INDIA'S AIR QUALITY ANALYSIS

Amanda Rachmani Artyan

BACKGROUND

At least **140 million people** in India breathe air that is 10 times or more over the WHO safe limit in 2016

30 most **polluted cities** in the world, 21 were in **India** in 2019.

- **51%** of the pollution is caused by **industrial pollution**
- 27% by vehicles
- 17% by crop burning
- 5% by other sources

OBJECTIVES

To minimize the air - pollution in India : Analysis features for gain more information & optimize metric ML to predicting data

DATA PREPROCESSING

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 707875 entries, 0 to 707874
Data columns (total 16 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   City        707875 non-null  object
 1   Datetime    707875 non-null  object
 2   PM2.5       562787 non-null  float64
 3   PM10        411138 non-null  float64
 4   NO          591243 non-null  float64
 5   NO2         590753 non-null  float64
 6   NOx         584651 non-null  float64
 7   NH3         435333 non-null  float64
 8   CO          621358 non-null  float64
 9   SO2         577502 non-null  float64
 10  O3          578667 non-null  float64
 11  Benzene     544229 non-null  float64
 12  Toluene     487268 non-null  float64
 13  Xylene      252046 non-null  float64
 14  AQI         578795 non-null  float64
 15  AQI_Bucket  578795 non-null  object
dtypes: float64(13), object(3)
memory usage: 86.4+ MB
```

cleaning

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 707875 entries, 0 to 707874
Data columns (total 22 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   City          707875 non-null  object
 1   PM2.5         707875 non-null  float64
 2   PM10          707875 non-null  float64
 3   NO            707875 non-null  float64
 4   NO2           707875 non-null  float64
 5   NOx           707875 non-null  float64
 6   NH3           707875 non-null  float64
 7   CO            707875 non-null  float64
 8   SO2           707875 non-null  float64
 9   O3            707875 non-null  float64
 10  Benzene       707875 non-null  float64
 11  Toluene       707875 non-null  float64
 12  Xylene        707875 non-null  float64
 13  AQI           707875 non-null  float64
 14  AQI_Bucket    707875 non-null  object
 15  year          707875 non-null  int64
 16  month         707875 non-null  int64
 17  time          707875 non-null  object
 18  week          707875 non-null  int64
 19  day           707875 non-null  int64
 20  city_lencode  707875 non-null  int64
 21  AQI_lencode   650420 non-null  float64
dtypes: float64(14), int64(5), object(3)
memory usage: 118.8+ MB
```

Theres **no duplicated data**

81,37% missing values and handled by **imputation of the median's data**

Source : https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india?rvi=1

DATASETS

Cleaned Datasets

707.875 rows
22 columns

1 Target
14 Features

2015 - 2020

Source : https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india?rvi=1

TARGET : Air Quality Index

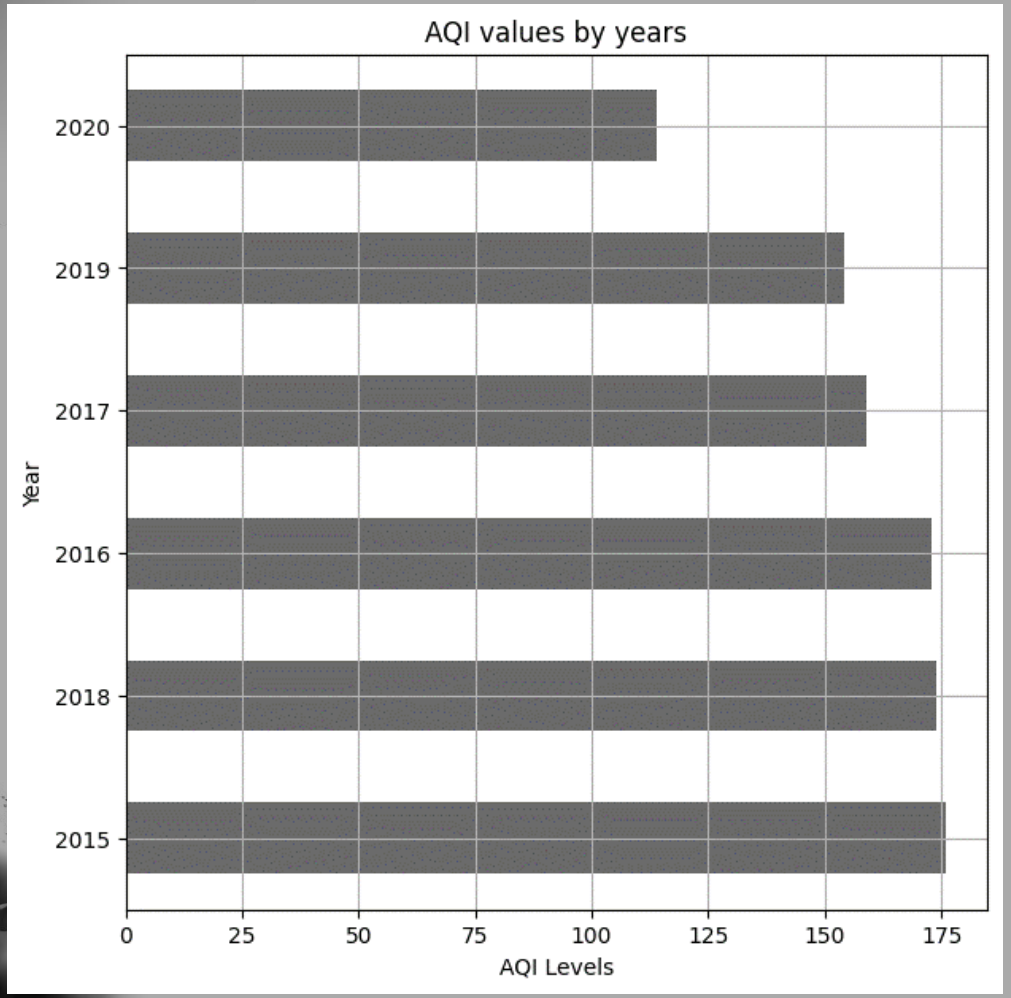
Elements Features

- PM 2.5
- PM 10
- NO
- NO2
- NOx
- NH3
- CO
- SO2
- O3
- Benzene
- Toluene
- Xylene

City Features

- ❑ Ahmedabad
- ❑ Aizawl
- ❑ Amaravati
- ❑ Amritsar
- ❑ Bengaluru
- ❑ Bhopal
- ❑ Bhajrajnagar
- ❑ Chandigarh
- ❑ Chennai
- ❑ Coimbatore
- ❑ Delhi
- ❑ Ernakulam
- ❑ Gurugram
- ❑ Guwahati
- ❑ Etc.

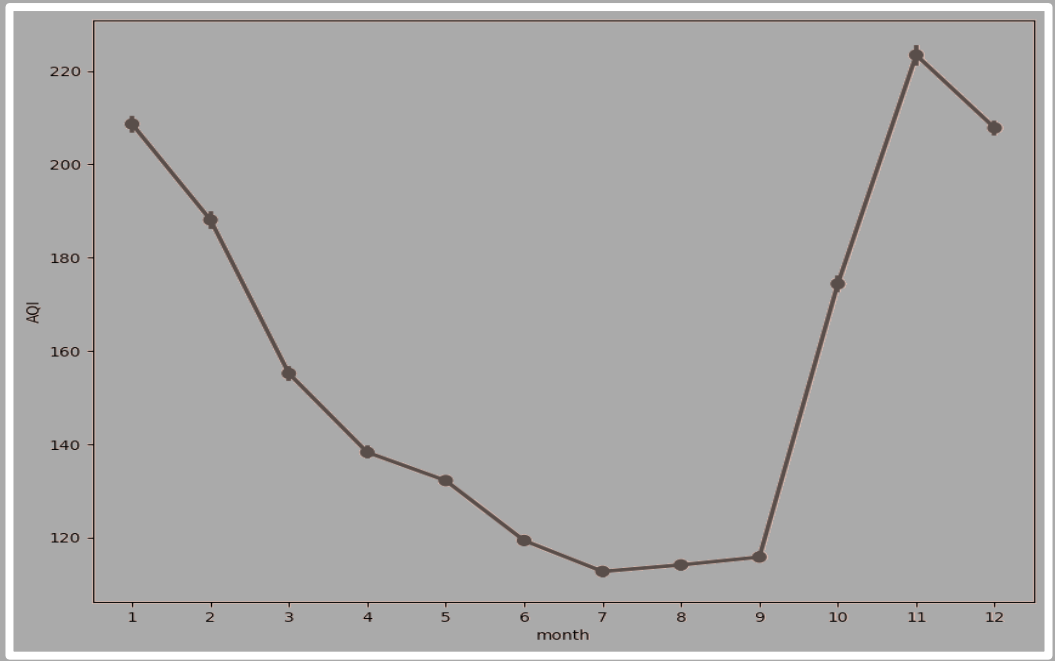
Exploratory Data Analysis



The higher the AQI value, the worse it is

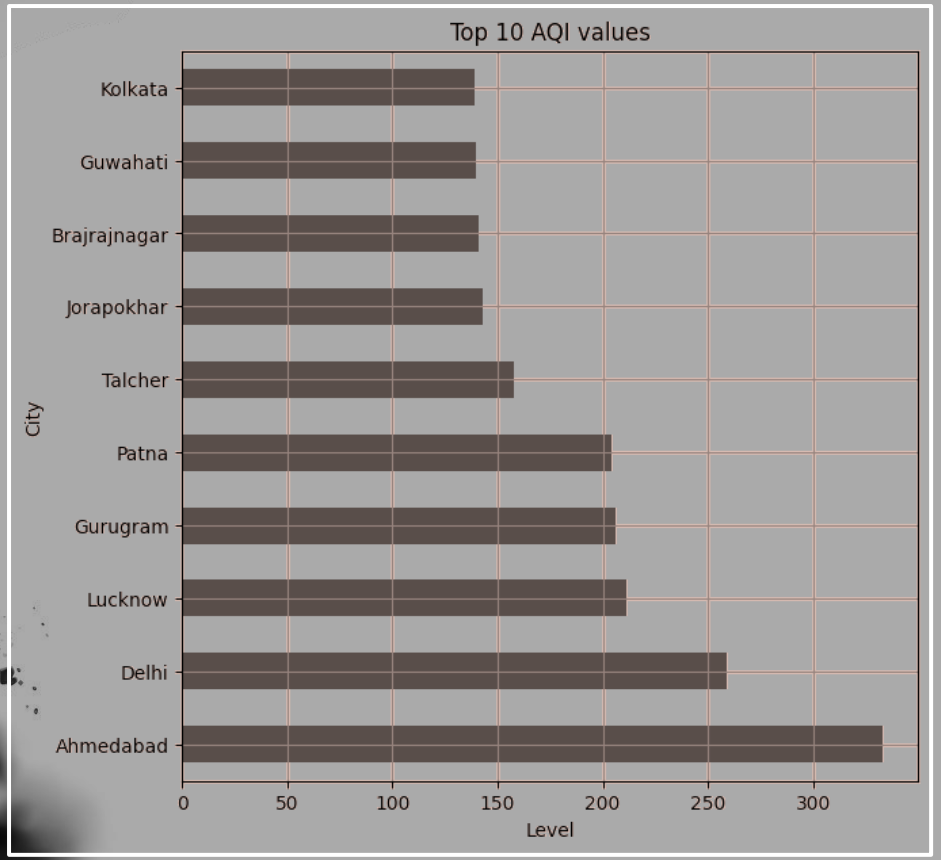
In the 5 years we can see the air quality in India is getting better. The overall average is at moderate level but with different values.

AQI by month wise



From October to February AQI values is genereally highest , might be because of Deewali and New Year fireworks

Which city is having high AQI values?



Ahmedabad having **high AQI value** which represent **worst air quality**. It can comes from vehicles and emissions from coal-fired power plants.

Source : https://www.iqair.com/india/gujarat/ahmedabad

What most affects the air quality in the city?

Ahmedabad

Max value
- NO = 498.58
- NO2 = 494.15
- NOx = 498.61
- CO = 498.57

Delhi

Max value
- NO = 497.40
- NO2 = 337.82
- NOx = 433.78
- CO = 47.42

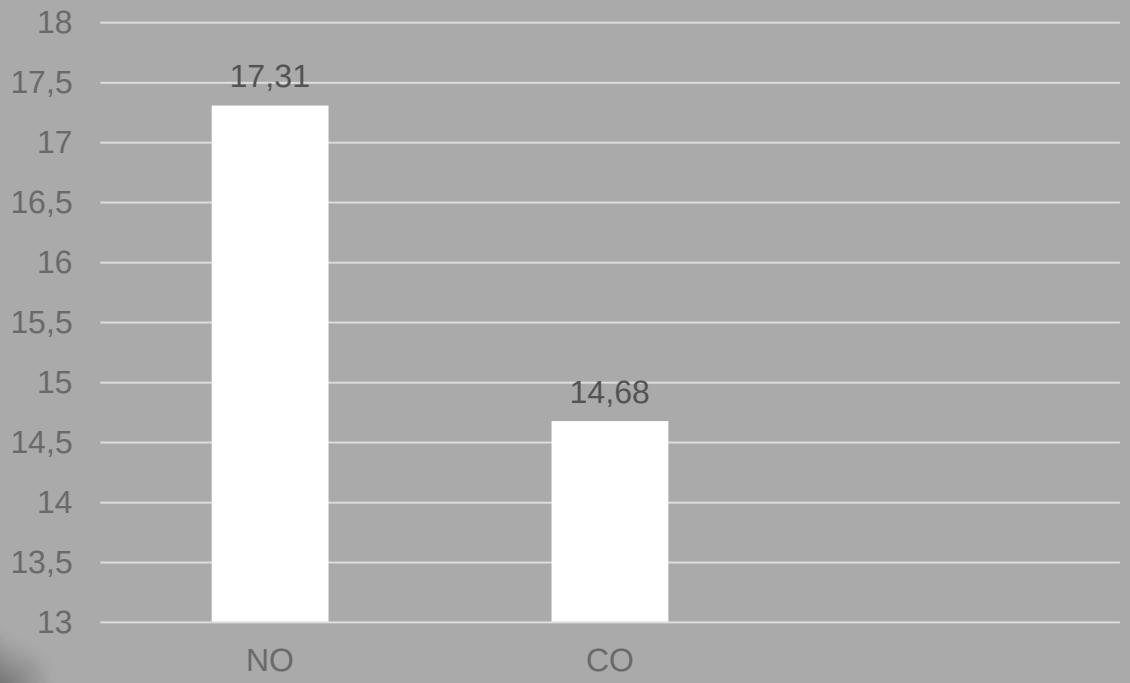
Lucknow

Max value
- NO = 489.27
- NO2 = 257.64
- NOx = 360.38
- CO = 50.00

Gurugram

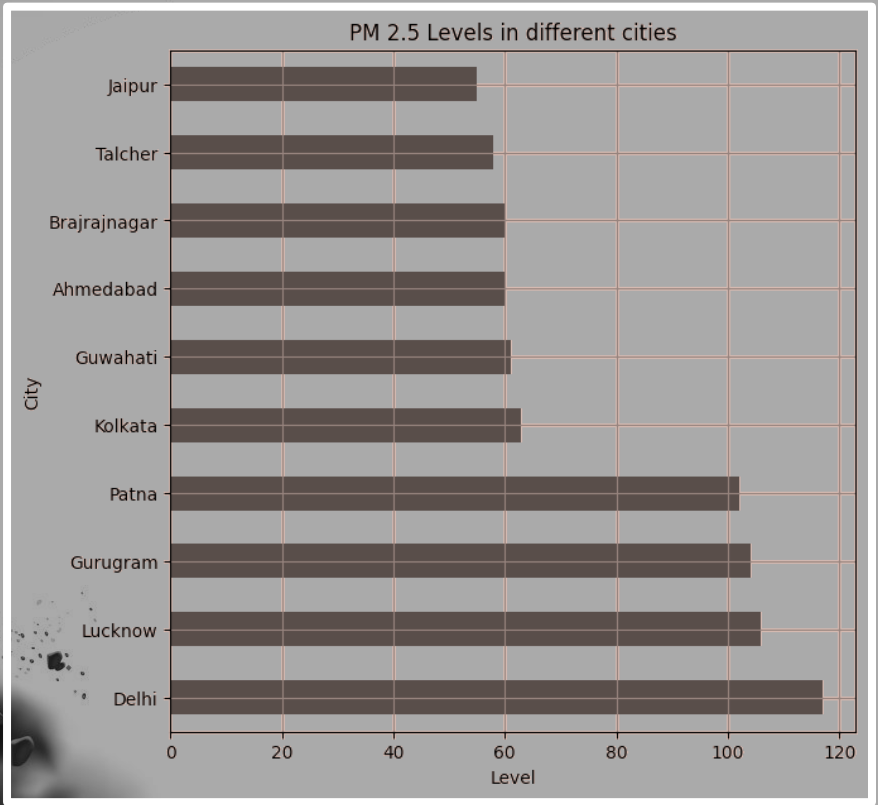
Max value
- NO = 499.99
- NO2 = 495.56
- NOx = 485.42
- CO = 49.27

The ratio of CO and NO in the air?



Based on the average of element's amount. We can see that **NO is higher** than CO.
NO is produced from vehicle fuel fumes.

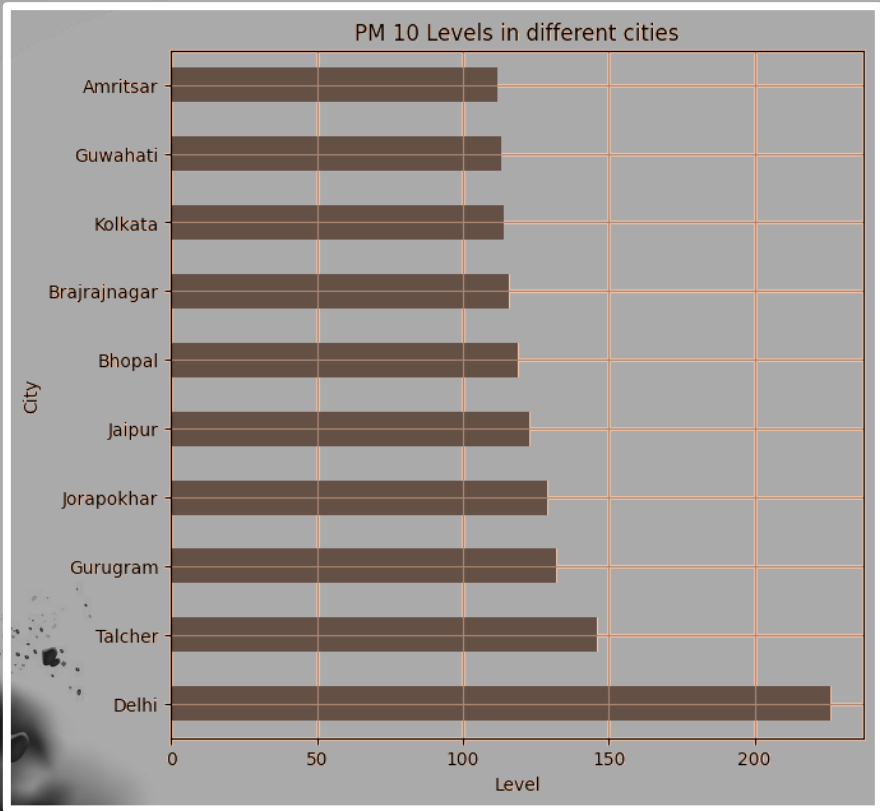
How about the particulate matter?



Delhi along few other cities having high average PM 2.5 levels. In 2019 it ranked in with a PM2.5 reading of 98.6 µg/m³

Source : https://www.iqair.com/india/delhi

How about the particulate matter?



There are an estimated **30.2 million people** registered living in Delhi as of 2020, More citizens mean **more vehicular emissions given off.**

Source : https://www.iqair.com/india/delhi

CLUSTERING USING K-MEANS

We are trying to figure it out if there's any cluster among the dataset based on AQI, and sample of elements

METHODOLOGY

1

Data Preprocessing

Cleaning data to handling missing values and duplicated values

2

Feature Standardization

Using StandardScaler

3

Elbow Method

To know how many clusters

4

K-Means Clustering

Assigning clusters to dataset

K – MEANS RESULTS

Clusters	Average CO	Average NO	AQI	AQI Level
0	0.93 mg/m3	8.9 ug/m3	247.71	3
1	0.60 mg/m3	5.15 ug/m3	67.57	5
2	1.17 mg/m3	15.16 ug/m3	106.55	4
3	0.59 mg/m3	6.69 ug/m3	126.21	4

1 = Severe
2 = Very poor
3 = Poor
4 = Moderate
5 = Satisfactory
6 = Good

ML MODELING

We want to predict AQI based on elements and city

Modeling

Logistic Regression

Decision Tree

Random Forest

PRE-PROCESSING STEPS

Aspects	Action
Prepare Dataset	Features: elements and city
Categorical Features	Encode categorical features into numeric using label encoding
Split Dataset	80% Training Data, 20% Test Data

ELEMENTS

No	Model	F1 SCORE				
		1	3	4	5	6
1.	Logistic Regression	40.29	0.4	71.91	28.73	0.2
2.	Decision Tree	59.62	43.19	77.45	67.95	58.63
3.	Random Forest	71.32	51.17	84.64	78.11	68.92

1 = Severe
2 = Very poor
3 = Poor
4 = Moderate
5 = Satisfactory
6 = Good

- Because the datasets are imbalance, then we can't use accuracy for predicting. we still another metric evaluation such as F1 score for predicting
- There’s no data contains level 2, so theres no predictions

ELEMENTS + CITY

No	Model	F1 SCORE				
		1	3	4	5	6
1.	Logistic Regression	47.41	0.0	73.73	39.69	0.04
2.	Decision Tree	60.31	44.80	78.20	69.07	60.00
3.	Random Forest	71.61	53.49	85.12	78.99	69.91

1 = Severe
2 = Very poor
3 = Poor
4 = Moderate
5 = Satisfactory
6 = Good

- Because the datasets are imbalance, then we can't use accuracy for predicting. we still another metric evaluation such as F1 score for predicting
- There’s no data contains level 2, so theres no predictions

CONCLUSION

- In the 5 years, the overall average is at moderate level but with different values.
- From October to February AQI values is genereally highest , might be because of Deewali and New Year fireworks
- Ahmedabad having high AQI value which represent worst air quality.
- Delhi is subject to a high level of pollution year-round. The levels of fine and coarse particulate matter, known respectively as PM2.5 and PM10
- Elements and location city are contribution to amount of Air Quality Index
- We can see the average of NO is higher than CO. NO is produced from vehicle fuel fumes, it means the air quality can be greatly influenced by vehicle pollution

HOW TO MINIMIZE AIR – POLLUTAN IN INDIA ?

- The government can allocate a certain amount of funds to develop environmentally friendly fuels
- Use Machine Learning Model to predict expected outcomes of Air Quality Index
- Focusing efforts to control air pollution from "The Big 4" by control population in those cities.

Thanks!

Do you have any questions?

Amandaartyan.aa@gmail.com
linkedin.com/in/amandaartyan/