Fighting COVID-19 using Data Science.

1. Introduction:

1.1. Background:

Coronavirus has now infected more than four million people globally, according to data collected by Johns Hopkins University. Many countries have been in lockdown since at least March, but some are beginning to ease restrictions. That must be done carefully, the World Health Organization warns, otherwise it risks a resurgence of infections. Nonetheless, billions remain largely at home, and many are struggling with the economic and social consequences. India remains in lockdown, first introduced on 24 March, but some restrictions have been eased. Around 122 million are believed to have lost their jobs in April, and many say they will starve if they cannot work. India's worst affected areas include the financial hub of Mumbai, capital New Delhi, southern state of Tamil Nadu and the western state of Gujarat

1.2. Business Understanding/Problem Description:

India says the number of total recoveries has outstripped active Covid infections for the first time. The health ministry said data showed that more people had been discharged than new infections recorded. The news is being cautiously welcomed in local media, but it comes amid concerns that picture is bleaker. A significant spike in infections in recent weeks has begun taking a toll on the healthcare system - and though India has ramped up testing, it is not uniform across the country, with some states testing much more than others.

To reduce the load on hospitals, the government of India intends to make COVID-19 tests easily available to the public, by opening temporary clinics for testing. For this project, we will be restricting ourselves to the worst affected region of Maharashtra: Mumbai. Our goal will be to identify the perfect locations for these test clinics, to ensure the ease of accessibility for the citizens of Mumbai. These clinics will be set up in locations which are devoid of Hospitals nearby.

Note: Certain assumptions have been made in this project. One of which being that every hospital in Mumbai provides the test for COVID-19.

1.3. Target Audience:

This project tends to serve the citizens of Mumbai, but the project can be extended to the entirety of India.

2. Data:

- Mumbai Neighbourhood Data: (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai)
- WHO COVID-19 Data: (https://covid19.who.int/info)
- India COIVID-19 Data: (https://github.com/imdevskp/covid-19-india-data/blob/master/complete.csv)
- Hospitals in Mumbai: (using Foursquare API)

2.1. Mumbai Neighbourhood Data:

This dataset contains all the neighbourhoods in Mumbai. It was scraped from the mentioned Wikipedia page. It has the following columns:

- Area
- Location
- Latitude
- Longitude

Below is a snapshot of the dataset:

	Area	Location	Latitude	Longitude
0	Amboli	Andheri,Western Suburbs	19.129300	72.843400
1	Chakala Andheri,	Western Suburbs	19.111388	72.860833
2	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373
3	Four Bungalows	Andheri,Western Suburbs	19.124714	72.827210
4	Lokhandwala	Andheri,Western Suburbs	19.130815	72.829270

2.2. WHO COVID-19 Data:

Dataset Description:

WHO Coronavirus Disease (COVID-19) Dashboard
Data last updated: 2020/6/27, 11:25am CEST

Back to top

Data Download

Daily aggregate case and death count data is available for download as a comma-separated values (CSV) file: Download Data

Users should note that, in addition to capturing new cases and deaths reported on any given day, updates are made retrospectively to correct counts on previous days as needed based on subsequent information received.

Table 1. Daily aggregate case and death count data dictionary

Short field name	Type	Description
Date_reported	Date	Date of reporting to WHO
Country_code	String	ISO Alpha-2 country code
Country	String	Country, territory, area
WHO_region	String	WHO regional offices: WHO Member States are grouped into six WHO regions -- Regional Office for Africa (AFRO), Regional Office for the Americas (AMRO), Regional Office for South-East Asia (SEARO), Regional Office for Europe (EURO), Regional Office for the Eastern Mediterranean (EMRO), and Regional Office for the Western Pacific (WPRO).
New_cases	Integer	New confirmed cases. Calculated by subtracting previous cumulative case count from current cumulative cases count.*
Cumulative_cases	Integer	Cumulative confirmed cases reported to WHO to date.
New_deaths	Integer	New confirmed deaths. Calculated by subtracting previous cumulative deaths from current cumulative deaths.*
Cumulative_deaths	Integer	Cumulative confirmed deaths reported to WHO to date.

** See "Daily aggregate case and death count data" above for further details on the calculation of new cases/deaths.*

After pre-processing, the result is:

	Date	Month	Country	Country Name	Region	Confirmed	Cumulative Confirmed	Deaths	Cumulative Deaths
0	2020-02-24	February	AF	Afghanistan	EMRO	1	1	0	0
1	2020-02-25	February	AF	Afghanistan	EMRO	0	1	0	0
2	2020-02-26	February	AF	Afghanistan	EMRO	0	1	0	0
3	2020-02-27	February	AF	Afghanistan	EMRO	0	1	0	0
4	2020-02-28	February	AF	Afghanistan	EMRO	0	1	0	0

2.3. India COVID-19 Data:

This dataset was compiled after scrapping and cleaning data on covid-19 in India from https://www.mohfw.gov.in/ website and API provided by https://www.covid19india.org/.

Note: This is pre-prepared data hosted on GitHub by imdevskp: (https://github.com/imdevskp)

Below is a snapshot of the dataset:

	Date	Name of State / UT	Latitude	Longitude	Total Confirmed cases	Death	Cured/Discharged/Migrated	New cases	New deaths	New recovered
76	2020-03-09	Maharashtra	19.7515	75.7139	2	0	0	0	0	0
88	2020-03-10	Maharashtra	19.7515	75.7139	5	0	0	3	0	0
100	2020-03-11	Maharashtra	19.7515	75.7139	10	0	0	5	0	0
113	2020-03-12	Maharashtra	19.7515	75.7139	11	0	0	1	0	0
126	2020-03-13	Maharashtra	19.7515	75.7139	11	0	0	0	0	0

2.4. Hospitals in Mumbai:

Foursquare API was employed to generate a dataset consisting of a list of hospitals in a 5km radius of every neighbourhood in Mumbai.

Below is a snapshot of the dataset:

	Name	Latitude	Longitude
0	Shivam Netram Eye Hospital	19.127722	72.843875
1	Ladies Clinic & Hospital	19.131188	72.841778
2	Ananda hospital	19.127574	72.847758
3	RG hospital	19.126204	72.839954
4	Shushurut Hospital	19.123895	72.842927

3. Methodology:

3.1. Data Collection:

Data was collected and pre-processed as shown above.

3.2. Analytic Approach:

I took two approaches in the project.

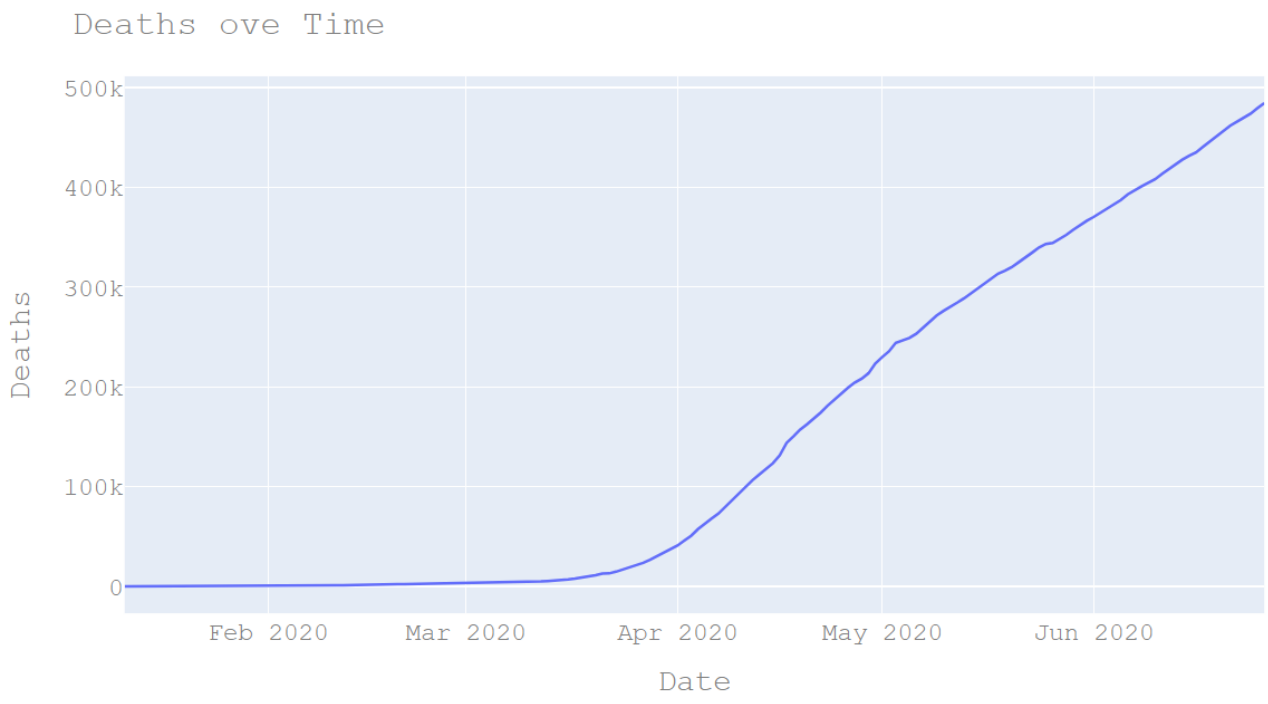
Firstly, I used exploratory data analysis (EDA) to uncover hidden properties of data and provide useful insights to the reader. To gain an insight in the present situation of the World and India, I tried visualizing the WHO COVID-19 Data and the India COVID-19 Data.

Secondly, I used prescriptive analytics to help the government decide a location for a test clinic. I will employ a model based on K-means Clustering. With the neighbourhoods as centroids, I propose to cluster each hospital based on a fixed radius. This will in turn show us the number of hospitals in each neighbourhood, which will help us identify neighbourhoods which lack medical facilities.

4. Analysis:

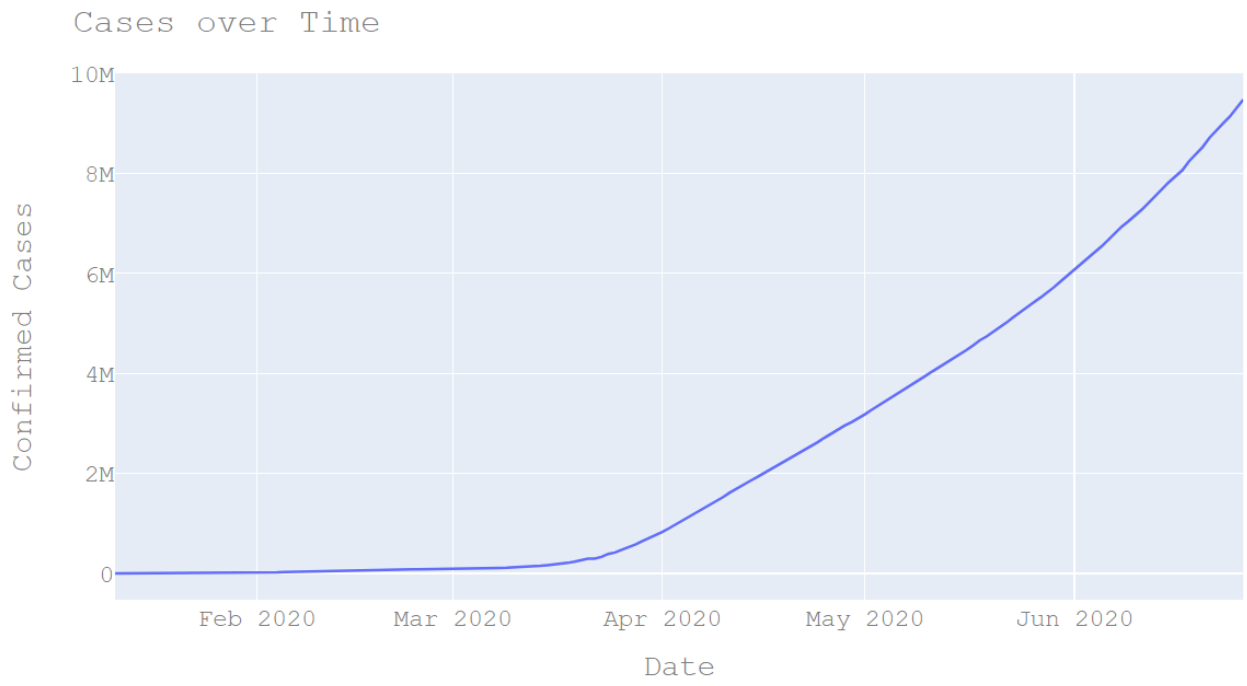
4.1. Exploratory Data Analysis:

Let's take a brief look on the state of the world. The WHO started documenting since the beginning of January. All my findings are true as of 27th June 2020. I used Plotly, a data visualization package which provides interactive tools to visualize data.



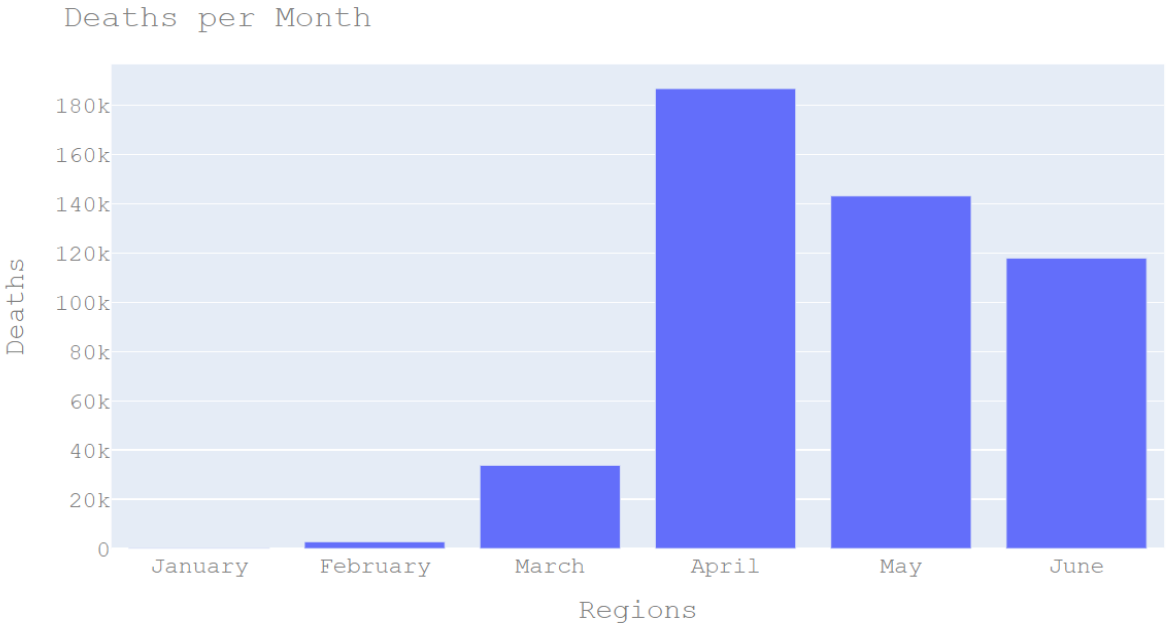
We can clearly see a rise in deaths and cases from the month of March. This article shows us how our world changed in just a month.

"BBC News - Coronavirus: The month everything changed" (https://www.bbc.com/news/stories-52066956)



As of 27th June 2020, the cumulative cases reported are close to 9.48 million, while the total deaths is 485,000. These numbers are staggering, showing us the severity of this pandemic. This article compares the current Coronavirus outbreak to Ebola and other major outbreaks.

"*How bad is Coronavirus versus the flu and Ebola?*" – (https://www.nationalgeographic.com/science/2020/02/graphic-coronavirus-compares-flu-ebola-other-major-outbreaks/)

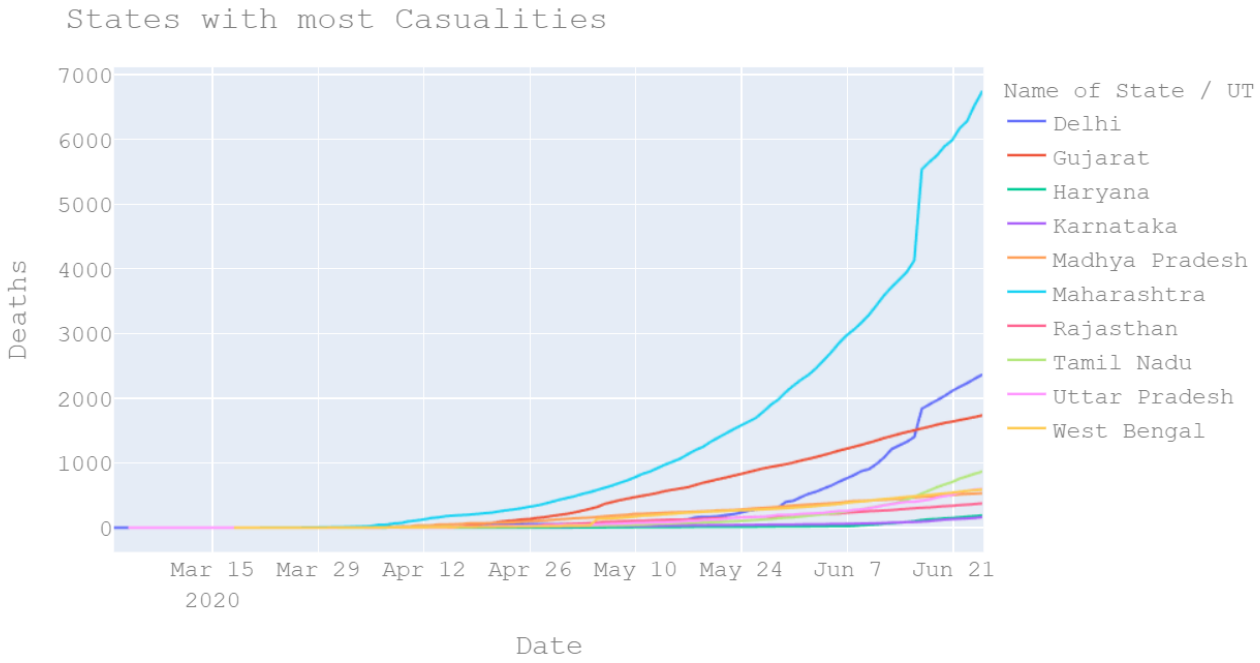


The month of April saw the most deaths around the world, with numbers close to 187,000. It was during this month when the world come to a stop. Almost every manufacturing and retail sector came to a halt. The whole world went into lockdown, to maintain social distancing. With the absence of a cure, the only sane decision was to stay locked up in your own homes.

During the months of April and May, India was in it's 3rd Lockdown, dubbed as Lockdown 3.0.

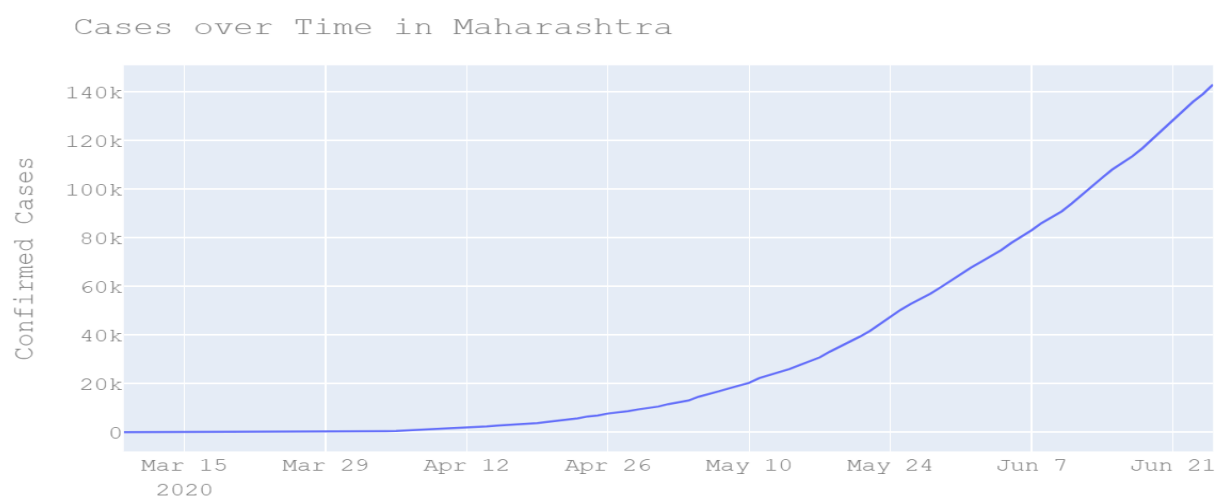
"Coronavirus: India lockdown extended for two more weeks": (https://www.bbc.com/news/world-asia-india-52505436)

Let's shift our focus to the country of interest: India.



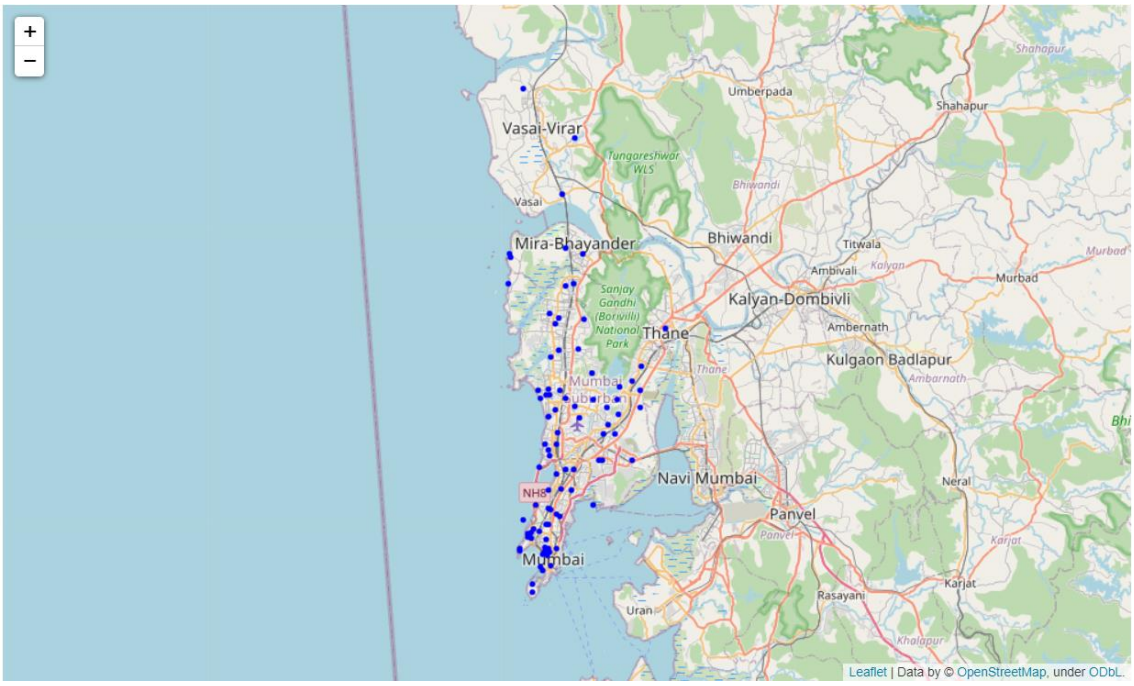
The state of Maharashtra has clearly surpassed others in number of casualties, with the number standing at 6739. It's capital, Mumbai has become the hotspot for new cases daily.

"Mumbai overtakes Wuhan peak as India Covid cases spike": (https://www.bbc.com/news/world-asia-india-52989452)



4.2. Neighbourhood Clustering:

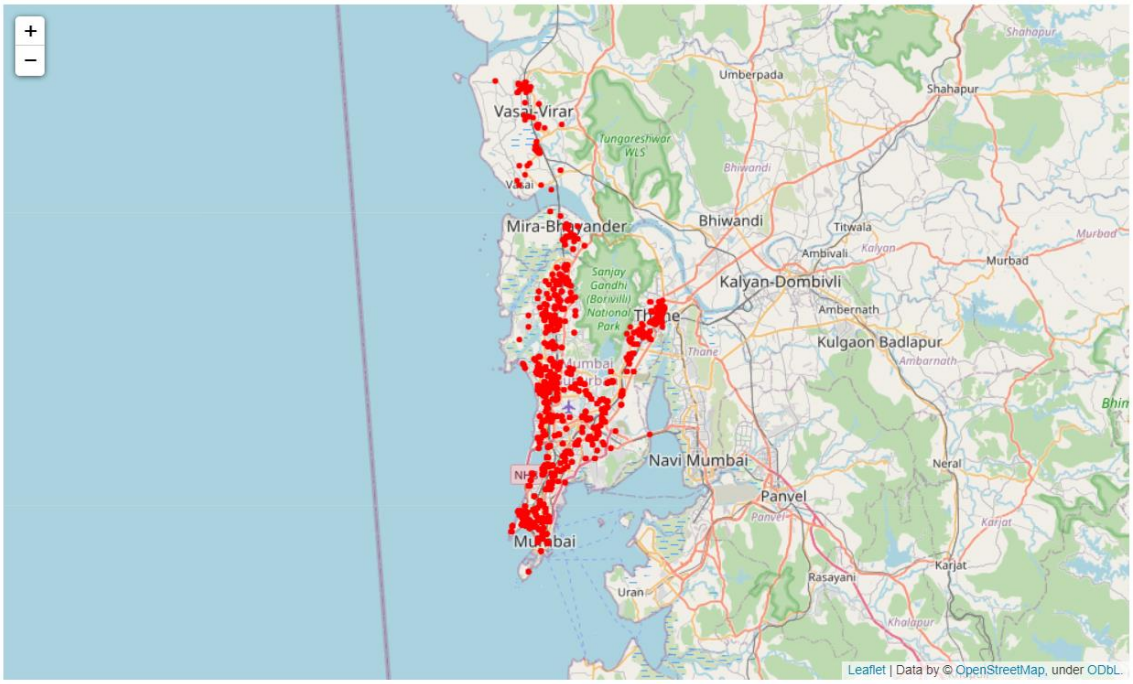
Before Clustering, lets take a look at all the neighbourhoods in Mumbai.



Mumbai is a very densely populated city. Hence, scarcity of medical facilities is always a threat during a pandemic. Take a look at this article:

*"Patients share beds in overrun Mumbai hospital" : (*https://www.bbc.com/news/av/world-asia-india-52818777/india-coronavirus-patients-share-beds-in-overrun-mumbai-hospital)

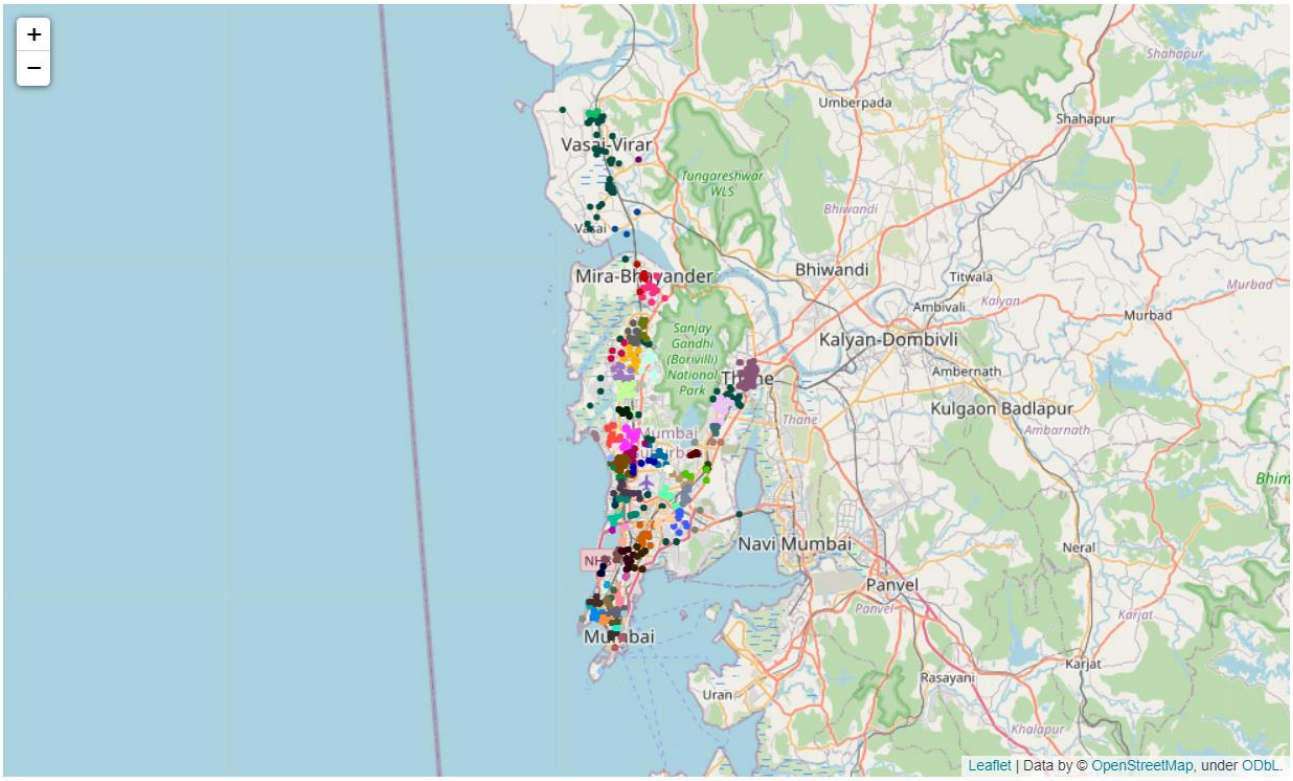
Hence certain measures must be taken to reduce the load on every hospital in Mumbai. Now let's take a look at all the hospitals in these neighbourhoods.



Our goal is to identify neighbourhoods with no/fewer medical facilities in a 2km radius from the city centre. I employed the following function to calculate the distance between a hospital and a city-centre using the *Haversine Formula*. Then I associated each hospital to a neighbourhood(s) if it was in a 2km radius around the city. This helped identifying neighbourhoods devoid of an accessible medical facility.

```
points = h[['Latitude','Longitude']]
centroids = neighborhoods[['Latitude','Longitude']]
cluster = []
R = 6373.0
for i in range(4243):
    m = 2000
    flag = -1
    for j in range(93):
        lat1 = math.radians(points['Latitude'][i])
        lon1 = math.radians(points['Longitude'][i])
        lat2 = math.radians(centroids['Latitude'][j])
        lon2 = math.radians(centroids['Longitude'][j])
        dlon = abs(lon2 - lon1)
        dlat = abs(lat2 - lat1)
        a = math.sin(dlat / 2)**2 + math.cos(lat1) * math.cos(lat2) * math.sin(dlon / 2)**2
        c = 2 * math.atan2(math.sqrt(a), math.sqrt(1 - a))
        distance = R * c * 1000
        if (distance <= m):
            flag = j
            m = distance
    cluster.append(flag)
```

```
h = h.join(pd.DataFrame({"Cluster":cluster}))
```



The biggest problem I faced till now was to depict all the 73 different clusters formed by my function, as it is visually impossible to differentiate them using colours. I found out this was a major problem many Data Scientists faced.

Here is a great article composed by Cambridge Intelligence pertaining to this problem : (https://cambridge-intelligence.com/choosing-colors-for-your-data-visualization/)

5. Results:

The results showed 7 neighbourhoods had no access to a medical facility in a 2km radius. During a pandemic, immediate medical attention is the key to fighting to fighting the virus.

"Why testing has been slow to take off in India" : (https://timesofindia.indiatimes.com/india/why-covid-19-testing-has-been-slow-to-take-off/articleshow/74859149.cms)

	Neighbourhood	No.Of Facilities
11	Uttan	0
16	Gorai	0
18	Aarey Milk Colony	0
41	Nehru Nagar	0
51	Mahul	0
64	Dongri	0
82	Hindu colony	0

By establishing temporary test clinics in these areas, the government could reduce the load on hospitals and also prevent secondary transmission of the virus. Adding to the list, we have 2 other neighbourhoods (Naigaon & Nalasopara) with less than 5 medical facilities.

6. Discussion:

One of the main assumptions made in this project was that every hospital provided a test for the COVID-19 virus. This is certainly not the case. Also a few drawbacks of this analysis are that it doesn't consider the feasibility of establishing a test centre in one of these neighbourhoods: terrain could play a big role. Population density could also help us in determining the number of test-centres each neighbourhood would need.

7. Conclusion:

In the above study, we explored and analysed the medical facilities in Mumbai, India using Data Science. We used an existing dataset and combined it with data collected from Foursquare API as well as data scraped from a website. We performed EDA and clustering on these datasets in our pursuit of solutions. We were able to find satisfactory answers to the questions we posed before the study. The study is based on limited data, but it is nevertheless a significant step in shedding light on the medical infrastructures in Mumbai. This study can be repeated easily for other cities in India.

"Coronavirus: The last 'normal' photo on your phone" – (https://www.bbc.com/news/uk-52622673)