Data Analysis - COVID-19

Aryan Garg

July 4, 2020

Abstract

This report mainly focuses on determining how is India coping up with the pandemic and tries to explore various possibilities if certain actions were undertaken or not, by mathematically analyzing the statistics provided by a volunteer-driven, crowd-sourced database for COVID-19 patient tracing. The report keeps the country's economy and geographical state in mind to predict how India is faring as compared to other nations and also raises a few questions for medical sciences. The study uses purely mathematical reasoning and conjectures are explicitly mentioned. The assumptions taken are also justified or proved analytically wherever they are brought up. We also look at an interesting case-study at the end. We use a point-based system to determine whether H1N1 or COVID-19 is more deadly by weighing the official statistics released till now with the technological and medical advancements in time till that year and the time of response to the virus by the world in both cases to get a fair and true estimate.
This study is broad in general and aims to make practical predictions and estimates.

1 Introduction

> "If life were predictable it would cease to be life, and be without flavor."
>
> –Eleanor Roosevelt

With that in mind, we begin our journey by predicting one of the most pivotal questions which determines if a country will revive soon enough or not, and that is, increasingly better recovery rates. We draw inferences and observations from the plots of the data points and try to reach a conclusion to validate or invalidate our proposed hypothesis. We also compare India with some countries who managed to control the infection spread and decrease the ever-growing cases. We assess if India is on the same path as them and where exactly, if it is. The hypothesis shows us some interesting factors and also, in a way, provides us with information to stay ahead of the outbreak.

Then we assess the current situation rather than the cumulative total stats in detail by observing different parameters and try to predict which age-groups are highly affected and which areas across the country are likely to be harder hit. Population statistics aren't up-to date but the variance in percentage is assumed to be fairly low and is a good estimate for mathematical computation.

Then we answer whether this is the deadliest epidemic that the human race has faced or not. We compare it with the H1N1 virus, which had caused some unrest in the world.

2 Better Recovery Rates?

2.1 Overall Cases

Now let's begin by looking at the total cases in India: date-wise.

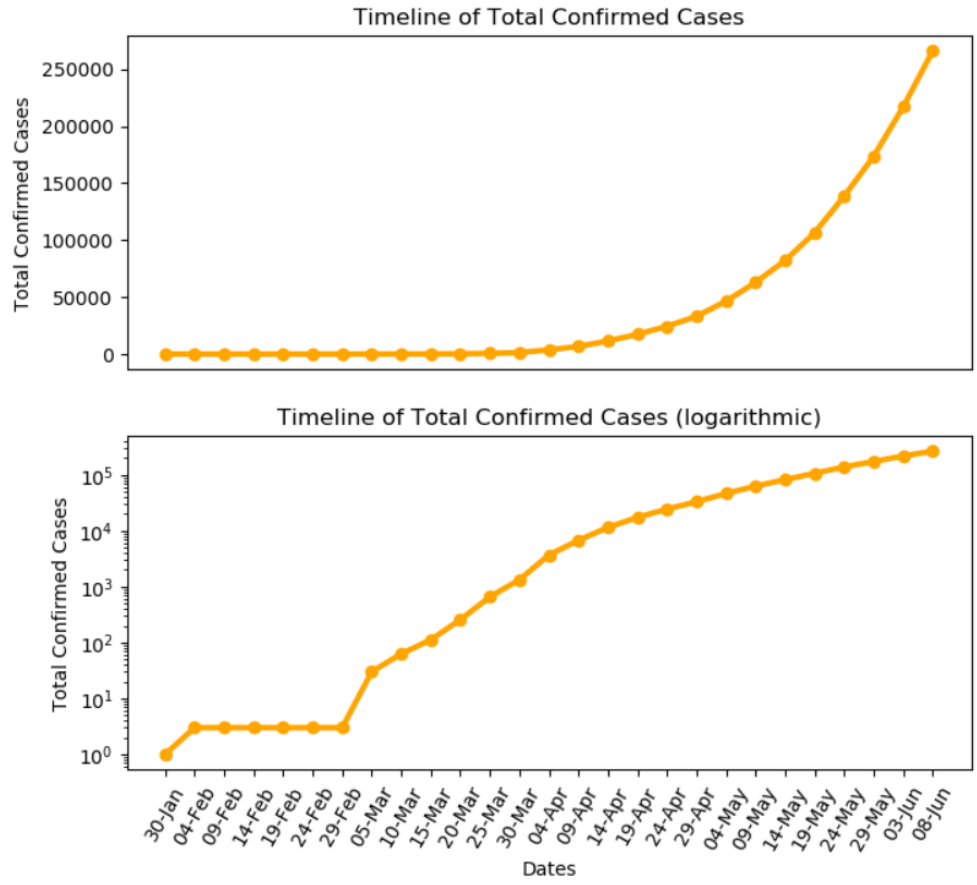


Figure 1: Total Confirmed Cases from 30th January to 8th June

`Observation: Total cases are rising exponentially.`

2.2 Total Deceased Count

Let's look at the actual and logarithmic plots (same as above):

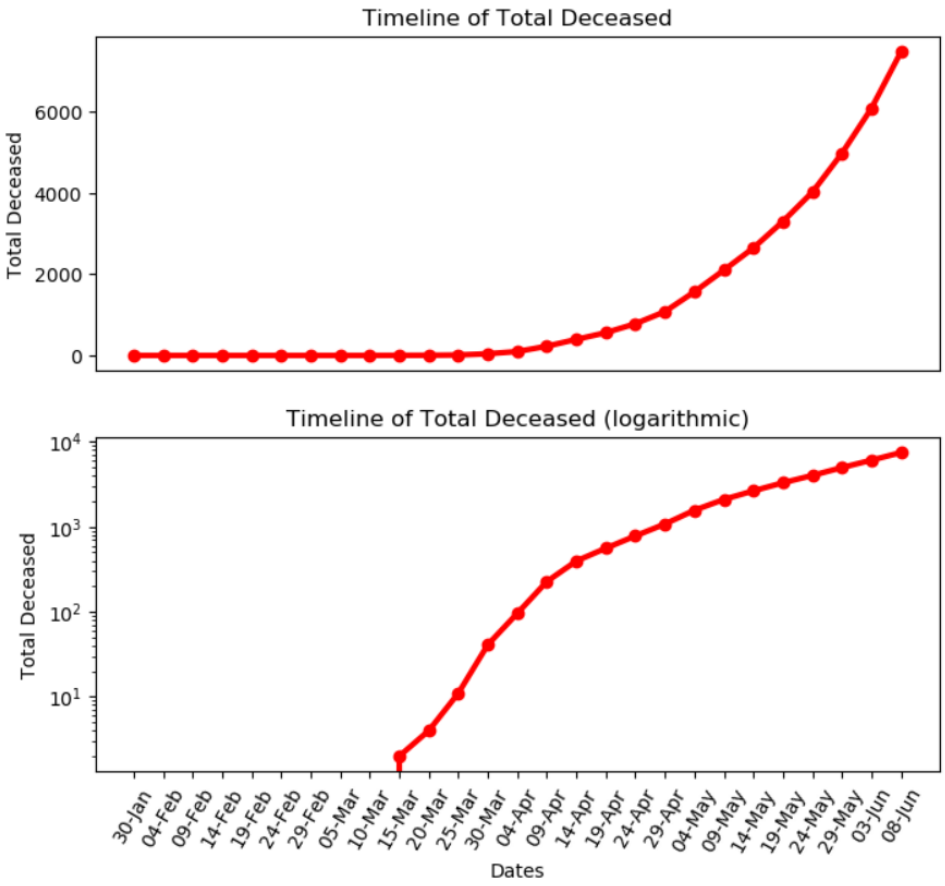


Figure 2: Total Deceased from 30th January to 8th June

`Observation: Total deceased are also rising exponentially. But somewhat at a slower rate than total cases`

2.3 The Hypothesis

Q: But how slow? What's the rate between the two?
A: We'll plot a timeline of total deceased per 100 cases to estimate that.

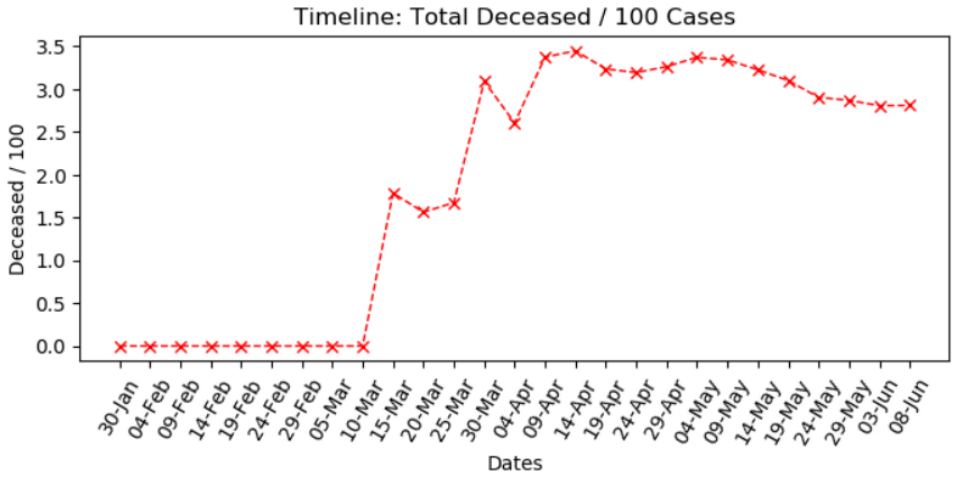


Figure 3: Timeline of Deaths per 100 cases:

`Observation 1: To say that the infection spread is slowing would be wrong by looking solely at Figure 3 ( Deceased / 100 cases ) because total confirmed cases are rising exponentially, but nonetheless, slower than some other countries like Brazil, USA when they were around the same cases as India.`

`Observation 2: The lock-down has certainly slowed the rate of cases in India and hence the deceased per 100 rates have stabilized.`

Verifying Observation - 1 : Growth rate of confirmed cases

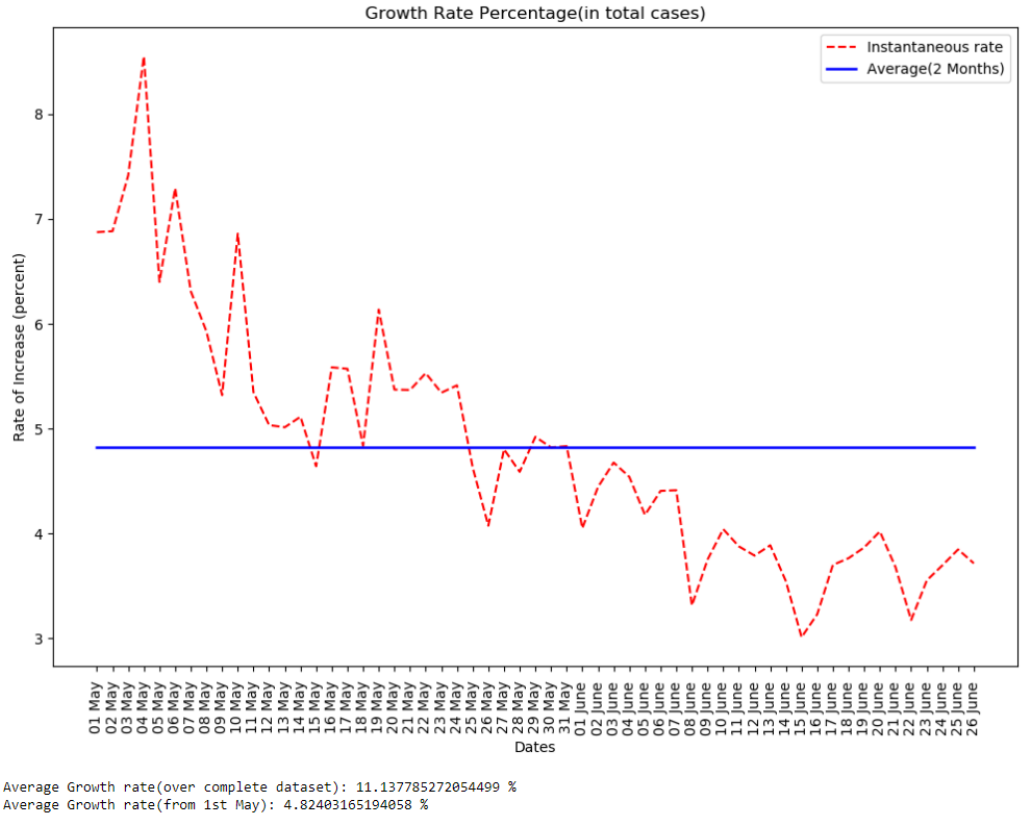


Figure 4: Growth Rate : Exponentially Decaying

Hypothesis :
India has better recovery rates which implies India is coping up with the virus and will recover from the pandemic.
Reasoning: The hypothesis is based on the above 2 observations as observation 1 clearly states that cases are NOT slowing down while observation 2 states that the deceased per hundred has somewhat stabilized. So, to achieve the net stabilization of deceased/100, when the denominator is rising exponentially (see formula below),

$$Deceased/100(on\ that\ date) = Total\ Deaths/Cases(till\ that\ date) * 100 \quad (1)$$

death rates must be also rising at the same rate. But if recovery rates rise at a steeper rate than the above two, the recovered plot will start equalling the number of cases and from there on-wards, everyone would recover, decreasing the spread eventually.

But before looking at recovery stats a few questions that come to mind after

seeing the 3 figures:

Q1 : Could there be a huge spike again? Just like there was at 10th March, 25th March, 5th April because the lock-down is lifted again and maybe the virus has evolved due to the new asymptomatic cases?

Q2: The rates for total deceased/100 have stabilized. Are people naturally developing immunity against the virus? Or people who are getting infected more than once are also included in the data who have some antibodies which are able to cope up with the virus? (A question for medical sciences)

2.4 Evidence: For or Against?

Now let's analyse the recovery statistics to give a concrete statement for or against the hypothesis!

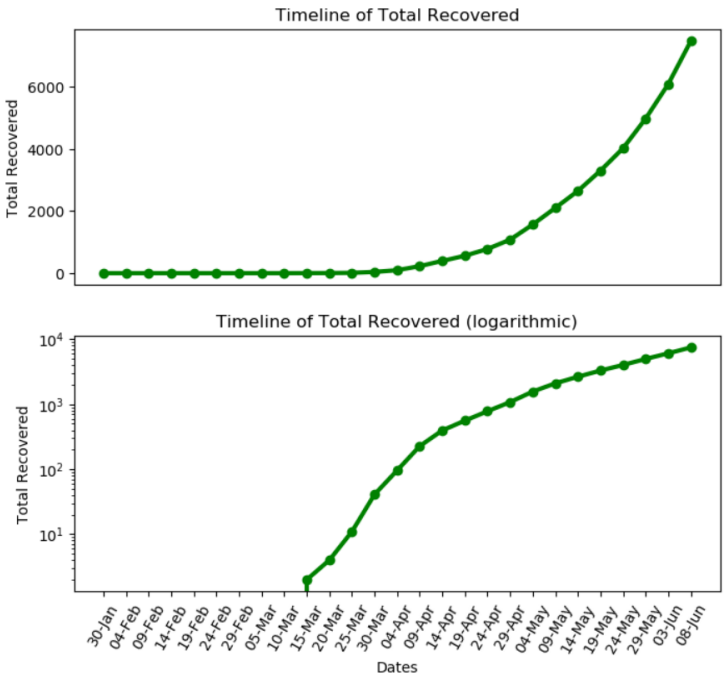


Figure 5: Timeline of Total Recovered

This is a good indication for our hypothesis but certainly doesn't prove anything yet. Let's look at Recovered per hundred cases:

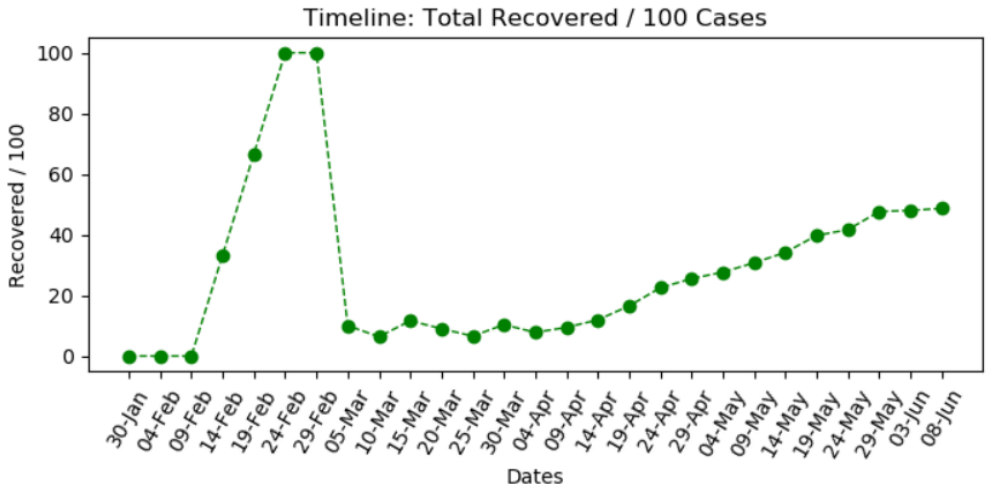


Figure 6: Timeline of Total Recovered per 100 cases

So our hypothesis could be true!

Reasoning:
Deaths per hundred have stabilized while the number of recovered patients in the same time frame are continuously rising. The total recovered timeline (Figure 4) also supports our initial reasoning of exponential recovery. So, far all the evidence is in support.

To get a yet better understanding, let's look at all the three stats together in the same plot and also the per hundred (percentage) plots and conclude.

2.5 Conclusion of the Hypothesis

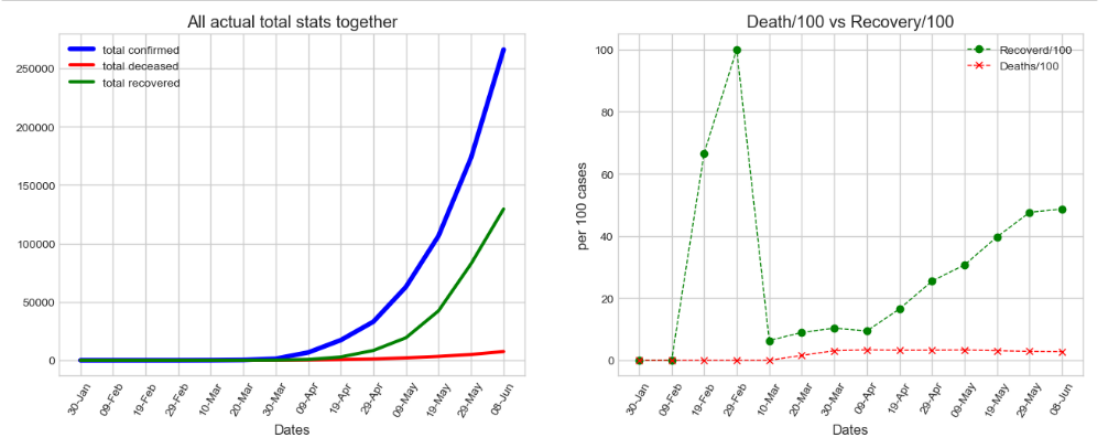


Figure 7: Total cases, deceased and recovered

`Observation 1: The cases are rising way faster than recovery and the total deaths have started rising slowly but it's a significant change.`

`Observation 2: There appears to be a slight plateau in the recovered per 100 cases plot`

Conclusion:
For now, our hypothesis holds true in all its might mathematically. But if the instantaneous trends stay the same and based on the two observations above(especially observation 2 - if the plateau continues); the number of cases, which are rising quicker, may at a point, outweigh the effects of recovery rates and the death rate might start catching up-to the recovery rate. In that case the hypothesis fails due to non-mathematical reasons.

Also, India has not reached it's peak yet like some other countries have. **Nonetheless, we know that the hypothesis is true until now so we'll build upon the hypothesis further. We'll try to find where is India on the path of revival? How much time will India take?**

2.6 The Path of Revival (Hypothesis Result)

So, let's look at Japan, a country which has managed to control the spread quite successfully.(In plots' x-axis: Days is wrongly typed as dates)
Japan:

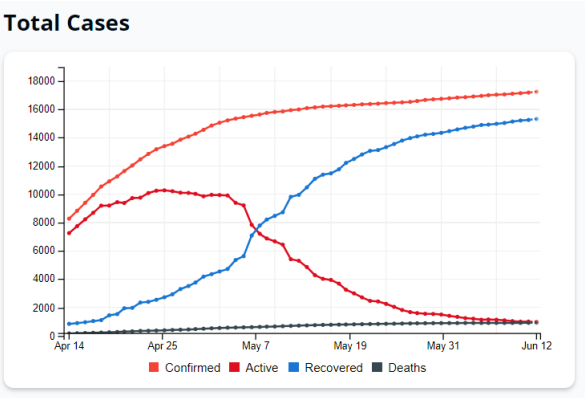


Figure 8: All stats of Japan together

Clearly, India is on a similar path and if we interpolate the recovery graph from stable rise (around 6th March) we can get number of days till total recovery!

Let's make a simple Linear Regression and a Non-Linear Regression model (SVR - RBF Kernel) to estimate time, from this date on-wards, required by India to achieve similar results like Japan.

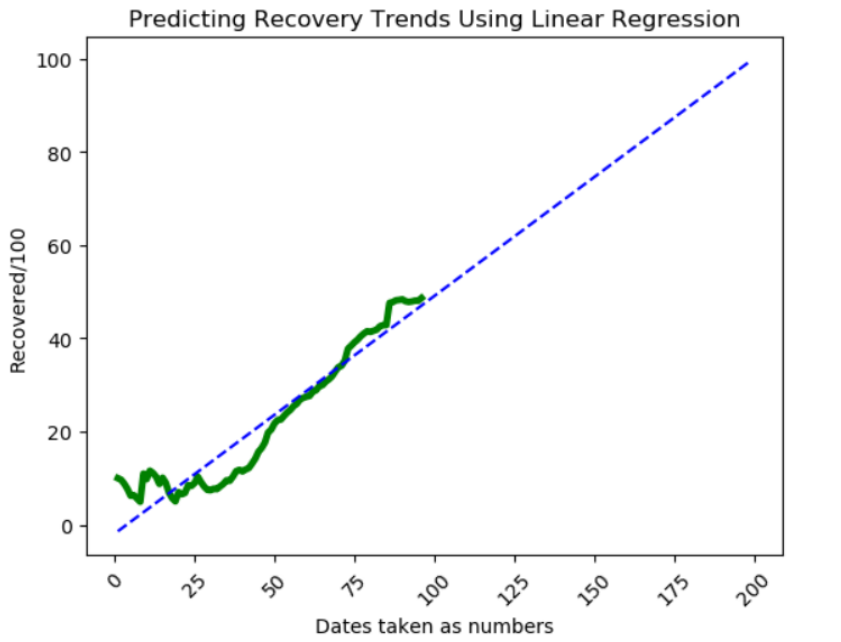


Figure 9: Predicting Number of Days Till Total recovery using Linear Regression :

`Observation 1: This model predicts about another 4 months(120 days) till total recovery in India.`

`Observation 2: It seems a little unreasonable considering the fact that recovery/100 cases is, by nature, a logarithmic function as seen in Japan's plots.`

So, let's build a more realistic predictive model: The Non-Linear Regressor –> SVR

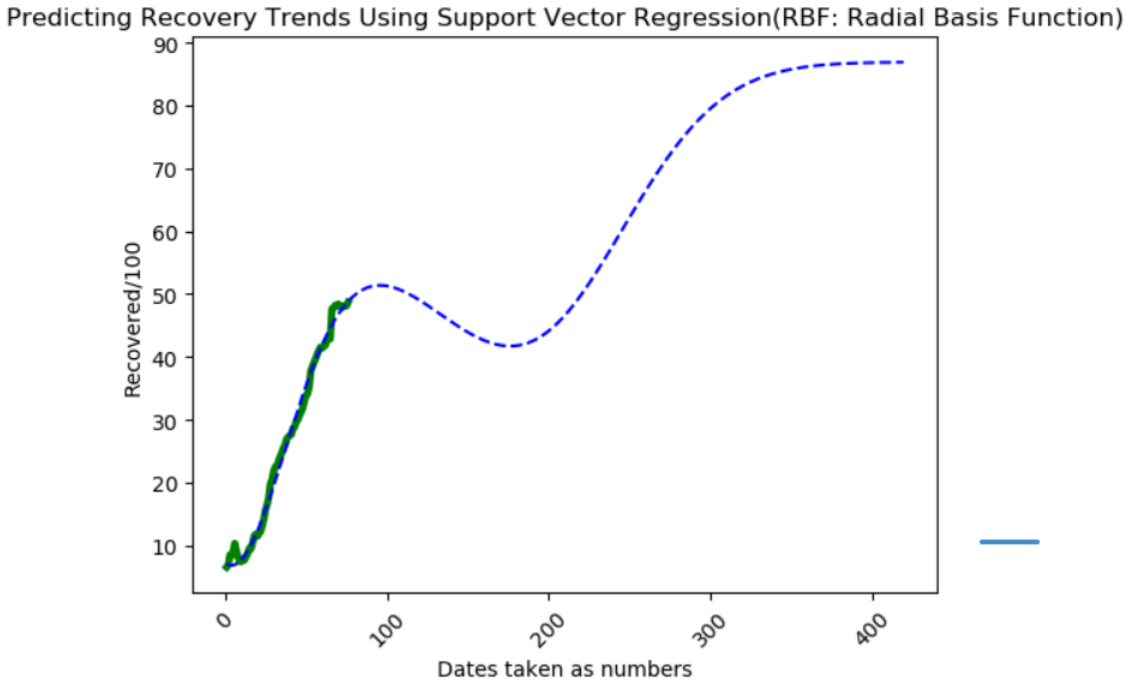


Figure 10: Predicting Number of Days Till Total Recovery using Support Vector Regression(params: gamma=40000, C=0.0001)

`Observation: This model(SVR) predicts another 300 days or so, i.e, 10 months. This seems reasonable considering the fact that it starts plateauing as it reaches the end, i.e, behaves like a logarithmic function.`

Finding: *So, the hypothesis predicts around 9-10 more months till total recovery!*

But how can we validate the model? Know whether it is accurate or not?

2.7 Validation

We started out with Japan's success in mind. If the model is truly predictive of the recovery rates, then it must trace Japan's actual curve.
To validate our model, we will run it (with exactly the same parameters) on 30% of Japan's data and plot the actual curve simultaneously to verify our model:

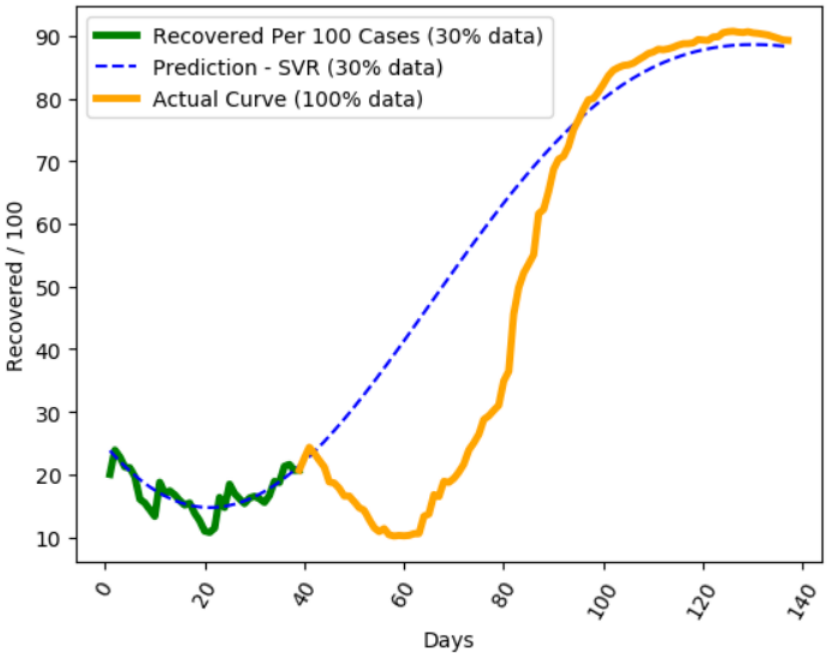


Figure 11: Running the model on 30% of Japan's Recovery per Hundred Data(params: gamma=40000, C=0.0001)

Result:
It is safe to say that our SVR is quite competent with what we started as reference in mind. And we can continue to build on the hypothesis till it is disproved or it starts to deviate from reality to an unacceptable extent.

3 Can The Hypothesis Fail?

In an ideal world, India would somewhat follow the predicted curve and recover but some major factors that could ruin those chances would be the lack of medical facilities and equipment, shortage of medical professionals or unexpected evolution of the virus for the worse which are unaccounted for in the hypothesis. We intend to see if the availability of medical facilities could throw India off the path of recovery, and if so, what is needed to get back on track.

3.1 Understanding Serious Cases in India

It is obvious that serious cases require special medical attention which is bound to be scarcely available for such a huge population. That is why patterns in serious/critical cases could give us an estimate as to what might be coming, or are we already prepared for it. (All the calculations are done from 22nd March on-wards.)
It is imperative to note that serious patients, if unattended, might become the concentrated carriers of the virus eventually turning hot-spots to sectors, cities and so on to an apocalyptic world.

We will be looking at Intensive Care Unit numbers and scale them up against serious cases, assuming every serious/critical case patient, regardless of the outcome, was admitted to an ICU and wasn't unexpectedly deceased.

Some vital information to calculate critical cases, as the critical cases and their outcome and duration data is not available:

*1) The ICNARC report(world report) suggests the median (commonest) duration of an ICU admission in patients with COVID-19 infections who survive is 4 days, but some stay 8 or more days.

2) Medical experts claim that patients who couldn't recover had to stay up to 2 weeks and sometimes even more.

3) Also, on an average, two-thirds of the ICU admitted patients survived.

4) The Union health ministry of India claimed a while ago that on average 4.16% of the total cases were serious in the country at a time. No data had been published regarding the claim but reports

`suggested the figures: 2.25 per cent were admitted to the ICU and 1.91 per cent needed oxygen support.`

Using the figures provided by ICNARC and India:

$$22.5\%\, of\, (Total\, Cases\, on\, that\, day) = 3 * (Total\, Deceased\, on\, that\, day) \quad (2)$$

$$=> People\, who\, were\, put\, in\, ICUs\, on\, that\, day \quad (3)$$

NOTE:
The fraction of those who survive in ICUs won't change very drastically for the next few months but would surely be on a slow and steady rise if the hypothesis is correct.

Let's look at serious cases per hundred confirmed cases and predict for a couple of months ahead using Linear and Non-Linear Regression:

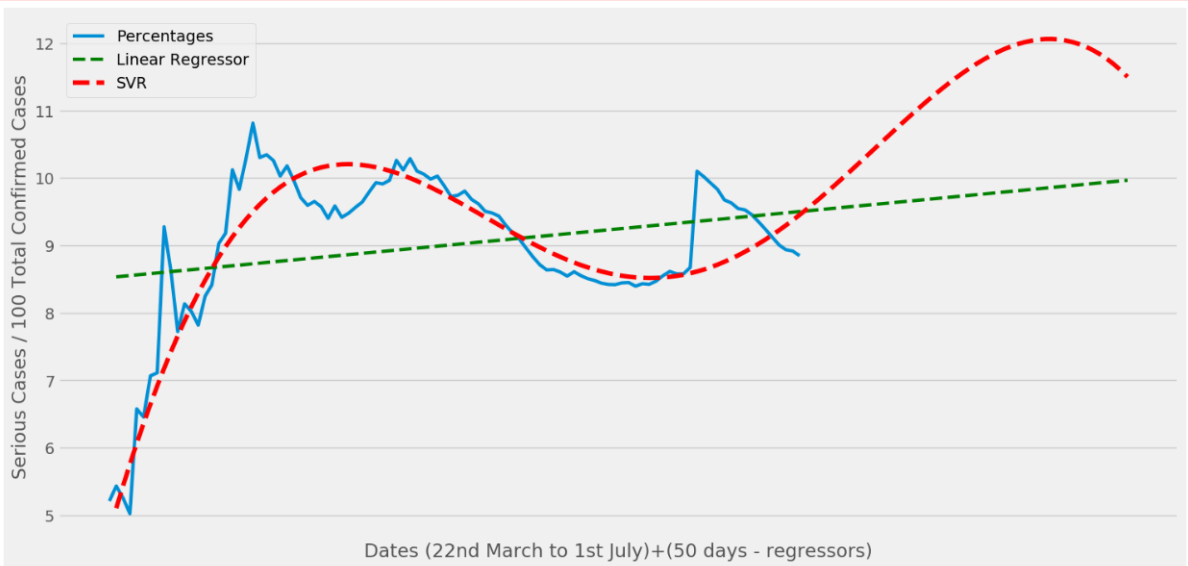


Figure 12: Serious Cases per Hundred Confirmed Cases

If our hypothesis is correct, then the SVR which predicted approx 90-95% recovery rates in 10 months, would also give us a good estimate of total cases

and the peak till the numbers stabilize.

3.2 The Peak: D-Day

If India can keep up the production or just manage to stay ahead of the rising number of serious cases that require serious medical attention (which requires patients to be kept in ICUs or on ventilator like equipment) till we reach the peak, then the nation will be on it's way to recovery or else it might be on it's way to disaster, thus failing our proposed hypothesis. If our hypothesis fails, we intend to calculate the approximate cost and time to implement certain things under the constraints already discovered.

So, our aim is to find the peak of total cases next.

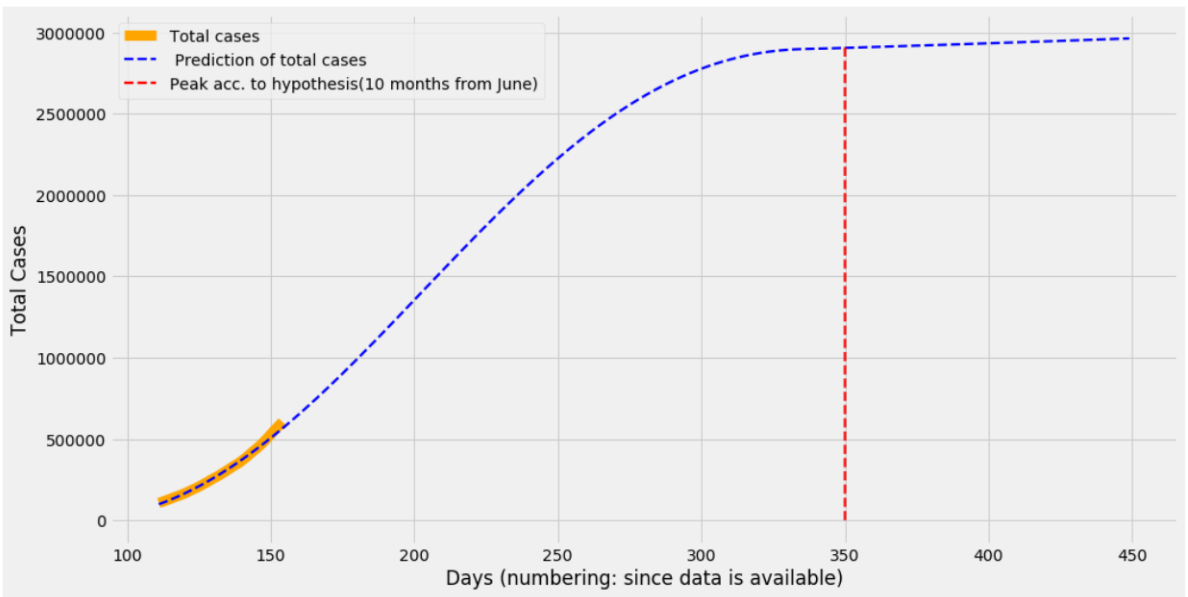


Figure 13: Serious Cases per Hundred Confirmed Cases using the same SVR which we built from our hypothesis and marking the approximate peak.

Findings:

`1) Peak occurs around 200 days, i.e., approximately 6 months, which ends up somewhere in mid-January.`

`2) Peak cases on D-Day:  2912416.059`

```
3) Serious cases on Peak Day:  65529.361
```

3.3 The Big Requirement

We have the number of serious cases on D-Day but this is not the true number of people in the ICUs at the moment as there might be older patients who are still recovering. We need to use ICNARC's findings as stated above(marked with *) to calculate average duration of stay of a critical patient.

Using ICNARC's findings, survivors get discharged in 4 days and the unfortunate ones in 14 days.

$$\frac{2}{3}\, of\ serious\ cases\ are\ survivors => 4 * \frac{2}{3} + 14 * \frac{1}{3} = 7.33\, days \simeq 8\, days \quad (4)$$

Implies total ICUs that would be in use on peak day would be the sum of serious cases from 8 days before till Peak Day.

We are assuming that a person admitted on day - 1 will still be in an ICU bed on day - 8. And will be relieved on the start of day - 9.

Calculation/Finding:
Total Serious Cases in ICUs on peak day: 523889.650

This implies we must have about 524 thousand fully operational ICUs and proportionate medical staff.

4 India's Health Sector

The data is taken from CCDP (A Princeton study which estimated the total number of hospitals and equipment in India)
Let's look at state-wise number of ICUs and total hospitals (both government and private):

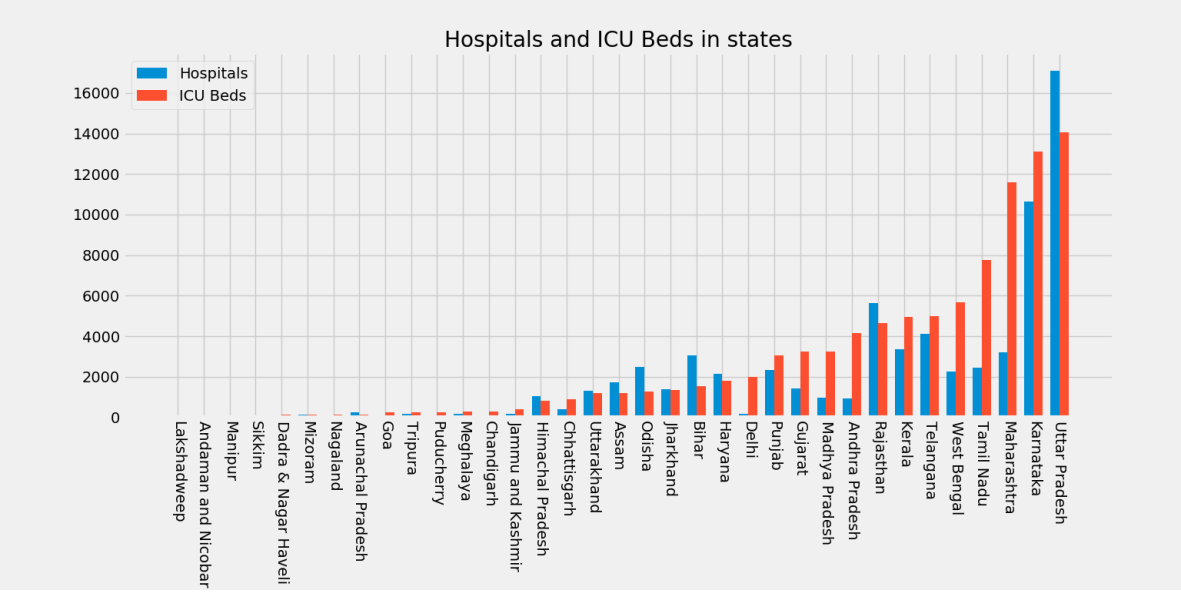


Figure 14: Hospitals and ICU beds

Total Hospitals in India (Private + Government) : 69,228
Total ICU Beds in India (Private + Government) : 94,963

Clearly, the numbers are too low to handle what is coming and our prediction of total recovery fails (the hypothesis fails).
India needs a huge and continuous investment in the medical and health sector to beat the virus.

4.1 Cost of Staying Ahead of the Outbreak

We are going to look at the monetary cost of setting up ICUs in already existing multi-functional hospitals in metropolitan cities because constructing new hospitals and then setting up ICUs in them is not a feasible plan of action.
It seems counter-intuitive to build new hospitals as firstly, the workforce itself might get infected, second of all, hiring doctors and medical staff from nurses

to ambulance drivers is a huge investment in time considering the fact we have only 200 days to spare and thirdly, it is just too expensive to construct a new hospital than converting pre-existing medical rooms to ICUs in big hospitals.
Diving straight in:

Cost of setting up an ICU (10 Beds) (minimal but necessary equipment) in Rupees:

1. Unmotorized Bed = 50,000/piece : 500,000
2. Patient Monitors = 200,000/piece : 2,000,000
3. Ventilators = 800,000/piece : 8,000,000
4. Infusion and Syringe Pumps(at least 18) : 40,000/piece =¿ 720,000
5. Central Monitoring Station : 100,000
6. Oxygen/Compressed Air Pipeline : 200,000 (if already available in hospital, else 600,000)
7. ABG Machine : 600,000
8. BP Apparatus, ECG Machine, etc. : 500,000
9. Air Mattress (Indian made) : 3000/bed =¿ 30,000
10. Defibrillator : 300,000
11. Medical Grade Flooring would cost about Rs. 100/sq. ft. minimum and would cost Rs. 300,000
12. Air Conditioning with sufficient fresh air exchange would cost about Rs. 750,000 for 25 tonnes of AC

NOTE: This is considering zero civil work and considering that the hospital is otherwise well equipped and has a proper Laboratory, Radiology, Pharmacy and OT.

Concluding Remarks:

1. Cost of setting up an ICU (in a fully operational hospital) which has 10 beds: Rs. 1,40,00,000 (1.4 crores or 14 million)

2. Required Number of Beds = 523889 - 94963 = 428,926

3. Cost of setting these ICU beds in operational hospitals is 60,049.64 crores (600,496.4 million Rs. which happens to be 0.35% of India's GDP)

4. We need to add 2,144 ICU beds everyday or simply put, create 214 ICU wards of capacity 10

5 Statistical Analysis at a Smaller Scale

5.1 Daily Stats

Before we dive into smaller regions in the country, let's look at the daily new cases, deaths and recovered cases.

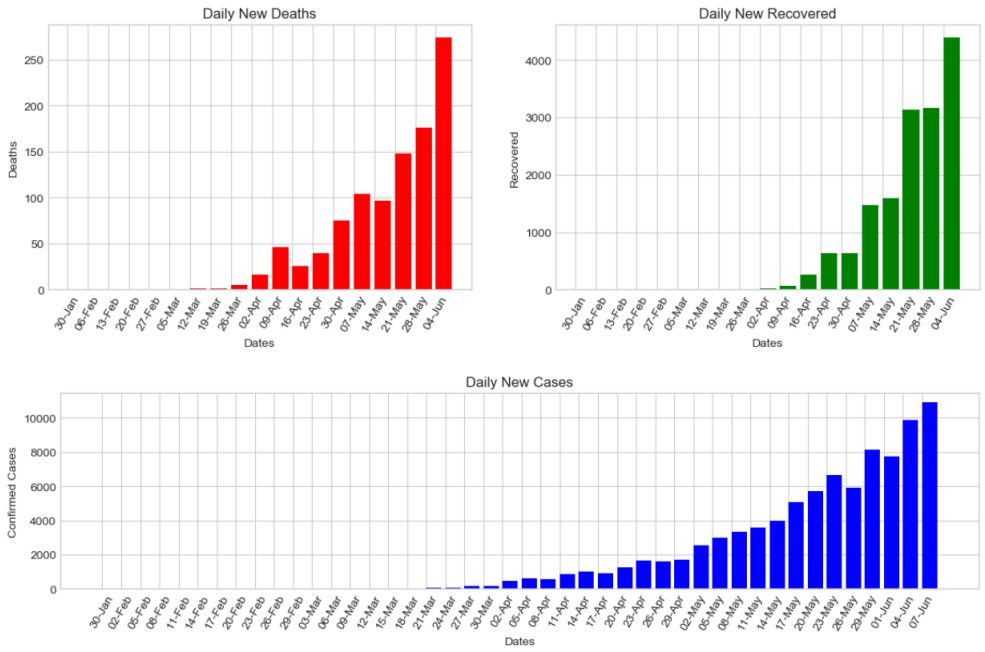


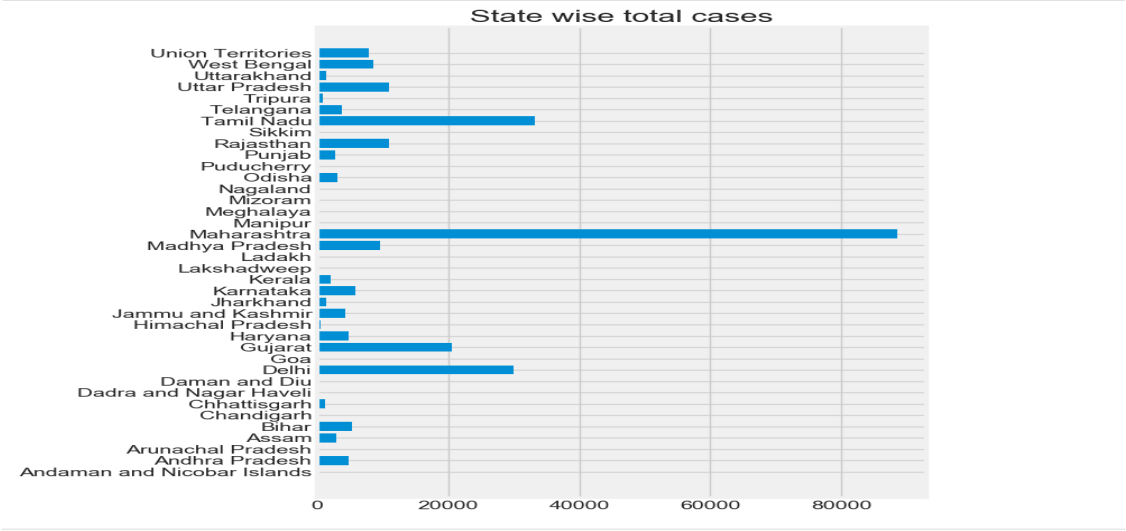
Figure 15: Daily Stats

Some noteworthy averages from this data:

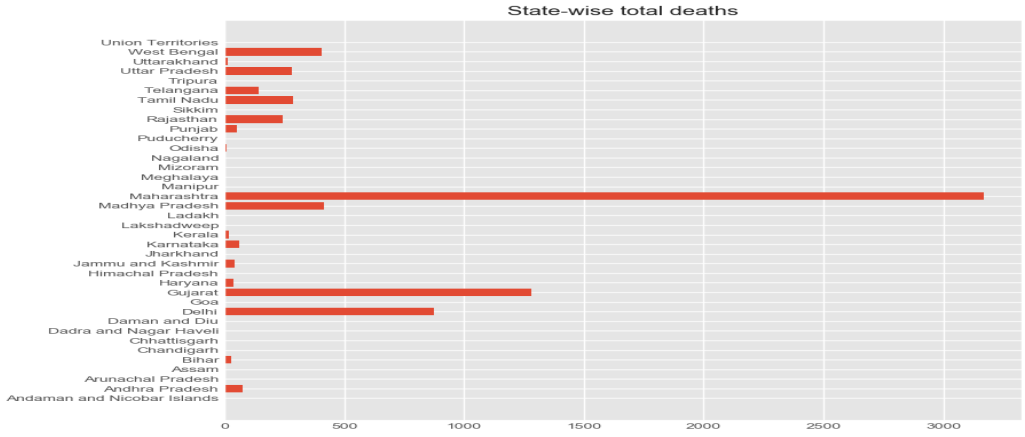
Number of new recovered cases per one new death	17.318175738932727
Number of new recovered cases per new case	0.4867585641735051
Number of new deaths per new case	0.02810680359821217

Is this on average the same? Let's look at the same parameters(Cases,Deaths and Recoveries) but at the state-level.

5.2 State-wise Stat Plots



(a) State-wise total cases



(b) State-wise Total deceased

Figure 16: All stats: state-wise

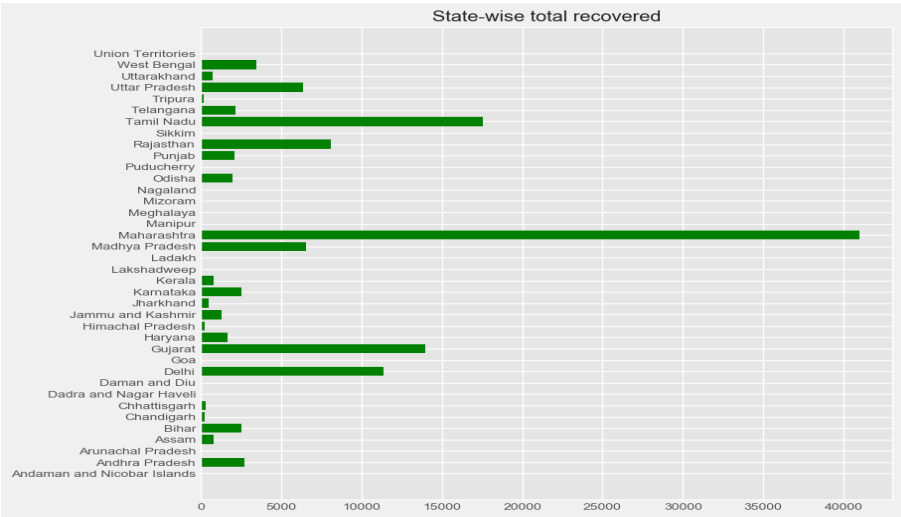


Figure 17: State-wise Total recovered

5.3 Demographics

States	Total Pop.(2019)	Total Cases	*Prediction	Pct 0-14yrs	Pct>60yrs
Uttar Pradesh	237,882,725	10947	1,04,973.158	30.2%	7.0%
Bihar	124,799,926	5247	61,585.727	35.0%	6.6%
Maharashtra	123,144,223	88529	50,251.023	25.1%	9.3%
West Bengal	99,609,303	8613	38,638.483	23.7%	9.0%
Madhya Pradesh	85,358,965	9638	37,768.556	30.1%	7.2%
Rajasthan	81,032,689	10876	36,334.941	30.1%	7.5%
Tamil Nadu	77,841,267	33229	29,640.624	21.6%	10.5%
Karnataka	67,562,686	5760	26,127.443	24.3%	8.3%
Gujarat	63,872,399	20574	26,291.484	26.1%	8.6%
Delhi	18,710,922	29943	7,080.407	25.0%	6.9%

Table 1: Demographics of India. *(The prediction column is based on the assumption mentioned below)

—> *Assuming one in eight hundred forty three gets infected:*

`Q. Why this assumption?`

`A. Population of India(2019) in the above mentioned states' populated`

cities with population density>10,000 per sq. Km / Total cases in these states

Cities considered and their population and population density:

State	Population	Population Density (per sq. Km)
Delhi	67,88,462	23,808.6
Mumbai	1,24,42,373	20,317
Chennai	4,64,673	24,963
Hyderabad	39,43,323	17,649
Kolkata	44,96,694	24,718

NOTE:
The above cities' data was given by the census of India in 2011 when India's population was 1.21 billion. Now the population is 1.36 billion. The compound annual growth rate(CAGR) from 2011 to 2019 is 1.11% which can be calculated from equation (2). Using the same equation let's calculate how many people are there per infected person.

Let population per infected in 2019 be $= y_1$
Let population per infecter in 2011 be $= y_2$
Let CAGR compund annual growth rate $= 1.11\% = R$
Time period = 2019 - 2011 = 8 years

$$y_1 = y_2\left(1 + \frac{R}{100}\right)^8 \tag{5}$$

$$\frac{Pct[(0-14yrs) + (>60yrs)] * pop.\,of\,that\,state}{843} = tot.\,cases\,in\,state \tag{6}$$

Observation: Tamil Nadu and Gujarat's predictions are pretty close to the actual data. While on the other hand states like Uttar Pradesh and Bihar have a huge loss value. It is clear that we have taken too few parameters and generalized the value(843) for all the states which can by no means be treated like this. We have under-fitted our data in some sense.

Implication:
Nonetheless, it doesn't disprove that the elderly and young are a higher risk. So, let's look at a new feature - Age-wise death shares

due to the virus.

5.4 Age-wise Death Share Percentage

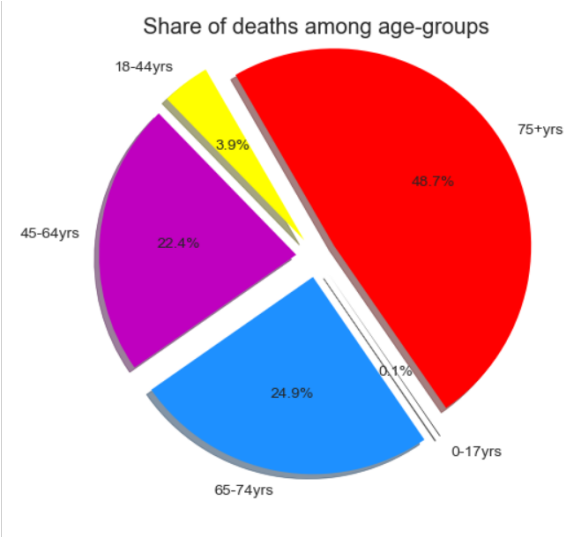


Figure 18: Death Share among Young and Elderly

Observation : Elderly are at the highest risk. And due to the joint family structures in India (contact structure in India),it might be difficult to contain the spread.

So, we've seen so far how many and the age-groups which are affected by this epidemic. But have we faced such a situation before or not?

6 Brief Case Study: COVID-19 vs. H1N1

We will look at: (World wide data)

1. Total cases for each virus
2. Total deaths for each virus
3. Mortality rate percentages
4. RO - Reproduction Number (No. of people infected from one infected individual)

And compare which one is the deadlier epidemic by measuring mortality rates and giving logical estimates of growth in medical sciences and technology.

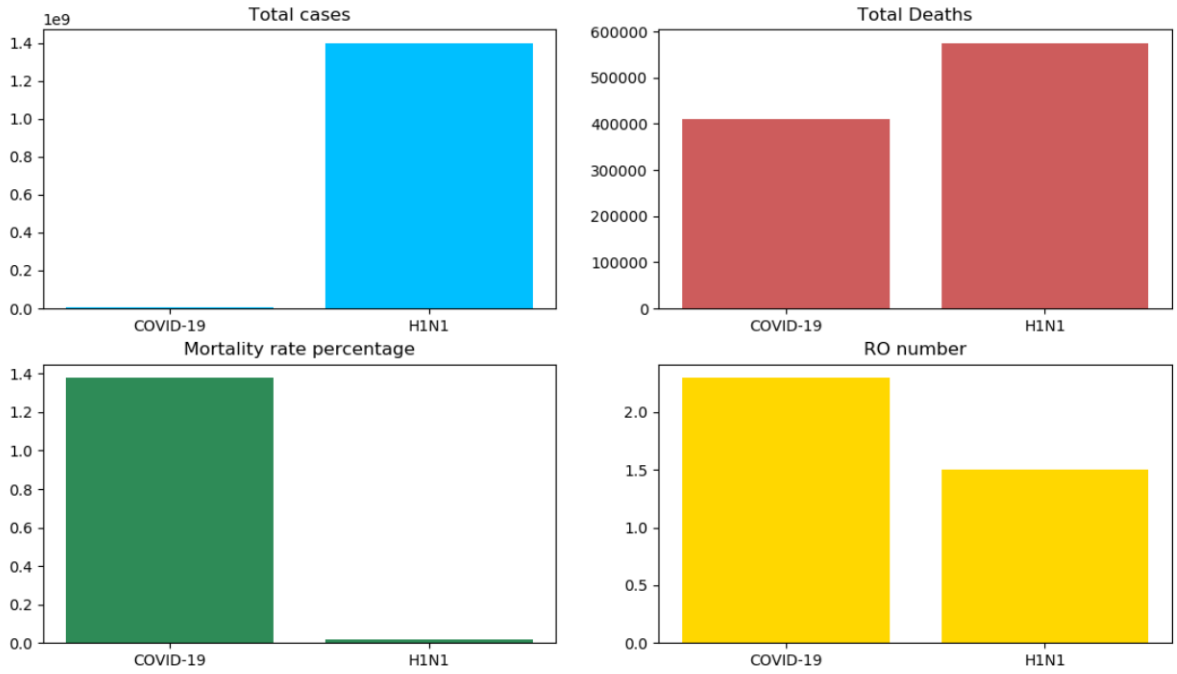


Figure 19: Case Study Parameters' Plots

Conclusion:
There is no doubt that COVID-19 is the deadlier virus among the two because:

1. The world was pretty quick to know that an infection was spreading and started taking preventive measures but this wasn't the case with H1N1 pandemic.

2. In every 10 years, the human race sees major developments in medical sciences and technology in general. But we are still not able to save many during this pandemic.

3. COVID-19 has possibly evolved into a deadlier virus over a short span of time. We are seeing asymptomatic cases very frequently. This was not the case with H1N1.

4. The plots are all in favour of COVID-19 being the deadlier one.

7 Summary

We proposed a hypothesis that India has a rising recovery rate which implies that the nation could revive soon. We proved that it was mathematically correct based on the past data and then after comparing India with Japan's example of successful recovery, we used the hypothesis and two simple predictive models to predict that if external factors such as availability of medical facilities, political actions, foreign relations and internal matters of the nation didn't affect the data to come, it'll lead to total recovery and revival of India within 10 months.
Then we added the factor of medical facilities in India to get more realistic measures of our hypothesis.
We went on to discover the state-wise stats so that we could understand the demographics of the affected and found that elderly, beyond the age of 60-65 were at the highest risk, followed by children. We didn't consider the fact that patients could have had a medical history which resulted in a different outcome of their case. This point reduces the accuracy of predictions of this study.
Seeing the devastating effects of the pandemic, we decided to do a brief case study of H1N1 vs. COVID-19 to find whether humans had faced a pandemic like this or not. And the results of the case-study were completely biased in favour of COVID-19 being the deadlier of the two.
We urge you to stay safe and follow WHO's guidelines on the same.