Air Quality and Covid-19: An exploratory study

Done by: Tracey Ha, Pooja Mallard, Alison Beer, Vineet Goyal

Our Motivation

The impacts of Covid-19 are widespread across all areas of the economy - the individual, business and government level. With restricted borders, social distancing rules and the “new normal” of working from home, there has been a reduction in our anthropogenic impact on the environment, more specifically, on air quality. Our question is, to what extent?

Our study is focused on *identifying a correlation between developments in Covid-19 active cases and the changes seen in air quality across the world.*

Key Research Questions

1. Are the changes in the number of Covid-19 active cases correlated to the changes observed in Australia's air quality in the first half of 2020?

2. Are the changes in the number of Covid-19 active cases correlated to the changes observed in India's air quality in the first half of 2020? What comparisons can be drawn between Australia and India?

3. Are the changes in the number of Covid-19 active cases correlated to the changes observed in the world's air quality in the first half of 2020?

Core Hypothesis

We expect to observe a strong, negative correlation between the number of Covid-19 active cases and the level of air pollutants, and more drastic improvements in air quality in developing countries.

How will air quality be defined?

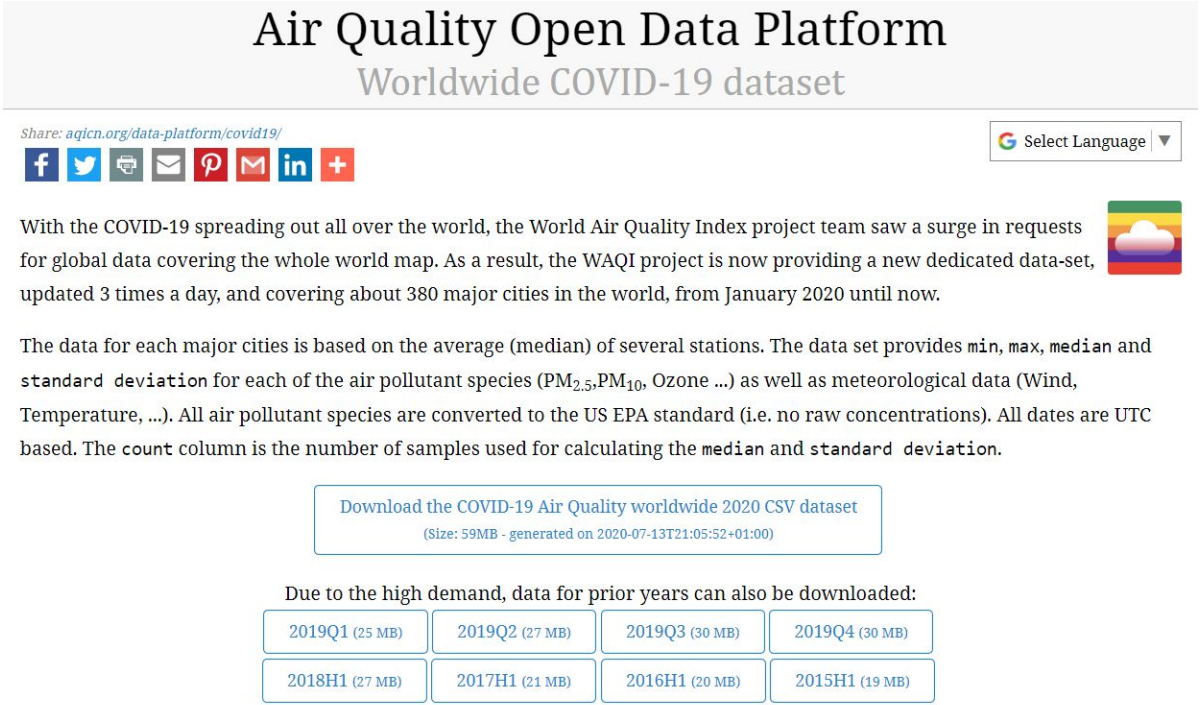
Air quality in our dataset is determined by the median level concentrations of various major air pollutants, of which we will focus on two:

1. Carbon Monoxide (CO) - colorless/odorless gas
    - Originates from vehicle exhaust, burning fossil fuels, smoke from wood fires
    - CO concentration measured through Parts Per Million (PPM)
2. Particulate Matter 2.5 (PM2.5)
    - Originates from natural sources (i.e. bush fires), mining, vehicle exhaust
    - PM2.5 concentration measured through microgram per cubic metre (μg/m3)

Concentrations of pollutants affected by weather changes or amount of pollutant generated (1).

Data Exploration & Clean-Up Process

Where is the air quality data from?



Air Quality Open Data Platform

Worldwide COVID-19 dataset

Share: aqicn.org/data-platform/covid19/

Select Language

With the COVID-19 spreading out all over the world, the World Air Quality Index project team saw a surge in requests for global data covering the whole world map. As a result, the WAQI project is now providing a new dedicated data-set, updated 3 times a day, and covering about 380 major cities in the world, from January 2020 until now.

The data for each major cities is based on the average (median) of several stations. The data set provides `min`, `max`, `median` and `standard deviation` for each of the air pollutant species ($PM_{2.5}$,PM_{10}, Ozone ...) as well as meteorological data (Wind, Temperature, ...). All air pollutant species are converted to the US EPA standard (i.e. no raw concentrations). All dates are UTC based. The `count` column is the number of samples used for calculating the `median` and `standard deviation`.

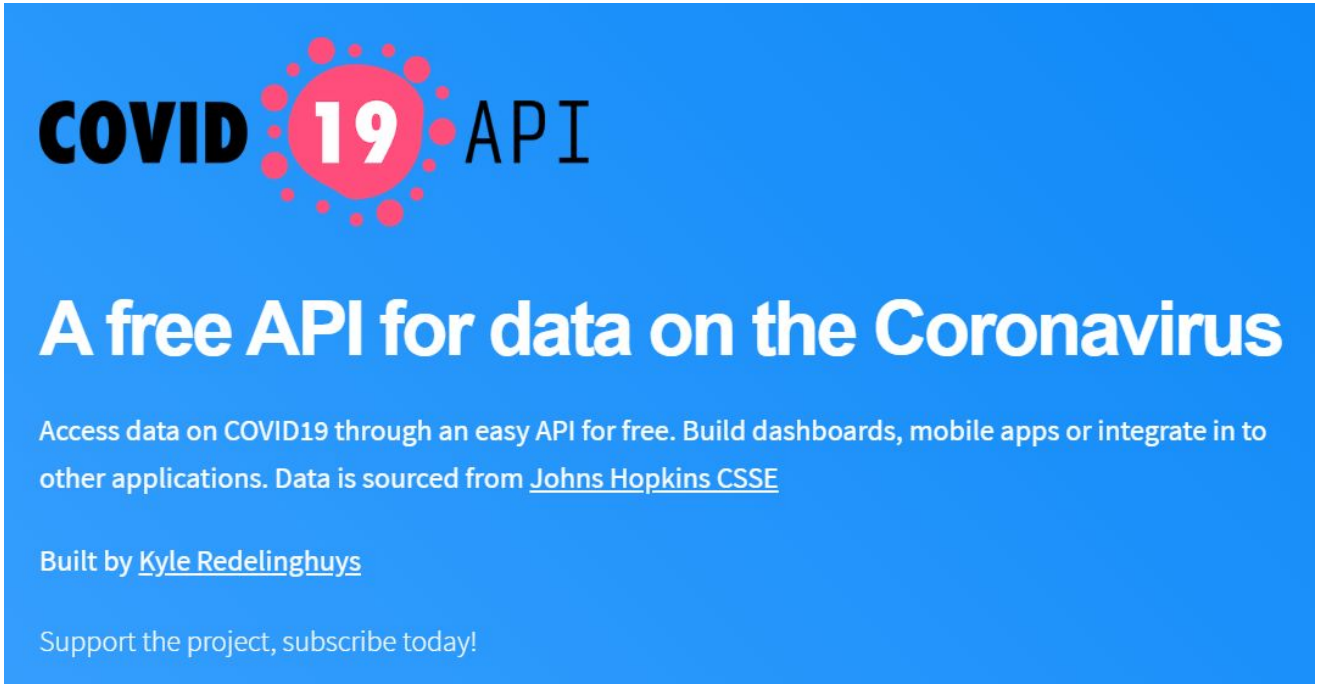
Download the COVID-19 Air Quality worldwide 2020 CSV dataset
(Size: 59MB - generated on 2020-07-13T21:05:52+01:00)

Due to the high demand, data for prior years can also be downloaded:

2019Q1 (25 MB)	2019Q2 (27 MB)	2019Q3 (30 MB)	2019Q4 (30 MB)
2018H1 (27 MB)	2017H1 (21 MB)	2016H1 (20 MB)	2015H1 (19 MB)

Source: https://aqicn.org/data-platform/covid19/

Where is the Covid-19 data from?



Source: https://covid19api.com/

Snapshot of the air data before cleaning

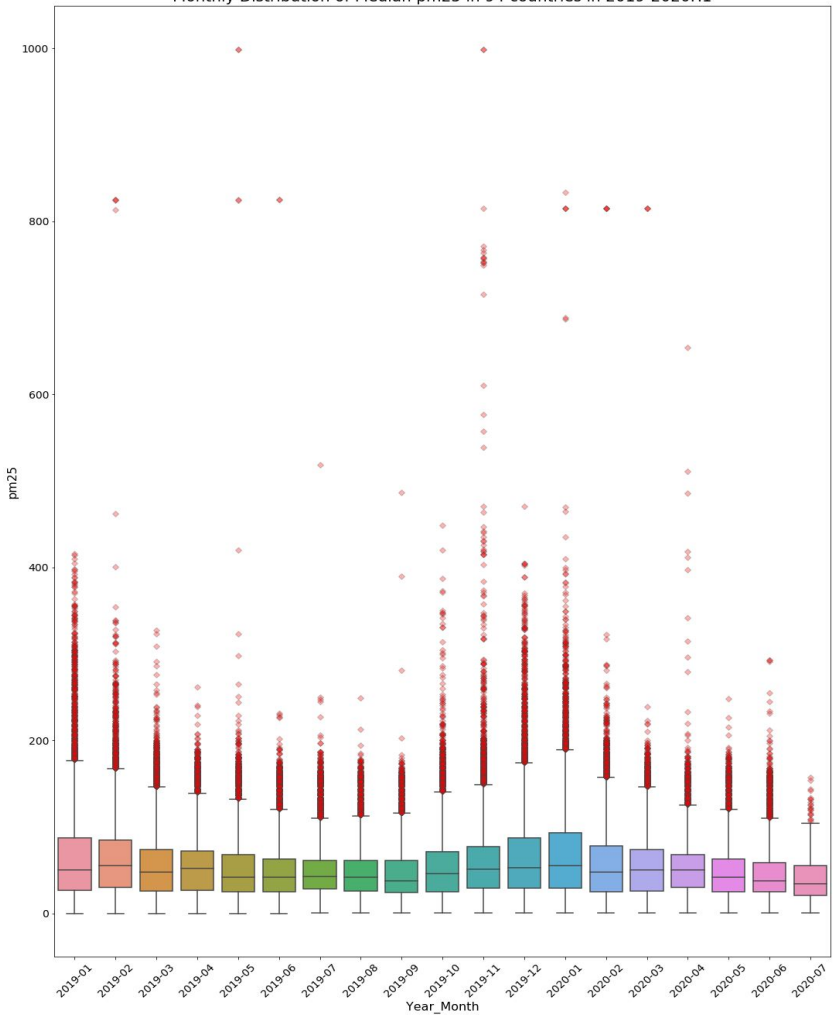
```
In [4]: airdf_2019_2020.head()
```

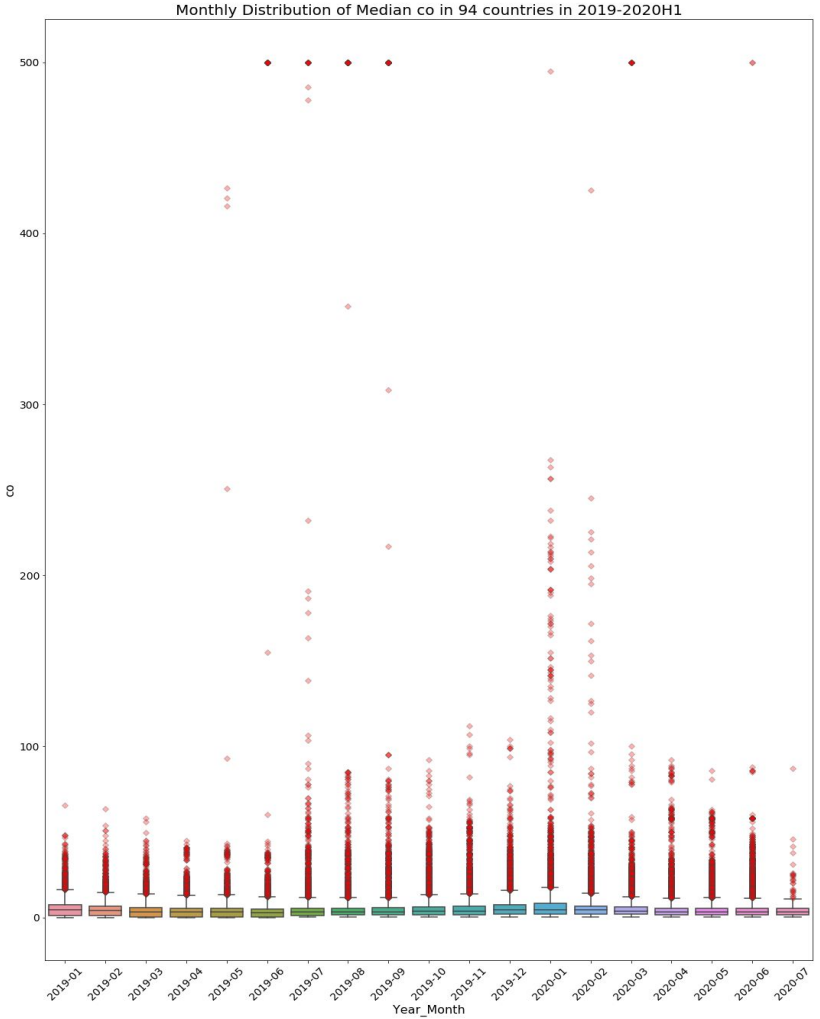
Out[4]:

	Date	Country	City	Specie	count	min	max	median	variance
0	31/05/2020	IR	Isfahan	temperature	120	17.5	35.0	27.5	331.51
1	13/06/2020	IR	Isfahan	temperature	144	16.0	36.5	27.5	488.74
2	3/07/2020	IR	Isfahan	temperature	67	19.0	33.0	24.0	128.08
3	28/03/2020	IR	Isfahan	temperature	240	3.0	14.0	9.5	136.68
4	23/04/2020	IR	Isfahan	temperature	168	6.0	25.5	16.0	400.79

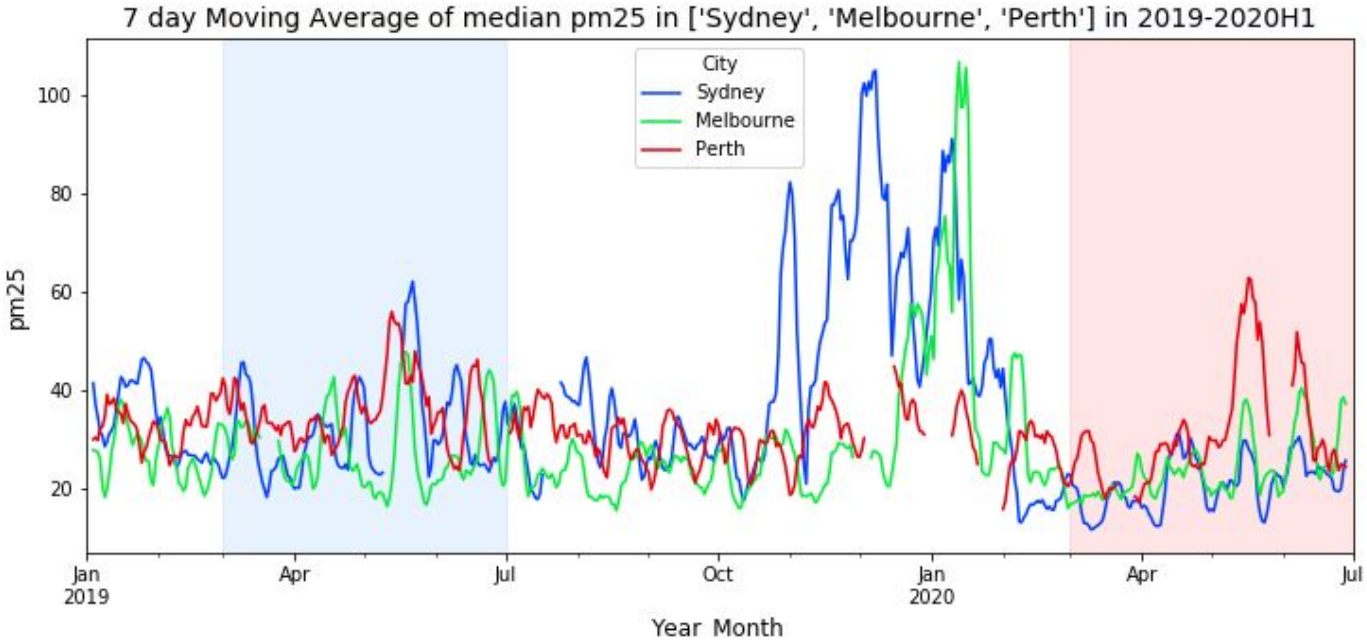
What did we do to clean up?

- Remove the meteorology-related species: **temperature, humidity, pressure, wind-speed, dew, wind-gust, wind speed, wind gust, precipitation, wd (wind direction), and uvi.**
- Remove species with the least number of available data points including **aqi, pol, pm1, mepaqi, and neph**
- Change date column into a DateTime format
- Set the start and end dates for the data (from 1/1/2019 to 1/7/2020).
- Draw boxplots to show monthly distributions of the median values of different air pollutants across 2019 and 2020H1.

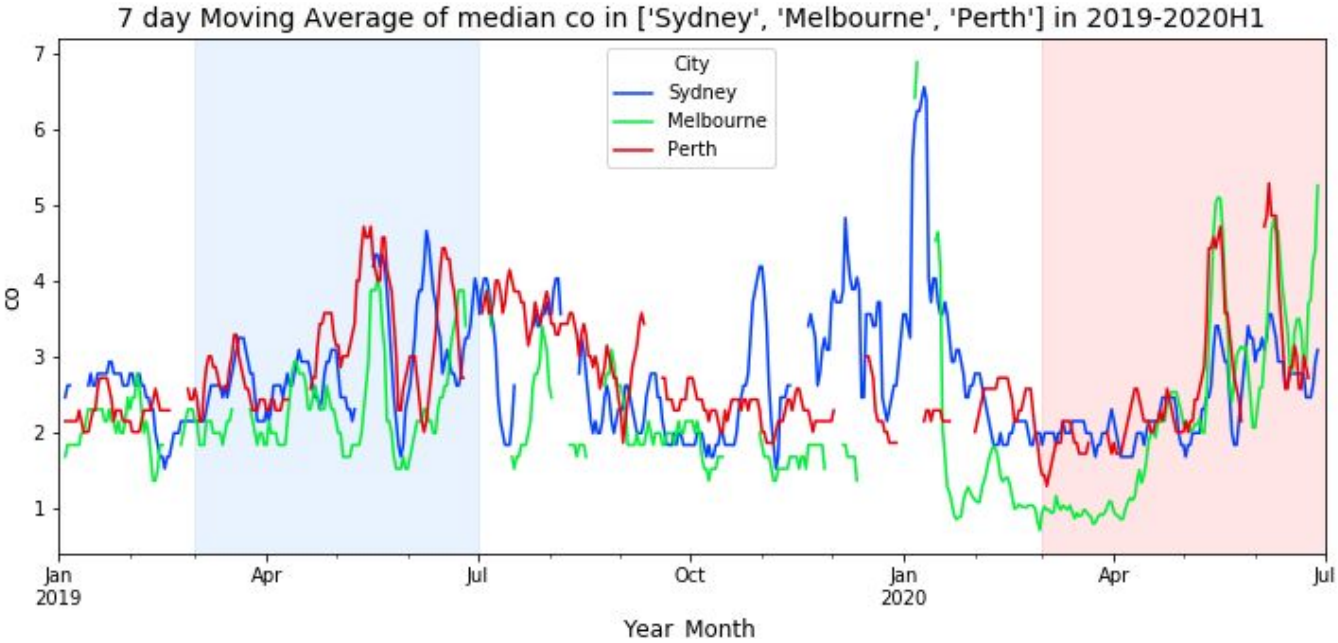

Monthly Distribution of Median pm25 in 94 countries in 2019-2020H1
pm25
Year_Month
0
200
400
600
800
1000
2019-01
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07
2019-08
2019-09
2019-10
2019-11
2019-12
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07


Monthly Distribution of Median co in 94 countries in 2019-2020H1
co
Year_Month
0
100
200
300
400
500
2019-01
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07
2019-08
2019-09
2019-10
2019-11
2019-12
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07

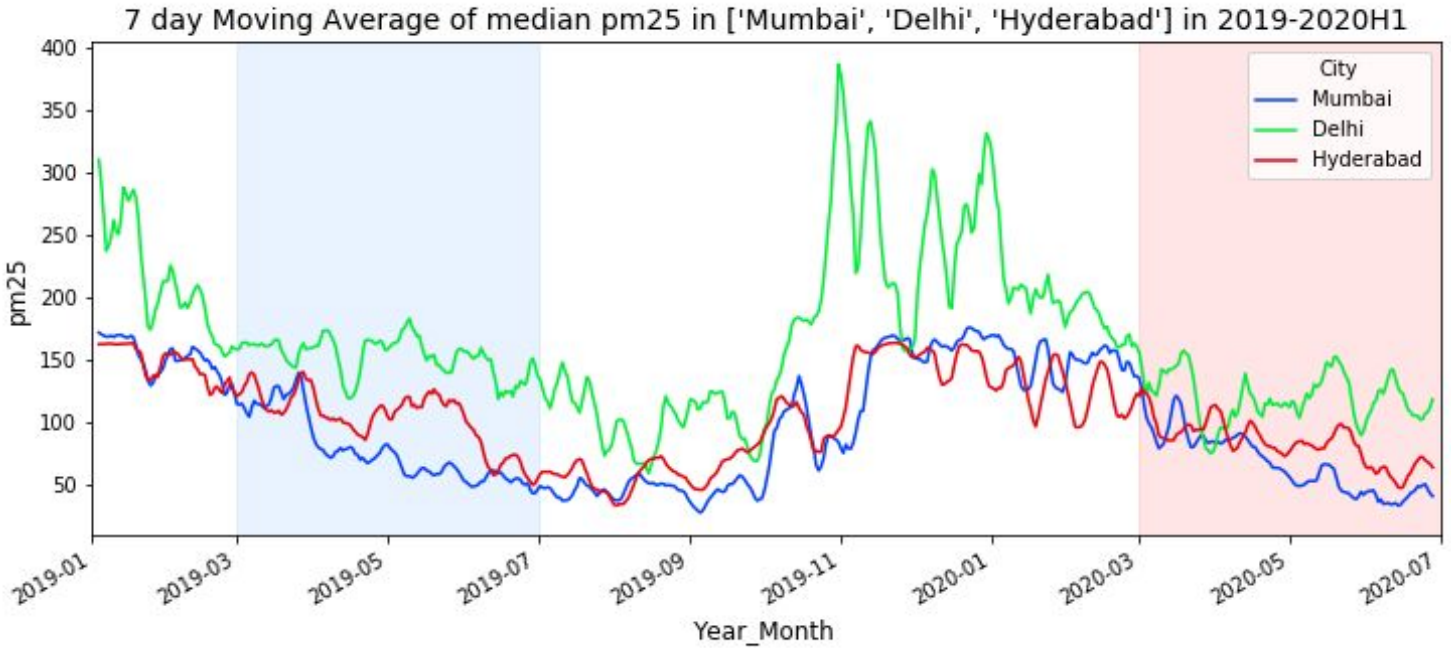
How has PM2.5 changed in Sydney, Melbourne and Perth?



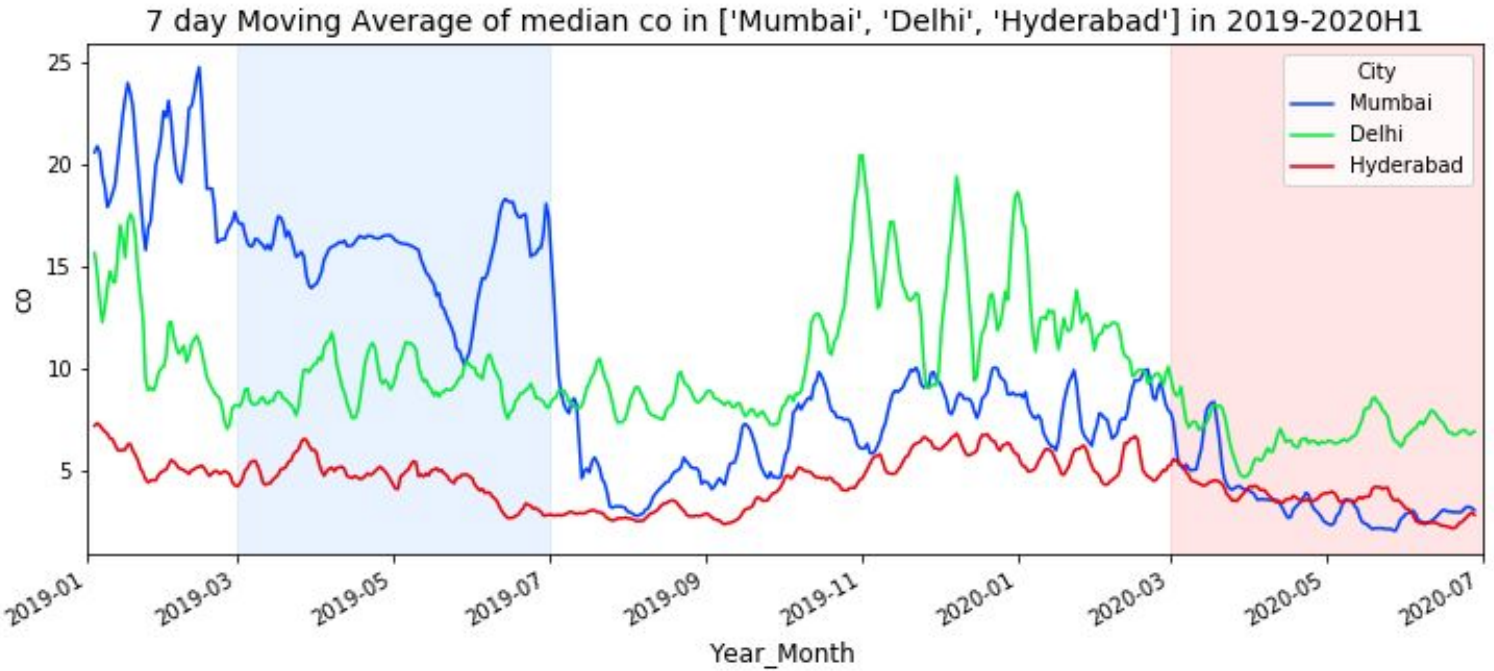
How has CO changed in Sydney, Melbourne and Perth?



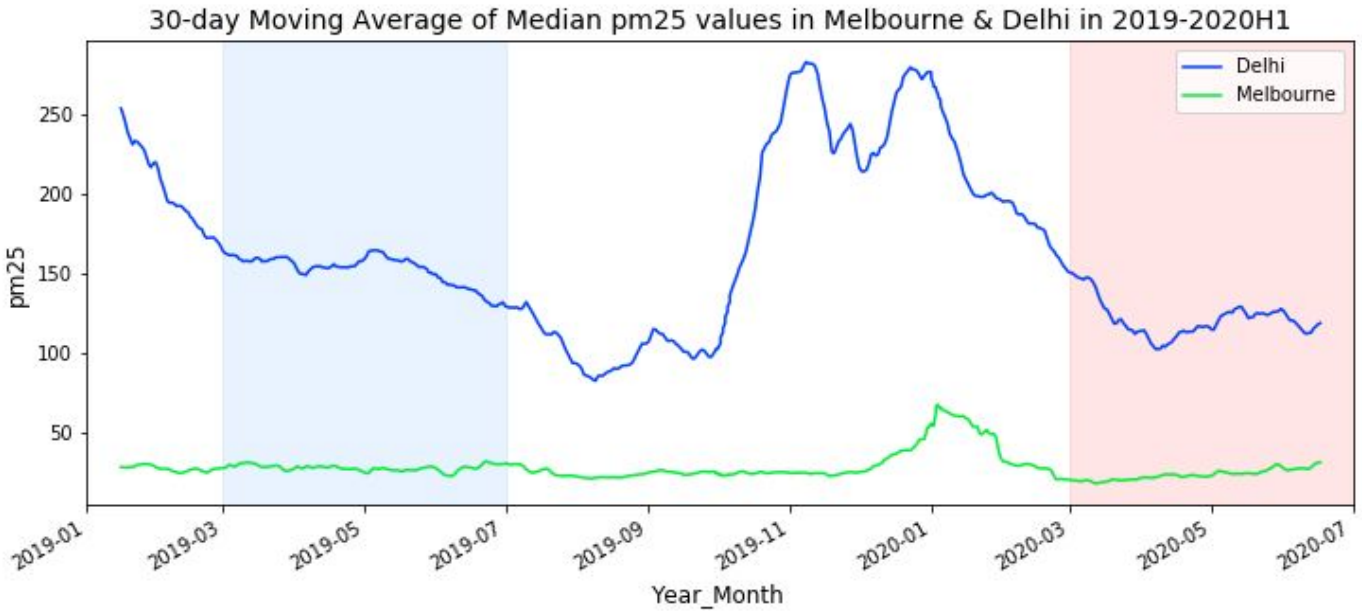
How has PM2.5 changed in Mumbai, Delhi and Hyderabad?



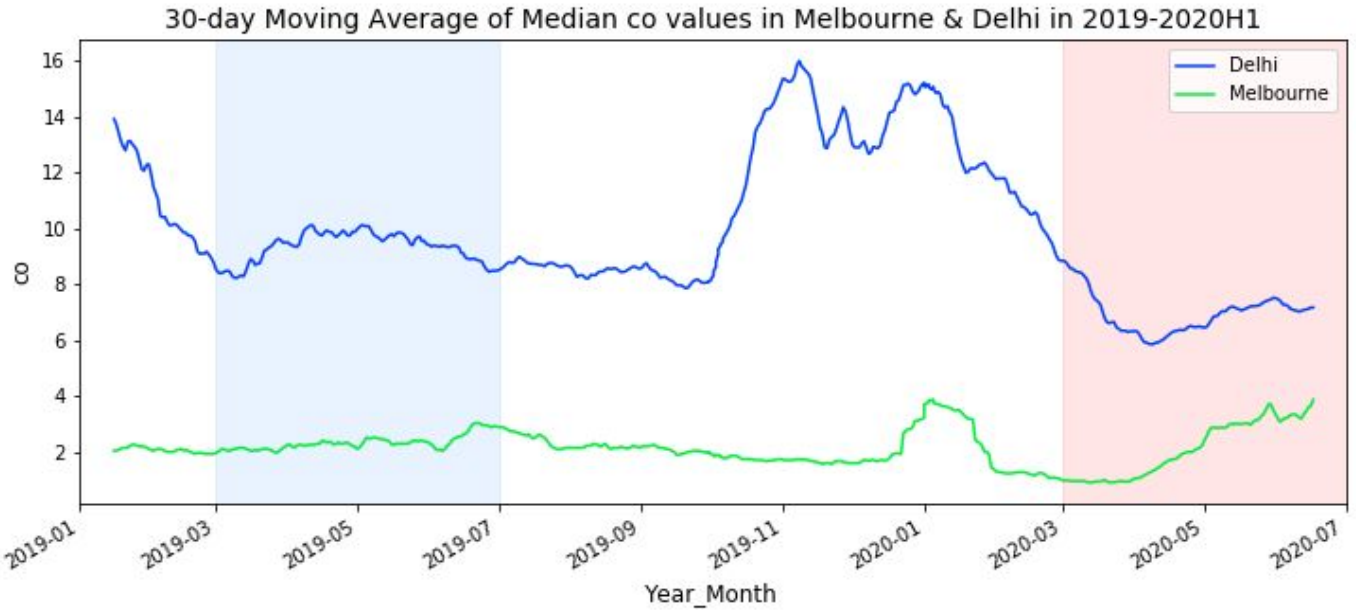
How has CO changed in Mumbai, Delhi and Hyderabad?



Comparing PM2.5 level between Melbourne and Delhi?



Comparing CO level between Melbourne and Delhi?



What did we do to get the Covid data?

```
In [65]: # Use the following api to get the country name, country code and slug
country_url = "https://api.covid19api.com/countries"
country_covid_data = requests.get(country_url).json()
pprint(country_covid_data)
```

```
[{'Country': 'Nigeria', 'ISO2': 'NG', 'Slug': 'nigeria'},
 {'Country': 'Slovenia', 'ISO2': 'SI', 'Slug': 'slovenia'},
 {'Country': 'Antarctica', 'ISO2': 'AQ', 'Slug': 'antarctica'},
 {'Country': 'Bolivia', 'ISO2': 'BO', 'Slug': 'bolivia'},
 {'Country': 'Guinea-Bissau', 'ISO2': 'GW', 'Slug': 'guinea-bissau'},
 {'Country': 'Monaco', 'ISO2': 'MC', 'Slug': 'monaco'},
```

What did we do to get the Covid data?

In [67]:
```
# The country code is used to merge with the air quality data (with country code only)
final_air_df = pd.merge(clean_airdf, country_covid_df, how="inner",
                        left_on="Country", right_on="ISO2", suffixes=("_code", "_name"))
final_air_df.head(10)
```

Out[67]:

	Date	Country_code	City	Specie	count	min	max	median	variance	Country_name	Slug	ISO2
0	2020-02-24	IR	Isfahan	pm25	129	54.0	194.0	126.0	10921.40	Iran, Islamic Republic of	iran	IR
1	2020-05-07	IR	Isfahan	pm25	168	17.0	168.0	91.0	14014.00	Iran, Islamic Republic of	iran	IR
2	2020-05-28	IR	Isfahan	pm25	127	17.0	115.0	72.0	3558.56	Iran, Islamic Republic of	iran	IR
3	2020-02-20	IR	Isfahan	pm25	113	26.0	181.0	76.0	11209.80	Iran, Islamic Republic of	iran	IR

What did we do to get the Covid data?

In [70]:
```
# Explore one covid API - By Country Total All Status
covid_url_example = "https://api.covid19api.com/total/country/australia"
covid_data_example = requests.get(covid_url_example).json()
pprint(covid_data_example)
```

```
[{'Active': 0,
  'City': '',
  'CityCode': '',
  'Confirmed': 0,
  'Country': 'Australia',
  'CountryCode': '',
  'Date': '2020-01-22T00:00:00Z',
  'Deaths': 0,
  'Lat': '0',
  'Lon': '0',
```

Snapshot of the Covid Dataframe

In [73]:
```
# Create the world covid dataframe
covid_df = pd.DataFrame({
    "Country": country_list,
    "Date": date_list,
    "Active cases": active_list,
    "Confirmed cases": confirmed_list,
    "Recovered cases": recovered_list,
    "Deaths": deaths_list
})
covid_df.head()
```

Out[73]:

	Country	Date	Active cases	Confirmed cases	Recovered cases	Deaths
0	Iran, Islamic Republic of	2020-01-22T00:00:00Z	0	0	0	0
1	Iran, Islamic Republic of	2020-01-23T00:00:00Z	0	0	0	0
2	Iran, Islamic Republic of	2020-01-24T00:00:00Z	0	0	0	0
3	Iran, Islamic Republic of	2020-01-25T00:00:00Z	0	0	0	0
4	Iran, Islamic Republic of	2020-01-26T00:00:00Z	0	0	0	0

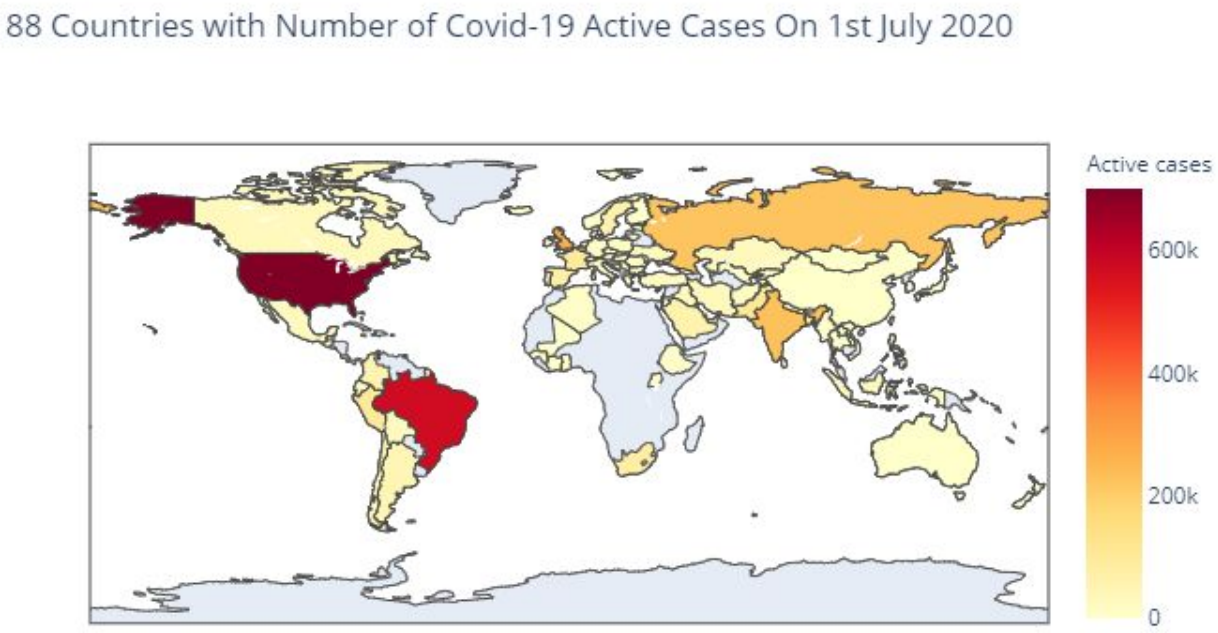
Snapshot of the air dataframe sharing the same country with the Covid dataframe

In [84]:
```
del final_clean_air_df["Slug"]
final_clean_air_df.head()
```

Out[84]:

	Date	Country_code	City	Specie	count	min	max	median	variance	Country_name
0	2020-02-24	IR	Isfahan	pm25	129	54.0	194.0	126.0	10921.40	Iran, Islamic Republic of
1	2020-05-07	IR	Isfahan	pm25	168	17.0	168.0	91.0	14014.00	Iran, Islamic Republic of
2	2020-05-28	IR	Isfahan	pm25	127	17.0	115.0	72.0	3558.56	Iran, Islamic Republic of
3	2020-02-20	IR	Isfahan	pm25	113	26.0	181.0	76.0	11209.80	Iran, Islamic Republic of
4	2020-02-23	IR	Isfahan	pm25	132	22.0	132.0	76.0	3209.67	Iran, Islamic Republic of

Overview of the 88 countries covered in our analysis

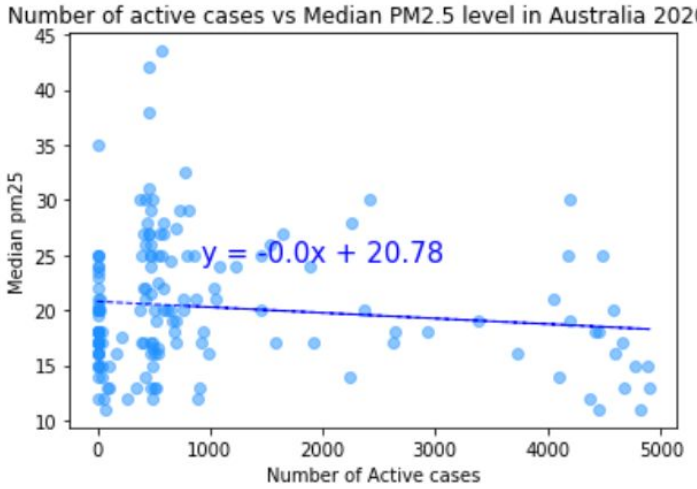


Analysis and Main Findings

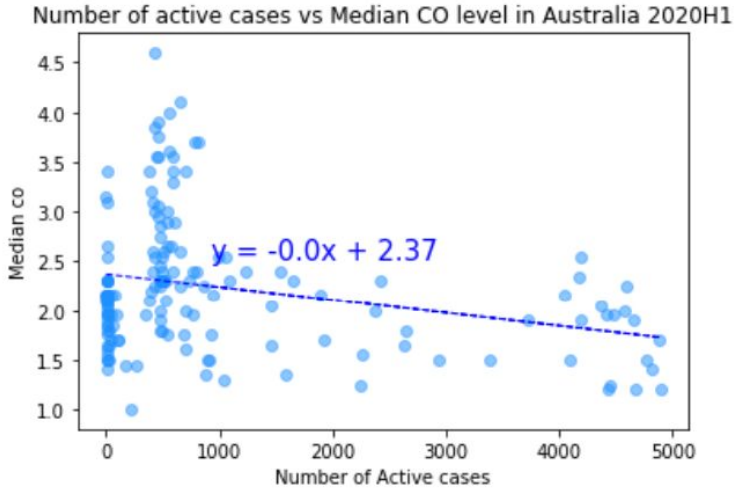
Australia

R-value indicates very weak to no strength of correlation for air quality measures PM2.5 and CO

```
The correlation coefficient is: -0.12186207920586578
The r-squared is: 0.014850366348376704
```



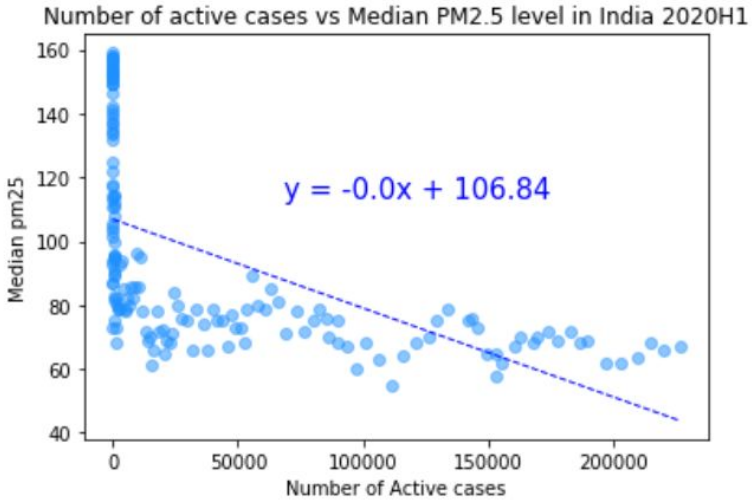

```
The correlation coefficient is: -0.27682230150022347
The r-squared is: 0.07663058660788062
```



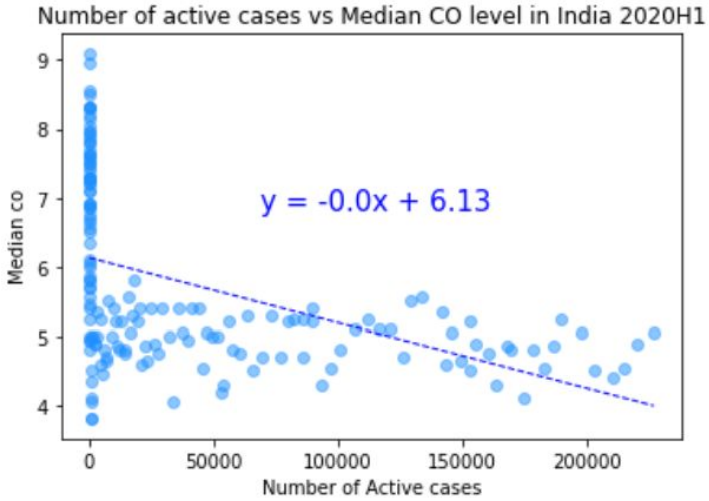
India

R-value indicates a weak strength of correlation for air quality measures PM2.5

```
The correlation coefficient is: -0.5590992737953021
The r-squared is: 0.3125919979584342
```



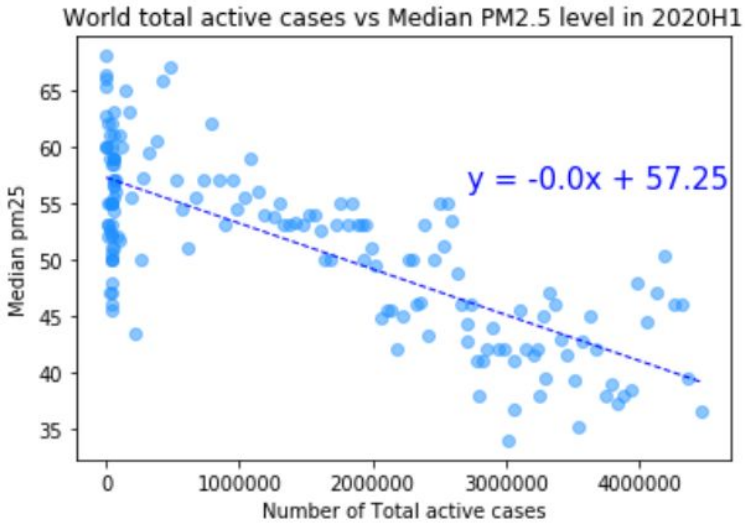

```
The correlation coefficient is: -0.46327863178099465
The r-squared is: 0.21462709066487043
```



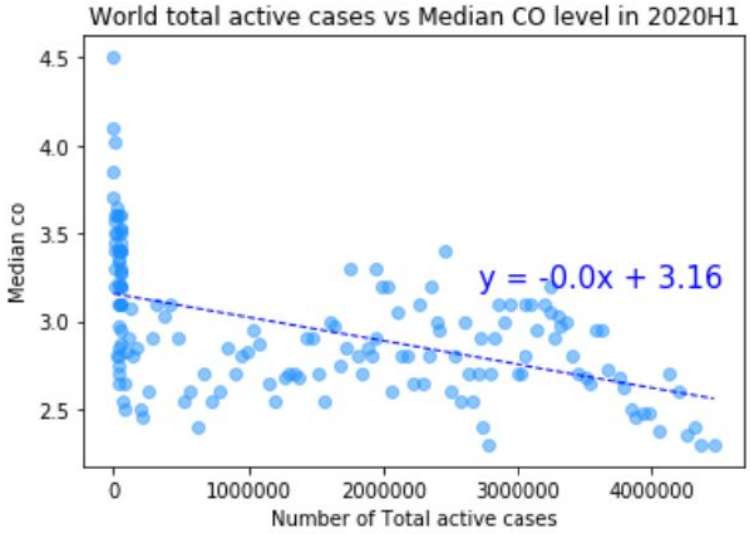
World

R-value indicates a moderate strength of correlation for air quality measures PM2.5

```
The correlation coefficient is: -0.7663059546216506
The r-squared is: 0.5872248160885992
```




```
The correlation coefficient is: -0.5036229211032536
The r-squared is: 0.253636046660574
```




Limitations &
Future Considerations

Project Limitations

- Limited meteorological knowledge prevents us from identifying root cause of outliers in air quality dataset
- Analysis is based on a short time frame (2019-2020H1) with no historical comparison
- Weather conditions and seasonal changes play a key role in air quality
  - Strong winds disperse air pollutants
  - Colder weather (stable atmospheric conditions) increases concentration of air pollutants
- Are active cases the most accurate predictor for the virus' impact on air quality? (i.e. cumulative case count, daily case count)

More Time = More Research

1. Narrow down analysis to countries that have similar seasonal cycles (Northern V. Southern Hemisphere)

2. Do a comparison between the impact of Covid-19 on high population cities of developed countries vs. developing countries (i.e. Europe V. Asia)

3. Include time series analyses (Pre Covid-19 till present) to allow us to account for the impact of seasonal weather changes

References

(1) https://learn.kaiterra.com/en/air-academy/how-does-the-laser-egg-measure-pm2.5#:~:text=How%20Is%20PM2..express%20the%20amount%20of%20PM2.

(2) https://katestone.global/air-quality-covid-19-update2/