DATA MINING PROJECT REPORT

Abhinav Kumar (01)

Ayush Kumar Singh (07)

Divesh Bhagat (10)

Vrinda Sharma (1722229, Dept. of Statistics)

Acknowledgement

We are extremely grateful to our mentor Dr Vasudha Bhatnagar, Head of the Department, Department of Computer Science for giving us the opportunity to work on this project. Her knowledge, ideas and experience helped us immensely right from the start till the very end. She patiently cleared all our doubts and apprehensions regarding the project with ease. Without her constant support and guidance, we could not have completed this project.

Dataset

Orissa Pollution Dataset

Dimensions of dataset : (2393, 13) Details of ambient air quality with respect to air quality parameters, like Sulfur dioxide, Nitrogen dioxide, Respirable Suspended Particulate Matter (RSPM) and Suspended Particulate Matter (SPM) etc. are given in the datasets.

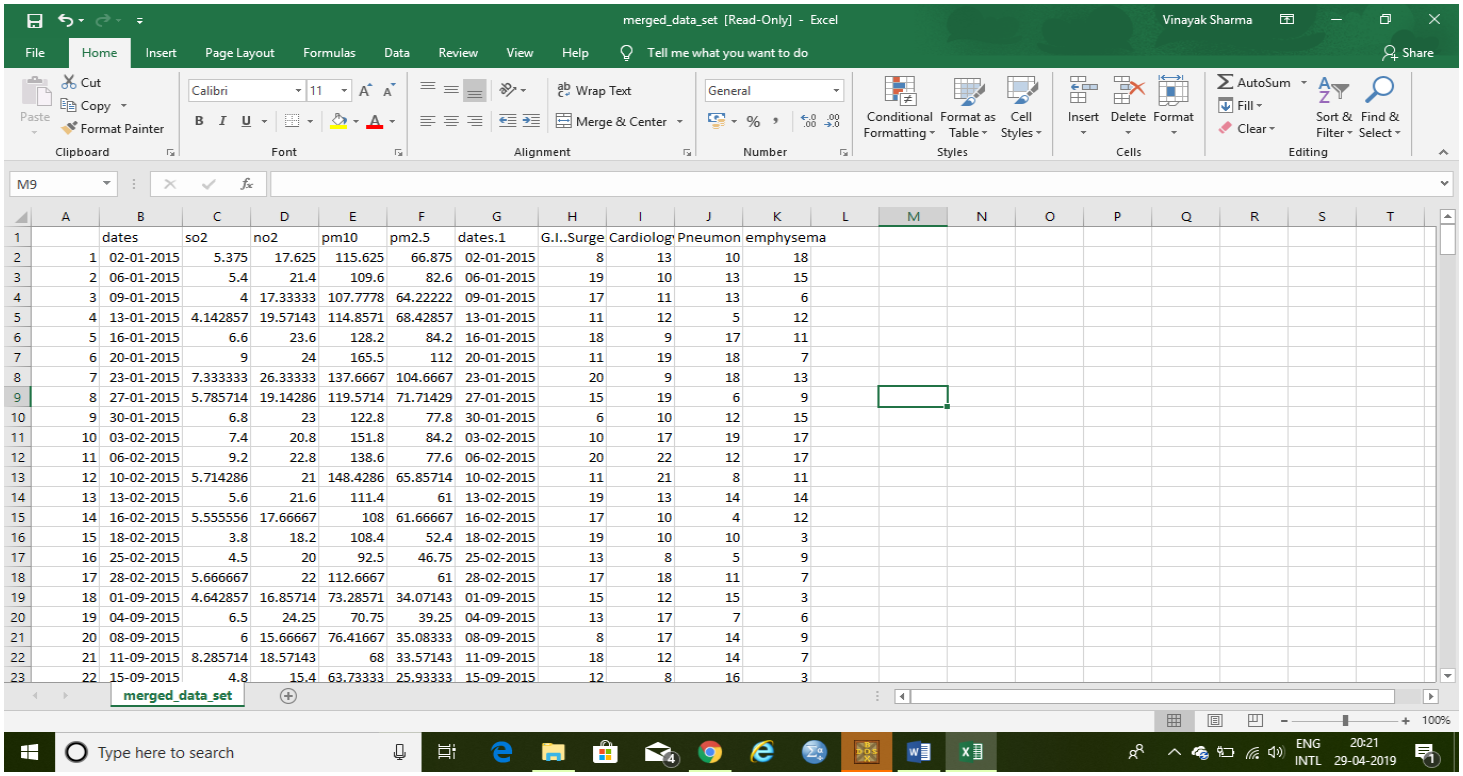
Source (main pollution dataset)

Data.Gov

Auxiliary Data Source (For OPD report)

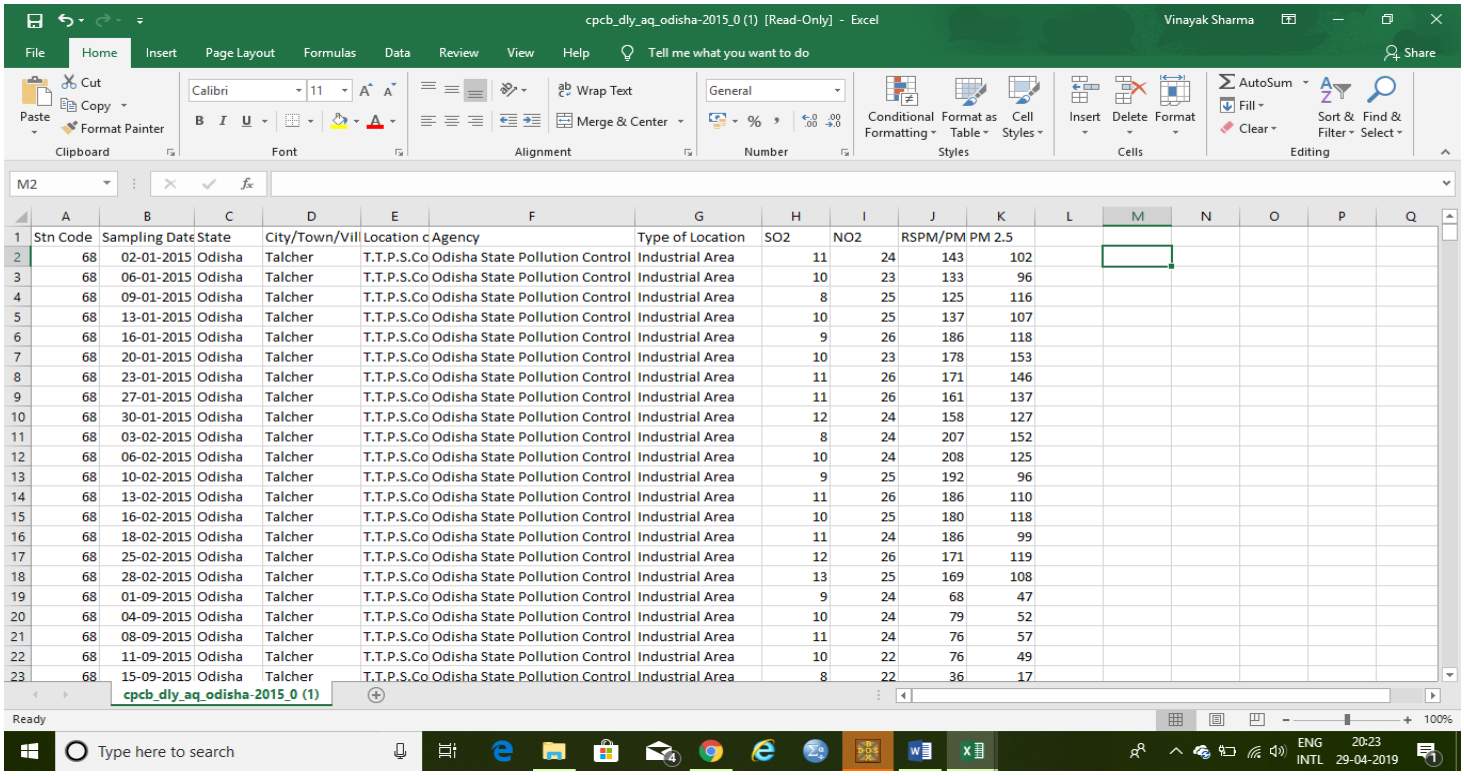
International Federation of Health Information Management Associations

Data Instance



merged_data_set [Read-Only] - Excel — Vinayak Sharma

	A	B	C	D	E	F	G	H	I	J	K
1		dates	so2	no2	pm10	pm2.5	dates.1	G.I..Surge	Cardiolog	Pneumon	emphysema
2	1	02-01-2015	5.375	17.625	115.625	66.875	02-01-2015	8	13	10	18
3	2	06-01-2015	5.4	21.4	109.6	82.6	06-01-2015	19	10	13	15
4	3	09-01-2015	4	17.33333	107.7778	64.22222	09-01-2015	17	11	13	6
5	4	13-01-2015	4.142857	19.57143	114.8571	68.42857	13-01-2015	11	12	5	12
6	5	16-01-2015	6.6	23.6	128.2	84.2	16-01-2015	18	9	17	11
7	6	20-01-2015	9	24	165.5	112	20-01-2015	11	19	18	7
8	7	23-01-2015	7.333333	26.33333	137.6667	104.6667	23-01-2015	20	9	18	13
9	8	27-01-2015	5.785714	19.14286	119.5714	71.71429	27-01-2015	15	19	6	9
10	9	30-01-2015	6.8	23	122.8	77.8	30-01-2015	6	10	12	15
11	10	03-02-2015	7.4	20.8	151.8	84.2	03-02-2015	10	17	19	17
12	11	06-02-2015	9.2	22.8	138.6	77.6	06-02-2015	20	22	12	17
13	12	10-02-2015	5.714286	21	148.4286	65.85714	10-02-2015	11	21	8	11
14	13	13-02-2015	5.6	21.6	111.4	61	13-02-2015	19	13	14	14
15	14	16-02-2015	5.555556	17.66667	108	61.66667	16-02-2015	17	10	4	12
16	15	18-02-2015	3.8	18.2	108.4	52.4	18-02-2015	19	10	10	3
17	16	25-02-2015	4.5	20	92.5	46.75	25-02-2015	13	8	5	9
18	17	28-02-2015	5.666667	22	112.6667	61	28-02-2015	17	18	11	7
19	18	01-09-2015	4.642857	16.85714	73.28571	34.07143	01-09-2015	15	12	15	3
20	19	04-09-2015	6.5	24.25	70.75	39.25	04-09-2015	13	17	7	6
21	20	08-09-2015	6	15.66667	76.41667	35.08333	08-09-2015	8	17	14	9
22	21	11-09-2015	8.285714	18.57143	68	33.57143	11-09-2015	18	12	14	7
23	22	15-09-2015	4.8	15.4	63.73333	25.93333	15-09-2015	12	8	16	3



cpcb_dly_aq_odisha-2015_0 (1) [Read-Only] - Excel — Vinayak Sharma

	A	B	C	D	E	F	G	H	I	J	K
1	Stn Code	Sampling Date	State	City/Town/Vil	Location c	Agency	Type of Location	SO2	NO2	RSPM/PM	PM 2.5
2	68	02-01-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	11	24	143	102
3	68	06-01-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	10	23	133	96
4	68	09-01-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	8	25	125	116
5	68	13-01-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	10	25	137	107
6	68	16-01-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	9	26	186	118
7	68	20-01-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	10	23	178	153
8	68	23-01-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	11	26	171	146
9	68	27-01-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	11	26	161	137
10	68	30-01-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	12	24	158	127
11	68	03-02-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	8	24	207	152
12	68	06-02-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	10	24	208	125
13	68	10-02-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	9	25	192	96
14	68	13-02-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	11	26	186	110
15	68	16-02-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	10	25	180	118
16	68	18-02-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	11	24	186	99
17	68	25-02-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	12	26	171	119
18	68	28-02-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	13	25	169	108
19	68	01-09-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	9	24	68	47
20	68	04-09-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	10	24	79	52
21	68	08-09-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	11	24	76	57
22	68	11-09-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	10	22	76	49
23	68	15-09-2015	Odisha	Talcher	T.T.P.S.Co	Odisha State Pollution Control	Industrial Area	8	22	36	17

Data Description

The dataset consists of 2393 rows and 13 columns. The description of the various columns are given below:

NO_2: Nitrogen dioxide is the chemical compound with the formula NO_2. It is one of several nitrogen oxides. NO_2 is an intermediate in the industrial synthesis of nitric acid, millions of tons of which are produced each year which is used primarily in the production of fertilizers. Breathing air with a high concentration of NO_2 can irritate airways in the human respiratory system. Such exposures over short periods can aggravate respiratory diseases, particularly asthma, leading to respiratory symptoms (such as coughing, wheezing or difficulty breathing), hospital admissions and visits to emergency rooms.

SO_2: Sulphur dioxide (SO_2), a colourless, bad-smelling, toxic gas, is part of a larger group of chemicals referred to as sulphur oxides (SO_x). These gases, especially SO_2, are emitted by the burning of fossil fuels — coal, oil, and diesel — or other materials that contain sulphur. Sources include power plants, metals processing and smelting facilities, and vehicles Sulphur dioxide, associated SO_x, and secondary pollutants can contribute to respiratory illness by making breathing more difficult, especially for children, the elderly, and those with pre-existing conditions. Longer exposures can aggravate existing heart and lung conditions, as well. Sulphur dioxide and other SO_x are partly culpable in the formation of thick haze and smog, which can impair visibility in addition to impacting health.

AQI: An **air quality index** (**AQI**) is used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become. As the AQI increases, an increasingly large percentage of the population is likely to experience increasingly severe

adverse health effects. Different countries have their own air quality indices, corresponding to different national air quality standards.

There are six AQI categories, namely Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe. The proposed AQI will consider eight pollutants (PM10, PM2.5, NO_2, SO_2, CO, O_3, NH_3, and Pb) for which short-term (up to 24-hourly averaging period) National Ambient Air Quality Standards are prescribed.

The AQI values and corresponding ambient concentrations (health breakpoints) as well as associated likely health impacts for the identified eight pollutants are as follows:

AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	**PM10 (24hr)**	**PM2.5 (24hr)**	**NO_2 (24hr)**	**O_3 (8hr)**	**CO (8hr)**	**SO_2 (24hr)**	**NH_3 (24hr)**	**Pb (24hr)**
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
Satisfactory (51-100)	51–100	31–60	41–80	51-100	1.1-2.0	41-80	201-400	0.5–1.0
Moderately polluted (101–200)	101-250	61–90	81-180	101-168	2.1-10	81-380	401-800	1.1–2.0
Poor (201–300)	251-350	91-120	181-280	169–208	10–17	381-800	801-1200	2.1–3.0
Very poor (301–400)	351-430	121–250	281–400	209–748	17–34	801-1600	1200-1800	3.1–3.5
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

AQI	Associated Health Impacts
Good (0–50)	Minimal impact
Satisfactory (51–100)	May cause minor breathing discomfort to sensitive people.
Moderately polluted (101–200)	May cause breathing discomfort to people with lung disease such as asthma, and discomfort to people with heart disease, children and older adults.
Poor (201–300)	May cause breathing discomfort to people on prolonged exposure, and discomfort to people with heart disease.
Very poor (301–400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases.
Severe (401–500)	May cause respiratory impact even on healthy people, and serious health impacts on people with lung/heart disease. The health impacts may be experienced even during light physical activity.

PM2.5: **Particulate matter** (**PM**), **particulates**, or **suspended particulate matter** (**SPM**) – are microscopic solid or liquid matter suspended in the atmosphere of Earth. PM2.5 refers to atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair. Commonly written as $PM_{2.5}$, particles in this category are so small that they can only be detected with an electron microscope. They are even smaller than their counterparts PM_{10}, which are particles that are 10 micrometers or less, and are also called **fine particles**. Since they are so small and light, fine particles tend to stay longer in the air than heavier particles. This increases the chances of humans and animals

inhaling them into the bodies. Owing to their minute size, particles smaller than 2.5 micrometers are able to bypass the nose and throat and penetrate deep into the lungs and some may even enter the circulatory system.

PM10: Particles come in a wide range of sizes. Particles less than or equal to 10 micrometers in diameter are so small that they can get into the lungs, potentially causing serious health problems. Ten micrometers is less than the width of a single human hair. **Coarse dust particles (PM10)** are 2.5 to 10 micrometers in diameter. Sources include crushing or grinding operations and dust stirred up by vehicles on roads.

AREA: This variable represents whether the area is industrial or residential (rural or industrial area).

City/Town/Village: This variable gives the name of the city/town/district.

Pneumonia: It is an infection that inflames the air sacs in one or both lungs. The air sacs may fill with fluid or pus (purulent material), causing cough with phlegm or pus, fever, chills, and difficulty breathing. A variety of organisms, including bacteria, viruses and fungi, can cause **pneumonia**.

Emphysema: It is a type of COPD involving damage to the air sacs (alveoli) in the lungs. As a result, your body does not get the oxygen it needs. **Emphysema** makes it hard to catch your breath. You may also have a chronic cough and have trouble breathing during exercise.

Cardiology: It is a branch of medicine that deals with the disorders of the heart as well as some parts of the circulatory system.

G.I. Surgery: Gastrointestinal surgery is a treatment for diseases of the stomach, intestines, and other parts of the body involved in digestion.

Objective

- To analyze the association between pollutants and area. For this exploratory data analysis was used to see the difference in the values of pollutants namely NO_2, SO_2, AQI, PM2.5 and PM10.

- Using correlation matrix, the correlation between the number of cases of health issues and the various pollutants level was obtained.

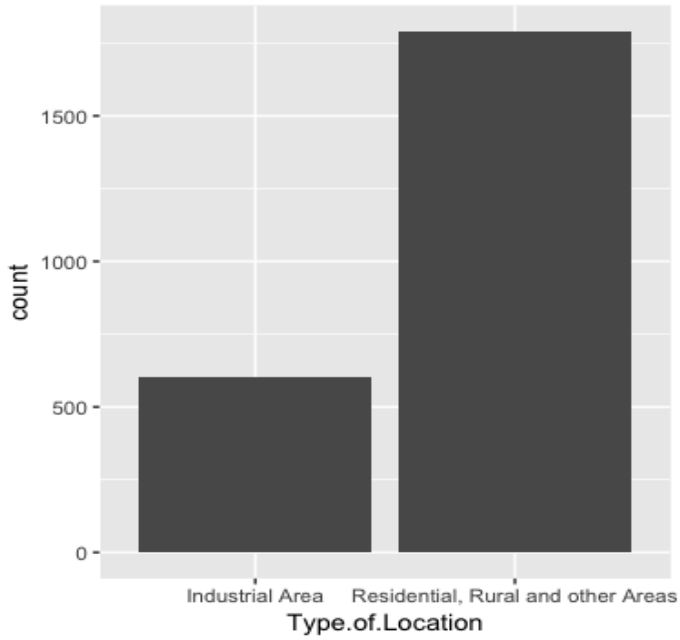
- Predicting the type of the location depending upon the AQI. A supervised learning technique like Logistic regression, Decision tree were used and their accuracy measures like confusion matrix, accuracy, precision, F1 score were compared to see which technique was better able to predict the location depending upon NO_2, SO_2, AQI, PM2.5 and PM10.

Preliminary Analysis

As the data is about the collection of pollution levels around many areas of Orissa and attributes like **Agency** is of no use for our analysis, it's a constant value attribute after examining the data we also found out the we could also make use the **year quarters** and **season** for our analysis, so we wrote the code in R to add those attributes to our dataset the we download from data.gov.

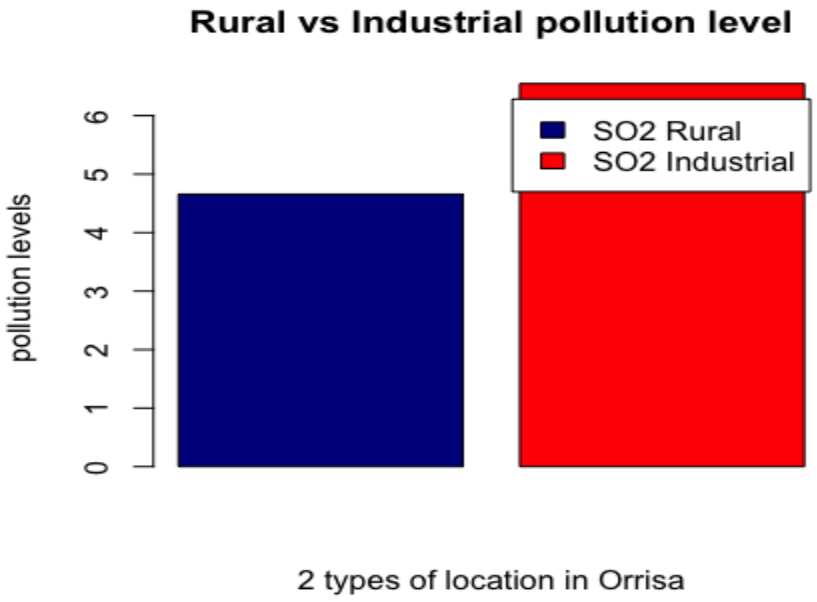
Pre-processing

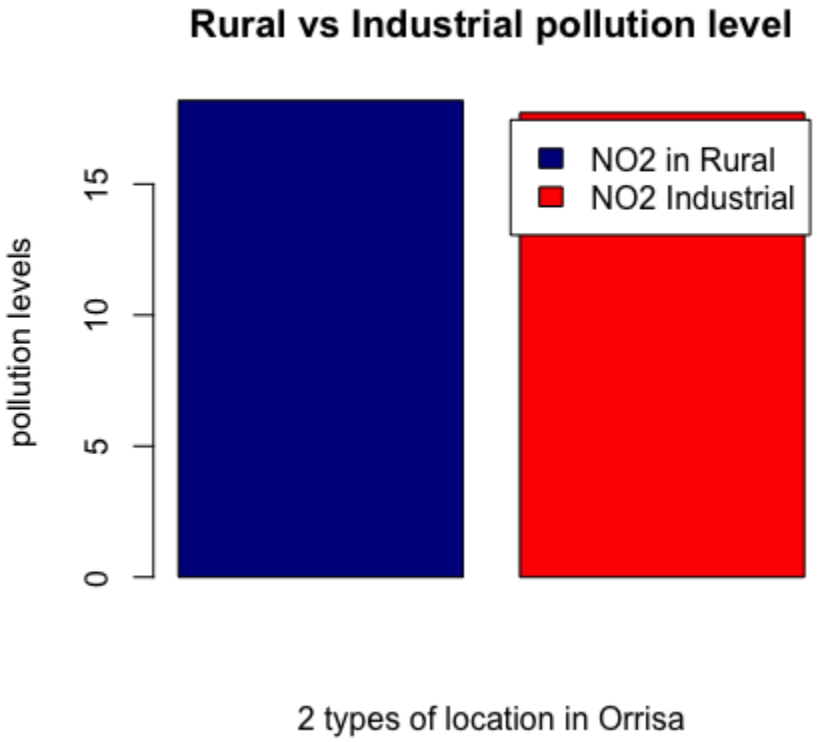
- Season column added according to the date in Orissa.
- Type of location ready for classification or regression.
- Outlier Capping has been done.

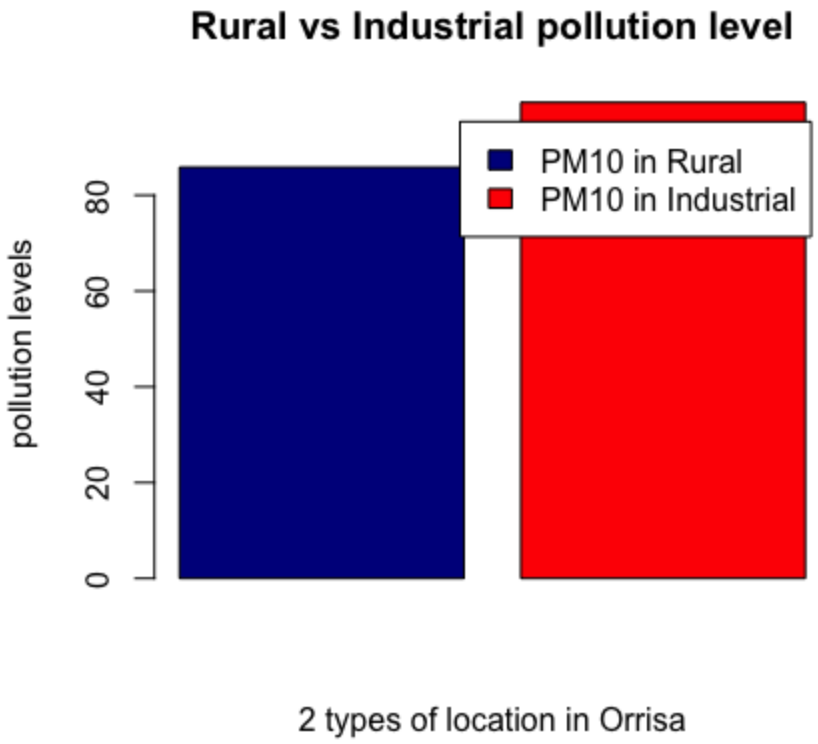


Classes in the dataset and their no. of instances

Visualization

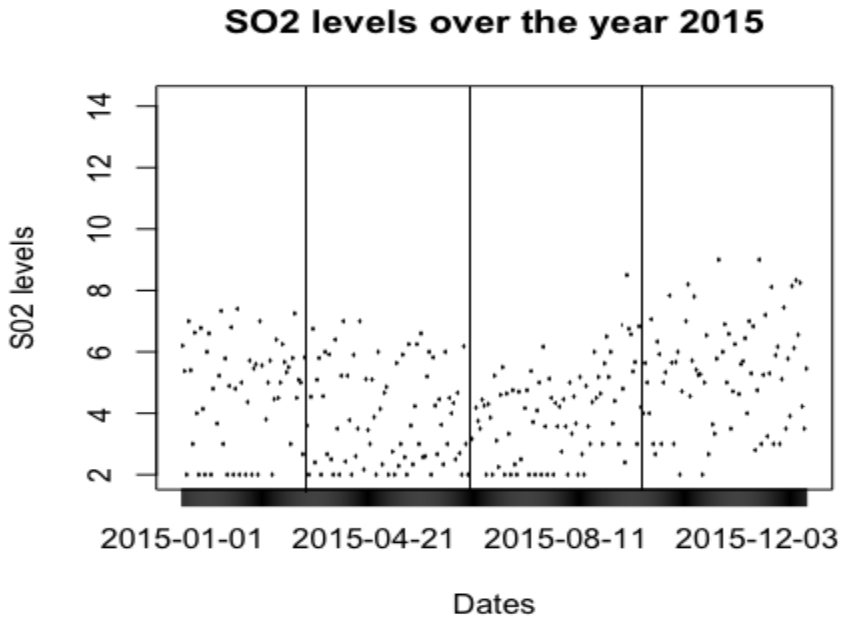


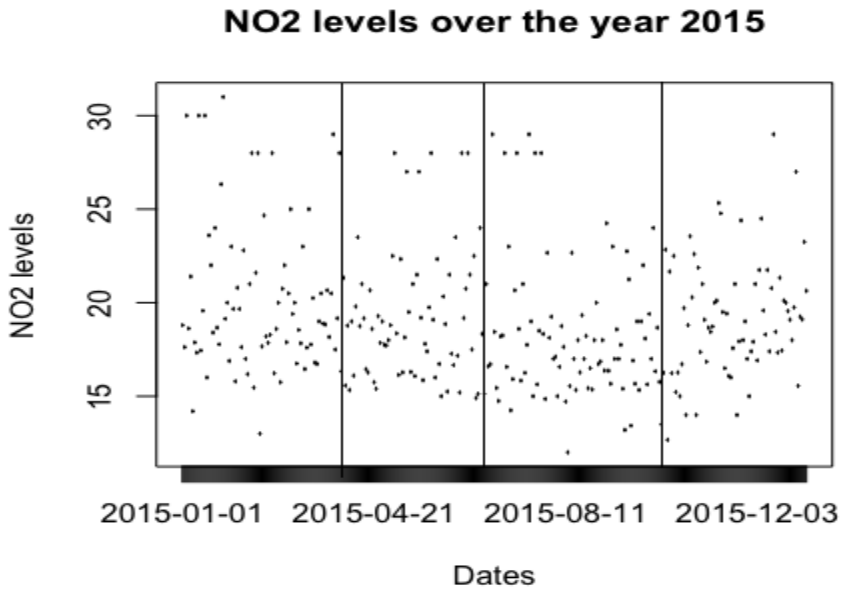




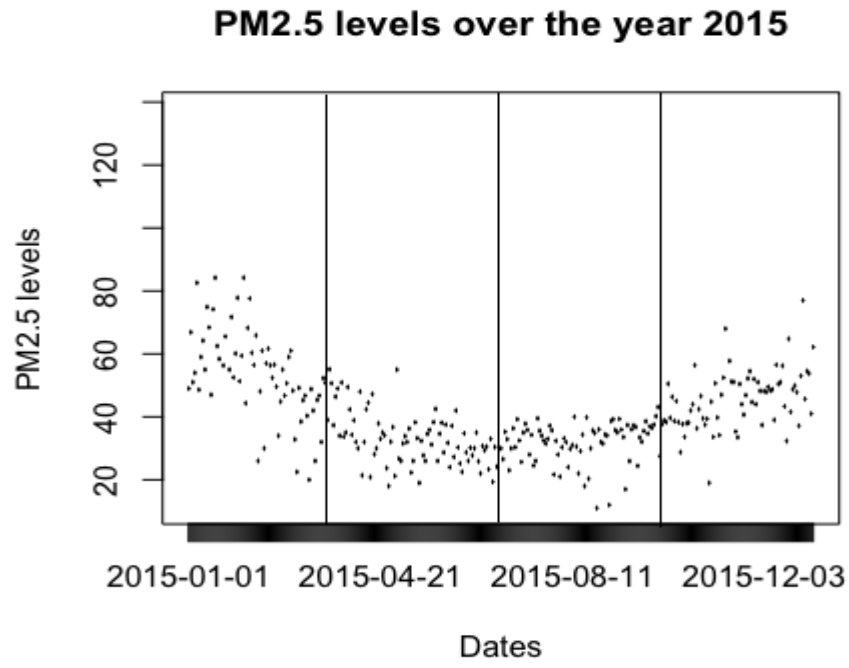
Here we can infer from the bar graphs that major factor for SO2 and NO2 pollutants is industrial pollution and RSPM.PM10 pollutants concentration is almost same for industrial and rural areas in Orissa.

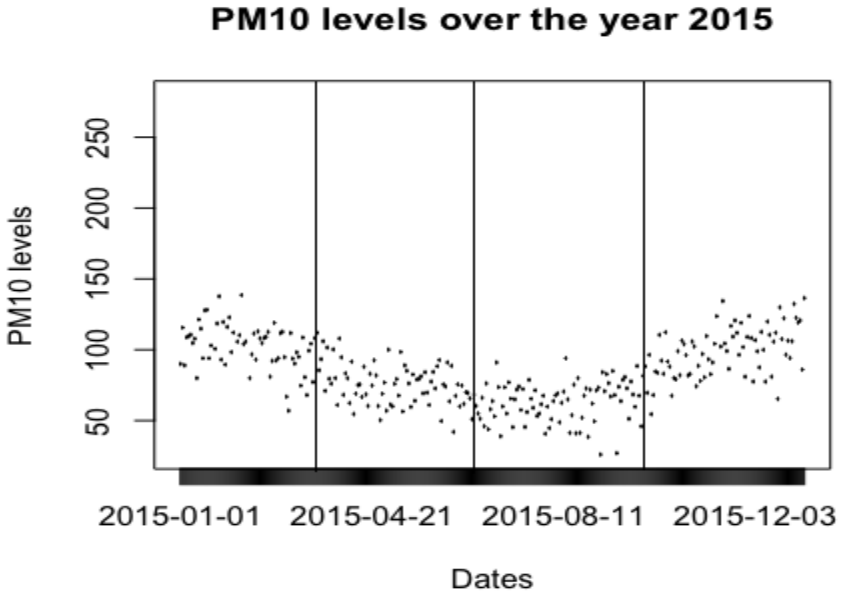
Scatter plots





Here nothing can be inferred from both the plots as the data points are scattered vastly all over the year.

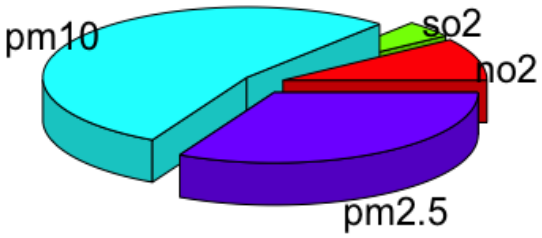




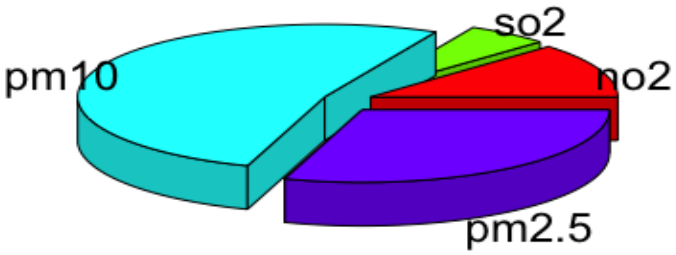
Here we can infer from the scatter plots of both PM2.5 and PM10 follow the normal trend as they should i.e. having high concentration in the winter months and low concentrations in the months of monsoon in the middle of the plots.

Pie-charts (District wise Pollutants)

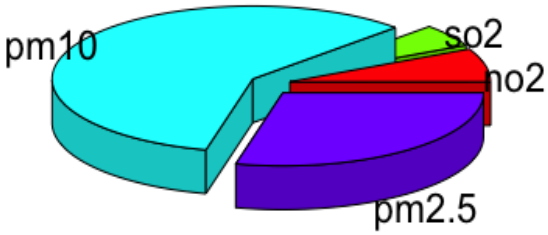
Talcher Pollutants



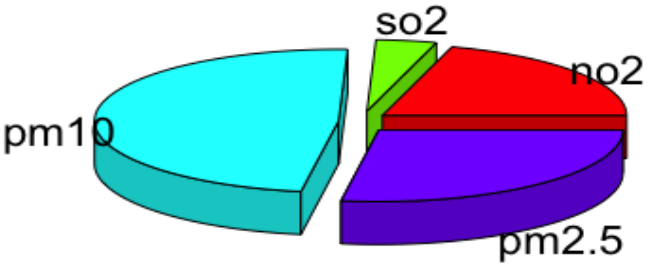
Angul Pollutants



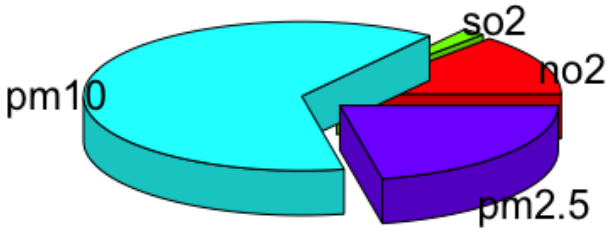
Rourkela Pollutants



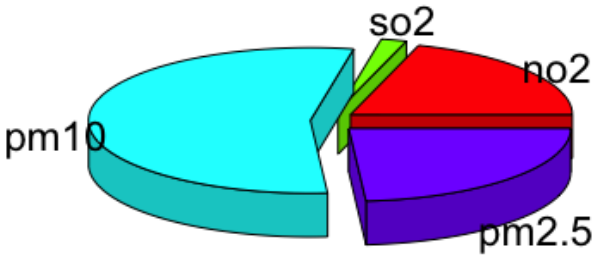
Rayagada Pollutants



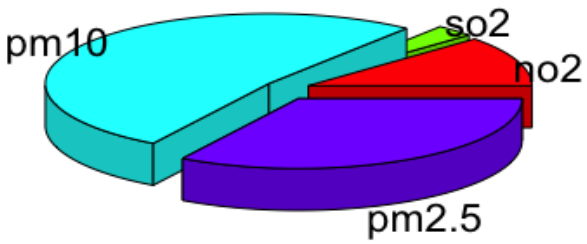
Bhubaneswar Pollutants



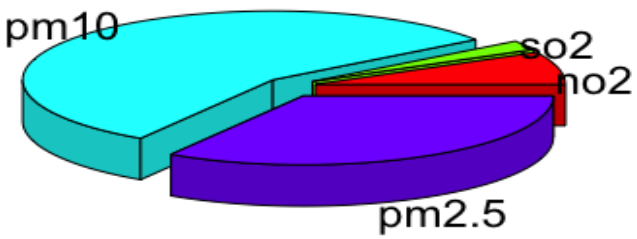
Cuttack Pollutants



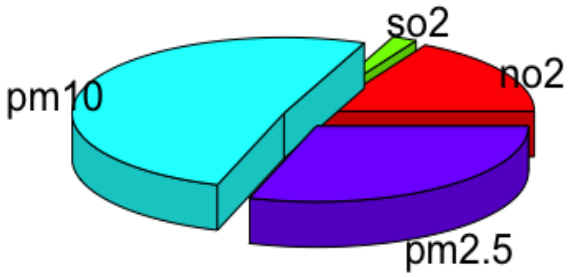
Sambalpur Pollutants



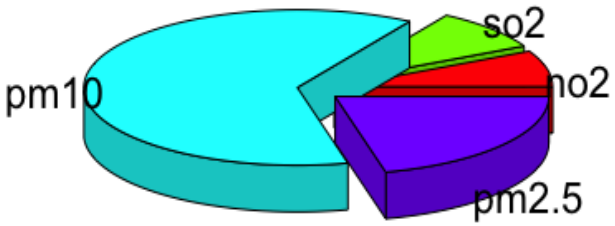
Balasore Pollutants



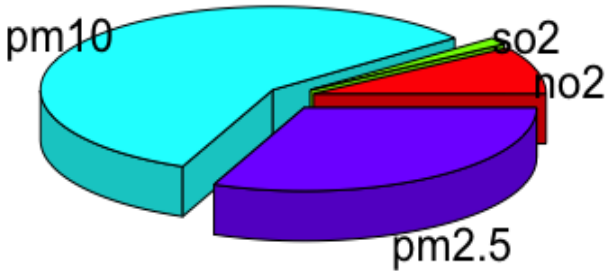
Berhampur Pollutants



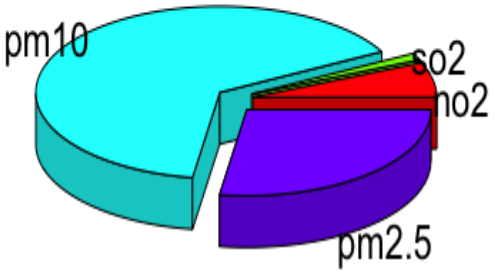
Paradeep Pollutants

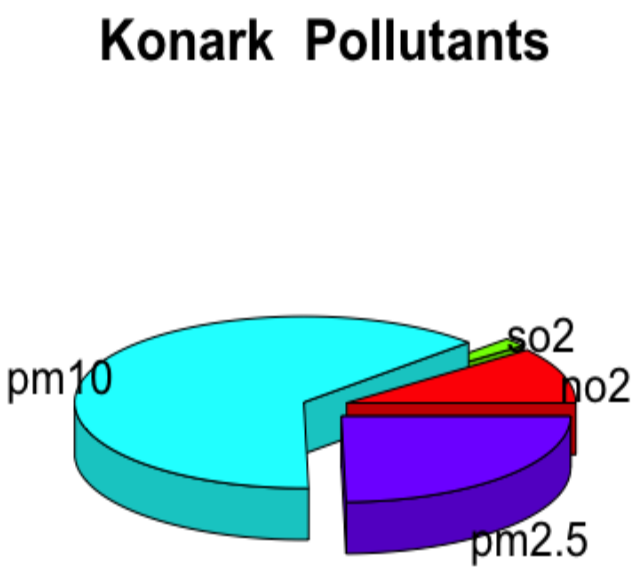


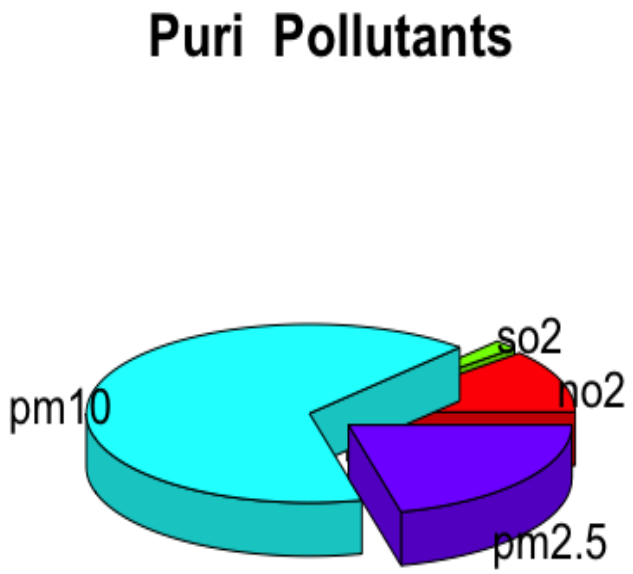
Keonjhar Pollutants



Kalinga Nagar Pollutants





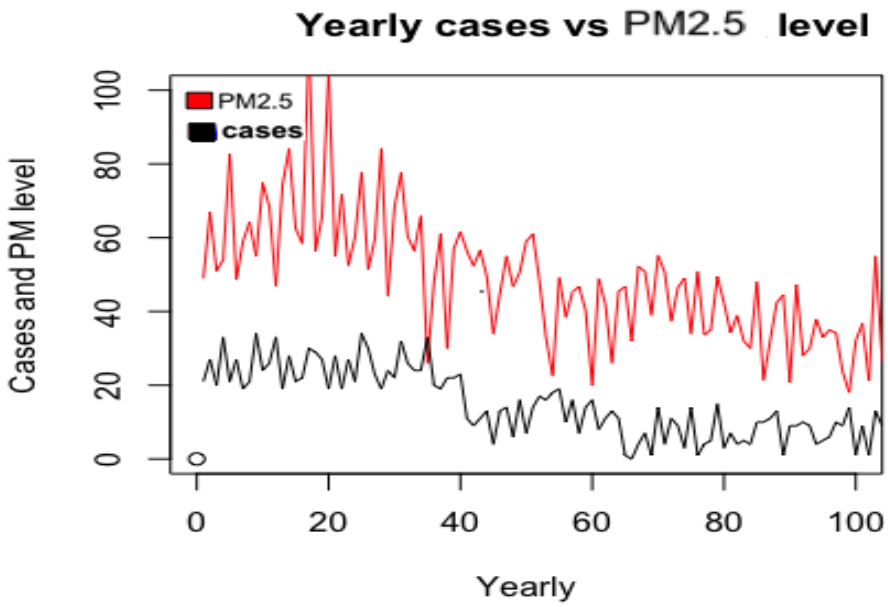


Here from the pie-charts it is evident that the major pollutants in almost all the districts of the state Orissa is PM10 followed by PM2.5 which are one of the major causes of respiratory diseases.

SO_2 is almost in negligible concentration when compared against all other pollutants in all the districts.

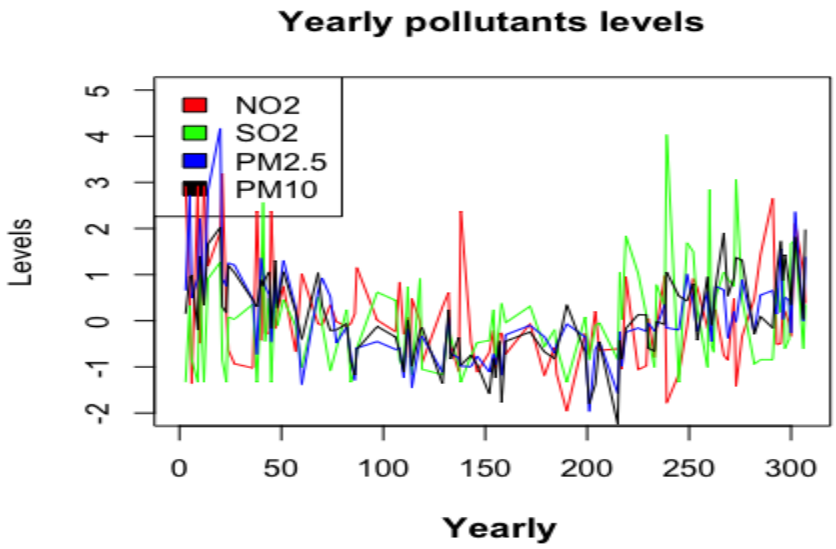
NO_2 has significant concentration in almost all the districts and it is one of the major contributor to air pollution and several ailments to human body and environment.

FIRST ANALYSIS

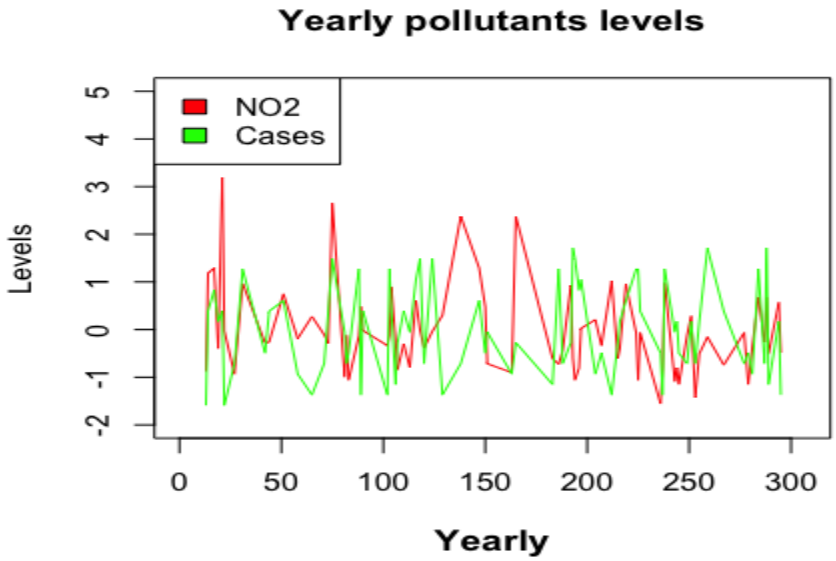


From here what we can deduce is more the pollutant PM2.5 level more is the number of cases of **Emphysema**.

Yearly Analysis



Here we had to normalize the data since the ranges for the pollutant levels differs vastly.

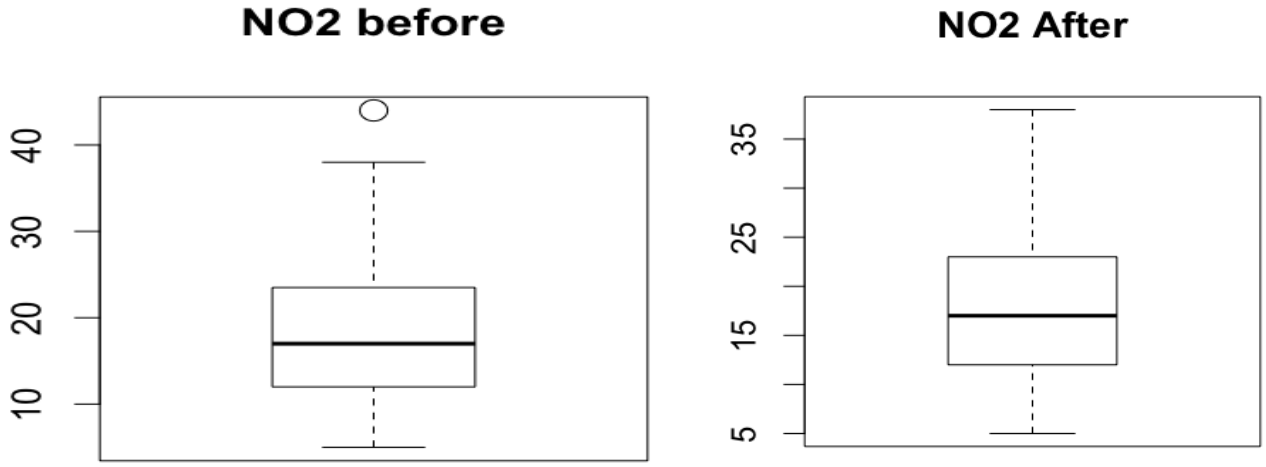


From here we can't say anything that whether the levels of NO2 are causing increase or decrease in the cases of **Pneumonia.**

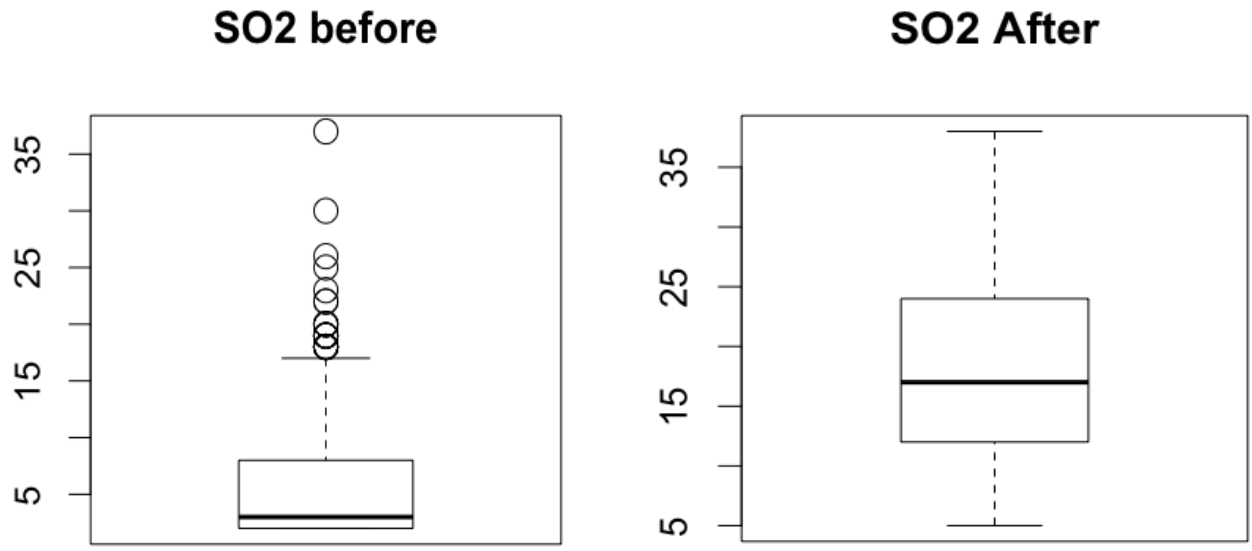
Boxplots (OUTLIER CAPPING)

Outlier capping (via means of the respective columns) have been done as number of data points is already less in our dataset. This is supposed to improve our dataset quality and help us to better visualize the dataset.

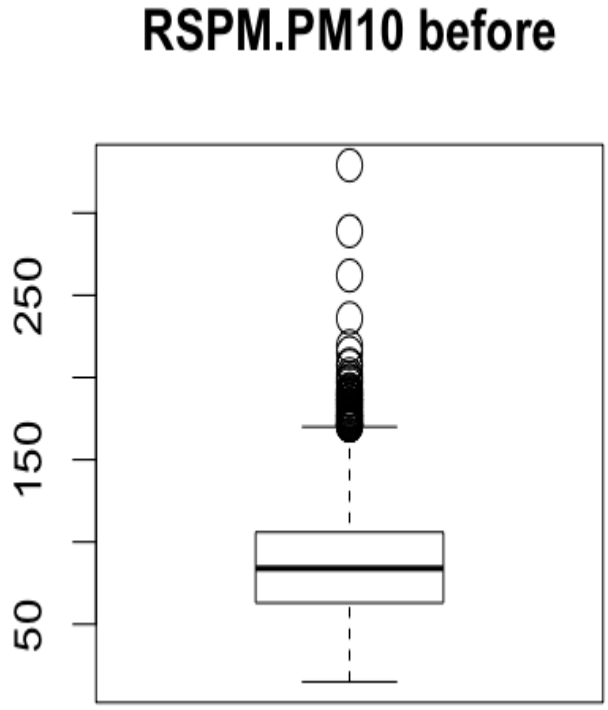
NO2

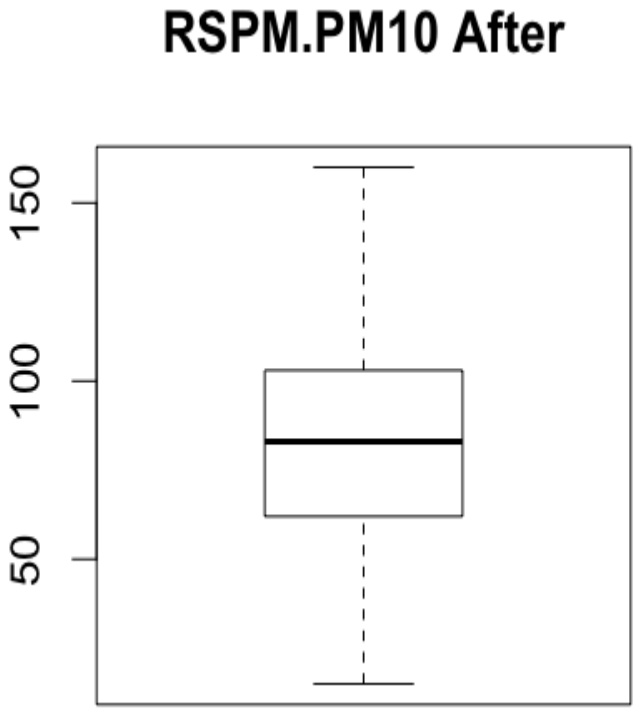


SO2

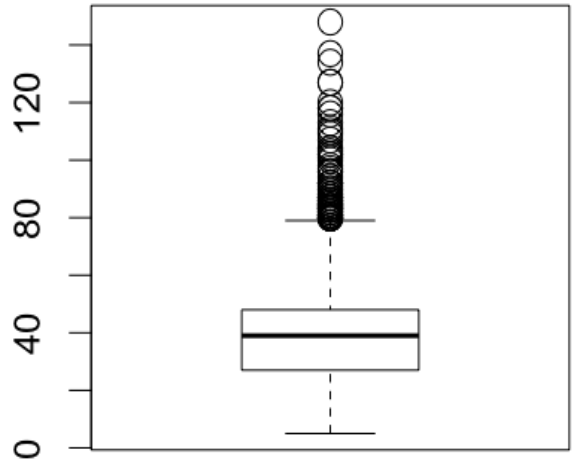


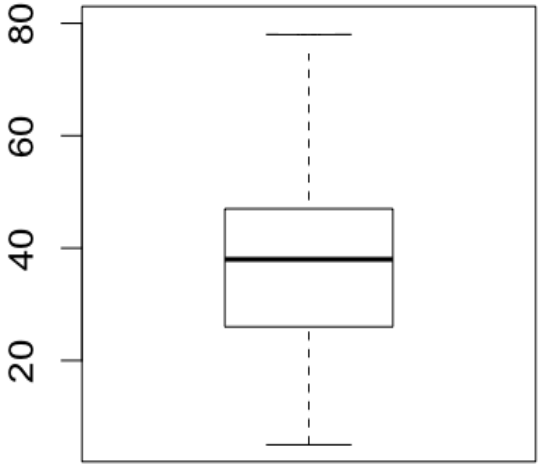
PM10


RSPM.PM10 before
50
150
250

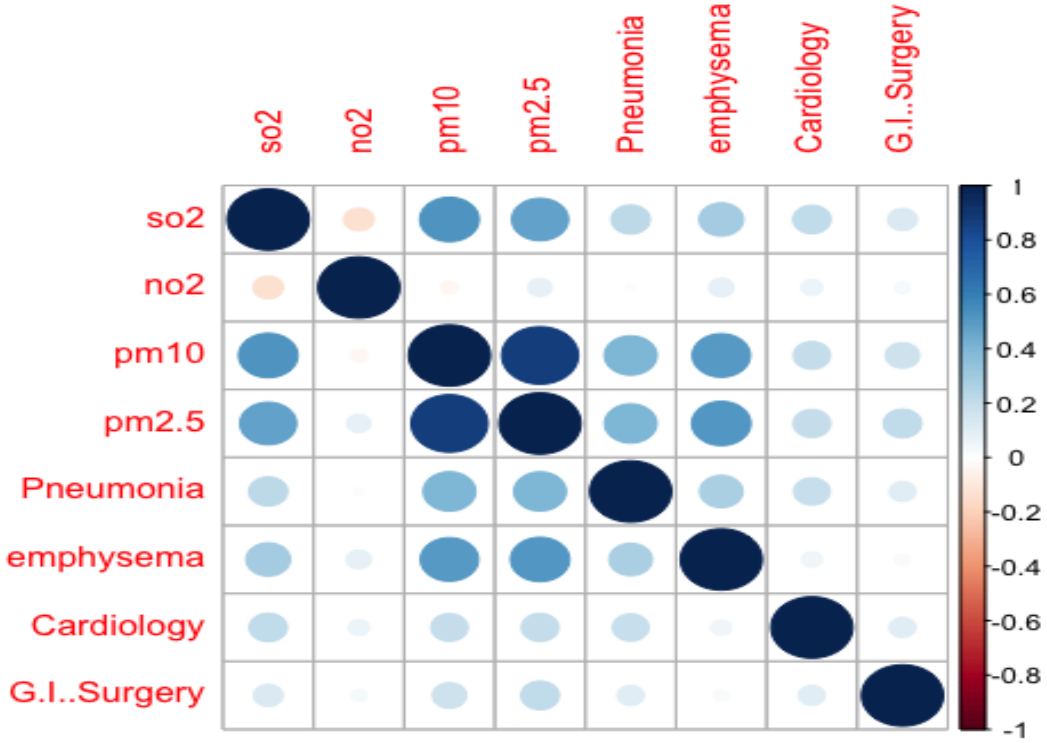

RSPM.PM10 After
50
100
150

PM2.5


PM.2.5 before
0
40
80
120


PM.2.5, After
20
40
60
80

Corrplot



From the corrplot we infer that the correlation of Emphysema, pneumonia with PM10 and PM2.5 is significant. Emphysema is more correlated with PM10 and PM2.5 more than pneumonia.

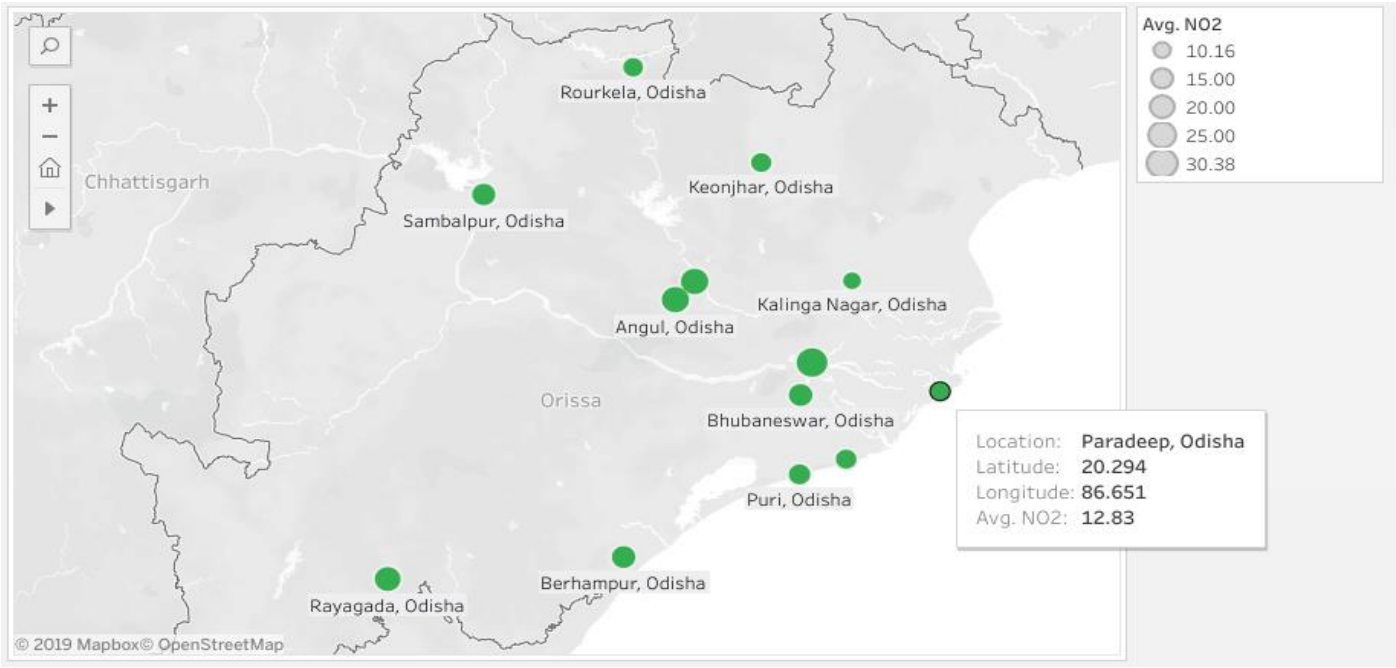
Since the correlation is positive as the level of PM10 and PM2.5 increases we will have more cases of these diseases and thus out of all the diseases we have in our dataset these two namely Emphysema and Pneumonia are significantly related with the pollutants level.

On further research we found that the pollutants could be one of the causes of these diseases.

Also the PM10 and PM2.5 have significant correlation with SO_2 thus increasing one would also increase the other. Even PM10 and PM2.5 are correlated.

Geo Mapping

NO_2 Geo Mapping



SO_2 Geo Mapping

PM2.5 Geo Mapping

PM10 Geo Mapping

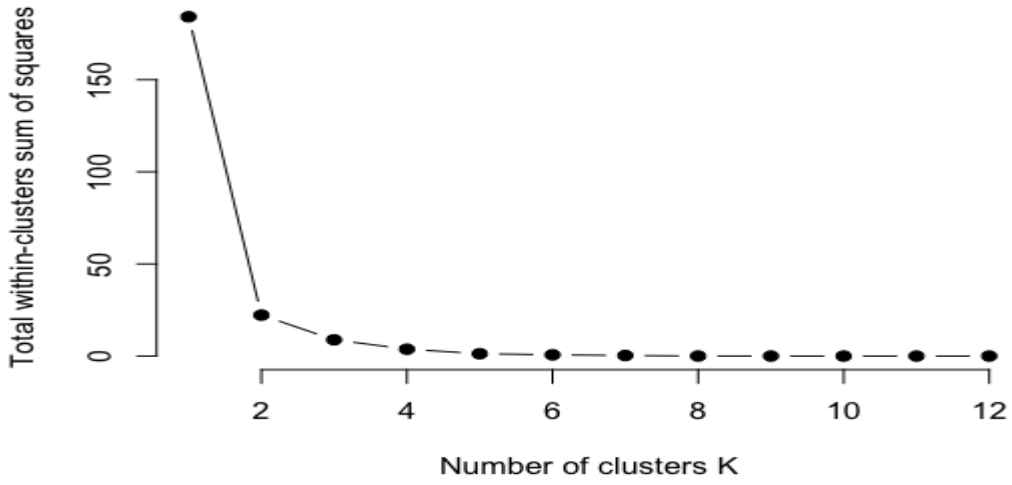
Clustering

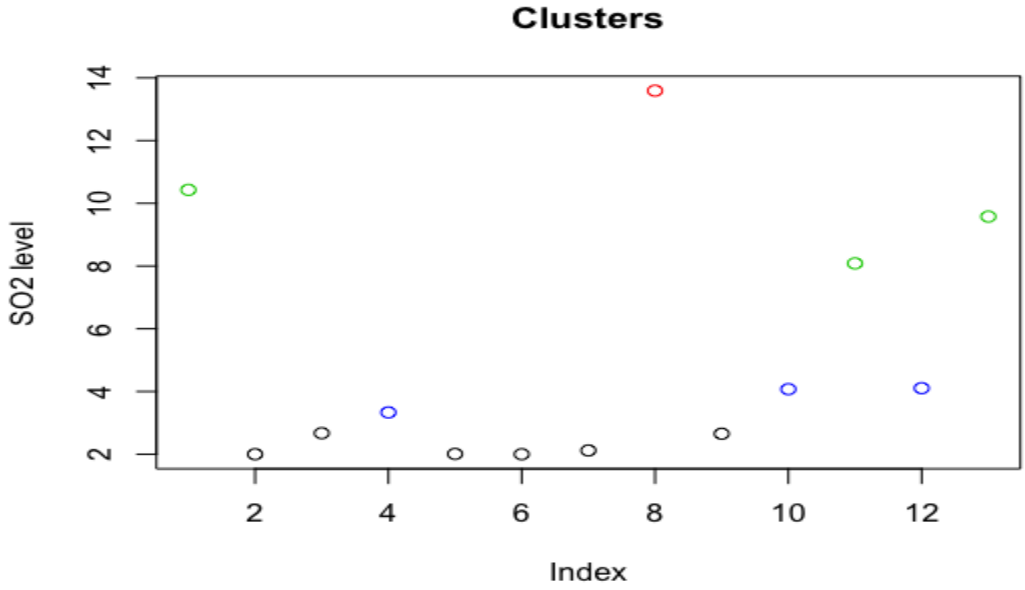
We have applied k-means clustering depending upon the values of the pollutants, so that we can get all those region that have similar pollutants level. We have used Elbow Curve method to determine optimal value of 'k'.

For SO_2

```
              1 2 3 4
Angul         0 0 1 0
Berhampur     1 0 0 0
Bhubaneswar   1 0 0 0
Cuttack       0 0 0 1
Kalinga Nagar 1 0 0 0
Keonjhar      1 0 0 0
Konark        1 0 0 0
Paradeep      0 1 0 0
Puri          1 0 0 0
Rayagada      0 0 0 1
Rourkela      0 0 1 0
Sambalpur     0 0 0 1
Talcher       0 0 1 0
```

```
K-means clustering with 4 clusters of sizes 6, 1, 3, 3

Cluster means:
       [,1]
1  2.242397
2 13.590909
3  9.363235
4  3.838002

Clustering vector:
 [1] 3 1 1 4 1 1 1 2 1 4 3 4 3

Within cluster sum of squares by cluster:
[1] 0.5393542 0.0000000 2.8165079 0.3826004
 (between_SS / total_SS =  98.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

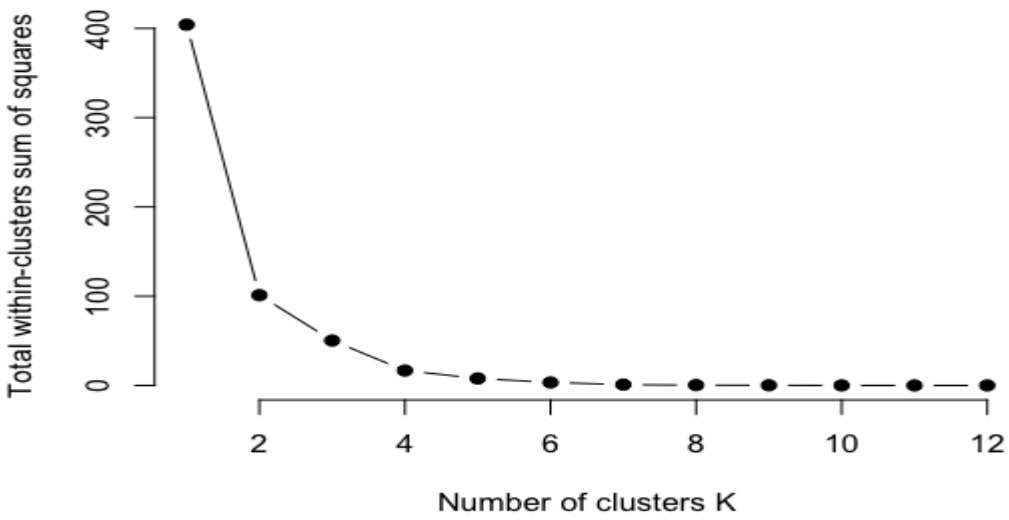


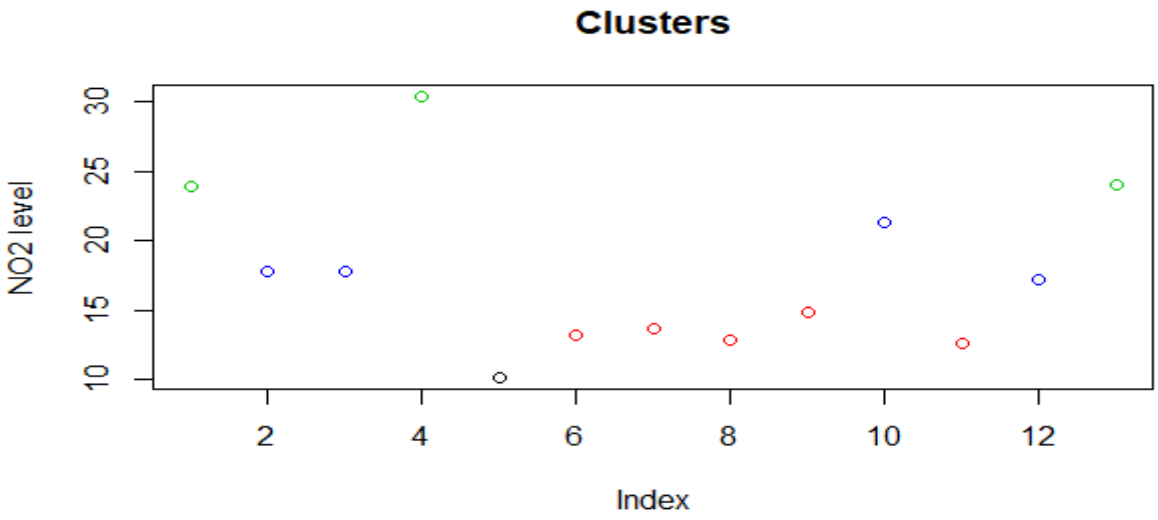


Above results show that SO_2 level is **highest in Paradeep** (**Red cluster**). Then comes **Angul, Rourkela** and **Talcher** (**Blue cluster**) with moderate pollution levels, followed by **Bhubaneswar, Rayagada** and **Sambalpur** (**Green Cluster**).
Lowest pollution levels are in **Berhampur, Bhubaneswar, Kalinga Nagar, Keonjhar, Konark** and **Puri** (**Black cluster**).

For NO_2

```
                1 2 3 4
  Angul         0 0 1 0
  Berhampur     0 0 0 1
  Bhubaneswar   0 0 0 1
  Cuttack       0 0 1 0
  Kalinga Nagar 1 0 0 0
  Keonjhar      0 1 0 0
  Konark        0 1 0 0
  Paradeep      0 1 0 0
  Puri          0 1 0 0
  Rayagada      0 0 0 1
  Rourkela      0 1 0 0
  Sambalpur     0 0 0 1
  Talcher       0 0 1 0

K-means clustering with 4 clusters of sizes 1, 5, 3, 4

Cluster means:
      [,1]
1 10.15686
2 13.41927
3 26.09392
4 18.53420

Clustering vector:
 [1] 3 4 4 3 1 2 2 2 2 4 2 4 3

Within cluster sum of squares by cluster:
[1]  0.000000  3.108065 27.523668 10.811507
 (between_SS / total_SS =  89.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```





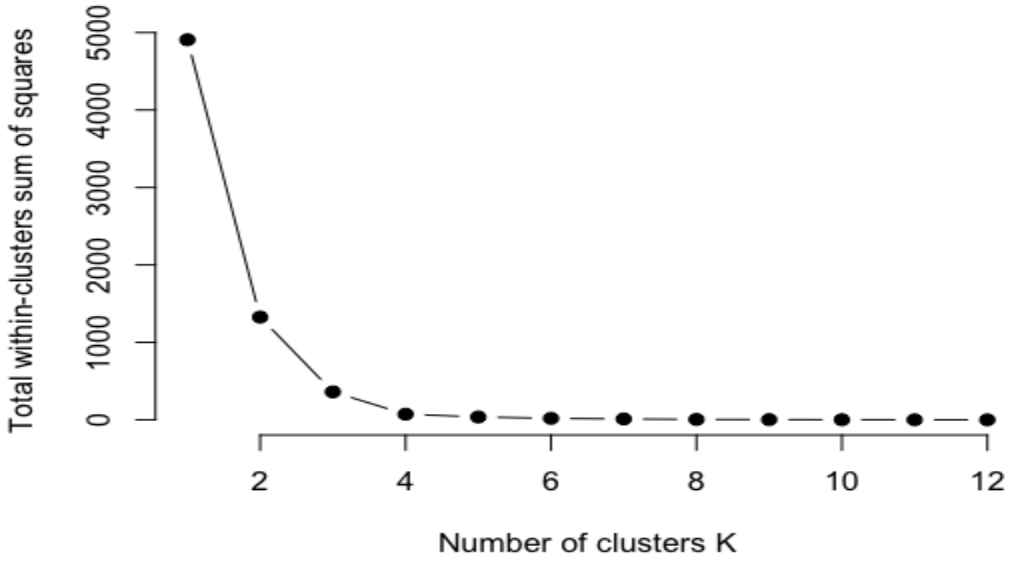
Above results show that NO_2 levels are highest in **Angul, Cuttack and Talcher** (**Green cluster**), then comes **Berhampur, Bhubaneswar, Rayagada** and **Sambalpur** (**Blue cluster**) , followed by **Keonjhar, Konark, Paradeep, Puri and Rourkela** (**Red cluster**).

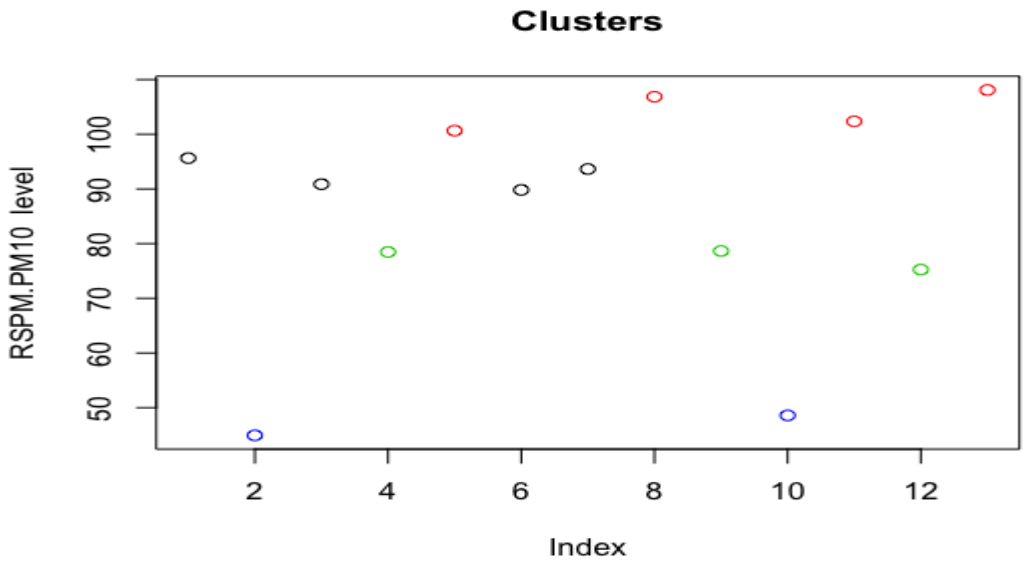
Lowest pollution levels are in **Kalinga Nagar** (**Black cluster**).

For PM10

```
               1 2 3 4
 Angul         1 0 0 0
 Berhampur     0 0 0 1
 Bhubaneswar   1 0 0 0
 Cuttack       0 0 1 0
 Kalinga Nagar 0 1 0 0
 Keonjhar      1 0 0 0
 Konark        1 0 0 0
 Paradeep      0 1 0 0
 Puri          0 0 1 0
 Rayagada      0 0 0 1
 Rourkela      0 1 0 0
 Sambalpur     0 0 1 0
 Talcher       0 1 0 0
```

```
K-means clustering with 4 clusters of sizes 4, 4, 3, 2

Cluster means:
       [,1]
1  92.49712
2 104.49872
3  77.46070
4  46.76471

Clustering vector:
 [1] 1 4 1 3 2 1 1 2 3 4 2 3 2

Within cluster sum of squares by cluster:
[1] 20.949639 37.757166  7.346222  6.650519
 (between_SS / total_SS =  98.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```



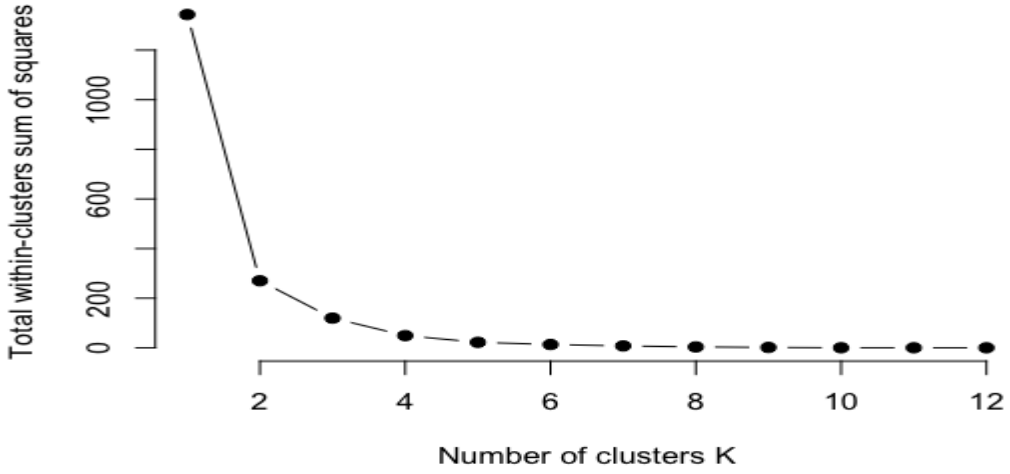


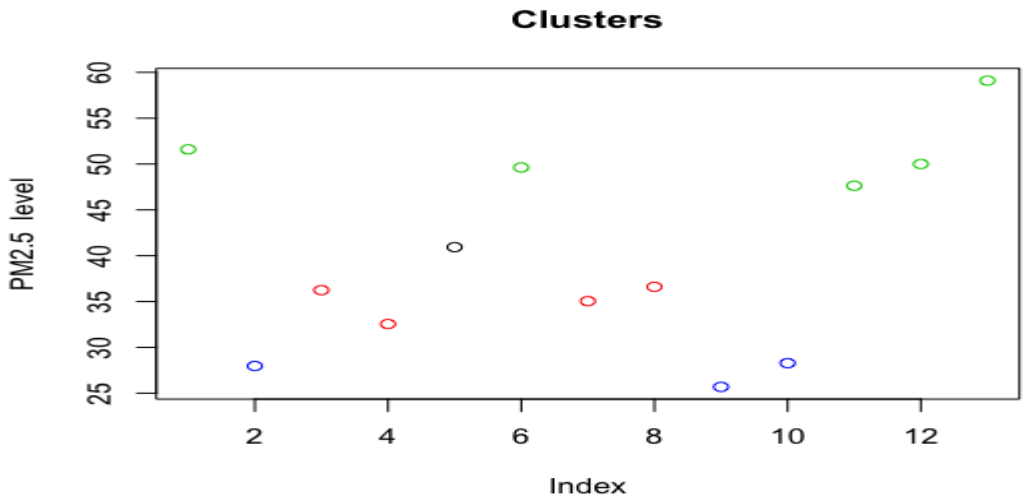
Above results show that RSPM10 level is **highest in Kalinga Nagar**, **Paradeep, Rourkela** and **Talcher** (**Red cluster**). Then comes **Angul, Bhubaneswar, Keonjhar** and **Konark** (**Black cluster**) with moderate pollution levels, followed by **Cuttack, Puri** and **Sambalpur** (**Green cluster**). Lowest pollution levels are in **Berhampur** and **Rayagada** (**Blue cluster**).

For PM2.5

	1	2	3	4
Angul	0	0	1	0
Berhampur	0	0	0	1
Bhubaneswar	0	1	0	0
Cuttack	0	1	0	0
Kalinga Nagar	1	0	0	0
Keonjhar	0	0	1	0
Konark	0	1	0	0
Paradeep	0	1	0	0
Puri	0	0	0	1
Rayagada	0	0	0	1
Rourkela	0	0	1	0
Sambalpur	0	0	1	0
Talcher	0	0	1	0

```
K-means clustering with 4 clusters of sizes 1, 4, 5, 3

Cluster means:
      [,1]
1 40.93137
2 35.11017
3 51.58968
4 27.31410

Clustering vector:
 [1] 3 4 2 2 1 3 2 2 4 4 3 3 3

Within cluster sum of squares by cluster:
[1]  0.000000 10.078722 78.417826  3.993014
 (between_SS / total_SS =  93.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```



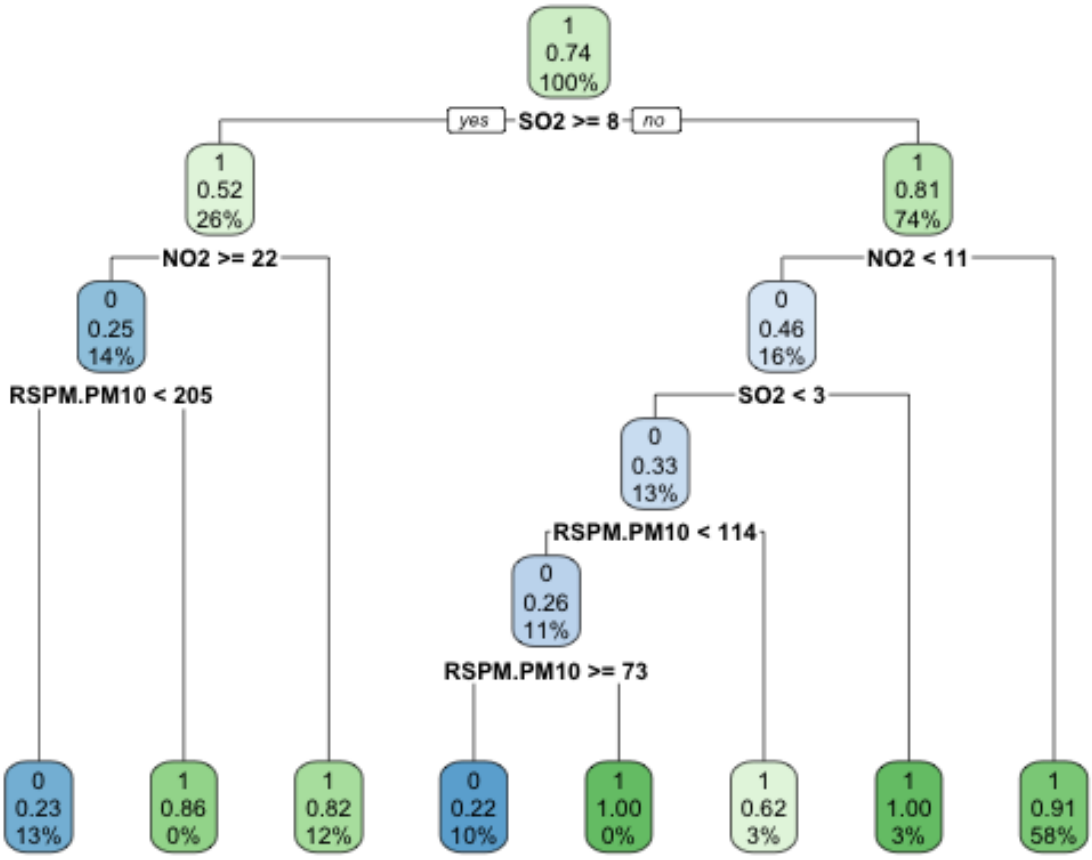


Above results show that PM2.5 level is **highest in Angul, Keonjhar, Rourkela, Sambalpur** and **Talcher** (**Green cluster**). Then comes **Kalinga Nagar** (**Black cluster**) with moderate pollution levels, followed by **Bhubaneswar, Cuttack, Konark** and **Paradeep** (**Red cluster**). Lowest pollution levels are in **Berhampur, Puri** and **Rayagada** (**Blue cluster**).

Thus by using above inferences and result, Orissa Pollution Control Board can take proper actions in those areas which are highly polluted.

Decision Tree

Decision tree for predicting the type of location from training set.



Confusion matrix

(0 - Industrial, 1 – rural)

	prediction_from_decision_tree	
	0	1
0	70	39
1	32	272

Confusion Matrix details

```
Confusion Matrix and Statistics

   prediction_from_decision_tree
      0   1
  0  70  39
  1  32 272

                Accuracy : 0.8281
                  95% CI : (0.7882, 0.8632)
    No Information Rate : 0.753
    P-Value [Acc > NIR] : 0.0001545

                   Kappa : 0.5482

 Mcnemar's Test P-Value : 0.4764221

            Sensitivity : 0.6863
            Specificity : 0.8746
         Pos Pred Value : 0.6422
         Neg Pred Value : 0.8947
             Prevalence : 0.2470
         Detection Rate : 0.1695
   Detection Prevalence : 0.2639
      Balanced Accuracy : 0.7804

       'Positive' Class : 0
```

Logistic Regression on Type Of Location

```
Confusion Matrix and Statistics

     y_pred
y_act   0   1
    0  11  98
    1   8 296

               Accuracy : 0.7433
                 95% CI : (0.6984, 0.7848)
    No Information Rate : 0.954
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1015

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.57895
            Specificity : 0.75127
         Pos Pred Value : 0.10092
         Neg Pred Value : 0.97368
             Prevalence : 0.04600
         Detection Rate : 0.02663
   Detection Prevalence : 0.26392
      Balanced Accuracy : 0.66511

       'Positive' Class : 0
```

From the results of Decision Tree analysis and Logistic Regression we concluded that the former shows better results and can be used to predict the type of location.