PREDICTIVE MEASURES FOR MITIGATION OF COVID-19 COMMUNITY TRANSMISSION

Manik Kapil

April 27, 2020

1. INTRODUCTION

1.1 BACKGROUND

Coronaviruses are family of viruses that cause diseases generally in birds and animals, which in some cases get transmitted to humans, like the one the whole world is currently facing. COVID-19 or Coronavirus Disease 19 is the name WHO (World Health Organization) coined for this new disease which as of 26 April 2020, consumed at least 203,705 lives and more than 2,935,876 confirmed cases. It causes respiratory tract infections that can range from mild to lethal. India being world's second most populous country with its high population density was at high risk if the contagion spread. A pre-emptive 21 days lockdown and then extended 19 days lockdown indeed helped in containing the rate of spread of the virus. IMA (Indian Medical Association) & ICMR (Indian Council of Medical Research) have been on their feet even before the first case was recorded in India, still as we speak there are at least 27.3k active cases and as many as 825 deaths and we are only in 2nd Stage of Transmission. To tackle this unprecedented situation, there were multiple novel techniques devised and used, therefore, it is imperative to estimate the existing cluster of cases in the country, investigate the treatment, recovery and deaths in those areas and to explore if there are any predictions can be made for new Quarantine and isolation facilities. A large number of daily wage workers suddenly boggled by this situation find them without food and shelter. We'll also try and predict the areas where migrant workers might take food and shelter while maintaining social distancing.

1.2 PROBLEM

The world has come to a halt amid COVID-19. All economic machineries have come to a standstill and the liquidity in market has dried out. Hence, the breakage of transmission chain and release in economic activity is one following the other. Hence, this project aims at predicting probable locations that can be used as quarantine facilities, food shelters and even testing facilities.

1.3 INTEREST

Every government in the world, every CEO of every company, every worker either daily-wage or salaried, every front-line COVID worker is interested in finding some way that can break the existing transmission chain and slowly the world returns to normalcy. Hence, a large number of people are stake holders in this.

2. DATA ACQUISITION AND CLEANING

2.1. DATA SOURCES

There were multiple datasets used in this notebook. First, the datasets were scrapped, cleaned, merged in some cases and then were used for analysis.

Datasets included:

2.1.1 Daily COVID-19 patient database across Indian states

2.1.2 Case time series database, which included patient data attributed by date

2.1.3 Covid-19 Daily tests dataset

2.1.4 Some other supporting datasets which included, Indian coordinates.

2.2. DATA PRE-PROCESSING

2.2.1 Extensive data cleaning was done as these datasets had a large number of missing values, values misplaced and improper and non-formatted values.

2.2.2 There were a large number of patients that had no information pertaining to their addresses apart from just the state name. Hence, it became difficult in keeping a track of these patients. There were some columns that had all the missing data except a few and we had to alter those columns to make our analysis easier. The only way to plot through folium is by the means of coordinates of a location. Hence, it became imperative to append the coordinates for each state.

2.2.3 The Data format was Timestamp in case time series dataset and there was no way to move forward with this format, hence, I had to alter it down to just the date part to move further. This was achieved with the help of a self-defined function. There were some outliers and some missing values in this dataset. For e.g. there were missing values for 'Daily Confirmed' and 'Daily Recovered' for some dates. Hence, we had to get rid of those to find our desired result.

2.2.4 The daily tests dataset was the most messed-up. With duplicate columns and very few countable values in those had to cleaned. The Time column in this dataset was again given in timestamp format and had to be formatted again for further use. The datatypes of some columns were misplaced and were creating problems in mathematical analysis. Hence, I had to assign proper datatype pertaining to that column. Multiple missing values had to be dealt with.

2.2.5 Finally after these datasets, I used Foursquare data for specific locations to get better insights about current scenario and them made recommendations. Even in this data, the proper address attributes like name of city, state were missing. But since, the coordinates were provided, I was able to draw a map with those locations as markers.

3. METHODOLOGY

The goal here is to aid the ongoing spread of COVID-19 and Mitigate its chances of going into **Community Transmission.** WHO has identified following steps to mitigate and control the spread of the virus:

i.   <u>Testing</u>: Testing! Testing! Testing! As much as possible has been mantra for controlling the spread of virus

ii.   <u>Contact Tracing</u>: Tracing the people that came in contact with the patient that tested positive

iii.   <u>Isolating</u>: Isolating the people contracted and back to step i

Hence, we are going to follow a similar strategy in tackling this situation. Our primary goal is going to be recommendation of location venues that can be used either as, Quarantine/Isolation Facilities, Testing Facilities, Food Production and Distribution centres.

We are assuming that, under extreme conditions, we will use, Hotels, Schools, College Hostels as Quarantine Centres; Restaurants as Food production centres; Community centres as Testing Facilities. Hence, our goal would be to identify these.

Second, the choice of these recommended locations will be based on the virus clusters as identified by Govt. of India. Our idea is as follows:

i.   We identify the districts worst hit by coronavirus and use them as hotspots

ii.   We use K-Means Clustering algorithm to cluster them into similar cluster

iii.   We find the centroid of these clusters and use those coordinates to identify the location of interest using Foursquare API

4. EXPLORATORY DATA ANALYSIS

4.1. <u>Total Confirmed Cases in Indian States</u>

The first step was to identify the worst hit states by COVID-19. Hence, it was imperative to identify the number of 'Confirmed' cases in each of state. This was understood with the help of 'bar graph'. We could see that Maharashtra was worst hit by coronavirus with over 8 thousand total infections.

Total Number of confirmed COVID-19 Cases State-Wise

Fig. 1 Total Confirmed Cases
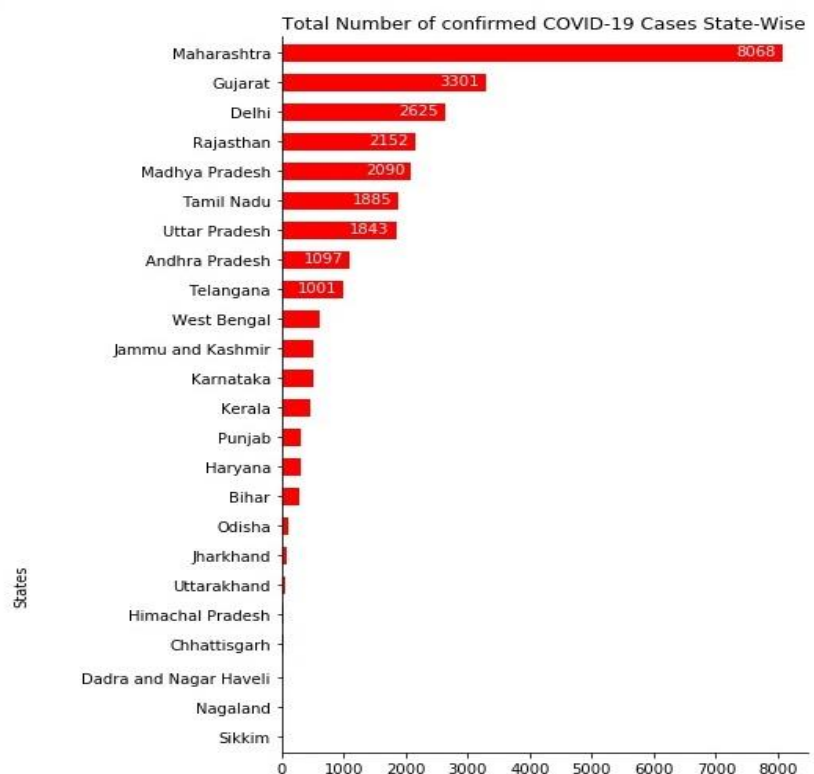
4.2. Total Recovered cases in Indian States

Next, we identified the 'recovery' status of patients across all the states. An important aspect as to how the individual states dealt with the issue and relief came to patients.
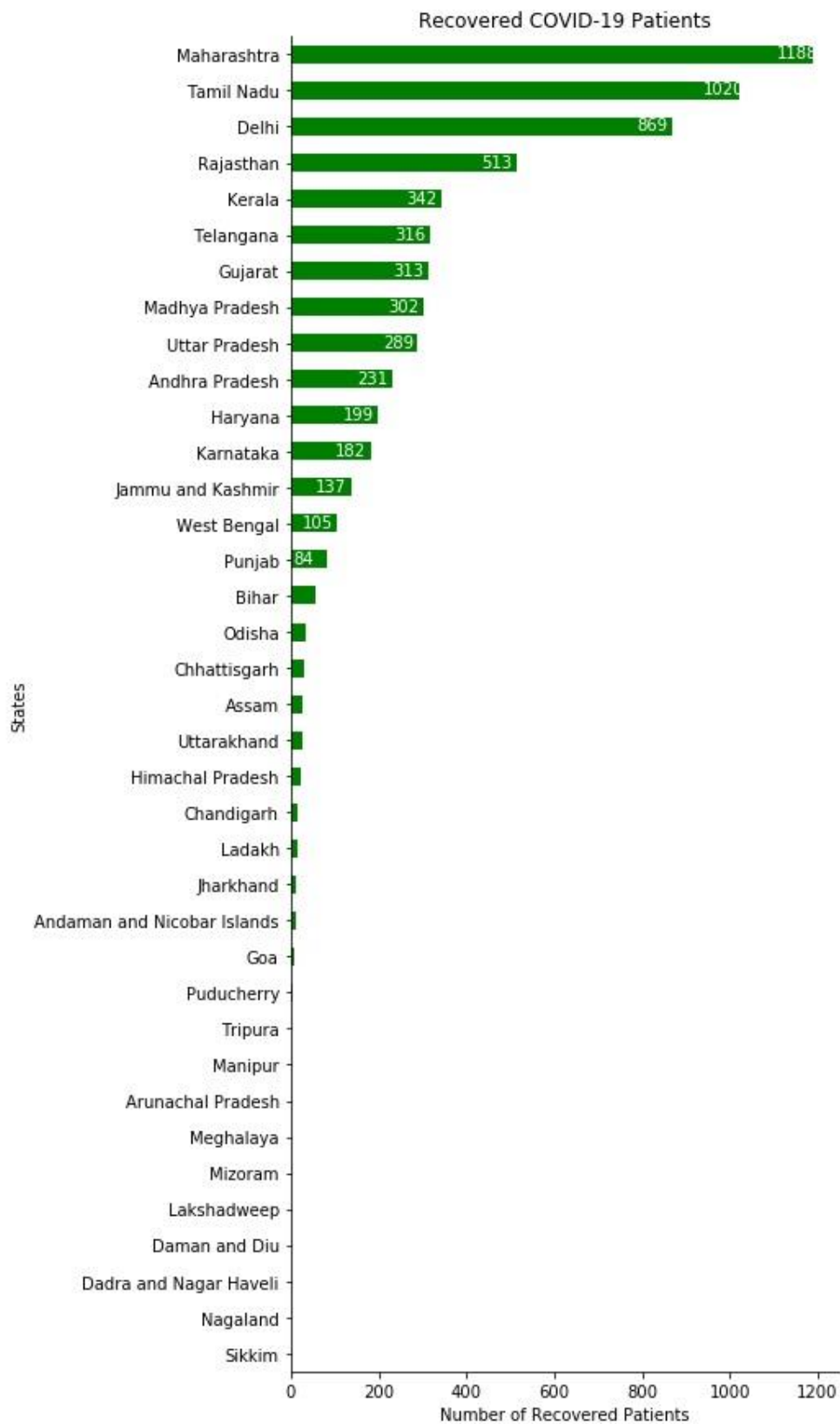

Fig. 2 Recovered Patient Numbers across Indian States

We could see that Maharashtra again has the most recovered cases, followed by Tamil Nadu and Delhi.

4.3. Total Active Cases in Indian States

The real scenario is much more understood by 'Active' patients that are undergoing diagnosis. These can be those that are in ventilators or even those asymptotic ones that show normal body functionality. The major application of looking at active patients is when we are trying to differentiate between multiple treatment techniques used by states to counter patients and their response. Hence, pick the one with best response. Recently, Delhi has recorded a better treatment facility by 'Plasma Therapy'. Maharashtra again records the most Active cases in India. Situation is really critical there.
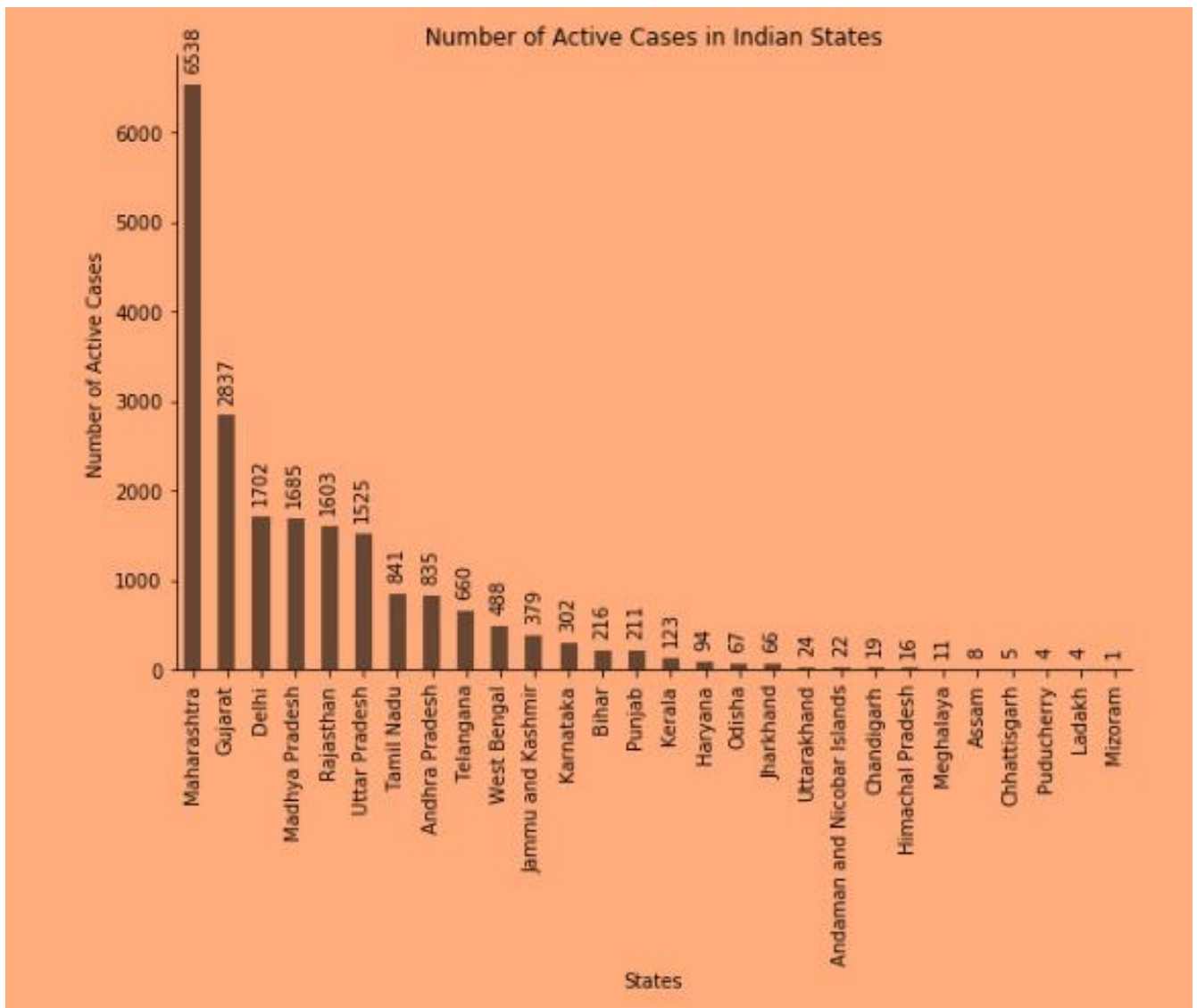

Fig. 3 Active Cases in Indian States

4.4. Total Deaths

Sadly, the whole world has lost a huge number of people to this virus and people are still dying as I write and you read this report because of this deadly coronavirus. From the above figure we can see how many deaths each state has recorded.
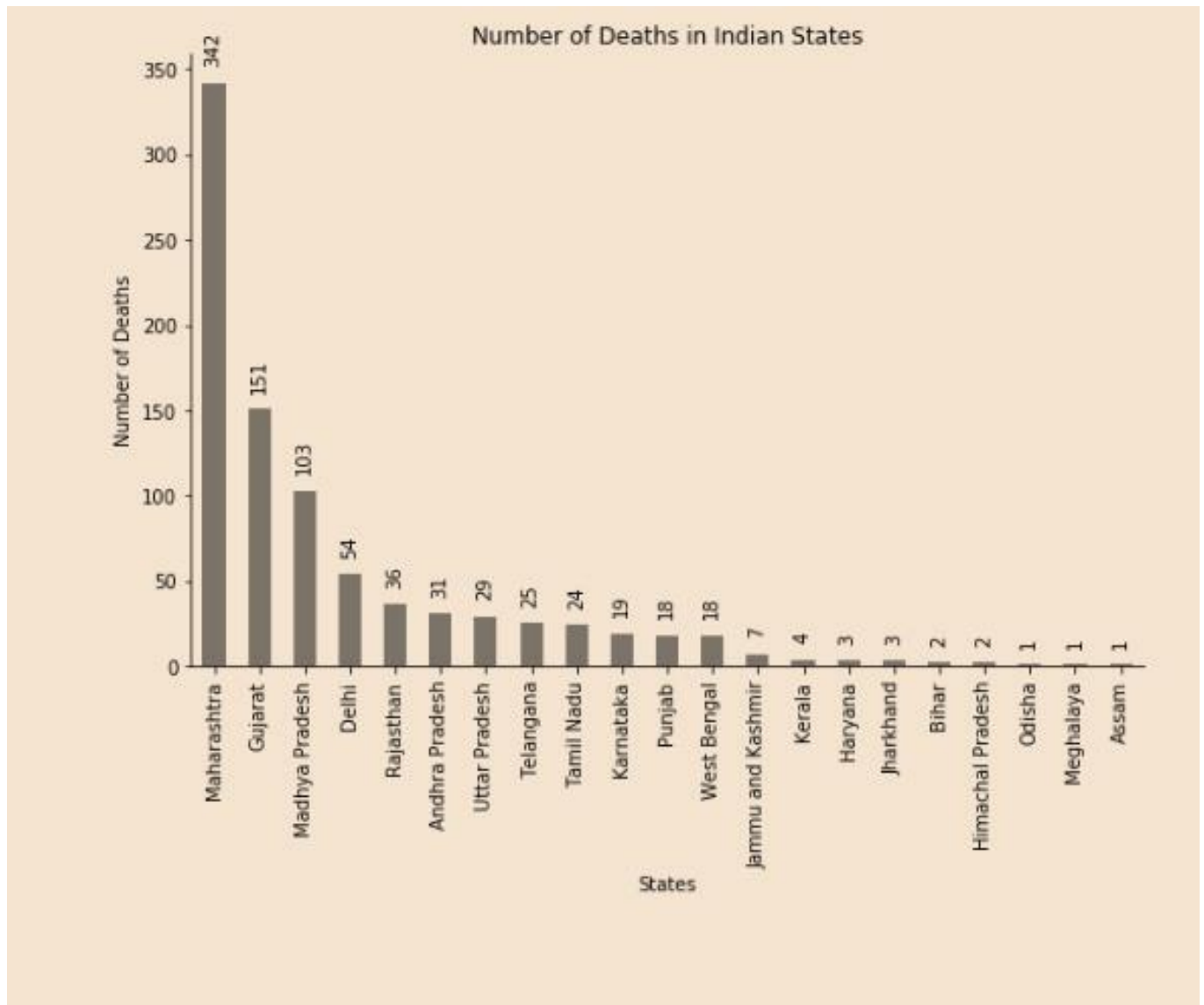

Fig. 4 Deaths in India

4.5. EFFECT OF LOCKDOWN

The Government of India has been pre-emptive in dealing with the crisis. Indian Council of Medical Research has been working on solutions far before the first case in India was recorded. As experts say, the lockdown was well in time and much needed for containment of the spread.

Hence, we are going to answer 2 very important questions:

i.   Is India still in '**Stage II'** of transmission?

ii.   Did lockdown have any impact on spread of coronavirus?

In order to see the current situation which is '**Stage II**' of transmission, where virus is contracted from someone who has a travel history, as per ICMR, we need to visualize with the help of 'Daily Cases' recorded. Also, we will estimate if the lockdown had any effect on spread of virus.
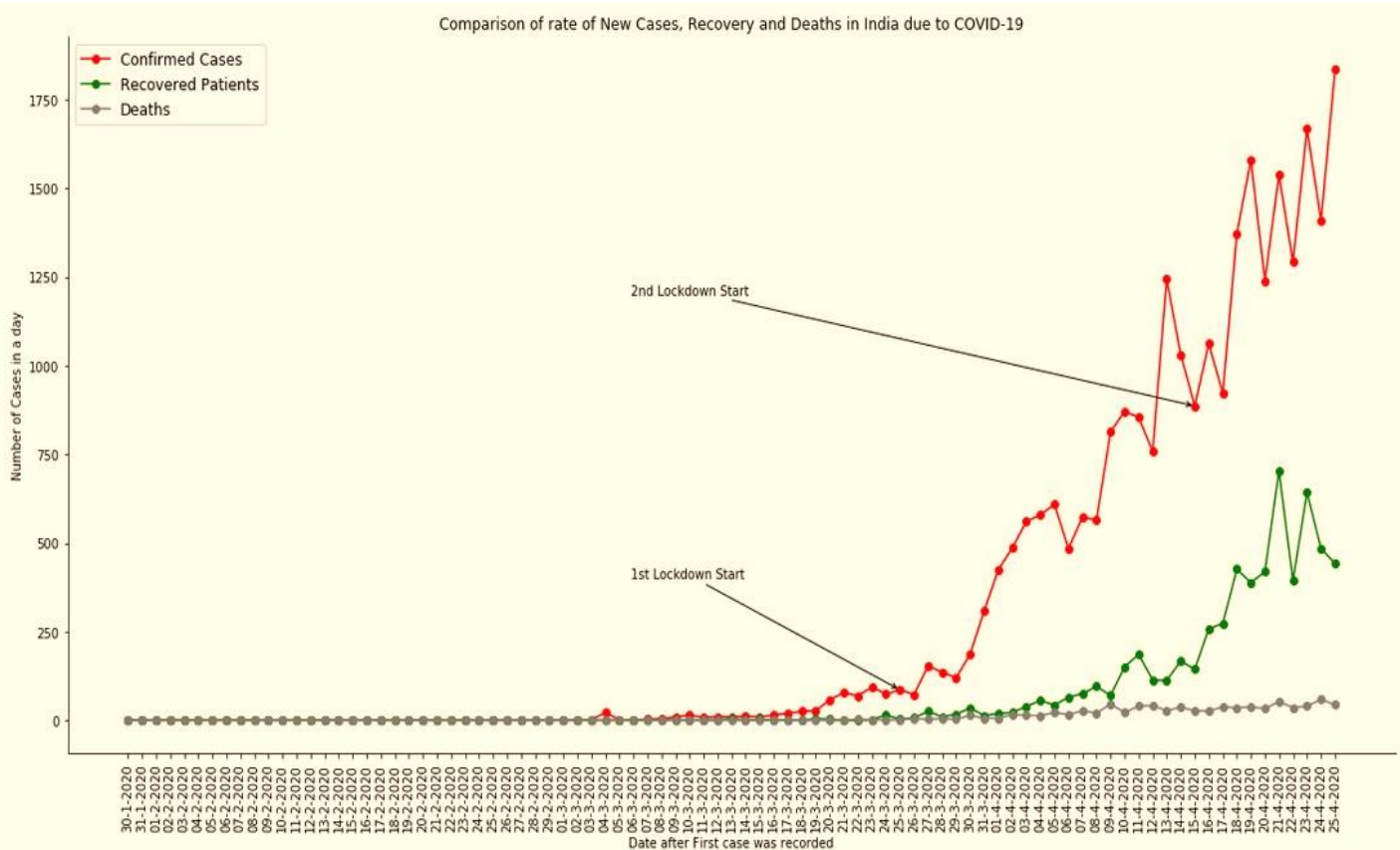

Fig. 5 Daily Coronavirus Cases

From the figure we can observe that post lockdown the daily confirmed cases rose. The sudden rise can be attributed a huge chunk of people found positive in congregation in Nizamuddin, Delhi, India. Post that, the cases rose because the testing before lockdown was not enough. As we can see the cases are still rising, this does not give us a complete wholesome idea as to what was the impact of lockdown. The fact that we capped the maximum daily case to 1800 something shows that we have somehow managed to restrain ourselves in Stage II only as of now.

To get a better view of what is going around we need to look at the testing data for all the dates. Let's visualize these and find the insights.
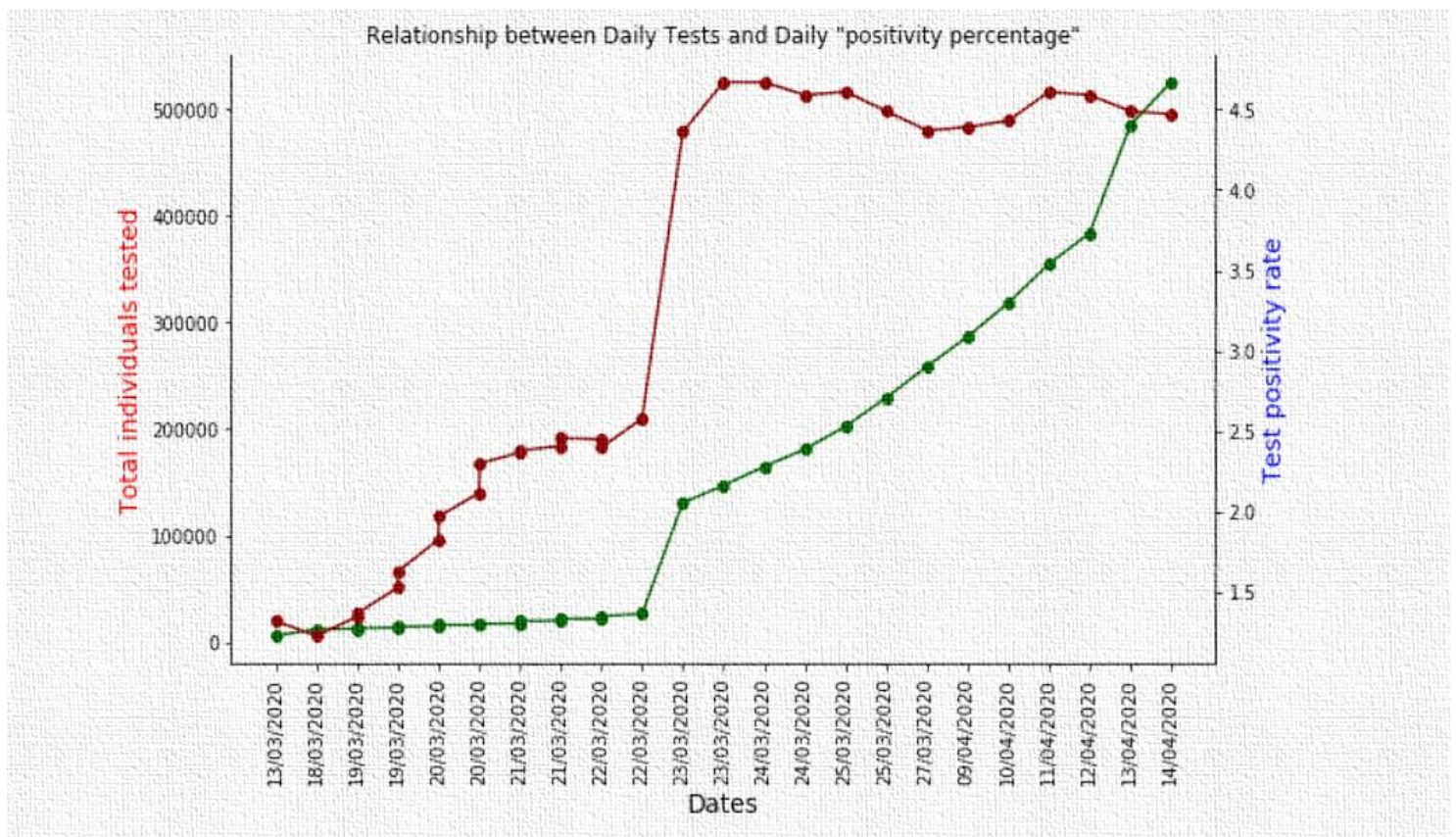

Fig. 6 Total Individuals Tested and Percent Positive

From this figure, we can conclude that testing numbers drastically improved immediately after imposing the first lockdown. So did the percent positive rate. But we can still observe that the increase is almost linear and not exponential. Hence, we can conclude that the lockdown did do some good and evaded the gloomy 'Stage III' of Community Transmission.

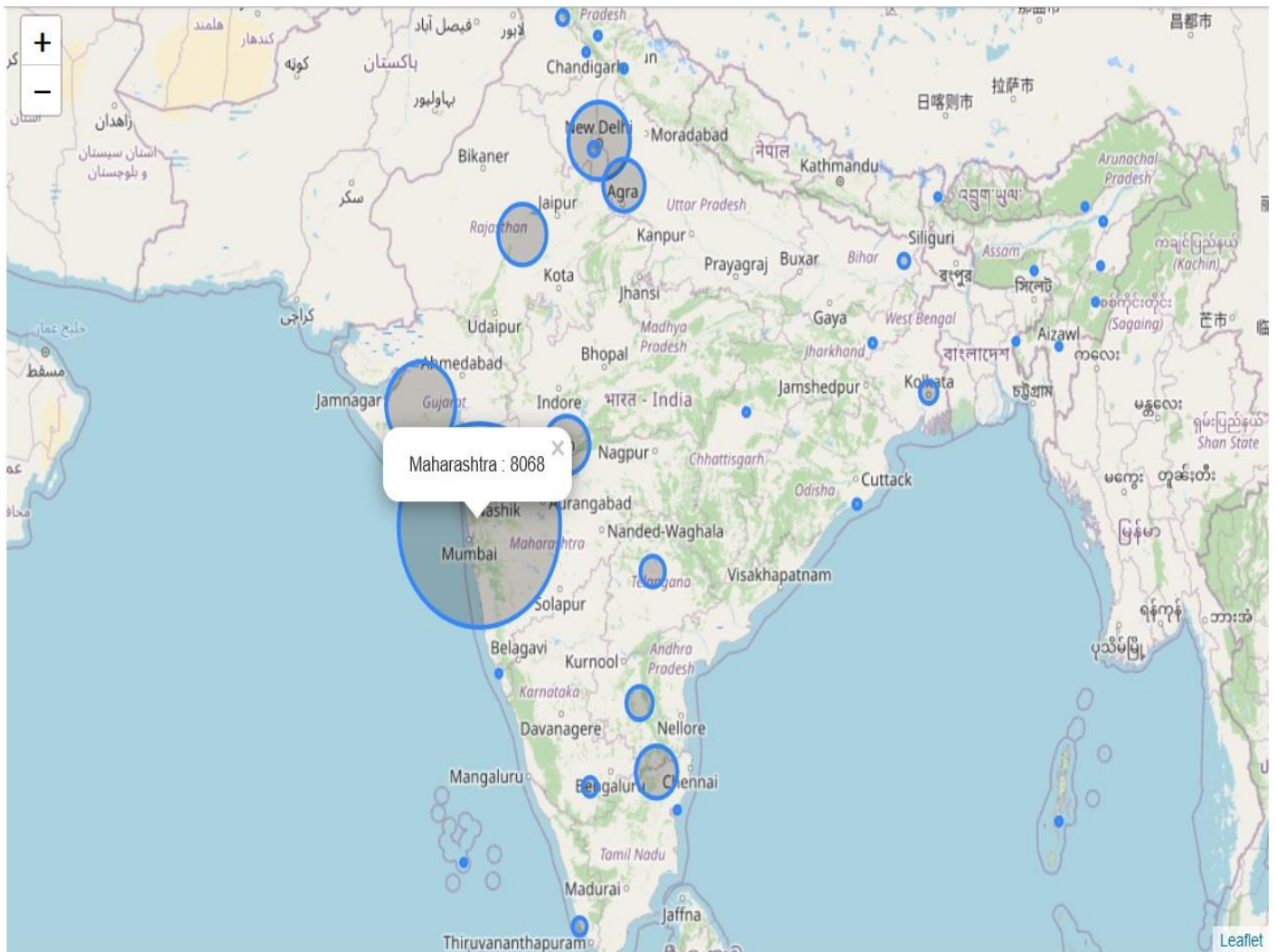Next, we are going to observe the severity of COVID-19 in Indian States, with the help of a map.

Fig 6 Indian States based on Severity of Cases

This figure shows the normalized number of confirmed cases in India States. The radius of the circular marker is in proportion to the number of normalized total confirmed cases in each individual state. Maharashtra is worst hit, in terms of total cases and followed by Gujarat and Delhi.

5. CLUSTERING AND MODELLING

5.1. The next step would be to identify coronavirus hotspots. This was made possible only by accessing a new json dataset, consisting records of, confirmed, active and recovered patients in each district.

5.2. The task of 'Feature Selection' needed location values appended to this dataset, hence, next task was to append the geographical coordinates of individual district.

5.3. The 'Confirmed' cases were normalized again to have remove the bias.

5.4. 'Confirmed Normalized', 'Latitude', 'Longitude' were selected as Features for K-Means Clustering algorithms.

5.5. I selected 100 to be the cluster number since, govt numbers have been showing a similar number of clusters in the country.

5.6. The pre-processing of these features was done by StandardScaler method present in sklearn's preprocessing library.

5.7. Finally, after the model was fit, the cluster numbers or the 'labels' were appended alongside location columns in hotspots dataset and it was observed through a map.

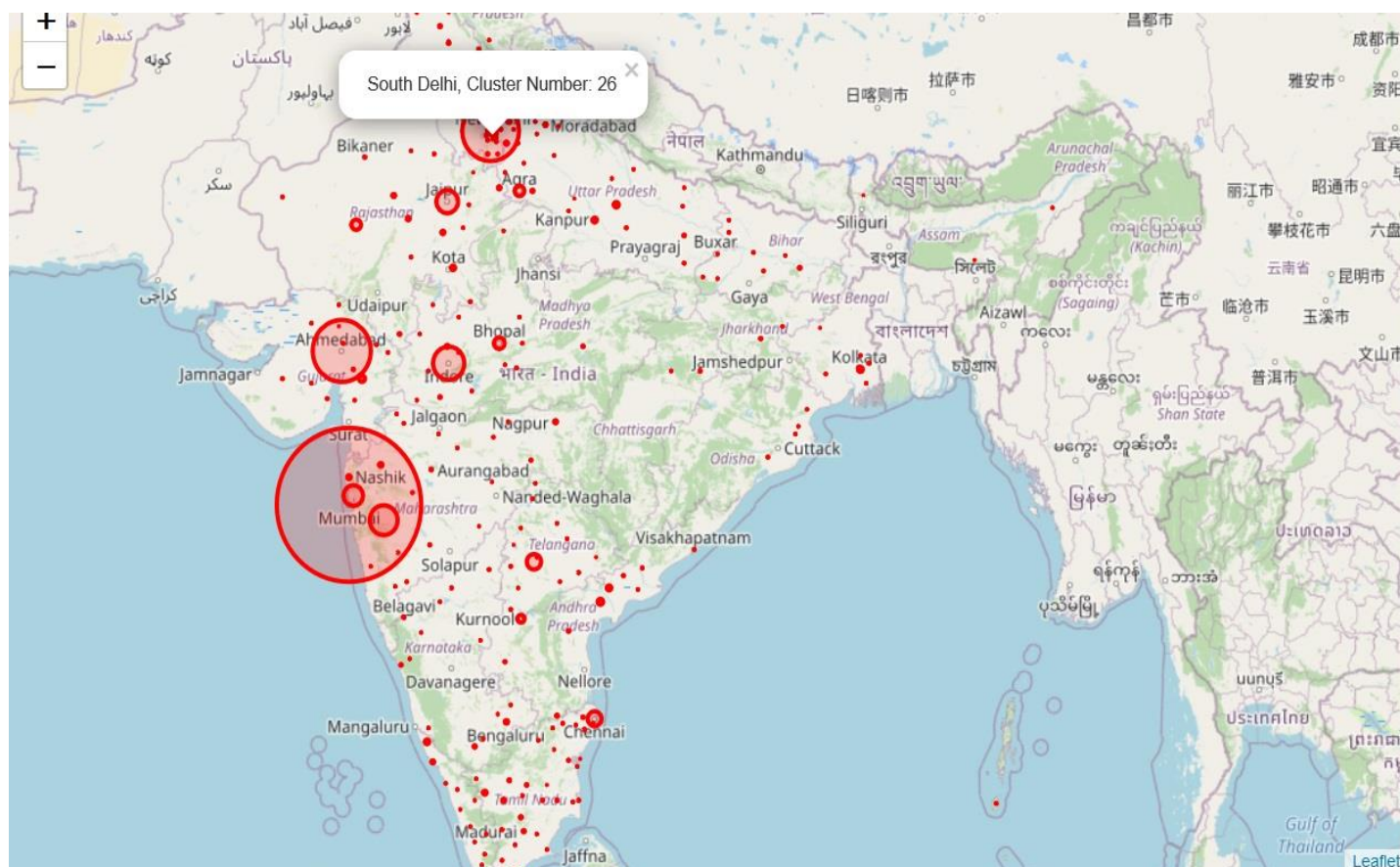5.8. The clusters predicted by our model can be visualized as follows:

Fig. 7 COVID-19 Clusters in India

5.9. Next step is the fulfilment of this project's goal that is to provide relief, recommendations in our case.

5.10. Now that we have our clusters and their coordinates, we need to find the centroids of these clusters to find an appropriate location which is more or less equidistant from each cluster point in that cluster.

5.11. For explanation part we are assuming a hypothetical situation where, crisis in Delhi deepens and there is an urgent need for more quarantine beds, so some hotel chains come forward and offer their spaces as quarantine facilities.

5.12. Also, the restaurants chains come ahead and volunteer for their kitchens being used as community kitchens so that food packets can be distributed among needy amid this crisis.

5.13. Hence, we are going to explore locations near this cluster number '26' that is 'Delhi' according to our K-Means Clustering Model.

5.14. Now, we made use of Foursquare location data API. To fetch the neighbouring localities near these coordinates.

6. CONCLUSION

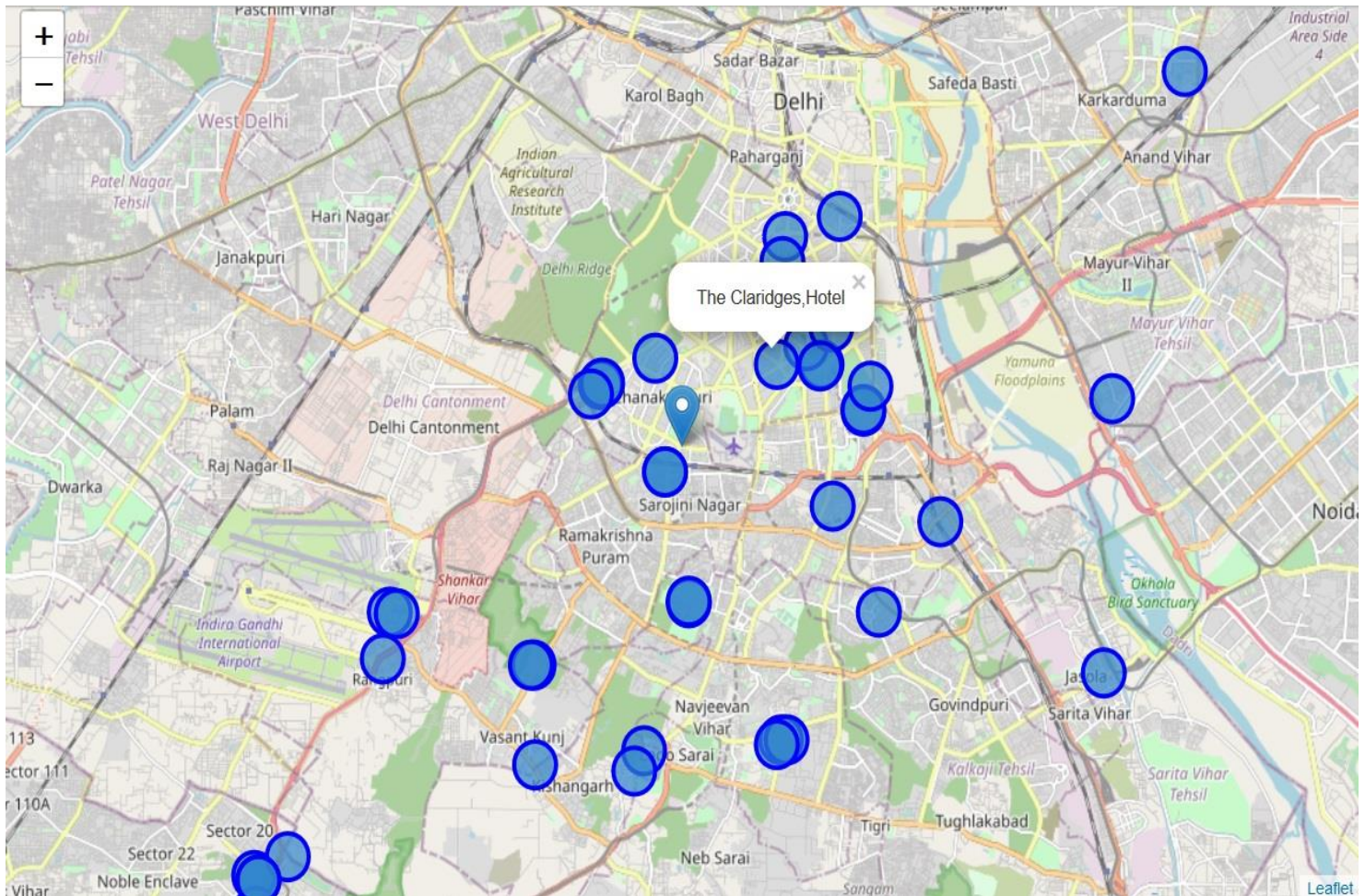6.1. The resulting recommendation for our selected cluster looks like below:

Fig. 8 Locations of Interest near Cluster Centre

6.2. We can observe that we can easily approach these venues, if god forbid crisis worsens and there is a need for more beds, food and other amenities.

6.3. This method was highly efficient in identifying the mean location of clusters, and we as we can observe from the figure, we are easily able to identify the locations of interest.

6.4. The new recommended locations for 'Testing Centres' was not provided as there were restrictions to the category of venues fetched by Foursquare API. Hence, Hospitals, big or small, Medical Dispensaries, were not listed in the API response.

6.5. As far as the analysis and real time coronavirus status in the country is concerned, there is definitely a need for more testing since, only after an increased number of daily tests can one identify any impact of new therapies or lockdowns on the virus spread. Without increased testing, analysis like these are just a play with numbers with no association to ground reality.

7. FUTURE SCOPE

7.1. The analysis can be further improved by better recording the data in patient database. More information about testing will provide better insights as to what is actually happening in the states and how the state governments are handling this crisis.

7.2. The prediction about the surge in cases in days to come can also be identified if all the existing data is refilled properly, checking values for each feature.

7.3. Foursquare API can be improved as more categories of venues can be added into it for better analysis.