Project Report (CS685A)

Group - 2

Group Details:

Name	Roll No.	Email
Nikhil Agarwal	180475	nikhilag@iitk.ac.in
Supreeth Baliga	180801	supbal@iitk.ac.in
Sarthak Kapoor	180675	sartkap@iitk.ac.in
Jayesh Shaw	180330	jayeshaw@iitk.ac.in
Chinmay Goyal	180206	chinmayg@iitk.ac.in

Problem Statement:

We aim to gain insights by analysing the Air Quality Index (AQI) derived from the various particulate concentrations present in the air. The data we are using is extracted from the public repository of the Central Pollution Control Board (CPCB) of India. This topic caught our eye because a few months back, the media was abuzz about the sudden drop in the quality of air due to the pandemic. Apart from that, it is crucial to monitor the AQI so as to strategize on how it can be controlled. Policy-makers can take critical decisions on where and when to deploy resources based on the insights gained. Thus, since analysing the AQI plays a role at the higher levels and given that it does have a direct impact on human health, we wanted to use our skills to derive various trends and relate them with various corresponding causes.

Data Extraction:

A public user interface provided by the CPCB of India (link here) allows us to download data corresponding to a lot of particulates suspended in the air. Using these particulate concentrations, we have calculated the AQI for each entry using the formula given in this link. The script which populates the data with corresponding AQI values is **scripts/data_populate.ipynb**. We have done this task first and stored the data before moving on to gathering insights.

We have chosen the following 25 cities for analysis: Ahmedabad, Amaravati, Amritsar, Bengaluru, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam,

Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, Visakhapatnam. These cities are well spread over India and have a relatively lower density of missing values as compared to others.

Also note, that we have used the data starting from 1st January 2015 and ending on 1st July 2021 because, before that time period, there are a lot of missing values making the data rather sparse which would lead to errors.

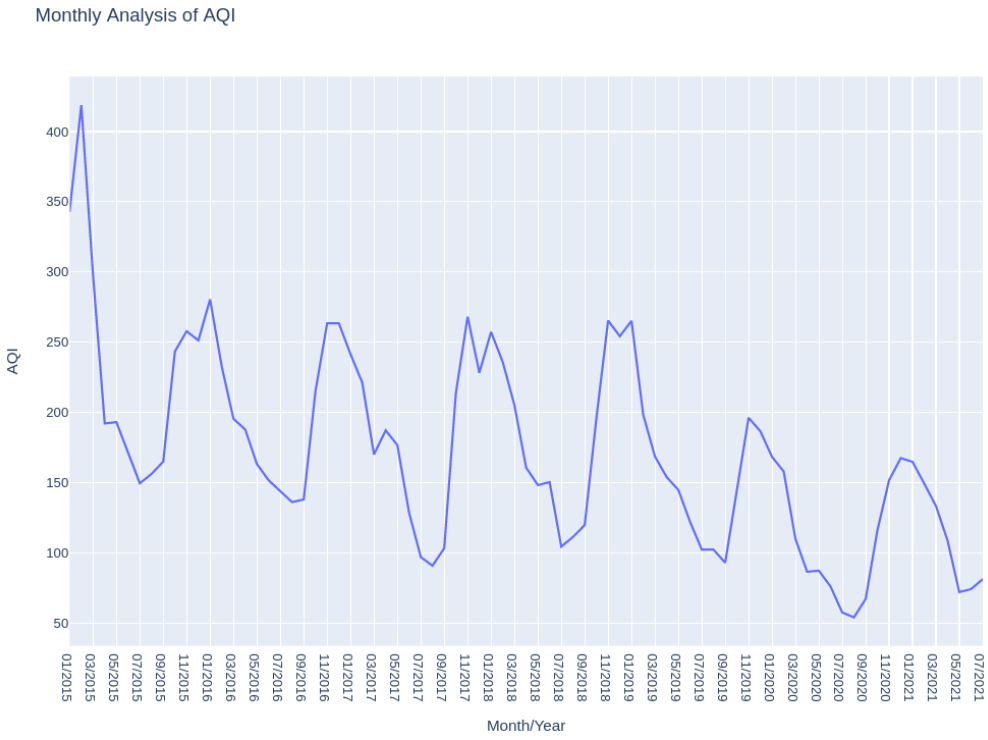
For each city, we have taken averages of the concentrations recorded by one or more stations in that city.

Insights:

Does AQI show a seasonality trend?

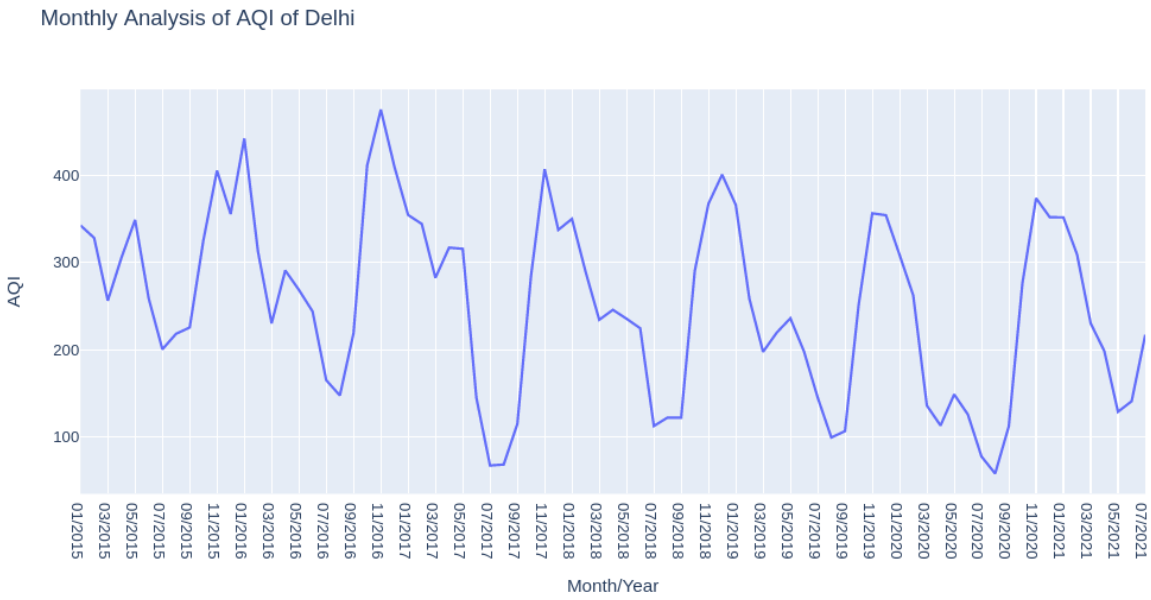
The first thing that comes to mind is whether AQI shows any periodic trend each year. Do the values show any abrupt changes in any season and if any reasons could be found for the same.

After getting rid of null values, we grouped the entries by (month, year) and took the average values. Plotting them on a graph gave us the following figure:



Results and their analysis:

It seems that the quality of air is rather pure during the summer months while we can see the poor quality of air during the winter months. The large peak at the start of 2015 is due to missing data because of which there is bias observed in that temporal region. The lower peaks in 2020 and 2021 are due to the pandemic (more details on this in further sections). For the rest of the years, the values are somewhat periodic. So what exactly causes this AQI to show such a seasonal trend? The answer to this is majorly the density of air. Cold air is denser than hot air. Higher density implies that the pollutant particles remain suspended in place for a longer time. Hence, during summer, the air pollutants are easily whisked away by the wind. Whereas during winter, these particles move rather slowly. There are other factors like ground-level ozone, temperature, humidity, pressure etc. which also play a role in such a seasonal trend. Although the values may be different for different regions, the graph should be more or less of a similar shape. We even tried to see if individual cities follow the same seasonal trend in AQI values too. Most of the cities indeed showed the same trend (picture for Delhi below). However, some of the cities showed completely lopsided graphs - the bigger reason for this was the missing values present in the dataset rather than the demographic of the city. Hence, no such comment can be made about individual cities, just the aggregate can be analysed. Note that the festival of Diwali which occurs at the brink of winter is also a major factor in the increase in air pollution levels.



Implication: Policy-makers can strategize on being ready to tackle the higher air pollution during the winter, whereas they can use the summer months to pull back the resources and re-organise them for the next phase. They can also place policies on various industries setting their output pollution levels according to the season they are operating in.

Which days show maximum and minimum AQI in different cities?

We found the dates for which various cities showed maximum AQI and tried to find reasons for the same. In the below tables, for cities with no significant data that year, we are marking the required dates as N/A.
Following are the dates on which we saw maximum AQI values for each city every year:

City	2015	2016	2017	2018	2019	2020	2021
Ahmedabad	23-02-2015	14-07-2016	14-11-2017	19-02-2018	03-01-2019	19-02-2020	20-03-2021
Amaravati	N/A	N/A	08-12-2017	10-12-2018	04-01-2019	27-12-2020	01-01-2021
Amritsar	N/A	N/A	11-05-2017	15-06-2018	28-10-2019	29-11-2020	31-01-2021
Bengaluru	06-07-2015	13-12-2016	06-02-2017	01-05-2018	07-11-2019	02-11-2020	31-03-2021
Bhopal	N/A	N/A	N/A	N/A	28-11-2019	11-11-2020	04-01-2021
Brajrajnagar	N/A	N/A	13-12-2017	12-03-2018	14-01-2019	07-12-2020	03-03-2021
Chandigarh	N/A	N/A	N/A	N/A	28-10-2019	21-12-2020	01-01-2021
Chennai	28-03-2015	21-09-2016	19-10-2017	13-01-2018	27-10-2019	14-01-2020	28-01-2021
Coimbatore	N/A	N/A	N/A	N/A	15-10-2019	25-12-2020	30-01-2021
Delhi	07-11-2015	07-11-2016	09-11-2017	15-06-2018	03-11-2019	09-11-2020	15-01-2021
Gurugram	N/A	03-11-2016	30-08-2017	14-06-2018	03-11-2019	09-11-2020	01-01-2021
Guwahati	N/A	N/A	N/A	N/A	13-08-2019	15-01-2020	05-03-2021
Hyderabad	02-06-2015	16-06-2016	13-01-2017	04-01-2018	07-12-2019	26-12-2020	01-01-2021
Jaipur	N/A	N/A	02-12-2017	21-04-2018	08-04-2019	10-11-2020	30-03-2021
Jorapokhar	N/A	N/A	27-04-2017	11-12-2018	05-01-2019	21-02-2020	17-01-2021
Kolkata	N/A	N/A	N/A	08-11-2018	23-01-2019	29-12-2020	03-01-2021
Lucknow	22-03-2015	08-11-2016	14-11-2017	08-11-2018	02-01-2019	08-07-2020	03-01-2021
Mumbai	N/A	N/A	N/A	10-11-2018	25-12-2019	02-01-2020	06-01-2021
Patna	31-12-2015	01-01-2016	02-12-2017	21-12-2018	02-01-2019	27-12-2020	12-03-2021
Shillong	N/A	N/A	N/A	N/A	05-09-2019	19-02-2020	28-03-2021
Talcher	N/A	N/A	N/A	27-11-2018	14-02-2019	27-01-2020	19-01-2021
Thiruvananthapuram	N/A	N/A	26-11-2017	12-12-2018	01-01-2019	10-02-2020	29-01-2021
Visakhapatnam	N/A	12-09-2016	26-12-2017	08-11-2018	14-01-2019	24-10-2020	13-01-2021

Following are the dates on which we saw minimum AQI values for each city every year:

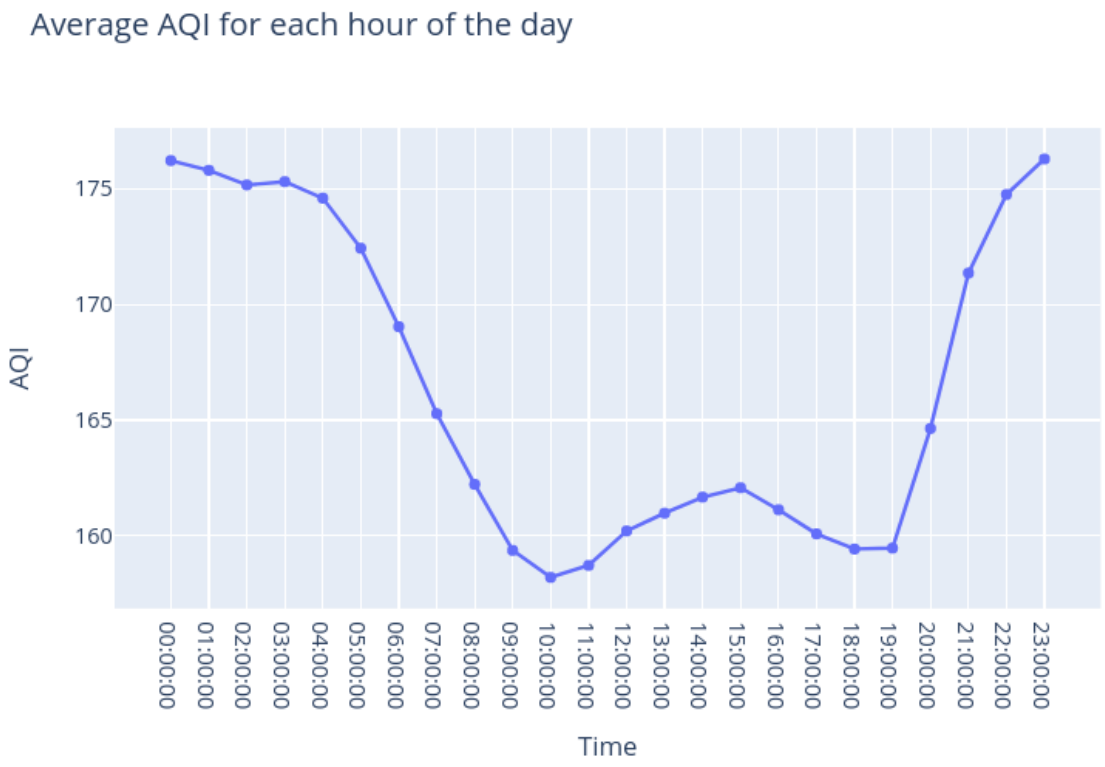
City	2015	2016	2017	2018	2019	2020	2021
Ahmedabad	24-07-2015	15-07-2016	23-10-2017	02-03-2018	23-12-2019	06-07-2020	17-05-2021
Amaravati	N/A	N/A	14-12-2017	16-07-2018	03-08-2019	04-09-2020	05-05-2021
Amritsar	N/A	N/A	29-05-2017	25-09-2018	18-08-2019	25-07-2020	04-02-2021
Bengaluru	18-09-2015	08-06-2016	02-12-2017	17-08-2018	20-07-2019	16-07-2020	30-05-2021
Bhopal	N/A	N/A	N/A	N/A	29-09-2019	23-08-2020	17-05-2021
Brajrajnagar	N/A	N/A	18-12-2017	16-08-2018	28-07-2019	21-08-2020	27-06-2021
Chandigarh	N/A	N/A	N/A	N/A	14-12-2019	08-07-2020	23-04-2021
Chennai	17-09-2015	24-09-2016	14-03-2017	16-08-2018	25-08-2019	03-12-2020	06-05-2021
Coimbatore	N/A	N/A	N/A	N/A	02-12-2019	16-07-2020	02-02-2021
Delhi	23-09-2015	22-08-2016	31-07-2017	24-09-2018	18-08-2019	31-08-2020	20-06-2021
Gurugram	N/A	02-08-2016	02-09-2017	28-07-2018	11-08-2019	31-08-2020	20-06-2021
Guwahati	N/A	N/A	N/A	N/A	03-07-2019	15-07-2020	11-05-2021
Hyderabad	25-07-2015	06-08-2016	17-09-2017	18-07-2018	05-09-2019	05-08-2020	15-06-2021
Jaipur	N/A	N/A	30-06-2017	23-09-2018	10-08-2019	31-08-2020	19-05-2021
Jorapokhar	N/A	N/A	24-07-2017	05-07-2018	10-09-2019	25-08-2020	14-02-2021
Kolkata	N/A	N/A	N/A	16-07-2018	24-09-2019	04-07-2020	17-05-2021
Lucknow	12-07-2015	31-07-2016	29-08-2017	01-08-2018	29-09-2019	07-07-2020	19-06-2021
Mumbai	N/A	N/A	N/A	31-07-2018	28-09-2019	06-08-2020	16-06-2021
Patna	30-10-2015	08-08-2016	03-10-2017	03-09-2018	29-09-2019	22-09-2020	28-05-2021
Shillong	N/A	N/A	N/A	N/A	23-11-2019	05-07-2020	11-05-2021
Talcher	N/A	N/A	N/A	13-10-2018	19-09-2019	23-06-2020	03-04-2021
Thiruvananthapuram	N/A	N/A	01-10-2017	15-07-2018	08-07-2019	17-07-2020	20-05-2021
Visakhapatnam	N/A	21-09-2016	19-11-2017	27-05-2018	23-09-2019	12-10-2020	16-05-2021

Results and analysis: It is quite clear from the above values that different cities have different days on which they have the worst quality of air (similarly for the best quality). Although we could not show it visually, looking at the values manually, you can see that cities closer to each other show closer maximum and minimum AQI dates compared to ones that are far apart. These dates are influenced by a number of factors including weather, population, the density of industries, the intensities with which various festivals are celebrated etc. Given that the dates are rather random and do not show a uniform trend, we could not find a common reason which could justify these dates. Giving some explanation for them would require much more extensive

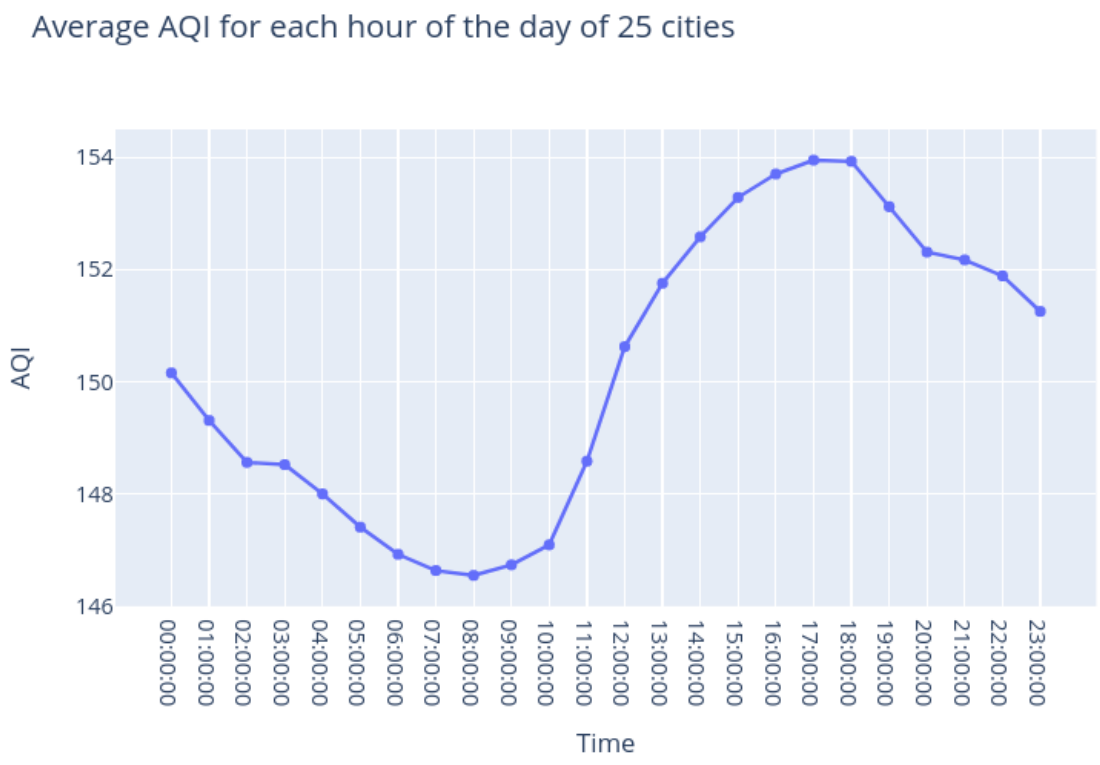
research for each of the cities and a lot more detailed data for not just the AQI but other controlling factors too.

Implications: Policy-makers can do further investigation regarding the factors that determine these dates and find corresponding solutions from the findings.

Is AQI high during the daytime or at night?



Results and their analysis:
As it can be seen that the air quality is worse in the night as compared to the daytime taking all the cities into consideration. However, the graphs for all cities should also be more or less similar to this in shape with their values being different. But we can observe that the graph for all cities other than Ahmedabad is not similar to the one observed. It is because Ahmedabad has very high values of average AQI values which is almost 4-5 times the average. So we need to perform an analysis again taking the other 25 cities.



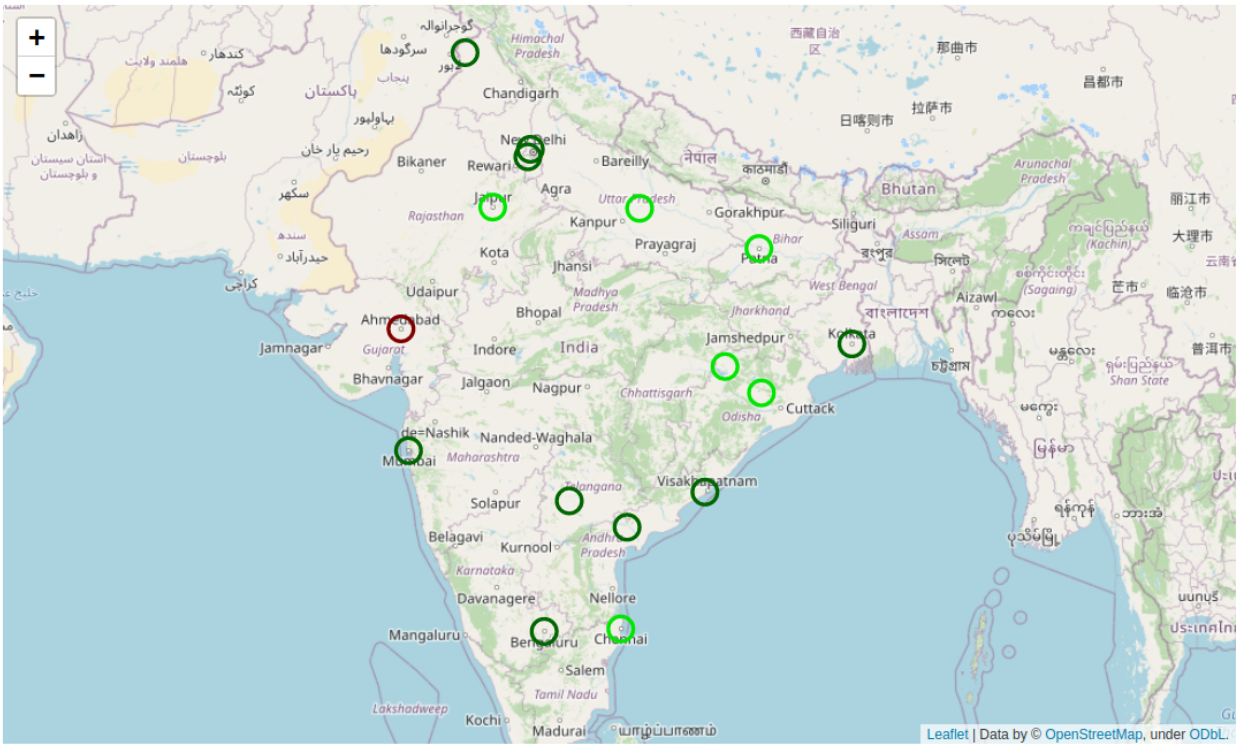
It can be observed that the Air Quality Index starts to rise from 09:00 am in a day since vehicles start to increase from that time and industries also usually start from that time. The air pollution rises continuously till 06:00 pm which marks the end of industrial pollution for the day as industrial pollution stops now for the day. However, until 10:00 pm it is gradually decreasing since the pollution due to vehicles is controlled after 10:00 pm when there are fewer vehicles on the road. During the night-time hours, the atmosphere traps car emissions, CO2, and other pollutants down near the ground. At night, the lack of cloud cover means the ground loses heat rapidly and the air in contact with the ground becomes colder. The warmer air rises and acts as a lid trapping the colder air close to the ground. Pollution, including that from road traffic, is also trapped, so the air layer closest to the ground becomes more and more polluted. This is the reason that during the night-time, although there are very few human activities involved, the pollution still gradually decreases and we cannot observe a sharp decrease in AQI. Now we can also see that most of the cities observe this trend only and some of the cities like Aizawl, Guwahati, Kochi, and Shillong which does not follow this trend is because we do not have much data for these cities and these cities are not much affected by human activities.

Implication: Policy-makers can strategize on controlling the number of vehicles during the daytime. Policies can be made for a few industries to operate at night rather than during the day to reduce pollution through daylight hours. This will help balance out the air pollution.

Can we get a spatial and temporal comparison of AQI among the cities?

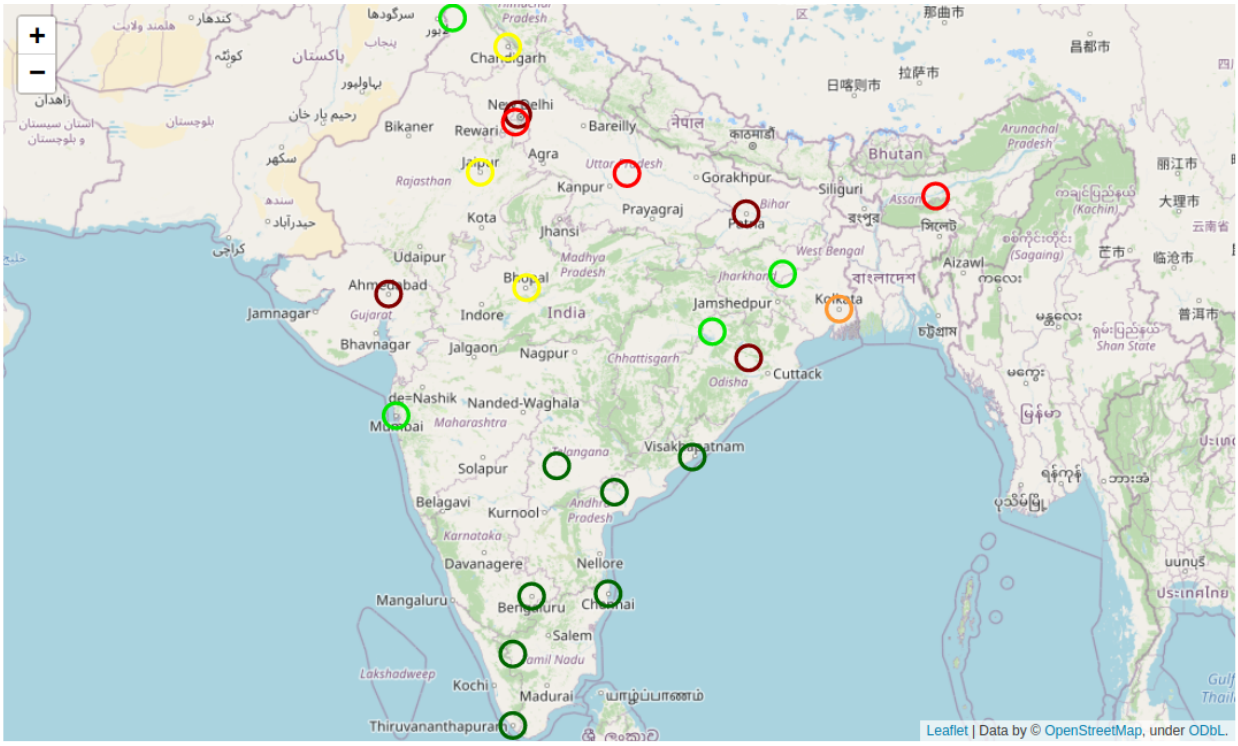
The next thing one would be intrigued about is to visualise how the AQI might have changed over time and if this effect is uniform throughout or if it can be seen more in certain parts of India but on a common scale. To answer this, we have compared two dates: 1st of July, 2018 and 31st of Dec, 2019. Note that we did not dive into 2019 in order to avoid the error due to the pandemic effect.

So, let us visualise the graph on 1st July 2018.



As can be seen, the only area with a high level of AQI is in Ahmedabad. This is because the AQI calculation has errors because of the fact that PM10 concentration for Ahmedabad was measured starting from 2019. Rest all cities have a relatively better quality of air.

Now, let us visualise the same for 31st Dec 2019.



You can see the drastic difference in levels of AQI just in a span of 1.5 years. Ahmedabad still has a very poor quality of air (this time not due to data errors) but now it is joined by cities like New Delhi, Gurugram, Lucknow, Patna and Talcher.

Results and their analysis:

1. The AQI quality has become poor in the northern parts of the country as compared to the southern side. A general reason for it is the fact that oceans act as natural air cleaners because of which the southern side, which is closer to the coast, has a better quality of air. This is also influenced by the fact that industries are present in greater density in the north as compared to southern India. Ahmedabad, Patna etc. are hubs of industrial activity.
2. You would usually expect that metropolitan cities have a greater probability of having poor quality of air. But as it turns out, the industries play a much more dominant role in deciding the quality of air. This is why Ahmedabad shows one of the worst qualities of air of all time even though it is not as populated as the metropolitan cities.
3. Overall, as expected, the quality of air has worsened over the years. Even if they remain in the same AQI bracket, their values have shown a drastic jump.

To conclude, in general, the value of AQI increases on the temporal scale with an increase in time. On the geographical scale, various factors like the presence of polluting industries, vicinity to the ocean etc. drive the AQIs in the respective locations.

Implications: Policy makers can take the factors of vicinity to the ocean into consideration. They can also mandate the pollution levels in areas with a high density of industries. Industries can be distributed among various locations to keep the density of industries low. This will help in balancing out the quality of air.

Correlation among particle concentrations and with AQI

The complete correlation table between all the particles as well as with AQI is provided below -

	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	AQI
PM2.5	1.000000	0.871494	0.469176	0.406199	0.475995	0.316530	0.108662	0.147238	0.181193	0.027294	0.140691	0.702553
PM10	0.871494	1.000000	0.519079	0.523979	0.556444	0.394602	0.132690	0.253499	0.258167	0.026689	0.197992	0.844010
NO	0.469176	0.519079	1.000000	0.500006	0.819155	0.220422	0.209814	0.183167	0.005071	0.041333	0.168642	0.462666
NO2	0.406199	0.523979	0.500006	1.000000	0.664503	0.299580	0.347848	0.399479	0.296289	0.027447	0.261889	0.548254
NOx	0.475995	0.556444	0.819155	0.664503	1.000000	0.228318	0.238680	0.255367	0.097633	0.043805	0.208113	0.503932
NH3	0.316530	0.394602	0.220422	0.299580	0.228318	1.000000	0.121293	-0.024005	0.109037	-0.002610	0.034683	0.294658
CO	0.108662	0.132690	0.209814	0.347848	0.238680	0.121293	1.000000	0.495671	0.051548	0.043718	0.237389	0.654849
SO2	0.147238	0.253499	0.183167	0.399479	0.255367	-0.024005	0.495671	1.000000	0.182535	0.031671	0.271102	0.465364
O3	0.181193	0.258167	0.005071	0.296289	0.097633	0.109037	0.051548	0.182535	1.000000	0.013854	0.108078	0.217084
Benzene	0.027294	0.026689	0.041333	0.027447	0.043805	-0.002610	0.043718	0.031671	0.013854	1.000000	0.774409	0.043167
Toluene	0.140691	0.197992	0.168642	0.261889	0.208113	0.034683	0.237389	0.271102	0.108078	0.774409	1.000000	0.257159
AQI	0.702553	0.844010	0.462666	0.548254	0.503932	0.294658	0.654849	0.465364	0.217084	0.043167	0.257159	1.000000

Analysis -

- We observe that almost all pairs of columns have a positive correlation.
- NOx has a very high correlation with NO and NO2, something which is expected since NOx is a mixture of these two gases.
- PM2.5 and PM10 have a high correlation. PM2. 5 is the subset of PM10 particles
- that have aerodynamic diameters less than or equal to 2.5 μm, thus such a result is very plausible.
- Benzene and toluene also have a high correlation. This is due to the fact that both these chemicals are very similar physically and chemically and thus are expected to exist under similar conditions.
- Apart from these, there is the absence of any pair of pollutants with a particularly high correlation.
- PM2.5, PM10, CO, SO_2, NO_2 have the most significant correlation with AQI while benzene, toluene, ozone and NH_3 have very little correlation with AQI.

Due to the above reasons, in the following part which is the comparison of pre and post lockdown air quality, we use only the particles which have a significant correlation with AQI.

For each city which particle has the highest concentration?

One of the next things which can be observed is that if the AQI is affected by the same particles in each city or for each city it is different due to some reasons.

Methodology: To check this, we first compare the concentration of each particle to find which is the highest.

PM10	24070
PM2.5	5277
O3	2348
NOx	1985
NH3	1309
NO2	691
CO	582
SO2	364
Benzene	241
NO	223
Toluene	108

Number of (City, Date) pairs for which a particle has the highest concentration

We notice that for most of the cities and dates, the PM10 concentration is the highest as a low concentration of PM2.5 is more dangerous than the higher concentration of PM10. As per the WHO, 15 μg/m3 PM2.5 concentration is equally hazardous as 45 μg/m3 PM10 concentration which is equivalent to 25 μg/m3 NO2 concentration. Hence, there must be some other way to compare concentrations of different particles with each other.

The AQI calculator comes in rescue for comparing these concentrations as it involves calculating subindices for different particles which are compared directly. The result obtained after comparison of subindices is as follows :

Cities and the proportion of pollution in each

Results and Analysis: As we can notice that Ahmedabad has the highest concentration of CO which is because of the high rate of increasing vehicles and a majority of power plants located there. These two only combine to around 60% of CO emission in Ahmedabad. However, as we can notice that for most of the cities, the highest concentration is of either PM2.5 or PM10 since PM may be either directly emitted from sources (primary particles) or formed in the atmosphere through chemical reactions of gases (secondary particles) such as sulfur dioxide (SO_2), nitrogen oxides (NO_X), and certain organic compounds. These organic compounds can be emitted by both natural sources, such as trees and vegetation, as well as from man-made (anthropogenic) sources, such as industrial processes and motor vehicle exhaust. However, the other gases which are a major part of air pollution are directly emitted from only some of the man-made processes but PM is emitted from every natural and man-made process. We can also observe that cities like Shillong, Aizawl, Ernakulam and Coimbatore have low pollution levels which can be explained as these are cities where no major industries are there and vehicle count is also not much as compared to metropolitan cities.

How has the lockdown during the pandemic impacted the air quality in India?

In March 2020, the coronavirus pandemic broke out in India, and by the end of March pretty much all of India went into lockdown. As a result, people could not travel as much as they usually did and neither did all factories around the country act at a similar level as they did, and a lot of people lost their jobs too. Since transportation is a huge factor influencing air pollution it is only natural that we investigate the impact COVID has had on the air quality of India.

Methodology: We have selected a few cities in our sample, Delhi, Mumbai, Kolkata and Bangalore. From another analysis of ours, we have found that PM10, PM25, SO2 and NO2 have a quite high correlation with the air quality index. For each of these particles, we attempt to analyse the trend of difference in concentration for days between 1st March and 31st May for the years 2019 and 2020.

Note: The plots have been presented in two pages in landscape mode, please refer to them if needed.

Analysis of the Data:

1. During the initial phase of the date range considered the difference in concentration for all the four particles considered does not differ by much. This is due to the fact that lockdown came into effect only at the end of March 2020.

2. For PM10 and PM2.5, across all the cities we can observe that the concentration was clearly higher in 2019 than in 2020 except for a few days. This is because the primary source of PM10 and PM2.5 in India is traffic. Traffic contributes to 34% of PM10 particles in the air in India and 37% of PM2.5 particles [5]. During lockdown most vehicles in India, both private and public were not operating and thus this trend is not surprising.

3. Sulphur Dioxide(SO2) emissions show an interesting trend among the particles under consideration. The primary source of SO2 emissions in India is coal plants generating electricity. In the lockdown, people were staying indoors at home and rarely had any other mode of entertainment than the internet, their mobile phones and other electronic media. Usage of this equipment started increasing. The decrease in consumption of electricity caused by the closure of offices was compensated by the increase in usage of electronic media at homes. Hence, we expect to observe a similar or increase in emission of SO2, which is the case in Kolkata and Bangalore. In Delhi, a few thermal plants were shut down temporarily during the lockdown resulting in slightly less emission of SO2 and hence the trend.

4. Similar to PM10 and PM2.5, the major source of NO2 emission in India is traffic and industrial plants. And since traffic fell down massively during the lockdown, levels of NO2

are expected to fall off. According to a study [6], NO_2 emission in urban Delhi reduced by 60% during Phase I of the lockdown (March 25 - April 13, 2020).

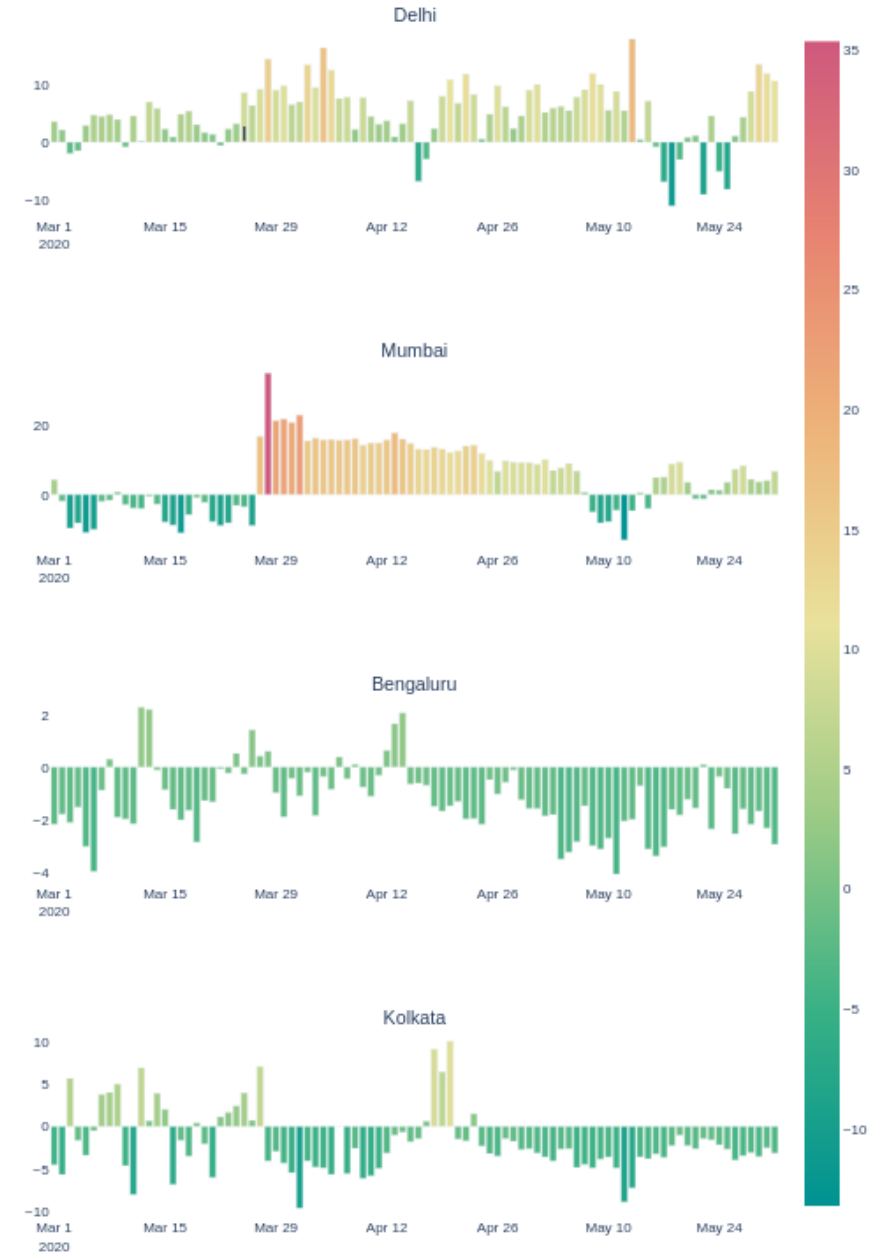
5. Here we point out that to answer the question “How long did it take for air quality to go back to its earlier values?”, we were unable to reach a satisfactory answer since there is no baseline that can be defined as “normal value”. Due to seasonality in AQI levels, we can only use the past year data of the same month to compare, but even this is inaccurate since due to prolonged lockdown in between, the conditions have not remained the same between a month in 2019 and in 2020. Moreover, it is generally expected that the pollution level will grow each year and not remain exactly the same. Thus, such an analysis would make more sense when a considerable amount of time has passed since normalcy.

6. We have carried out the analysis for both the first wave (end March 2020 - May 2020) as well as the second wave (April 2021 - June 2021). While the above arguments have been made for the data regarding the first wave, we obtained similar results for the second wave analysis as well. Thus, the same reasonings seem to work consistently for both the waves. The graphs of both the waves have been provided below.

Implications: Clearly, the lockdown has had a huge impact on the air quality of India. Due to the shutdown of Industrial plants and reduction in traffic, the concentration of most pollutant matters has shown a decline from 2019. It goes to show us that we can achieve a better and healthier environment if the usage of vehicles is somehow regulated and monitored. Also, for a reduction in the emission of SO_2, it is important that we shift from a thermal-based electricity source to a renewable source of energy.

The first four graphs given in the following 2 pages correspond to the first wave, while the 4 subsequent graphs correspond to the second wave. (Timeline is present in the x-axis in each graph)

SO2


Delhi
Mumbai
Bengaluru
Kolkata
Mar 1
2020
Mar 15
Mar 29
Apr 12
Apr 26
May 10
May 24

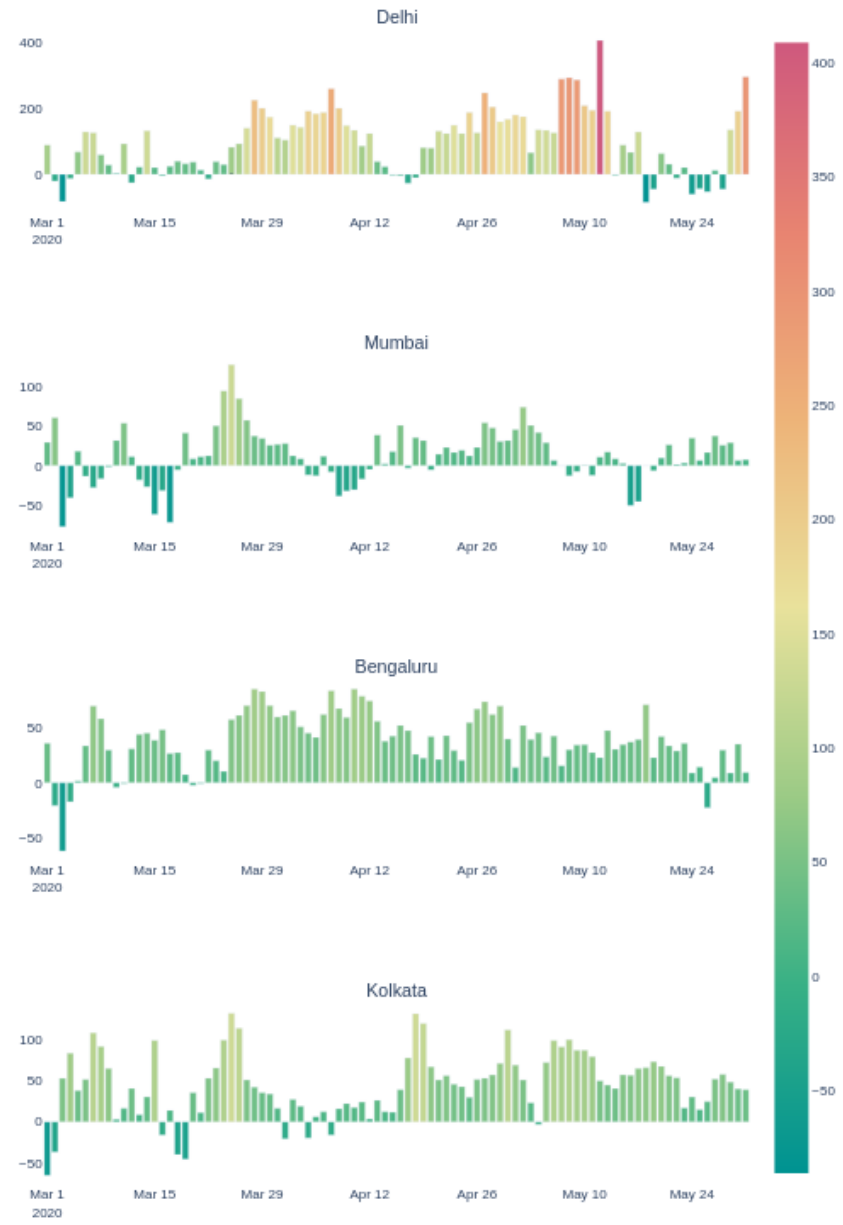
NO2


Delhi
Mumbai
Bengaluru
Kolkata
Mar 1
2020
Mar 15
Mar 29
Apr 12
Apr 26
May 10
May 24

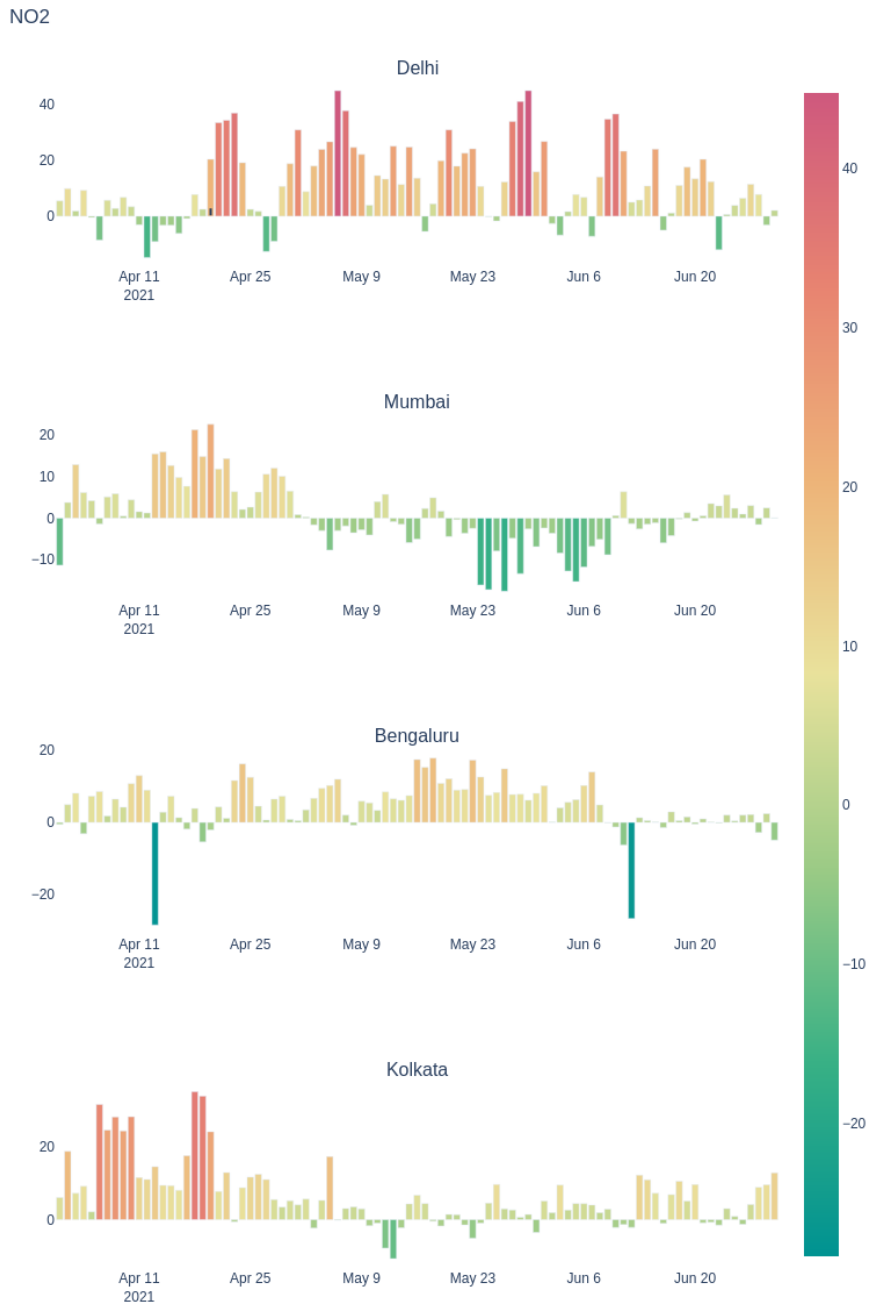
PM25


Delhi
Mumbai
Bengaluru
Kolkata
Mar 1
2020
Mar 15
Mar 29
Apr 12
Apr 26
May 10
May 24

PM10


Delhi
Mumbai
Bengaluru
Kolkata
Mar 1
2020
Mar 15
Mar 29
Apr 12
Apr 26
May 10
May 24

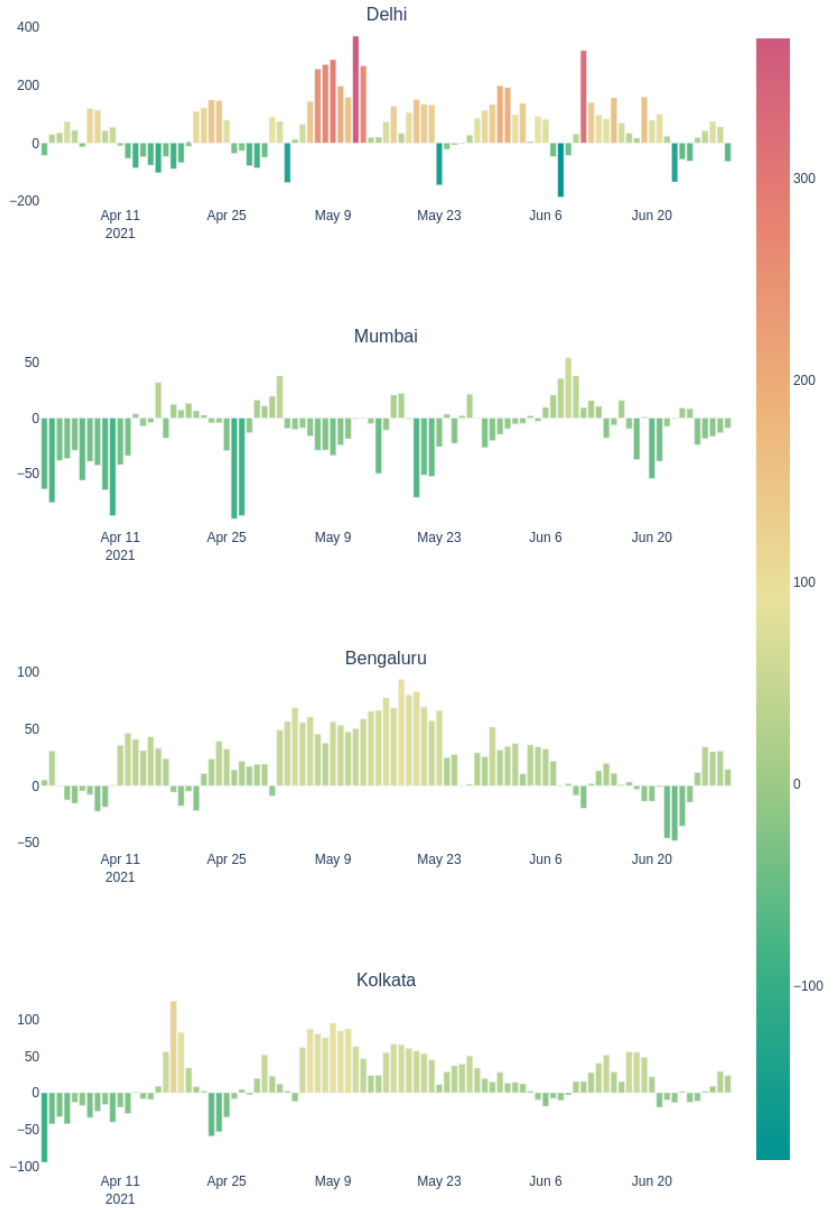

SO2
Delhi
Mumbai
Bengaluru
Kolkata
Apr 11
2021
Apr 25
May 9
May 23
Jun 6
Jun 20


NO2
Delhi
Mumbai
Bengaluru
Kolkata
Apr 11
2021
Apr 25
May 9
May 23
Jun 6
Jun 20

PM25


Delhi
100
0
−100
Apr 11
2021
Apr 25
May 9
May 23
Jun 6
Jun 20
Mumbai
0
−20
−40
Bengaluru
40
20
0
−20
Kolkata
50
0
100
50
0
−50
−100

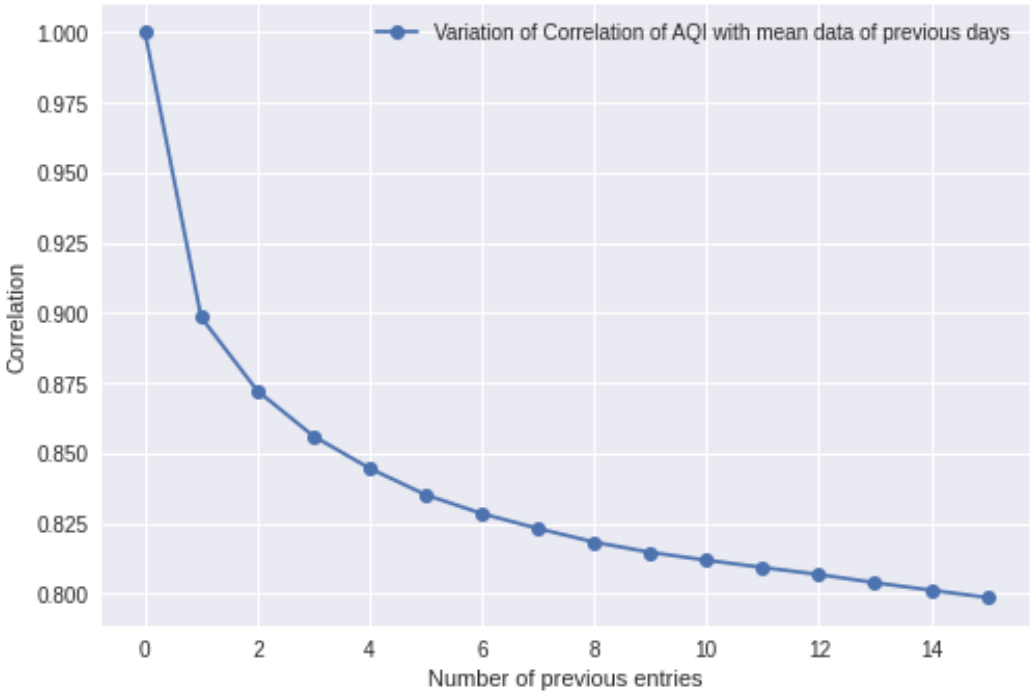
PM10


Delhi
400
200
0
−200
Apr 11
2021
Apr 25
May 9
May 23
Jun 6
Jun 20
Mumbai
50
0
−50
Bengaluru
100
50
0
−50
Kolkata
100
50
0
−50
−100
300
200
100
0
−100

Prediction of AQI based on current pollutant concentration

It would be really helpful for the government if there is some method to predict the AQI of the upcoming days based on the concentration of the pollutants given for the past few days.

To study the relation between AQI of consecutive days, we took the mean AQI for previous n days (1 <= n <= 15) and studied the correlation of the those mean AQIs with the AQI of the current day



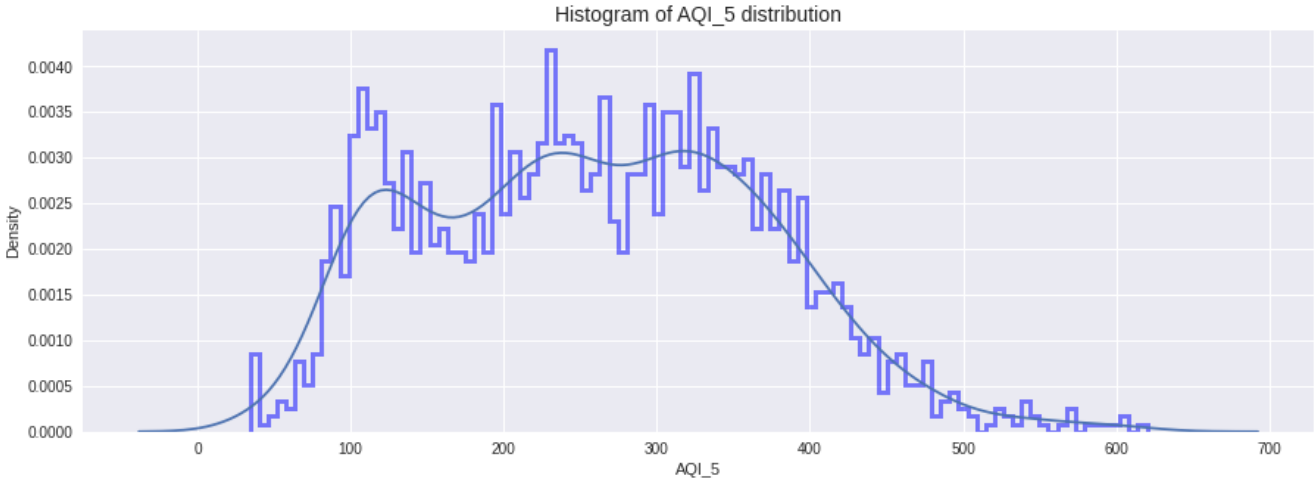
This graph clearly shows that there is significant correlation between AQIs of consecutive days. Going by the same reasoning we can say that we can to an extent correlate the different pollutant correlation of consecutive days. These make out initial set of features. To keep the number of initial features reasonable, the average values for previous {3,5,10,15} days were chosen. Selecting the given values for the feature set generated a total of 28 features.

Filling NAN values

The data had some NAN values. In order to tackle the problem, the mean of the column was taken and used in place of the NAN values. Other methods like median value selection or linear interpolation did not result in much different values for the placeholders, hence in order to keep the model simple, mean values were taken.

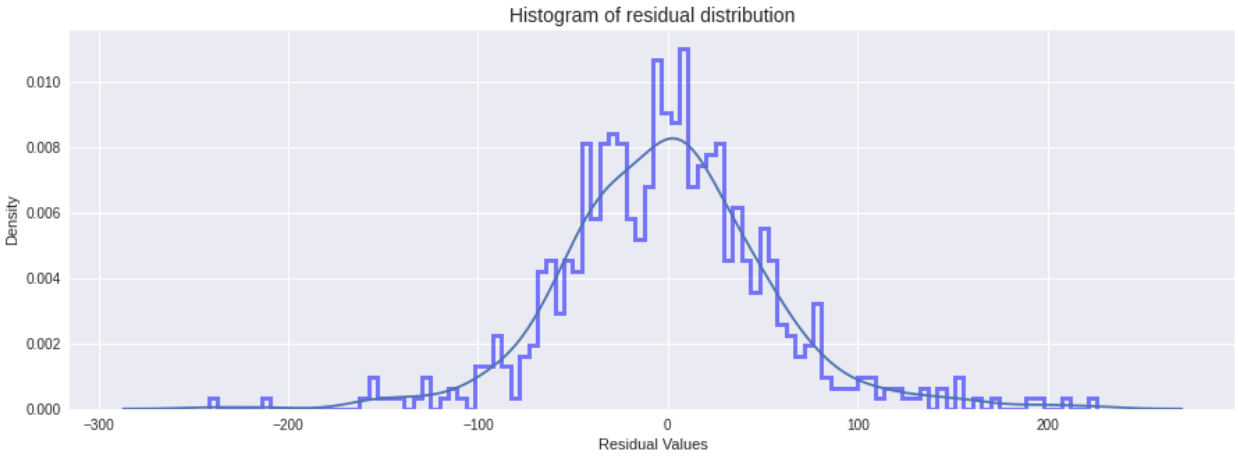
Normalization

The data ranges of the different pollutants vary greatly, hence in order to have a meaningful model some normalization techniques need to be used. The data of the pollutants does not have a mean distribution as shown below for one of the features. Hence Standard Scaler is not used, in place RobustScaler is being used which does well for data which does not have a normal distribution and might have outliers.



Mapping the residuals

In order to see the effectiveness of a linear model with the given feature set, we can plot the distribution of residuals.



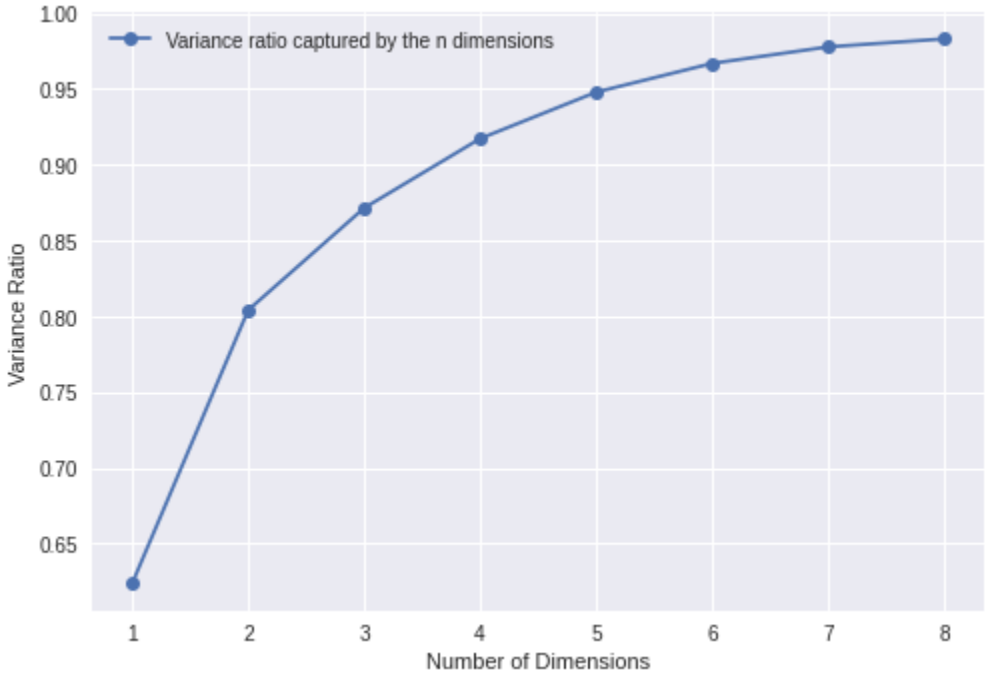
The distribution of residuals show that it has a normal distribution, hence a linear model would function well for the given dataset.

Feature Selection

Though the feature set is small (only comprising of 24 features), we can still perform feature selection method or dimensionality reduction based techniques to further improve the model and make the model faster.

PCA

One of the methods used for dimensionality reduction is PCA.

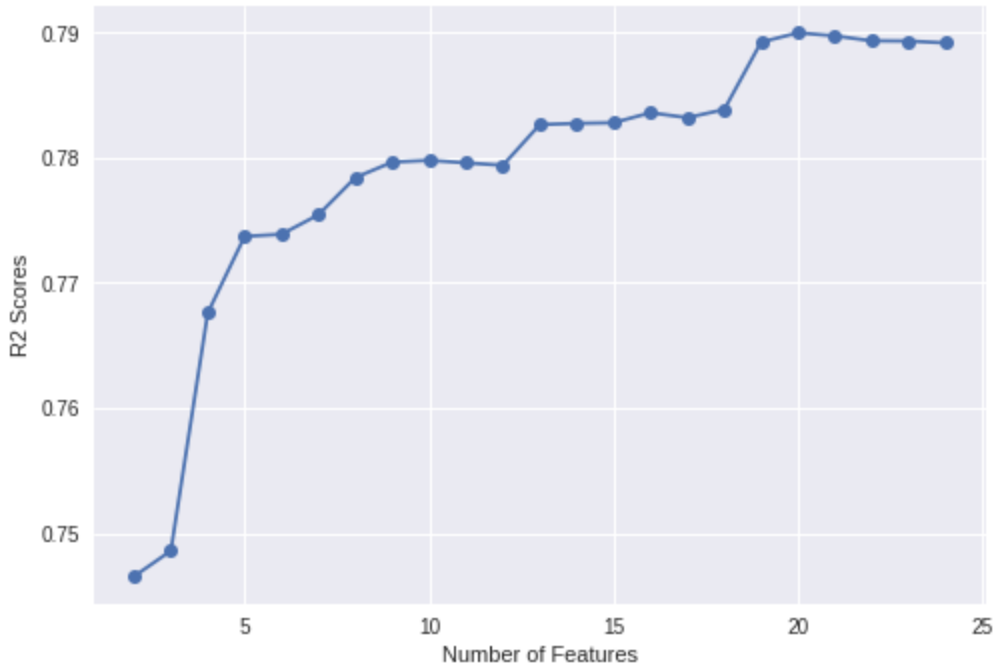


The above graph clearly shows that even 8 dimensions are able to capture the majority of the variance of the input data. Running a Linear Model on these 8 dimensions does not result in good results in comparison to a model with no feature selection.

Recursive Feature Elimination (RFE)

In this method we try to remove features 1 at a time recursively till we get the desired number of features.

Using Lasso Regression based method for feature selection, and then running a Ridge Regression based model on the selected features gives the following graph:



This graph shows that the R2 scores are more or less the same on changing the number of features. The best results are obtained when we have 20 features. This model is not prone to overfitting as the results on training and test data are almost the same.

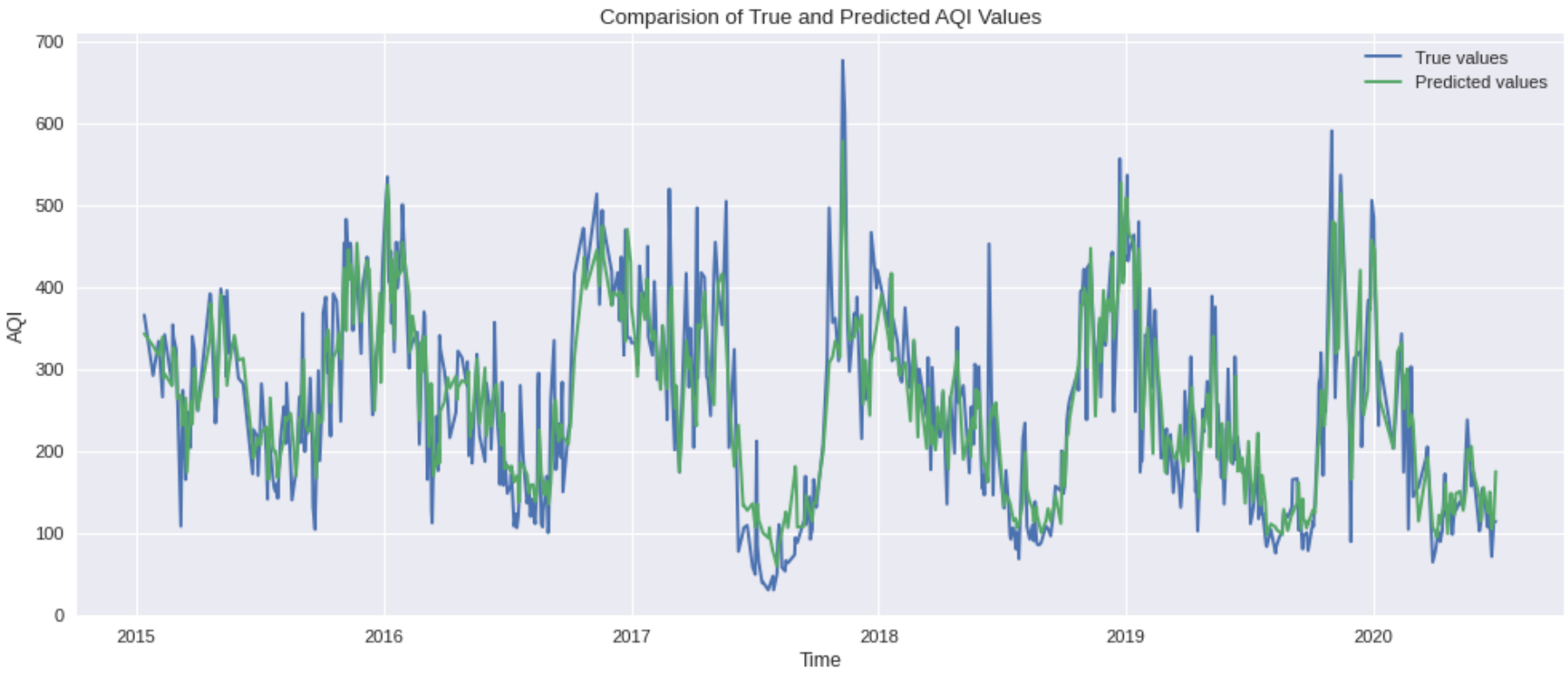
Trying Nonlinear Models

In order to check whether non linear models might perform better than linear models certain models based on Decision Trees Ensemble and Boosting methods were used. These models were prone to extreme overfitting as can be seen in the graph below.

This is a very bad model as it highly overfits the training data. In order to create non linear models, some polynomial features of order 2 were included and then Ridge Regression was applied on them after performing feature selection. The result of the model did not improve by much but the number of features exploded, hence polynomial features were not used for the final model.

Final Prediction

The graph on the next page shows the comparison of the true and the predicted values of the AQI.


Comparision of True and Predicted AQI Values
True values
Predicted values
AQI
Time
0
100
200
300
400
500
600
700
2015
2016
2017
2018
2019
2020

References:

1. Why air pollution is worse in winter? - Accuweather
2. Latitude and Longitude of Indian Cities
3. Ocean effect on AQI
4. Urban air pollution – what are the main sources across the world?
5. Ambient (outdoor) air pollution
6. Nitrogen oxides concentration and emission change detection during COVID-19 restrictions in North India
7. Temporal variations of atmospheric CO2 and CO at Ahmedabad in western India