Prediction of Air Quality Index using Time Series

Drishti Chulani

Ovi Jadhav

Sahil Raorane

Time Series Report

Submitted to the

School of Data Science and Business Intelligence

For the

Post Graduation Degree

In

Data Science and Business Analytics

May 2023

Abstract

The Air Quality Index (AQI) is a number that is used to quantify and convey the state of the air at a given time and place. It offers details on the quantity of contaminants in the air and their possible effects on human health.

Monitoring the AQI helps individuals, communities, and policymakers understand the air quality in their surroundings, make informed decisions regarding outdoor activities, and take appropriate measures to protect their health. It also assists in evaluating the effectiveness of air pollution control measures and implementing necessary interventions to improve air quality.

During the COVID-19 pandemic, there were significant changes in mobility patterns and economic activity due to lockdowns, travel restrictions, and work-from-home policies. Analyzing the impact of these changes on AQI can provide insights into how reduced industrial activities, transportation, and human movement affected air pollution levels.

We use SARIMA (Seasonal Autoregressive Integrated Moving Average) to do time series forecasting and it can be applied to predict future AQI values based on historical data. According to this model, we can predict the AQI values for 2021 and according to the trend, it seems to be reducing quite well.

Introduction

The air quality index (AQI) is an index for reporting air quality on a daily basis. It helps assess the health impact of air pollution on individuals, allowing them to take necessary precautions. A low AQI indicates good air quality and low levels of pollution while a higher AQI suggests increased concentrations of pollutants in the air which is extremely detrimental to human health.

The air quality index is composed of 8 pollutants ((PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb)

The AQI is calculated based on the average concentration of a particular pollutant measured over a standard time interval. AQI scores and categories: Good (0–50) Satisfactory (51–100) Moderately polluted (101–200) Poor (201–300) Very poor (301–400) Severe (401–500).

The dataset is a collection of pollutant readings across cities in India recorded between 2015 and 2020. The data consists of 26 cities in India, and is split into the following categories: -

Date - daily readings between 2015 and 2020
PM2.5 - Particulate Matter 2.5-micrometer in ug / m3
PM10 - Particulate Matter 10-micrometer in ug / m3
NO - Nitric Oxide in ug / m3
NO2 - Nitric Dioxide in ug / m3
NOx - Any Nitric x-oxide in ppb
NH3 - Ammonia in ug / m3
CO - Carbon Monoxide in mg / m3
SO2 - Sulphur Dioxide in ug / m3
O3 - Ozone in ug / m3
Benzene - Benzene in ug / m3
Toluene - Toluene in ug / m3
Xylene - Xylene in ug / m3
AQI - Air Quality Index
AQI Bucket - Air Quality Index Bucket (ranging from 'very poor' to 'good')

We monitor AQI as it aids in evaluating environmental impact, identifying pollution sources, and protecting ecosystems. During emergencies, the AQI assists in issuing timely alerts and implementing emergency measures. AQI predictions enable emergency response operations, assist authorities in designing pollution management plans, assist people in making educated decisions to safeguard their health, and contribute to research on the patterns and effects of air pollution.

Purpose of this Study

In this report, we will be using the day-wise AQI dataset which contains information regarding the daily level of pollutants and AQI in around 26 Indian cities from 2015-2020.

The data is in the form of a csv file and a combination of Python and Pandas libraries has been used to read the file and clean the data (i.e. drop any duplicated data and handle any missing values).

Cities used: Ahmedabad, Aizawl, Amaravati, Amritsar, Bengaluru, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, Visakhapatnam

Our notebook has two major parts:

1. Finding the most polluted cities in recent years and analyzing the levels of pollutants here. Understanding the impact of COVID-19 induced lockdowns on Air Quality in some of the major cities : analyzing which cities underwent the most drastic improvement in Air Quality and which cities showed a spike in AQI levels in spite of a stringent lockdown
2. We do a time-series analysis of the data and fit a SARIMA model with computed orders to forecast India's AQI in 2021. This helps us to see if the claim of air pollution reduction is actually true and can be backed by data.

The time-series analysis of AQI data serves the purpose of understanding historical patterns, forecasting future trends, detecting anomalies, identifying correlations, evaluating policies, and supporting decision-making related to air quality management and public health.

Data Exploration and Cleaning

We find that there are 16 variables in our data with 29531 records. We have daily and hourly city data. We begin by analyzing the various cities' daily data to get a big picture. We begin by importing the dataset and the necessary libraries for the analysis. We check for nas (NULL values) in our dataset, and notice that apart from city and date, we have nas values in the rest of the variables.

City	0
Date	0
PM2.5	4598
PM10	11140
NO	3582
NO2	3585
NOx	4185
NH3	10328
CO	2059
SO2	3854
O3	4022
Benzene	5623
Toluene	8041
Xylene	18109
AQI	4681
AQI_Bucket	4681

Overall, 18.73% of the data is missing. The reason for missing values could be a combination of records missed and data not existing. It is clear that the dataset has increased over time to include more cities across India. We filled the missing values with zero and converted the date column into the DateTime format. We did basic checks and clubbed similar pollutants into single columns to make it easier to handle. For example, we created a new ‘PM’ column that has both of the columns ‘PM2.5’ and ‘PM10’ added. We have used NO, NO2 and NOx to make Nitric column and similarly we have combined Benzene, Toluene and Xylene to create a BTX. we have then dropped the separate columns.

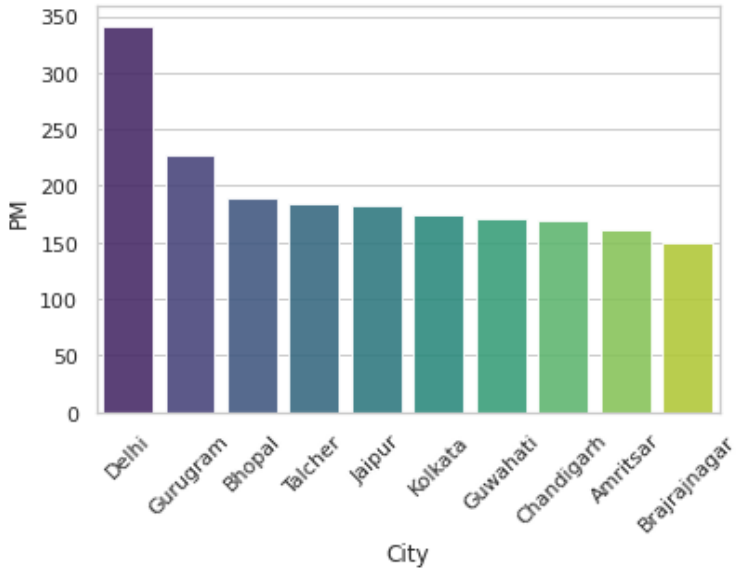
The AQI bucket has 6 categories with the following records

Moderate	8829
Satisfactory	8224
Poor	2781
Very Poor	2337
Good	1341
Severe	1338

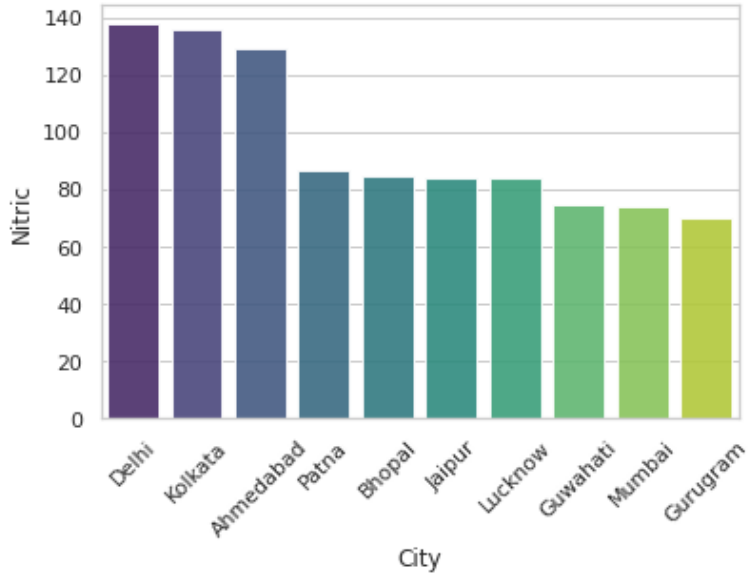
Exploratory Data Analysis

We take years 2017-2019 as our reference years to understand the general trend of pollutants types prevailing in some of the most polluted Indian cities:

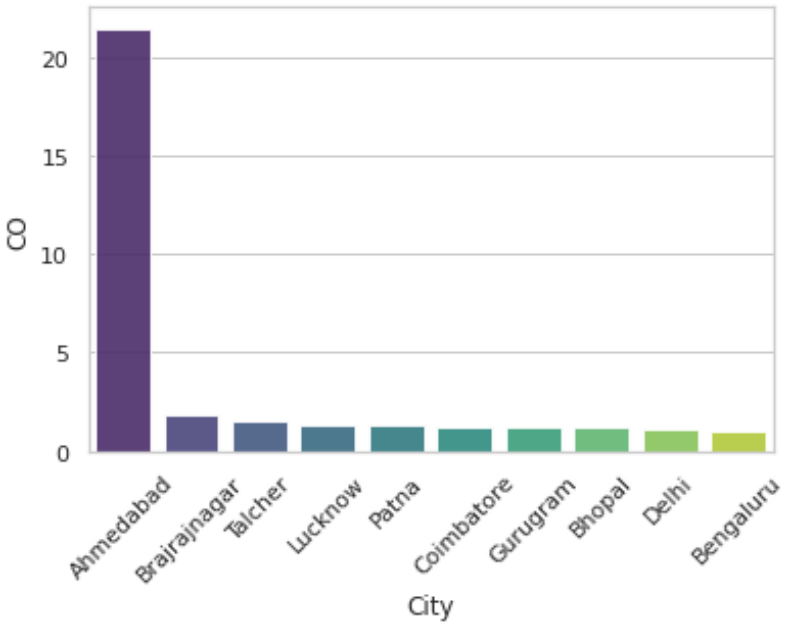
1) Cities with highest PM pollutant



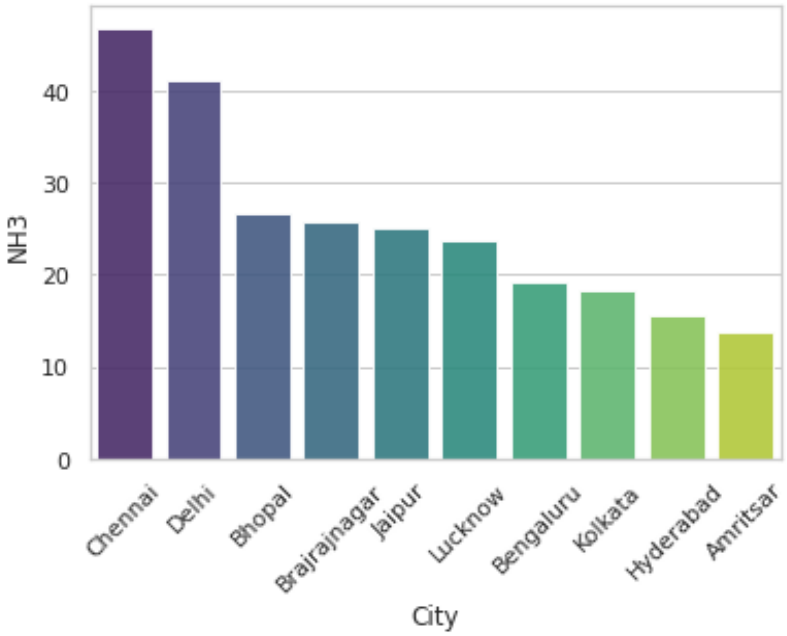
2) Cities with highest Nitric pollutant



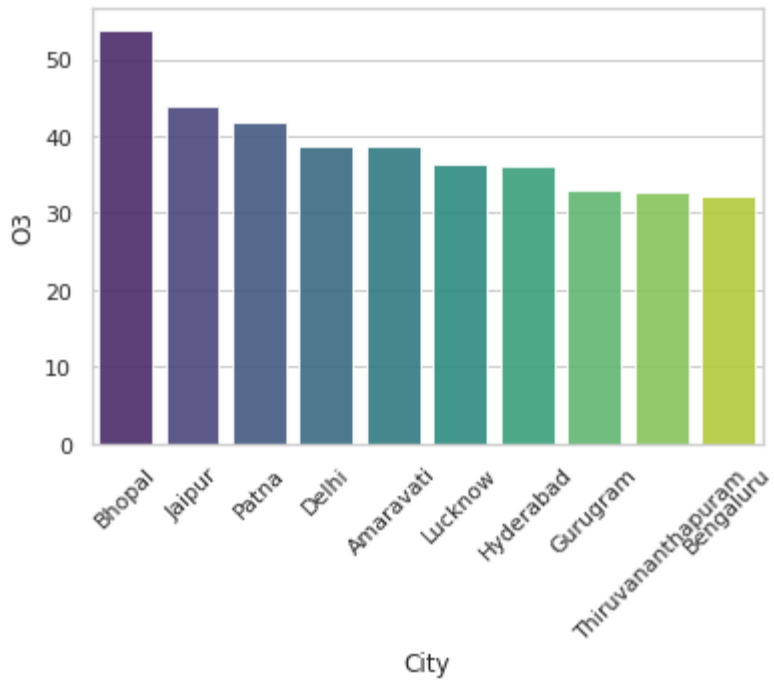
3) Cities with highest CO pollutant



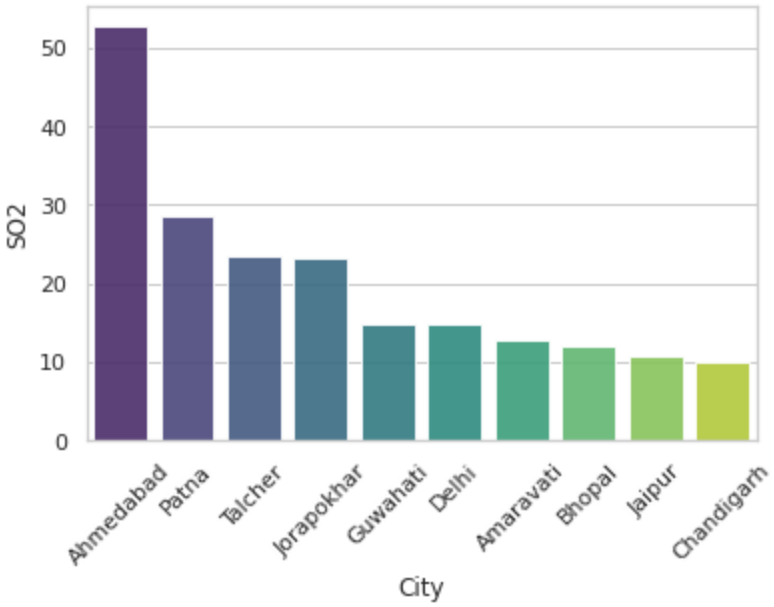
4) Cities with highest NH3 pollutant



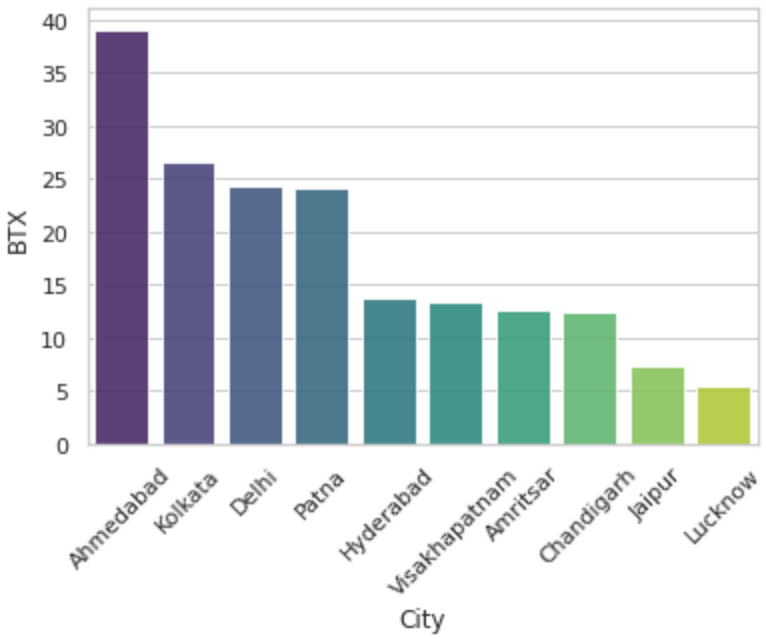
5) Cities with highest O3 pollutant



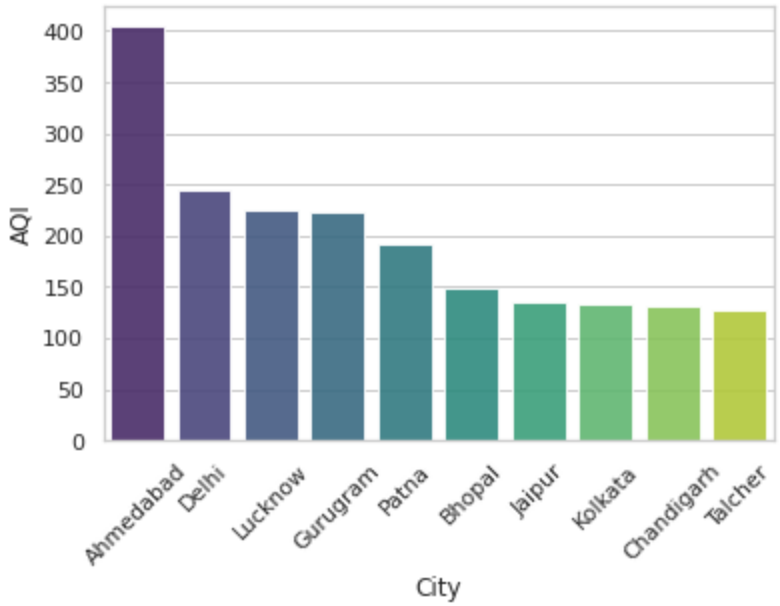
6) Cities with highest SO2 pollutant



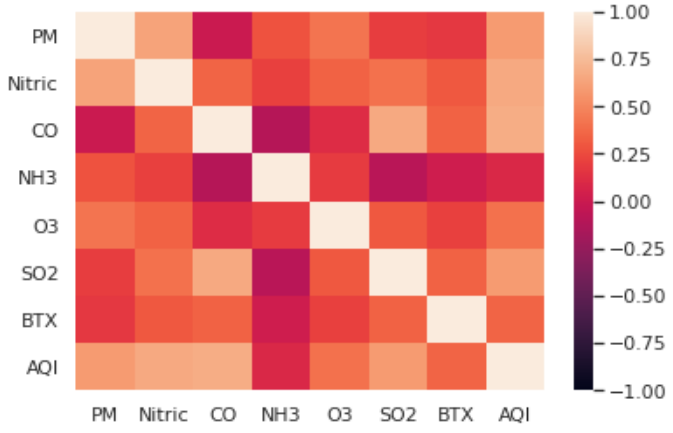
7) Cities with highest BTX pollutant



8) Cities with highest AQI

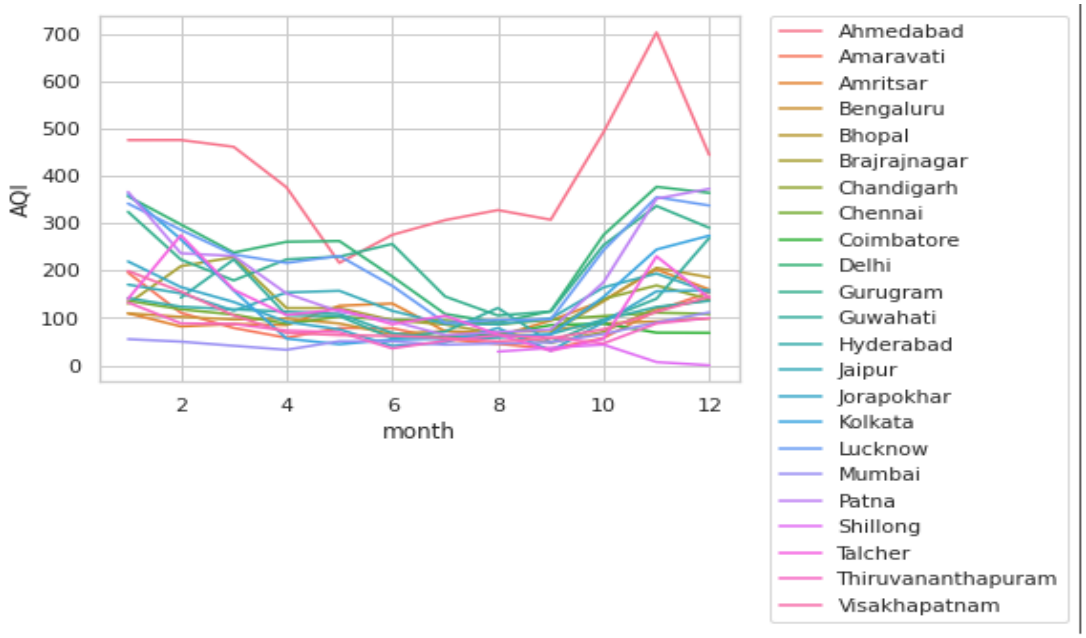


Next, we try to find how each individual pollutant is related to the AQI.



We see that BTX has the lowest correlation with AQI- which is perfectly in sync with the AQI calculation formula. The air quality index is composed of 8 pollutants ((PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb), but does not directly account for BTX.

Next, we study the general AQI trend over the months from the year 2017-19.

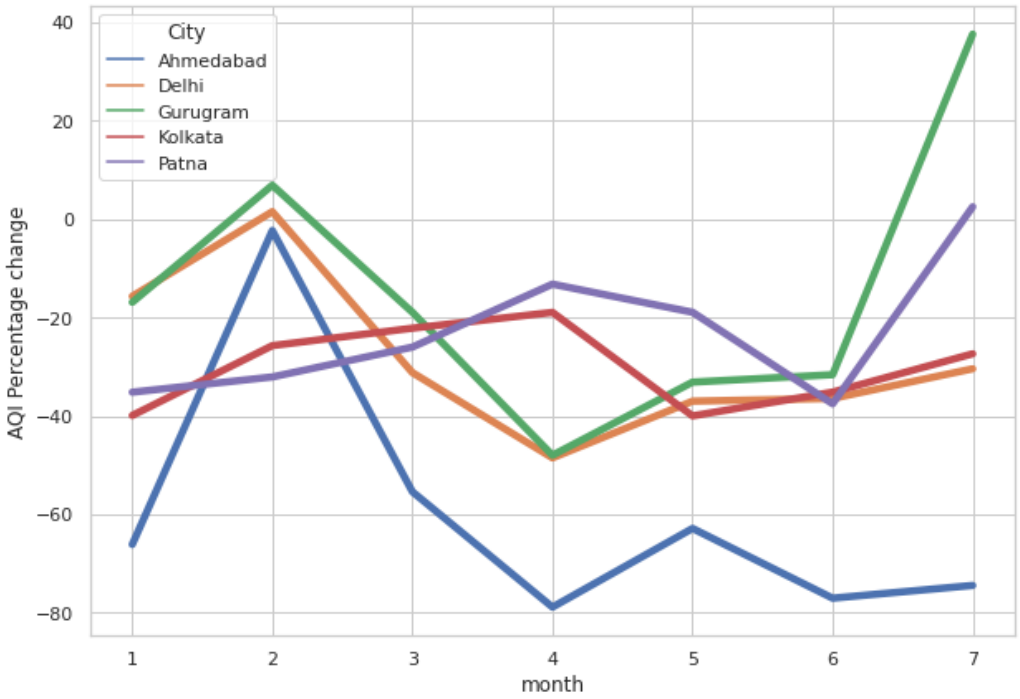


We see that there is a clear pattern which emerges here. AQI decreases in the summer months, the higher temperatures and increased wind speed help disperse pollutants, leading to lower concentrations in the air. Which in turn means that air quality improves over these months.

Analyzing the Impact of Covid-19 induced Lockdown on AQI

To start off, we will be picking out some cities from the most polluted ones(as inferred above) and try to visualize how their AQI changed in 2020 as compared to 2019.

Here from now on, we will be doing our analysis and inference based on percentage change in AQI from 2019 to 2020.



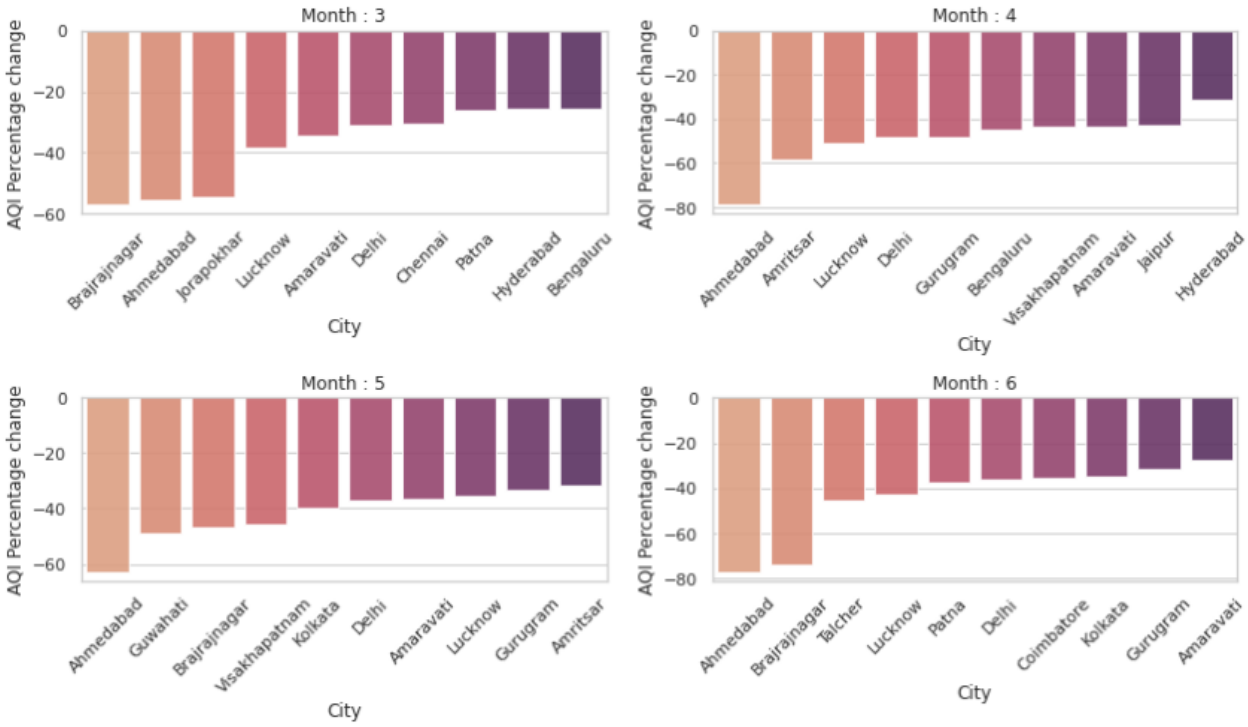
The above graph depicts the 5 most polluted cities in India.

The general trend shows that the AQI indeed decreased for the lockdown months, this can be attributed to the significant reduction in human activities, such as industrial operations, transportation, and other sources of pollution, due to the restrictions imposed during the COVID-19 pandemic.Thus, signifying a major improvement in Air quality with reduced pollution levels.

The decrease in AQI during the lockdown months provided an opportunity to observe the potential air quality improvements that can be achieved by reducing human activities and adopting cleaner practices.

However, we will now investigate the cities which fared the best in these 4 months and also the ones which showed anomalies with a spike in AQI.

Cities that underwent most drastic improvements in Air Quality:



We can see that there has been a significant improvement in the air quality for these cities over the four months.

Cities which showed an increased AQI as compared to 2019 in the lockdown-months:

	City	month	AQI_x	AQI_y	AQI Percentage change
83	Jorapokhar	6	0.000000	136.533333	inf
125	Thiruvananthapuram	6	28.266667	45.400000	60.613208
60	Guwahati	4	105.933333	127.833333	20.673379
82	Jorapokhar	5	113.709677	135.580645	19.234043
31	Brajrajnagar	4	101.633333	119.533333	17.612332
116	Talcher	4	118.466667	127.733333	7.822172
81	Jorapokhar	4	113.833333	121.400000	6.647145

It is important to note that air quality changes during the lockdown period can vary depending on various factors specific to each region, but to the contrary, there were few areas that had their AQI increase:

- Guwahati: we see that AQI for April 2020 is more than 20% higher as compared to April 2019. Particulate matter and NH3 were the increased contributing factors
- Jorapokhar: we see that AQI for May 2020 is substantially higher as compared to May 2019. Concentration of almost all pollutants have increased
- Brajrajnagar: Higher AQI in April 2020 as compared to April 2019. Can be attributed to increased O3 and PM levels
- Talcher: Higher AQI in April 2020 as compared to April 2019. Can be attributed to increased CO,O3 and NH3 levels

While we do not know the exact reasons for what were the reasons, weather conditions changes, air pollution emission increase or seasonal factors were the main cause, we are not sure. It's important to conduct more detailed analysis and examine local factors to understand the specific reasons for the observed increase in AQI during the lockdown period in these regions.

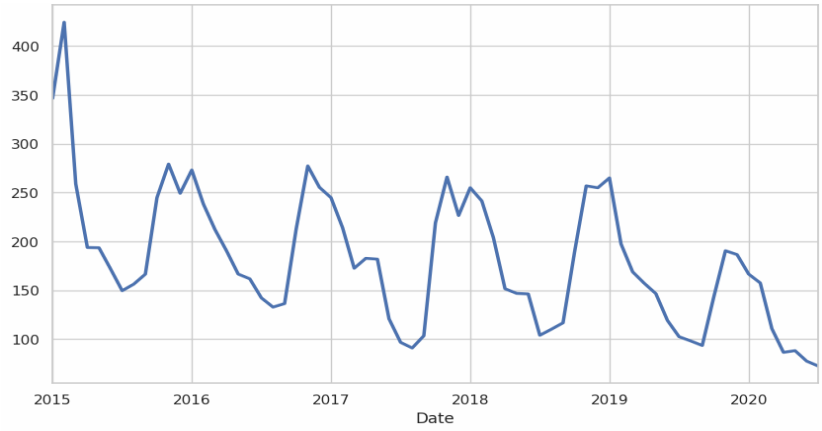
Methodology

City Date	Ahmedabad_AQI	Aizawl_AQI	Amaravati_AQI	Amritsar_AQI	Bengaluru_AQI	Bhopal_AQI	Brajrajnagar_AQI	Chandigarh_AQI	Chennai_AQI	Coimbatore_AQI	...
2015-01-01	350.333333	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
2015-02-01	520.640000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
2015-03-01	418.571429	NaN	NaN	NaN	130.545455	NaN	NaN	NaN	363.800000	NaN	...
2015-04-01	308.640000	NaN	NaN	NaN	113.733333	NaN	NaN	NaN	175.862069	NaN	...
2015-05-01	263.466667	NaN	NaN	NaN	102.774194	NaN	NaN	NaN	176.129032	NaN	...

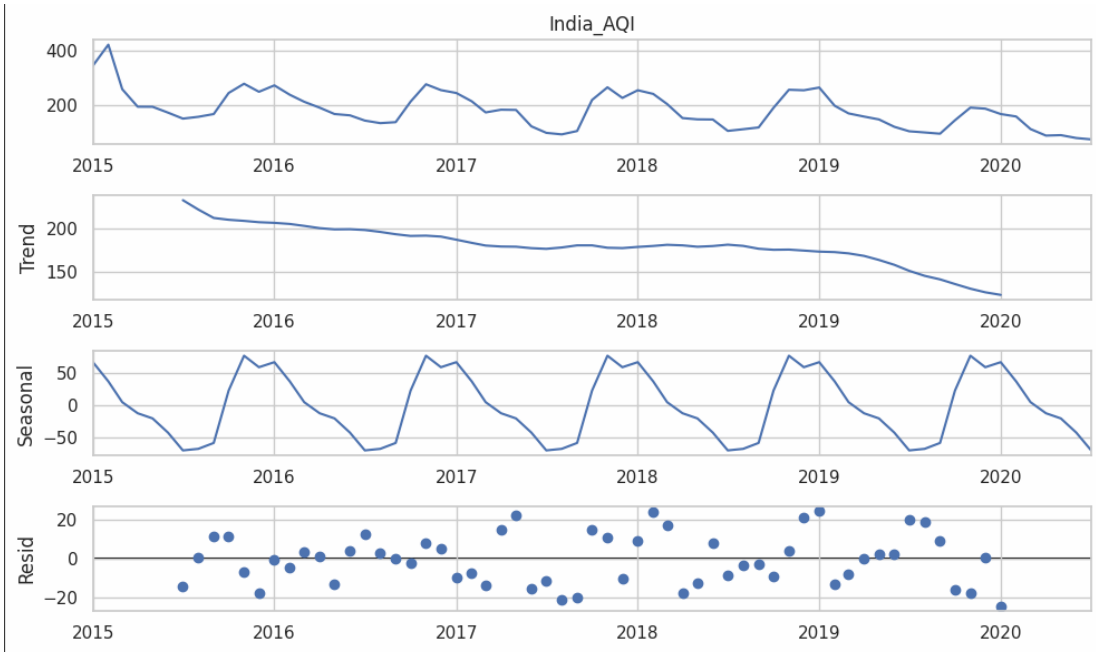
Taking just the AQI column for all the cities, so that we can have a comparative view of the value of every city's AQI through every day.

City Date	Ahmedabad_AQI	Aizawl_AQI	India_AQI
2015-01-01	350.333333	NaN	346.311828
2015-02-01	520.640000	NaN	424.284286
2015-03-01	418.571429	NaN	258.875688
2015-04-01	308.640000	NaN	193.815995
2015-05-01	263.466667	NaN	193.556272
...	...	...	...
2020-03-01	277.466667	65.350000	110.777963
2020-04-01	120.733333	39.233333	86.532043
2020-05-01	128.677419	24.193548	88.262751
2020-06-01	97.357143	20.862069	77.532108
2020-07-01	119.000000	20.000000	72.500000

Taking the AQI column for all the cities and creating a column India_AQL which takes the mean of all the cities for each month.



From the plot above, we can visually see that there is a slight downward trend and a seasonality present. However, we will decompose the plot into trend, seasonality and residuals to get a clearer picture. The trend component reveals long-term direction and changes, while seasonality highlights recurring patterns within specific time periods. Analyzing the residuals uncovers irregular fluctuations not accounted for by trend or seasonality. This decomposition aids in identifying factors impacting air quality, evaluating interventions, and making informed decisions.



We can see a clear seasonality and trend present here. The AQI decreases towards mid-year before rising again.

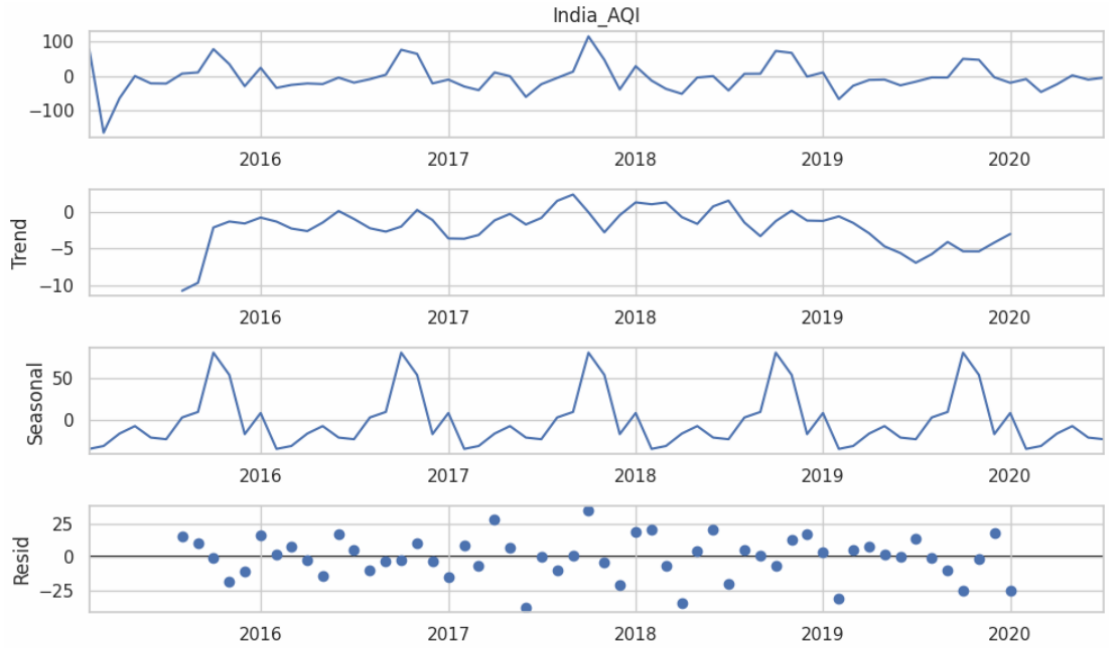
We ran an Augmented Dicky Fuller Test to check if data is stationary or not.

Test Statistic	-0.114224
p-value	0.948003
#Lags Used	10.000000
Number of Observations Used	56.000000
Critical Value (1%)	-3.552928
Critical Value (5%)	-2.914731
Critical Value (10%)	-2.595137

The p-value is 0.94, which means that this time series is not stationary.

The Augmented Dickey-Fuller (ADF) test is a statistical test commonly used to determine the stationarity of a time series.

We perform a first order differencing to remove the trend and then perform the ADF test again.

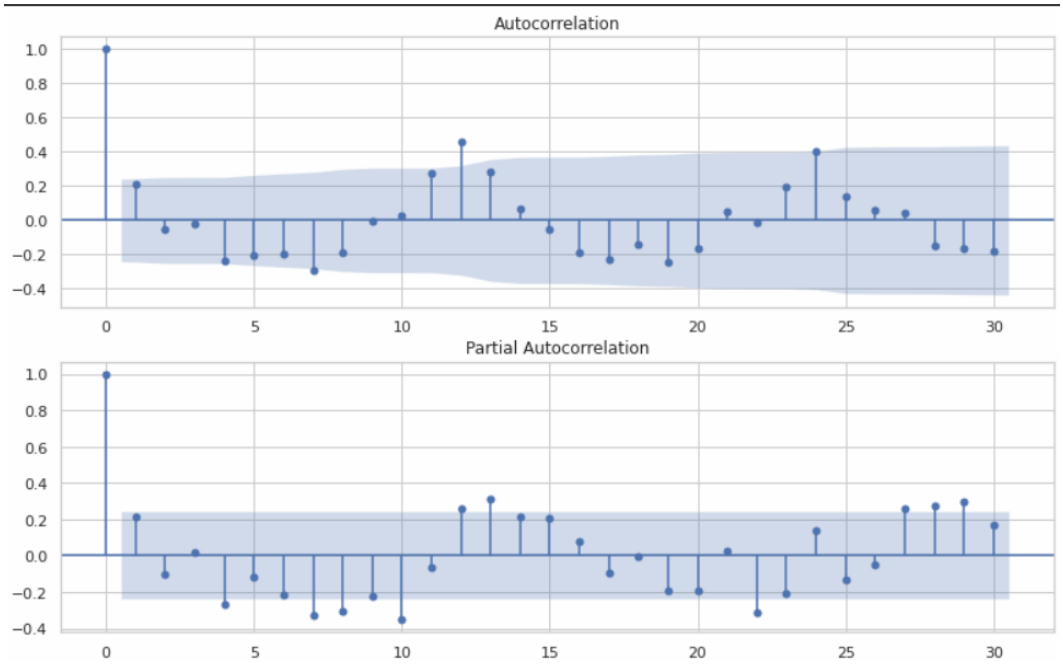


Test Statistic	-8.385232e+00
p-value	2.448599e-13
#Lags Used	9.000000e+00
Number of Observations Used	5.600000e+01
Critical Value (1%)	-3.552928e+00
Critical Value (5%)	-2.914731e+00
Critical Value (10%)	-2.595137e+00

Here, P-value from the Augmented Dickey-Fuller (ADF) test is 2.448599e-13 (a very small value), which is significantly less than the common significance level of 0.05. In this case, the extremely small p-value suggests strong evidence to reject the null hypothesis of non-stationarity. Thus, you can conclude that the time series is stationary.

From the p-value and the Test Statistics, we can conclude that with one differencing, the time series becomes stationary. Therefore, d=1.

After performing the ADF test, examining the ACF and PACF plots helps determine the appropriate orders of the ARIMA model's AR and MA components. These components capture the dependencies between the current observation and the previous observations, accounting for autocorrelation in the time series data, allowing for a more accurate modeling and forecasting of the time series.



We are using auto-arima to determine the parameters of the SARIMA model.

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	67
Model:	SARIMAX(0, 1, 2)x(1, 0, [1], 12)	**Log Likelihood**	-316.908
Date:	Tue, 16 May 2023	**AIC**	643.816
Time:	19:01:24	**BIC**	654.765
Sample:	01-01-2015 - 07-01-2020	**HQIC**	648.143

Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	0.0189	0.059	0.320	0.749	-0.097	0.135
ma.L2	-0.8363	0.069	-12.077	0.000	-0.972	-0.701
ar.S.L12	0.9444	0.062	15.221	0.000	0.823	1.066
ma.S.L12	-0.5623	0.229	-2.458	0.014	-1.011	-0.114
sigma2	694.3701	142.982	4.856	0.000	414.130	974.610

Ljung-Box (L1) (Q):	0.95	**Jarque-Bera (JB):**	2.99
Prob(Q):	0.33	**Prob(JB):**	0.22
Heteroskedasticity (H):	0.38	**Skew:**	-0.52
Prob(H) (two-sided):	0.03	**Kurtosis:**	2.99

We are using auto_arima to get values of p,d,q and P,D,Q,m for SARIMA.

SARIMAX Results

Dep. Variable:	India_AQI	**No. Observations:**	48
Model:	SARIMAX(0, 1, 2)x(1, 0, [1], 12)	**Log Likelihood**	-229.813
Date:	Sat, 16 Oct 2021	**AIC**	469.625
Time:	07:17:45	**BIC**	478.876
Sample:	01-01-2015 - 12-01-2018	**HQIC**	473.106

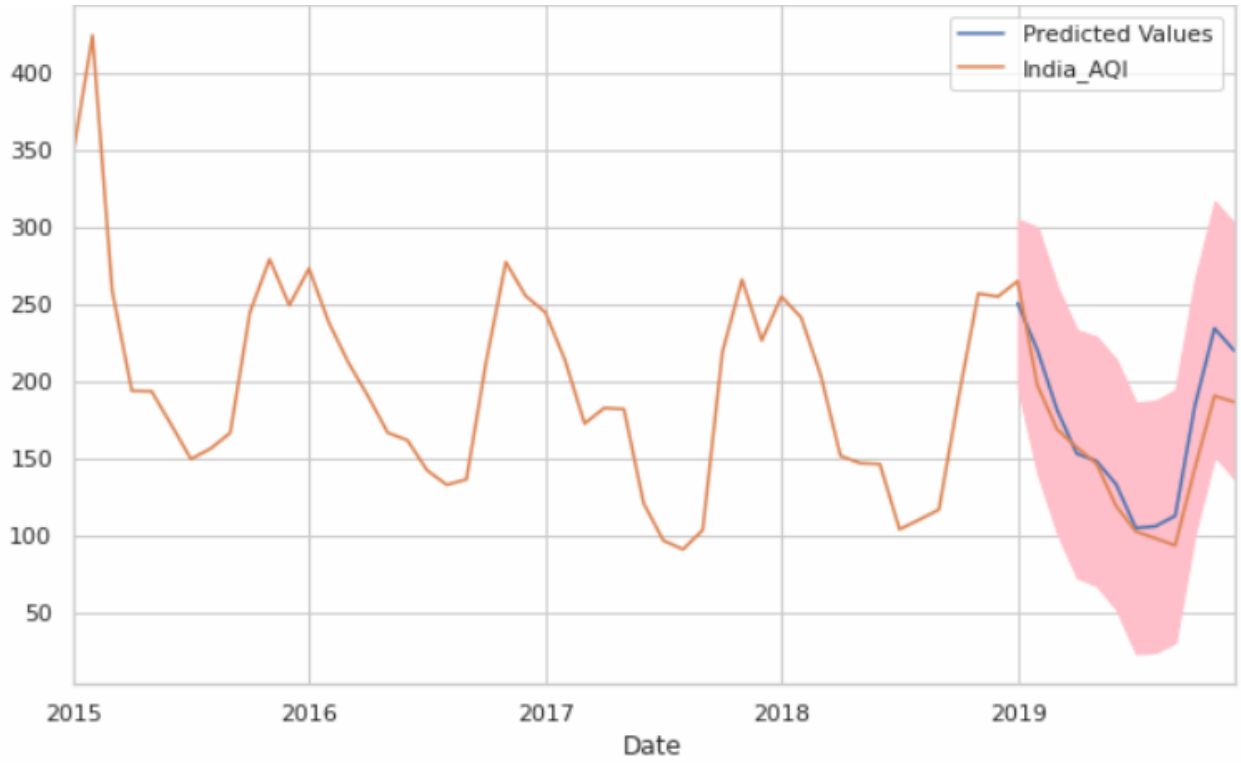
Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	0.0644	0.555	0.116	0.908	-1.024	1.152
ma.L2	-0.9323	0.547	-1.703	0.089	-2.005	0.141
ar.S.L12	0.9183	0.097	9.440	0.000	0.728	1.109
ma.S.L12	-0.4473	0.301	-1.485	0.138	-1.038	0.143
sigma2	767.1435	413.405	1.856	0.064	-43.116	1577.403

Ljung-Box (L1) (Q):	0.23	**Jarque-Bera (JB):**	3.61
Prob(Q):	0.63	**Prob(JB):**	0.16
Heteroskedasticity (H):	0.24	**Skew:**	-0.66
Prob(H) (two-sided):	0.01	**Kurtosis:**	3.28

Making use of p,d,q and P,D,Q,m in SARIMA which we found with the help of auto_arima and creating a model.

Prediction of the next 12 months values



Checking how accurate our model is on the known data.

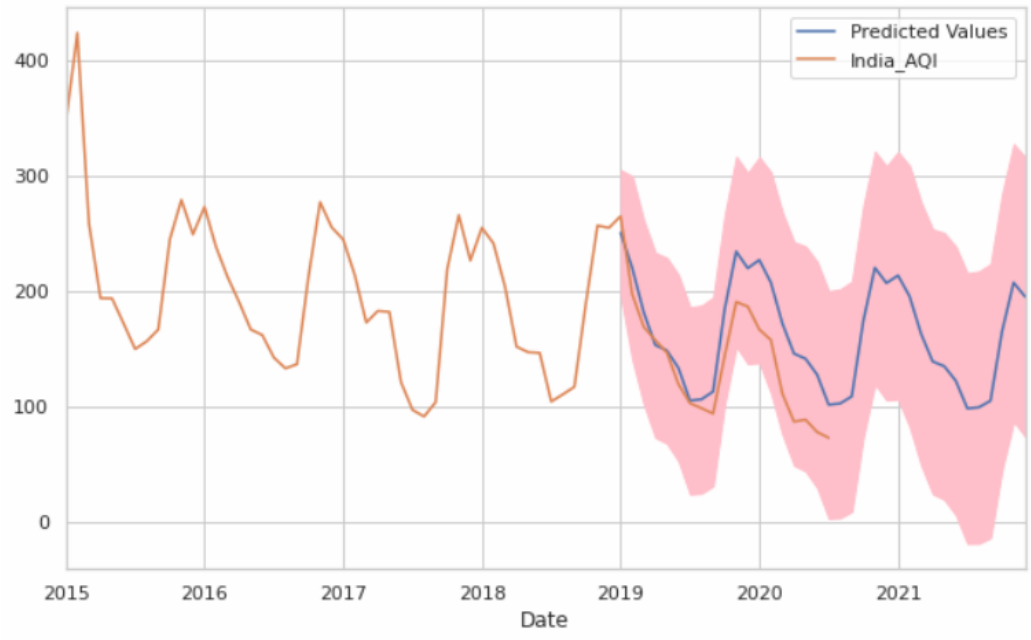
```
RMSE =  22.7551159490353
MAPE =  11.640057775135464
```

We see that the model has an RMSE of 22.75 on the test data set. Now, we can use this model to predict values into the future.

We'll be forecasting AQI values for 2021. However, 2020 yielded unexpected AQI values owing to the lockdown imposed due to COVID-19, as we saw earlier. So our prediction might have a wider margin of error that needs to be considered. Factors to consider include the impact of COVID-19, shifts in trends, data availability and quality, model adjustments, and communicating uncertainty.

Predicting AQI of India for 2021 with code as reference

```
fig, ax= plt.subplots(figsize=(10,6))

forecasts = results.get_forecast(steps=36, dynamic=True)

confidence_intervals=forecasts.conf_int()
lower_limits = confidence_intervals.loc[:,'lower India_AQI']
upper_limits = confidence_intervals.loc[:,'upper India_AQI']

#plot the forecasted data
forecasts.predicted_mean.plot(legend=True, ax=ax, label ='Predicted Values')

#plot the confidence interval as the shaded area
plt.fill_between(confidence_intervals.index, lower_limits, upper_limits, color='pink')

#Plot India's AQI Data
cities['India_AQI'].plot(legend=True, ax=ax)
```



Conclusion

History has shown us that reducing air pollution is a key component to protecting public health. During the industrial revolution, although there were many economic benefits, there was a severe impact on air quality from the increase in pollutants. Currently, a lot of the factors contributing to air pollution in our cities are due to sectors such as energy, urban planning, transport and agriculture. The findings from the data analysis should urge policy makers to implement tighter controls in India and for the global community to come together to innovate cleaner solutions.

In this study, our objective was to forecast AQI values for the upcoming year and gain insights into the factors influencing air quality. We employed a time-series modeling approach using ARIMA, considering historical AQI data and relevant environmental variables.

There are few models we thought we could use but linear regression is not a good way and SARIMA seemed to be the one that suited our model better.

While our analysis provides valuable insights, there are limitations to consider. Data availability and quality, as well as assumptions made during the modeling process, can introduce uncertainties in the predictions. Additionally, the forecasted AQI values are subject to potential changes in human activities and emission sources, which may impact the actual air quality levels.

Since it’s a time series data it's gonna follow a particular pattern which is evident from our predictions for the year 2021 but will slowly be reducing, which is a good sign keeping in mind that there were lots of missing values.

In conclusion, this study's AQI predictions offer valuable insights into air quality dynamics, supporting decision-making processes, and promoting public health. By understanding the factors influencing air quality and accurately forecasting AQI, we can strive towards effective pollution control measures and a healthier environment for all.