# Winter In Data Science

Covid-19 Data Analysis


Submitted By-
200100105
Niteesh Singh

Mentor-
19b080003
Akshat Agrawal

# Abstract

Coronavirus is lasting much longer than expected. Since its first case was reported on December 19 in Wuhan, China it has spread to more than 223 countries and Territories with over 300+ million cases worldwide. It has a huge impact on the economy, health, mental wellness, and education both at the individual and the national level.
In this study, we will take a brief look at its impact and spread over different countries then we will look at three countries' data namely India, China, and the USA to observe and analyze which country was more effective in flattening the curves and curbing the peaks. In the second part of the analysis, we will focus only on India and we will see a state-wise distribution of cases, death rates, recovery rates, and vaccination status.
The third part will be where we will be implementing a simple machine learning model to predict daily new cases using linear regression and curve fitting.

# Introduction

The Coronavirus disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was declared a pandemic on 11th March 2020 Owning to more than 250 million cases and more than 5 million deaths. The fast spread of the phenomenal Coronavirus pandemic has placed the world in peril and changed the worldwide viewpoint out of the blue. Numerous nations have taken on different ways to manage the pandemic. Using resources carefully to adequately control and confine the spread of cases.

That's where comes the role of Data Analysis to analyze the efficiency of the different strategies adopted worldwide and within a nation, to further decide which strategies should be adopted. It helps governments to channel the use of resources and decide on strategies. It helps to predict the peaks of coronavirus cases and helps the government to decide lockdown rules and durations.

It also helps Scientists and Doctors to analyze the trend virus spread is following and get some idea of which conditions are appropriate for virus spread and came up with strategize to tackle it.
In the final section, we will implement a machine learning model to predict the daily new number of cases rising using time series analysis using the time-step method.

We will make use of some standard libraries of python like pandas, NumPy, and sci-kit-learn. We will be using the google-colab notebook for coding and analysis.

# Datasets

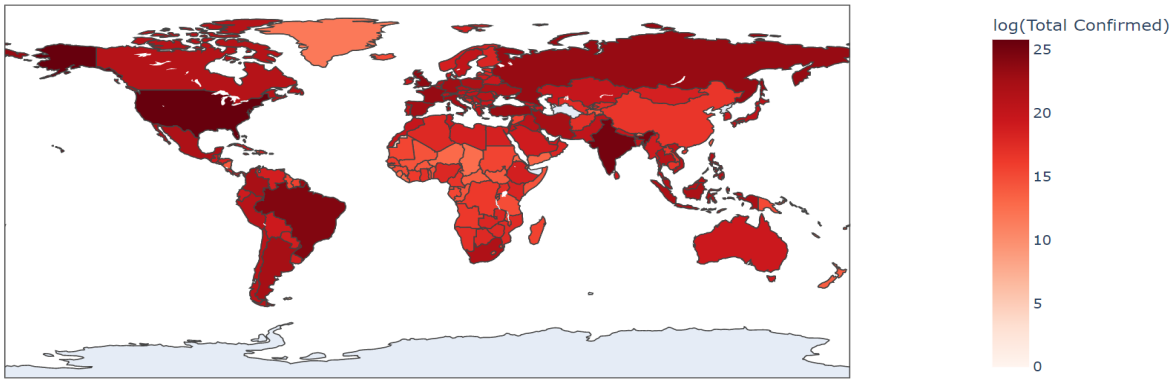
- ★ A global dataset that contains a country-wise summary of total deaths, total cases, total recovery, and population. Link to this dataset
- ★ A date-wise dataset of daily new cases, daily deaths for different countries, and cumulative total cases. Link to this dataset
- ★ Statewise dataset for India. Link to this dataset

# Analyzing the Impact at Global Level

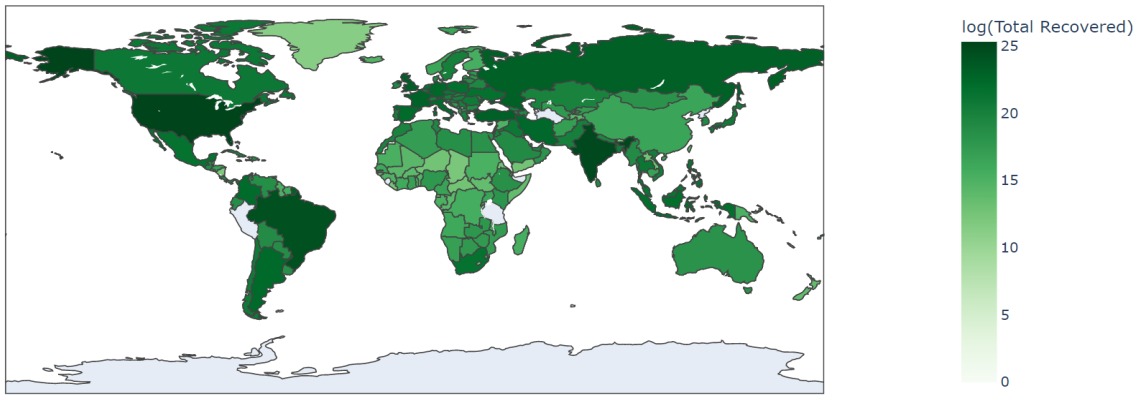
Looking at the dataset we get the total number of confirmed cases, total deaths, and recovered cases around the globe as 299040599, 5485341.0, 255368603.0. While the total population in which the data was calculated was 7873291916.

The recovery rate at the global level (recovered cases/confirmed cases) is approximately 85% whereas the death rate was approximately 1.8%. Total confirmed cases out of the total test conducted to give the positivity rate that is nearly 6%.

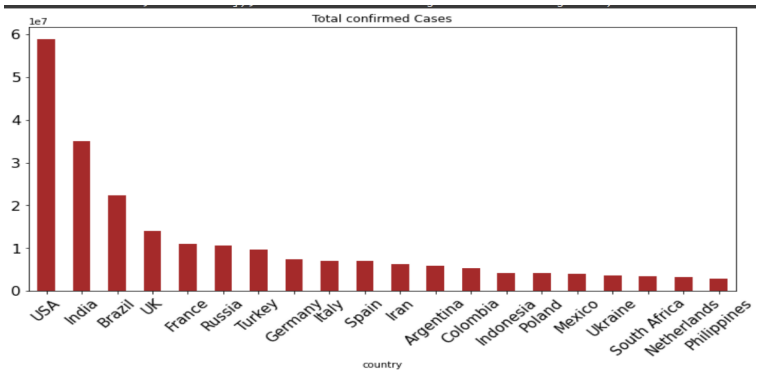
**Confirmed Cases Around The Globe**




**Recovered Cases Around The Globe**



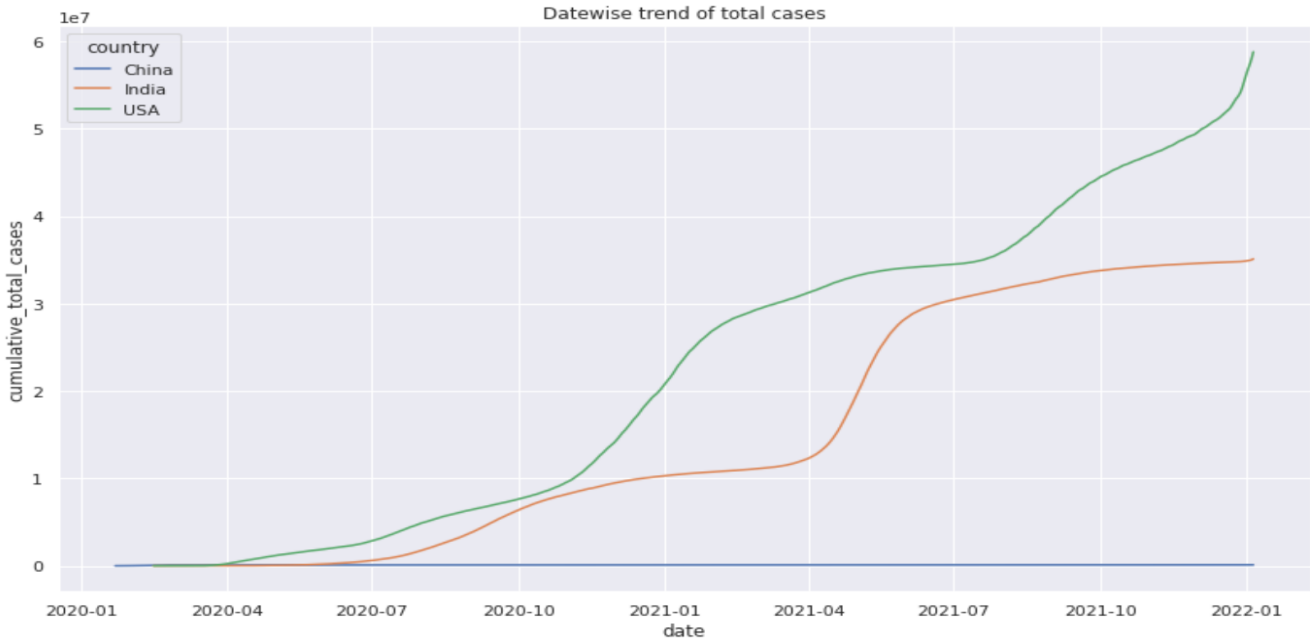

It is visible how widespread this disease is and also that the (SARS-CoV-2) virus can withstand and spread in every type of geographical and environmental condition. From the dataset, we get the top affected countries are the USA, India, and Brazil.



# Trend In USA, India, China

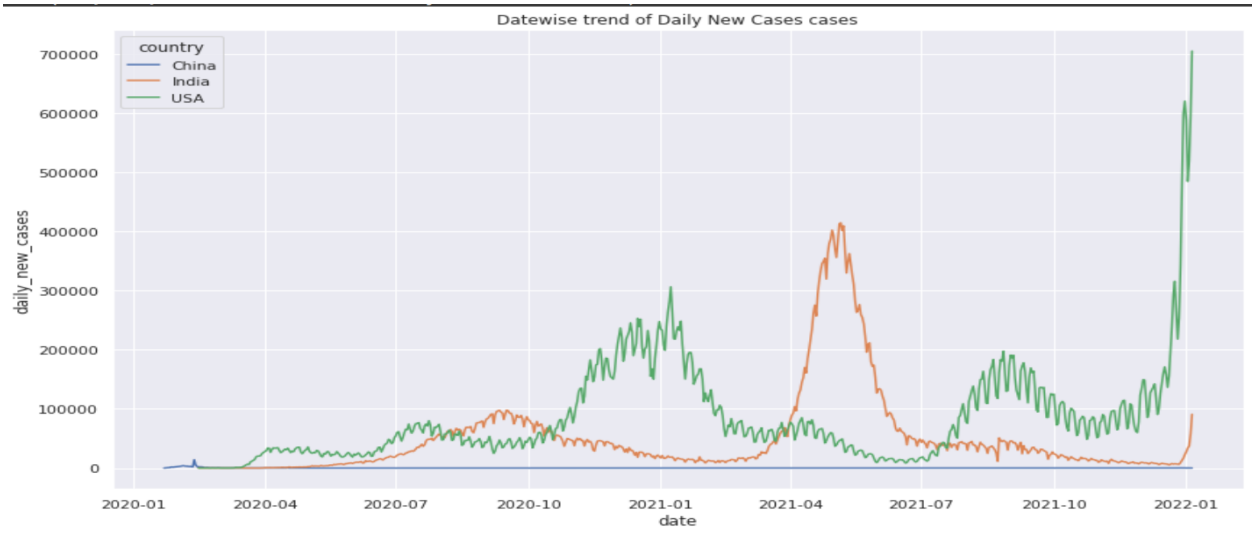
Now we will focus on three countries mainly china from where the virus started, the USA developed country with available resources and economic and political importance, and India -a developing country with a large population.

***The datewise trend of total confirmed cases in these countries***



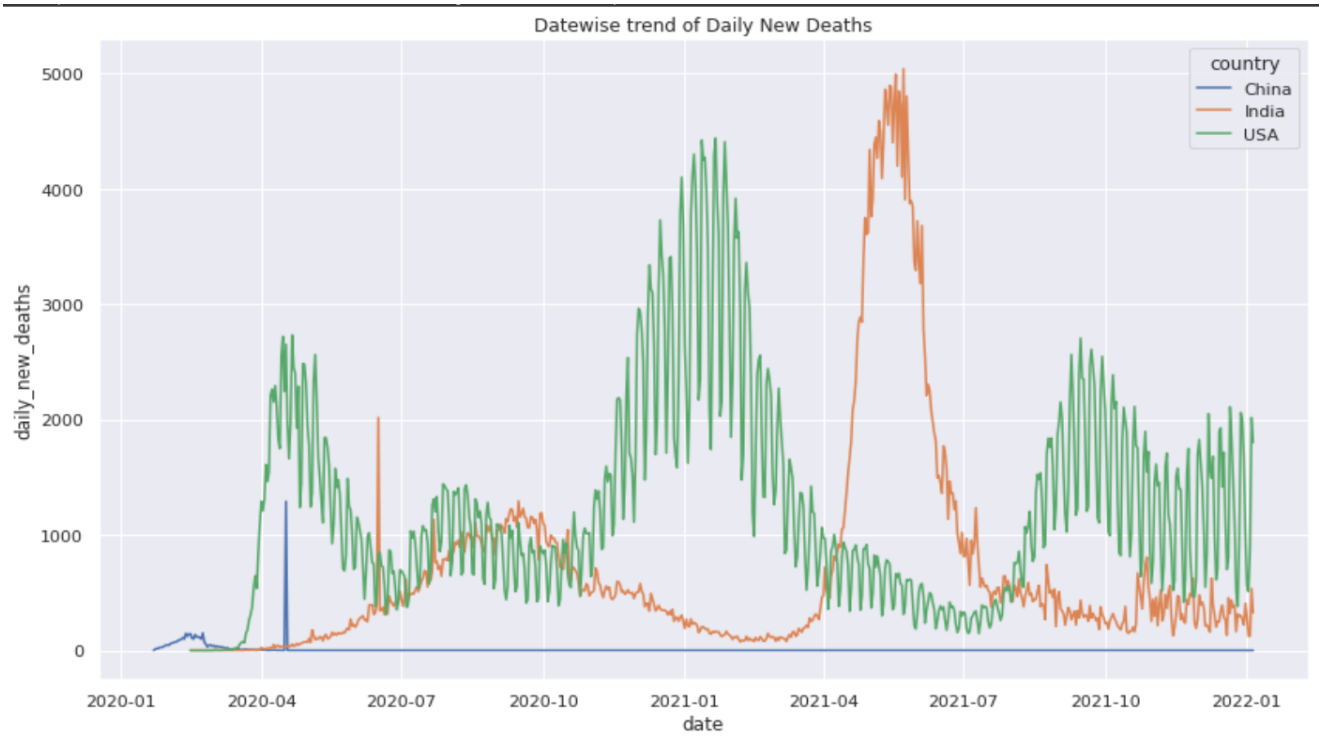

It is observable that the cases were more in china at starting of the pandemic but the cases reduce significantly after some time the cases were negligible as compared to the USA and India which shows that china was quite successful in preventing the spread of the virus. In India, the number of cases increases very rapidly between 2020/07-2020/10 and 2021/04-2021/06 while in the USA the total number of cases rises very rapidly throughout the year.

***The datewise trend of daily new cases in these countries***



The peaks in the chart represent the first second and third waves respectively. In China, this peak is very small and at the beginning supports our claim that covid started in china but china managed to suppress the spread quite effectively.
While in India first wave hit nearly in aug-sep 2020 while the second wave was in April-march 2021 and more people get affected during the second wave it is observable that the second wave reduces quite fast (not widespread as the first) it may be due to starting of vaccination. A similar trend is followed in the case of the USA.

***The datewise trend of daily deaths in these countries***

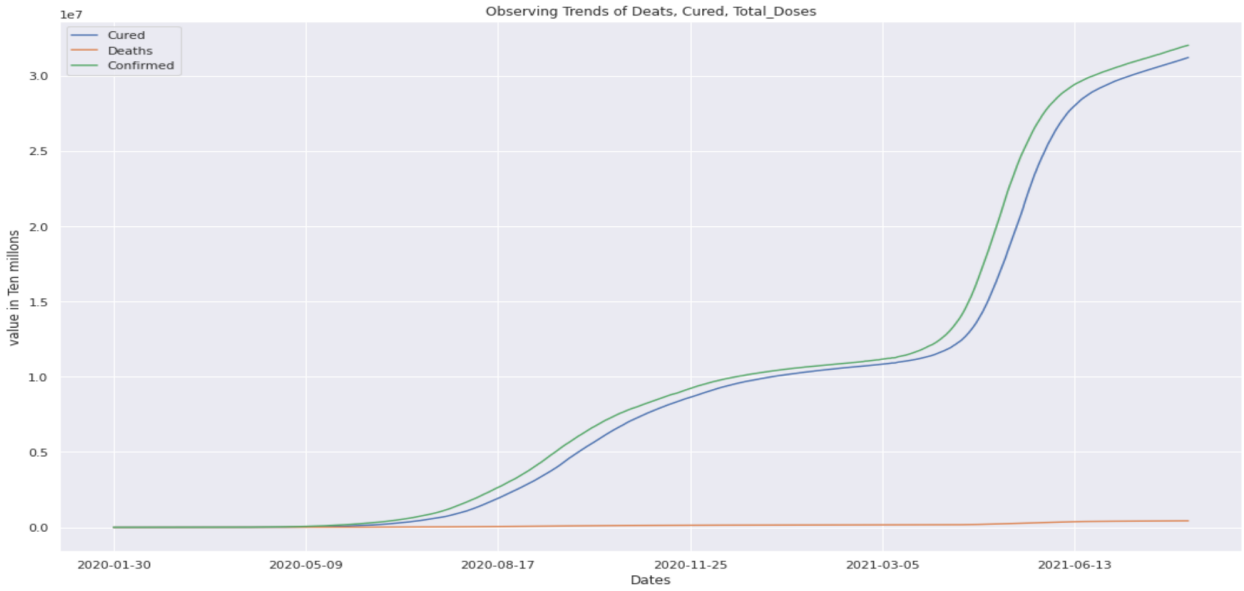



The daily deaths follow a similar trend as daily new cases and it is quite obvious that as the number of cases rises the number of deaths also rises. The number of death cases is more in the second wave as compared to in the first wave. We can also observe that the lockdown imposed was quite effective because after the peaks flatten the number decreases to a very low value which means the lockdown suppressed the number of cases effectively.

# Analysis of covid trends in INDIA

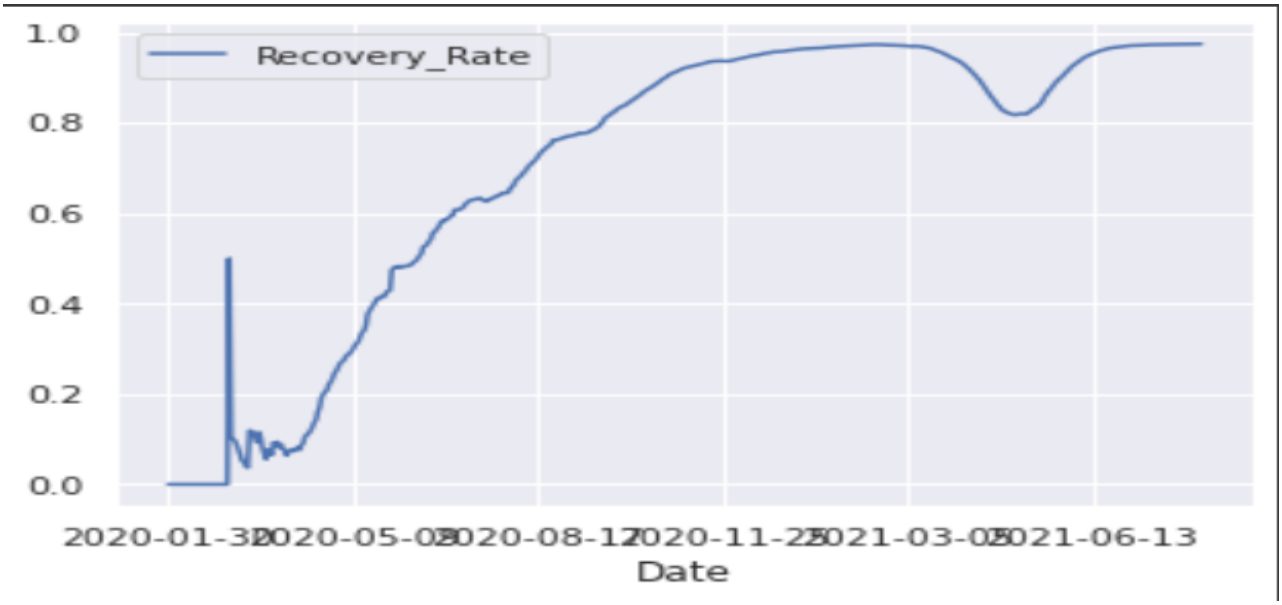
***The trend of total cured, recovered, deaths with time***

The first few cases in India came when few students return from Wuhan china. The first case was registered on 30th Jan 2020 the other few cases came only due to the return of infected people from other countries like China etc. Daily cases peaked in mid-September 2020 with over 90,000 cases reported per day, dropping to below 15,000 in January 2021. A second wave beginning in March 2021 was much more devastating than the first, with shortages of vaccines, hospital beds, oxygen cylinders, and other medical supplies in parts of the country. By late April, India led the world in new and active cases. On 30 April 2021, it became the first country to report over 400,000 new cases in 24 hours.




***The trend of the Recovery rate in India***

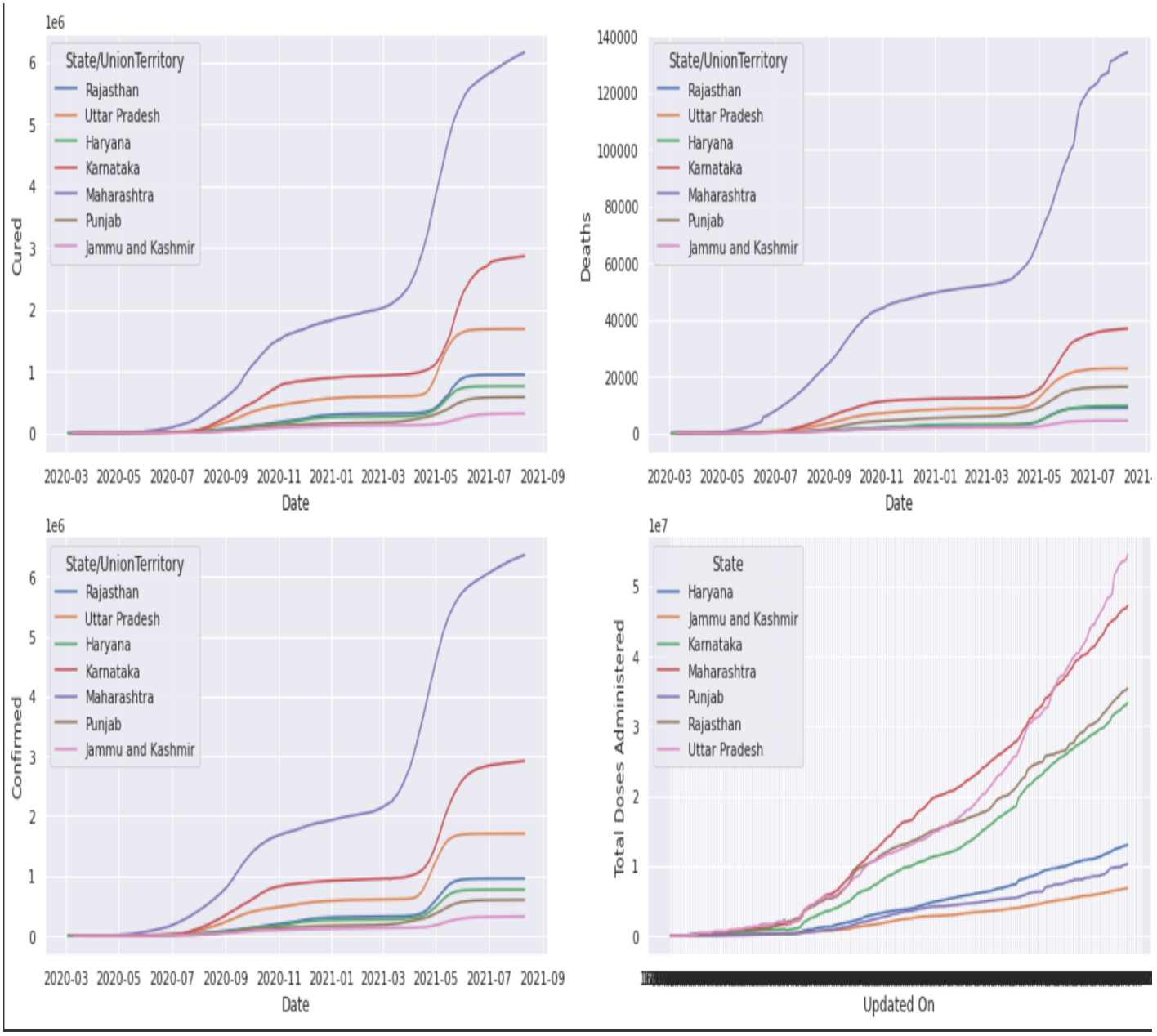
Recover rate is defined as the total number of cured cases to the total number of confirmed cases.




As it can be interpreted that at the starting due to a lack of proper knowledge of a cure for the disease recovery rate was quite low but as the time passed we get the proper cure and vaccination for covid due to which recovery rate saturated to 1. The dip in the flattened region in the graph is due to a huge increase in confirmed cases during the second wave.

***Statewise Trends***

Let's look at the state-wise statistics of corona cases in India which will tell which states were most affected. We will compare death rates, daily cases, and vaccination status to see the trend and effect of vaccination over decreasing daily new cases in different states. The top leading states were Maharashtra, Karnataka, and Uttar Pradesh in the total number of active and confirmed cases.




Kerala was the first state to report a corona case whereas it controlled the rise of cases. It depends upon the policies followed by a particular state and how far the restriction was imposed on and followed. Mumbai being one of the biggest cities and being the financial capital it was hard to impose a full lockdown and hence as a result the number of cases is high in Maharashtra. The other thing to notice is that Uttar Pradesh is the third most affected state which may be due to its large population.

# Predictive Models

***Multiple Regression Model***

Multiple regression is simply fitting a line to data points which can be phrased as an optimization by defining a cost function. In general, if our feature space contains features x1, x2, …, xn. Then the linear line y = w1x1 + w2x2 + … + wnxn. Can be assumed a fit to data where wi denotes the weight of the particular feature. The cost function will be the square of the difference between the predicted and actual value.

In our case, we have taken features such as No_of_days, State/UnionTerritories, Total_Deaths, and Cured cases and we will be predicting the number of total confirmed cases for a particular day. And we get the results as -
Mean Absolute Error: 16876.032053323263
Mean Squared Error: 1483529834.4635236
Root Mean Squared Error: 38516.61764048764

***Polynomial Regression***

Polynomial regression is similar to linear regression but we expand the degree of the independent variable. Which models and follows the dependent variable more accurately. If the feature variable is x then we extrapolate it and make a few more feature variables as x^2, x^3, …, x^n. Then we fit a Multivariable Regression Model with features as x, x^2, x^3,.., x^n while the hypothesis function is given by y = w1x + w2x^2 + … + wnx^n. The cost function remains the same i.e. it calculates the squared error between the predicted and actual value and tries to minimize that error and we get optimal values of w1, w2, …, wn.

In our case feature x was the number of days passed by and we extrapolated it to degree four thus we have featured as x, x^2, x^3, x^4.then we fitted a linear line to predict the number of cases that day. And we get the results as-
Mean Absolute Error: 282787.40165808925
Root Mean Squared Error: 532264.4074885099