Covid-19 Pandemic Data Analysis and Forecasting using Machine Learning Algorithms

Sohini Sengupta (sohini.forensics@gmail.com)

Sareeta Mugde (saritamugde54@gmail.com)

Abstract

India reported its first case of covid-19 on 30th Jan 2020. Though we did not notice a significant rise in the number of cases in the month of February and like many other countries, this number escalated like anything from March 2020. This research paper will include analysis of covid-19 data initially at a global level and then drilled down to the scenario of India. Data is gathered from multiple data sources from several authentic government websites. The paper will also include analysis of various features like gender, geographical location, age using Python and Data Visualization techniques. Getting insights on Trend pattern and time series analysis will bring more clarity to the current scenario as analysis is totally on real-time data(till 19th June). Finally we will use some machine learning algorithms and perform predictive analytics of the near future scenario. We are using a sigmoid model to give an estimate of the day on which we can expect the number of active cases to reach its peak and also when the curve will start to flatten sigmoid model gives us a count of date which is a unique feature of analysis in this paper. We are also using certain feature engineering techniques to transfer data into logarithmic scale for better comparison removing any data extremities or outliers. Based on the predictions of the short-term interval, our model can be tuned to forecast long time intervals. Needless to mention there are a lot of factors responsible for the cases to come in the

upcoming days. It depends on the people of the country and how strictly they obey the rules and restriction imposed by the Government.

Keywords – Covid 19, Data Analysis, Forecasting, Predictive Models

Introduction

The Covid-19 pandemic in India is a part of the world Covid-19 which is caused due to severe acute respiratory syndrome coronavirus 2. Unfortunately India currently reports the highest number of cases in Asia. However the fatality rate is relative lower (2.8%) as compared to the world(6.1%) as of 3rd June 2020. On 12th January 2020 the World Health Organization (WHO) declared the novel coronavirus was responsible for the respiratory illness of a community of people in Wuhan, China. First confirmed case in India was reported on 30th January and the first confirmed death was reported on 12th March. Eventually all Indians were addressed by the Government of India to maintain social distancing as a preventive measure. People were supposed to follow this till 31st March, however before that only the nationwide lockdown was announced. Only essential services were kept open. Major cities and some state made wearing face masks compulsory. Central armed forces came into action. Helpline numbers was set up. Various research agencies along with Government of India took the initiative to gather data on Covid-19 and a database was setup that included real time data of number of confirmed cases, deaths, recovery date cases based on age, gender, geographical location, and position of India with respect to other countries. Also data was collected that included information on the number of diagnostic tests that were happening at a state and district level. The nation was divided into three zones namely a) Red zones (Hotspots) b) Orange zones (non-hotspots) and c)

Green zone (districts without confirm cases for three consecutive weeks). The nation witnessed huge economic downfall in these months. Thousands of people lost jobs. Retail sector became the biggest casualty of lockdown. Also the tourism, hospitality and aviation sector faced use losses. The pandemic allowed near to zero inflows of tourists and visitors. This largely impacted travel agents, hotel and aviation. According to sources, at least 4200 companies across the country have been forced to shutdown. Global supply chain has been heavily disturbed. According to data reported by Centre for Monitoring the Indian economy (CMIE) the unemployment rate had just reached to 30% in urban and 21% in rural areas bringing the total unemployment rate of the country to 23.8% . However there are also some potential winners in this lockdown. Video conferencing apps witnessed its biggest gains as online meetings become the key to existing business. We can also count to medical supply sector, Information and Communication Technology (ICT) supermarkets and e-commerce sector as other potential winners. Not to forget telecom sector has also reported increased demand to internet and voice services in the lockdown period. The ministry of Electronics and IT has successfully launched an application called Aarogya Setu. Its wonderfully designed database provides necessary information about contact tracing. The government has also announced an amount of RS 20.97 crore package which will be infused in several sectors. This package accounts nearly 10% of GDP of the country. Researchers are not sure where the situation will lead us to. There is a lot of uncertainty in every predictive model given by data scientists. As of now, all what we can do is stay at home, follow the instructions and build a strong immunity system within our body.

Review of Related Literature

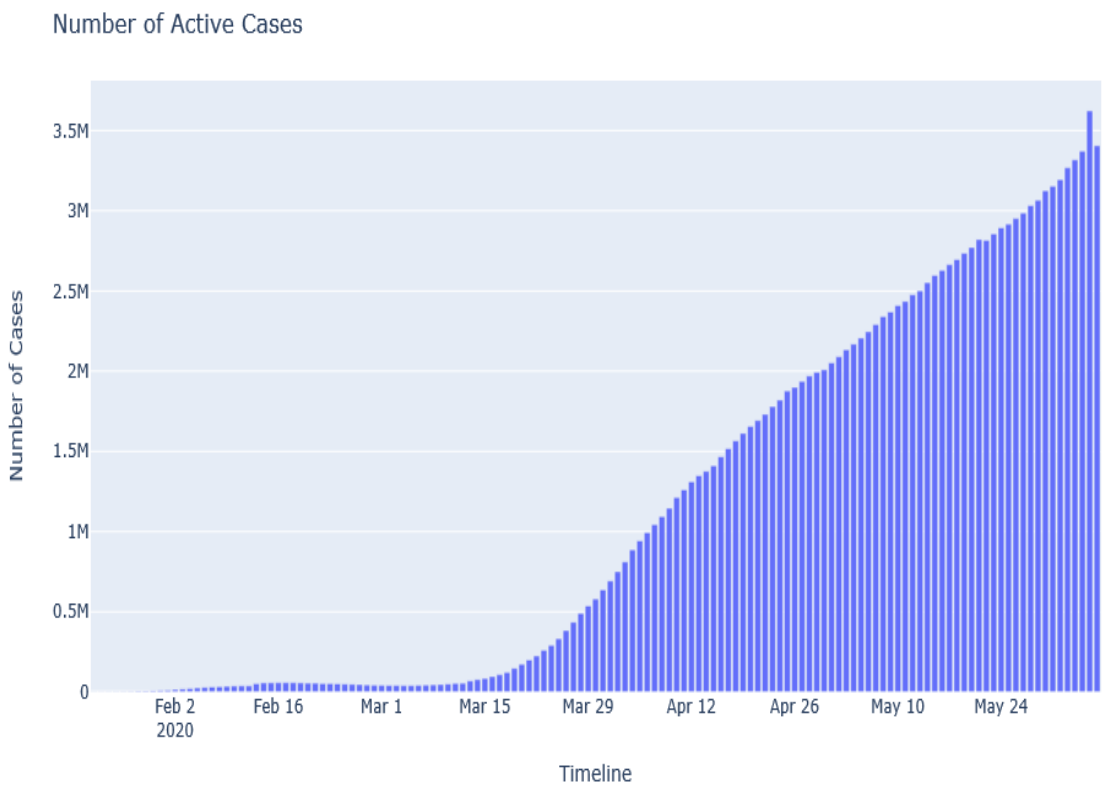
There are a lot of research papers published that are related to covid-19. Some of them to name can be research work related to vaccine or other medical drugs that can help to recover. Deep analysis is done on the people who recovered which can shed some light on how to deal with the active cases. Data scientists all over the world are busy in making sense out of the available data and predict the near future. Finding trend pattern, feature selection, forecasting techniques are being applied in and out to come to a conclusion. Rajan Gupta and Saibal Kumar Pal, in their research paper 'Trend analysis and forecasting of covid-19 break in India' used exploratory data analysis to report the situation in the time period of January to March in India. They use time series forecasting methods to predict the future trends. A very famous machine learning model- Arima model prediction was used and the inferred a result that predicted huge range of the number of likely covid-19 positive cases in April and May. The average that was forecasted was a detection of approximately 7000 patients in a span of 30 days in April. However in reality the figures were pretty higher. Another research paper by the department of CSE and IT of Northcap University, India in collaboration with Defence Research and Development Organization (DRDO) India also covered data from January 30 to March 30, 2020. They used regression models for forecasting. According to them, expected cases may rise to about 5000 in a two week time period. This was far more accurate however actual scenario showed a bigger upsurge. This research paper can also be of help to several other sectors or other branches of healthcare as immunity power is very related to fighting with Covid-19.

According to healthcare experts, people having a less developed immunity system are more likely to be a victim of Covid-19. A research paper by Gail Dovey-Pearce , Ruth Hurell , Carl May in the journal of 'Health Social Care in the community', has revealed certain suggestions for providing developmentally appropriate diabetic services and they have focused on the opinion of young adults (16 - 25 years). We all know people suffering from diabetes are more prone to many other diseases with the reduced immunity power. So it's high time we gather information from healthcare workers regarding how to stay fit and healthy during this pandemic. Another research paper in the same journal by Susan Kerr, Hazel Watson, Debbie Tolson gives a qualitative exploratory research study on older current smoker's view on smoking and stopping smoking. This research paper helps to pick certain facts that common people mostly unaware of. A detailed consumer behavior analysis along with the experiences reveals the danger of Covid-19 to smokers. Another very relevant paper in the journal of 'Chaos, Solitons & Fractals' uses clustering methods to analyze countries on the basis of most affected patients and how they are reacting to it. This paper shows the world scenario first and thereby drilling down to the country of Mexico. Speaking of a country like India, where approximately 65% are belonging to rural population, accessibility and utilization of primary health care during the current pandemic situation is a big question. A similar study has been done in UK and we can find its related paper in the journal- 'Health Social Care in the community'. This inspires me to conduct a similar study in India. However it is quite and qualitative research and is beyond the scope of this paper. Another recently published paper in the journal - 'Chaos Solitons & Fractals' captures the trend of cases and also a prediction using Fourier Decomposition Method. This

paper also collects data till 1st week of June and predicts the expected number of cases and deaths in the upcoming days.

Analysis

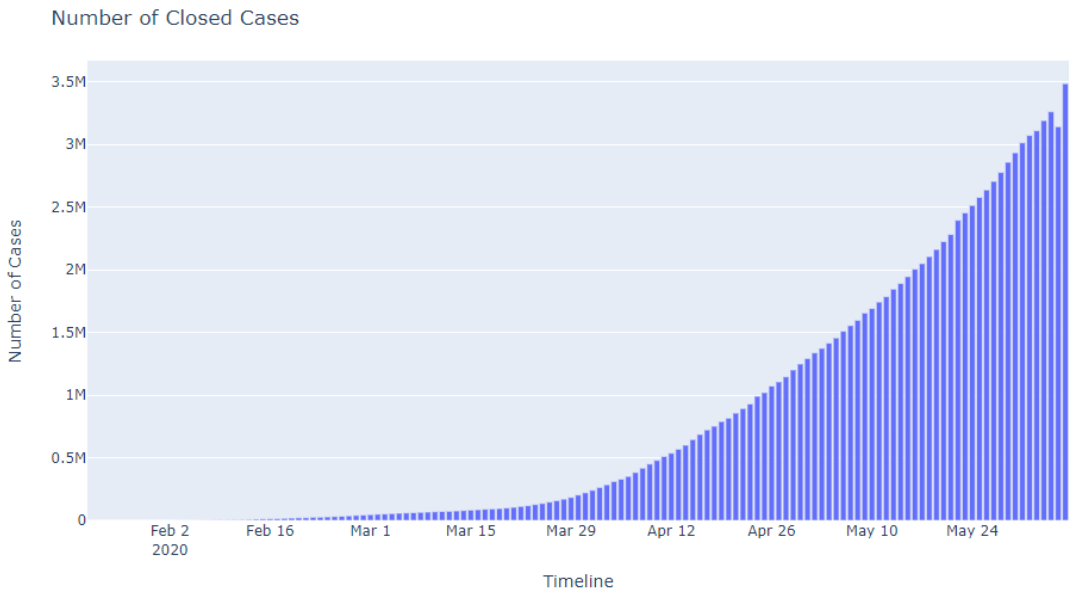
Visualizations had always been easy to understand the raw data. Here we are going to compare the growth of Covid-19 confirmed, death and recovered cases of India to other major countries that have also been heavily infected. The visualizations are created in Python where using the matplotlib, seaborn, plotly libraries and also datetime library for time series data analysis. Date of India and world is gathered from multiple data sheets from kaggle.



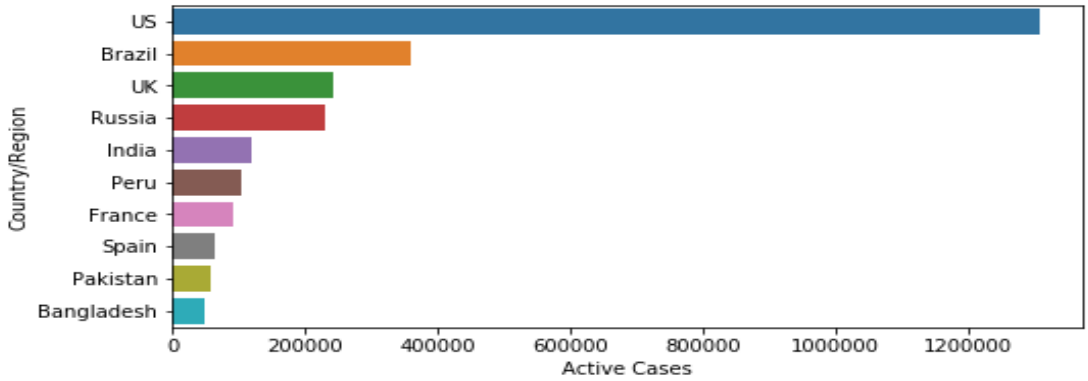
The above bar graph represents the timeline of how the number of active cases had been increasing over the world. (till 19th June).

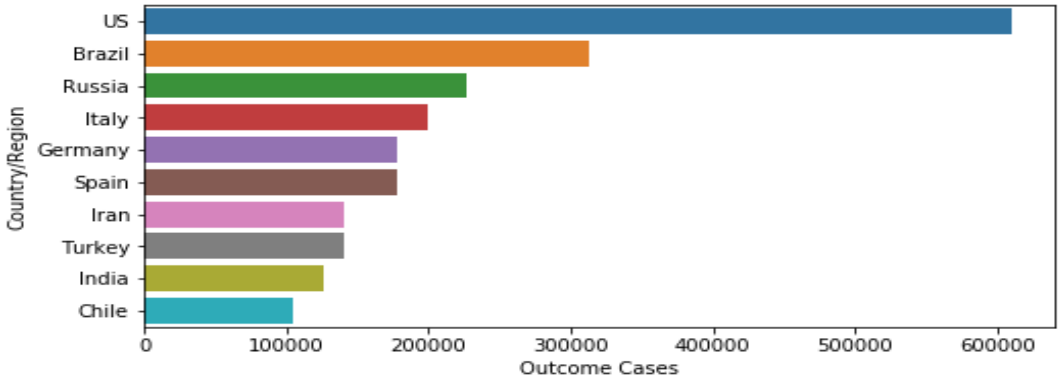
Active cases= confirmed cases- recovered cases- death cases

An increase in number of active cases signifies a considerable drop in the number of recovered and death case with respect to the number of confirmed cases. To further confirm this we can visualize the number of closed cases as well.

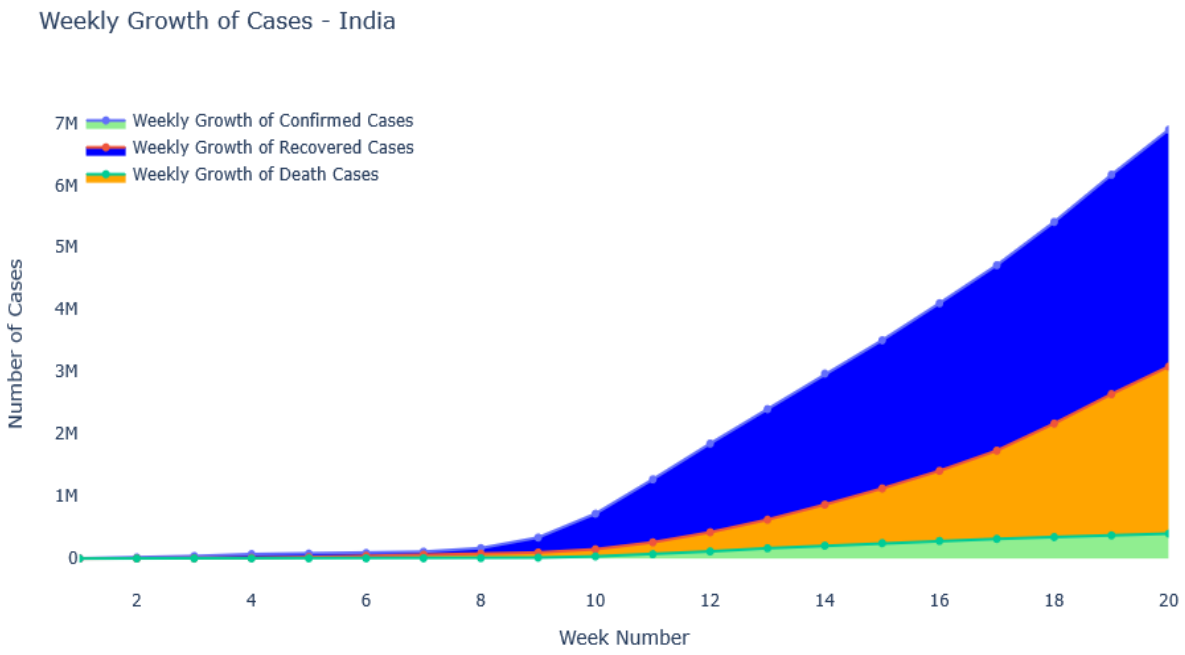


The above graph supports the fact that either more patients are getting recovered from the disease or more patients are dying due to Covid-19.

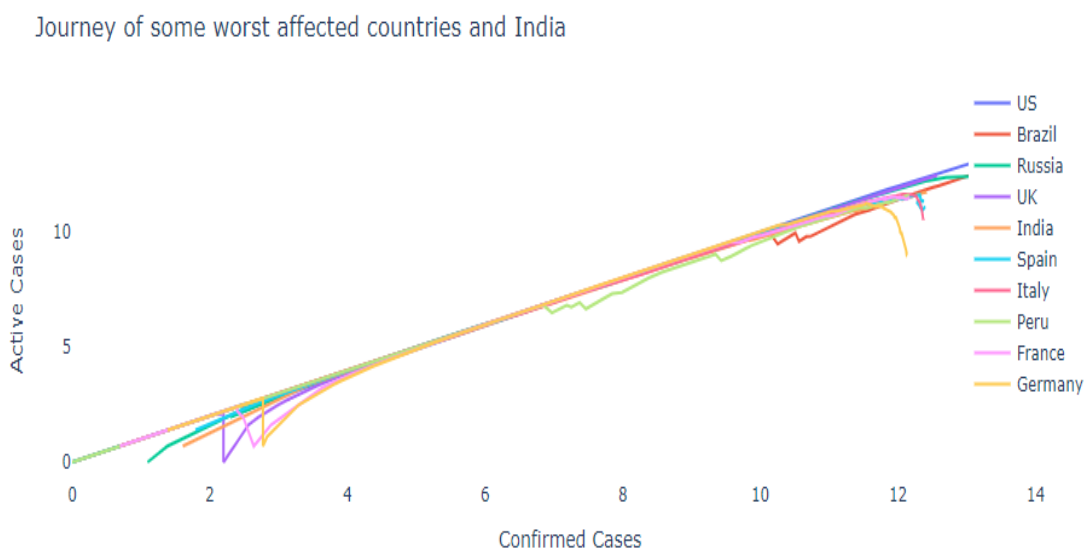




Here's a bar graph to compare the active and closed cases of different countries as of 8 June 2020 India's active cases remain at v position in the world it comes to the ninth position in the number of closed cases.



Weekly growth of cases worldwide starting from January 2020 cases considerable increase from week 9 that is March and experience sharp growth.

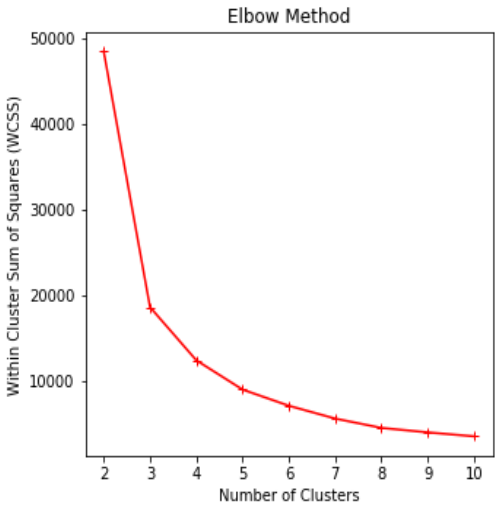


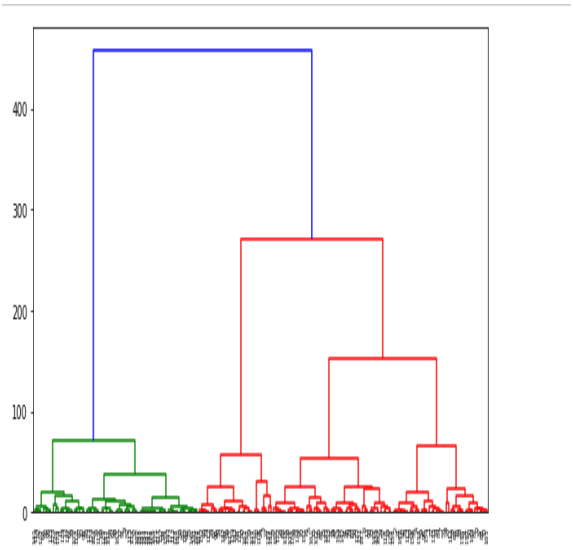
The visualization is done on a logarithmic scale. It is quite evident that the pandemic has spread

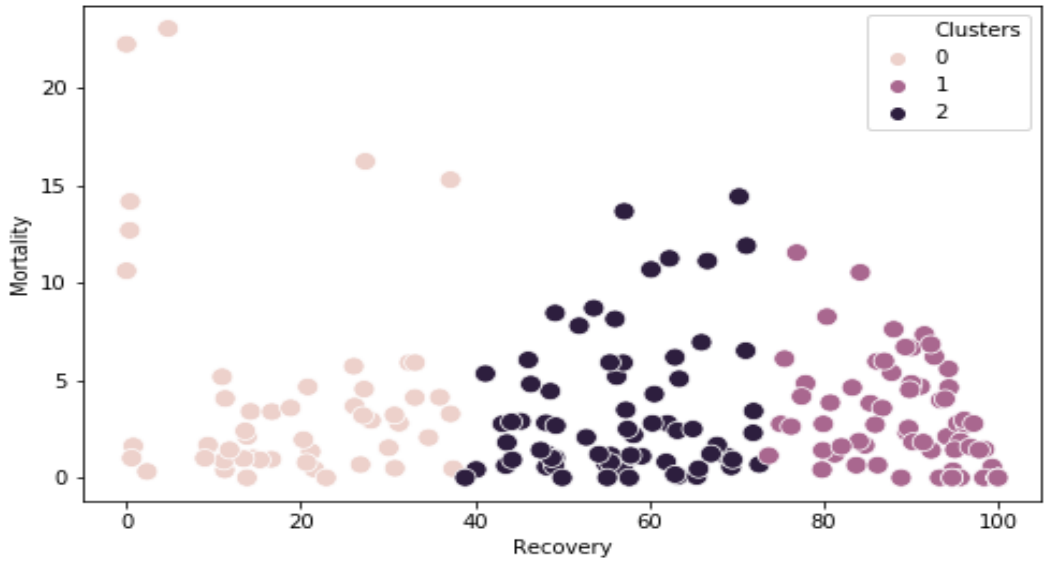
in all the countries in same manner, however certain countries are practicing controlling procedures rigorously and its evident from the graph. Most of the countries are following the same trajectories of US i.e uncontrolled exponential growth while a few countries like Germany Spain Italy have started showing a dip indicating signs of control over Covid-19.

Clustering –

Machine learning provides an excellent feature of clustering which will help us to categorize countries on the basis of severity of the pandemic. Severity can be measured on several features, here I am considering the mortality and recovery rate of countries. We are using k-means clustering and hierarchical clustering methods both of which suggests that suitable number of clusters will be 3.







Therefore considering k= 3, we get the following.

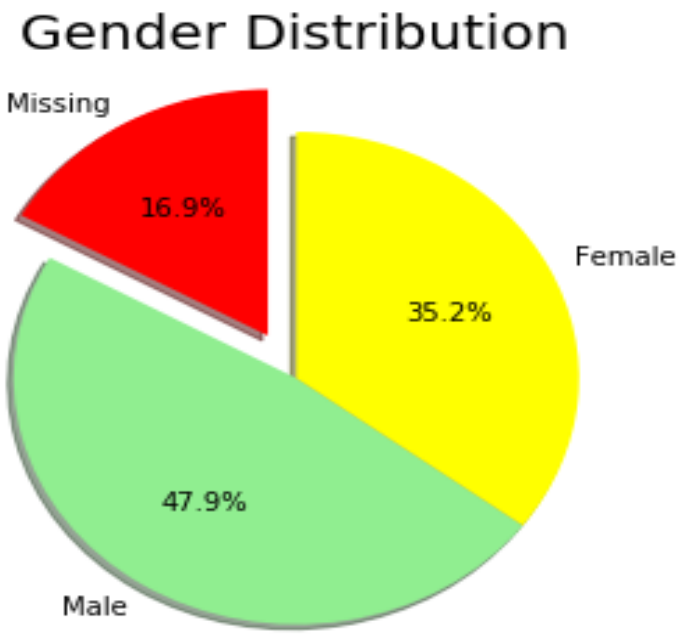
```
In [34]: print("Avergae Mortality Rate of Cluster 0: ",countrywise[countrywise["Clusters"]==0]["Mortality"].mean())
         print("Avergae Recovery Rate of Cluster 0: ",countrywise[countrywise["Clusters"]==0]["Recovery"].mean())
         print("Avergae Mortality Rate of Cluster 1: ",countrywise[countrywise["Clusters"]==1]["Mortality"].mean())
         print("Avergae Recovery Rate of Cluster 1: ",countrywise[countrywise["Clusters"]==1]["Recovery"].mean())
         print("Avergae Mortality Rate of Cluster 2: ",countrywise[countrywise["Clusters"]==2]["Mortality"].mean())
         print("Avergae Recovery Rate of Cluster 2: ",countrywise[countrywise["Clusters"]==2]["Recovery"].mean())

Avergae Mortality Rate of Cluster 0:  4.287289547311022
Avergae Recovery Rate of Cluster 0:  19.114863396176958
Avergae Mortality Rate of Cluster 1:  2.9385115321751485
Avergae Recovery Rate of Cluster 1:  89.85890893106854
Avergae Mortality Rate of Cluster 2:  3.4025888997387757
Avergae Recovery Rate of Cluster 2:  56.79028093432937
```

CLUSTER	Mortality Rate	Recovery Rate
0	High	Low
1	Low	High

2	Medium	Medium

We can see countries belonging to cluster 1 are at a comparatively safer zone with low mortality rate and high recovery rate. Countries belonging to cluster 1 are India, Peru, Chile, Pakistan, Bangladesh, USA and Russia.

A gender distribution analysis reveals that males are more likely to be diagnosed with Covid-19

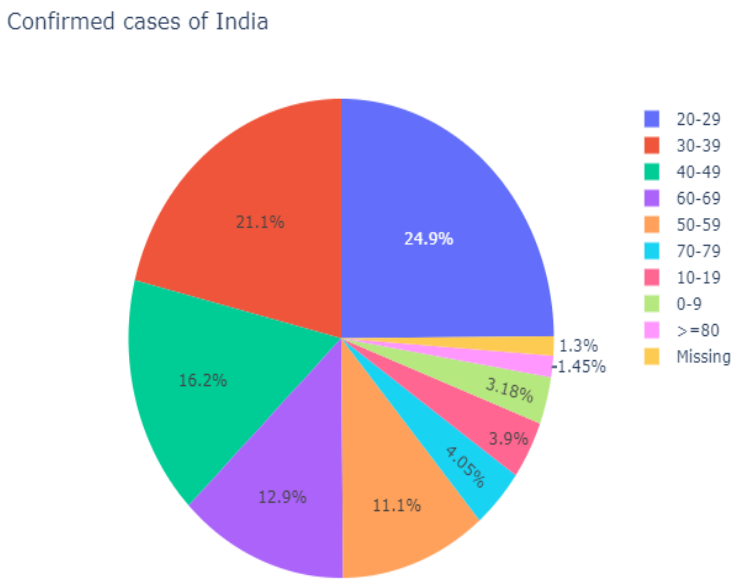


Indian Scenario

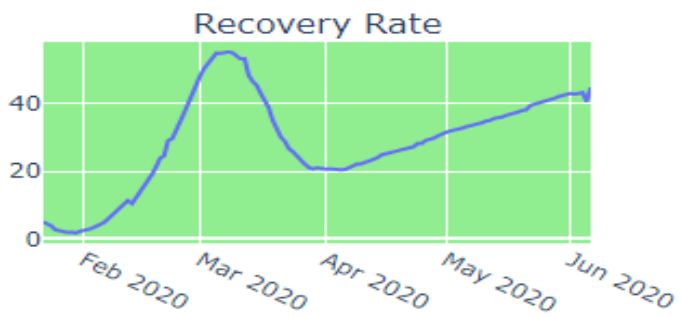
Drilling down to the present scenario of India- data available from various sources reveals that covid-19 reached India at a little later stage. The first case was diagnosed on 30th January and a

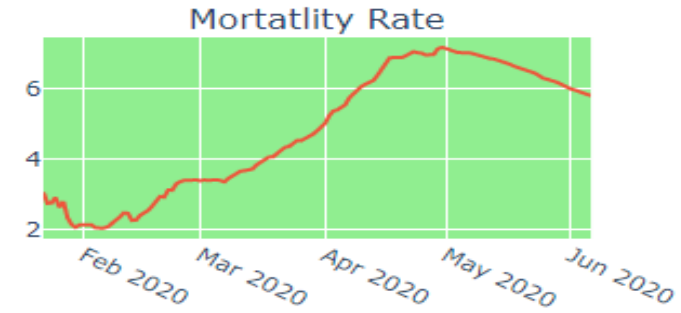
few cases were reported on February, all of them being students returning from Wuhan,China to Kerala, India.

Age wise distribution of confirmed cases till 19th June 2020 -



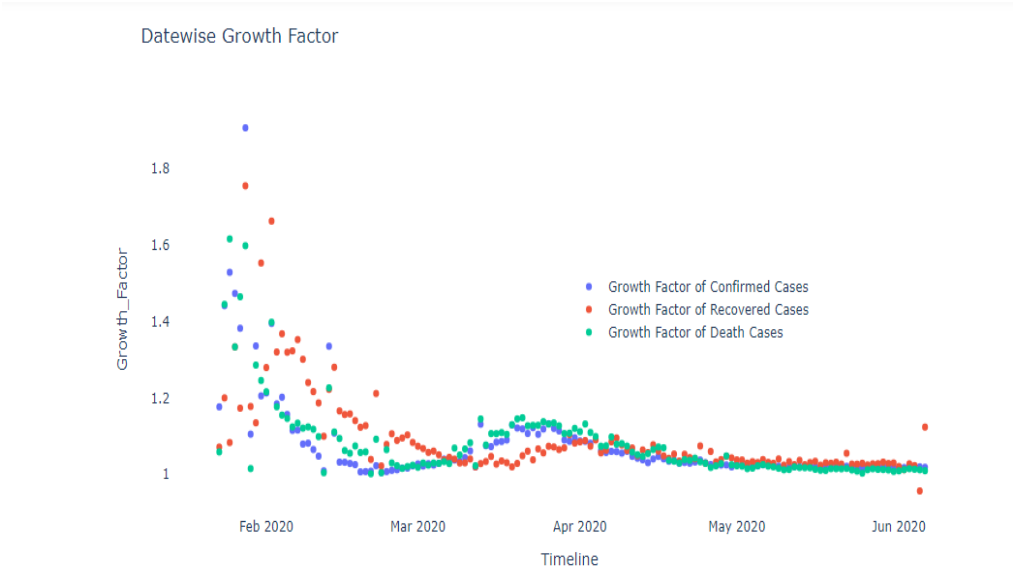
Calculating the average mortality rate and recovery rate-



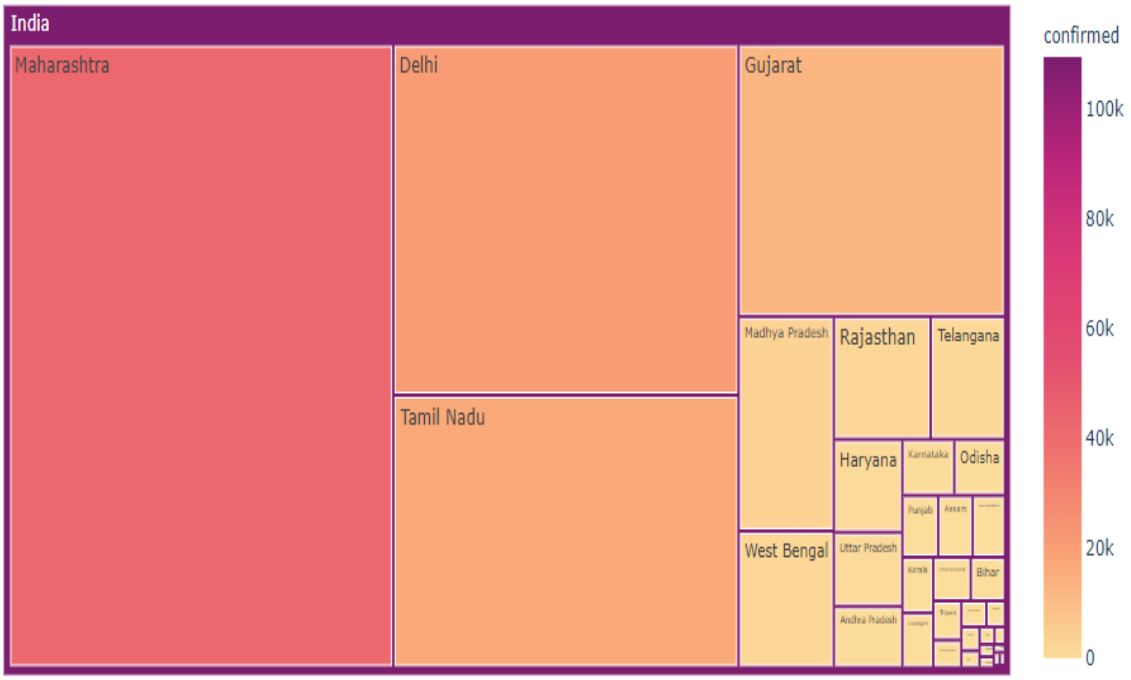


Growth factor (Graph in log scale) -

It is calculated by the number of cases confirmed yesterday divided by that of today. Using the shift function of python, we can find out a continuous trend of cases increasing with respect to the previous day. A growth factor of 1 indicates cases of yesterday and today was same while growth factor of 2 indicates doubling of new confirmed cases.

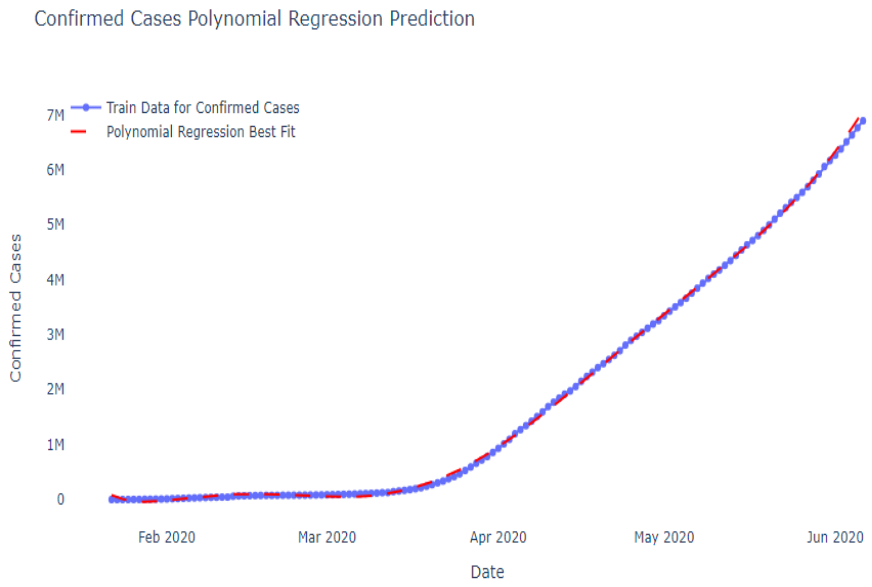


Treemap visualization of cases according to states. As of June 19th, 2020, Maharashtra, Delhi, Tamil Nadu, Gujarat where were at the top of the list.



Forecasting

The below figure shows that polynomial regression is the best fit for our data. We had also tried with linear regression and support vector machine regression which did not give accurate results.



Forecasting using sigmoid model -

I am using sigmoid function to forecast the near future scenario in India. Even China's data also resemble does sigmoidal shape.

It was only from mid-march when India witnessed the continuous upsurge of cases, therefore in this forecasting model I am considering data from 16th March 2020.

Fitting sigmoid function in the data –

$$y = \frac{c}{1+np^{-a(-b)}}$$

a - sigmoidal shape progress of infection (better if small in our case)

b - The point where sigmoid star to flatten from steepening - midpoint of segment when rate of increment of cases will slow down.

c - The maximum value (maximum number of infected patients)

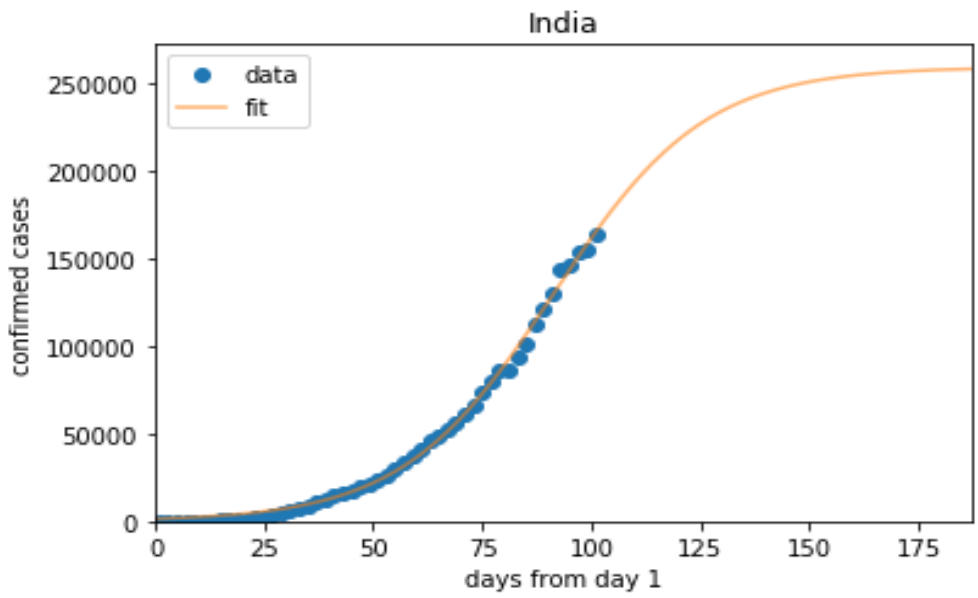
Results –

Python calculation and overview of the data -

```
model fitted max Active at: 258846
model sigmoidal coefficient is: 0.058
model curve stop steepening, start flattening by day: 91
model curve flattens by day: 191.1
```

	Date	Confirmed	Deaths	Recovered	Active	day_count	increase	rate
48	2020-03-10	56	0	4	52	1	NaN	NaN
49	2020-03-11	62	1	4	57	2	5.0	0.087719
50	2020-03-12	73	1	4	68	3	11.0	0.161765

	Date	Confirmed	Deaths	Recovered	Active	day_count	increase	rate
147	2020-06-17	366946	12237	194325	160384	100	5157.0	0.032154
148	2020-06-18	380532	12573	204711	163248	101	2864.0	0.017544
149	2020-06-19	395048	12948	213831	168269	102	5021.0	0.029839



The model predicts maximum active cases at 258846.

The curve flattens by day 154 i.e. 25th September and after that the curve goes down and the number of active cases eventually will decrease.

Conclusion -

There are a lot of research works going on with respect to vaccines, economic dealings, precautions and reduction of Covid-19 cases. However currently we are at a mid-Covid situation. India along with many other countries are still witnessing upsurge in the number of cases at alarming rates on a daily basis. We have not yet reached the peak. Therefore cuff learning and downward growth are also yet to happen. Each day comes out with fresh information and large amount of data. Also there are many other predictive models using machine learning that beyond the scope of this paper. However at the end of the day it is only the precautionary measures we as responsible citizens can take that will help to flatten the curve. We can all join hands together and maintain all rules and regulations strictly. Maintaining

social distancing, taking the lockdown seriously is the only key. This study is based on real time data and will be useful for certain key stakeholders like government officials, healthcare workers to prepare a combat plan along with stringent measures. Also the study will help mathematicians and statisticians to predict outbreak numbers more accurately.

References –

1) Rajan Gupta, Saibal Kumar. Trend Analysis and Forecasting of COVID-19 outbreak in India. Rajan Gupta, Saibal Kumar PalmedRxiv, 2020. doi: https://doi.org/10.1101/2020.03.26.20044511

2) https://www.kaggle.com/neelkudu28/covid-19-visualizations-predictions-forecasting

3) Kundapur, Rashmi; Rashmi, Anusha; Sachin M.; Falia, Karishma; Remiza R. A.; Bharadwaj, Shambhavi. COVID 19 -- Observations and speculations -- A trend analysis. Source: Indian Journal of Community Health. 2020 Special issue, Vol. 32 Issue 2, p300-305. 6p.

4) Kok Hoe Chan Divya Thimareddygari Amr Ramahi Liana AtallahClinical Characteristics and Outcome in Patients with Combined Diabetic Ketoacidosis and Hyperosmolar Hyperglycemic State Associated with COVID-19: A Retrospective , Hospital-Based Observational Case SeriesDiabetes research and clinical practice · June 2020 *with* 13 Reads. DOI: 10.1016/j.diabres.2020.108279

5) Amit Singhal, Pushpendra Singh, Brejesh Lall, Shiv Dutt Joshi, Modeling and prediction of COVID-19 pandemic using Gaussian mixture model, Chaos, Solitons and Fractals (2020), doi: https://doi.org/10.1016/j.chaos.2020.110023

6) PatriciaMelinJulio, CesarMonica, DanielaSanchez,OscarCastillo. Analysis of Spatial Spread Relationships of Coronavirus (COVID-19) Pandemic in the World using Self Organizing Maps. https://doi.org/10.1016/j.chaos.2020.109917

7) Dovey-Pearce, Gail, Hurrell, Ruth , May, Carl , Walker, Caron , Doherty, Yvonne. Young adults' (16-25 years) suggestions for providing developmentally appropriate diabetes services: A qualitative study. Article *in* Health & Social Care in the Community 13(5):409-19 · October 2005. *DOI - 10.1111/j.1365-2524.2005.00577.x*

8) Yan Bai, Lingsheng Yao, Tao Wei, Fei Tian, Dong-Yan Jin, Lijuan Chen, Meiyun Wang , Jama Presumed asymptomatic carrier transmission of COVID-19. 323 (14), 1406-1407, 2020

9) Sourish Das. Prediction of covid-19 disease progression in india: Under the effect of national lockdown. arXiv preprint arXiv:2004.03147, 2020

10) Anuradha Tomar, Neeraj Gupta. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Science of the Total Environment, 138762, 2020

11) Gaurav Pandey, Poonam Chaudhary, Rajan Gupta, Saibal Pal. SEIR and Regression Model based COVID-19 outbreak predictions in India. arXiv preprint arXiv:2004.00958, 2020

12) Qiurong Ruan, Kun Yang, Wenxia Wang, Lingyu Jiang, Jianxin Song. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Qiurong Ruan, Kun Yang, Wenxia Wang, Lingyu Jiang, Jianxin Song. Intensive care medicine 46 (5), 846-848, 2020

13) FA Binti Hamzah, CH Lau, H Nazri, DV Ligot, G Lee, CL Tan, MKM Shaib, UH Zaidon, A Abdullah, MH Chung, CH Ong, PW Chew, RE Salunga. CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. Bull World Health Organ 1, 32, 2020

14) Muhammad Adnan Shereenab, Suliman Khana , Abeer Kazmic, Nadia Bashira , Rabeea Siddiquea. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. https://doi.org/10.1016/j.jare.2020.03.005

15) Julie Byles Lyn Francis Mary McKernon. The experiences of non-medical health professionals undertaking community-based health assessments for people aged 75 years and over. Health & Social Care in the Community 10(2):67-73 · April 2002. DOI: 10.1046/j.1365-2524.2002.00343.x

16) Kathryn Roberts. Across the health-social care divide: Elderly people as active users of health care and social care. Health & Social Care in the Community 9(2):100-7 · April 2001 DOI: 10.1046/j.1365-2524.2001.00286.x

17) Awadhesh Kumar Singh, Akriti Singh, Altamash Shaikh, Ritu Singh, Anoop Misra. Chloroquine and hydroxychloroquine in the treatment of COVID-19 with or without diabetes: A systematic search and a narrative review with a special reference to India. (Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 2020.)

18) Pranab Chatterjee, Nazia Nagi, Anup Agarwal, Bhabatosh Das, Sayantan Banerjee, Swarup Sarkar, Nivedita Gupta, Raman R Gangakhedkar. The 2019 novel coronavirus disease (COVID-19) pandemic: A review of the current evidence. Indian Journal of Medical Research 151 (2), 147, 2020

19) Ronojoy Adhikari. Age-structured impact of social distancing on the COVID-19 epidemic in India. Rajesh Singh, arXiv preprint arXiv:2003.12055, 2020

20) Naresh Dhanwant, V Ramanathan. Forecasting COVID 19 growth in India using Susceptible-Infected-Recovered (SIR) model. Jay arXiv preprint arXiv:2004.00696, 2020

21) Eili Klein, Gary Lin, Katie Tseng, Emily Schueller, Geetanjali Kapoor, R Laxminarayan. COVID-19 for India updates. Princeton University, 2020

22) Murat Yıldırım Abdurrahim Güler. Factor analysis of the COVID-19 Perceived Risk Scale: A preliminary study. Death Studies · June 2020. DOI: 10.1080/07481187.2020.1784311

23) Novel Coronavirus (COVID-19) Cases Data, https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases,

24) WHO COVID-19 Dashboard, covid19.who.int

25) www.mygov.in/covid-19