Data Analysis - COVID-19

Aryan Garg

June 13, 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.

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 one of the most deadly viruses ever known, the H1N1 virus. And then conclude with our observations, conclusions and hypothesis' results (conclusive or inconclusive).

2 Better Recovery Rates?

2.1 Overall Cases

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

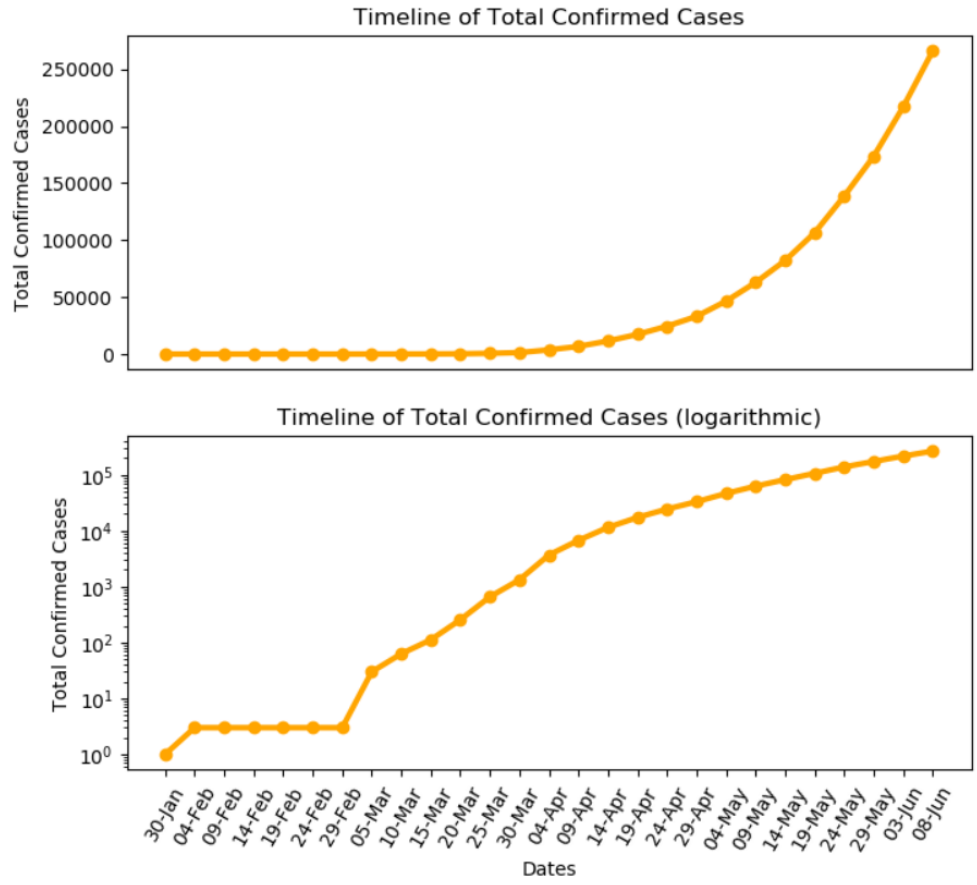


Figure 1: Total Confirmed Cases from 30th January to 8th June

`Observation: Total cases are rising exponentially.`

2.2 Total Deceased Count

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

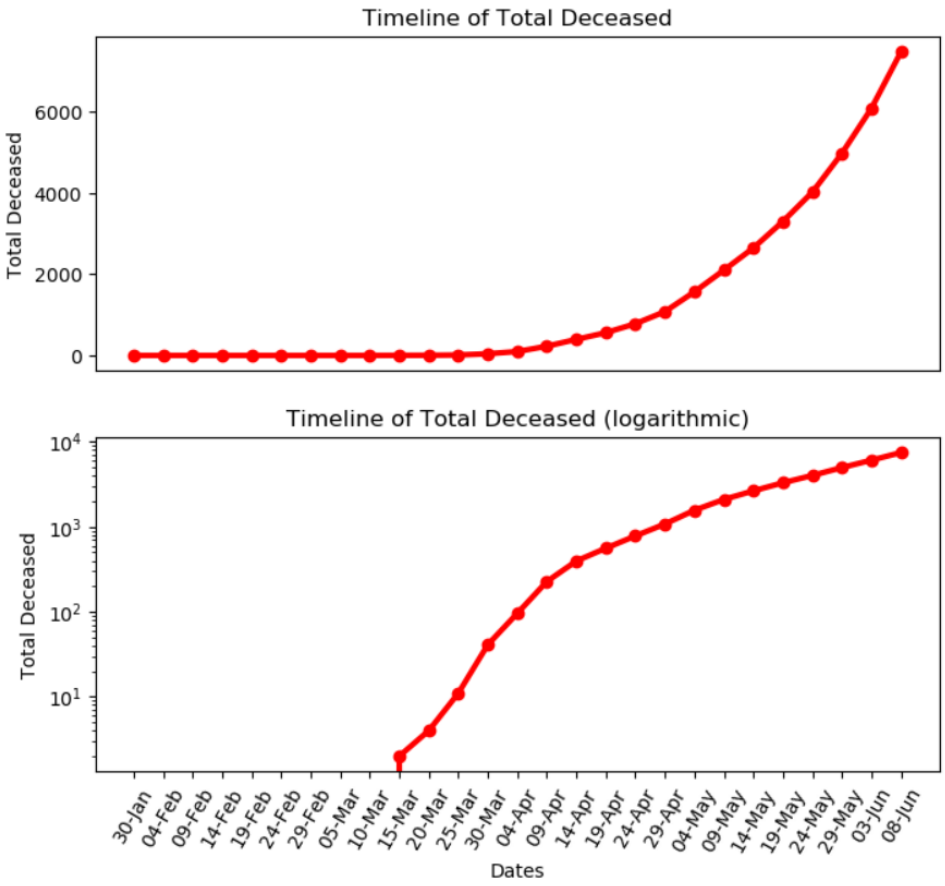


Figure 2: Total Deceased from 30th January to 8th June

`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.

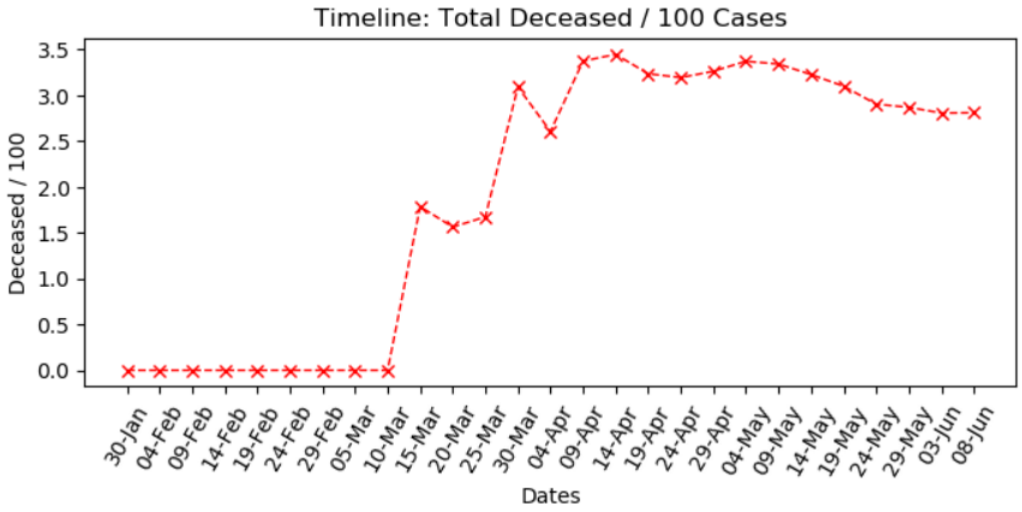


Figure 3: Timeline of Deaths per 100 cases:

`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.`

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!

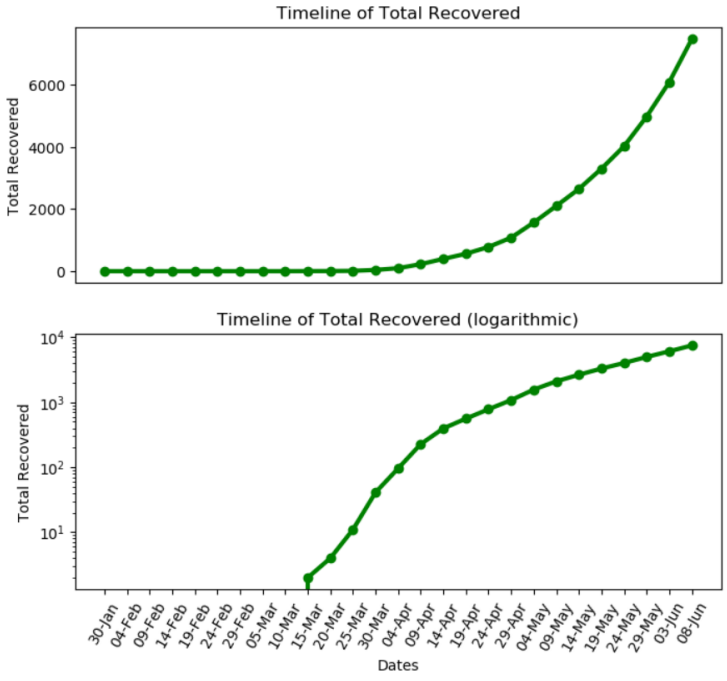


Figure 4: Timeline of Total Recovered

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

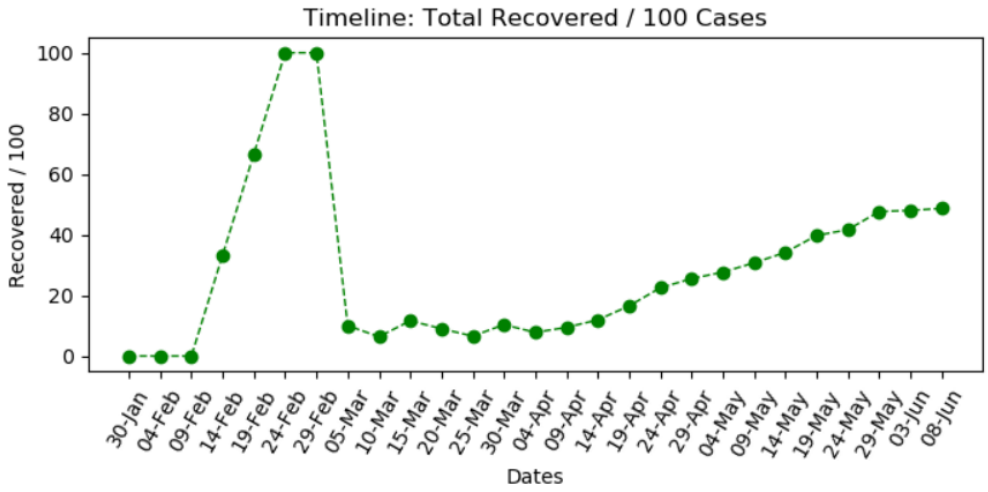


Figure 5: Timeline of Total Recovered per 100 cases

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

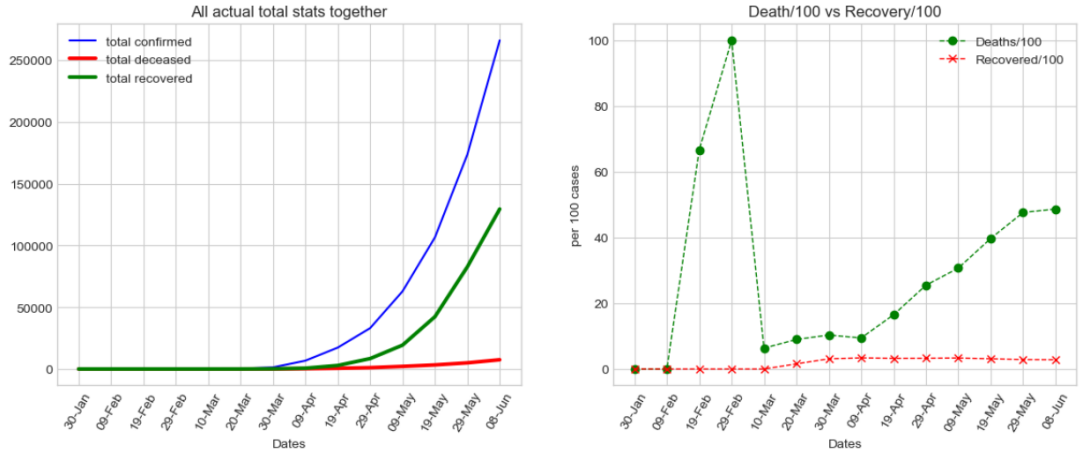


Figure 6: Total cases, deceased and recovered

`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:

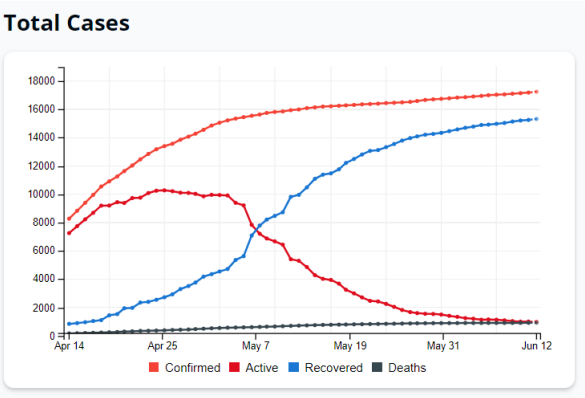


Figure 7: All stats of Japan together

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.

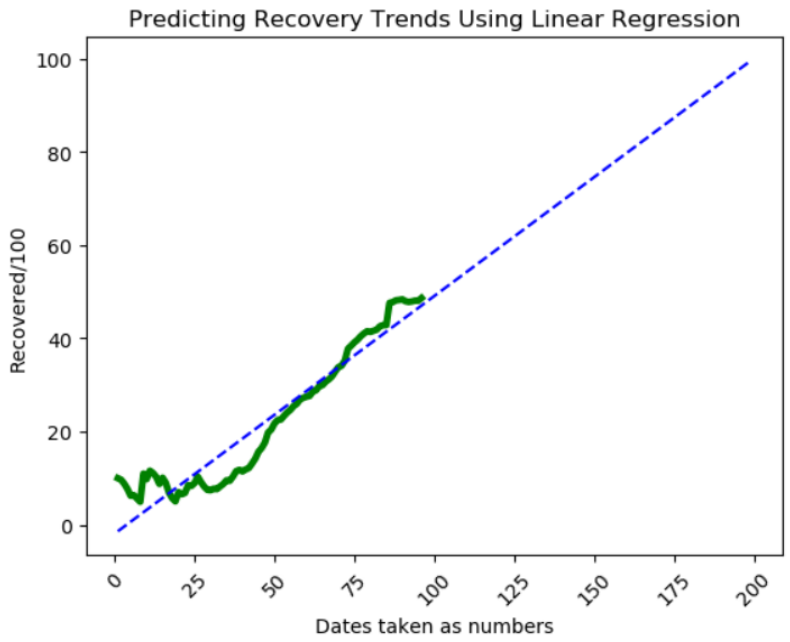


Figure 8: Predicting Number of Days Till Total recovery using Linear Regression :

`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

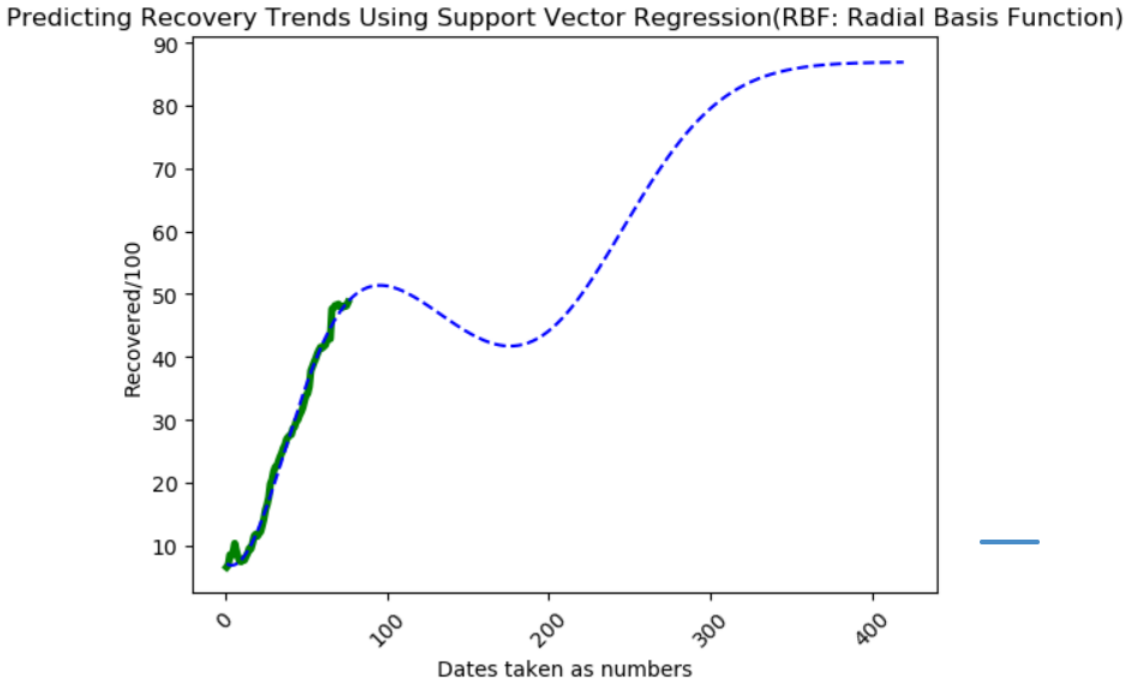


Figure 9: Predicting Number of Days Till Total Recovery using Support Vector Regression(params: gamma=40000, C=0.0001)

`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!*

Q. Can overall stats give us finer details as to who are at risk or not? Who is it that will recover as per the finding above?
—>So let's analyze the stats at a smaller level,i.e, at the state-level.

3 Statistical Analysis at a Smaller Scale

3.1 Daily Stats

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

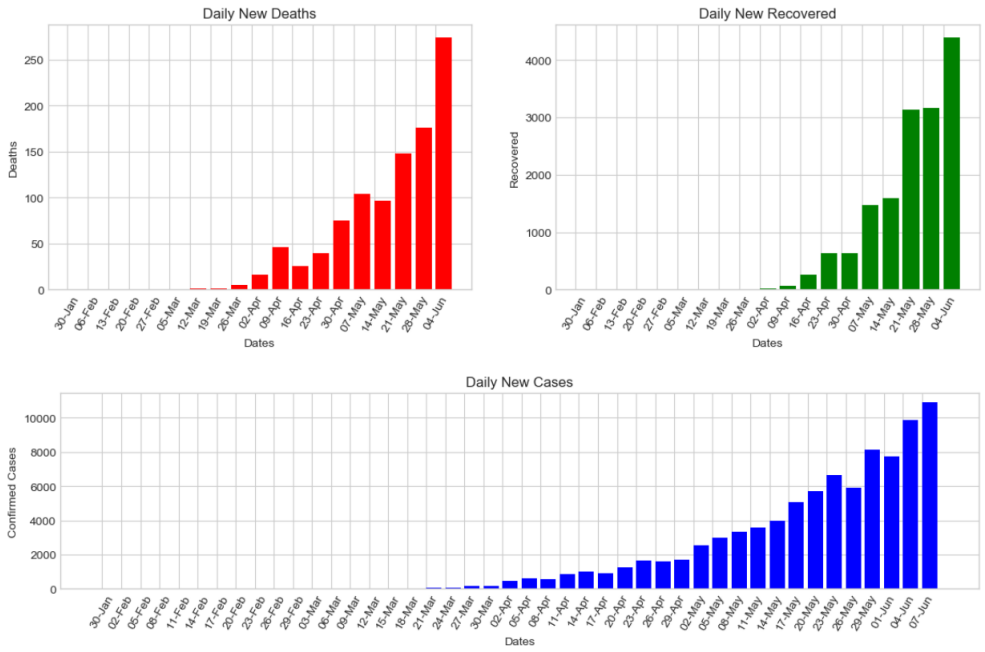


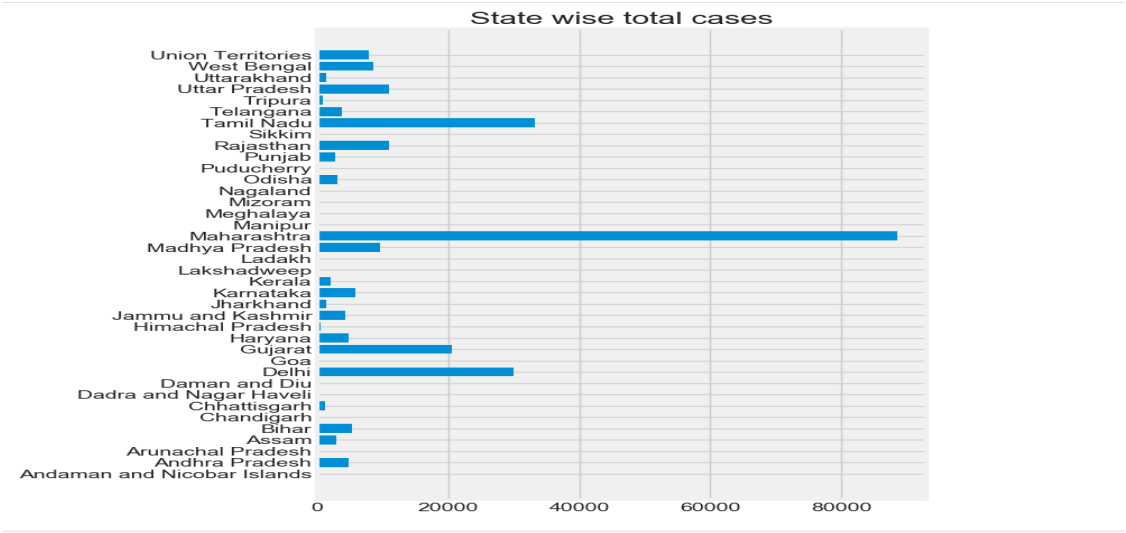
Figure 10: Daily Stats

Some noteworthy averages from this data:

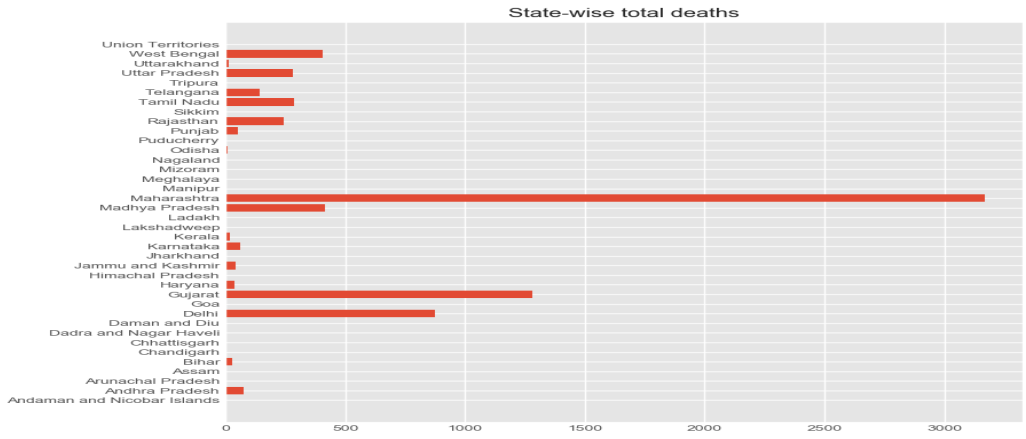
Number of new recovered cases per one new death	17.318175738932727
Number of new recovered cases per new case	0.4867585641735051
Number of new deaths per new case	0.02810680359821217

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

3.2 State-wise Stat Plots



(a) State-wise total cases



(b) State-wise Total deceased

Figure 11: All stats: state-wise

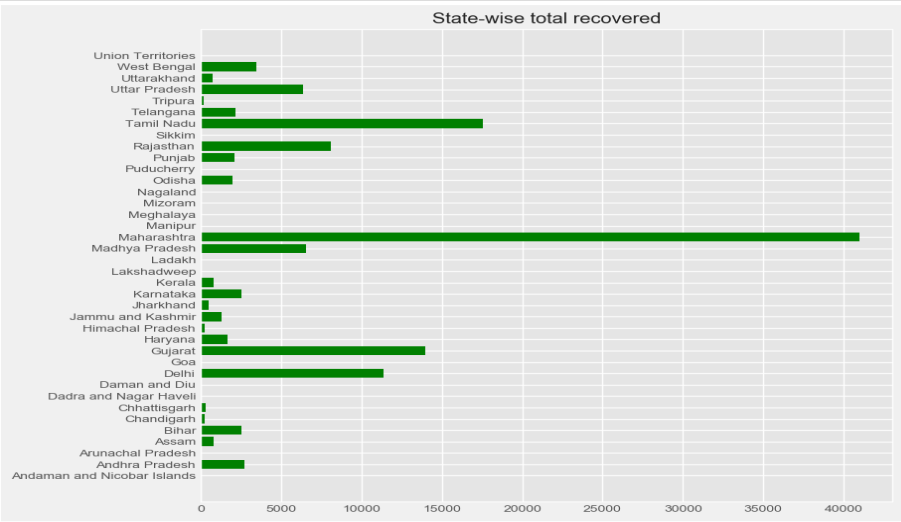


Figure 12: State-wise Total recovered

3.3 Demographics

States	Total Pop.(2019)	Total Cases	Pct 0-14yrs	Pct 15-59	Pct>60yrs
Uttar Pradesh	237,882,725	10947	30.2%	62.8%	7.0%
Bihar	124,799,926	5247	35.0%	58.4%	6.6%
Maharashtra	123,144,223	88529	25.1%	65.6%	9.3%
West Bengal	99,609,303	8613	23.7%	67.4%	9.0%
Madhya Pradesh	85,358,965	9638	30.1%	62.7%	7.2%
Rajasthan	81,032,689	10876	30.1%	62.4%	7.5%
Tamil Nadu	77,841,267	33229	21.6%	67.9%	10.5%
Karnataka	67,562,686	5760	24.3%	67.4%	8.3%
Gujarat	63,872,399	20574	26.1%	65.3%	8.6%
Delhi	18,710,922	29943	25.0%	68.1%	6.9%

—> *Assuming one in five hundred gets infected:*

[

```
Q. Why this assumption?

A. Population of India in the above mentioned states(populated cities
with population density>7000 per sq.  Km) / Total cases in these
states
```

]

It is quite evident that:

$$Pct[(0-14yrs)+(>60yrs)]*pop.\,of\,that\,state/500 = tot.\,cases\,in\,state \quad (2)$$

Implication: Children and elderly are at a higher risk.

```
'''Proof (Python Script)'''
pop_maha = 123144223
pct_children_maha = 25.1
pct_elders_maha = 9.3

LHS = ((pct_children_maha+pct_elders_maha)/100)*pop_maha/500
RHS = total_cases_st_wise[-17] # Maharashtra's index in dataset

print("Our assumption and calculation ->",LHS)
print('Real data available to us      ->',RHS)
```

Output:
Our assumption and calculation -> 84723.22542400002
Real data available to us ->88529

This somewhat proves our implied statement. The data is bound to vary to a greater extent. We just got lucky!

3.4 Age-wise Death Share Percentage

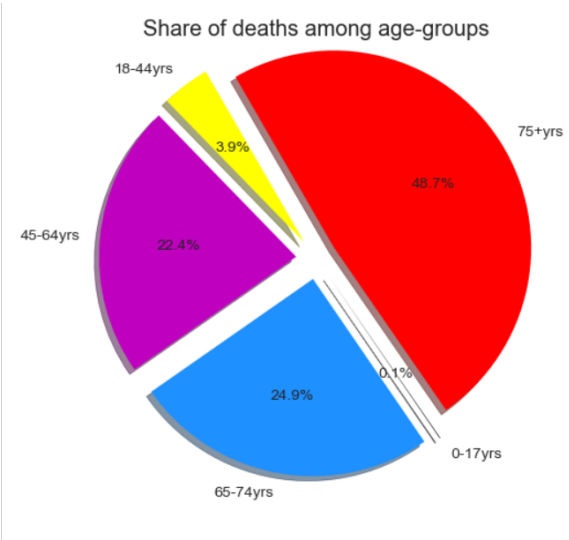


Figure 13: Death Share among Young and Elderly

`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?

4 Brief Case Study: COVID-19 vs. H1N1

We will look at:

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.

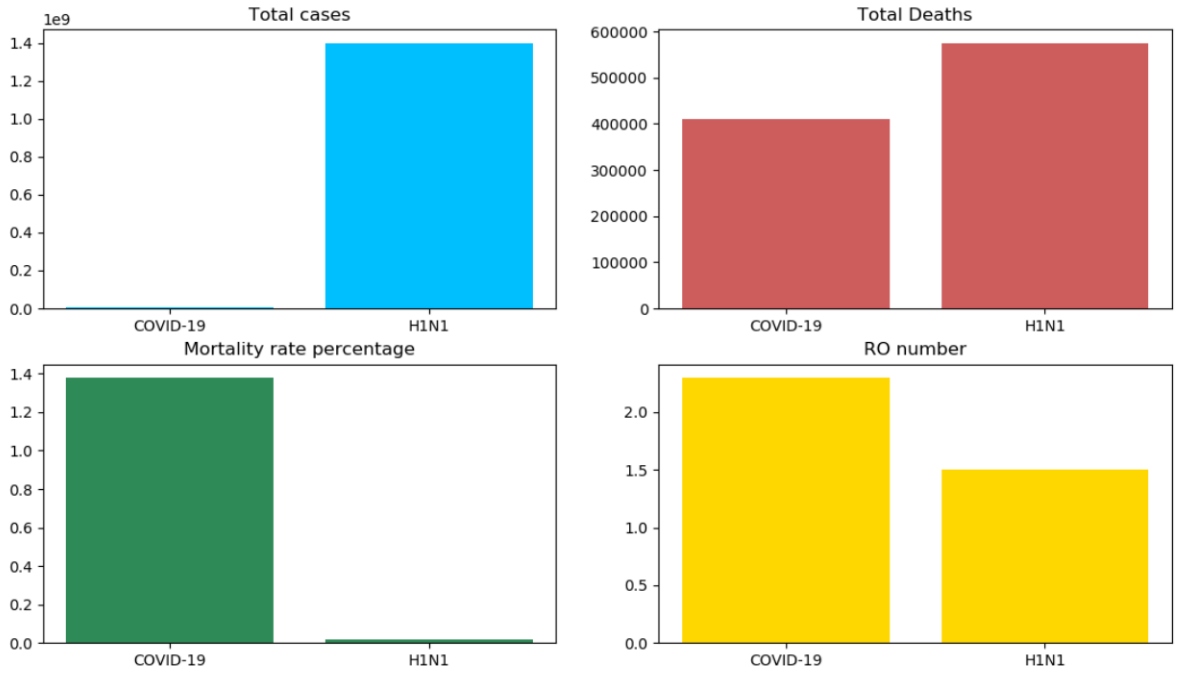


Figure 14: Case Study Parameters' Plots

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.

5 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.

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.