CASE STUDIES IN DATA SCIENCE

(COSC2669)

FINAL REPORT

WIL PROJECT & DATATHON

BY

- **VISHAL D. PATNAIK** (s3811521@student.rmit.edu.au)
- **S V MANIDEEPU M** (s3820822@student.rmit.edu.au)
- **PHANI P** (s3798488@student.rmit.edu.au)
- **SUJAY KAMAL M** (s3794983@student.rmit.edu.au)

AT

RMIT UNIVERSITY

Date: 21st October 2020

Table of Contents

Abstract

In late December 2019, a cluster of unexplained pneumonia cases has been reported in Wuhan, China. A few days later, the causative agent of this mysterious pneumonia was identified as a novel corona virus. And the relevant infected disease has been named as COVID-19 by the World Health Organization. The COVID-19 pandemic is spreading all over the world now and almost 10% of the world population is now affected by this ongoing pandemic. The purpose of our report is primarily to assist the government agencies with a feasible approach in containing the spread of this virus through a web application that will assess and provide the viability of social distancing and lock down features as the key factors in preventing the further spread of this virus, and also by providing government experts and health experts with viable approaches to deal with the spread through futuristic SIR modelling of the existing data.

1. Introduction

1.1. Purpose:

The Fundamental purpose of this report is to describe in detail the analysis, scope, and futuristic prediction that we developed as a visualisation-based web application and to further discuss and expand the data analytics through SIR modelling.
The extended purpose of this report is to discuss in detail our collaborative approach as a team in coming up with a feasible solution.

1.2. Worldwide Scenario:

Covid-19 infects humans affecting their respiratory, digestive, liver, and central nervous systems. There have been 39,801,612 confirmed cases of Covid-19, including 1,110,908 deaths globally in 213 countries according to the world health organization as of 19th October 2020. The rise in the number of cases is because there are no targeted medicines and vaccines available. And the fact that it usually takes 2 to 14 days after the viral infection to detect Covid-19 [1].

1.3. India amid the pandemic:

In India from the 3rd of January till 19th of October there have been 7,550,273 confirmed cases of Covid-19 with 114,610 deaths. Thereby making this country the largest and the worst affected country by this pandemic. So we as a team went ahead in showcasing a working SIR model and web application for visualising the data in India. However we do not want our work to be restricted so in this report though we will be discussing about India and various factors surrounding India as a whole but our actual model is powered enough to predict and visualise the data of several different countries based on the given input data.

2. Problem Statement

Government agencies and health officials have so many dashboards and statistics around the Covid-19 spread throughout the internet. However what they lack is in visualising the future metrics and in adopting the right strategy to prevent the spread of the Covid-19 through external factors such as social distancing and lockdown.
The end result of this project report will address the following:

- As a government official, I want to strategise a viable approach to stop the spread of Covid-19 through visualisation and SIR model.

3. Data

CSV format data for the simulation of model is the one provided by the John Hopkins University's Centre for Systems Science and Engineering (JHU CSSE) [2].
Data for visualisation is fetched from the API provided in the postman website [3].

4. Methodology

SIR Model:

The SIR model is a simple model which comes under compartmental models. This model is mainly used to predict the infectious disease reasonably in a large population where there is a transmission of disease from human to human.
This model is divided into three compartments where each compartment consists of population number. These compartments are denoted with the letters *S, I* and *R*. Where,

- ***S***: indicates the number of susceptible individuals, who are in contact with the infected individuals and become suspects of carrying disease/virus after contacting them.
- ***I***: indicates the number of infected individuals, who are having the disease/virus and can transmit to susceptible individuals.
- ***R***: indicates the number of recovered individuals, who are recovered from the disease/virus or died.

The number of susceptible, infectious and recovered individuals may vary over time even though the total population size remains constant, we represent these numbers as a function of t (time) i.e., S(t), I(t) and R(t).

This model predicts the future scenario of any epidemic of an infectious
disease like COVID-19 over a period as implied by the function of t. During an epidemic like COVID, the number of susceptible individuals decreases as they will be infected by a disease and moved to infected or recovered compartments [4].

Each member in a population passes through Susceptible, Infectious and Recovered is shown in below diagram.

Susceptible individuals move to Infectious compartment because of contact between infected person and suspected person. The rate of new infections is represented as β *S I* where β is a constant [5].

- There is no way that individuals can enter or leave the susceptible compartment. Therefore, the first differential equation is,

$$\frac{d}{dt} S = -\beta I S$$

- One more process is that the infected individuals can only enter to recovered or removed compartment. The other two differential equations are,

$$\frac{d}{dt} I = \beta I S - \gamma I$$

$$\frac{d}{dt} R = \gamma I$$

Where γI the rate of individuals entering into recovered compartment.

- The total population i.e., S + I + R is always constant, therefore

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = -\beta I S + (\beta I S - \gamma I) + \gamma I = 0$$

5. Final Visual Interpretation

5.1. Initial Analysis:

Even though our model is powered keeping in mind that it should not be restricted geographically. However, for the better understanding of informative statistics, visualisation and in regard with this report our below examples are confined with the Indian sub-continent and several provincial terrains that are in relation with India as a whole.

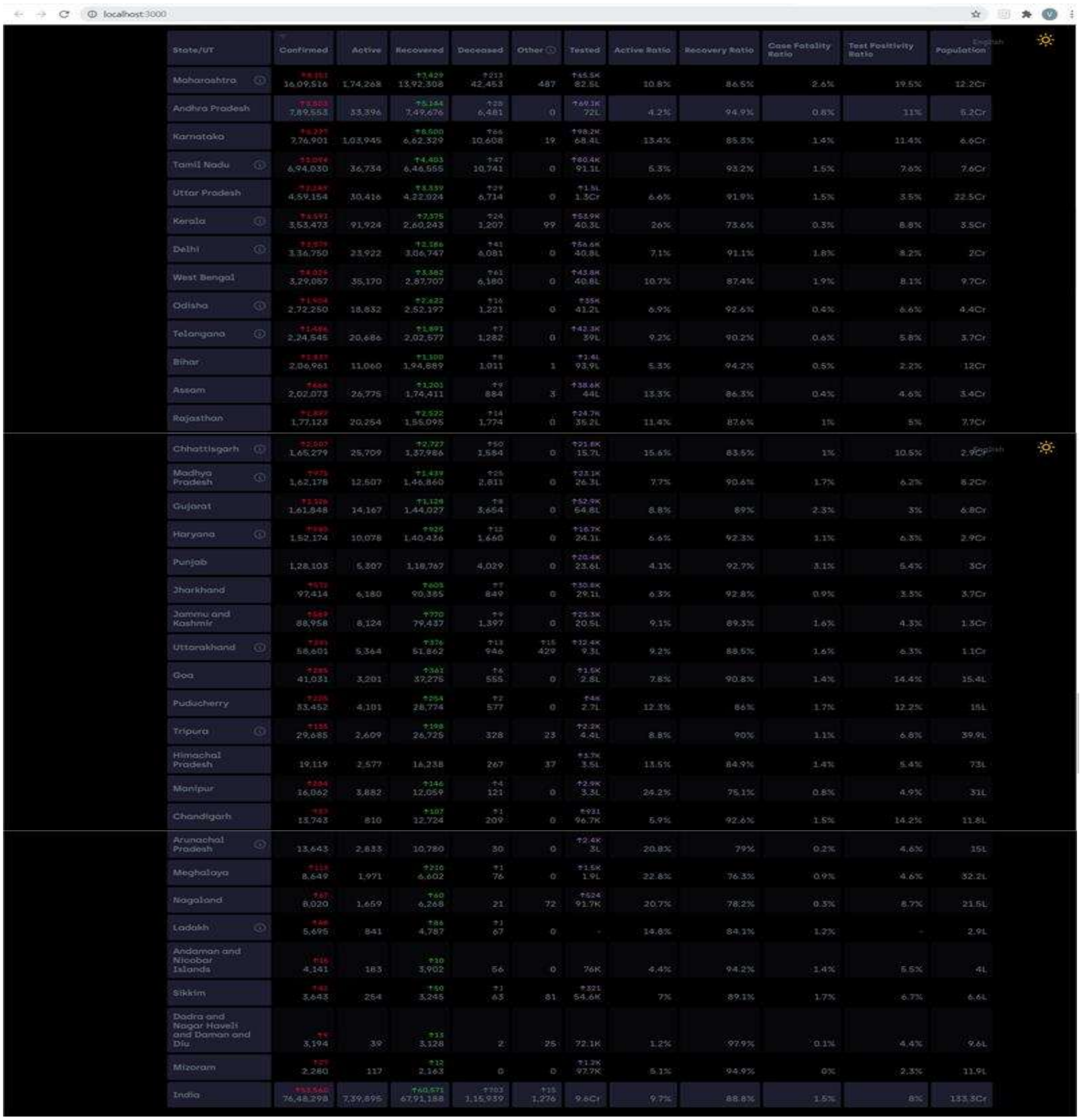
5.2. Features of Our Web Application:

- The top four tabs show the results with respect to confirmed, active, recovered, and deceased cases. Upon user selection the heat map visualises the same.
- The concentration of the above features can be seen as circles, coloured borders and area based layered colouring.
- Our web app also enables to see the end user several different attributes over a period of time either linearly or logarithmically based on the user point if view.
- Also it enables the end user if he wants to see the data based on a cumulative approach over a period of time as in 3 months, 1 month or daily.
- Apart from the above our web app also provides a search bar for the end user to search for a region to get the data insights of that particular region.

- And on top of the visualisation our end user will always be able to find the cumulative number of samples tested that will be updated daily.

5.3. Descriptive Analysis:

Below given visualisations are developed exclusively based on our web based application and explains in detail several descriptive statistics namely informative statistics such as confirmed cases, active cases, recovered cases, number of deaths, tested, active ratio , population with respect to different states within India.



State/UT	Confirmed	Active	Recovered	Deceased	Other	Tested	Active Ratio	Recovery Ratio	Case Fatality Ratio	Test Positivity Ratio	Population
Maharashtra	16,09,516	1,74,268	↑7,429 13,92,308	↑213 42,453	487	↑65.5K 82.5L	10.8%	86.5%	2.6%	19.5%	12.2Cr
Andhra Pradesh	7,89,553	33,396	↑5,144 7,49,676	↑25 6,481	0	↑69.1K 72L	4.2%	94.9%	0.8%	11%	5.2Cr
Karnataka	7,76,901	1,03,945	↑8,500 6,62,329	↑66 10,608	19	↑98.2K 68.4L	13.4%	85.3%	1.4%	11.4%	6.6Cr
Tamil Nadu	6,94,030	36,734	↑4,403 6,46,555	↑47 10,741	0	↑80.4K 91.1L	5.3%	93.2%	1.5%	7.6%	7.6Cr
Uttar Pradesh	4,59,154	30,416	↑3,339 4,22,024	↑29 6,714	0	↑1.5L 1.3Cr	6.6%	91.9%	1.5%	3.5%	22.5Cr
Kerala	3,53,473	91,924	↑7,375 2,60,243	↑24 1,207	99	↑53.9K 40.3L	26%	73.6%	0.3%	8.8%	3.5Cr
Delhi	3,36,750	23,922	↑2,166 3,06,747	↑41 6,081	0	↑56.6K 40.8L	7.1%	91.1%	1.8%	8.2%	2Cr
West Bengal	3,29,057	35,170	↑3,382 2,87,707	↑61 6,180	0	↑43.8K 40.8L	10.7%	87.4%	1.9%	8.1%	9.7Cr
Odisha	2,72,250	18,832	↑2,622 2,52,197	↑16 1,221	0	↑35K 41.2L	6.9%	92.6%	0.4%	6.6%	4.4Cr
Telangana	2,24,545	20,686	↑1,891 2,02,577	↑7 1,282	0	↑42.3K 39L	9.2%	90.2%	0.6%	5.8%	3.7Cr
Bihar	2,06,961	11,060	↑1,100 1,94,889	↑8 1,011	1	↑1.4L 93.9L	5.3%	94.2%	0.5%	2.2%	12Cr
Assam	2,02,073	26,775	↑1,201 1,74,411	↑9 884	3	↑38.6K 44L	13.3%	86.3%	0.4%	4.6%	3.4Cr
Rajasthan	1,77,123	20,254	↑2,522 1,55,095	↑14 1,774	0	↑24.7K 36.2L	11.4%	87.6%	1%	5%	7.7Cr
Chhattisgarh	1,65,279	25,709	↑2,727 1,37,986	↑50 1,584	0	↑21.8K 15.7L	15.6%	83.5%	1%	10.5%	2.9Cr
Madhya Pradesh	1,62,178	12,507	↑1,439 1,46,860	↑25 2,811	0	↑23.1K 26.3L	7.7%	90.6%	1.7%	6.2%	8.2Cr
Gujarat	1,61,848	14,167	↑1,128 1,44,027	↑8 3,654	0	↑52.9K 64.8L	8.8%	89%	2.3%	3%	6.8Cr
Haryana	1,52,174	10,078	↑925 1,40,436	↑12 1,660	0	↑16.7K 24.1L	6.6%	92.3%	1.1%	6.3%	2.9Cr
Punjab	1,28,103	5,307	1,18,767	4,029	0	↑20.4K 23.6L	4.1%	92.7%	3.1%	5.4%	3Cr
Jharkhand	97,414	6,180	↑605 90,385	↑7 849	0	↑30.8K 29.1L	6.3%	92.8%	0.9%	3.5%	3.7Cr
Jammu and Kashmir	88,958	8,124	↑770 79,437	↑9 1,397	0	↑25.3K 20.5L	9.1%	89.3%	1.6%	4.3%	1.3Cr
Uttarakhand	58,601	5,364	↑376 51,862	↑13 946	↑15 429	↑12.4K 9.3L	9.2%	88.5%	1.6%	6.3%	1.1Cr
Goa	41,031	3,201	↑361 37,275	↑6 555	0	↑1.6K 2.8L	7.8%	90.8%	1.4%	14.4%	15.4L
Puducherry	33,452	4,101	↑254 28,774	↑2 577	0	↑46 2.7L	12.3%	86%	1.7%	12.2%	15L
Tripura	29,685	2,609	↑198 26,725	328	23	↑2.2K 4.4L	8.8%	90%	1.1%	6.8%	39.9L
Himachal Pradesh	19,119	2,577	16,238	267	37	↑1.7K 3.5L	13.5%	84.9%	1.4%	5.4%	73L
Manipur	16,062	3,882	↑146 12,059	↑4 121	0	↑2.9K 3.3L	24.2%	75.1%	0.8%	4.9%	31L
Chandigarh	13,743	810	↑107 12,724	↑1 209	0	↑931 96.7K	5.9%	92.6%	1.5%	14.2%	11.8L
Arunachal Pradesh	13,643	2,833	10,780	30	0	↑2.4K 3L	20.8%	79%	0.2%	4.6%	15L
Meghalaya	8,649	1,971	↑210 6,602	↑1 76	0	↑1.5K 1.9L	22.8%	76.3%	0.9%	4.6%	32.2L
Nagaland	8,020	1,659	↑60 6,268	21	72	↑524 91.7K	20.7%	78.2%	0.3%	8.7%	21.5L
Ladakh	5,695	841	↑86 4,787	↑1 67	0	-	14.8%	84.1%	1.2%	-	2.9L
Andaman and Nicobar Islands	4,141	183	↑10 3,902	56	0	76K	4.4%	94.2%	1.4%	5.5%	4L
Sikkim	3,643	254	↑50 3,245	↑1 63	81	↑321 54.6K	7%	89.1%	1.7%	6.7%	6.6L
Dadra and Nagar Haveli and Daman and Diu	3,194	39	↑13 3,128	2	25	72.1K	1.2%	97.9%	0.1%	4.4%	9.6L
Mizoram	2,280	117	↑12 2,163	0	0	↑1.3K 97.7K	5.1%	94.9%	0%	2.3%	11.9L
India	76,48,298	7,39,895	↑60,571 67,91,188	↑703 1,15,939	↑15 1,276	9.6Cr	9.7%	88.8%	1.5%	8%	133.3Cr

Fig.no-1: Data of different Indian States through our web application.

The below given visualisation from our web application shows the data of India as a whole with several different informative statistics including confirmed cases, recovered cases, and active cases being spread out geographically, along with their respective descriptive graphs over a period of time on the right thereby providing valuable data insights to the end user.

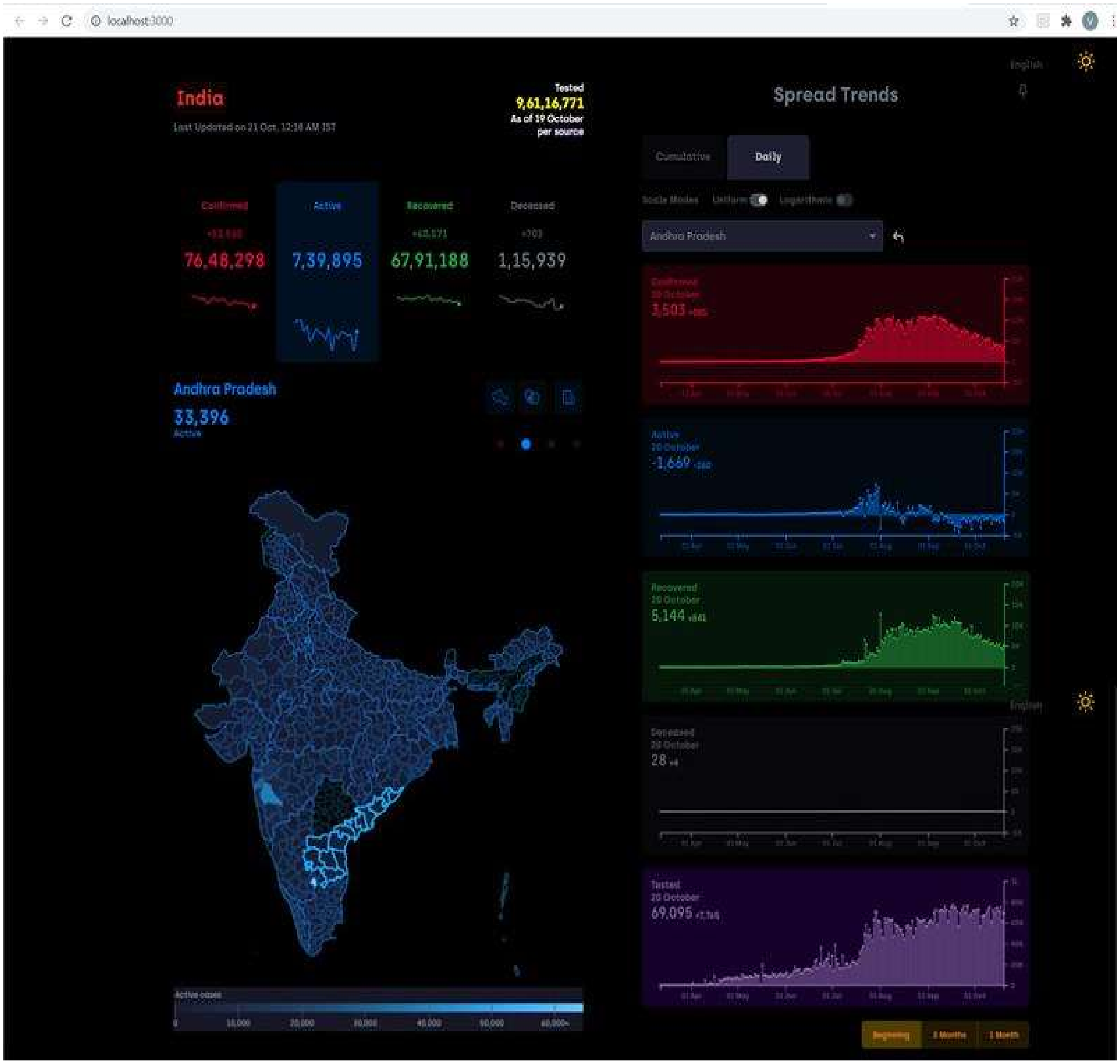


Fig.No-2: Visualisation of Active Cases in India through our web application.

In addition to the above visualisation, this tab below enables the user to better visualise the number of recovered cases in India as a whole and further have a look at the descriptive statistics as graphs over a period of time as shown in the right side of the image same as in case of the above image. Further explaining the descriptive statistics they are the linear representation of different tabs such as recovered cases, confirmed cases etc. Over a period, which here is from the start till the current day, which is the 20th of October.

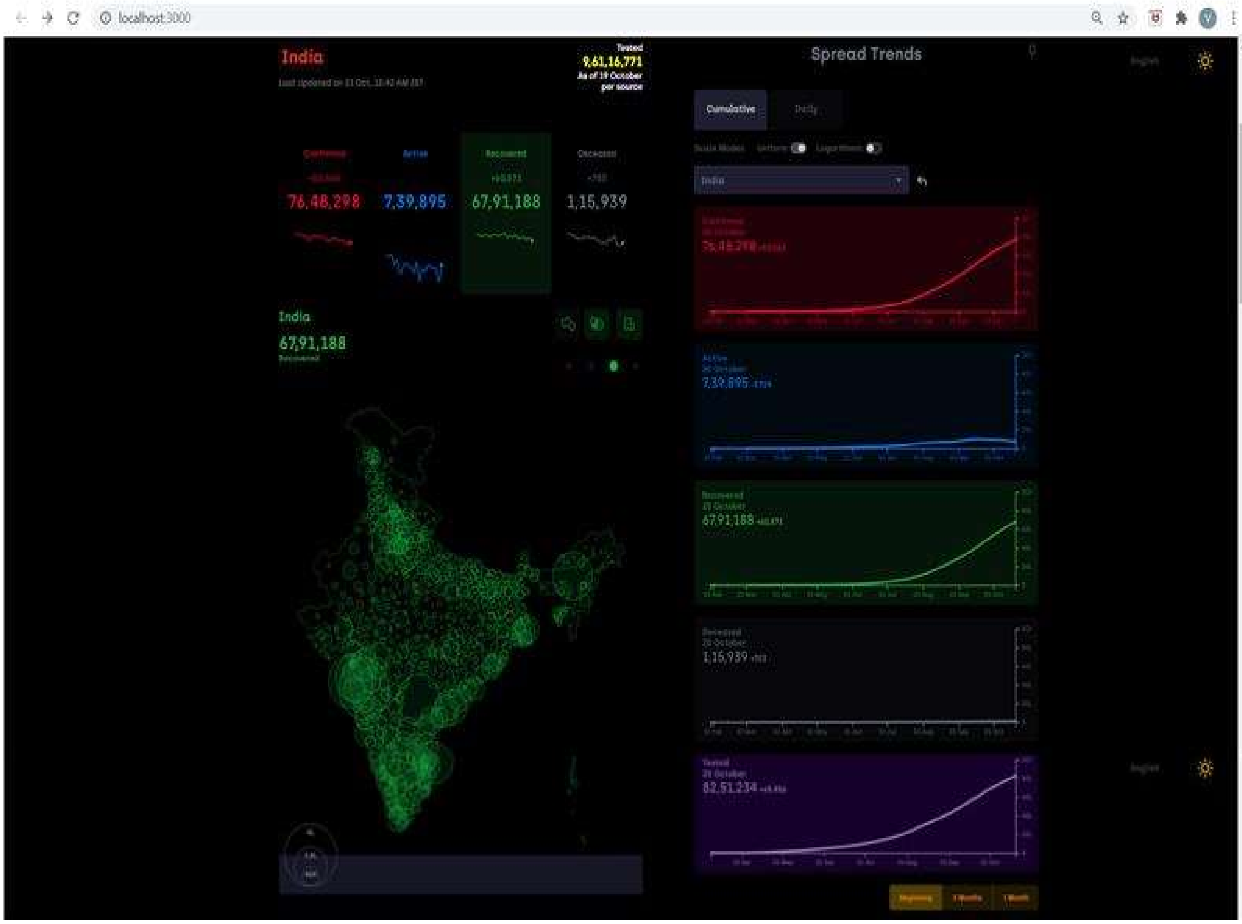


Fig.No-3: Visualisation of Recovered Cases in India through our web application.

Now if the end user wants to further analyse the data with respect to states or districts within the states, Our web application also enables that option and by clicking the state icon over the map or by selecting the state desired within the dashboard the user will be able to gain access to the below given visualisation.

This visualisation as a whole is a collection of several data insights that was designed keeping in mind the simpler way to communicate valuable data insights to the end user. This visualisation is a combination of several descriptive statistics, bar graphs, Heat map, and linear graphs where each of them clearly represent the data insights with respect to their attribute. And the visualisation is as follows.

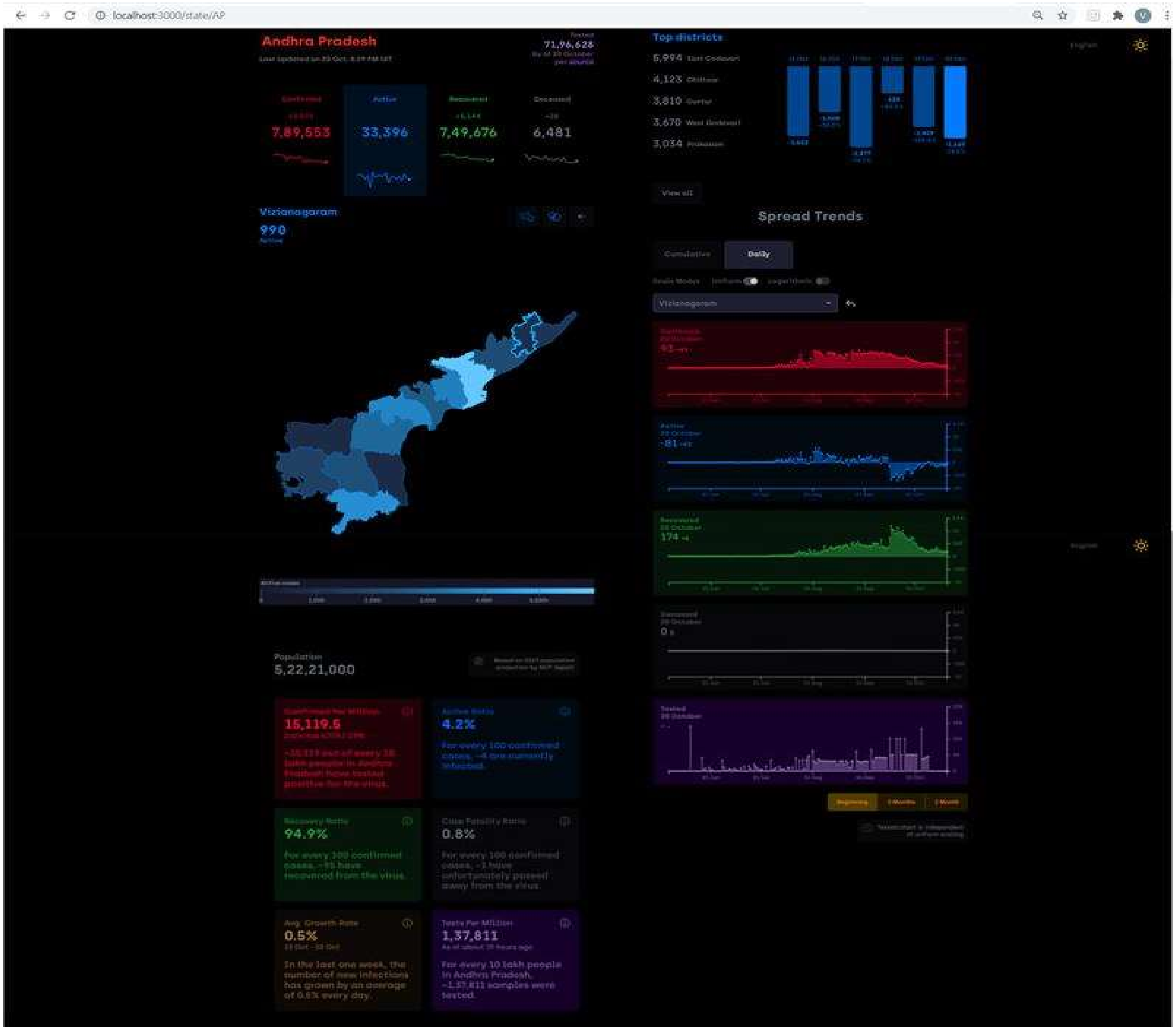


Fig.No-4: Visualisation of a State in India through our web application.

5.4. SIR model interpretation

Let us consider the infection is spreading with certain rate and to control these spread certain actions are implemented like social distancing and lockdown. Due to these actions, susceptible individuals do not expose to the infectious individuals.

Now consider the number of susceptible persons, an infected person infects on an average per timestamp is, $\beta = 0.25$ and the fraction of people that recover per timestamp is, $\gamma = 0.083$.

The current disease-spread parameter R_0 is 3, this value describes how many people an infected person infects on an average over of a period.

The numerical scheme (step length) computes new values for the number of susceptible, infected, and recovered patients for every time step. Typically, you can set this value to 1, meaning an updated will be computed for every day. If you only want to compute updates for every week, it would be 7. Some numerical errors can occur if this value is too large, but the runtime will increase if it is very small.

The scheme will keep computing steps until this value is reached. If the time step is 2 and this value is 40, 20 steps would be computed. Let us consider the time span of 300 days.

Since, there is no vaccine for COVID-19, so considering Immunity/vaccination rate as 0 and the base mortality as 0.006, it is the fraction of the cases, in which the infected person dies from the disease while the health care system is still operating normally (non-critical case).

For diseases, which require intense medical care, it can become relevant how many people are infected to determine mortality. For the corona virus, urgent cases require a ventilator to be treated. So, if there are more urgent cases than ventilators, the mortality increases dramatically because the people can no longer be treated. This situation is called critical. Hence, consider the critical mortality rate 1. Once the critical stage is reached, how high is the mortality? If, for example 5% of all cases require a certain device and with that device 80% of critical cases survive, the base mortality is 1% (5% * 0.2), the critical threshold is the number of devices per population and the critical mortality is 5% (all critical cases above the threshold die.

Critical mortality threshold describes, the infection rate must be passed to activate the critical mortality rate. For corona virus, this would be the number of ventilators per population. Let us consider this as 0.01.

Initially the pandemic started in China and spread within china at certain rate and gradually transmitted throughout the world. For instance, consider initial infected individuals in Chain as 1000 and run the SIR simulation with this sample data to predict the spread of virus and how long it lasts over time.

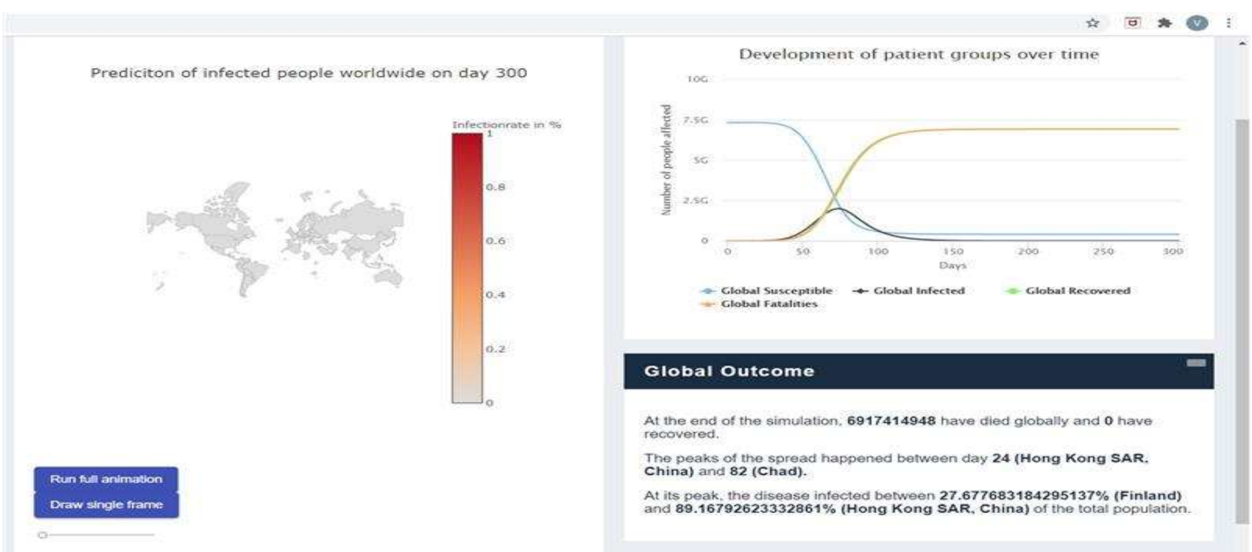


Fig.No-5: SIR model simulation globally

On running the simulation, we got above outcome as shown in figure. Over the time, Number of people infected, and susceptible individuals has decreased gradually and reached 0 at 140 days. At the end of the simulation i.e., on reaching 300 days the recovered number reaches to zero. Since, there are no infectious individuals from 140 days.

The same SIR stimulation can also be done on individual countries. As an example, we have taken India and ran the stimulation with the initial susceptible individuals 1.35 billion.

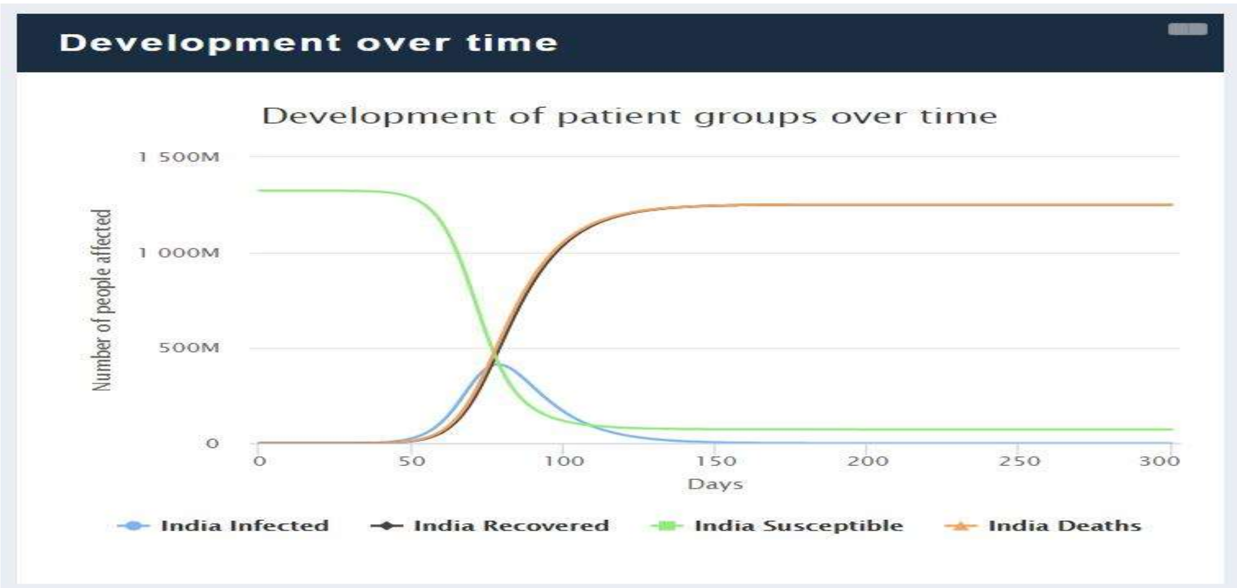


Fig.No-6: SIR model simulation on India

In need of controlling the infectious disease/virus, let us observe in depth based on individual measures such as social distancing and lock down.

Simulation with different social distancing factors,

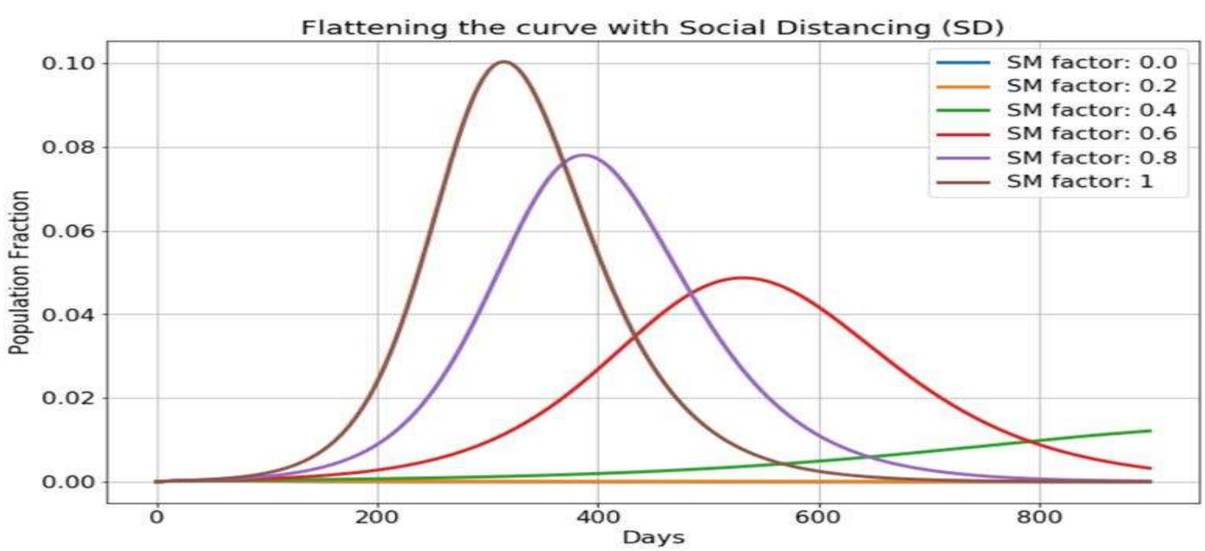


Fig.No-7: SIR model simulation based on social distancing factor

When the crowd violates the social distancing measures i.e., 100% of social mixing. There will be more infected cases. When social mixing factor is 0 i.e., strict social distancing is followed. In this case there are no infected cases. Therefore, on following strict social distancing measures will reduce the spread of virus.

Now, let us consider a scenario where there is an early lift in lock down and reduce in social mixing. Initially, consider certain value as a social distancing factor and run for certain time and increase the value of social mixing factor.

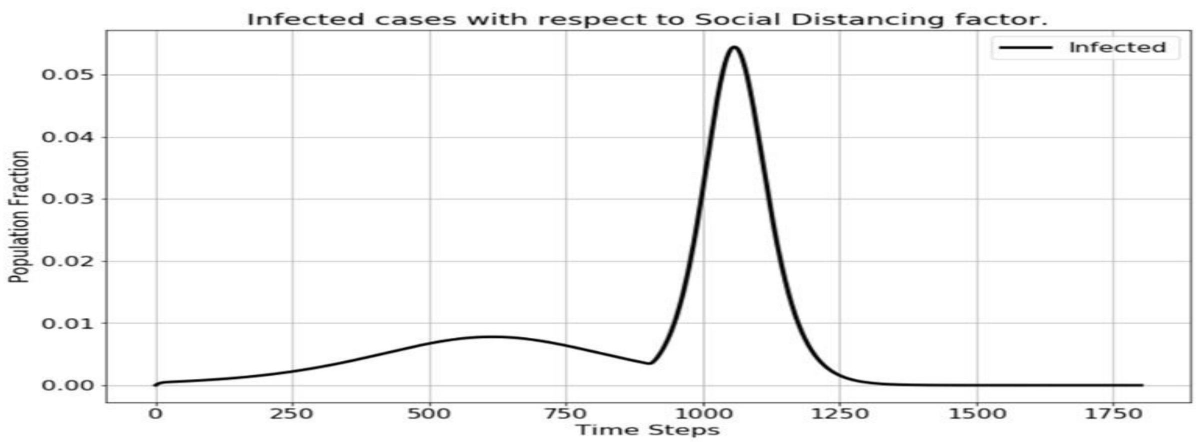


Fig.No-8: Infected cases w.r.t Social distancing factor on two runs.

As shown in above figure, on first run, the curve has less spike due to the implementation of lock down and social distancing measures i.e., less social mixing factor. On second run, when the lockdown and social distancing measures relaxed i.e., high social mixing factor, there is a huge spike in the curve which implies there are more infected people.

Lockdown period vs infected cases,

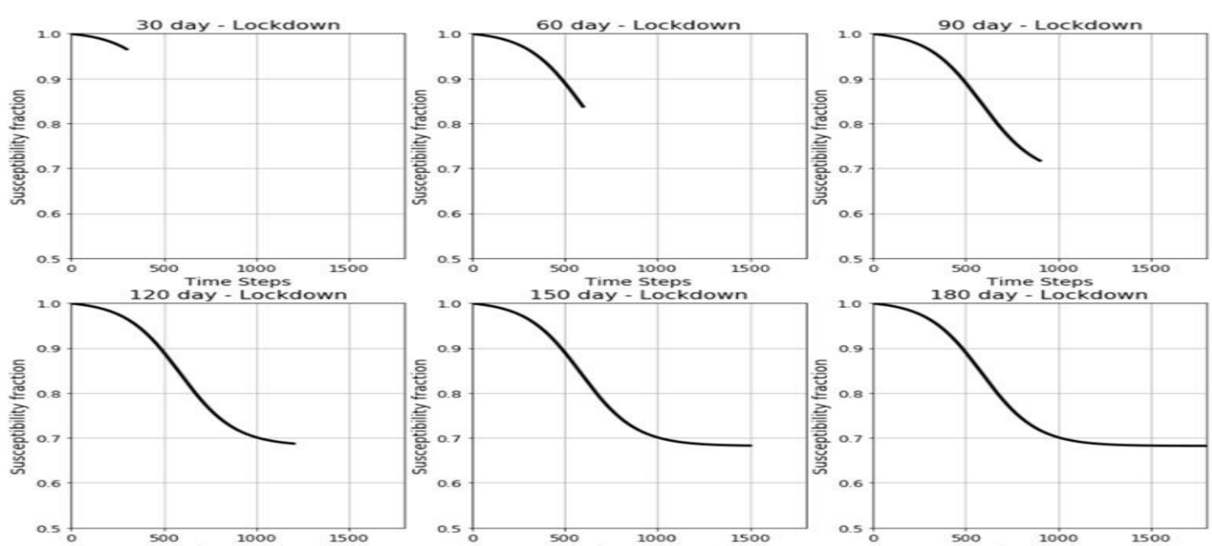


Fig.No-9: Lockdown simulations.

With a longer period of lockdown, the susceptible individuals reduce significantly and observed that after 150 days of lockdown there is no decrease in the susceptible fraction. The same has observed in figure 5 and 6. Therefore, again running the model two times with increasing social mixing factor and lockdown up to 150 days, the simulation of infected cases will be as shown below,

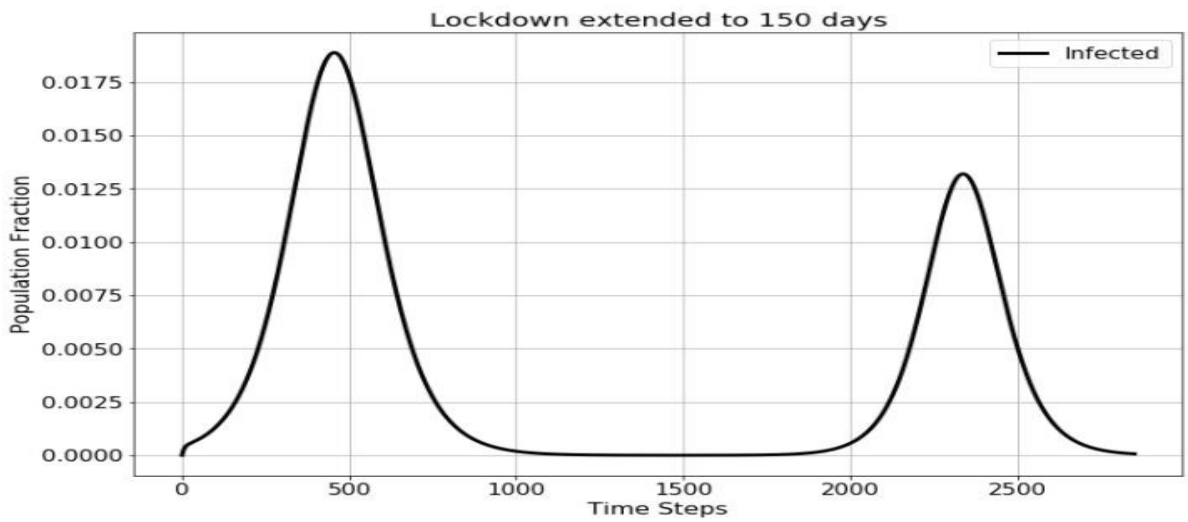


Fig.No-10: Infected cases when first lockdown of 150 days.

Now let us see the infected cases simulation when the lockdown is extended for 250 or greater than 150 days. There will be only 1.8% or 0.018 fraction increase in lockdown days as shown below,

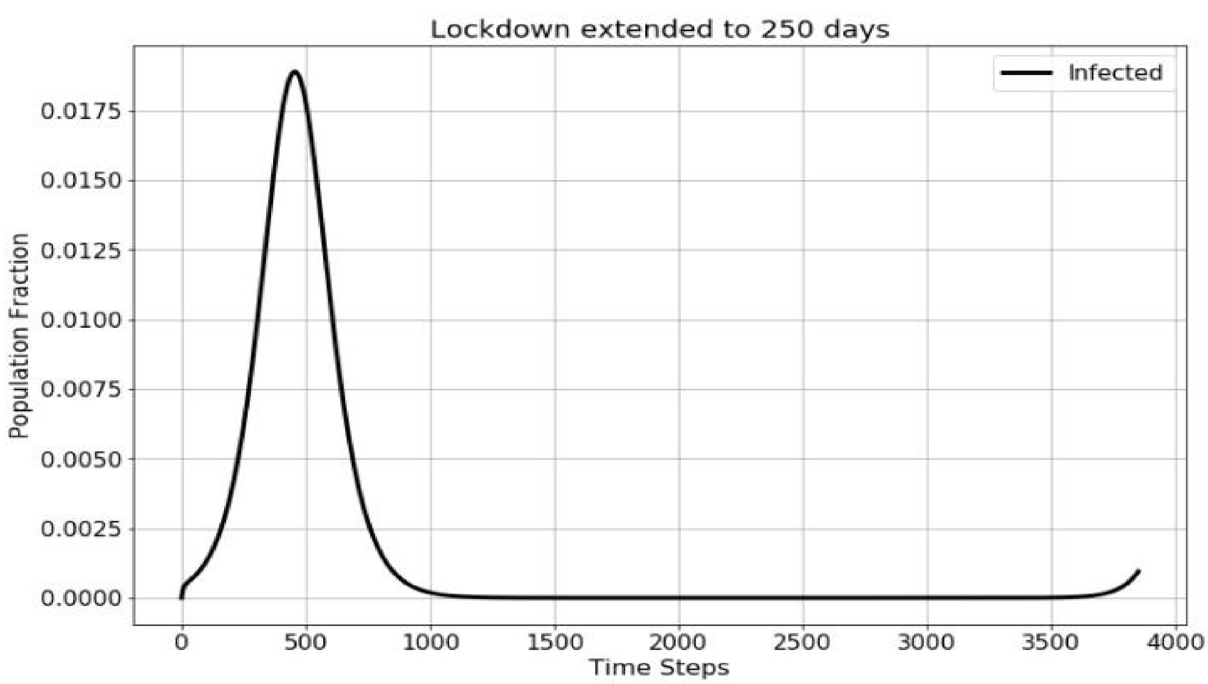


Fig.No-11: Infected cases when first lockdown of 250 days.

6. Interesting Findings

6.1. Effect of Social Distancing:

The idea of social distancing was proposed vastly by several different health organisations along with government officials. Also, our SIR model suggests that when there is a greater number of cases a 100% social distancing factor is required. So it is pretty much clear that social distancing factor as a whole plays a key role in stopping the spread of the virus. Further iterations when the number of cases decreases, the predictions of our SIR model and visualisations should be considered in easing the restrictions with respect to social distancing factor.

6.2. Effect of Lockdown:

A 21-day lockdown was announced by the Government of India banning all the movement and restricting Indians to stay at home. The citizens were allowed to step out only in emergency situations. All these instructions were given in the hope of flattening the curve of infected cases and to restrict the exponential growth of cases within the country. However, after the 21 days lockdown the country saw a steep increase in the number of cases.

7. Conclusion

Apart from the interesting findings the 21-day lockdown proved to be effective only for a temporary period. However our web based application along with the SIR model powers the government organisations in strategising an even better approach in devising the spread of the virus. As our SIR model suggests that a lock down over a longer period of time, 250 days as in case of India would have definitely made a long term effect on the nation in reducing and completely bringing the spread of the COVID-19 virus to a halt. Thereby our application along with our model has been proven successful in determining the futuristic trends as well as in assisting in strategizing a COVID safe plan.

8. Project Management

(darkest)	100%
	60%
	40%
(lightest)	20%

Project Planner

Project Phase	Sub Task	Resources	10-08-2020 - 14-08-2020	17-08-2020 - 21-08-2020	24-08-2020 - 28-08-2020	07-09-2020 - 11-09-2020	14-09-2020 - 18-09-2020	21-09-2020 - 25-09-2020	28-09-2020 - 02-10-2020	05-10-2020 - 09-10-2020	12-10-2020 - 21-10-2020
Requirement and Data Gathering	Collecting COVID 19 Global data.	Eswar	100%								
	Collecting individual country data.	Vishal, Manideep	100%								
	Setup Jupyter Notebook, Eclipse and NPM	All	100%								
Data Preprocesing and Exploration	Data cleaning	Eswar		100%							
	Data Integration	Manideep		100%							
	Exploring Data in jupyter	Vishal, Sujay		100%							
Modeling	Analysis of SIR model	ALL			40%	40%	20%				
	Implementation of SIR model	Vishal, Manideep				20%	20%	60%			
	Implementation of SIR model based on different measures	Vishal, Eswar					60%	60%			
Web Apps Development	Develop HTML and CSS code to visualise India COVID data	Eswar, Sujay						60%			
	Develop Javascript	Vishal, Manideep						20%	40%	40%	
	Develop python code and typescript for SIR model	Vishal, Manideep						20%	40%	40%	
	Develop HTML and CSS code to visualise SIR simulation	Eswar, Sujay						60%			
Integration	Integrate HTML and CSS code with Java script	Eswar,Vishal								20%	100%
	Integrate frontend and backend of SIR	Manideep, Sujay								20%	100%
Presentation	Design of slides - 1 to 3	Vishal							100%		
	Design of slides - 4 to 5	Manideep							100%		
	Demo preparation	All							100%		
	Video Presentation	Eswar							100%		
	Review and provide feedback on presentation slides	All							100%		
Final Report	Abstract, Introduction and problem statement	Eswar									100%
	Data and Methodology	Manideep									100%
	Results interpretation	Vishal									100%
	Discussion and conclusion	Sujay									100%

As a team we have divided the project into different phases where each phase involves every individual in a team (refer above Project planner table). We have maintained our work progress in Microsoft Excel and organized stand-up calls on every Monday to track individual tasks and used Microsoft teams as a medium of communication for sharing the work files.

Everyone in a team did tremendous job to finish the Project on time even though there are few delays in Modelling and Web App development. Starting from data gathering, every member in a team participated in stand-up calls and shared their ideas and plans on future tasks and moreover everyone is available to each other at any time. There are few tasks which are dependent on other tasks where members should coordinate and complete together, these kinds of tasks are shorted out by discussing and agreeing on one view.

Each member contributed equally to a team and gained knowledge on every bit of the project.

Student ID	**Student Name**	**Contribution (%)**
s3811521	Vishal Patnaik Damodarapatruni	25
s3798488	Eswar Phani Paruchuri	25
s3820822	Sri Venkata Manideepu Maddipati	25
s3794983	Sujay Kamal Madisetty	25

9. References

[1] [Online]. Available: https://covid19.who.int/region/searo/country/in.

[2] Johns Hopkins University, “GitHub,” [Online]. Available: https://github.com/CSSEGISandData/COVID-19.

[3] Post Man, [Online]. Available: https://documenter.getpostman.com/view/10808728/SzS8rjbc#b07f97ba-24f4-4ebe-ad71-97fa35f3b683.

[4] Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology.

[5] [Online]. Available: https://www.maplesoft.com/applications/download.aspx?SF=127836/SIRModel.pdf.

[6] [Online]. Available: Johns Hopkins University.