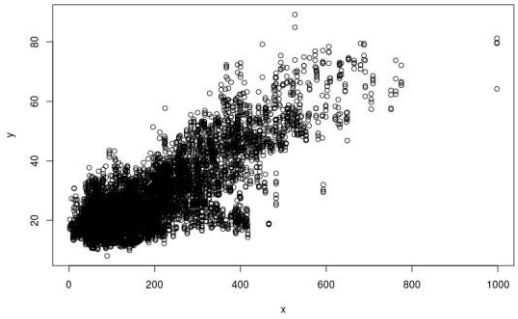
1. Relations between two pollutants
2. Relation between a pollutant and a meteorologic attributes

Pollutant-Pollutant:

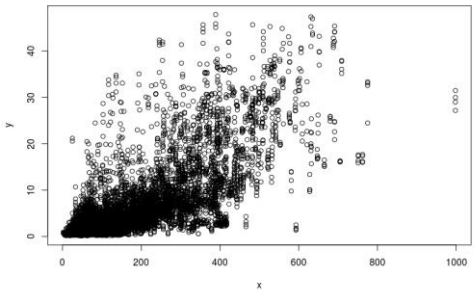
1. PM10- PM2.5- 0.7619542

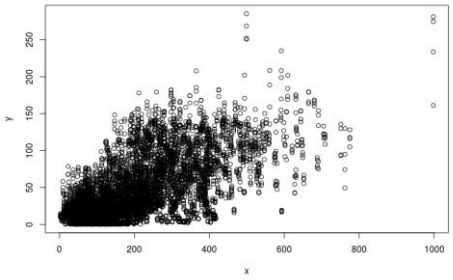


2. PM10- NH3- 0.7846695



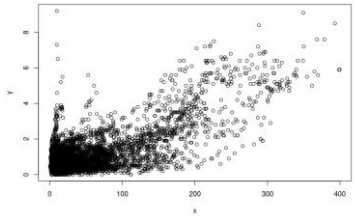
3. PM10- Benzene- 0.7220362



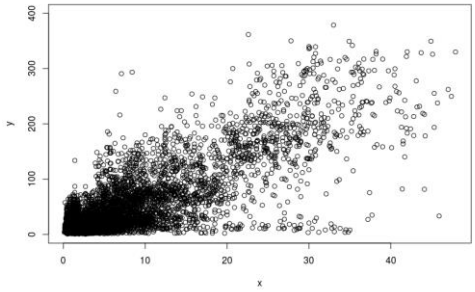


4. PM10- NO2- 0.6681658

5. CO- NOx- 0.7445582

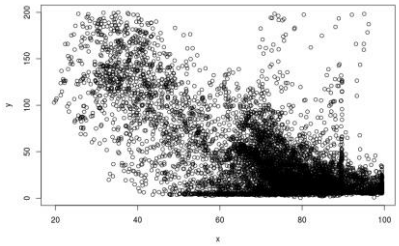


6. Benzene- Nox- 0.8083927

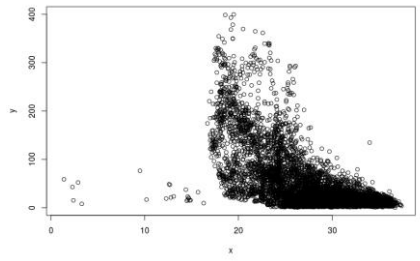


Pollutant-Meteorologic attribute:

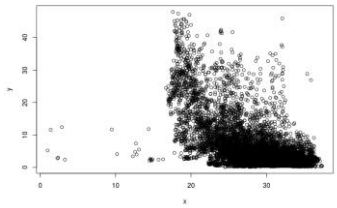
1. RH- Ozone- -0.705501



2. AT- Nox- -0.617112



3. AT- Benzene- -0.561772



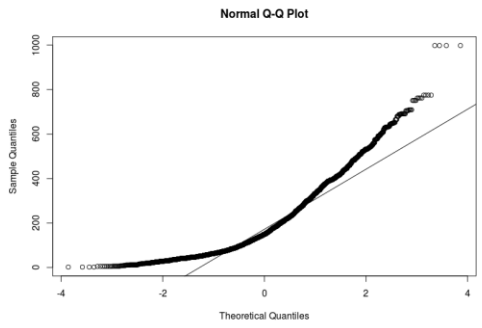
DATA PREPARATION

Missing Values, Errors, and Outliers:

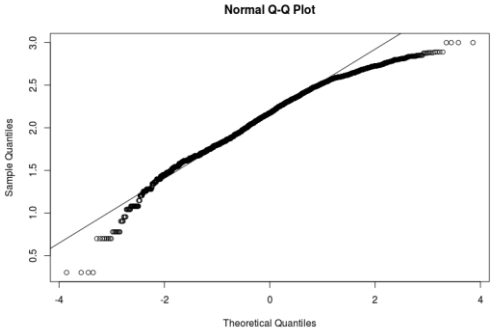
- No specific missing value treatment was used. The tuples with missing values were removed from analysis.
- The outliers don't seem to be affected by some external parameter, they are all erroneous values, and hence, have been removed.

Scaling Features:

- All the attributes were transformed using the logarithmic(base 10) function for the multiple regression model.



Splitting the Data:



- The dataset was divided into training and test sets in the ratio 3:1 (Training Set = 75%, Test Set = 25%)

ALGORITHM IMPLEMENTATION

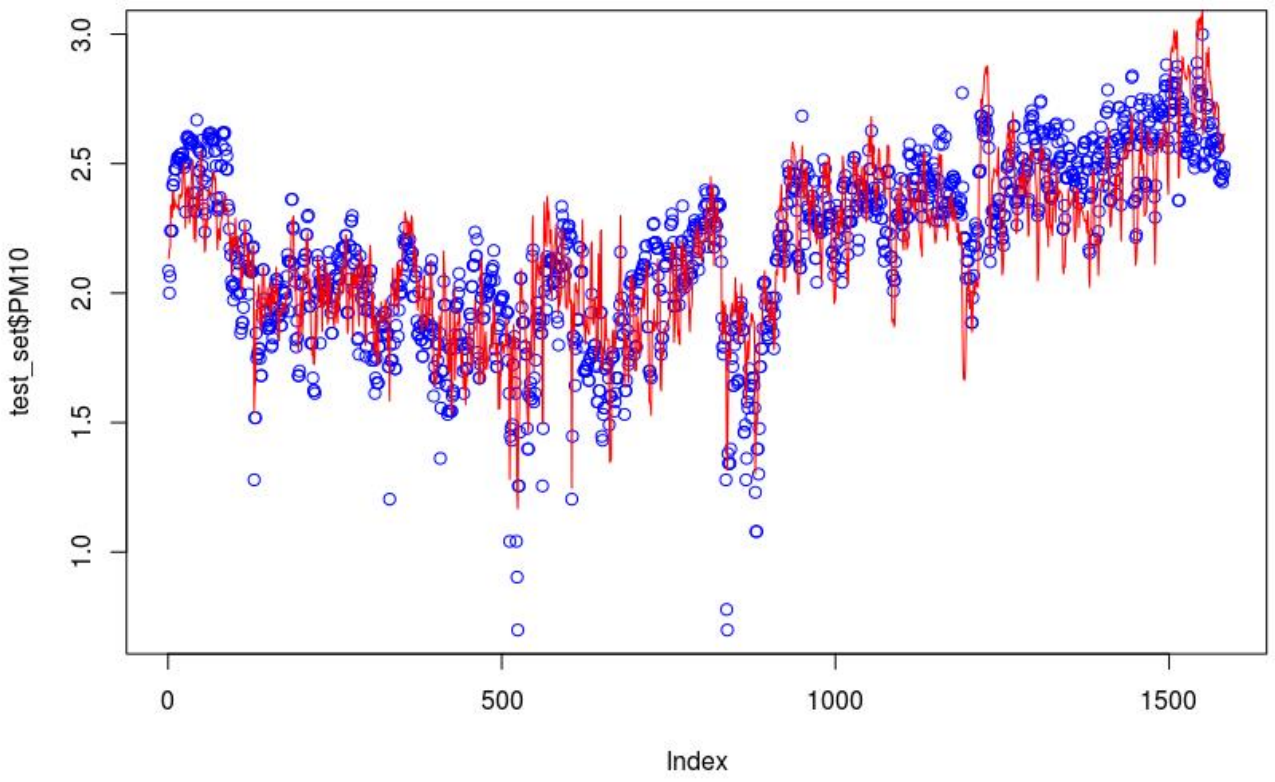
Multiple Regression:

- The goal is to predict the values of key pollutants like PM10, Ozone, etc.
- We made use of the Multiple Regression model using backward elimination since more than one attribute appeared to have effect on the other.

Prediction of PM10 levels:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.26128 -0.10881  0.00728  0.11399  0.80025

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.636414   0.021188   30.04   <2e-16 ***
Benzene     0.145842   0.006109   23.88   <2e-16 ***
NH3         0.309225   0.017586   17.58   <2e-16 ***
PM2.5       0.615273   0.008672   70.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.155 on 4659 degrees of freedom
Multiple R-squared:  0.7926,    Adjusted R-squared:  0.7925
F-statistic:  5937 on 3 and 4659 DF,  p-value: < 2.2e-16
```

- Confidence level: 79.26
- Prediction for PM10 levels has been made using Benzene, NH3, and PM2.5 pollutant levels.
- This shows that the pollutants have a strong relationship among themselves.

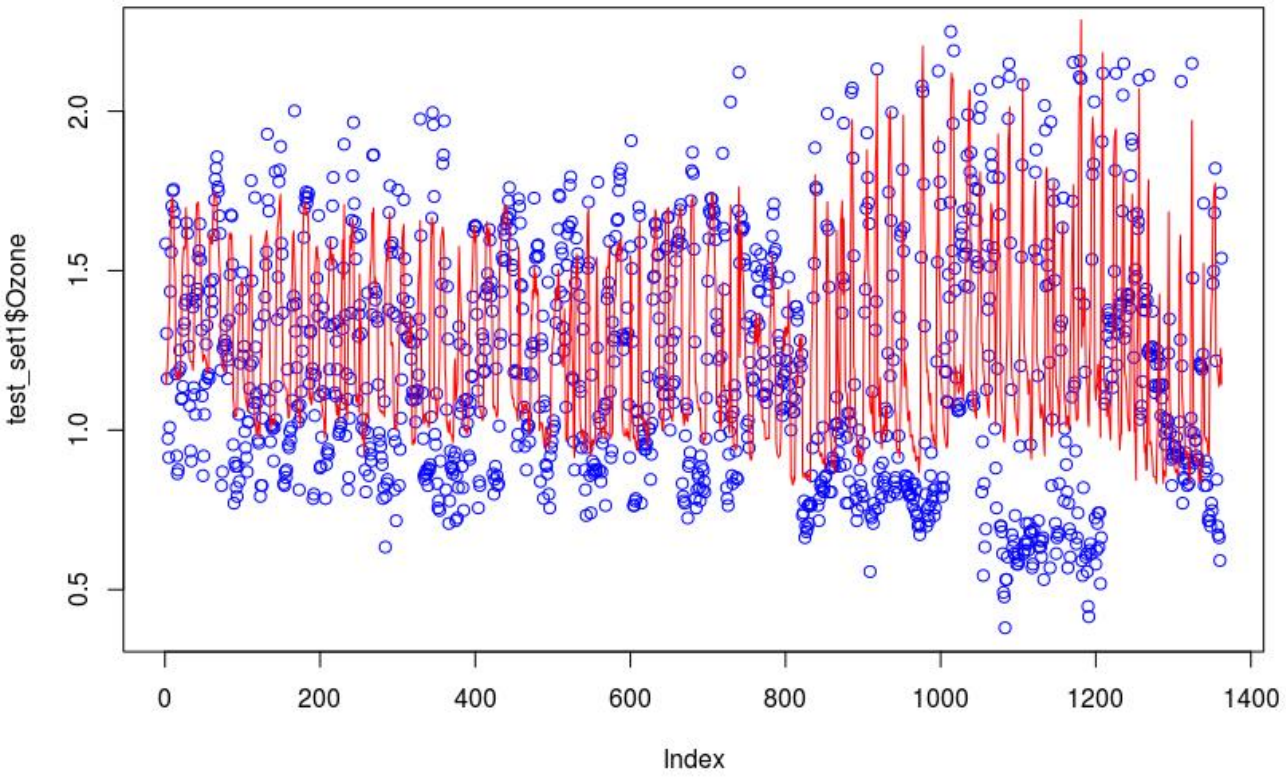


Prediction of Ozone levels:

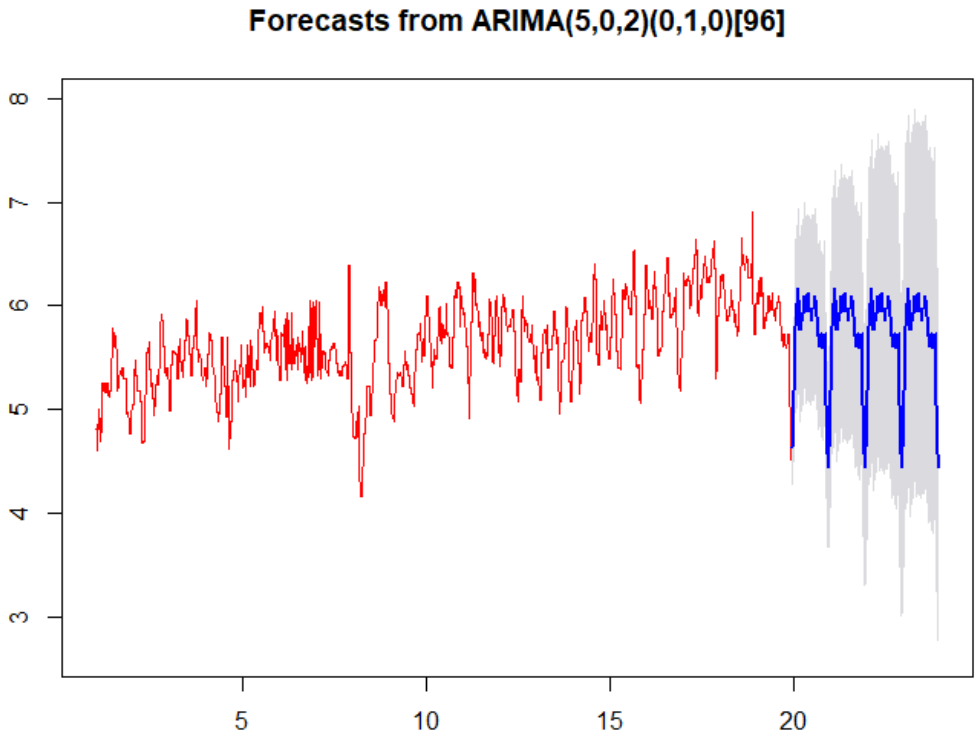
```
lm(formula = Ozone ~ SR + AT + RH, data = training_set1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.70994 -0.19715  0.00253  0.19540  1.37034

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.894822   0.131549  22.006   <2e-16 ***
SR           0.173718   0.004815  36.076   <2e-16 ***
AT           0.551618   0.064849   8.506   <2e-16 ***
RH          -1.384930   0.042512 -32.577   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3078 on 4880 degrees of freedom
Multiple R-squared:  0.5542,    Adjusted R-squared:  0.5539
F-statistic:  2022 on 3 and 4880 DF,  p-value: < 2.2e-16
```

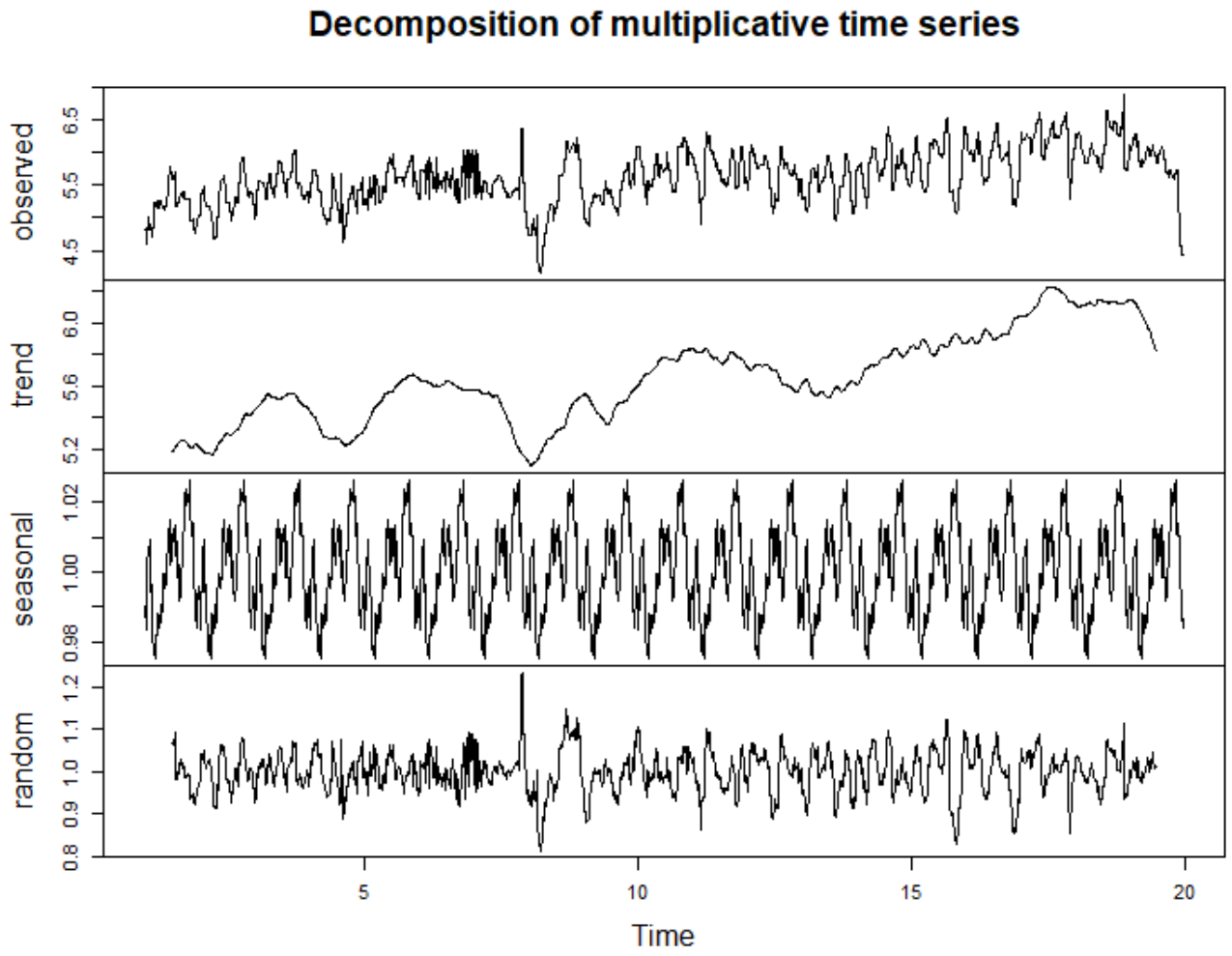
Confidence level: 55.42
Ozone levels have been affected by SR, AT, and RH

Increase in solar radiation leads to formation of ozone due to the availability of nascent oxygen because of splitting of Oxygen molecule with the solar radiation.



Time Series Analysis:




Decomposition of multiplicative time series
observed
trend
seasonal
random
4.5
5.5
6.5
5.2
5.6
6.0
0.98
1.00
1.02
0.8
0.9
1.0
1.1
1.2
5
10
15
20
Time