Comprehensive Report

on

"Data Analytics and Data Visualisation with Tableau"

(Summer Course)

Submitted by:

Name: N R Samhitha Rao

SRN: PES2UG19CS243

Semester: 7

Faculty Handling the Course:

Dr.Sudeepa Roy Dey
Dr.Prajwala TR
Prof.Ruby Dinakar

4th - 9th July 2022

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
PES UNIVERSITY, EC CAMPUS
(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, Ind

Table of Contents

Dataset chosen and software requirements:

The data for the purpose of this case study has been made publicly available by the Central Pollution Control Board: https://cpcb.nic.in/ which is the official portal of Government of India.

Sheet1 | 16 fields 18007 rows | 100 rows

Table Details

F1	City	Date	Pm2.5	NO	NO2	N Ox	CO	SO2	O3	AQI	AQI Bucket
1	ahmedabad	01/01/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
2	ahmedabad	01/02/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
3	ahmedabad	01/03/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
4	ahmedabad	01/04/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
5	ahmedabad	01/05/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
6	ahmedabad	01/06/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
7	ahmedabad	01/07/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
8	ahmedabad	01/08/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
9	ahmedabad	01/09/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
10	ahmedabad	01/10/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate
11	ahmedabad	01/11/2017	67.8545	22.4434	59.0546	47.3883	22.2086	55.2739	39.0842	166	moderate

Snippet of the dataset used

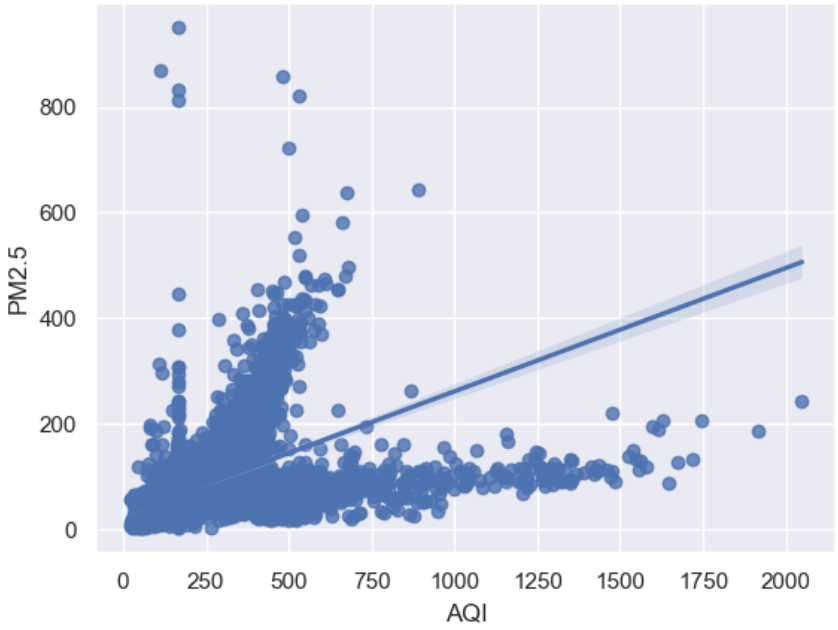
Software requirements and tools used:

Language used: Python 3.8

Libraries: pandas, numpy, scipy, sklearn

Tool for visualisation: Tableau Public 2020.3

Identification of the nature of the Dataset:



Scatter plot between PM2.5 and AQI which shows linearity

```
Samhithas-MacBook-Pro:~ samhitharao$ /usr/bin/python3 "/Users/samhitharao/Library/Mobile Documents/com~apple~CloudDocs/Summer co
urse/Monotonicity.py"
Monotonicity of AQI is: False
spearman coefficient

          PM2.5        NO       NO2       NOx        CO       SO2        O3       AQI
PM2.5  1.000000  0.478061  0.402351  0.507552  0.340585  0.358562  0.222891  0.820688
NO     0.478061  1.000000  0.538265  0.723697  0.334509  0.336543 -0.013247  0.499022
NO2    0.402351  0.538265  1.000000  0.666819  0.375037  0.299302  0.263643  0.458044
NOx    0.507552  0.723697  0.666819  1.000000  0.426863  0.331741  0.035814  0.522418
CO     0.340585  0.334509  0.375037  0.426863  1.000000  0.241733  0.117150  0.440773
SO2    0.358562  0.336543  0.299302  0.331741  0.241733  1.000000  0.253954  0.457944
O3     0.222891 -0.013247  0.263643  0.035814  0.117150  0.253954  1.000000  0.266221
AQI    0.820688  0.499022  0.458044  0.522418  0.440773  0.457944  0.266221  1.000000
```

Monotonicity and Spearman coefficient for all parameters

The given dataset is a non monotonic dataset since the values drastically decrease in the year 2020.

The correlation coefficients are all mostly positive. The purpose of this case study is to check the relation of AQI with respect to various parameters which are all positive, which means that there's a direct relation of AQI with all the pollution parameters.

The given dataset cannot be tested with Pearson's coefficient due to the given reasons:

- The dataset has outliers
- The distribution is not normal.

Hence Spearman's coefficient has been used for analysis. However it is observed that the correlation

coefficient values are not very high due to the dataset being an aggregate of values which have considerably dipped in the year 2020 and have hence resulted in a value less than expected.

Testing accuracy of Machine Learning models used for predicting values:

The learning model used for this dataset is Linear Regression which is suitable for continuous data and the model's accuracy has been determined as well. Furthermore, statistical parameters such as MSE, MAE and RMSE have been calculated to confirm the accuracy of the model in predicting AQI values.

```
Samhithas-MacBook-Pro:~ samhitharao$ /usr/bin/python3 "/Users/samhitharao/Library/Mobile Documents/com~apple~CloudDocs/Summer co
urse/training models.py"
/Users/samhitharao/Library/Python/3.8/lib/python/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following vari
ables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments w
ithout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
Accuracy of your prediction is 94.5512106541337%
MAE: 9.432192049074084
MSE: 140.51395862515042
RMSE: 11.853858385570094
Samhithas-MacBook-Pro:~ samhitharao$
```

Using linear regression to predict AQI values and calculating statistical values

MAE: Mean Average Error- Since the values of AQI on an average, are in the range of 100's, a MAE value of 9.43 is a good range for deviation for the predicted value.

MSE: Mean Square Error- The value is quite high due to the high range of values in AQI across various cities.

RMSE: Root Mean Square Error- The value is also relatively high compared to the ideal value of 0, due to the large difference between the highest and lowest values.

Creating a hypothesis test:

A hypothesis has been proposed for the purpose of this case study which states:
"H0 statement: The AQI has reduced by nearly more than 35% during lockdown"

The purpose of this test is to validate the observations made graphically and prove the assumed hypothesis. This is also an interesting insight to the case study which is conducted.

To do so, first the desired Dataset was segregated to reflect values only during lockdown.

```
Samhithas-MacBook-Pro:~ samhitharao$ /usr/bin/python3 "/Users/samhitharao/Library/Mobile Documents/com~apple~CloudDocs/Summer co
urse/hypothesis.py"
H0 statement: The AQI has reduced by nearly more than 35% during lockdown
H1 statement: The AQI has not reduced more than 35% during lockdown
65% of population mean:  109.38878769367469
Alpha:  0.05
Sample size (total observations in 2020):  1559
hypothetical z value:  16.25592243665579

Failed to reject hypothesis
H0 : μ >= 109.38878769367469
H1 : μ < 109.38878769367469
P value was found to be(in %) 100.0
Samhithas-MacBook-Pro:~ samhitharao$
```

Results of the hypothesis test

```
a_pop_mean = data['AQI'].mean()
pop_mean = 0.65*data['AQI'].mean()
```

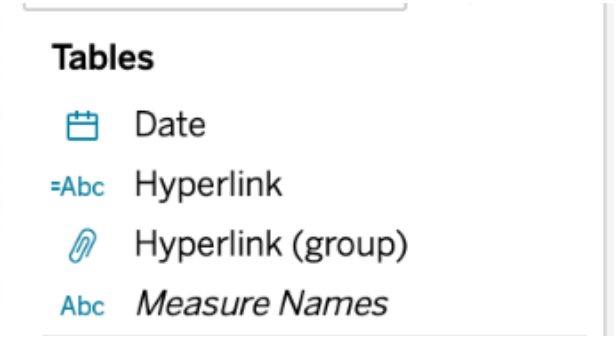
Calculating 65% of mean

The degree of statistical significance generally varies depending on the level of significance.

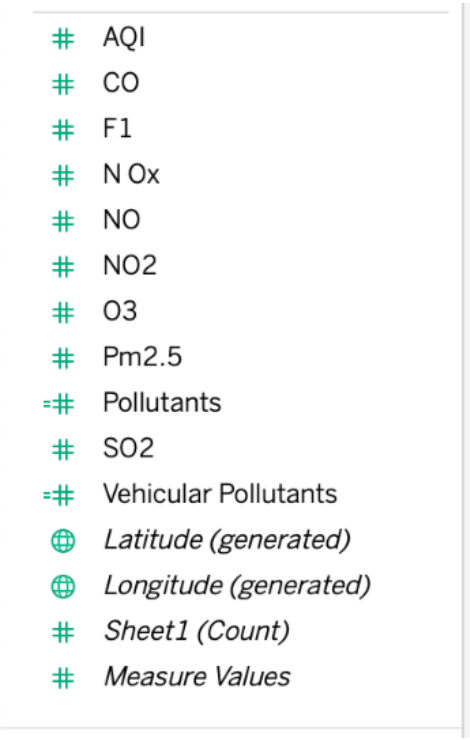
As P value is 98.90% which means that the probability of the assumption being correct was 98.9%, it can be safe to accept the hypothesis or at least it is plausible that our statement is correct with accuracy of 98.9%.

Graphical Analysis:

Measures and Dimensions present in the dataset:



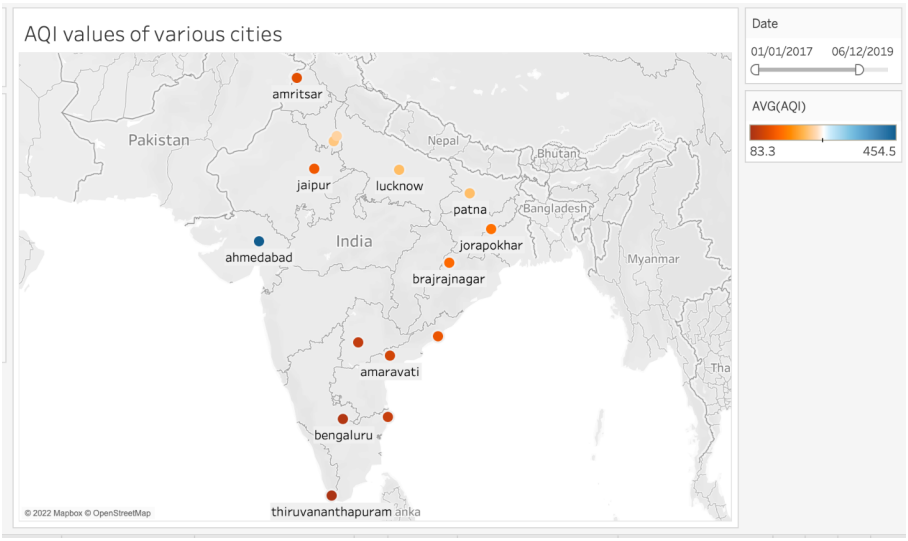
Measures of the dataset



Dimensions of the dataset

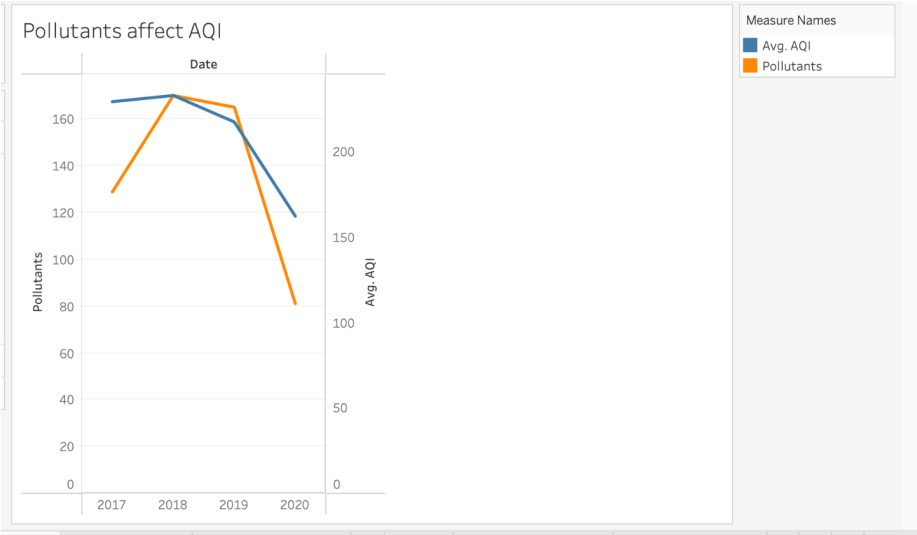
No appropriate hierarchies could be created for this dataset as it was not applicable and the number of dimensions were inadequate.

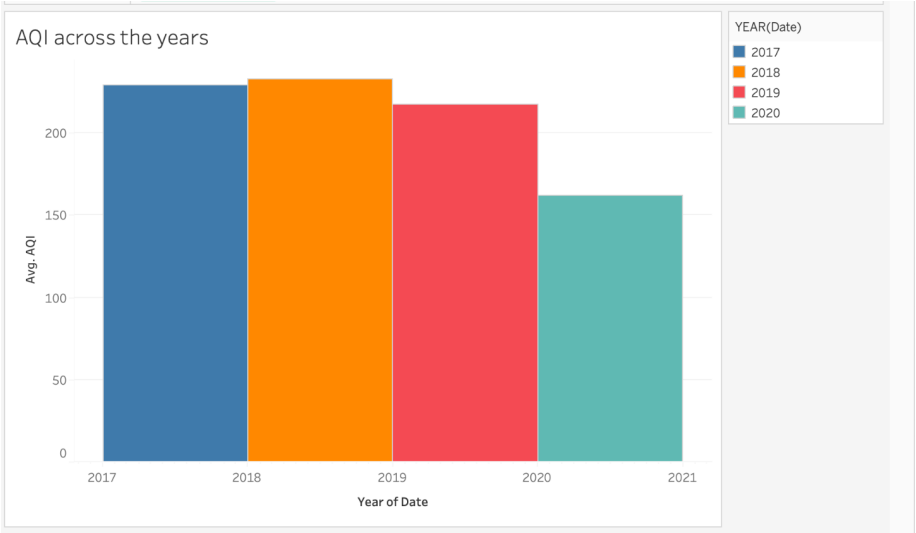
Charts with Insights:



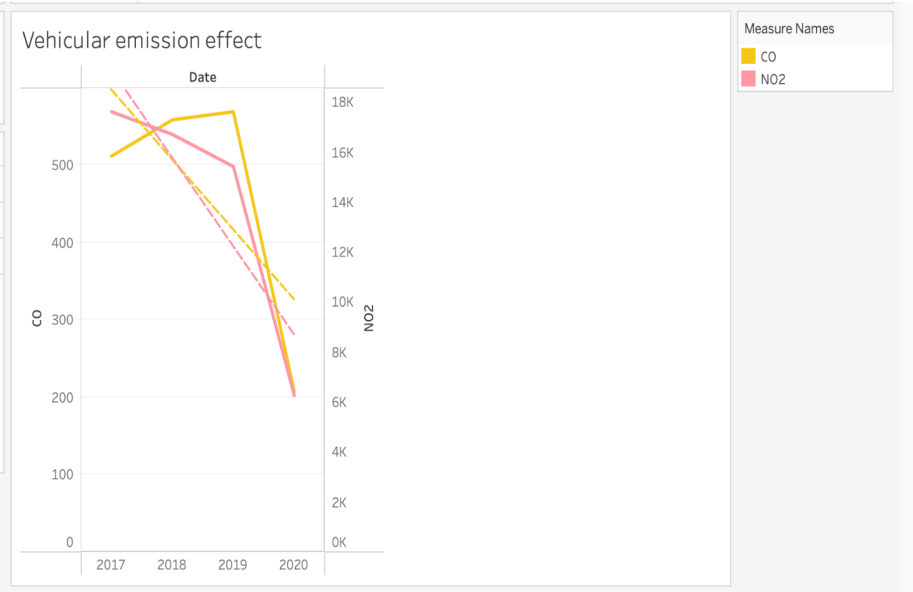
The adjecent chart shows the AQI values of various cities in India pre-covid, and it can be observed that AQI was not very good with the least value at 83.3, whereas the max approved level is 100.

This chart provides an insight as to how pollutants are really affecting AQI and the two axes nearly show similar patterns which indicate that the 2 are closely related and hence the values taken are realistic and reliable.





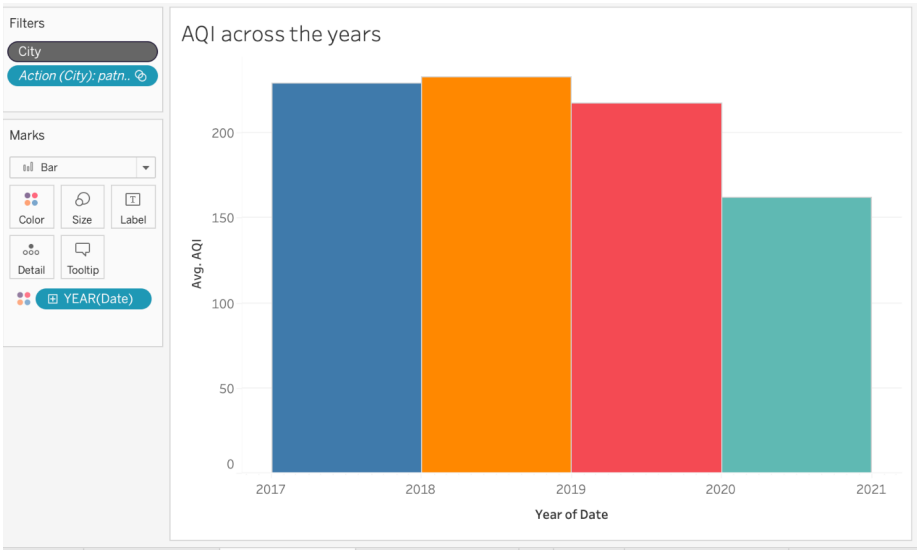
This chart shows the pollution rate across all the years and clearly 2020 has a steep decrease in pollution.



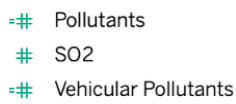
This graph shows the effect of vehicular pollutants throughout the years and a very steep decline indicated by the trend line shows that lack of vehicles during lockdown may have impacted the decline.

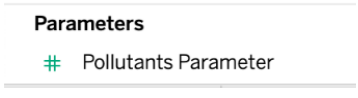
Context filter addition:

Context filter on the *City* dimension has been added in this graph.



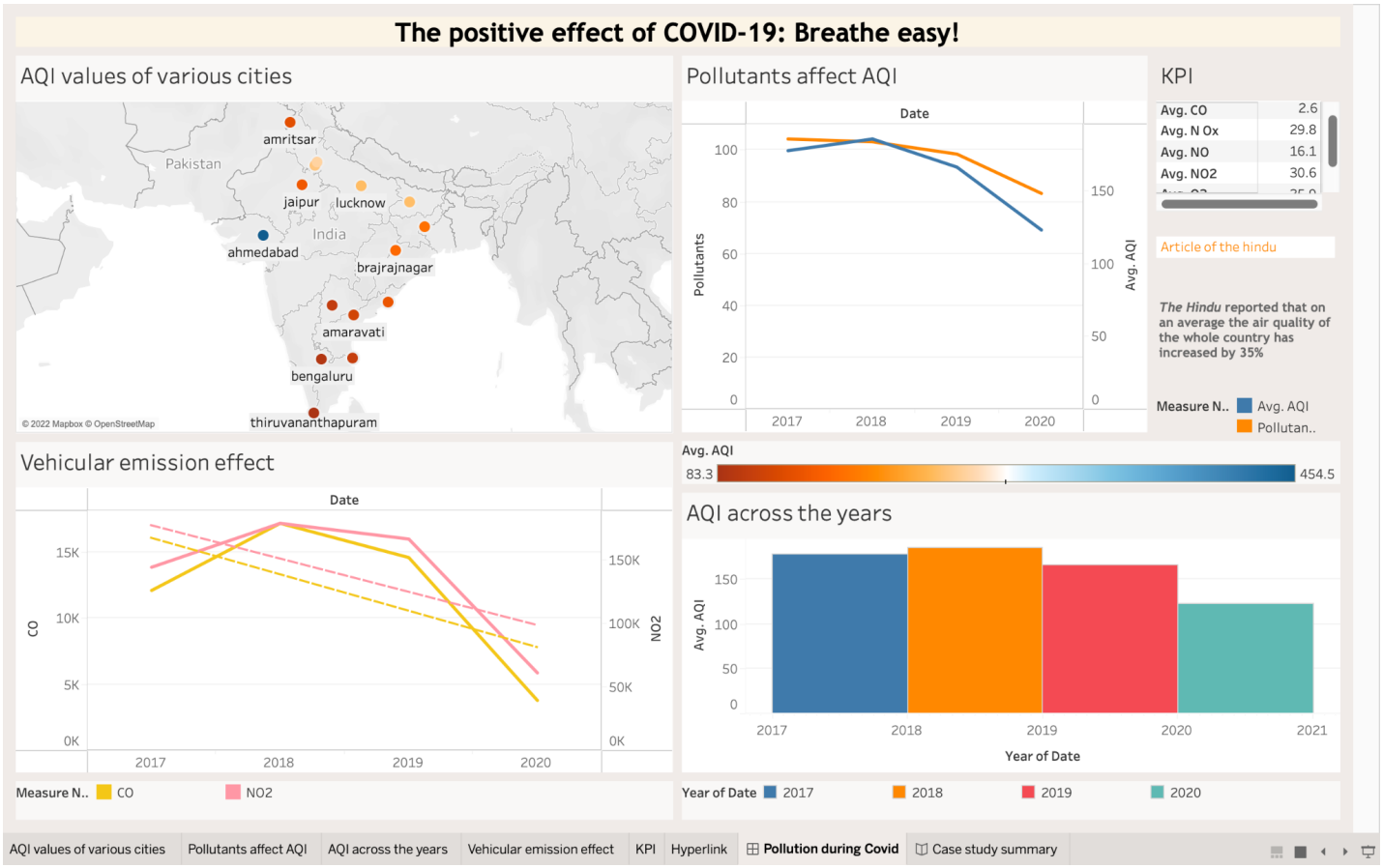
Calculated field and parameters included:





Pollutants and Vehicular pollutants were the calculated fields created.

Final Dashboard:



Conclusion:

The case study focused on the effect of COVID-19 on the pollution in India. Initially it was found that the pollution levels measured using AQI, were very high for all cities selected in the dataset. During the onset of the lockdown, the values began to reduce and soon the AQI value turned out to be really low during 2020, with nearly 35% reduction in the overall AQI of all cities in the dataset.

Throughout the report multiple examples to display the effect of lockdown on air quality in India have been demonstrated. The lessons learnt from the COVID-19 pandemic can be utilised to target source specific actions leading to maximum improvement in ambient air quality. In conclusion, results of the present study indicate a sharp decline in overall air quality indexes and in the concentration of primary air pollutants. However, such a lockdown cannot be planned in normal conditions in such a huge country and therefore it cannot be considered as a permanent solution. Nevertheless, taking lessons from the present lockdown and emerging and foreseen environmental scenarios, similar prospective measures that are feasible can adopted in future to mitigate air pollution.