DATA ANALYTICS

Unit 1:Case Study "Identifying Patterns in New Delhi's Air Pollution"

Mamatha.H.R

Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1:Case Study
“Identifying Patterns in New Delhi’s Air Pollution”

Mamatha H R
Department of Computer Science and Engineering

Overview

- The rate at which urban air pollution has grown across India is alarming.
- A vast majority of cities are caught in the toxic web as air quality fails to meet health-based standards.
- Almost all cities are reeling under severe particulate pollution while newer pollutants like oxides of nitrogen and air toxics have begun to add to the public health challenge
- Exposure to particulate matter for a long time can lead to respiratory and cardiovascular diseases such as asthma, bronchitis, lung cancer and heart attack.

Overview

- The Global Burden of Disease study pinned outdoor air pollution as the fifth largest killer in India, after high blood pressure, indoor air pollution, tobacco smoking, and poor nutrition.

- In 2010, about 620,000 early deaths in India occurred from air pollution-related diseases.

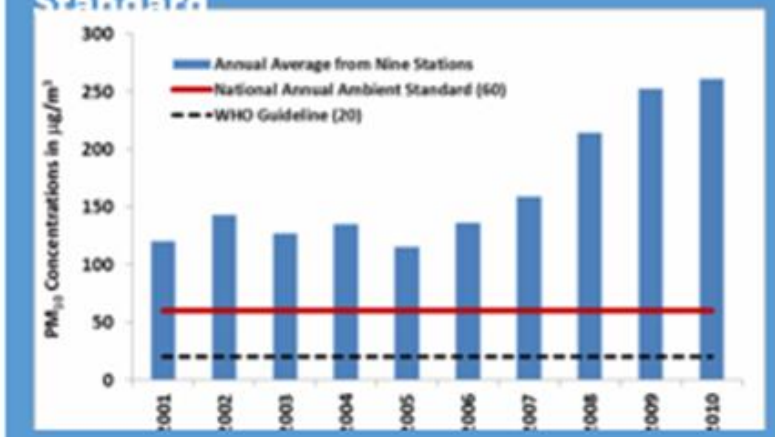
- The Central Pollution Control Board (CPCB) sponsored the study that links the pollutants, pm 10 (particulate matter smaller than 10 microns), the cause of these diseases.

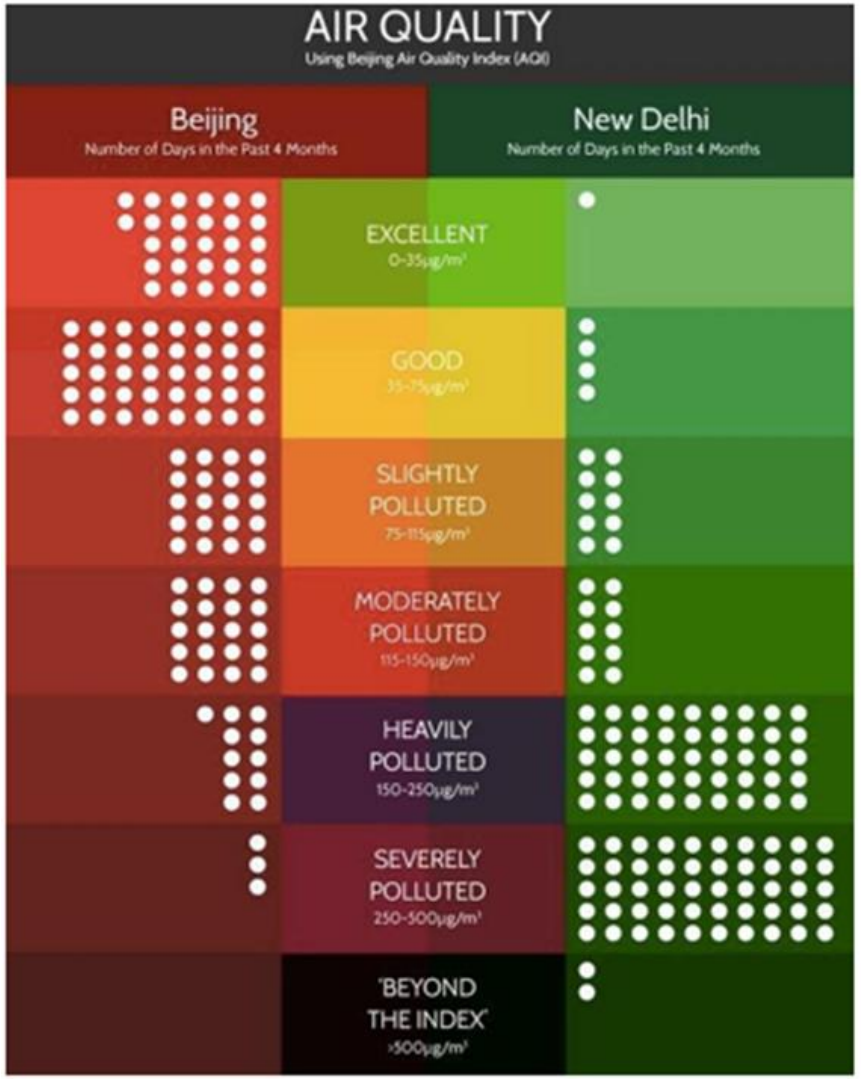
DATA ANALYTICS

Overview

WHO says India ranks among the world's worst for its polluted air. Delhi is among the most polluted cities in the world today.

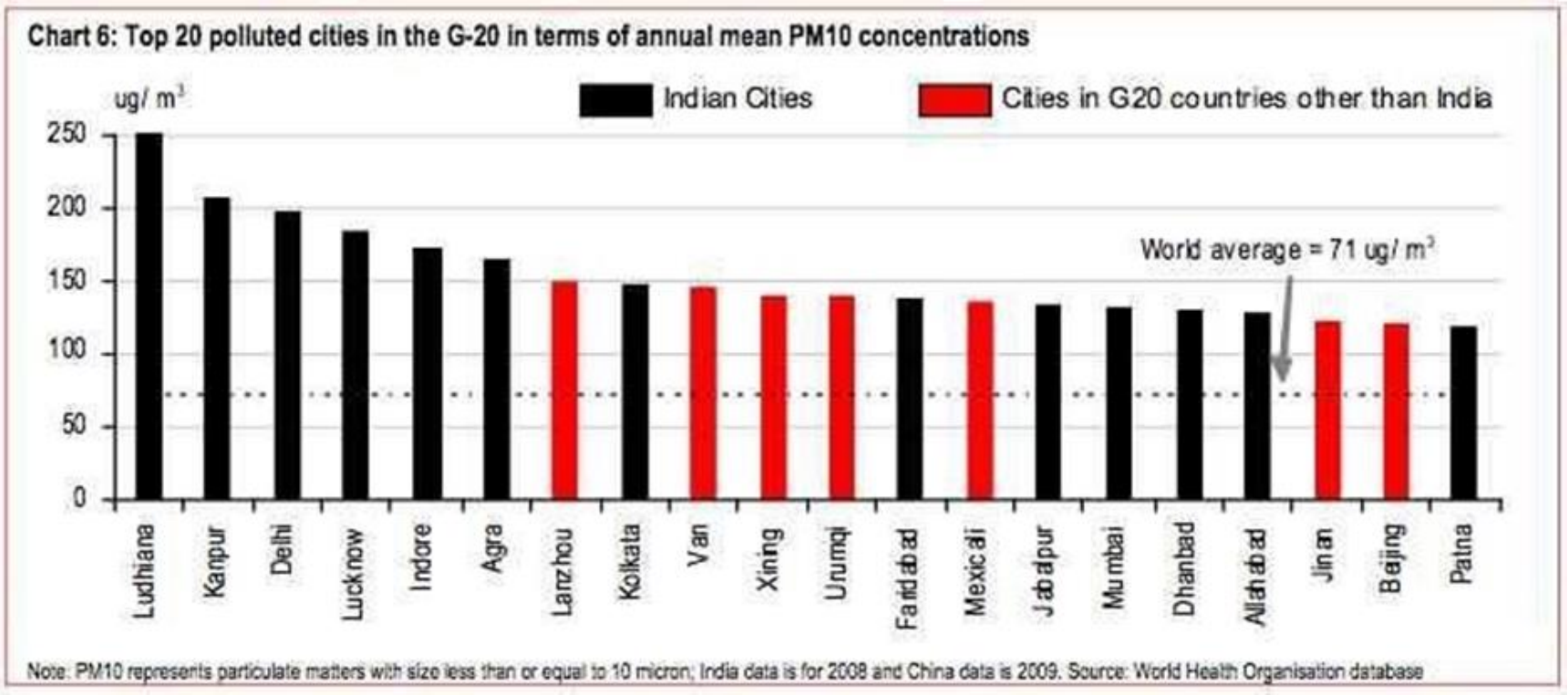
Figure 2: Chart showing New Delhi's PM10 Levels over a 10-year period against Indian Standard & WHO Standard





DATA ANALYTICS

Overview



DATA ANALYTICS

Problem Statement

To develop some insights that can help organizations (State / Central Pollution Control Boards & NGOs) to advocate more stringent policies to control air pollution.

Objective

The primary objectives of the study are:

- Study Air Pollution Data for various locations in New Delhi to identify patterns of spike in Air Pollution levels w.r.t to various monitored parameters
- Identify the Meteorological factors that correlate with the air pollution levels for the respective locations
- Explore the possibility of developing a Predictive Model for predicting the levels for key pollutants like PM 5
- Study the Odd-Even Pilot Project (Phase II) and its impact on air pollution levels in New Delhi. As part of this, also study the people's response to this by studying the social conversation around 'Odd-Even'

DATA ANALYTICS

Scope

- The scope of the study covers 3 major polluting centers in New Delhi
- The study covers one-year Data starting from 1st April'15. This is done to ensure seasonality factors are covered
- The Study's focus is on factors for which authentic secondary data are available that can be used for Statistical Analysis

Out of Scope

- Experimental measures like developing first-hand data are not considered i.e. factors like Vehicle density during the given period at each location, measuring & monitoring level of road dust, Industrial pollution
- The scope of the study will cover 3 to 4 major cities in India and will include 2-3 key monitoring stations per city (depending on the data availability)
- The study will cover up to one year data starting 1st April’15 to 31st March’16. This is done to ensure seasonality factors are covered

DATA ANALYTICS

Data Source

The data for the Project was obtained from the website of Central Pollution Control Board (CPCB). Currently, CPCB tracks the Air Pollution levels across 23 dimension (variables). Day wise, hour wise (for some variables). Data is available on-line across the following dimensions:

1. Nitric Oxide (NO)
2. Carbon Monoxide(CO)
3. Suspended Particulate Matter/RPM/PM10
4. Nitrogen Dioxide (No2)
5. Ozone
6. Sulphur Dioxide (SO2)
7. PM 2.5 (DUST 5)
8. Toluene
9. Ethyl Benzene (Ethylben)
7. M & P Xylene
8. Oxylene
9. Oxides of Nitrogen (Nox)
10. PM10 DUST
11. PM10 RSPM
12. Ammonia NM3
13. Non Methane Hydro Carbon (NMHC)
17. Total Hydro carbon (THC)
18. Relative Humidity (RH)
19. Temperature
20. Wind Speed (Wind speed S)
21. Vertical Wind speed (Wind speed V)
22. Wind Direction
23. Solar Radiation

DATA ANALYTICS

Tools and Techniques

The following Analytical techniques / methodology are used for analyzing the Data :

1. Summary of Statistics for each variable
2. Identification of frequency of standard violation for each of the factors
3. Using Graphs and Box Plots to visually represent them
4. Identification of significant Metrological factors through correlation and regression methodology
5. Using Multiple Linear Regression & Neural Network for Model Development

DATA ANALYTICS

Tools and Techniques

1. Tools used: R, Tableau & Excel
2. Techniques: Box Plot, Histogram, Bar Chart, Line Chart, Infographics, Visual Clues, Correlation Matrix, Multiple Linear Regression, Artificial Neural Network
3. We have used R Programming environment and Microsoft Excel for our analysis and Tableau for data

Analytics approach

The Analytical Approach will involve the following (not necessarily in the order) activities:

- Data extraction from Primary Data source as well as secondary data sources
- Data quality check
- Data cleaning and data preparation
- Study each of the variables by exploring the data
- Study the variables for its relevance for the study
- Identifying Y variable(s).

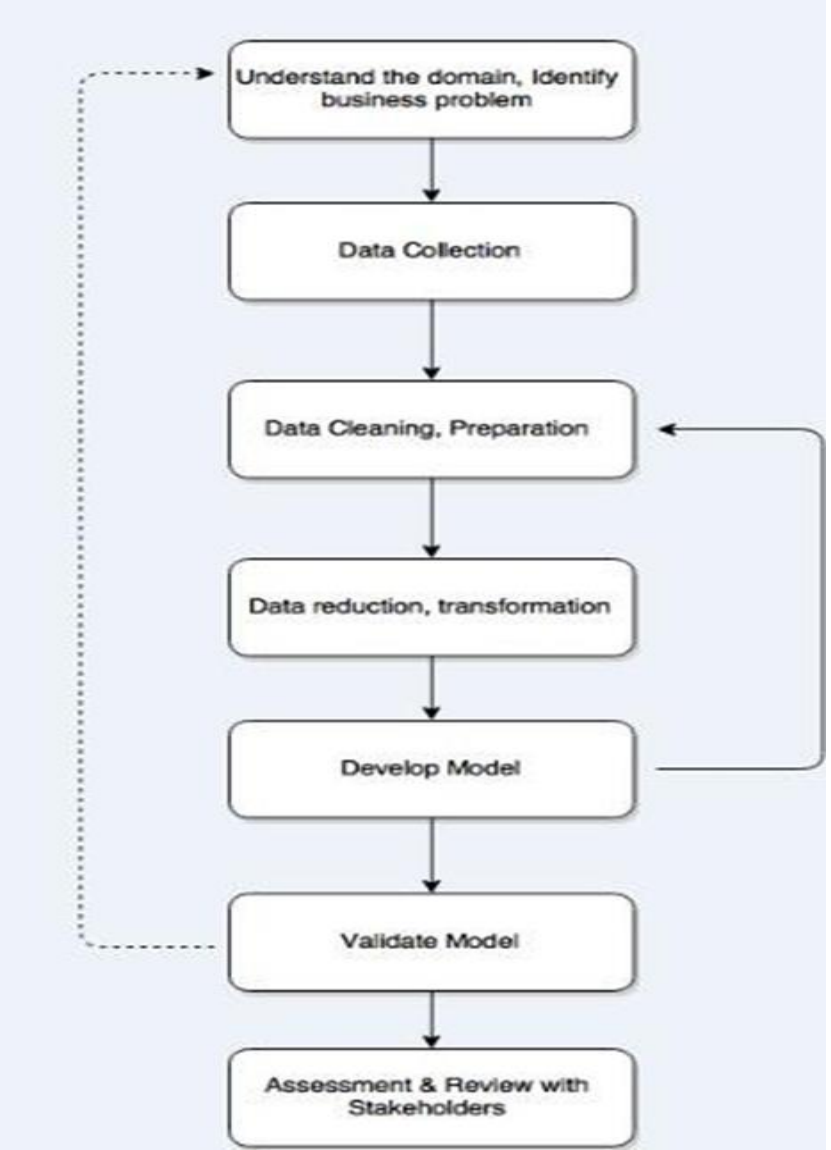
DATA ANALYTICS

Analytics approach

- Performing Univariate analysis for all variables
- Division of data into train and test
- Model Development
- Final Model
- Model Validation & Model Validation on Test
- Intervention Strategies and recommendations

DATA ANALYTICS

Seven Step Analytical Approach



Limitations

There are few limitations that this study has w.r.t data and the methodology that can be used.

- Due to time and cost constraints we could not deploy a primary source for data collection. We were not in a position to deploy primary pollution data collection by deploying near ground level monitoring system that are typically used in advanced countries for such Air Pollution studies. They help accurately capture the road level air pollution contributed maximum by the automobiles.
- Due to a very short window of 15 days for the Odd-Even Campaign, we had to live with a very small data size rendering the data unusable for any kind of rigorous statistical analysis.

Limitations

- Since the Analysis & Models were built specifically for a particular location, the insights and the Models cannot be used for other locations in New Delhi or for other locations outside New Delhi.
- Since the Models were built on rather small data size (about a year), the models need to be strengthened with data of at least another year or two. Till then, the Models are likely to work in a larger range,i.e. the variance is likely to be higher.

DATA ANALYTICS

Data Description and Preparation -Data Management

Table: List of Variables and Their Type				
Variable Abbreviation	Variable	Variable type	Unit of Measurement	Data Type
NO	Nitric Oxide	Pollutant	µg/m3	Continuous
CO	Carbon Monoxide	Pollutant	mg/m3	Continuous
NO2	Nitrogen Dioxide	Pollutant	µg/m3	Continuous
OZONE	Ozone	Pollutant	µg/m³	Continuous
SO2	Sulphur Dioxide	Pollutant	µg/m3	Continuous
NOx	Oxides of Nitrogen	Pollutant	µg/m³	Continuous
RSPM	Respiratory Suspended Particulate Matter	Pollutant	µg/m³	Continuous
PM2.5	Particulate Matter less than 2.5 Micrometer	Pollutant	µg/m³	Continuous
PM10	Particulate Matter less than 10 Micrometer	Pollutant	µg/m³	Continuous
Benzene	Benzene	Pollutant	µg/m³	Continuous
Toluene	Toluene	Pollutant	µg/m³	Continuous
Ethylene	Ethyl Benzene	Pollutant	µg/m³	Continuous
M_P_Xylene	M & P Xylene	Pollutant	µg/m³	Continuous
O_Xylene	O Xylene	Pollutant	µg/m³	Continuous
P_Xylene	P Xylene	Pollutant	µg/m³	Continuous
NH3	Ammonia	Pollutant	µg/m³	Continuous
CH4	Methane	Pollutant	µg/m³	Continuous
NMHC	Non Methane Hydro Carbon	Pollutant	µg/m³	Continuous
THC	Total Hydro Carbon	Pollutant	µg/m³	Continuous
RH	Relative Hydrocarbon	Meteorological	%	Continuous
Temp	Temperature	Meteorological	°C	Continuous
WS	Wind Speed	Meteorological	m/s	Continuous
VWS	Vertical Wind Speed	Meteorological	m/s	Continuous
WD	Wind Direction	Meteorological	°	Continuous
SR	Solar Radiation	Meteorological	W/m2	Continuous
Bar Pressure	Bar Pressure	Meteorological	mmHg	Continuous

DATA ANALYTICS

Data Quality

1. Pollutant level data for certain days was missing. Some days had data for only few of the variables. Data for the days where there was no data for key variables like PM 2.5, PM 10, NO2, SO2, CO were removed. There was no data available for few of the days on the source system itself.
2. Especially for Odd-Even Campaign, data was not reported for a few days (already on a short window of 15 days pre-campaign and 15 days post campaign) on the source system. After plummeting all such variables and observations, the data was merged.
3. There were 26 variables with 284 records for Anand Vihar; 289 records for Punjabi Bagh & 345 records for R.K. Puram

Data Preparation

Variables Transformation

1. For building the Multiple Linear Regression Model, all the variables were transformed using logarithm
2. For Neural Network, no data transformation .

Data Preparation

Missing values and Outliers

1. No specific missing value treatment was used
2. Days for which no data was available for the key variables, then that day's record was removed from analysis
3. Only days where observations were recorded for the key variables were included in the analysis
4. Days in which outliers were present, the day's record was removed from the data

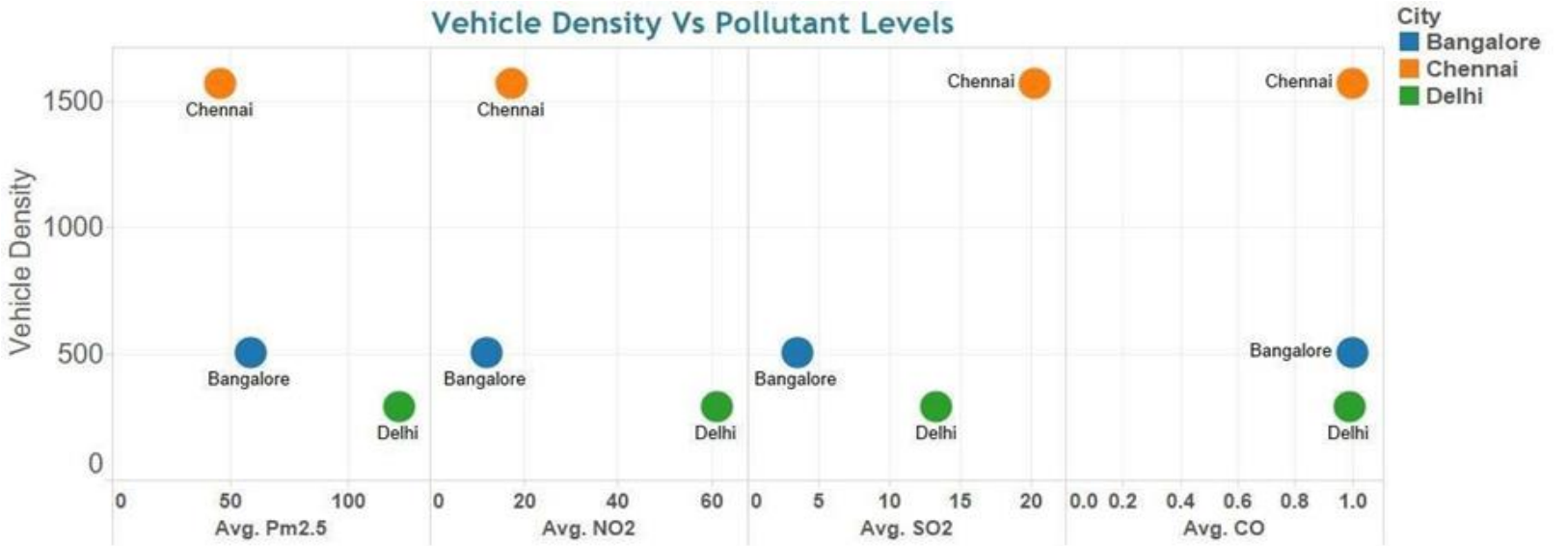
Exploratory Data Analysis

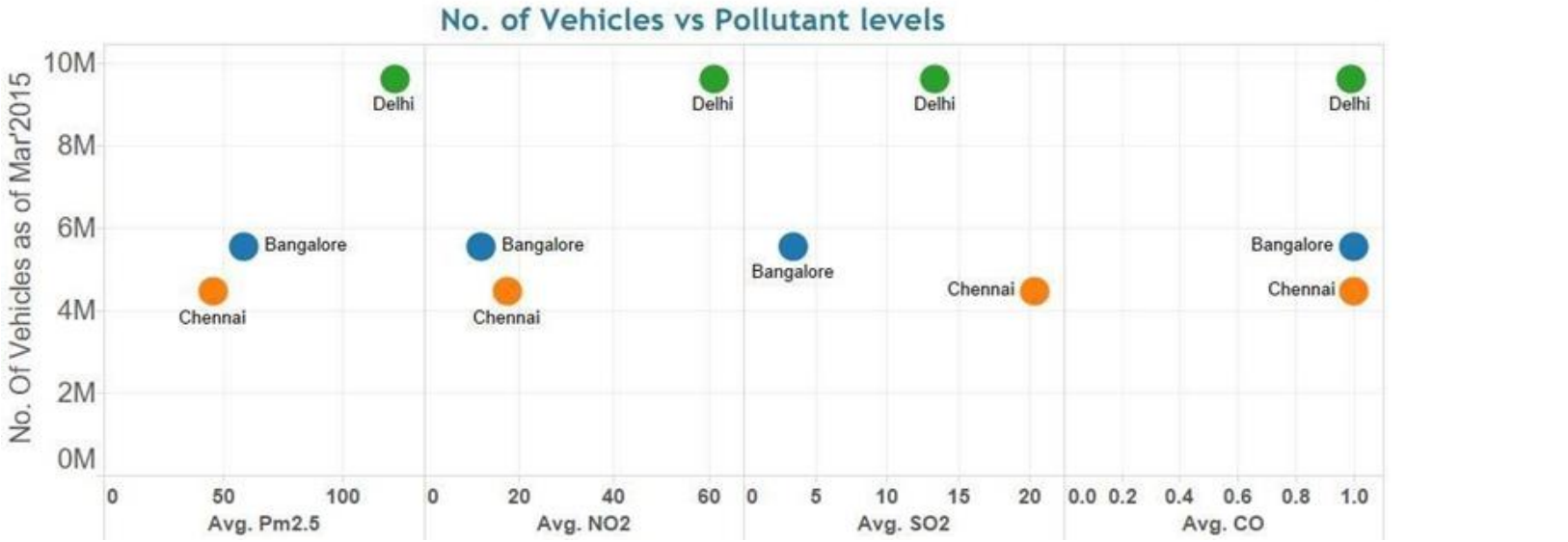
The Exploratory Data Analysis is divided into three parts. They are:

1. Analyzing three cities air pollution data and check whether the number of vehicles & vehicle density have any impact on air pollution levels
2. Analyzing the data of three locations of New Delhi across various factors and find out any correlation exists between the factors
3. Analyzing the New Delhi Data to find out the impact of 'Odd-Even' experiment on the pollution levels (i.e. measured across 4/5 key parameters). Also, explore the social data and do a sentimental analysis for gauging people's reaction to the experiment

DATA ANALYTICS

Analyzing the impact of Vehicle Density & Vehicle Population





Insights

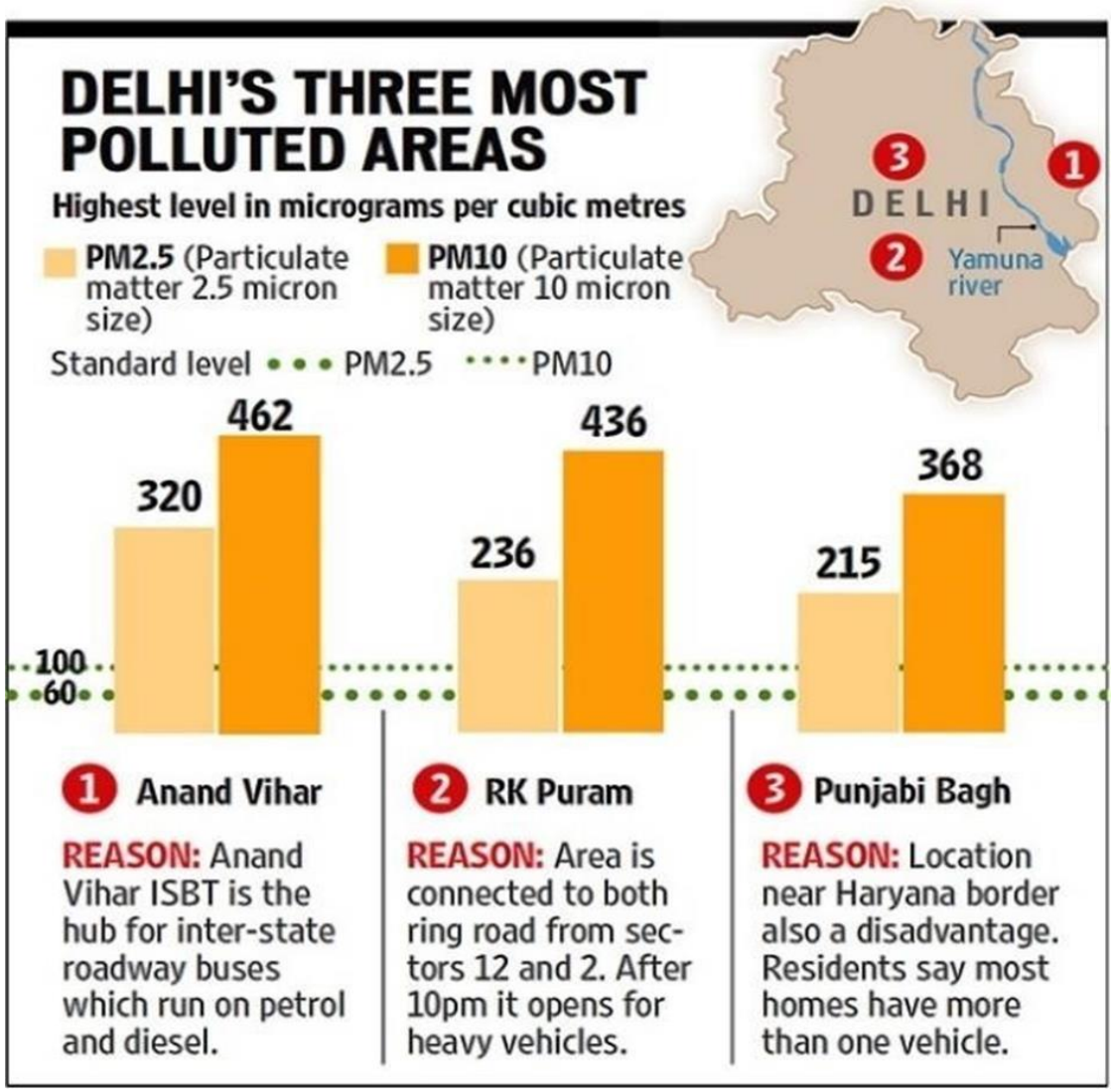
- Vehicle density (measured as vehicles/km of road) does not have any impact on the air pollution.
- New Delhi has the least vehicle density amongst the three cities we have considered for the study, but the PM 2.5 levels are significantly higher in New Delhi as compared to Bangalore and Chennai.
- Though Chennai has the highest density of vehicles, but has a lower pollution levels for (PM 2.5)

DATA ANALYTICS

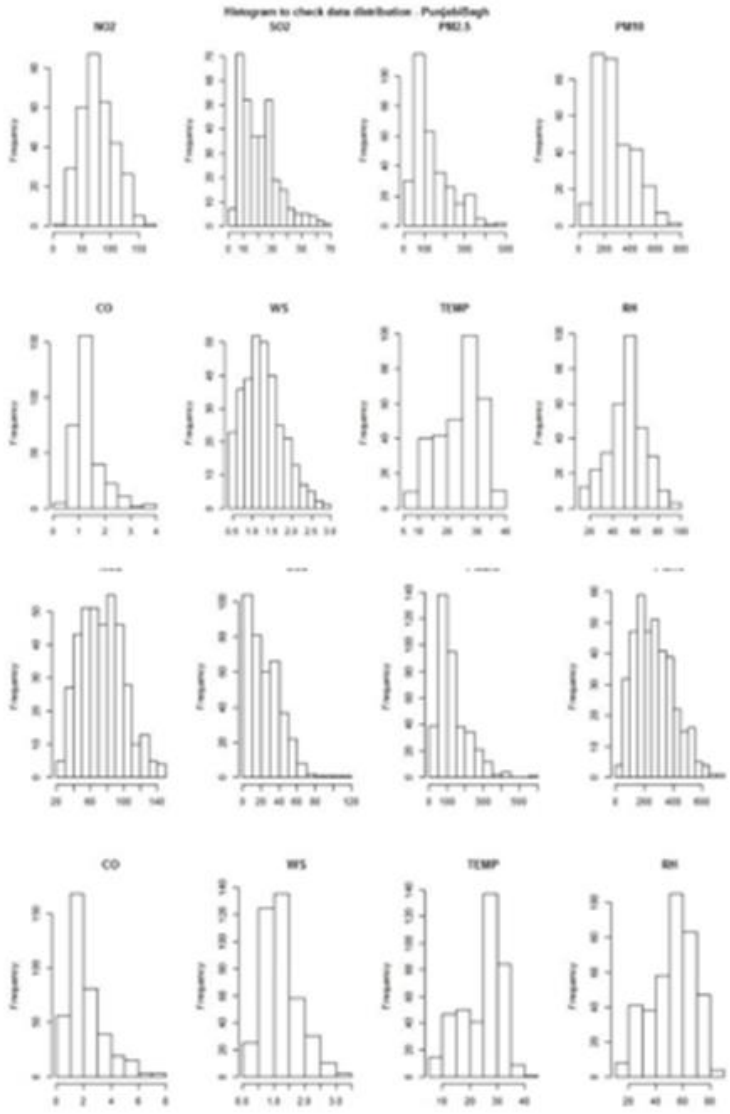
Insights

1. If you consider the absolute vehicle population, then there seem to be a positive correlation between the number of vehicles and the air pollution levels of PM 2.5 & to a lesser extent on NO2.
2. CO levels does not seem to have any correlation with either vehicle density or with vehicle population as the levels of CO are almost at same levels across the 3 cities.
3. The result indicates the factors, other than vehicular pollution, contributing to the overall air pollution in the three cities are almost equal.
4. New Delhi has wide roads, so the vehicle density tends to get averaged out to a lower number.
5. But, there is a high probability that the vehicle density in many of the observatory locations are high and contributing to higher air pollution levels

Identifying Patterns for Air Pollution in New Delhi

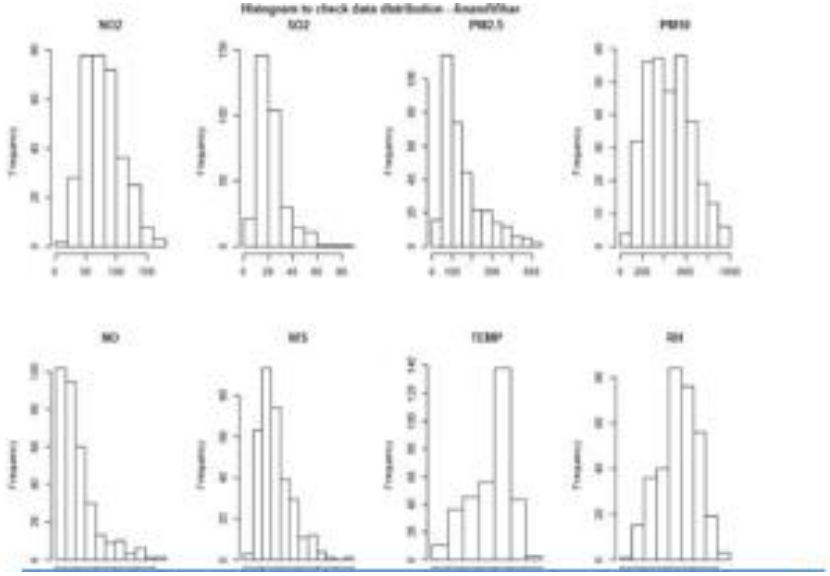


Histogram for Various Pollutants



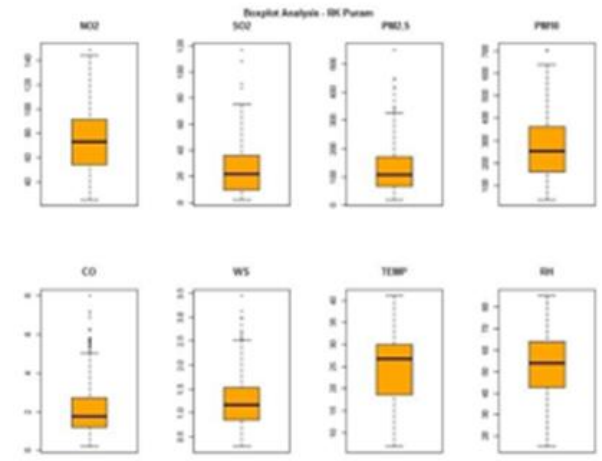
The histogram shows a few key attributes about the distribution of the different pollutants.

- Distribution is asymmetric – Left or right skewed
- Distribution is Unimodal in most pollutant data.
- There are some Outliers near the low and high ends

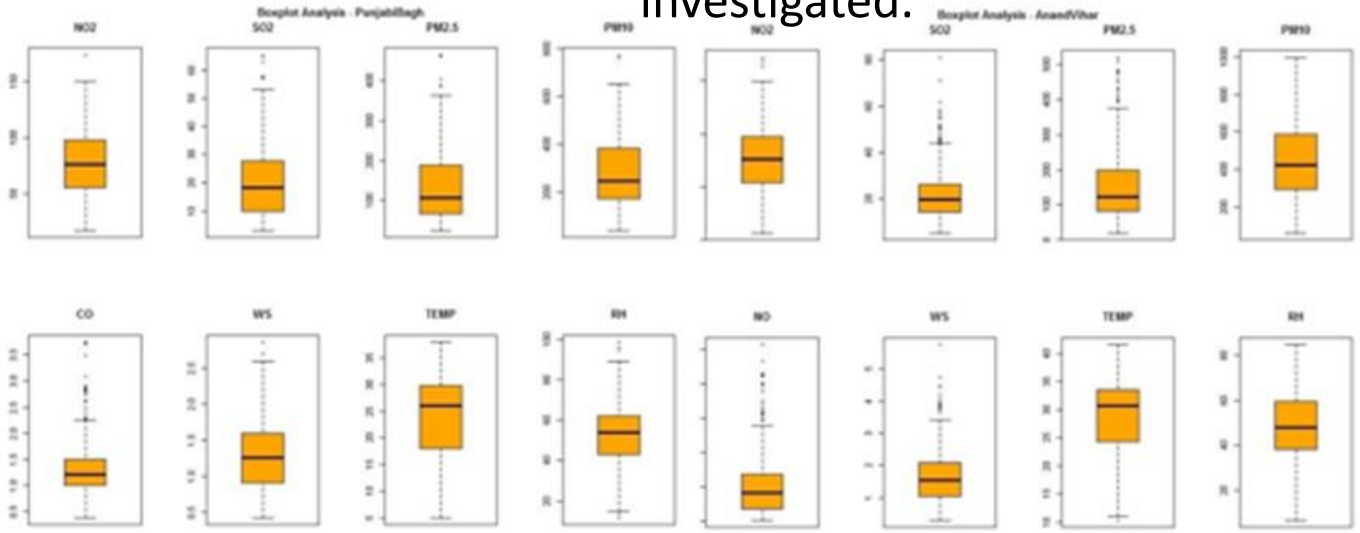


DATA ANALYTICS

Box Plot for Various Pollutants – All Locations



1. All the pollutants are almost at the same level in the 3 areas (Centres and spreads are equally likely for all 3 areas).
2. Indicating the areas between Anand Vihar, Punjabi Bagh and RK Puram are equally polluted.
3. The data has outliers caused by external factors and that needs to be investigated.



DATA ANALYTICS

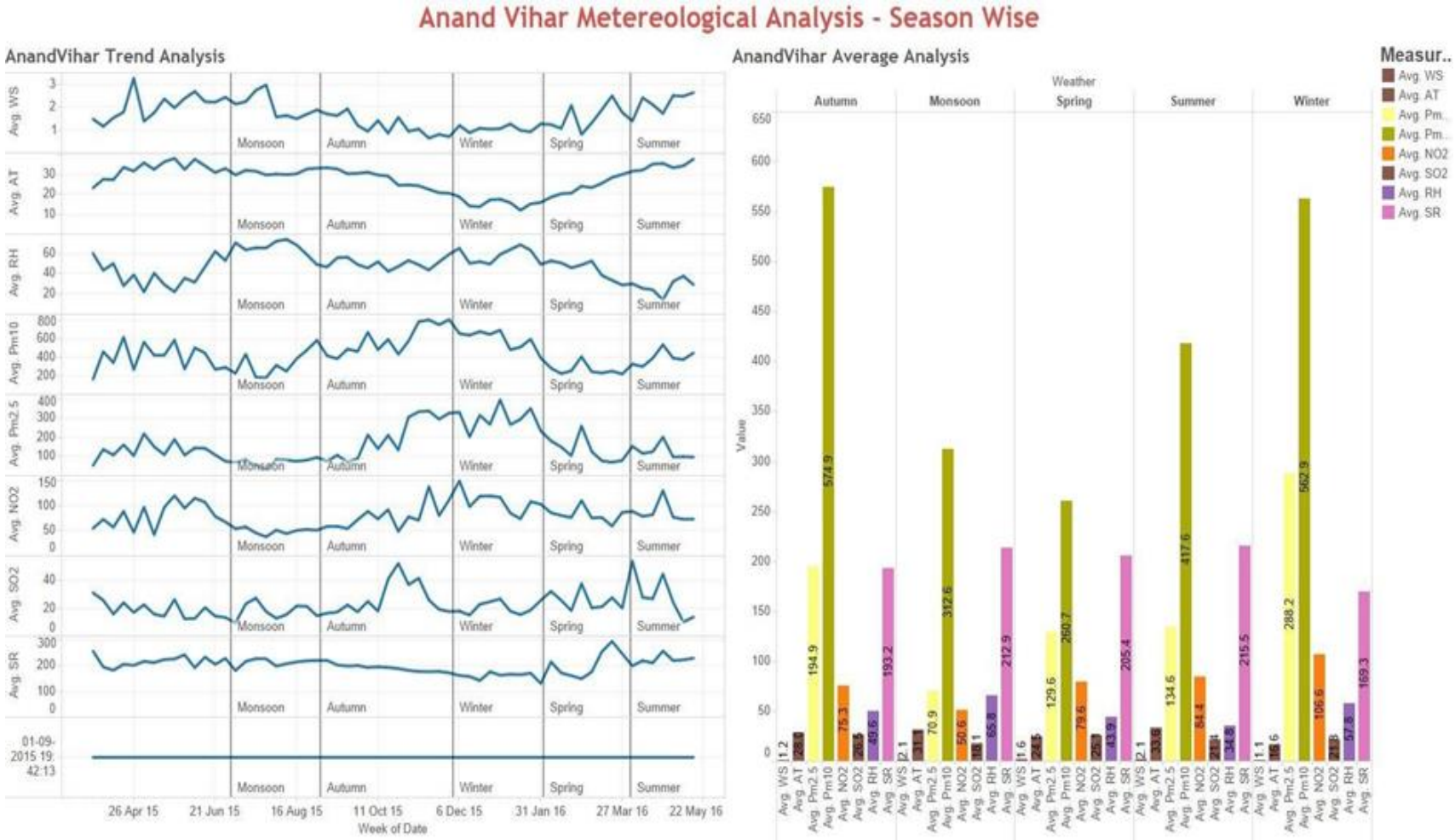
Summary of Data for Key Variables for each Location

WS	TEMP	WD	RH	SR
Min. :0.300	Min. :10.30	Min. : 63.74	Min. : 6.52	Min. : 12.29
1st Qu.:1.040	1st Qu.:22.89	1st Qu.:133.81	1st Qu.:39.13	1st Qu.:176.89
Median :1.520	Median :30.41	Median :194.77	Median :49.20	Median :204.90
Mean :1.699	Mean :28.21	Mean :189.52	Mean :48.43	Mean :201.25
3rd Qu.:2.060	3rd Qu.:33.37	3rd Qu.:247.50	3rd Qu.:60.23	3rd Qu.:221.54
Max. :5.760	Max. :41.54	Max. :287.03	Max. :84.86	Max. :429.69

Bar.Pressure	NO2	SO2	PM2.5	PM10
Min. :739.0	Min. : 6.55	Min. : 5.33	Min. : 19.51	Min. : 60.65
1st Qu.:740.0	1st Qu.: 54.88	1st Qu.: 14.51	1st Qu.: 83.11	1st Qu.:297.94
Median :740.0	Median : 76.29	Median : 19.56	Median :128.58	Median :429.78
Mean :739.9	Mean : 78.65	Mean : 22.54	Mean :161.93	Mean :450.64
3rd Qu.:740.0	3rd Qu.: 97.14	3rd Qu.: 25.91	3rd Qu.:213.56	3rd Qu.:586.75
Max. :740.0	Max. :279.51	Max. :101.10	Max. :519.68	Max. :996.62

DATA ANALYTICS

Seasonality Analysis



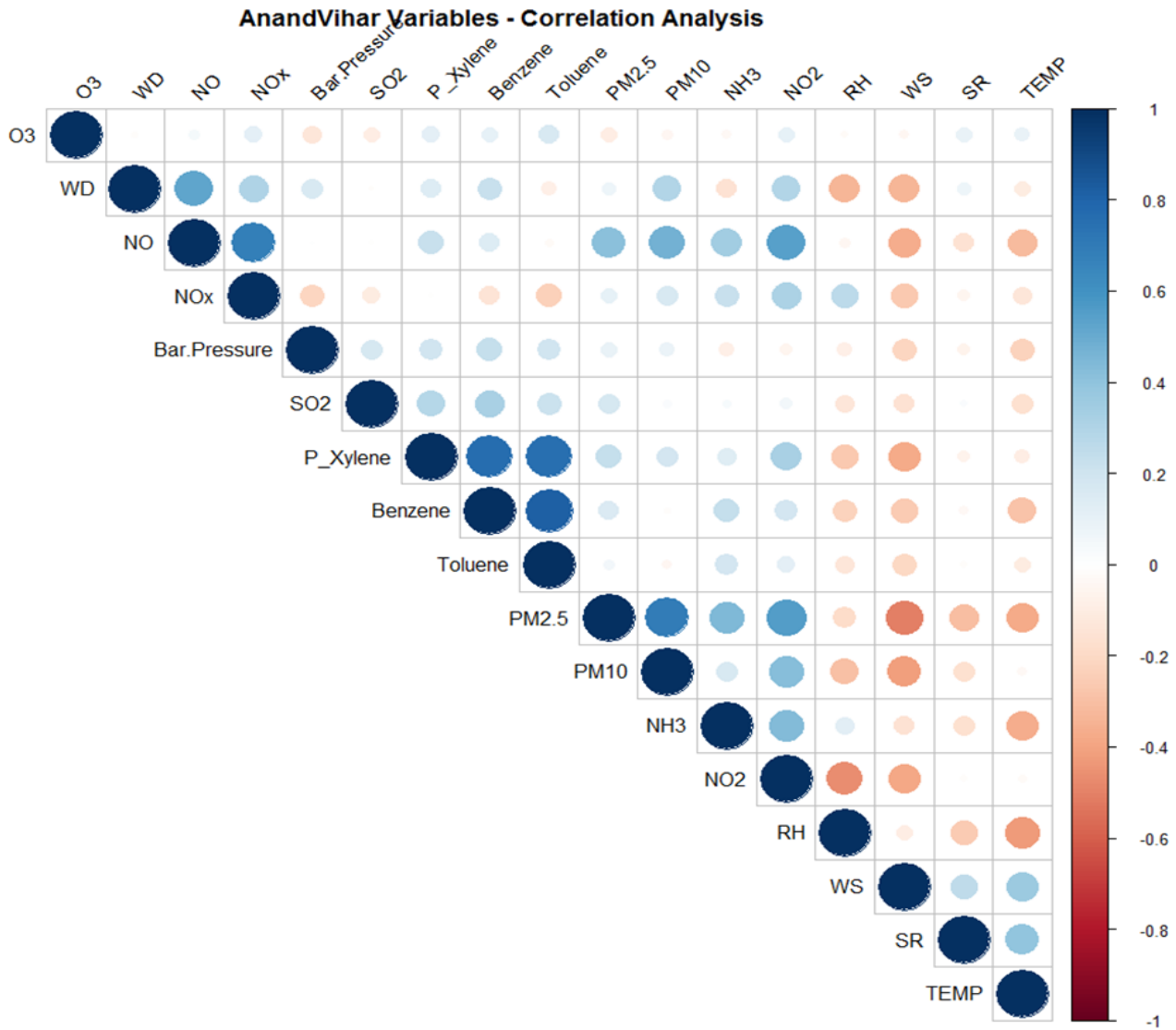
DATA ANALYTICS

Seasonality Analysis : Conclusion

1. Concentration of Particulate matter known as PM2.5 and PM10 are lower during Monsoon (July-August)
2. 5 and PM10 averages are exceeding its permissible values of 60 µg/m3 and 100 µg/m3 during WINTER (November-January) followed by AUTUMN (September-October), SUMMER (April-June) and to a lesser extend during SPRING (February-March)
3. Some kind of association between PM 2.5/PM 10 levels and Wind Speed as well as Temp can be seen in the graph
    - Relatively lower Pollution levels seem to be associated with higher Wind Speed
    - Very low Atmospheric Temperature is associated with relatively higher Pollution levels of PM 2.5/PM 10
4. Other pollutants data remains significantly same throughout the year except for NO2, peaks during winter and is at its lowest during monsoon

DATA ANALYTICS

Correlation Matrix & Analysis: Anand Vihar



DATA ANALYTICS

Insights

- PM 2.5 & 10 have a strong negative correlation with Wind Speed
- Temp has a negative correlation with PM 2.5, NH3 & Relative Humidity
- PM 2.5 also has a positive correlation with NO2
- Xylene, Toluene & Benzene are positively correlated with each other

DATA ANALYTICS

Exercise

Analysis and visualization of COVID-19 Pandemic spread.
-COVID-19 Open Research Dataset (CORD-19).

References

https://www.analyticsvidhya.com/blog/2016/10/complete-study-of-factors-contributing-to-air-pollution/

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834