DS250 Project Report

Study of Air Quality Index across India

Submitted by:
Aryan Jain (11840210)
Blobhit Behera (11840320)
Naved Koser Ansari (11840750)
Pothukuchi Siddartha (11840800)
Pratham Mittal (11840840)

Study of Air Quality Index across India

Link to drive folder : ***Our work***

PROJECT INTRODUCTION AND OBJECTIVE:

The air quality index has been a major issue since the Industry revolution and impacts the world significantly. This project gives an insight to people of all domains, such as the Investors, the Government, and even the common people.

To make it easier, our model finds out the cause of the pollution (Factories, Vehicles..etc), which brings out more specific methods to improve the air quality.

AIR QUALITY INDEX (AQI):

The AQI is an index that reports daily air quality. It tells us how clean or polluted the air in a particular area is, and the associated health effects that might be a concern for the people in the area. Air contains smaller amounts of many other gases and particles. AQI tracks five major air pollutants:

- Ground level ozone
- Carbon monoxide
- Sulfur dioxide
- Nitrogen dioxide
- Airborne particles, or aerosols

Computation of the AQI requires the air pollutant concentration of places over a specified averaging period, obtained from an air monitor or model. Taken together, concentration and time represent the dose of the air pollutant.

The AQI can increase due to an increase of air emissions (for example, during rush hour traffic or when there is an upwind forest fire) or from a lack of dilution of air pollutants. Stagnant air, often caused by an anticyclone, temperature inversion, or low wind speeds lets air pollution remain in a local area, leading to high concentrations of pollutants, chemical reactions between air contaminants and hazy conditions.

BUSINESS OBJECTIVE: The study makes it easy for the investor/Government to find out the regions across the country with high pollution along with the major pollutant so that they can invest in reducing pollution. It generates data that gets updated dynamically along with a trend set for a week for a better understanding of the regions.

DATA SOURCES:

URL: [Real time air quality index of various locations.]

Filters: API obtained from the above URL,
offset=0,
format=json,(For easy usage as dictionaries in python),
limit=2000,(Maximum number of data-points obtained in one hour=~1400)

```python
import requests
import pandas as pd
import datetime
import calendar
import os
r = requests.get('https://api.data.gov.in/resource/3b01bcb8-0b14-4abf-b6f2-c1bfd384ba69?api-key=579b464db66ec23bdd0000010383c9072fc6459b4bbb9817c4d522c6&format=json&offset=0&limit=2000')
x=r.json()
data=pd.DataFrame(x['records'])
date=data['last_update'][0]
date_x=date.split(" ")[0]
date_pa=datetime.datetime.now().strftime("%p")
def findDay(date):
    born = datetime.datetime.strptime(date, '%d-%m-%Y').weekday()
    return (calendar.day_name[born])
day=findDay(date_x)
path='/home/siddharth/Semester_V/DS250/DS250_Project/CSV/'
name_csv=date+date_pa+" "+day
data.to_csv(os.path.join(path,(name_csv+'.csv')))
```

DATA COLLECTION:

In this process of gathering and measuring information on targeted variables, it enables one to evaluate to answer relevant questions and evaluate the desired outcomes. To answer the research question more accurately, and to provide qualitative research we started collecting data on September 22nd.

This project consists of a wide range of problem statements, which require hourly, daily, weekly and monthly data. To make our work easier we have used **Crontab** for running our python code automatically.

The following steps were used for **Crontab.**

```
#EXECUTE IN LINUX TERMINAL

1. crontab -e

#AT THE END OF CRONTAB, WRITE THE FOLLOWING LINE

2. * */1 * * * <path of python interpretor> <path of python file>
   [This indicates every one hour all the days.]
```

DATA CLEANING:

In raw data we had every station with every pollutant in different rows, so fixed it with increasing the number of columns and reducing the number of rows also dropped some of the columns which had Nan values, here are the screenshots for the same.

RAW DATA:

Out[3]:

id	country	state	city	station	last_update	pollutant_id	pollutant_min	pollutant_max	pollutant_avg	pollutant_unit
1	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	15-10-2020 09:00:00	PM2.5	13.0	52.0	24.0	NaN
2	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	15-10-2020 09:00:00	PM10	16.0	53.0	25.0	NaN
3	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	15-10-2020 09:00:00	NO2	22.0	53.0	31.0	NaN
4	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	15-10-2020 09:00:00	NH3	2.0	3.0	3.0	NaN
5	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	15-10-2020 09:00:00	SO2	12.0	47.0	23.0	NaN
6	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	15-10-2020 09:00:00	CO	19.0	57.0	40.0	NaN
7	India	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	15-10-2020 09:00:00	OZONE	27.0	61.0	48.0	NaN
8	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	15-10-2020 09:00:00	PM2.5	12.0	63.0	27.0	NaN
9	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	15-10-2020 09:00:00	PM10	15.0	81.0	29.0	NaN
10	India	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	15-10-2020 09:00:00	NO2	12.0	64.0	23.0	NaN

CLEANED AND MODIFIED DATA :

In [946]:
```
result_df
```

Out[946]:

Station	Date_time	City	State	PM10_min	PM10_max	PM10_avg	PM2.5_min	PM2.5_max	PM2.5_avg	NO2_min	...	SO2
Secretariat, Amaravati - APPCB	16-10-2020 02:00:00	Amaravati	Andhra_Pradesh	17.0	60.0	39.0	13.0	52.0	32.0	22.0	...	
Anand Kala Kshetram, Rajamahendravaram - APPCB	16-10-2020 02:00:00	Rajamahendravaram	Andhra_Pradesh	24.0	87.0	44.0	12.0	70.0	35.0	8.0	...	
Tirumala, Tirupati - APPCB	16-10-2020 02:00:00	Tirupati	Andhra_Pradesh	17.0	70.0	34.0	15.0	45.0	27.0	8.0	...	
GVM Corporation, Visakhapatnam - APPCB	16-10-2020 02:00:00	Visakhapatnam	Andhra_Pradesh	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
Railway Colony, Guwahati - APCB	16-10-2020 02:00:00	Guwahati	Assam	33.0	174.0	76.0	NaN	NaN	NaN	8.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
Jadavpur, Kolkata - WBPCB	16-10-2020 02:00:00	Kolkata	West_Bengal	22.0	64.0	35.0	7.0	37.0	21.0	14.0	...	
Rabindra Bharati University, Kolkata - WBPCB	16-10-2020 02:00:00	Kolkata	West_Bengal	44.0	106.0	72.0	23.0	87.0	55.0	11.0	...	
Rabindra Sarobar, Kolkata - WBPCB	16-10-2020 02:00:00	Kolkata	West_Bengal	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
Victoria, Kolkata - WBPCB	16-10-2020 02:00:00	Kolkata	West_Bengal	19.0	110.0	48.0	5.0	83.0	32.0	29.0	...	
Ward-32 Bapupara, Siliguri - WBPCB	16-10-2020 02:00:00	Siliguri	West_Bengal	86.0	153.0	110.0	79.0	244.0	126.0	17.0	...	

216 rows × 25 columns

Observe that we have dropped the country column as all the locations are within India and also dropped the pollutant unit column as we received NaN for that. Also, we have merged these dataframes to keep all the data of a given day in a single dataframe. This reduced the size of one csv from (~140KB) to (~36KB).
Now we can select data from the dataframe as per our need.
Also, the missing values were replaced with a NaN.

DATA ANALYSIS:

The visualisations provided various insights into the data and we can derive various results from the plots.

1. We can clearly identify the most polluted cities.
2. We can analyze the change in pollutant levels to learn which hours of the day are most polluted.
3. We can see that the number of regions/states/stations affected by pollutants increases over time.
4. We can determine the most significant pollutant in a city and based on it we can try to determine the main source of pollution.
5. We can compare the AQI post and pre lockdown to determine the effect of lockdown on pollution levels.

EDA :

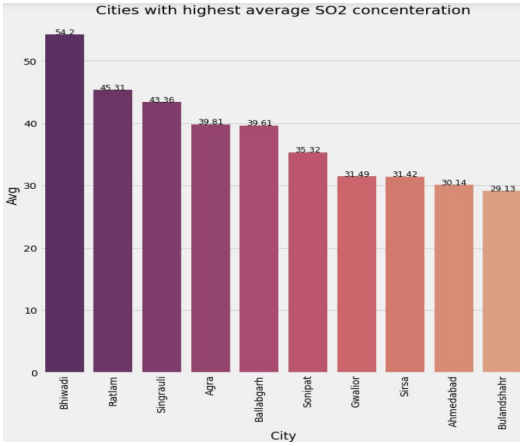
Libraries/Software Used: matplotlib, seaborn, folium,Tableau

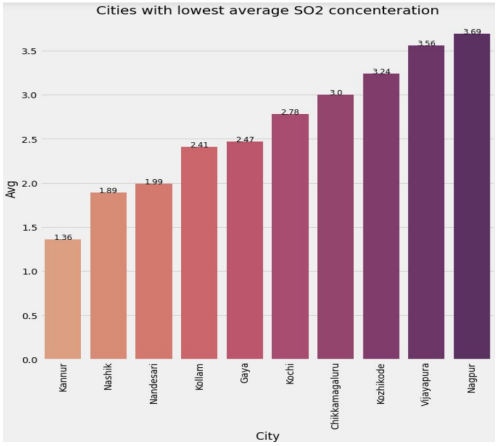
We have used various plots to study the following:

- Change in pollutant levels in a city as the day goes by.
- Change in pollutant levels in a city over a week.
- Variation in pollutant concentration across different cities.
- The worst and the least polluted cities across the country.
- Plotted the cities on the Indian map with the radius of each city representing the pollutant concentration.

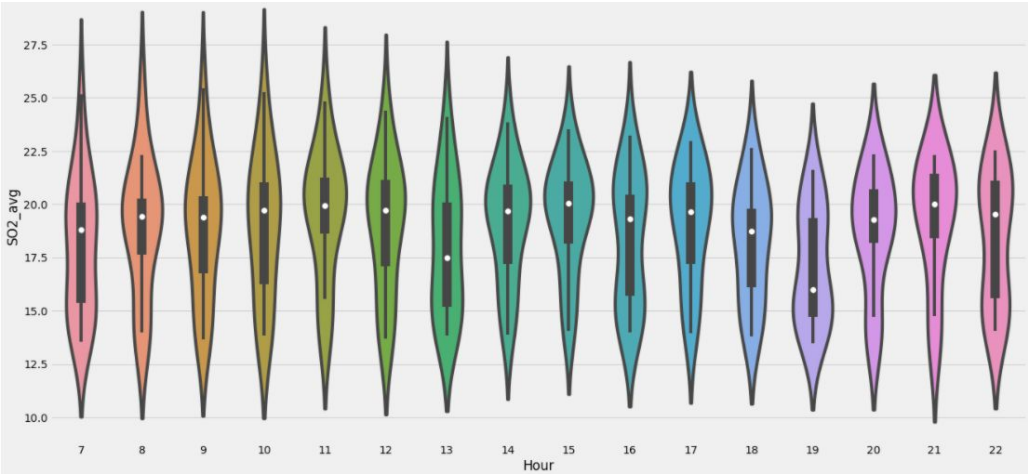
Plots for SO_2 *(Similar plots have been made for each of the 7 pollutants)*

<u>Worst and Least polluted cities in India</u>

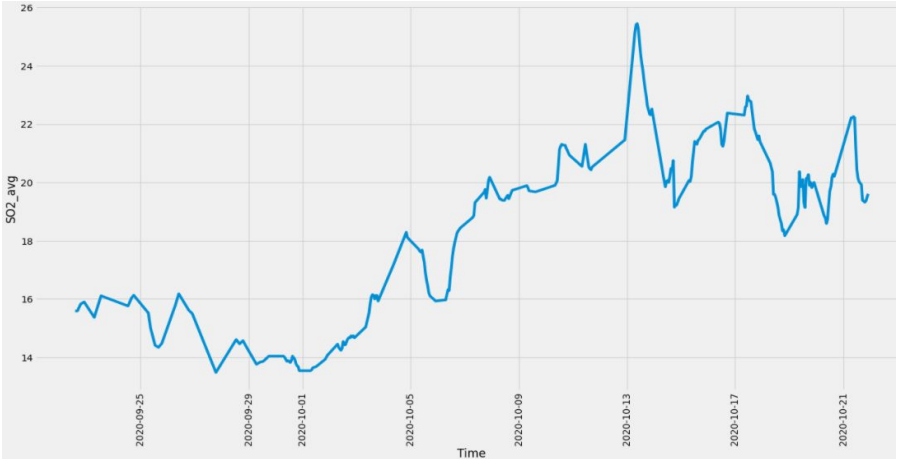




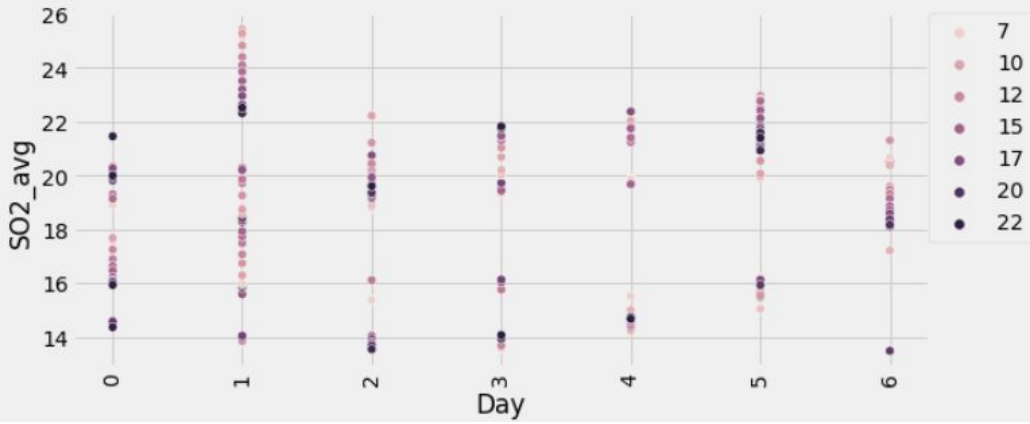
CITY : DELHI



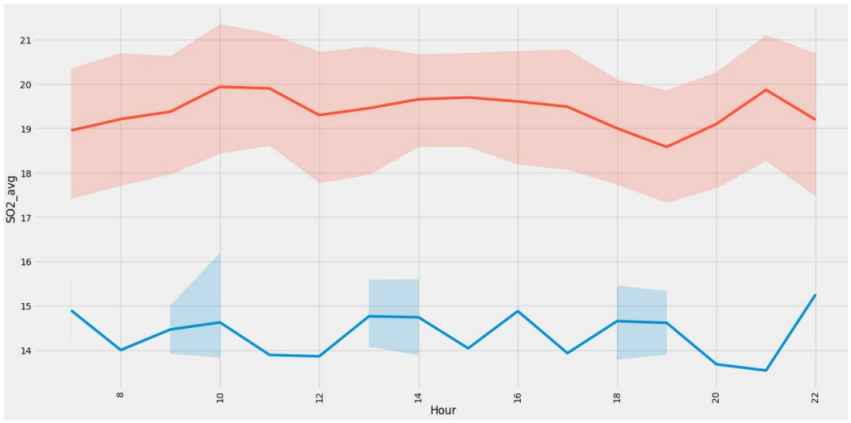
Variation in average SO2 levels during a day



Variation in average SO2 levels during the period of our data collection in Delhi

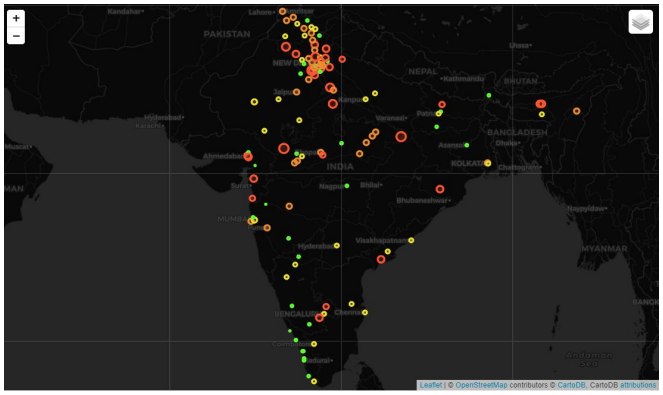


Variation in average SO2 levels in Delhi according to day of the week colored with respect to hour of the day

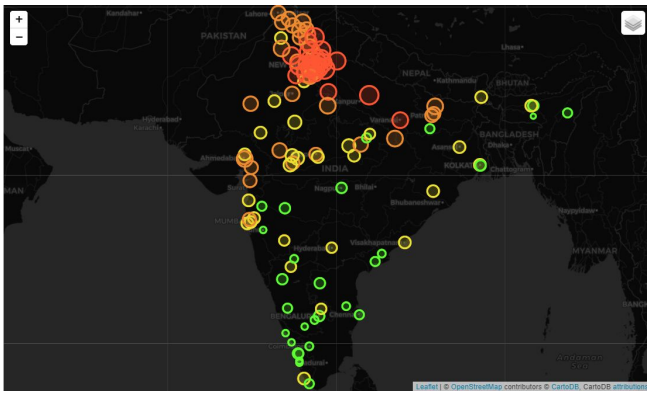


SO2 average levels in Delhi (orange for october, blue for september)

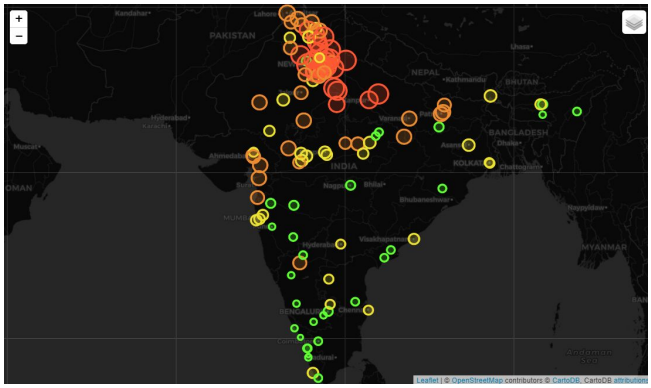
MAPS



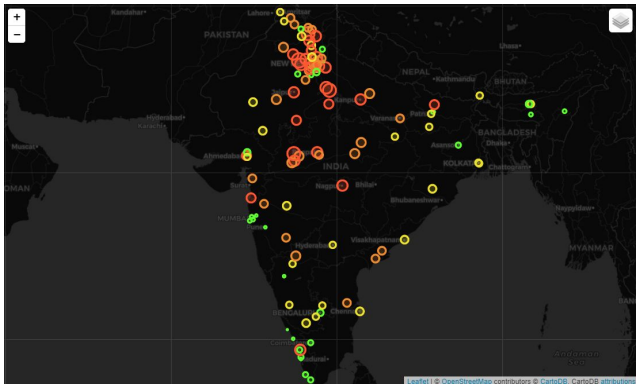
SO2



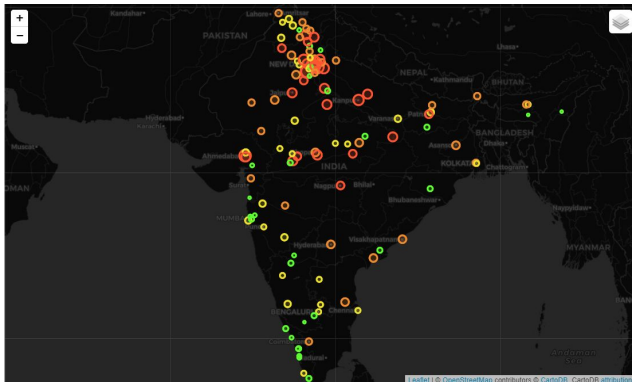
PM10



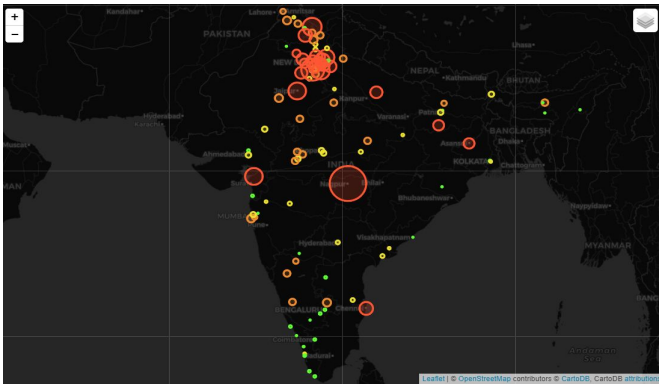
PM2.5



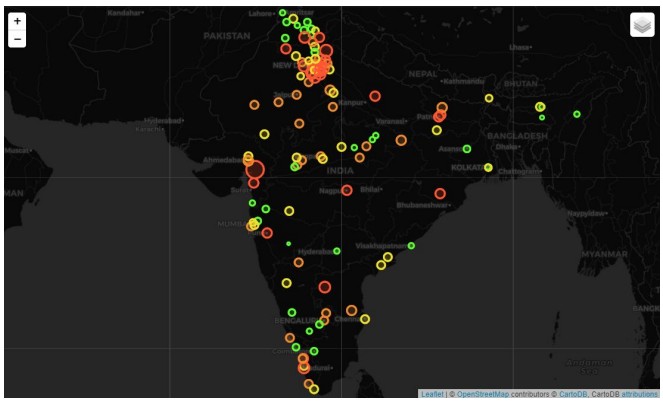
OZONE



NO2



NH3



CO

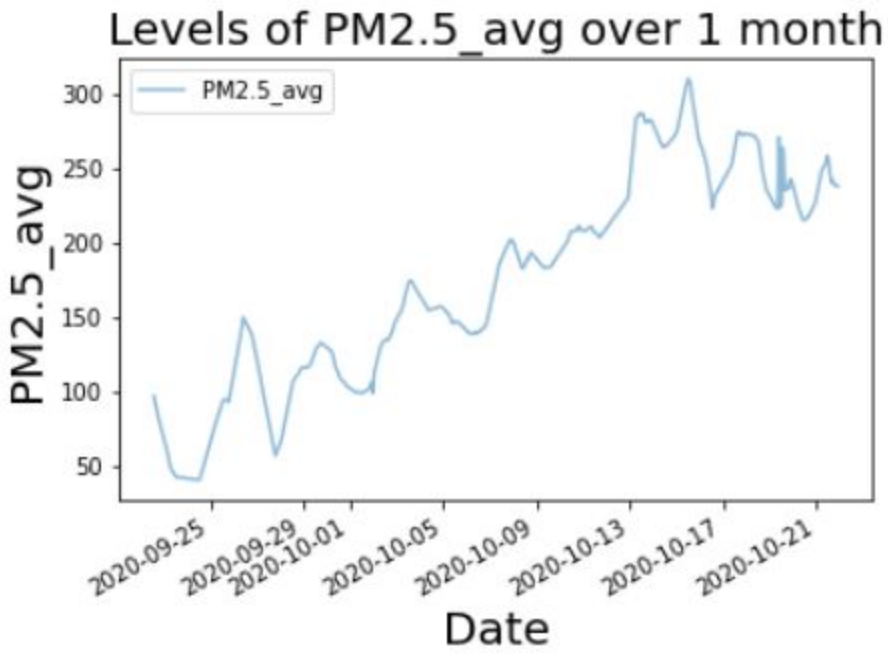
OBSERVATIONS FROM EDA

- There aren't many AQI stations in India.
- Many areas don't have one while some cities have AQI stations in abundance.
- In case of cities like Ahmedabad, there is only one station. This gives a reading which shows that Ahmedabad has clean air, while in reality it is highly polluted.
- In various cities like Varanasi, AQI stations measure concentration of only a couple of pollutants.
- In cities like Visakhapatnam, there are stations, but they don't give readings.
- We can see a rise in pollutant concentration from 1st of October. This is due to the fact that Lockdown regulations were made less stringent.
- Also, observed was the huge difference between PM2.5 particles whose primary source is vehicles.
- Pollution across a week varies greatly across different cities and for different pollutants . In the case of Delhi and pollutant SO2, Saturday was the most polluted and Friday the least while for PM2.5 Sunday was the most polluted and Friday still being the least polluted.

DATA MODELING:

Primarily we have used the ARIMA model for the modeling purpose as the data we wanted to forecast was a univariate time series .
Later after observing the data, we have also worked with SARIMA models and those gave better accuracy than ARIMA models in most cases.
We have used a 90/10 split for the validation of our model.

We have done modeling for the pollutant PM2.5_avg and for four cities in this project.
First city that we have worked on was DELHI
Let's see what is in there in modeling:
First we have taken out our target city from the final dataframe and then converted the dataframe as Univariate Time series
After this we have done Resampling of this data at 12H
This results in our final time series



After this we have checked if our time series is stationary or not by performing AD fuller test

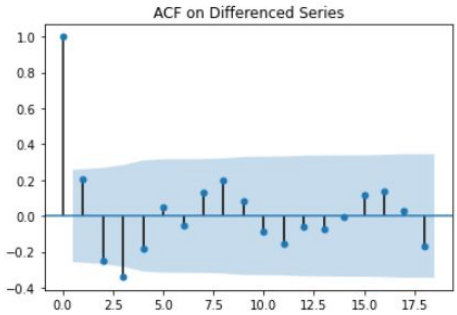
```
In [15]: from statsmodels.tsa.stattools import adfuller
         print("p-value:", adfuller(b['PM2.5_avg'].dropna())[1])

         p-value: 0.473341139756345

In [16]: print("p-value:", adfuller(b['PM2.5_avg'].diff().dropna())[1])

         p-value: 1.0422639955747807e-08
```

After this we have checked the seasonality in the data and found its value to be 7

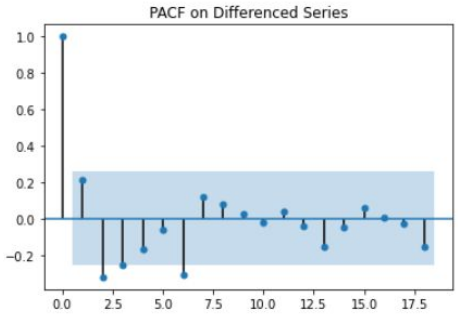
Then we have plotted ACF and PACF plots to find out q and p for the ARIMA model and after inferring these values we have also altered some values to get better accuracy for the model and this will be more clear when you see the notebook

So from the ACF and PACF plots , have selected p=2 and we have selected q=1 to keep our model simple, and also it yields better accuracy.

```
In [38]: fig = plot_acf(b['PM2.5_avg'].diff().dropna(),
                        title="ACF on Differenced Series")
```




```
In [39]: fig = plot_pacf(b['PM2.5_avg'].diff().dropna(),
                         title="PACF on Differenced Series")
```



Out[40]:

ARIMA Model Results

Dep. Variable:	D.PM2.5_avg	**No. Observations:**	51
Model:	ARIMA(2, 1, 1)	**Log Likelihood**	-216.546
Method:	css-mle	**S.D. of innovations**	16.269
Date:	Wed, 18 Nov 2020	**AIC**	443.092
Time:	17:46:48	**BIC**	452.751
Sample:	09-23-2020	**HQIC**	446.783
	- 10-18-2020		

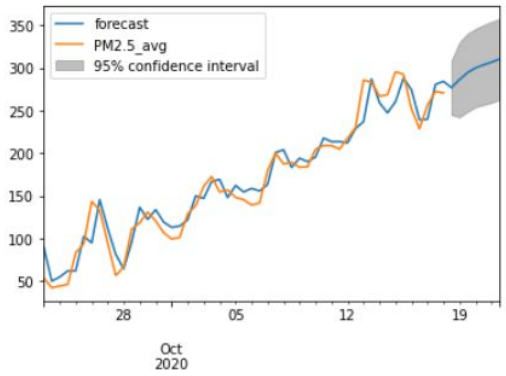
	coef	std err	z	P>|z|	[0.025	0.975]
const	4.4538	0.276	16.152	0.000	3.913	4.994
ar.L1.D.PM2.5_avg	0.9501	0.123	7.694	0.000	0.708	1.192
ar.L2.D.PM2.5_avg	-0.5028	0.121	-4.147	0.000	-0.740	-0.265
ma.L1.D.PM2.5_avg	-1.0000	0.074	-13.430	0.000	-1.146	-0.854

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.9448	-1.0470j	1.4103	-0.1332
AR.2	0.9448	+1.0470j	1.4103	0.1332
MA.1	1.0000	+0.0000j	1.0000	0.0000

This is the model summary for the ARIMA(2,1,1) model .

```
In [44]: results3.plot_predict(1, 58)
         plt.show()
```




```
In [45]: print('Mean Absolute Percent Error:', round(np.mean(abs(residuals3/test_data)),6))
```

```
Mean Absolute Percent Error: PM2.5_avg    0.26321
dtype: float64
```

We have got 26.32% MAPE for this model and every P>|z| is also 0.00
After this we have tried modeling with SARIMAX and for selecting the parameters for the SARIMAX model we have used auto arima
This is what auto arima gave
order = (0,1,1)
seasonal_order = (0, 0, 0, 7)
Then we have built the model with these parameters and got the following summary:

```
In [51]: #summary of the model
         print(model4_fit.summary())
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              PM2.5_avg   No. Observations:                   52
Model:               SARIMAX(0, 1, 1)   Log Likelihood                -225.350
Date:                Wed, 18 Nov 2020   AIC                            454.699
Time:                        17:46:58   BIC                            458.563
Sample:                    09-22-2020   HQIC                           456.176
                         - 10-18-2020
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1          0.3549      0.133      2.677      0.007       0.095       0.615
sigma2       402.0953     74.114      5.425      0.000     256.835     547.356
===================================================================================
Ljung-Box (Q):                       43.35   Jarque-Bera (JB):                 1.90
Prob(Q):                              0.33   Prob(JB):                         0.39
Heteroskedasticity (H):               0.75   Skew:                             0.47
Prob(H) (two-sided):                  0.57   Kurtosis:                         3.07
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

```
In [54]: print('Mean Absolute Percent Error:', round(np.mean(abs(residuals4/test_data)),5))

         Mean Absolute Percent Error: PM2.5_avg    0.1643
         dtype: float64

In [55]: print('Root Mean Squared Error:', np.sqrt(np.mean(residuals4**2)))

         Root Mean Squared Error: PM2.5_avg    39.412945
         dtype: float64
```

So we are getting MAPE = 16.43% with our SARIMAX model and P>|z| values are also well within limits so this is what we have selected as our final model for the city DELHI and for pollutant PM2.5.

Like this we have done modeling for four more cities and their respective notebooks are attached. We have done such analysis for Mumbai, Bangalore and Agra.

CHALLENGES FACED + THINGS LEARNT :

We got to learn many things from this project, a part of which is because we also faced many difficulties throughout the project.

One of the major challenges faced was collecting the data hourly throughout the month. To overcome this challenge, we have automated the process of running our python script, using Crontab. But there were times of power failures (causing inability to keep the Crontab running) so we also collected data through the government api manually for several turns.

While Researching about the AQI data and its classification in the country, we came across standards of other countries. Their rules and standard values for different pollutants clearly indicated the difference in the level of pollution as compared to our country. Also, we came across the official documentation releases (2010) of Indian Govt. related to Air Pollution which gave us an idea as to how can we better implement the data, what more aspects can we analyse and visualize, regarding the calculation of the AQI, and how the environment has changed from 2010 as compared to 2020, looking just at the values corresponding to the pollutants.

We ourselves learnt more about the AQI and became aware of the measures taken by the government to keep track of the pollution while working for this project.

Talking about modeling, sometimes the model parameters couldn't just be selected by looking at the ACF and PACF plots and also they might not give the best performance and summary of the data, so it has to be tuned and checked accordingly. We could have done more modeling with other cities and other pollutants to forecast the particular action that could be taken to resolve the issue but there were time constraints.

All in all it was a good experience, working on a team project in the lockdown.

REFERENCES :

- https://www.kaggle.com/parulpandey/breathe-india-covid-19-effect-on-pollution
- https://www.kaggle.com/frtgnn/clean-air-india-s-air-quality
- https://towardsdatascience.com/geocode-with-python-161ec1e62b89
- https://machinelearningmastery.com/time-series-data-visualization-with-python/
- https://towardsdatascience.com/a-complete-guide-to-an-interactive-geographical-map-using-python-f4c5197e23e0
- https://www.kaggle.com/sumi25/understand-arima-and-tune-p-d-q
- https://data.gov.in/resources/real-time-air-quality-index-various-locations/api
- https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/
- https://www.breeze-technologies.de/blog/what-is-an-air-quality-index-how-is-it-calculated/
- https://visualize.data.gov.in/?page=my-data

Duties of team members

Team Coordinator: Aryan Jain

Member Name	Data Collect/Clean	Data Model	Analysis	Data Viz.	Report
Aryan	yes	no	yes	yes	yes
Pratham Mittal	yes	yes	yes	no	yes
Siddhartha Pothukuchi	yes (wrote the script)	no	yes	no	yes
Naved Koser Ansari	yes	no	yes	yes	yes
Blobhit	yes	no	yes	no	yes