India Air Quality Index - Data Analysis

Rishi Dey Chowdhury

7/9/2021

INTRODUCTION

In this Data Analysis Project, I am going to work with **Air Quality Index (AQI) Data of India**. I will be using several Statistical Tools to Analyze the Data which includes Exploratory Data Analysis, Techniques and methodologies used for Inference and Modelling.

MOTIVE

- To understand the relationship that holds between different parameters which we measure as a part of AQI Data of India.
- To find out if COVID-19 Nation-wise Lockdowns, Social Distancing and Closure of Industries, Factories and Suspension of movement via private vehicles and public transport had a significant impact on India's AQI.

UNDERSTANDING THE DATA

Let's take a look at the Data

- This Dataset contains 263890 rows and 10 columns.
- The Year ranges from 2014 to 2021 (till June), with observations recorded on each of the 30 /31 days of the month for 12 months for the last 3 years.
- The Data is generated from the 22 cities from various Stations located near that Cities. The Cities include:

- The parameters which we measure at the different Stations are given under the Specie Column and it includes -

Table 1: First few rows of the Air Quality Index Data

Year	Month	Day	City	Specie	count	min	max	median	variance
2014	12	29	Delhi	pm25	24	296.0	460.0	394.0	27226.40
2014	12	29	Hyderabad	pm25	13	159.0	162.0	161.0	8.59
2014	12	29	Delhi	pm10	82	79.0	999.0	218.0	634717.00
2014	12	29	Delhi	o3	79	0.1	87.4	3.2	2324.38
2014	12	29	Delhi	so2	91	0.3	21.2	4.2	231.83
2014	12	29	Delhi	pm25	83	139.0	747.0	307.0	215149.00

Table 2: Last few rows of the Air Quality Index Data

Year	Month	Day	City	Specie	count	min	max	median	variance
2021	6	24	Kolkata	o3	48	2.9	105.7	8.4	4611.99
2021	6	24	Kolkata	pm25	48	45.0	104.0	63.0	1398.61
2021	6	24	Kolkata	pressure	56	996.9	1007.5	999.3	67.94
2021	6	24	Kolkata	wind-speed	56	0.1	4.2	1.1	10.87
2021	6	24	Kolkata	dew	37	28.0	28.0	28.0	0.00
2021	6	24	Kolkata	co	48	1.0	5.2	2.3	16.41

Table 3: City Stations

State	City	Number of Stations
Andhra_Pradesh	Visakhapatnam	1
Arunachal_Pradesh	Visakhapatnam	1
Bihar	Patna	6
Chandigarh	Chandigarh	1
Delhi	Delhi	40
Kerala	Thiruvananthapuram	2
Kerala	Thrissur	1
MadhyaPradesh	Bhopal	1
Maharashtra	Mumbai	21
Maharashtra	Nagpur	1
Maharashtra	Nashik	1
Meghalaya	Shillong	1
Rajasthan	Jaipur	3
Tamil_Nadu	Chennai	8
Telangana	Hyderabad	6
Uttar_Pradesh	Lucknow	6
Uttar_Pradesh	Muzaffarnagar	1
West_Bengal	Kolkata	7

Table 4: Specie Description

Parameters	Description	Units
pm25	Particle pollution/particulate matter(particles less than or equal to 2.5 micrometers in diameter)	microgr
pm10	Particle pollution/particulate matter(particles less than or equal to 10 micrometers in diameter)	microgr
o3	Ground-level ozone	microgr
so2	Sulphur dioxide	microgr
no2	Nitrogen dioxide	microgr
co	Carbon Monoxide	miligra
temperature	Temperature	Celcius
pressure	Air Pressure	Torr
wind-gust	Wind Gust/Force	kmph
humidity	Relative Humidity	No Uni
wind-speed	Wind Speed	kmph
dew	Dew Point	Celcius
precipitation	Precipitation	milimet

Table 5: Significance of the AQI Values

AQI Values	Level of Health Concern
0-50	Good
51-100	Moderate
101-150	Unhealthy for sensitive group
151-200	Unhealthy
201-300	Very Unhealthy
301-500	Hazardous

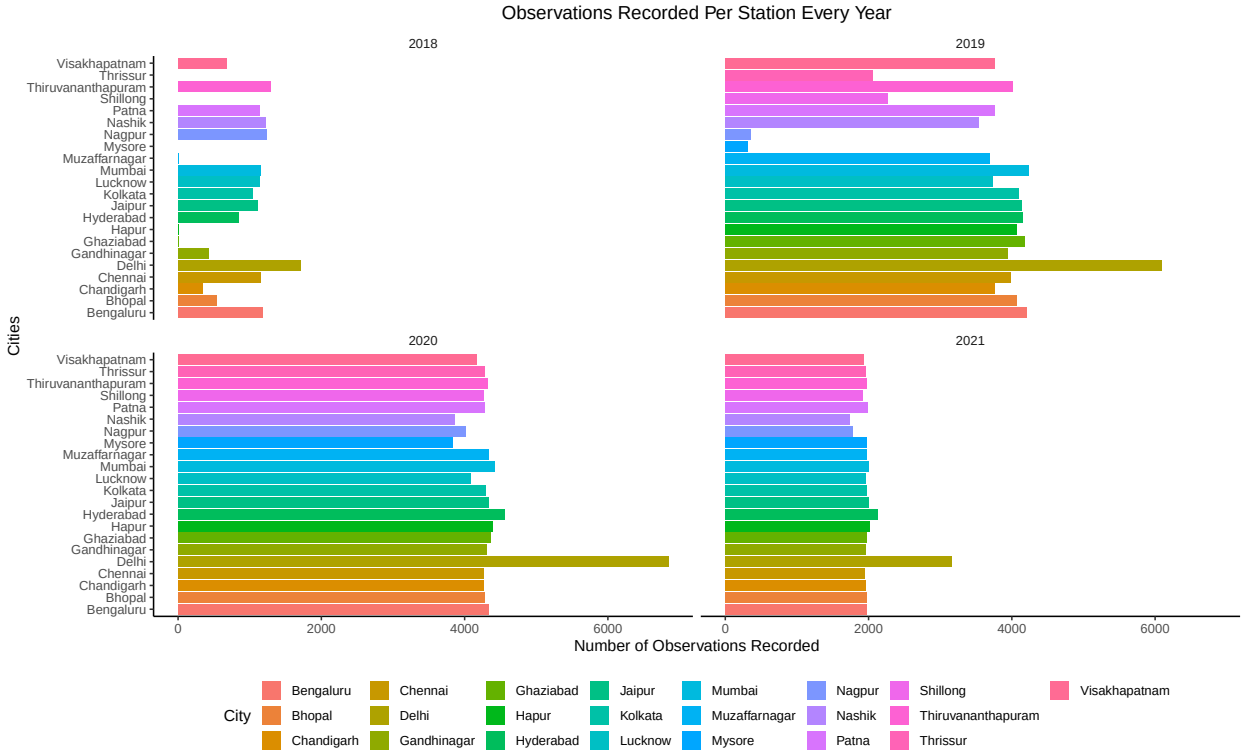
AQI is a comparable and communicable way of measuring the parameters in the Air. It is calculated when atleast 3 of the top 6 parameter's data is available of which one must be pm10 or pm25. It is the max of the parameters recorded given they satisfy the above condition.

- It also helps in identifying faulty standards and inadequate monitoring programmes.
- AQI helps in analysing the change in air quality (improvement or degradation).
- Comparing air quality conditions at different locations/cities.
- It can be easily interpreted by anyone, without knowing about background details.

In further Analysis we will refer *pm25, pm10, o3, so2, no2 and co2 as pollutants* and the *remaining weather parameters as non-pollutants.*

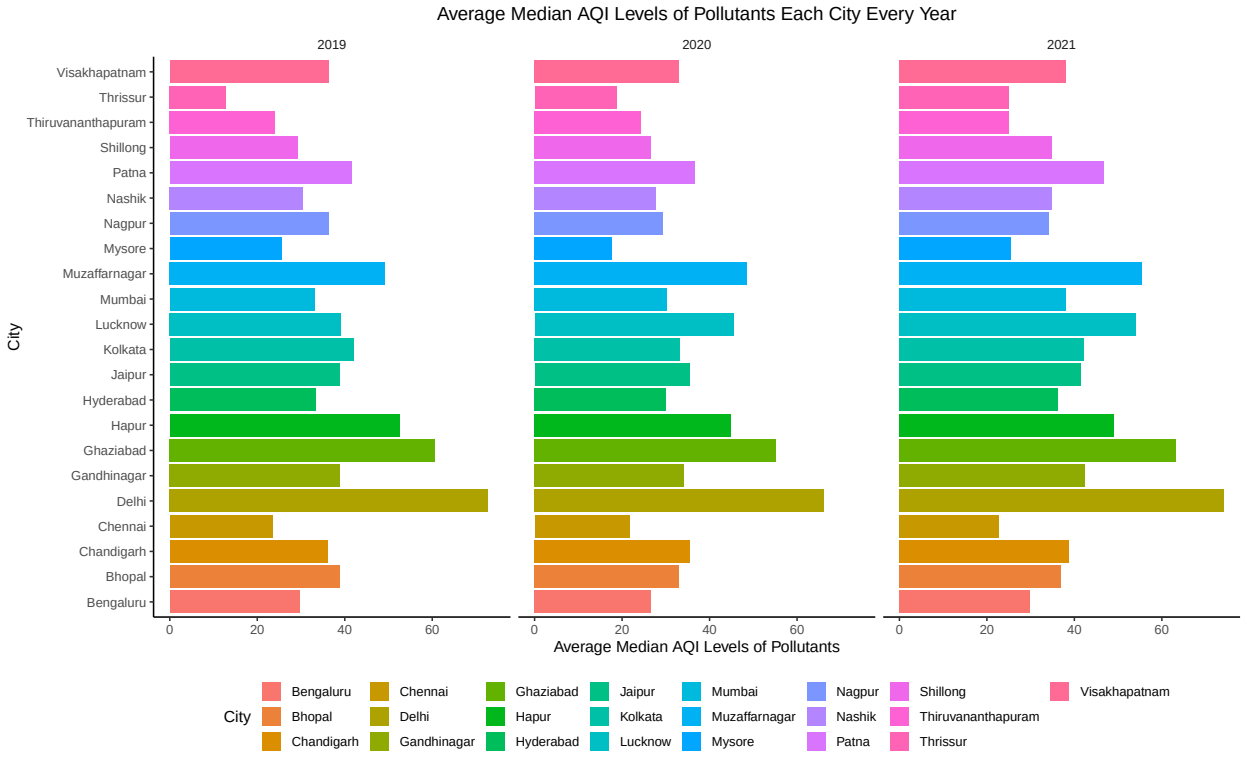
TOP CITY STATIONS

Here we look at *which city records how many observations per year.*



It seems **Delhi is the most monitored cities among the others**. We see all the other stations have almost equal number of observations per year. Our data seems to have lot of missing values for the year 2018 and the years before that; Hence we will only consider the data of the year 2019, 2020 and 2021.

Speculating on the reason *why Delhi is so heavily monitored* we look at *how the Average Median AQI Levels of pollutants at each city every year*.

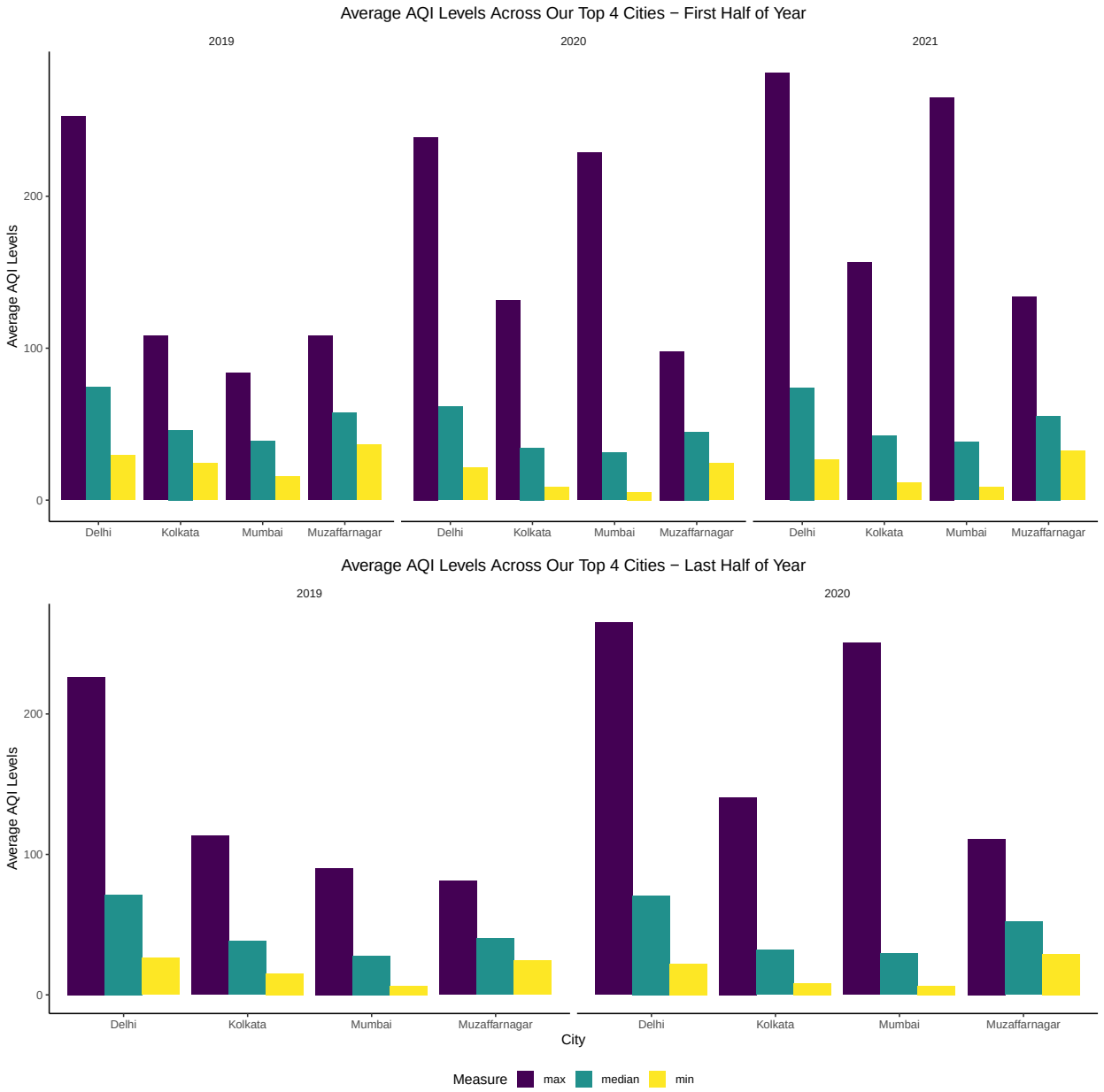


Indeed it is clearly visible that **Delhi's AQI Levels of Pollutants is higher** than all other cities. This makes Delhi our city of main focus in all the further analysis. Every city owing to it's location at different parts of the country, different development status, different population, different local weather conditions has different measured values of AQI. So, it makes sense to look at them separately. So, whenever we will look at City-wise Analysis, we will look at these Cities as our **Top 4 Cities - Delhi, Muzaffarnagar, Kolkata, Mumbai.**

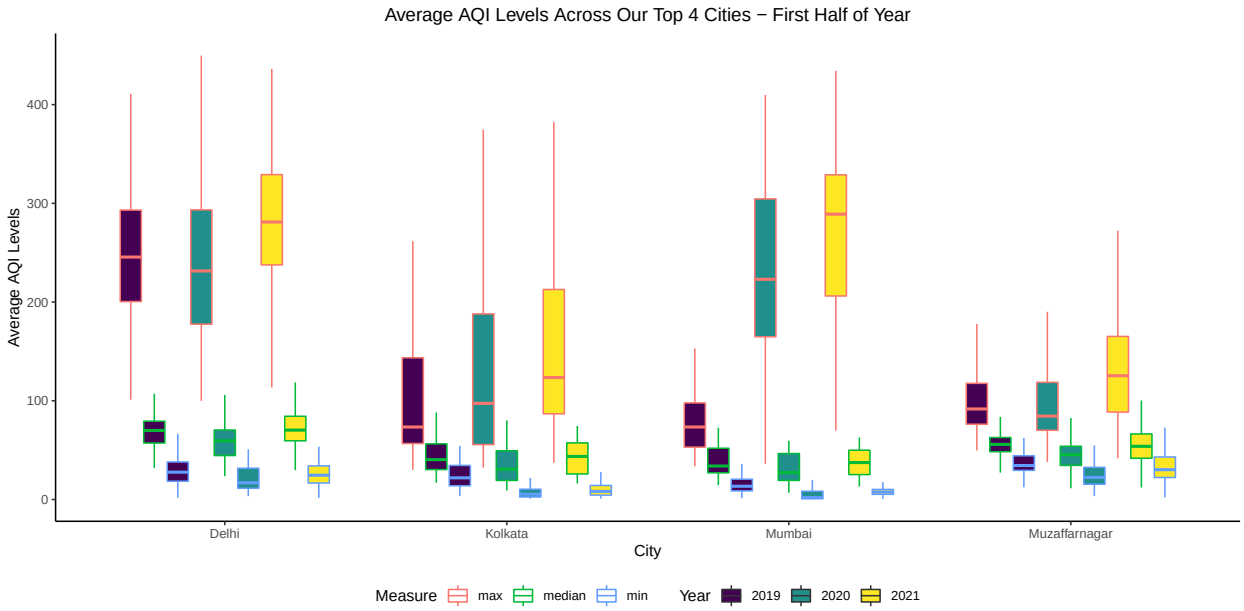
VISUAL OVERVIEW

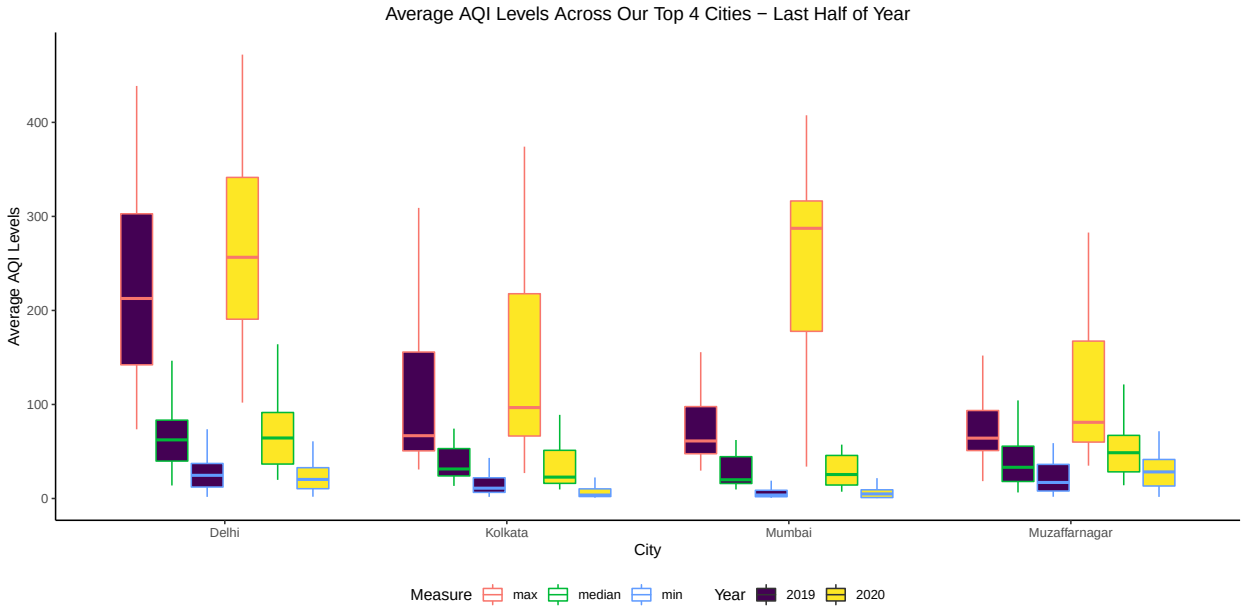
Here we will take a visual tour of the Data through Exploratory Data Analysis.

- Let's see the AQI Levels in the Top 4 across all 3 measures - Max, Median and Min splitting the year into two halves.



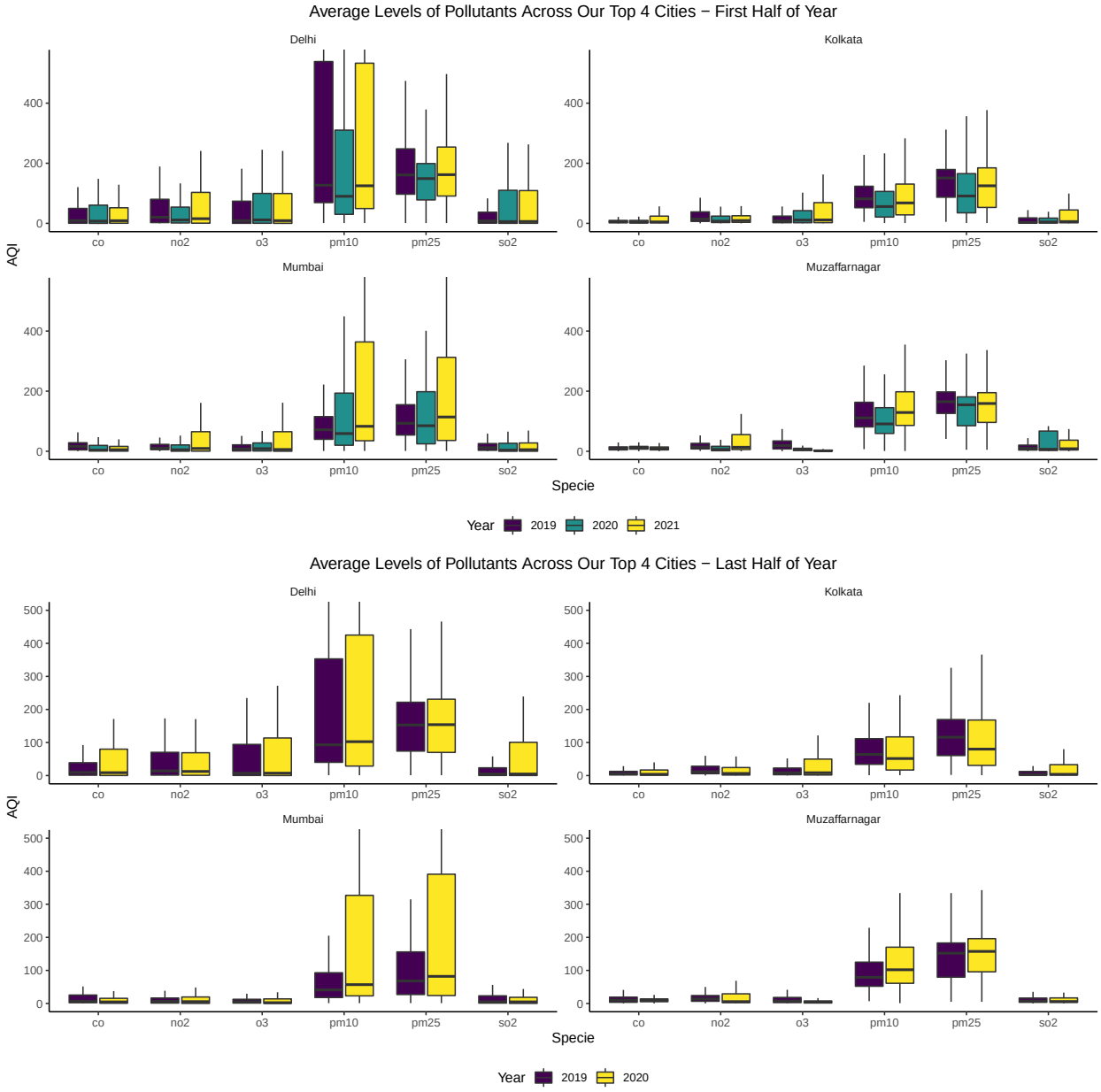
The first bit of observation is the **Median & Min AQI Levels of all 4 cities have dropped in 2020 and 2021 is similar to 2019**. But the **Max AQI Levels have increased in 2020 and even more in 2021** even though we had so many restrictions in 2020!





In the first half of the year 2020 Min and Median AQI Levels have dropped whereas in the second half of the year it either increased or remained same as that of 2019. 2021's observed values are pretty much the same as that of 2019 if not more. Max AQI Levels kept rising across the years irrespective of the part of the years. **The spread(Variance) of the resp. measures look the same across the years**.

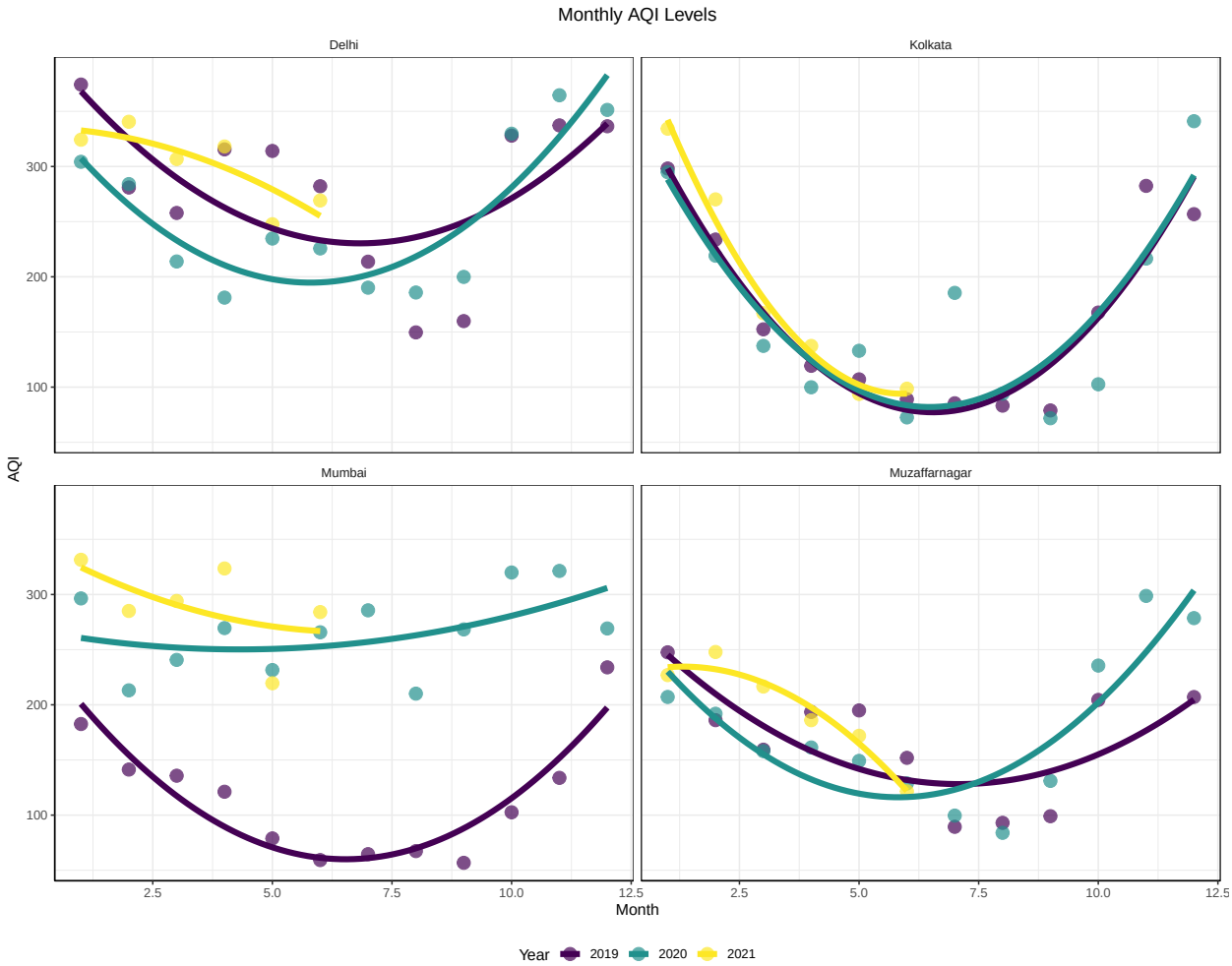
- Let's take a look at the pollutant-wise Levels in our Top 4 Cities.



Here we see an interesting effect of Lockdowns. **All the pollutants have decreased considerably in the First half of 2020 whereas in the Second half it has bumped up again to match the Levels of 2019 or even more! While the Levels of 2021 & 2019 are very similar**.

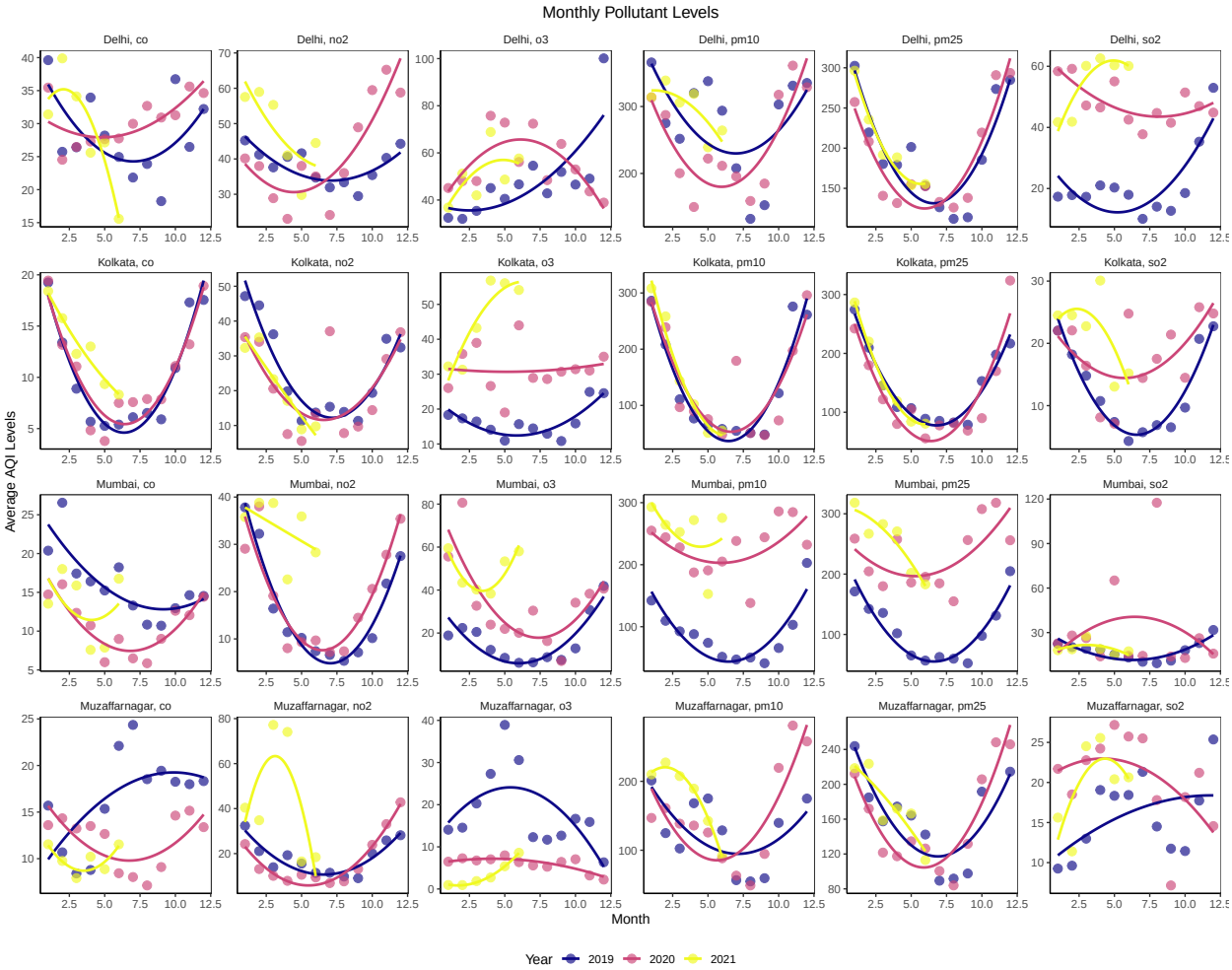
RELATIONSHIPS - PATTERNS & TRENDS

- Pollutant Gases and Particulate Matter, whose levels despite being affected by Human Intervention and Industrial Activities, are nothing but Natural Gases which follow nature's process. Here we will take a look at *how the AQI Levels change with seasons throughout the year.*



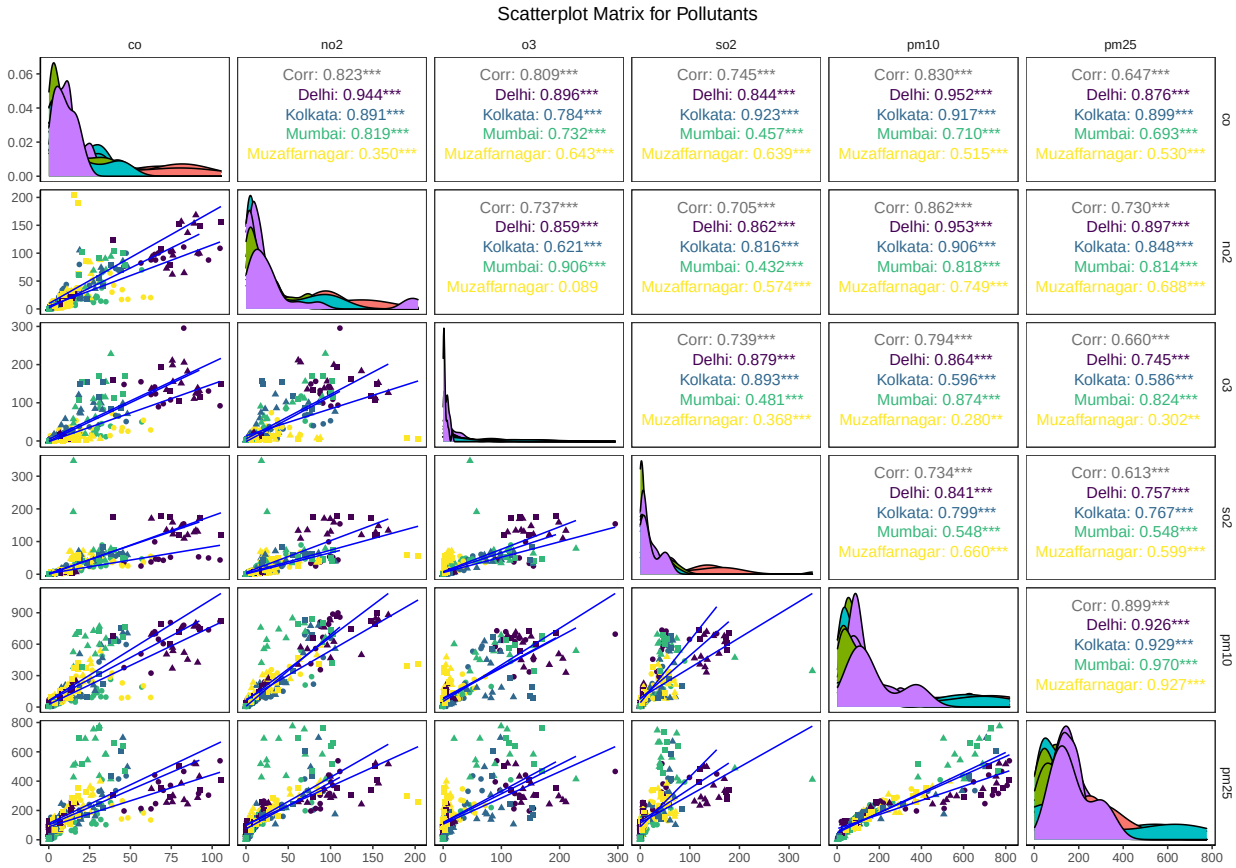
Indeed it looks like the **AQI Levels have dropped in the first half of 2020 compared to 2019, with the exception of Mumbai. But it has increased to levels more than 2019 in the later half. The AQI Levels of 2021 is more that both the previous years**. I have fit here a Quadratic Model which seems to model the situation well though we shouldn't focus on the model fit to 2021 as we only have the data pertaining to first 6 months.

Now, we will see the above observed AQI Levels in terms of Levels of each Pollutants.

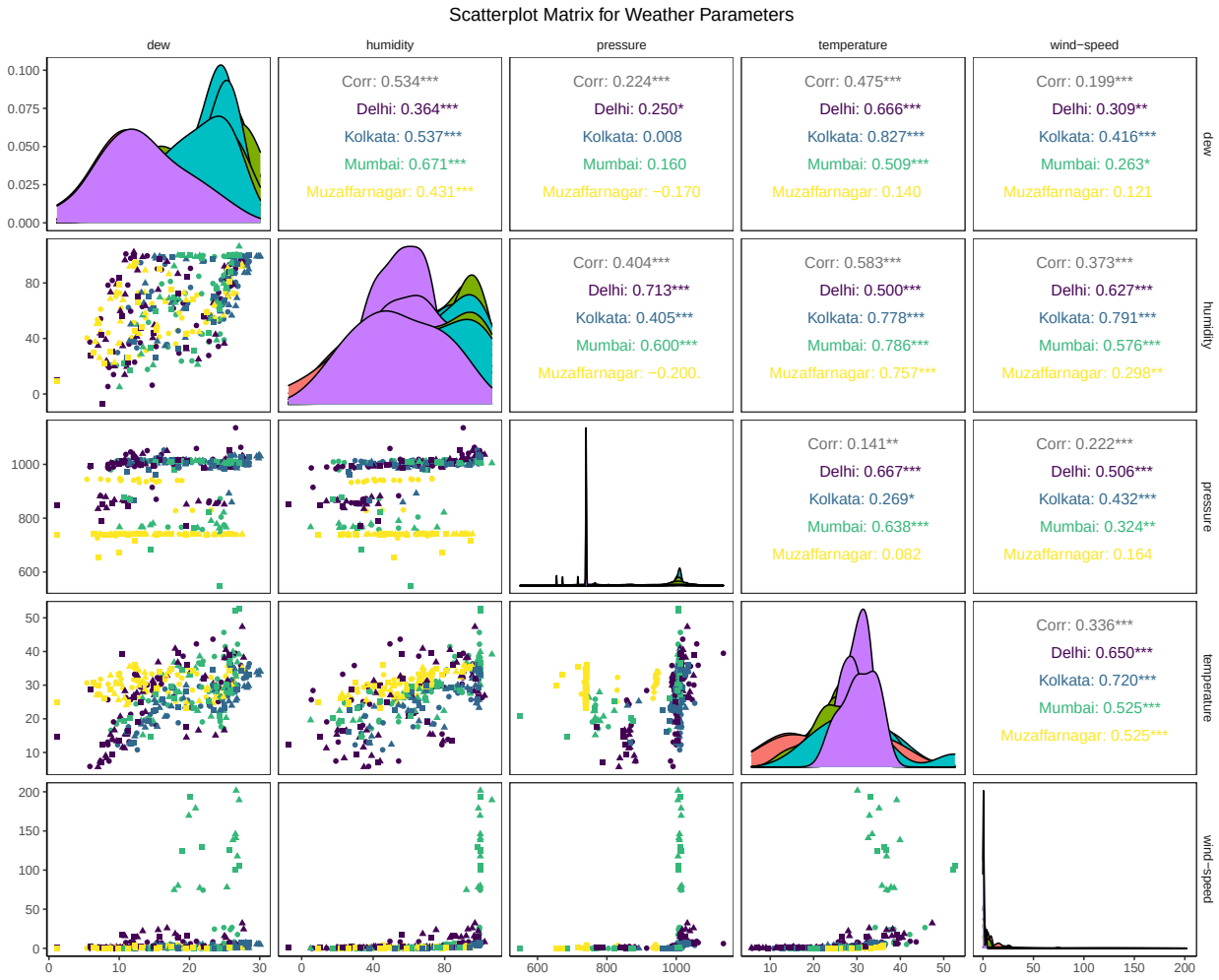


Except for Ground-Level Ozone and Sulphur Dioxide most of the Pollutant Levels have decreased during the first half of 2020 which saw an increase in the second half when compared to 2019.This increase can be attributed to increased Stubble Burning that took place in the later half of the Year. The surprising thing is that even **particulate matter increased to much higher levels in Mumbai** even though it saw a massive lockdown and transportation as well as industrial work were suspended.

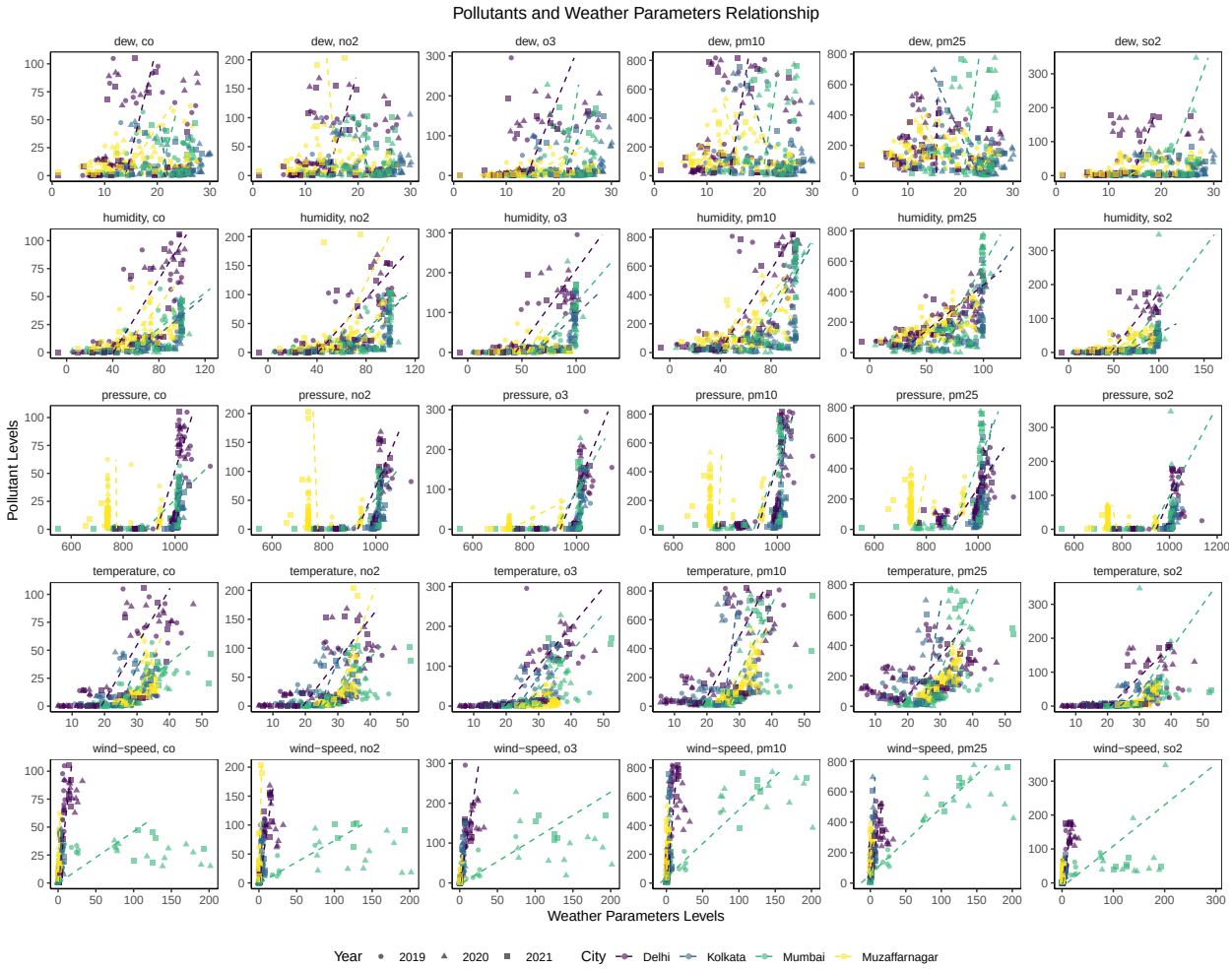
- The Pollutant Levels are affected to a great deal by the weather conditions of a place. Here we try to look at *how the Pollutant Levels are related to other Pollutant Levels* and *how the weather parameters are related among themselves* followed by a dependency of *Pollutant Levels of weather parameters*.



We see a **strong linear relationship between the pollutants** among themselves across the past 3 years at our Top Cities.

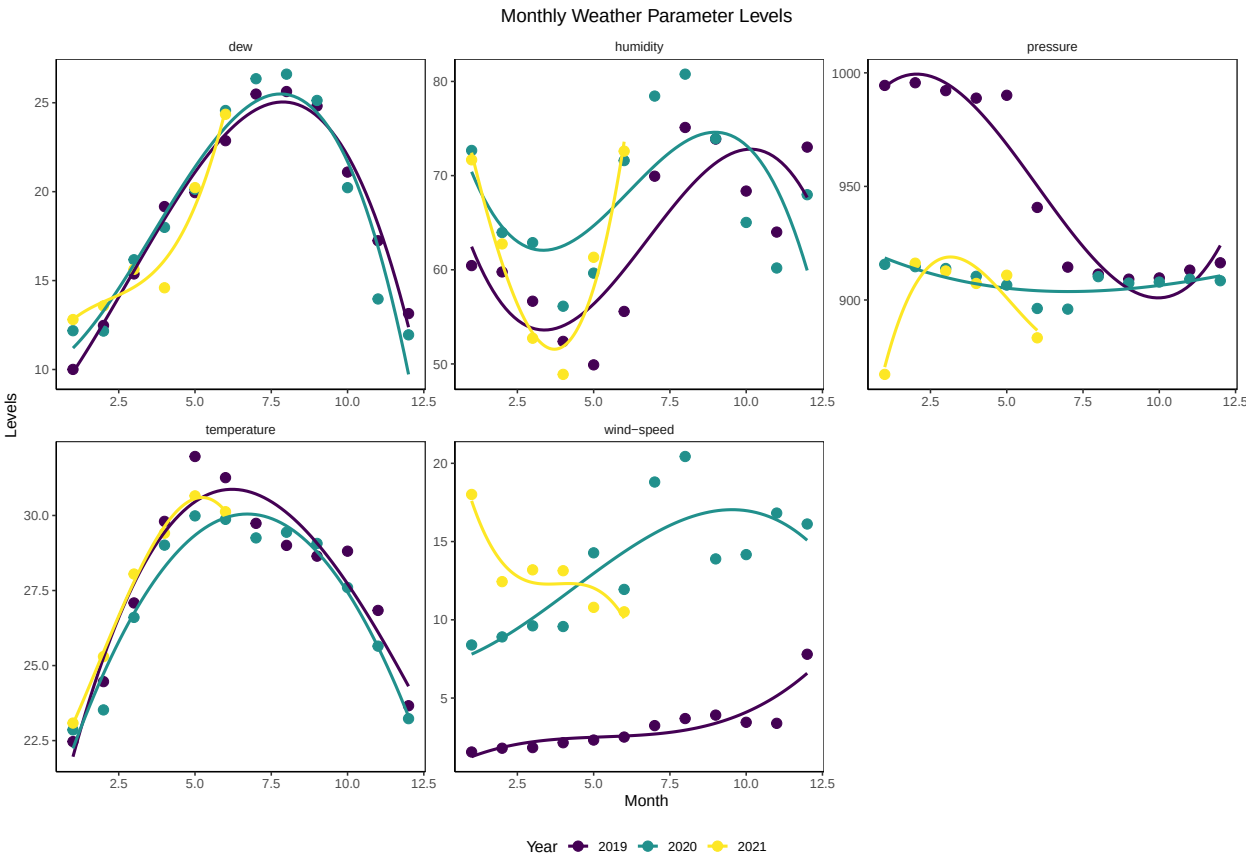


We see a **strong linear relationship between the weather parameters** across the past 3 years. One important thing that gives us a hint towards *why the pm10 and pm25 levels were higher in Mumbai*, as we see **Mumbai has experienced really high wind-speeds in 2020 and 2021. The pm content was thus increased due to movement of lots of dust by wind**.



This suggests that there is a **positive linear relationship between Temperature & Humidity and all the Pollutant Levels. Wind speed is indeed positively correlated with pm10 and pm25 which might be the cause for higher AQI in Mumbai**.

- Another Important thing to keep in mind is *how the weather parameters change with seasons.*



It suggests that the **wind-speeds and humidity were higher in 2020, whereas pressure and temperatures were lower in 2020**.

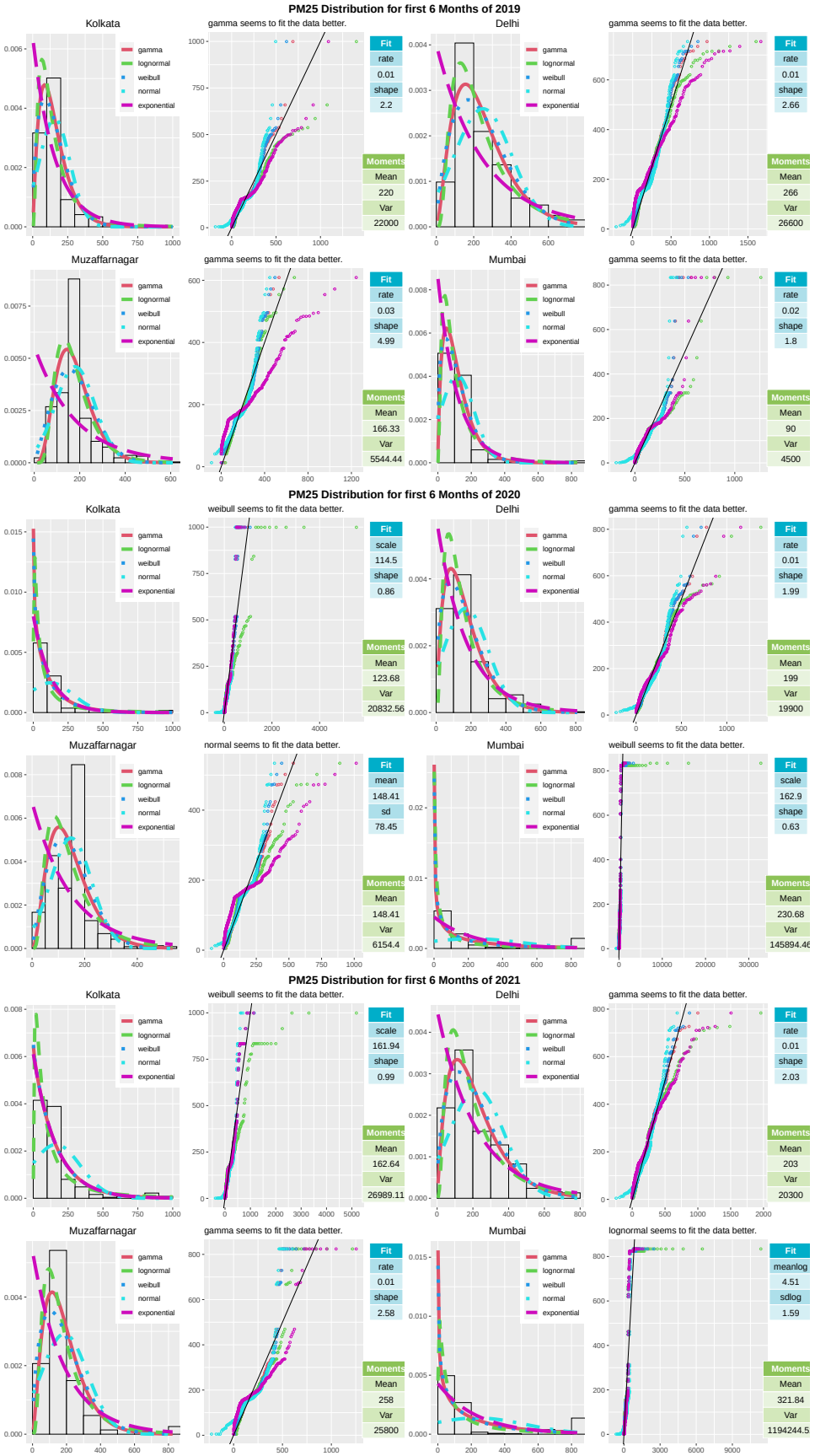
STATISTICAL INFERENCE

ESTIMATION

DISTRIBUTIONS & PROBABILITY MODELS

To perform any kind of analysis we need to model the data we have in order to get an idea about the expected value, variance and other properties related to the data. It also helps us get rid of the noise we get from the data we collect.

- Here we will look at *how the different pollutants are distributed and how different are the models for different year & location and which probability model fits our data and what are the parameter estimates of the fitted model along with Mean and Variance.*

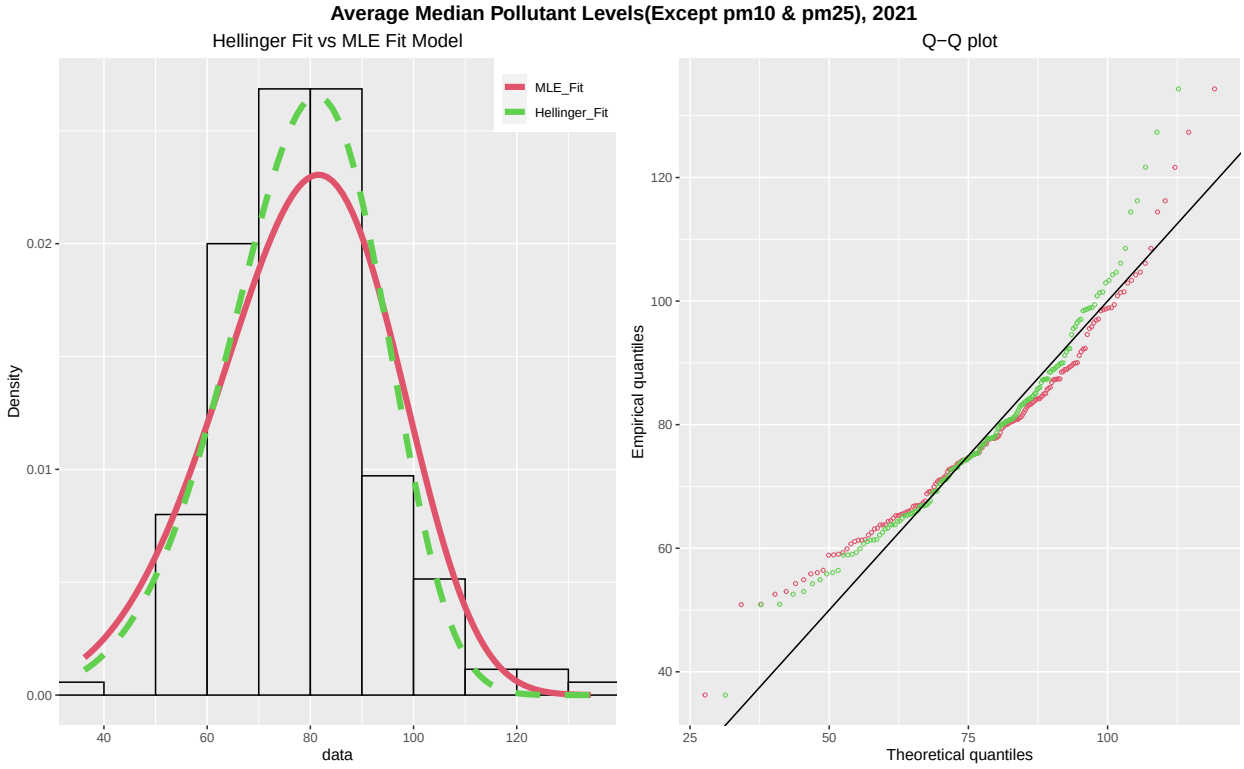

PM25 Distribution for first 6 Months of 2019
Kolkata
gamma seems to fit the data better.
Fit
rate
0.01
shape
2.2
Moments
Mean
220
Var
22000
Delhi
gamma seems to fit the data better.
Fit
rate
0.01
shape
2.66
Moments
Mean
266
Var
26600
Muzaffarnagar
gamma seems to fit the data better.
Fit
rate
0.03
shape
4.99
Moments
Mean
166.33
Var
5544.44
Mumbai
gamma seems to fit the data better.
Fit
rate
0.02
shape
1.8
Moments
Mean
90
Var
4500
PM25 Distribution for first 6 Months of 2020
Kolkata
weibull seems to fit the data better.
Fit
scale
114.5
shape
0.86
Moments
Mean
123.68
Var
20832.56
Delhi
gamma seems to fit the data better.
Fit
rate
0.01
shape
1.99
Moments
Mean
199
Var
19900
Muzaffarnagar
normal seems to fit the data better.
Fit
mean
148.41
sd
78.45
Moments
Mean
148.41
Var
6154.4
Mumbai
weibull seems to fit the data better.
Fit
scale
162.9
shape
0.63
Moments
Mean
230.68
Var
145894.46
PM25 Distribution for first 6 Months of 2021
Kolkata
weibull seems to fit the data better.
Fit
scale
161.94
shape
0.99
Moments
Mean
162.64
Var
26989.11
Delhi
gamma seems to fit the data better.
Fit
rate
0.01
shape
2.03
Moments
Mean
203
Var
20300
Muzaffarnagar
gamma seems to fit the data better.
Fit
rate
0.01
shape
2.58
Moments
Mean
258
Var
25800
Mumbai
lognormal seems to fit the data better.
Fit
meanlog
4.51
sdlog
1.59
Moments
Mean
321.84
Var
1194244.5
gamma
lognormal
weibull
normal
exponential

I have mostly tried **fitting one of Gamma, Lognormal, Normal, Weibull and Exponential Distribution by the Method of Maximum Likelihood Estimation** after looking at the histograms of the data for all the Pollutants. The **better model out of all these 5 models was chosen based on Chi-Squre test for goodness of fit**.

Looking at the moments of the fitted distribution we see one of the most important factor contributing to AQI Calculation i.e. **pm25 Levels during the first 6 months were much lower in all the cities except Mumbai in 2020 compared to 2019, whereas these levels increased back to a much higher value in 2021**.

Similar **Fitted Probability Models suggests almost all the Pollutant Levels were lower in 2020 compared to that of 2019, but the values of 2019 and 2021 are very similar, if not that of 2021 is higher**. I will attach all the relevant Plots and figures for other Pollutants at the end, if interested in looking at them.

There were also some cases were **outliers degraded the fit of the model, I tried minimizing the Hellinger Distance to fit the model in that case**. Below give is one such case -

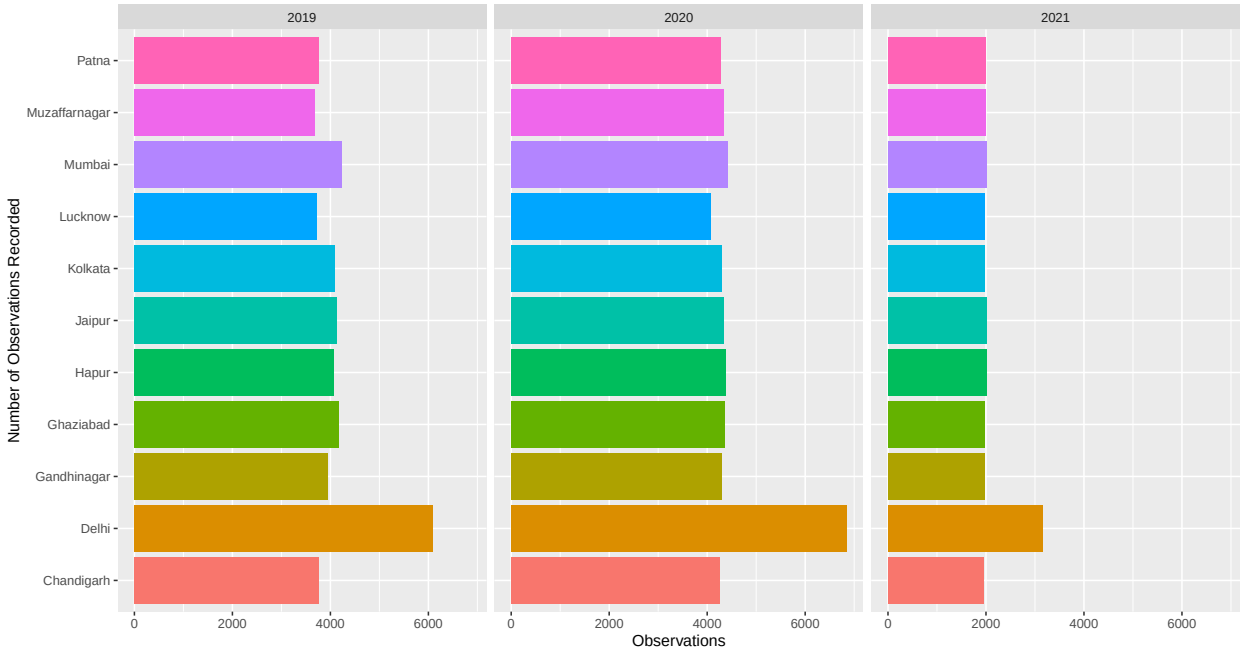


It is clear that the **Hellinger fit is much better in fitting the part of the distribution where there are maximum observations**.

HYPOTHESIS TESTING

Here we would try to Test some of out believes about the data, which will help us get a better understanding of the Pollutant and AQI Levels.

- In the Introduction part we saw how the number of observations recorded by different cities looked similar except for Delhi which was obviously very high and need not be tested. We would like to *test our null Hypothesis of the number of observations recorded by all cities are equal against all other possible alternatives.* We perform a **Chi-square Equality of proportion Test**



	Test Statistic	df	p−value
1	98.52099	9	0

Chi−sq Test for Equality of Proportion

	Test Statistic	df	p−value
1	17.61247	9	0.03995

Chi−sq Test for Equality of Proportion

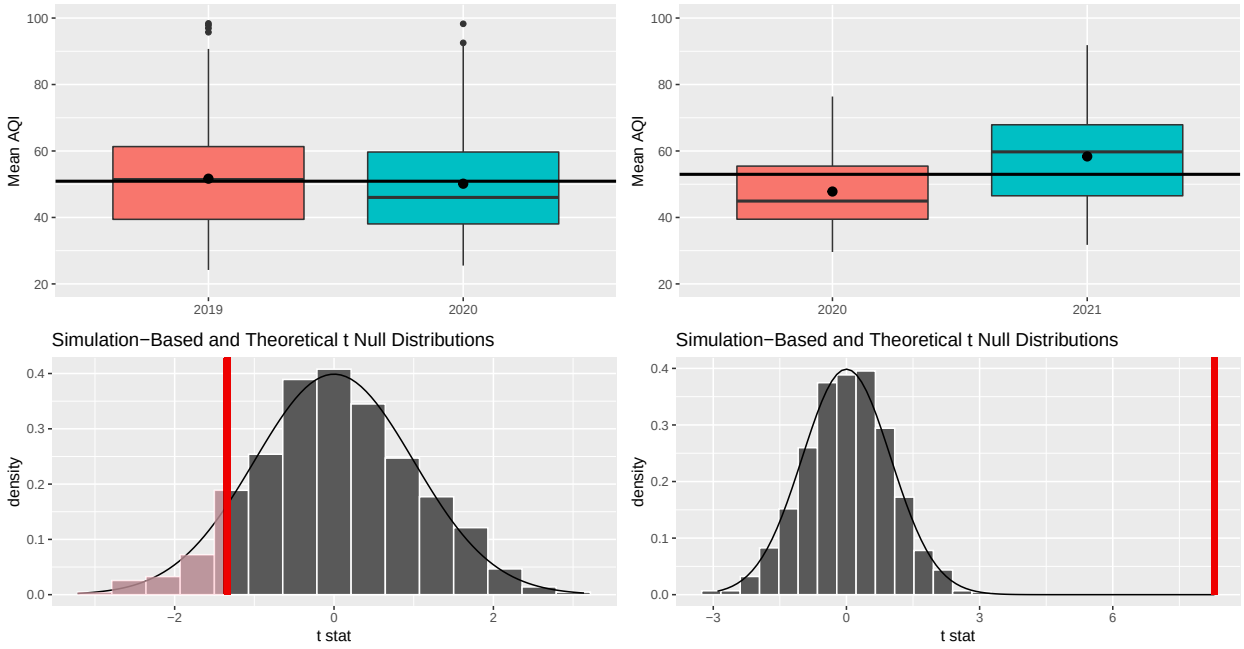
	Test Statistic	df	p−value
1	2.08042	9	0.99013

Chi−sq Test for Equality of Proportion

Hence, it is clearly visible **all other cities except Delhi are equally monitored in 2019 & 2020, where Delhi is more heavily monitored**.

- Now the Question that rises is *whether the Mean Level of Pollutants in 2020 Less than that of 2019?* Along with it we will also find out *whether the Mean Level of Pollutants in 2021 More than that of 2020* (Considering the Data of the First 6 months of 2020 and 2021)?

So, to Test our Hypothesis of the Mean Pollutant Level of 2020 is equal to that of 2019 against the Alternative that the Mean Pollutant Level has Dropped in 2020, we will perform a **two sample t-test for equality of mean**. Here after looking at the boxplots, the variances for each year looks more or less the same, and I took variances of both the samples to be same.



	estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
1	−1.513788	50.14927	51.66305	−1.342248	0.08996705	728	−Inf	1.115658	Two Sample t−test	less

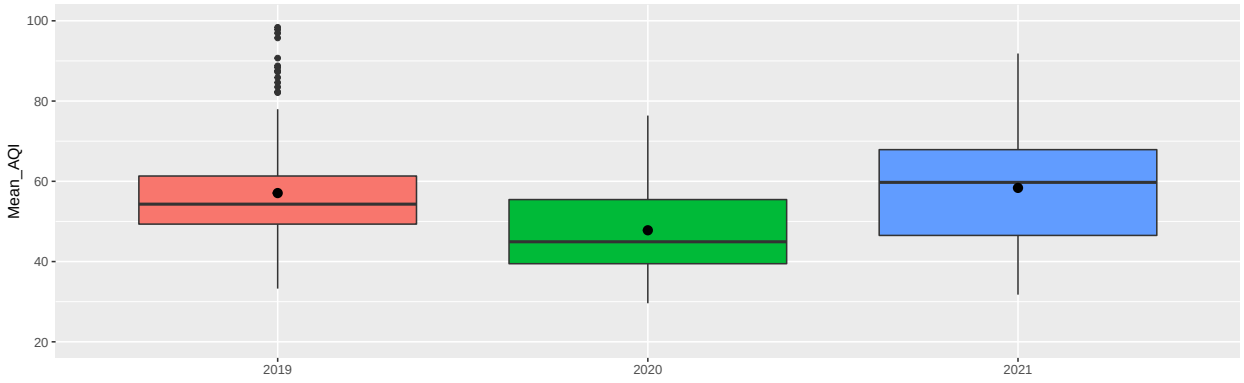
two−sample t−test for 2019 & 2020

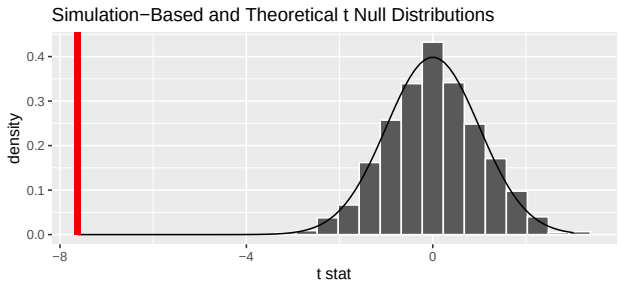
	estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
1	10.55898	58.35932	47.80034	8.325928	9.17533e−16	355	7.595315	Inf	Two Sample t−test	greater

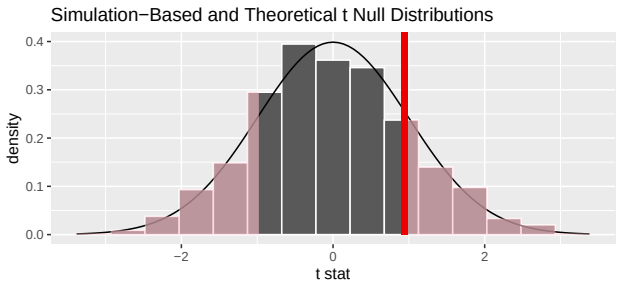
two−sample t−test for 2021 & 2021

So, **it is true that 2021's Pollutant Levels are more than that of 2020 to be precise it is atleast 7.5 points higher with 99% Confidence** when comparing the first 6 months. But it is surprising that **2020's Pollutant Levels haven't dropped significantly compared to 2019** despite Lockdowns and all other measures which were favorable for decrease of Air Pollutant Levels.

Now, we try to look at the first half and later half of the Years separately, cause the Nation-wide Lockdown was there during the First Half of 2020. We test the same hypothesis as above, but now on split data.







	estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
1	−9.273084	47.80034	57.07343	−7.628027	1.074962e−13	360	−Inf	−6.432391	Two Sample t−test	less

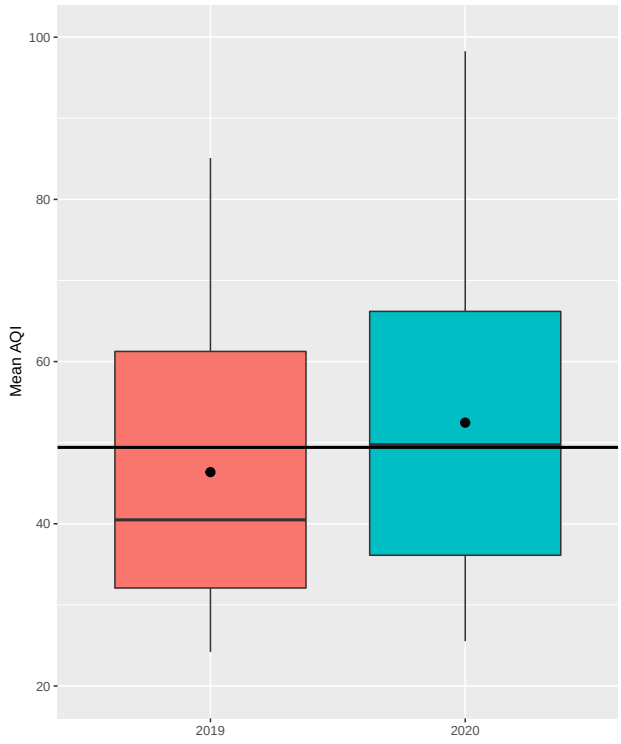
two−sample t−test for 2019 & 2020

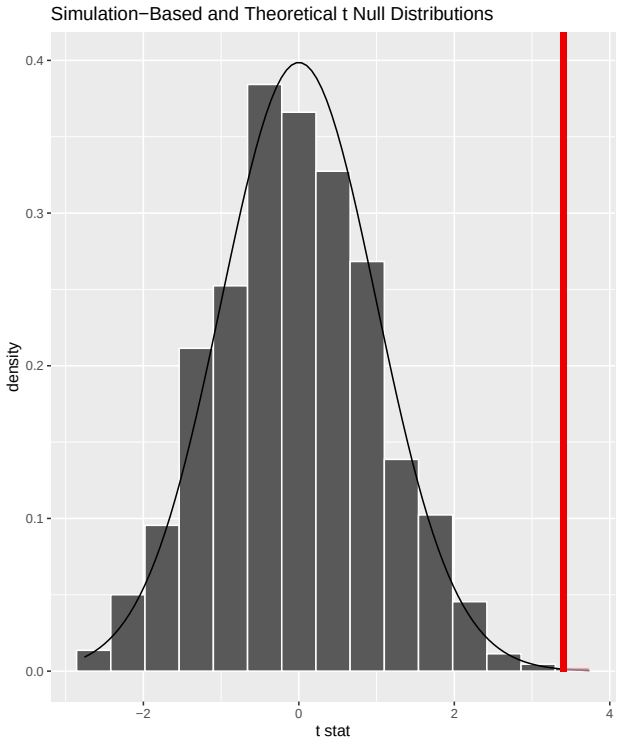
	estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
1	1.285895	58.35932	57.07343	0.9424858	0.3465886	353	−2.247578	4.819369	Two Sample t−test	two.sided

two−sample t−test for 2021 & 2019

Indeed our guess was right here. The **Pollutant Levels have decreased during the first half of 2020 by atleast 6.4 points with 99% confidence and we reject our Null Hypothesis**.

An Interesting look at the second half of 2020 and performing an *Hypothesis testing for the Mean Level of Pollutants for 2019 and 2020 are equal against that of 2020 was higher* reveals that **2020's Level was higher in the second half**.





	estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
1	6.10236	52.47266	46.3703	3.400758	0.0003730604	366	1.909575	Inf	Two Sample t−test	greater

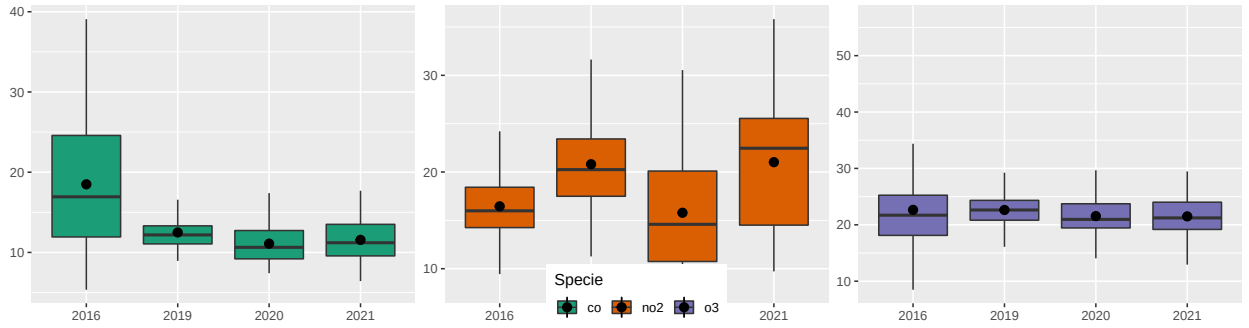
two−sample t−test for 2019 & 2020

- Now we turn to finding out *how different are the Mean AQI Levels of Each Pollutants.*

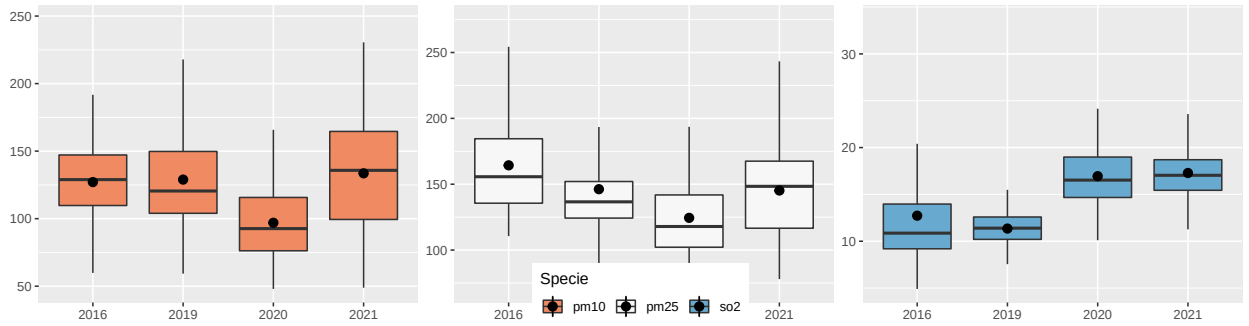
Considering the First Half of 2019, 2020 and 2021:

Here we will perform **two-sample t-tests** and will find **99% confidence interval for difference of two mean assuming same variance**. Our Tests are to be performed with *Null Hypothesis being 2020 and 2019 Mean Level of Pollutants are same against the alternative that the Mean Pollutant Level of 2020 being less than 2019.* Similarly we will perform a test for *2021's Mean Level being higher than 2020's.* And we will test *whether Mean of 2021 and 2019 are equal, which will be a both-sided test.*

The first row of each table corresponds to the first test, second row to second test and third row to third test from the tests mentioned above.



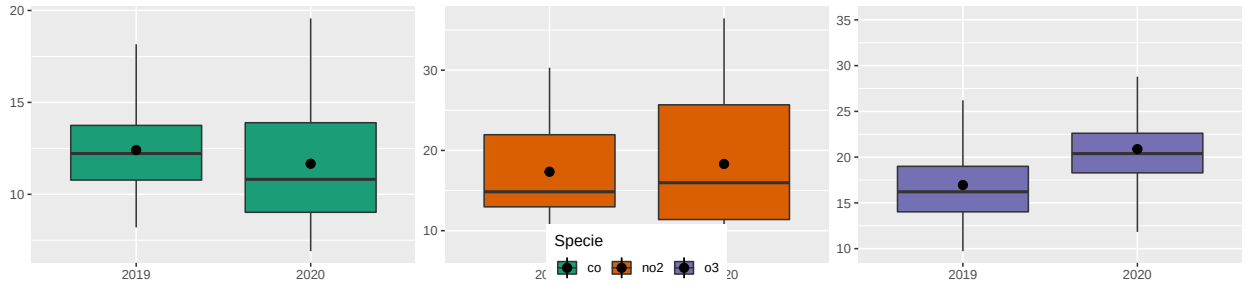
estimate	p.value	conf.low	conf.high	alternative	estimate	p.value	conf.low	conf.high	alternative	estimate	p.value	conf.low	conf.high	alternative
−1.399	0.000	−Inf	−0.844	less	−5.022	0.000	−Inf	−3.778	less	−1.084	0.000	−Inf	−0.337	less
0.471	0.038	−0.146	Inf	greater	5.234	0.000	3.790	Inf	greater	−0.055	0.557	−0.948	Inf	greater
−0.928	0.000	−1.579	−0.276	two.sided	0.211	0.705	−1.234	1.656	two.sided	−1.138	0.001	−2.022	−0.254	two.sided



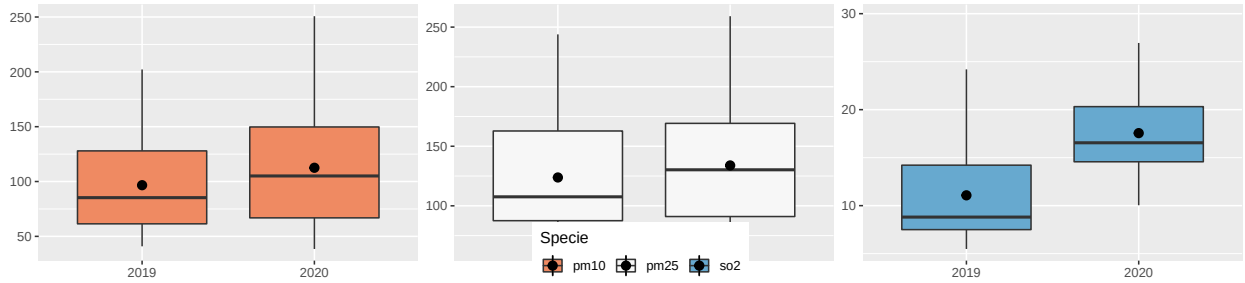
estimate	p.value	conf.low	conf.high	alternative	estimate	p.value	conf.low	conf.high	alternative	estimate	p.value	conf.low	conf.high	alternative
−31.968	0.000	−Inf	−24.164	less	−21.758	0.000	−Inf	−13.610	less	5.592	1.000	−Inf	6.227	less
36.602	0.000	28.290	Inf	greater	20.761	0.000	12.658	Inf	greater	0.340	0.148	−0.420	Inf	greater
4.635	0.244	−5.655	14.924	two.sided	−0.996	0.793	−10.832	8.839	two.sided	5.932	0.000	5.288	6.576	two.sided

Indeed! In case of most of the Pollutants our guess seem to hold! And **2020's Pollutant Levels dropped when compared to 2019 and 2021, whereas the means of 2021 and 2019 are similar**.

A similar Test Considering the second half of 2020 and 2019, where we *test the null against the alternative of 2020's Mean Level of Pollutant being higher than that of 2019*, reveals -



estimate	p.value	conf.low	conf.high	alternative	estimate	p.value	conf.low	conf.high	alternative	estimate	p.value	conf.low	conf.high	alternative
−0.744	0.996	−1.384	Inf	greater	0.97	0.081	−0.649	Inf	greater	3.922	0	2.893	Inf	greater



estimate	p.value	conf.low	conf.high	alternative	estimate	p.value	conf.low	conf.high	alternative	estimate	p.value	conf.low	conf.high	alternative
15.881	0.001	4.414	Inf	greater	10.103	0.017	−0.965	Inf	greater	6.481	0	5.386	Inf	greater

That **almost all the Pollutant's Levels have either gone up or remained same as that of 2019's**. A possible cause for this to happen **maybe due to a 40% increase in Stubble Burning during the later half of 2020 as per what various reports suggests**.

MODELING

In this section we will try to analyze the data and will provide a quantitative way to summarize the trends and patterns. We will mostly be using concepts and techniques pertaining to **Multiple Linear Regression(MLR)**.

REGRESSION

Here we will try to fit a MLR to the strong linear relationship we saw in the Visual Overview Section. Because of the difficulty in visualization, residuals will play an important role in deciding model fit.

- We will look at *what is the linear relationship between each Pollutant Level and weather parameters*, *what is the linear relationship between each Pollutant with other Pollutant* We will also look at the *standard error of our the estimates and their confidence intervals* and *how much the weather parameters explains the variation in the pollutant levels.*
- Some assumptions made during Modelling are-
  - The weather parameters are independent of the Level of pollutants.

- The weather parameters are similarly distributed across 2019, 2020 and 2021 [Except for Mumbai where we see an increased wind-speed in the later two years]

In most of the models which we will try to fit we will look at the first half and the second half of the year separately.

1. Model Based on 2019

2019 Models

Kolkata

t Half of the Year (Pollutant Levels On Weather Parameters)

rm	SO2	NO2	CO	O3	PM10	PM25
rcept)	−369.15*** [−682.7, −55.59]	−448.75* [−913.45, 15.94]	207.15* [−43.7, 458.01]	4.98 [−316.22, 326.18]	1138.03 [−2592.92, 4868.98]	−1766.28* [−3675.66, 143.1
ew	−2.58*** [−3.06, −2.1]	−5.03*** [−5.74, −4.32]	−2.24*** [−2.63, −1.86]	−2.27*** [−2.76, −1.78]	−30.95*** [−36.67, −25.24]	−16.62*** [−19.55, −13.7]
idity	0.3*** [0.21, 0.39]	0.7*** [0.57, 0.83]	0.38*** [0.31, 0.45]	0.33*** [0.24, 0.42]	4.07*** [3.03, 5.11]	2.81*** [2.28, 3.34]
sure	0.36*** [0.05, 0.66]	0.44* [−0.01, 0.9]	−0.21* [−0.45, 0.04]	−0.02 [−0.34, 0.29]	−1.11 [−4.77, 2.55]	1.85* [−0.02, 3.72]
erature	2.11*** [1.68, 2.53]	3.31*** [2.68, 3.95]	1.25*** [0.9, 1.59]	2.09*** [1.65, 2.53]	19.66*** [14.57, 24.75]	8.47*** [5.86, 11.07]
speed`	0.61 [−0.56, 1.78]	0.2 [−1.53, 1.94]	−1.02* [−1.95, −0.08]	1.01* [−0.19, 2.21]	−29.22*** [−43.15, −15.29]	−11.98*** [−19.11, −4.85]
R2	0.74	0.81	0.67	0.71	0.59	0.73

*** p < 0.001; ** p < 0.01; * p < 0.05; 99% confidence Intervals

First Half of the Year (Pollutant Levels on Other Pollutant Levels)

term	SO2	NO2	CO	O3	PM10	PM25
(Intercept)	−1.94* [−3.79, −0.08]	0.66 [−2.37, 3.69]	−2.58*** [−3.68, −1.49]	4.89*** [2.46, 7.32]	−41.2*** [−57.61, −24.79]	74.11*** [66.73, 81.48]
SO2	NA	0.84*** [0.69, 1]	0.1*** [0.03, 0.17]	0.35*** [0.2, 0.49]	0.42 [−0.6, 1.45]	−0.17 [−0.84, 0.5]
NO2	0.32*** [0.26, 0.38]	NA	0.06*** [0.02, 0.11]	0.16*** [0.07, 0.25]	−0.06 [−0.7, 0.57]	0.83*** [0.42, 1.23]
CO	0.27*** [0.09, 0.45]	0.45*** [0.16, 0.75]	NA	0.53*** [0.29, 0.77]	6.55*** [5.03, 8.07]	2.17*** [1.09, 3.24]
O3	0.2*** [0.11, 0.28]	0.23*** [0.1, 0.37]	0.11*** [0.06, 0.16]	NA	−0.57 [−1.34, 0.2]	−0.25 [−0.75, 0.26]
PM10	0.01 [−0.01, 0.02]	0 [−0.02, 0.02]	0.03*** [0.02, 0.04]	−0.01 [−0.03, 0]	NA	0.32*** [0.26, 0.38]
PM25	0 [−0.02, 0.01]	0.06*** [0.03, 0.09]	0.02*** [0.01, 0.03]	−0.01 [−0.04, 0.01]	0.75*** [0.6, 0.9]	NA
R2	0.83	0.85	0.87	0.67	0.84	0.82

*** p < 0.001; ** p < 0.01; * p < 0.05; 99% confidence Intervals

Second Half of the Year (Pollutant Levels on Weather Parameters)

term	SO2	NO2	CO	O3	PM10	PM25
(Intercept)	−6.23 [−46.98, 34.52]	4.11 [−30.53, 38.74]	3.31 [−19.59, 26.21]	−25.08 [−62.4, 12.24]	319.14 [−92.31, 730.6]	70.43 [−145.97, 286.83]
dew	−3.5*** [−4.13, −2.88]	−4.12*** [−4.66, −3.59]	−2.29*** [−2.64, −1.94]	−3.46*** [−4.03, −2.88]	−43.33*** [−49.64, −37.01]	−20.58*** [−23.91, −17.26]
humidity	0.31*** [0.17, 0.45]	0.47*** [0.35, 0.6]	0.29*** [0.21, 0.37]	0.4*** [0.27, 0.53]	4.04*** [2.6, 5.48]	2.35*** [1.59, 3.11]
pressure	0.02 [−0.02, 0.06]	0.01 [−0.02, 0.04]	0 [−0.02, 0.02]	0.03* [−0.01, 0.06]	−0.13 [−0.52, 0.26]	0.07 [−0.13, 0.27]
temperature	2.07*** [1.47, 2.66]	2.38*** [1.88, 2.88]	1.26*** [0.93, 1.59]	2.25*** [1.71, 2.8]	24.34*** [18.36, 30.32]	10.38*** [7.23, 13.52]
wind−speed`	1.86*** [0.62, 3.11]	1.55*** [0.5, 2.61]	1.1*** [0.4, 1.8]	2.51*** [1.37, 3.65]	−2.37 [−14.94, 10.2]	6.45* [−0.16, 13.06]
R2	0.48	0.65	0.6	0.61	0.49	0.52

*** p < 0.001; ** p < 0.01; * p < 0.05; 99% confidence Intervals

Second Half of the Year (Pollutant Levels on Other Pollutant Levels)

term	SO2	NO2	CO	O3	PM10	PM25
(Intercept)	−2.79*** [−4.99, −0.6]	3.52*** [2.18, 4.87]	−1.34*** [−2.23, −0.46]	1.11 [−0.89, 3.12]	−50.66*** [−68.65, −32.67]	53.96*** [44.07, 63.86]
SO2	NA	0.12*** [0.05, 0.19]	0.03 [−0.01, 0.08]	0.18*** [0.08, 0.28]	1.61*** [0.68, 2.54]	−0.07 [−0.65, 0.51]
NO2	0.3*** [0.12, 0.47]	NA	0.35*** [0.29, 0.41]	0.42*** [0.26, 0.57]	1.13 [−0.36, 2.62]	1.37*** [0.46, 2.27]
CO	0.2 [−0.07, 0.47]	0.85*** [0.71, 1]	NA	0.36*** [0.12, 0.61]	3.42*** [1.13, 5.72]	2.31*** [0.91, 3.72]
O3	0.22*** [0.1, 0.34]	0.2*** [0.13, 0.28]	0.07*** [0.02, 0.12]	NA	1.48*** [0.45, 2.51]	−0.19 [−0.83, 0.45]
PM10	0.02*** [0.01, 0.04]	0.01 [0, 0.01]	0.01*** [0, 0.01]	0.02*** [0.01, 0.03]	NA	0.22*** [0.16, 0.29]
PM25	0 [−0.02, 0.02]	0.02*** [0.01, 0.03]	0.01*** [0.01, 0.02]	−0.01 [−0.02, 0.01]	0.59*** [0.42, 0.76]	NA
R2	0.69	0.89	0.87	0.77	0.78	0.72

*** p < 0.001; ** p < 0.01; * p < 0.05; 99% confidence Intervals

Delhi

irst Half of the Year (Pollutant Levels On Weather Parameters)

term	SO2	NO2	CO	O3	PM10	PM25
ntercept)	10.83 [−18.16, 39.82]	−60.93* [−117.13, −4.73]	−19.1 [−95.96, 57.77]	−11.09 [−102.54, 80.37]	−205.89 [−764.27, 352.48]	−308.94*** [−580.61, −37.26]
dew	−0.54*** [−1.03, −0.04]	−2.36*** [−3.32, −1.4]	−1.21* [−2.52, 0.1]	1.05 [−0.51, 2.61]	−6.1 [−15.62, 3.42]	−8.58*** [−13.21, −3.95]
umidity	0.12*** [0.04, 0.19]	0.38*** [0.25, 0.52]	0.31*** [0.12, 0.5]	0.21* [−0.02, 0.43]	1.6*** [0.22, 2.98]	2.39*** [1.72, 3.07]
ressure	−0.02 [−0.05, 0.01]	0.04* [−0.01, 0.1]	0 [−0.07, 0.08]	−0.03 [−0.12, 0.07]	0.14 [−0.42, 0.7]	0.35*** [0.07, 0.62]
mperature	0.73*** [0.46, 1.01]	1.86*** [1.32, 2.39]	0.93*** [0.21, 1.66]	0.98*** [0.11, 1.85]	7.46*** [2.17, 12.75]	4.42*** [1.85, 6.99]
nd−speed`	3.71*** [3.02, 4.41]	6.72*** [5.37, 8.06]	6.58*** [4.74, 8.42]	8.41*** [6.22, 10.6]	56.52*** [43.16, 69.87]	14.51*** [8.01, 21.01]
R2	0.79	0.8	0.6	0.67	0.67	0.59

*** p < 0.001; ** p < 0.01; * p < 0.05; 99% confidence Intervals

First Half of the Year (Pollutant Levels on Other Pollutant Levels)

term	SO2	NO2	CO	O3	PM10	PM25
(Intercept)	1.6 [−0.78, 3.97]	−0.16 [−4.85, 4.52]	−5.89* [−12.32, 0.54]	3.86 [−3.42, 11.14]	−43.38*** [−79.16, −7.61]	101.99*** [84.15, 119.84]
SO2	NA	0.82*** [0.58, 1.05]	0.32* [−0.03, 0.68]	1.2*** [0.83, 1.57]	2.44*** [0.48, 4.41]	−0.22 [−1.44, 1.01]
NO2	0.21*** [0.15, 0.27]	NA	0.36*** [0.19, 0.54]	0.08 [−0.13, 0.28]	1.59*** [0.6, 2.58]	1.19*** [0.58, 1.79]
CO	0.04* [0, 0.09]	0.19*** [0.1, 0.28]	NA	0.13* [−0.02, 0.28]	0.24 [−0.49, 0.97]	0.49*** [0.04, 0.93]
O3	0.13*** [0.09, 0.17]	0.03 [−0.05, 0.12]	0.1* [−0.01, 0.22]	NA	1.85*** [1.25, 2.46]	−0.77*** [−1.16, −0.38]
PM10	0.01*** [0, 0.02]	0.03*** [0.01, 0.04]	0.01 [−0.02, 0.03]	0.08*** [0.05, 0.1]	NA	0.28*** [0.21, 0.35]
PM25	0 [−0.02, 0.01]	0.05*** [0.03, 0.08]	0.04*** [0, 0.08]	−0.08*** [−0.12, −0.04]	0.75*** [0.56, 0.93]	NA
R2	0.83	0.84	0.67	0.76	0.84	0.68

*** p < 0.001; ** p < 0.01; * p < 0.05; 99% confidence Intervals

Second Half of the Year (Pollutant Levels on Weather Parameters)

term	SO2	NO2	CO	O3	PM10	PM25
(Intercept)	15.17 [−31.05, 61.39]	−35.58*** [−65.79, −5.37]	0.38 [−39.9, 40.66]	−14.33 [−89.54, 60.88]	201.44* [−36.03, 438.9]	100.3* [−26.11, 226.71]
dew	−4.11*** [−5.2, −3.02]	−3.25*** [−3.96, −2.54]	−2.72*** [−3.66, −1.77]	−6.1*** [−7.87, −4.33]	−38.95*** [−44.54, −33.37]	−19.41*** [−22.38, −16.44]
humidity	0.38*** [0.12, 0.64]	0.54*** [0.37, 0.71]	0.4*** [0.18, 0.63]	0.85*** [0.43, 1.27]	3.55*** [2.22, 4.88]	2.03*** [1.32, 2.74]
pressure	−0.03 [−0.09, 0.03]	0 [−0.04, 0.04]	−0.03 [−0.08, 0.02]	−0.06 [−0.16, 0.03]	−0.37*** [−0.66, −0.07]	0.01 [−0.15, 0.17]
temperature	3.32*** [2.32, 4.32]	3.36*** [2.71, 4.02]	2.67*** [1.81, 3.54]	6.54*** [4.92, 8.16]	34.16*** [29.04, 39.28]	11.85*** [9.12, 14.57]
`wind−speed`	1.15 [−0.53, 2.84]	2.73*** [1.63, 3.83]	3.22*** [1.75, 4.69]	6.44*** [3.7, 9.18]	8.7* [0.05, 17.35]	0.21 [−4.39, 4.82]
R2	0.33	0.7	0.45	0.55	0.63	0.55

*** p < 0.001; ** p < 0.01; * p < 0.05; 99% confidence Intervals

Second Half of the Year (Pollutant Levels on Other Pollutant Levels)

term	SO2	NO2	CO	O3	PM10	PM25
(Intercept)	−4.82* [−10.6, 0.97]	5.31*** [1.58, 9.04]	1.37 [−3.88, 6.62]	2.65 [−6.87, 12.16]	−56.86*** [−81.53, −32.2]	89.21*** [76.06, 102.35]
SO2	NA	0.12*** [0.05, 0.2]	−0.15*** [−0.25, −0.04]	0.54*** [0.36, 0.71]	0.37 [−0.13, 0.87]	0.21 [−0.11, 0.54]
NO2	0.29*** [0.12, 0.47]	NA	0.52*** [0.37, 0.66]	0.98*** [0.72, 1.24]	1.94*** [1.2, 2.68]	0.18 [−0.33, 0.68]
CO	−0.18*** [−0.3, −0.05]	0.27*** [0.19, 0.34]	NA	0.07 [−0.14, 0.27]	1.25*** [0.71, 1.79]	−0.24 [−0.6, 0.12]
O3	0.2*** [0.13, 0.27]	0.15*** [0.11, 0.2]	0.02 [−0.04, 0.08]	NA	0.41*** [0.11, 0.72]	−0.2* [−0.39, 0]
PM10	0.02 [−0.01, 0.05]	0.04*** [0.03, 0.06]	0.05*** [0.03, 0.08]	0.06*** [0.02, 0.1]	NA	0.4*** [0.34, 0.46]
PM25	0.03 [−0.01, 0.07]	0.01 [−0.02, 0.04]	−0.02 [−0.06, 0.01]	−0.06* [−0.13, 0]	0.94*** [0.8, 1.08]	NA
R2	0.49	0.77	0.55	0.65	0.8	0.62

*** p < 0.001; ** p < 0.01; * p < 0.05; 99% confidence Intervals

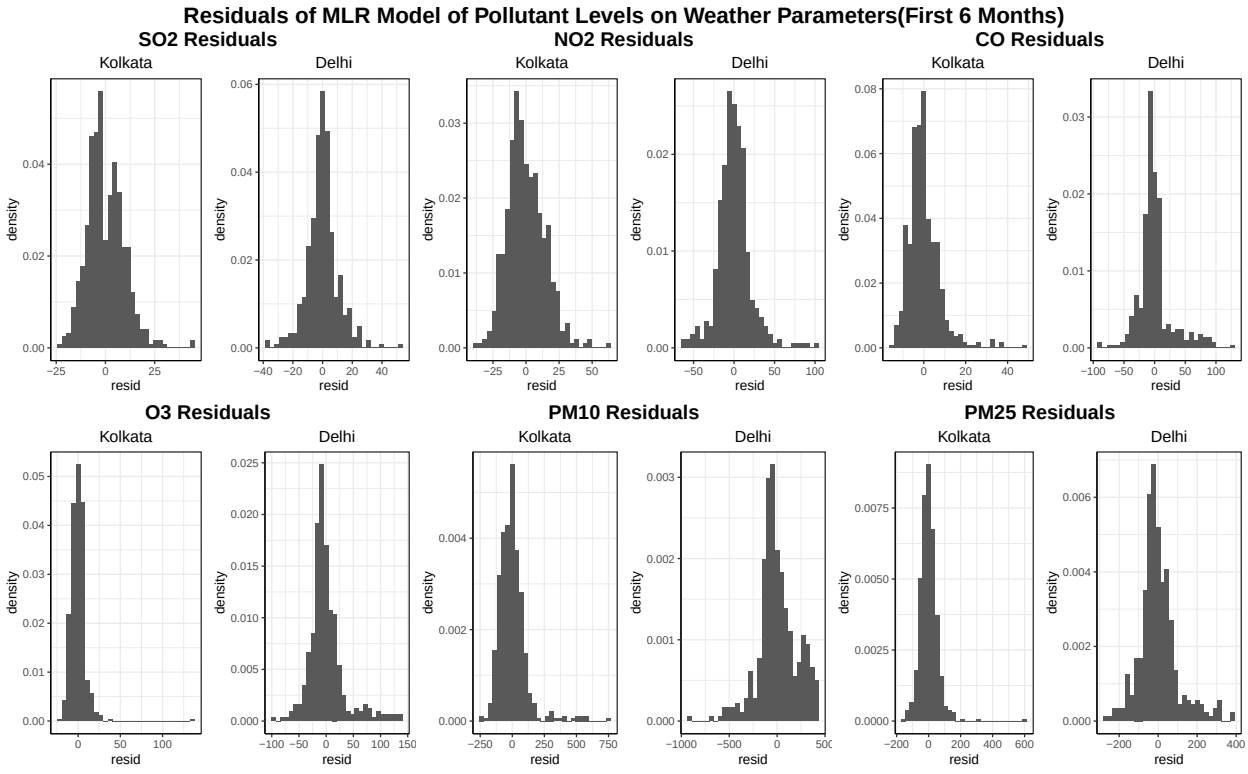
I have presented here the models for Kolkata and Delhi only. Other models for other cities will be attached at the end if interested to take a look at them. We have fit a **MLR model with each pollutants as**

the response and the weather parameters as the predictors and similarly to figure out the strong relationship between the pollutants among themselves we have fit a separate **MLR model considering only the other pollutant's Levels**. The reason behind splitting the year into 2 parts are as can be seen from the visual overview section the **first half of the year has weather parameter levels much different from that of the second half** and the second reason being **we have data for the year 2021 till first 6 months only so fitting a model for the first half of the year will help us compare it with the year 2021**.

We see an interesting thing that is in the **first half of the year, the weather parameters are able to explain the pollutant Levels much better compared to that of the second half**. The **strength of the relationship of Pollutant Levels on other Pollutant Levels remains almost similar during both halves of the year. Temperature, Humidity and Dew have high significant effect throughout the year in explaining the Pollutant Levels throughout the year, which can be due to the intricate physical relationship that exists between this quantities i.e. concentration of gas being directly proportional to temperature, etc.**

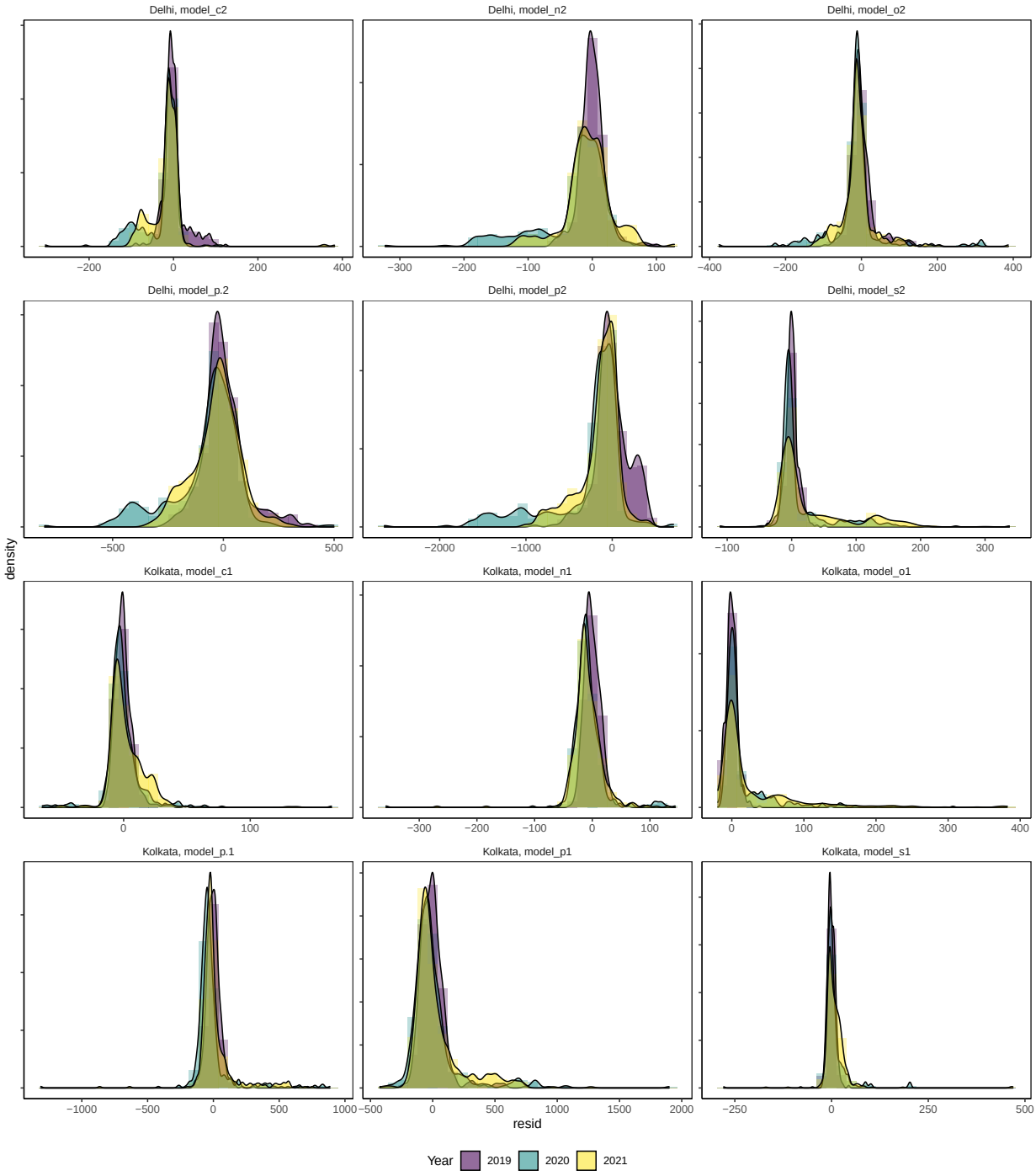
The confidence intervals are mentioned under each coefficient giving us information about it's significance of being present in the model. Any coefficient which includes 0 in its confidence interval can be thought of being less important in explaining the response.

- How well will take a look at *how the residuals of the above mentioned models are spread*



Indeed it seems that the **residuals are normally distributed with mean around 0**. Therefore our model doesn't seems to have an issue in assuming that the **relationship is linear between Pollutant Levels and weather parameters**.

- Now since, we have assumed that the weather parameters remains similar in all the years. It would be interesting to see *how well our model fitted with 2019's weather parameter performs when used on 2020's first 6 Months Data & 2021's first 6 Months Data.*



Model R2 on 2019 Data

Pollutant	Kolkata	Delhi
so2	0.74	0.79
no2	0.81	0.80
co	0.67	0.60
o3	0.71	0.67
pm10	0.59	0.67
pm25	0.73	0.59

Model R2 on 2020 Data

Pollutant	Kolkata	Delhi
so2	0.13	0.68
no2	0.34	0.56
co	0.09	0.71
o3	0.40	0.51
pm10	0.07	0.45
pm25	0.18	0.47

Model R2 on 2021 Data

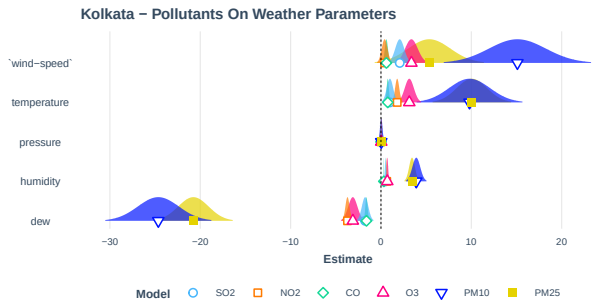
Pollutant	Kolkata	Delhi
so2	0.22	0.69
no2	0.47	0.71
co	0.18	0.47
o3	0.39	0.62
pm10	0.30	0.75
pm25	0.37	0.67

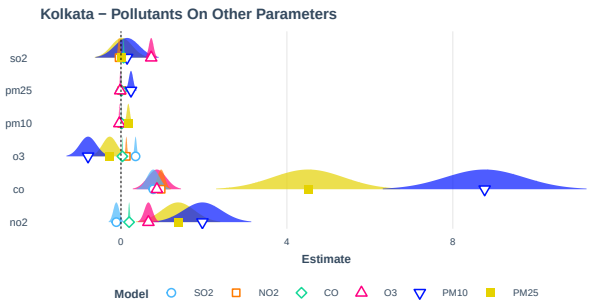
We see an interesting thing that out **Model fitted with 2019 Weather parameters performs less well on 2020's Data as compared to 2021's. The histogram-density plots for residuals also shows our Model predicts higher values for the weather parameters in 2020 compared to the actual values observed which suggests that 2020's Pollutant Levels were lower during first 6 Months.**

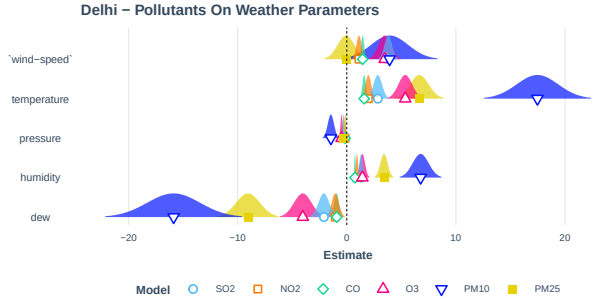
2. Model Based on 2020. We will take a Visual Look instead of the tables here, along with the estimate Distributions.

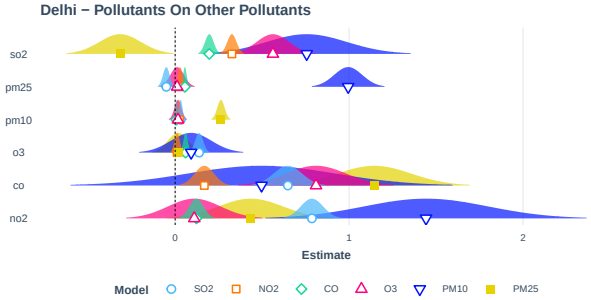
Models−2020

First 6 Months









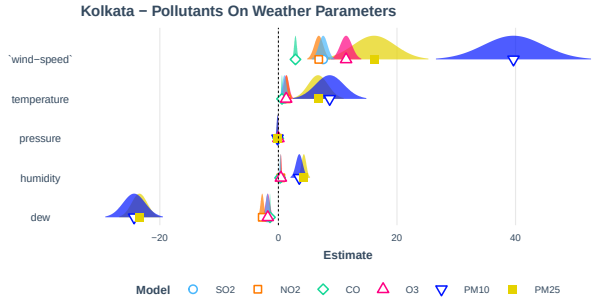
Model R2

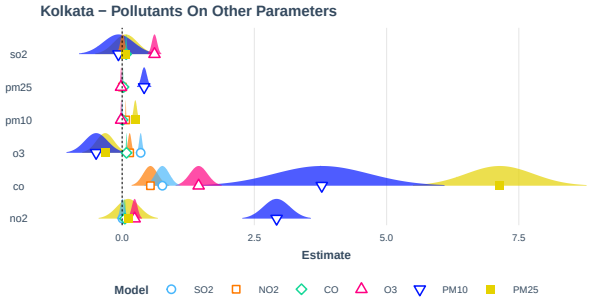
	so2	no2	co	o3	pm10	pm25
Kolkata	0.31	0.56	0.53	0.42	0.42	0.44
Delhi	0.72	0.64	0.78	0.57	0.55	0.56

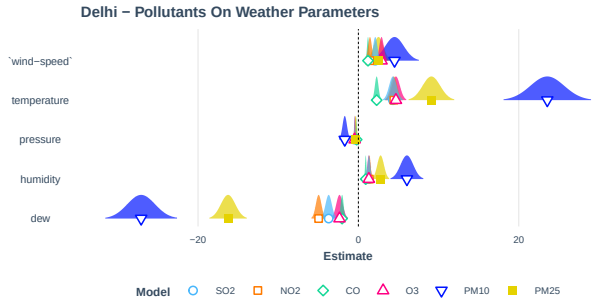
Model R2

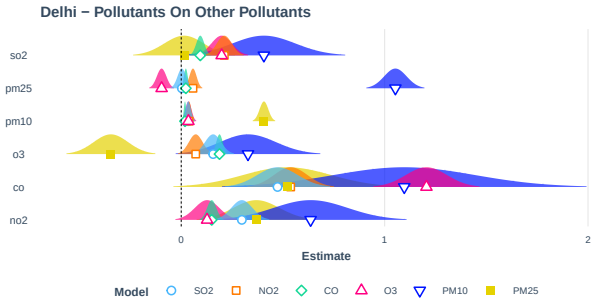
	so2	no2	co	o3	pm10	pm25
Kolkata	0.55	0.73	0.79	0.61	0.61	0.49
Delhi	0.84	0.81	0.79	0.62	0.75	0.68

Last 6 Months









Model R2

	so2	no2	co	o3	pm10	pm25
Kolkata	0.44	0.47	0.70	0.50	0.48	0.56
Delhi	0.61	0.66	0.74	0.66	0.63	0.66

Model R2

	so2	no2	co	o3	pm10	pm25
Kolkata	0.58	0.60	0.76	0.64	0.62	0.61
Delhi	0.67	0.74	0.81	0.67	0.82	0.76

CONCLUSION

- In 2020, During Lockdown i.e. in the **first Half of the Year there was a drop in the Pollutant Levels and AQI Levels** across most of the Major Cities, our study included Kolkata, Delhi, Muzaffarnagar and Mumbai. Though the **Particulate Matter Levels in Mumbai were much higher which can be seen due to Cyclonic Impact of *Nisarga***.
- Despite there being several restrictions on travel, transportation, industrial work which are major contributors to pollutant Levels, we saw an **increased Level of Pollutants in the second half of the year 2020 which surpassed the Levels of 2019** which maybe due to the **40% increase in the Stubble Burning during the end of the Year**.
- **Coastal Regions tend to have lower AQI Levels and Pollutant Levels** mostly because of the influence of **winds and more Cyclonic Conditions**.
- **Major Contributor to AQI Levels are particulate matters**. So reducing there levels will greatly help reducing the overall Air Quality.

SUGGESTIONS

- Introduction of Electric Powered Vehicles can greatly combat Air Pollution reducing Levels of NO2, SO2, CO which are released by Diesel and Petrol Powered Engines.
- Controlled Stubble Burning can stabilize Air Quality throughout the later half of the Year.