COVID-19 Data Analysis

Anand Narasimhan
Dept of Computer Science and Engineering
IIT Bombay
210051001@iitb.ac.in

Abstract—The goal of this study is to compare the variations in COVID-19 cases, deaths and vaccine doses in a select group of states, with the objective of understanding which state was better in controlling the spread of the virus.

I. INTRODUCTION

According to the World Health Organization (WHO, 2020), the coronavirus (COVID-19) outbreak which emerged from central China in late December 2019 has spread to over 218 countries, areas or territories, and has resulted in over 367 million confirmed cases as well as over 5.5 million deaths across the globe as of January 2022. Given the widespread and ongoing transmission of the novel coronavirus worldwide, the WHO officially declared it a pandemic on March 11, 2020.
The rapid spread of the unprecedented COVID-19 pandemic has put the world in jeopardy and changed the global outlook unexpectedly. Many countries and states have adopted various ways to deal with the pandemic. Some regions were able to effectively control or restrict the spread of cases, while others, because of a large population, poor infrastructure, or lack of structured policies in general, performed poorly in these testing times.

Analytics techniques played a huge role in predicting the future course of the pandemic and helped governments worldwide to channelize their resources and impose restrictions, including lockdowns. Data Science algorithms also helped in predicting the rate of mutation of the coronavirus strains and helped assess the severity of disease caused by different variants.

In this analysis, I attempt to analyse the COVID-19 situation in some of India's states, by analysing their statistics like Cases, Deaths and Vaccine Doses in order to justify and debunk a few common thoughts and misconceptions regarding the data.

II. DATASETS

In this report, I have obtained and used datasets on COVID-19 cases in four states of India : Maharashtra, Kerala, Tamil Nadu and Uttar Pradesh. I have performed an Exploratory Analysis on COVID cases and the vaccination campaign in each of those states.

Procedure :

For the EDA part, I obtained datasets detailing the total case count in the four states (from March 2020 to August 2021), as well as the vaccination records (from January 2021 to August 2021). SInce both the datasets only had the total cases/doses up to that particular date, I calculated the daily cases/doses as well. I then cleaned the data by removing NaN values and correcting some inconsistencies, and plotted different kinds of graphs to help better understand the data. We then drew conclusions based on the graphs/ plots obtained, trying to explain the nature of and variations in the plots.

A representation of few tables used :

	Date	Time	Region	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Total_Cases	Daily_Cases	Daily_Deaths	Population
0	2020-03-07	6:00 PM	Tamil Nadu	1	0	0	0	1	1.0	0.0	83697770
1	2020-03-08	6:00 PM	Tamil Nadu	1	0	0	0	1	0.0	0.0	83697770
2	2020-03-09	6:00 PM	Tamil Nadu	1	0	0	0	1	0.0	0.0	83697770
3	2020-03-10	6:00 PM	Tamil Nadu	1	0	0	0	1	0.0	0.0	83697770
4	2020-03-11	6:00 PM	Tamil Nadu	1	0	0	0	1	0.0	0.0	83697770
5	2020-03-12	6:00 PM	Tamil Nadu	1	0	0	0	1	0.0	0.0	83697770
516	2021-08-05	8:00 AM	Tamil Nadu	-	-	2513087	34197	2567401	1949.0	38.0	83697770
517	2021-08-06	8:00 AM	Tamil Nadu	-	-	2515030	34230	2569398	1997.0	33.0	83697770
518	2021-08-07	8:00 AM	Tamil Nadu	-	-	2516938	34260	2571383	1985.0	30.0	83697770
519	2021-08-08	8:00 AM	Tamil Nadu	-	-	2518777	34289	2573352	1969.0	29.0	83697770
520	2021-08-09	8:00 AM	Tamil Nadu	-	-	2520584	34317	2575308	1956.0	28.0	83697770
521	2021-08-10	8:00 AM	Tamil Nadu	-	-	2522470	34340	2577237	1929.0	23.0	83697770
522	2021-08-11	8:00 AM	Tamil Nadu	-	-	2524400	34367	2579130	1893.0	27.0	83697770

	Date	Region	Total Doses Administered	Sessions	Sites	First Dose Administered	Second Dose Administered	Male (Doses Administered)	Female (Doses Administered)	Transgender (Doses Administered)	...
0	2021-01-16	Maharashtra	5726.0	179.0	174.0	5726.0	0.0	3668.0	2057.0	1.0	...
1	2021-01-17	Maharashtra	6521.0	269.0	216.0	6521.0	0.0	3953.0	2566.0	2.0	...
2	2021-01-18	Maharashtra	6521.0	772.0	320.0	6151.0	0.0	3569.0	2581.0	1.0	...
3	2021-01-19	Maharashtra	13699.0	1196.0	340.0	13699.0	0.0	6328.0	7367.0	4.0	...
4	2021-01-20	Maharashtra	23880.0	1547.0	347.0	23880.0	0.0	9658.0	14205.0	17.0	...
5	2021-01-21	Maharashtra	24148.0	1820.0	357.0	24148.0	0.0	9784.0	14347.0	17.0	...

IV. ANALYSIS PIPELINE

Parameters

A. Comparison Between Absolute and Relative

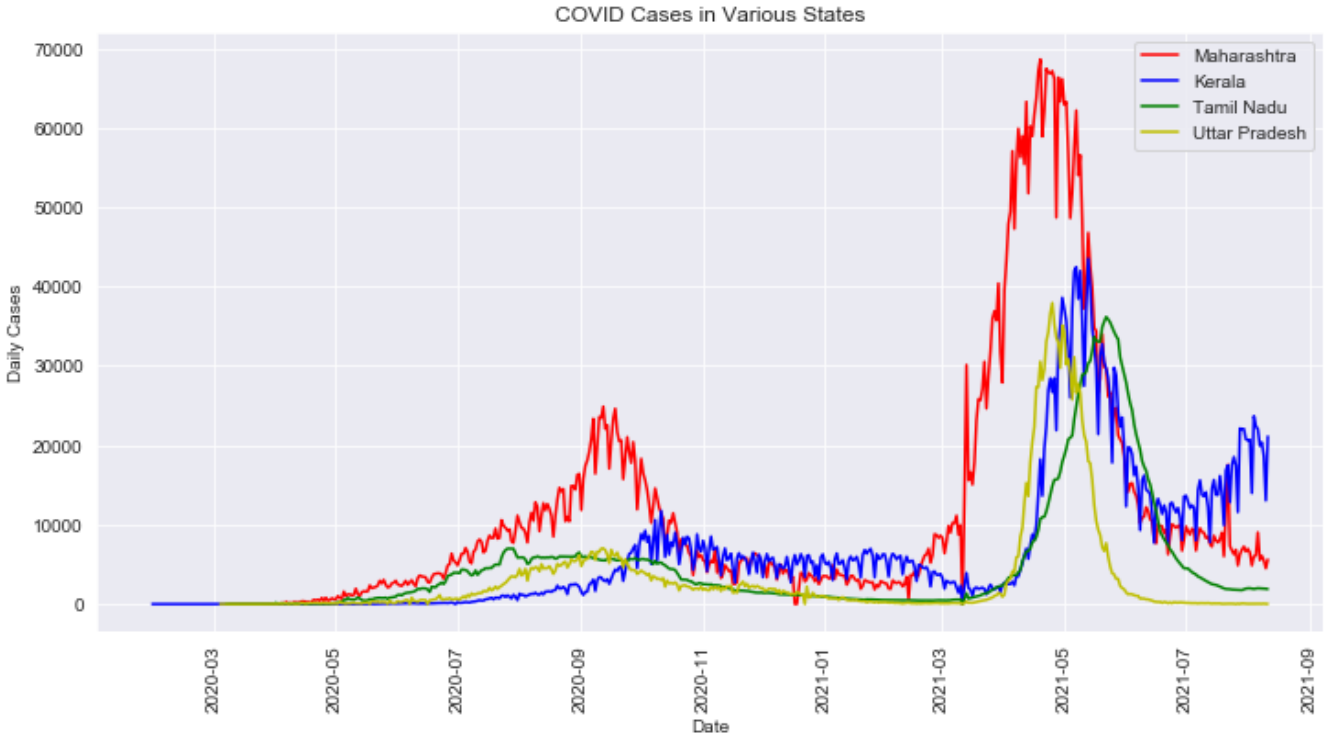

Fig. 1. Absolute number of daily cases in different states

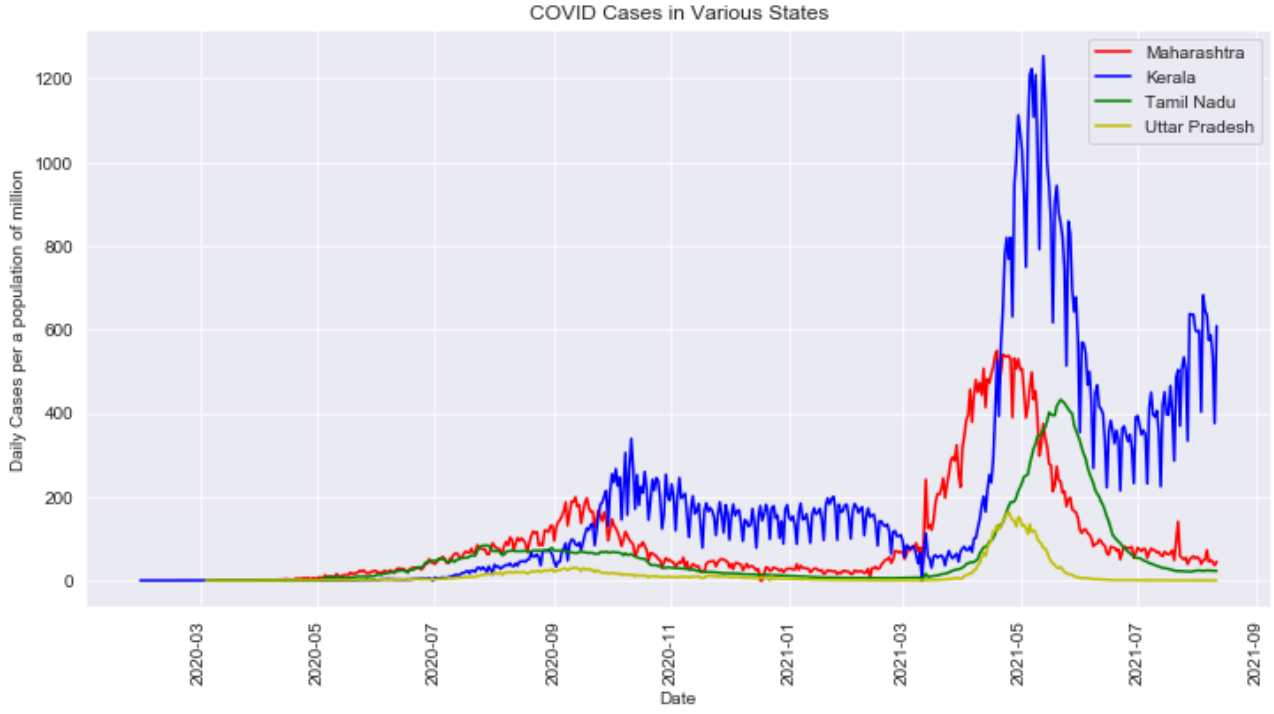

Fig. 2. Relative daily cases in different states(per 1m pop)

For comparing the spread of COVID-19 virus in different countries, there are two trains of analysis to follow. We observe that the absolute number of cases in different countries depicts the magnitude of spread of the virus, taking into consideration the population. This gives us an idea about the stress on the healthcare system of the country, the estimate about health expenditure of the country as well as the number of patients in the country. On the other hand, the relative number of cases gives us the ability to compare the spread of COVID-19 in different countries, irrespective of the population. Dividing the cases by population, and scaling it to per 1m population gives us the ability to compare and contrast the efficiency of any country in controlling the spread of COVID-19. The relative scaling of parameters gives us an understanding of the quality and relative quantity of healthcare facilities in different countries.

First of all, we see that there are two peaks in the graph, the first corresponding to around September/October 2020 and the second at May 2021, the peak of the Delta variant in India. Less distinct is the fact that Kerala was the first state in India to get affected by COVID, albeit with a low number of cases. The second wave originated in Maharashtra and then spread to the rest of India.

Also, the peaks for the Delta variant were far higher than that of the first wave, clearly signifying that it had a higher potency and a higher mortality rate.

A common theme in the graph (and newspapers in real life) is the fact that Maharashtra's cases hit the highest peak, at least in the first figure (absolute cases). However, the second figure paints a different story, showing that it was in fact Kerala, with it's small population, where COVID spread the most relatively.

Looking at Tamil Nadu and Uttar Pradesh, we see that both countries had very similar absolute cases, however, we have to take into account the population difference between the two states, which results in Uttar Pradesh's spread being lesser.

B. Relative Cumulative Cases in Different Countries

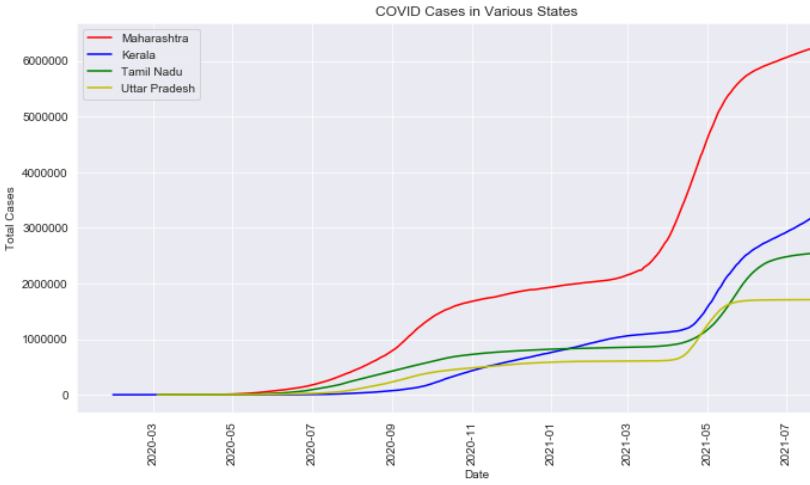

Fig. 3. Absolute cumulative cases in different states

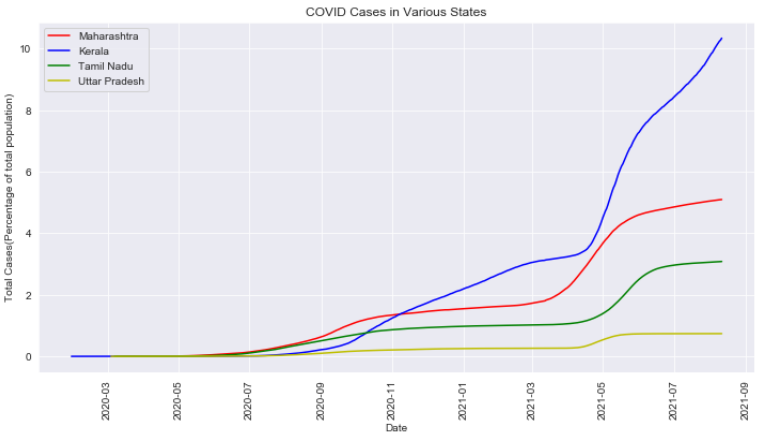

Fig. 4 Relative cumulative cases in different states (percentage)

The line plots for relative cumulative cases(percentage population) depict that the relative spread of COVID-19 in Kerala has been the greatest out of all the countries so far. By the end of the time duration in consideration, more than 10% of the population in Kerala had been infected by the coronavirus at least once.

Again, the same trends are shown : Maharashtra leading the absolute case metric but dropping below Kerala relatively, Tamil Nadu and Uttar Pradesh having similar absolute cases but UP having far lesser relative cases. The slopes of the graph sharply rise during the second wave, confirming that the second wave was far more contagious than the first.

C. Daily Deaths

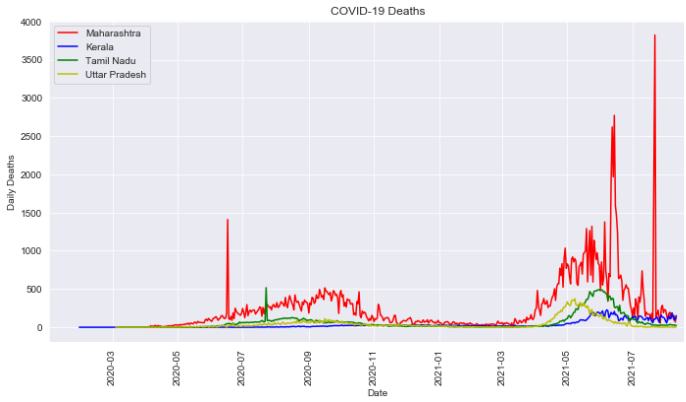

Fig. 5 Absolute Daily Deaths in different states

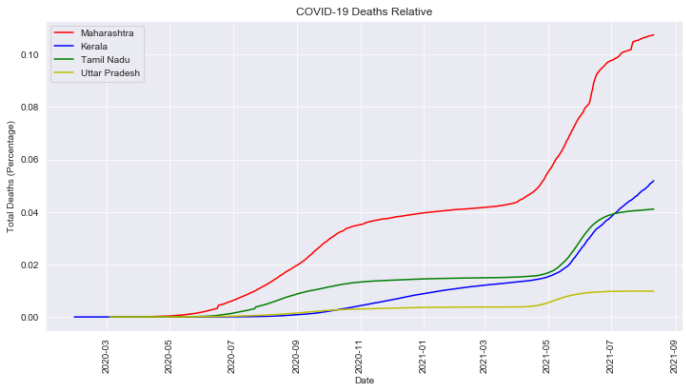

Fig. 6 Relative Total Deaths in different states (Percentage)

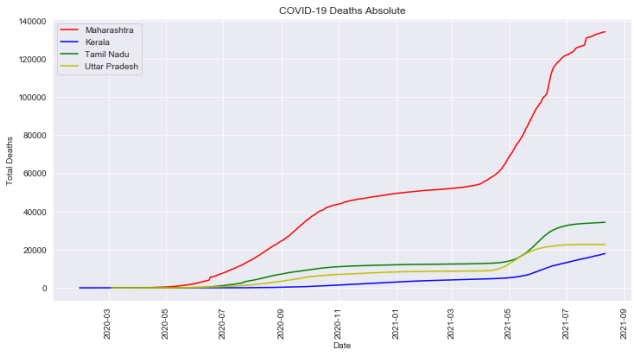

Fig. 7 Absolute Total Deaths in different states

These graphs about the Deaths due to COVID 19 tell us that Maharashtra has by far the most amount of deaths, whether cumulative or daily, and it has the worst spikes as well, with close to 4000 deaths on a particular day. Kerala's small population again masks it's spread, it's absolute cases being lower than it's relative percentage of people.

Behind Maharashtra, Tamil Nadu has the second most absolute deaths, whereas Uttar Pradesh has a surprisingly low number of deaths, similar to it's cases.

One thing to note here is that the peaks in the deaths occurred about 1.5 to 2 months after the peaks for the cases. This can be partially explained by the fact that medical centres and hospitals were overworked due to the peak and thus didn't have adequate resources to treat everyone. Another possible explanation could also be that the reporting of deaths was not done properly and was compiled later.

D. Distribution of Daily Cases

The violin plot describes the distribution of the daily cases in different countries during the time duration in consideration. For Uttar Pradesh, the majority of the days had seen a very low number of new cases. But Maharashtra having a high and lean violin plot suggests that Maharashtra had constantly seen new cases almost everyday, and the maximum number of daily new cases was also much higher than everywhere else. Tamil Nadu has a distribution between the ones of Kerala and Uttar Pradesh, with the lowest peak.

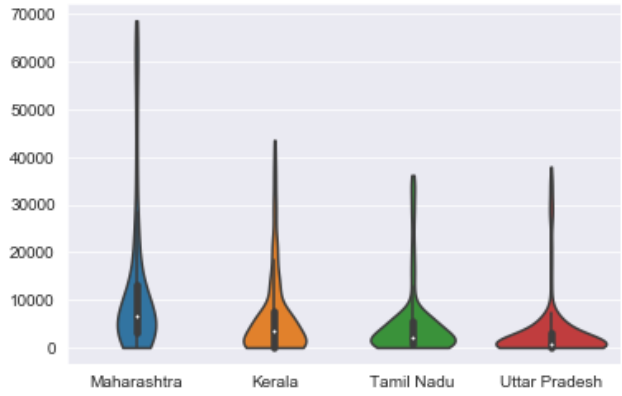

Fig. 8. Violin plot for distribution of daily cases

E. Effect of the vaccines

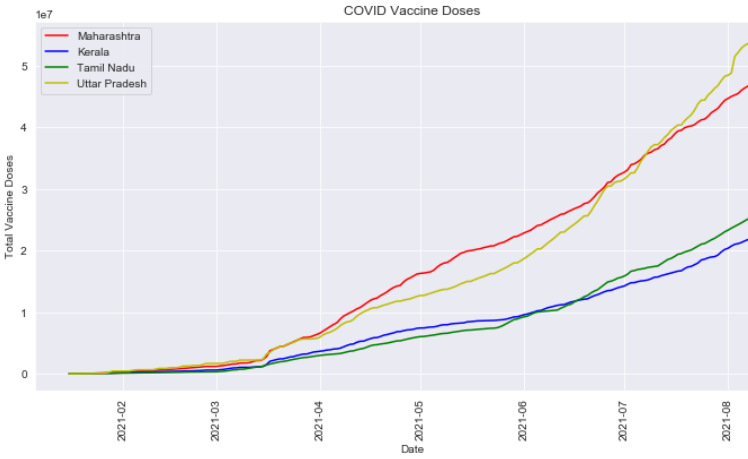

Fig. 9. Absolute Total Vaccine doses in different states (/10^7)

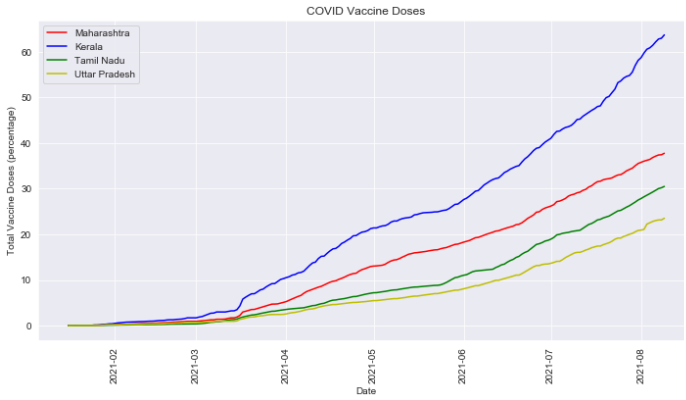

Fig 11 represents a side by side comparison of the daily cases and the daily doses of the 4 states. We clearly see that when the daily doses are high (mostly the later dates), the deaths also generally fall in number (except for a few outliers). So there is a general trend of less deaths after a period of high vaccine dosage, which fits the accepted thought of the vaccine preventing the fatality of the virus.

But a surprising result is the fact that although, by the cases and deaths metric, Uttar Pradesh has controlled the virus most efficiently, the percentage of the population vaccinated is the lowest, whereas Kerala, which had the highest percentage of covid infections, has the highest percentage of vaccinations as well. This illustrates the point of vaccines not being sufficient to prevent you from getting the virus, it is supposed to help prevent the fatality rate of the virus.

Fig. 10 Percentage of population vaccinated

We see from these vaccination charts that Uttar Pradesh has administered the most doses of the Vaccine, closely followed by Maharashtra, which is not a surprise, given the high populations of both the states.

Tamil Nadu and Kerala are also relatively similar in the total vaccinations metric, but diverge in the percentage.

Fig. 11 Comparison between the deaths and the doses for the four states

ACKNOWLEDGEMENT

I would like to thank my mentor Akshat, for constantly clarifying my doubts and giving me a path to approach this project. I would also like to thank the WIDS team for providing me and many others with the opportunity to learn and apply what we have learnt about Data Analysis. This first project will give me a foundation on which I build on in my later projects

REFERENCES

https://www.youtube.com/watch?v=GPVsHOlRBBI
https://pandas.pydata.org/docs/
https://matplotlib.org/stable/users/index.html
https://seaborn.pydata.org/