Analysing and Predicting the COVID-19 Trend in India

Ritika Bhole

April 29, 2020

1. Introduction

1.1. Background

Coronaviruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19. COVID-19 is an infectious disease caused by the most recently discovered coronavirus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019. COVID-19 is now a pandemic affecting many countries globally.

1.2. Problem

The spread of COVID-19 has been rapid. The government is collecting plenty data from every state in an attempt to understand the spread and act accordingly. There is insufficient information available to analyse the current and predict the future scenarios. The spread of COVID-19 has to be analysed thoroughly to predict its expanse and when its growth will be curbed.

1.3. Interest

Currently, everyone is interested in knowing how and when this pandemic will stop. If we focus specifically on the specialists, then the analysts will be interested in this as to figure out ways to handle the situation. As the country is currently in lockdown and on the situation where it may be extended, this will help the strategists to estimate the duration of lockdown needed to curb the pandemic in India.

2. Data acquisition & cleaning

2.1. Data sources

The data source is a dataset from kaggle which is scraped from the authentic data source of the Indian government's COVID dashboard and website. The data has the following information:

- Date - Date of cumulative report

- Name of State / Union Territory / National Capital Region
- Total Confirmed cases (Indian National) - Cumulative count of Indian national confirmed with COVID-19
- Total Confirmed cases (Foreign National) - Cumulative count of foreign national confirmed with COVID-19
- Cured/Discharged/Migrated - Cumulative count of cured/ discharged cases
- Latitude - Latitude of the location
- Longitude - Longitude of the location
- Death - Cumulative count of deaths reported
- Total Confirmed cases

3. Methodology

3.1. Feature Extraction

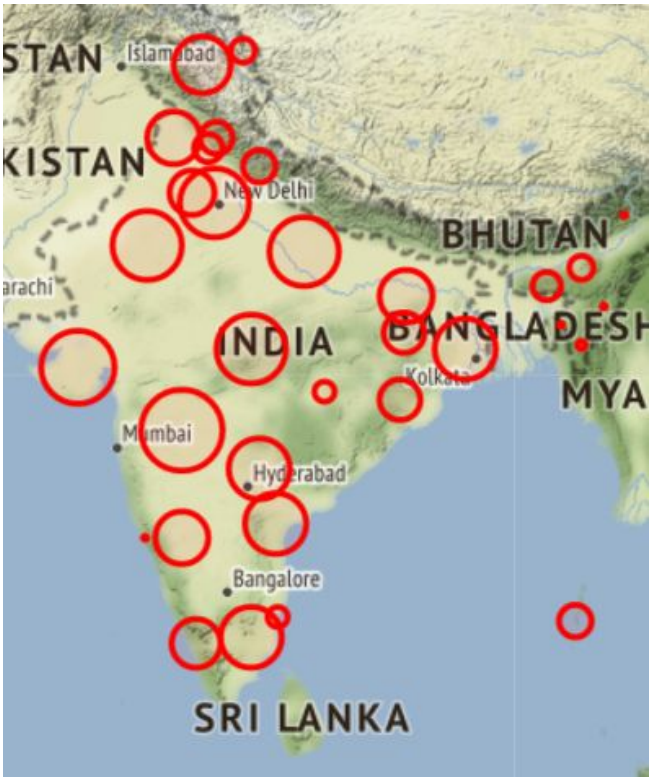
- The date attribute in the data is of 'object' data type. Conversion of this to datetime datatype helps us to extract more features.
- From the date attribute, the days from start can be enumerated, which helps in the plotting of the graph.
- 'Active' attribute is generated, which indicates the number of active cases. It is calculated simply as follows:

$$Active = Confirmed - Recovered - Deaths$$

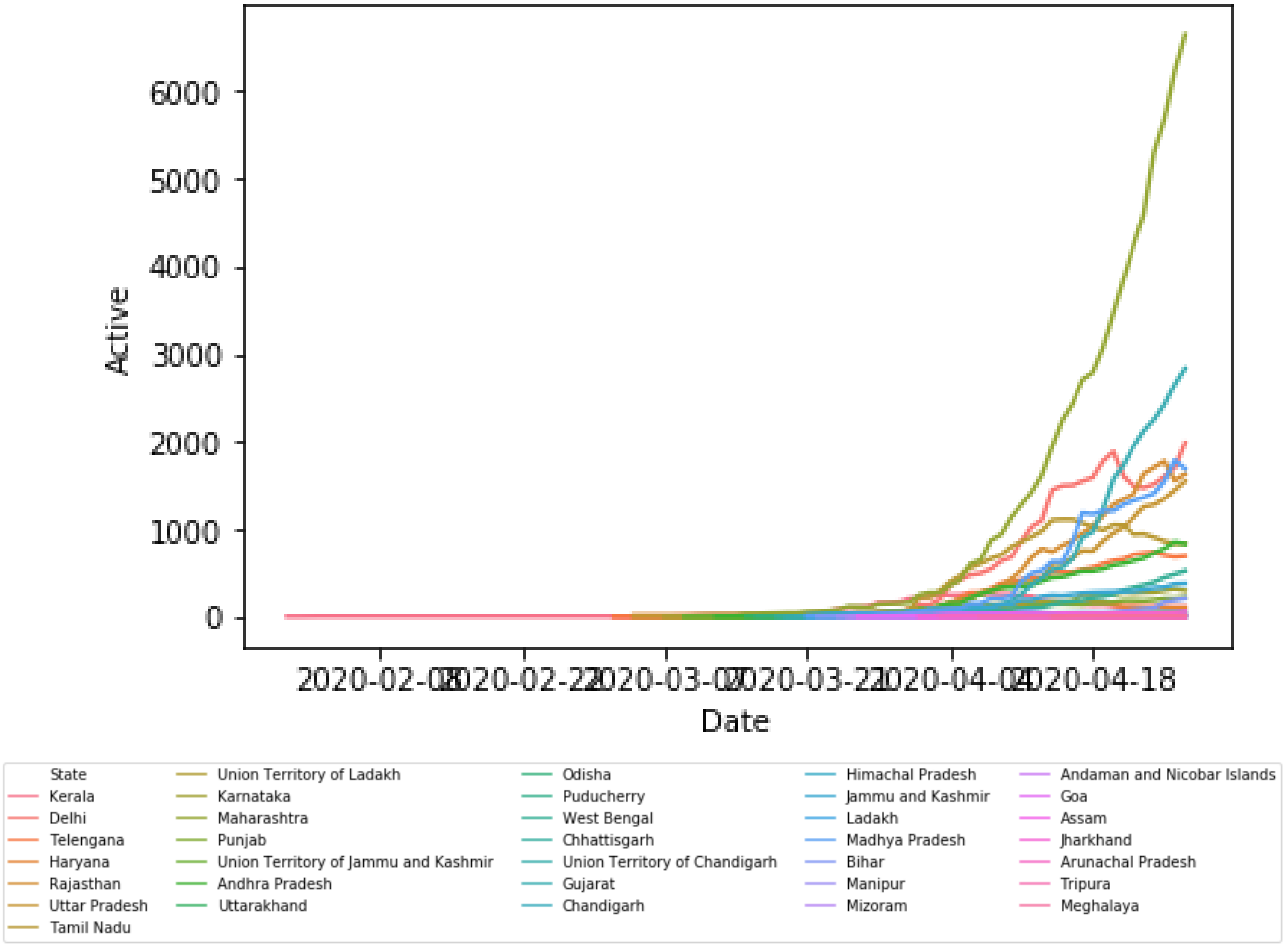
4. Exploratory Data Analysis:

4.1. Cases Till Date

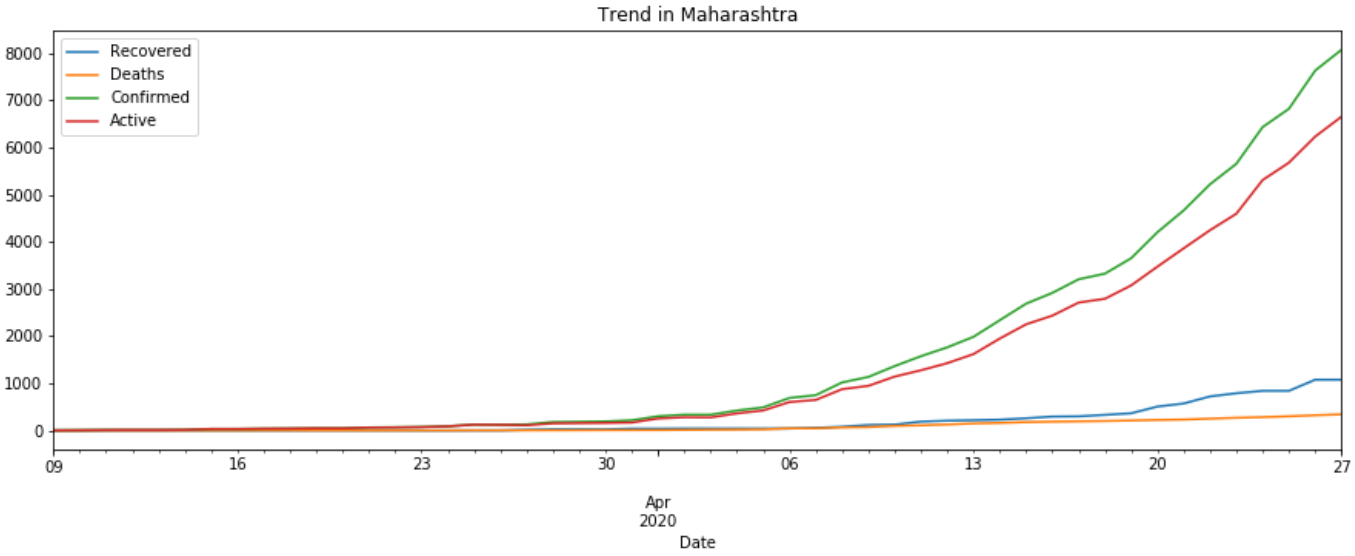
Here, the active cases in each state is showed on map, using Folium and Staten Terrain tiles.



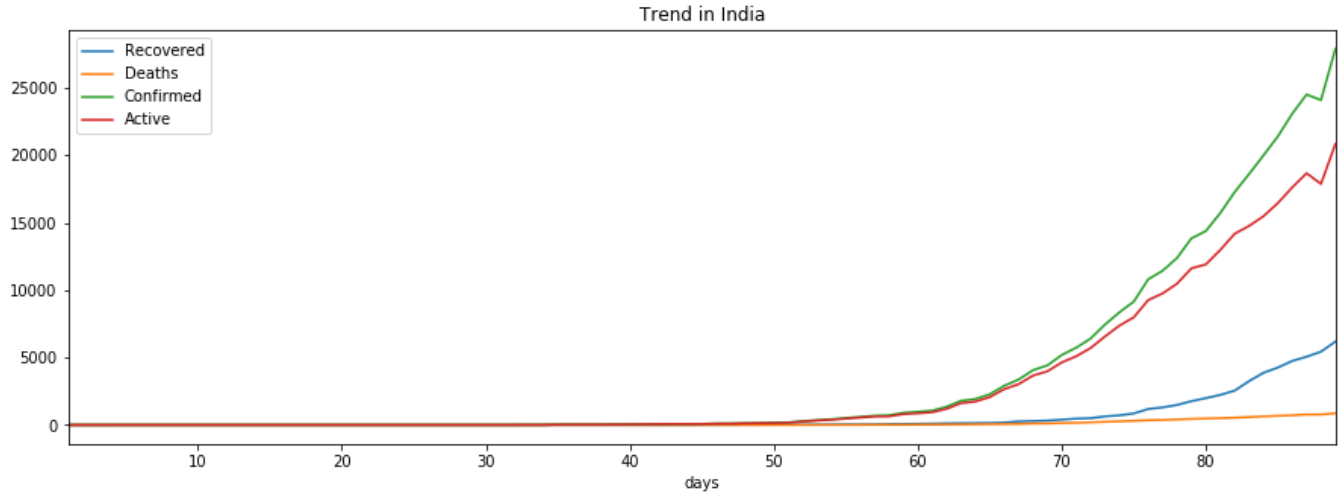
4.2. Cases timeline



4.3. EDA for Maharashtra State

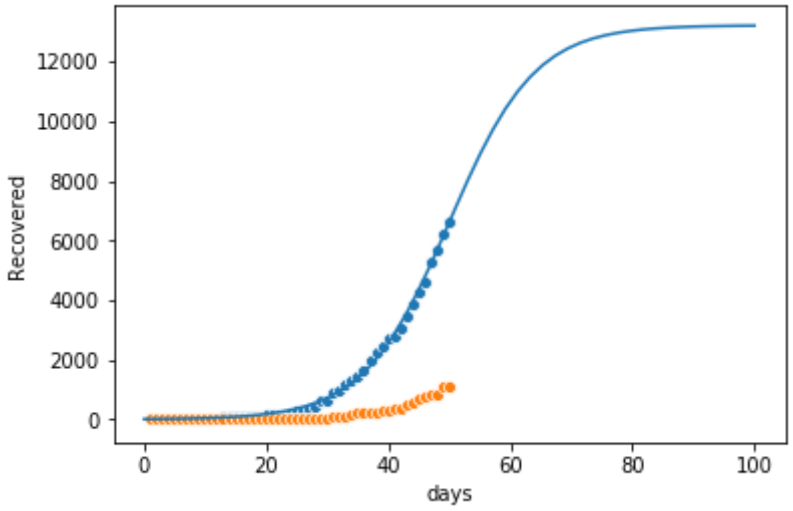


4.4. EDA for India

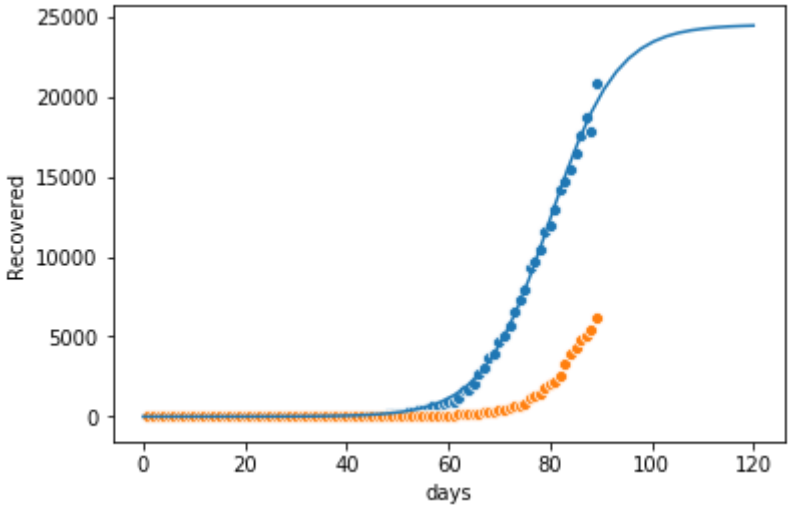


5. Results

On analysis of data, a sigmoidal curve fits the data better. Hence, first, we fit Maharashtra state data on a sigmoid. The figure is as follows:



Here, we can see that the sigmoidal curve flattens at around day 75 for Maharashtra
Similarly, we fit the sigmoid function for India's data, whose figure is as follows:



Here, we can see that the sigmoidal curve flattens at around day 115 for India.

6. Discussion

- Although we see that the sigmoid curve fits the data well, we're still unsure of how the factors affect the spread.
- Adding more features could be the future scope of the analysis

7. Conclusion

- Curve flattens at around day 75 for Maharashtra and at around 115 for India.
- As seen the duration of the dates is when lockdown had started and is continued. Hence, continuation of it till the recommended days will help to curb this pandemic.