# The COVID-19 Project

**Introduction:-**

COVID-19 commonly known as coronavirus was first identified in Wuhan, China in 2019, since then it has spread worldwide and caused a pandemic. It is an infectious disease caused by severe acute respiratory syndrome with a fatality rate of nearly 1 percent.

COVID-19 entered US through following possibilities

1. Imported cases in explorers
2. Cases among close contacts of a known case
3. Community-procured situations where the wellspring of the disease is obscure.



**Coronavirus Cases in USA till 25th of April 2020**

The infection that causes COVID-19 is thought to spread for the most part from individual to individual, predominantly through respiratory beads created when a contaminated individual hacks or wheezes. These beads can land in the mouths or noses of individuals who are close by or potentially be breathed in into the lungs.

Spread is more probable when individuals are in close contact with each other (inside around 6 feet).

The coronavirus pandemic is influencing each part of life in the United States now, and with that effect come some hard decisions. Who gets money related assistance from the national government and what amount? By what method will specialists choose who gets treatment and who doesn't if medical clinic assets are inadequate to treat everyone who needs consideration? The share market has fallen, a lot of businesses are closed.

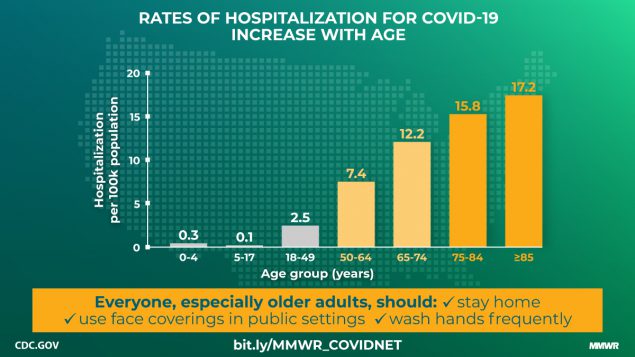
**Agencies responsible for pandemic disease like COVID 19 In US**

* **White House Coronavirus Task Force**

The White Houseand Centers for Disease Control and Prevention (CDC)Coronavirus Task Force is a [United States Department of State](https://en.wikipedia.org/wiki/United_States_Department_of_State) task force that "coordinates and oversees the Administration's efforts to monitor, prevent, contain, and mitigate the spread" of the [coronavirus disease](https://en.wikipedia.org/wiki/Coronavirus_disease_2019) (COVID-19).

Dr. Anthony Fauci the "face of the federal government's response". **Anthony Stephen Fauci** is an American physician and [immunologist](https://en.wikipedia.org/wiki/Immunology) who has served as the director of the [National Institute of Allergy and Infectious Diseases](https://en.wikipedia.org/wiki/National_Institute_of_Allergy_and_Infectious_Diseases) (NIAID) since 1984. Since January 2020, he has been one of the lead members of the [Trump Administration's](https://en.wikipedia.org/wiki/Presidency_of_Donald_Trump) [White House Coronavirus Task Force](https://en.wikipedia.org/wiki/White_House_Coronavirus_Task_Force) addressing the [2019–20 coronavirus pandemic](https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic) [in the United States](https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_the_United_States).

* **Centers for Disease Control and Prevention** (**CDC**)

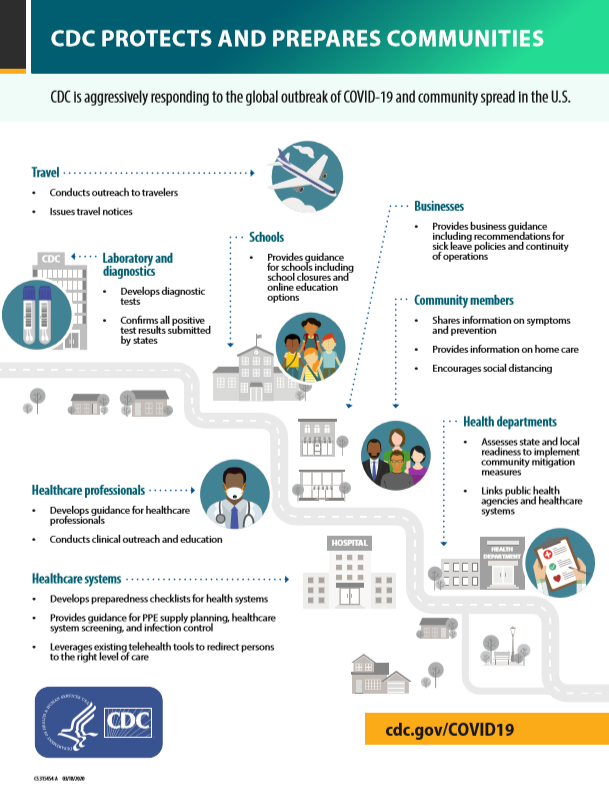


The **Centers for Disease Control and Prevention** (**CDC**) is the leading national public health institute of the United States. Its main goal is to protect [public health](https://en.wikipedia.org/wiki/Public_health) and safety through the control and prevention of disease, injury, and [disability](https://en.wikipedia.org/wiki/Disability) in the US and internationally.

**Role of CDC** : CDC is responsible for controlling the introduction and spread of infectious diseases, and provides consultation and assistance to other nations and international agencies to assist in improving their disease prevention and control, environmental health, and health promotion activities.

**Summary of CDC’ action plans:-**

**Dataset website** - <https://covidtracking.com/data>



The COVID Tracking Project collects its data from state/district/territory public health authorities—or, occasionally, from trusted news reporting, official press conferences, or (very occasionally) tweets or Facebook updates from state public health authorities or governors.

**Description of the data :**

COVID-19 dataset comprises the total number of tests conducted, breaking out positive, negative, and hospitalized patients. The data is collected from ”The COVID Tracking Project” collects its data from state/district/territory public health authorities—or, occasionally, from trusted news reporting, official press conferences, or tweets or Facebook updates from state public health authorities or governors.

**Description of columns -**

* State - State or territory postal code abbreviation.
* Positive - Total cumulative positive test results.
* Negative - Total cumulative negative test results
* Death - Total cumulative number of people that have died.
* Total test - Total no of tests conducted.
* Hospitalized - Total cumulative number of people hospitalized.

**Limitations** - Not all the states consistently report their test results and regularly. In such cases, they use other reporting tools like directly asking state officials, watching news conferences, gleaning information from trusted news sources. Moreover, since the symptoms are not visible until 14 days, the number of actual positive cases may be more than reported to the state/district/territory public health authorities, etc.

**Data Analysis Tools: R-studio, Tableau**

The tool that we will use to manipulate the data will be R-studio for data analysis and Tableau for visualizations.

**R-Studio:**

RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. It provides free and open-source tools for R and enterprise-ready professional software for data science teams to develop and share their work at scale. RStudio makes it easy to set your working directory and access files on your computer.

**R-Studio for Data Analysis:**

R is very important in data science because of its versatility in the field of statistics. R is usually used in the field of data science when the task requires a special analysis of data for standalone or distributed computing.R is also perfect for exploration. It can be used in any kind of analysis work, as it has many tools and is also very extensible. Additionally, it is a perfect fit for big data solutions. Following are some of the highlights which show why R is important for data science:

* Data analysis software.
* Statistical analysis environment
* Open-source

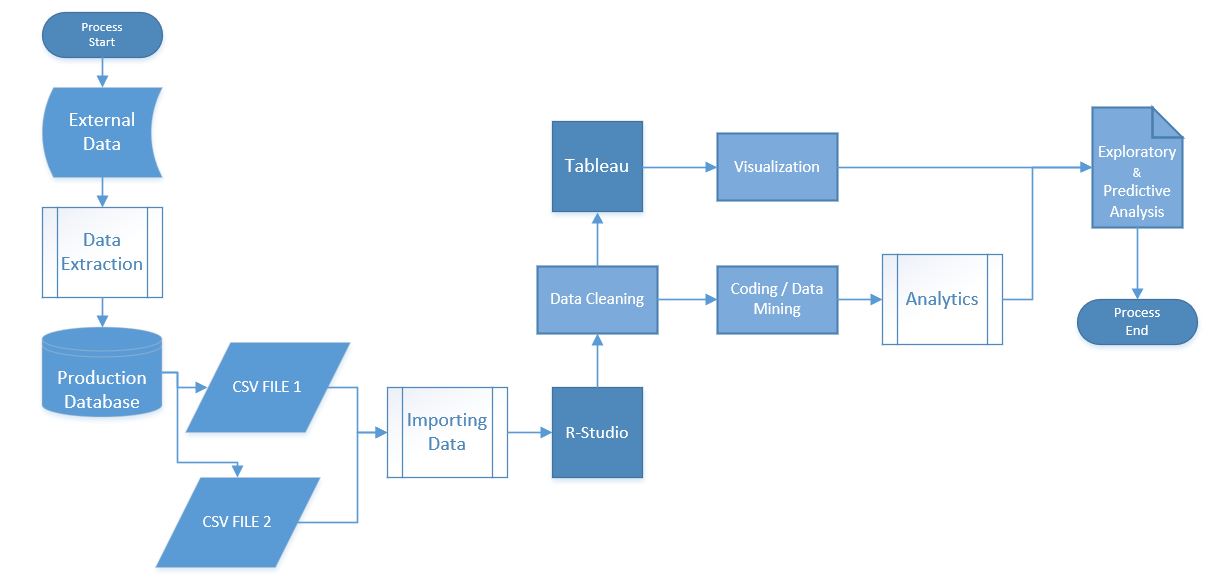
So, most of the development of the R language is done by keeping data science and statistics in mind. As a result, R has become the default choice for data science applications and data science professionals.

**Tableau:**

Tableau is a powerful and fastest-growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data into a very easily understandable format. Tableau is an interactive, self-service reporting and analytics tool that enables faculty and staff to integrate and combine data from multiple sources into visualizations and be accessed in a single desktop environment using Tableau Desktop or through a shared dashboard.

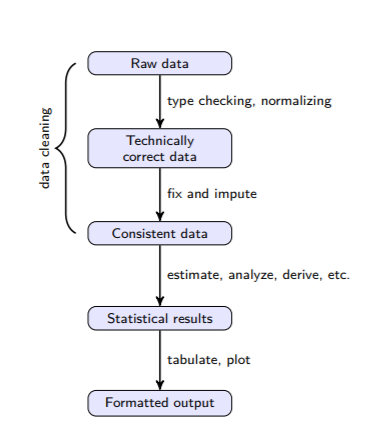
**Data Flow Chart**

The software Microsoft Visio is used to prepare a process flow chart of how the data moves throughout different phases of the project. Firstly, the data is extracted from covidtracking.com and stored in the production database that is the local server. The data from the Production database is imported into data analytics tools (R-studio and Tableau) for sandboxing, where data cleaning, manipulation, and analysis is carried out. Finally, the Information and insights from the data analysis are documented.



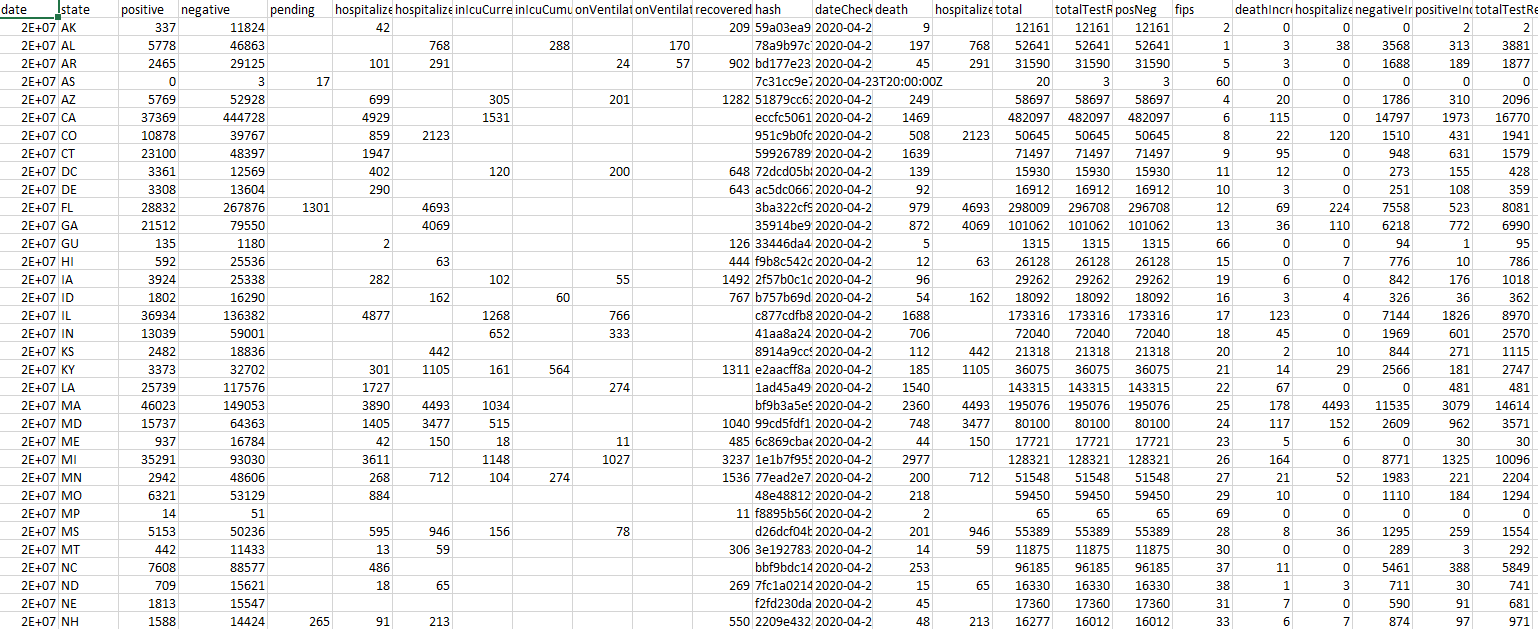
**Data cleaning and manipulation**

Firstly, the required packages are installed and recalled to load the data into R-studio. Then the columns in the data are analyzed to eliminate the redundant columns from the dataset. Secondly, the data is cleaned. Then data is checked for any missing values in the columns and subsequently, the missing values are removed from the dataset and by replacing them with 0.



**Production data set:**

Our data set is taken from <https://covidtracking.com/data/us-daily> and it contains positive, negative, pending, death, and hospitalized cases of covid-19 in different states of the US.



**Exploration of data :**

While exploring the data, the top 5 states with highest positive cases are in New York, New Jersey, Massachusetts, California and Illinois. On the other hand, the bottom five states with least positive cases are Wyoming, Guam, Virgin Islands, Northern Marianas and American Samoa. The top 5 states with highest death cases are in New York, New Jersey, Michigan, Massachusetts and Illinois. On the other hand, the states with the least death cases are in Wyoming, Guam, Virgin Islands, Northern Marians and American Samoa. The top 5 states with highest recovery cases are New York, Texas, Tennessee, South Carolina and Michigan while the bottom 5 states with least recovery number are Oregon, Pennsylvania, Puerto Rico, Washington and Wisconsin.

**Geo-Mapping & Color-Coding in Tableau :**

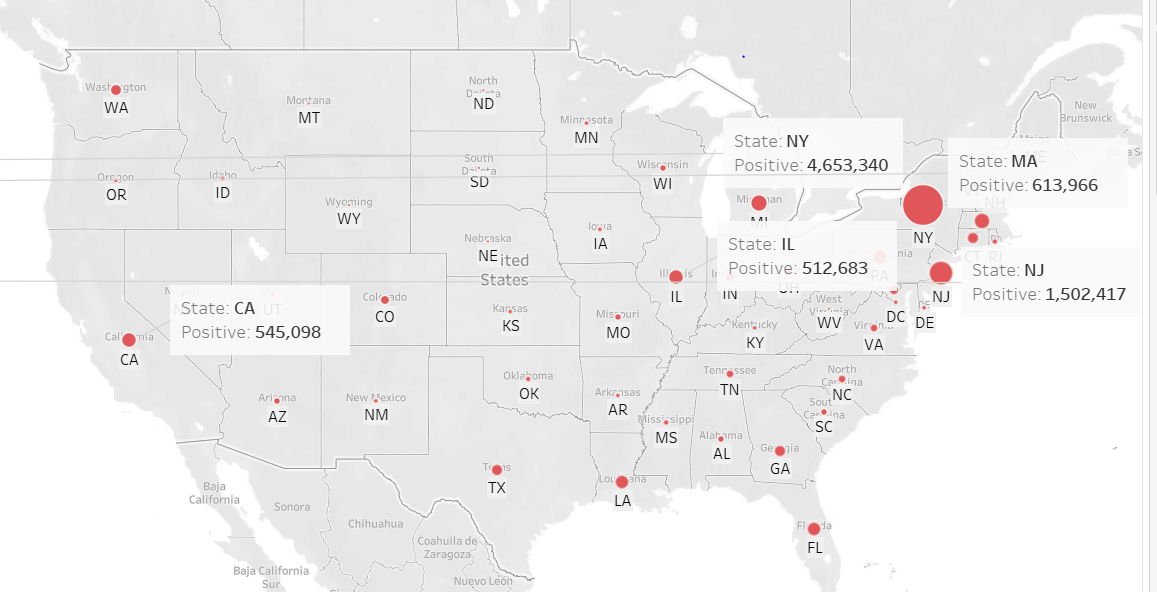
We have obtained insights from the data set about the top 5 states that have the highest positive cases, highest death rates and the recovery of COVID-19 using Tableau.

**Analysis of the data:**

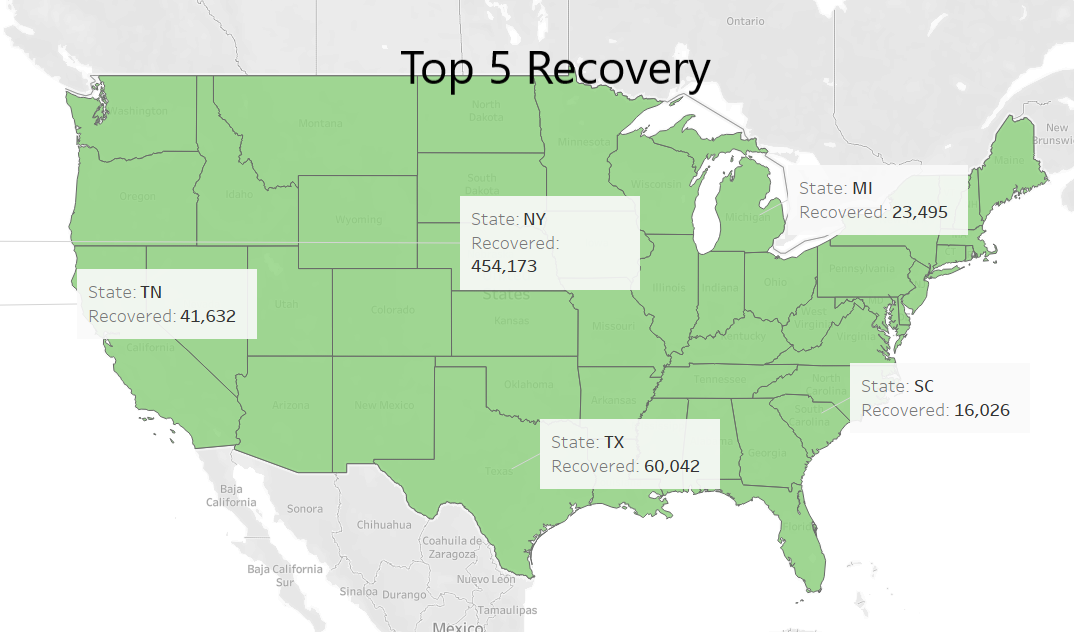
Data analysis is a process of inspecting, [cleansing](https://en.wikipedia.org/wiki/Data_cleansing), [transforming](https://en.wikipedia.org/wiki/Data_transformation), and [modeling](https://en.wikipedia.org/wiki/Data_modeling) [data](https://en.wikipedia.org/wiki/Data) with the goal of discovering useful information, informing conclusions, and supporting decision-making. Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a [process](https://en.wikipedia.org/wiki/Process_theory) for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses, or disprove theories.

Here is the analysis that has been done to obtain useful information from the huge dataset of USA COVID-19.We have analyzed the following insights from the COVID-19 datasets:

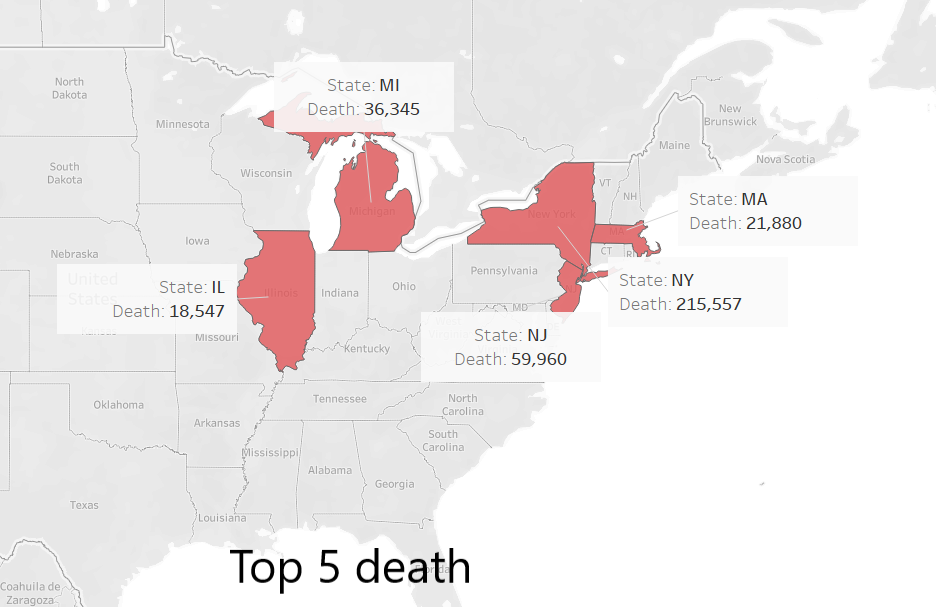
* **Studying the trendline to get more insights about the top 5 states.**

**Figure:- -** Top 5 states that have the highest positive

**Figure:-** Top 5 states that have the highest recovery cases

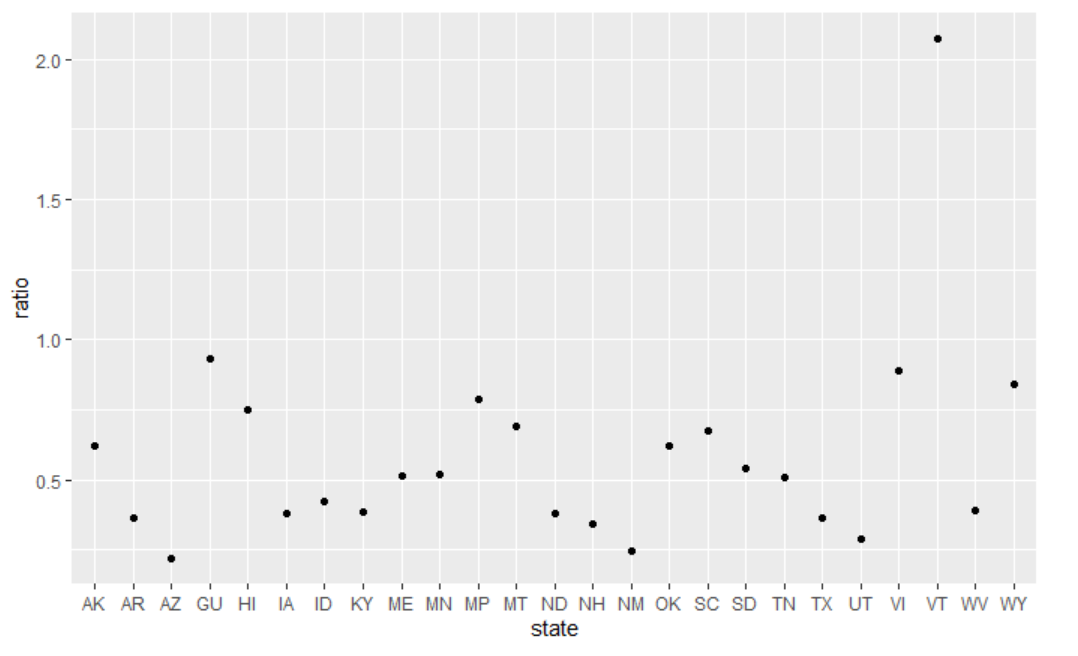


**Figure:-** Top 5 states that have the highest death cases



* **Which state will be the first to get out of this pandemic.**

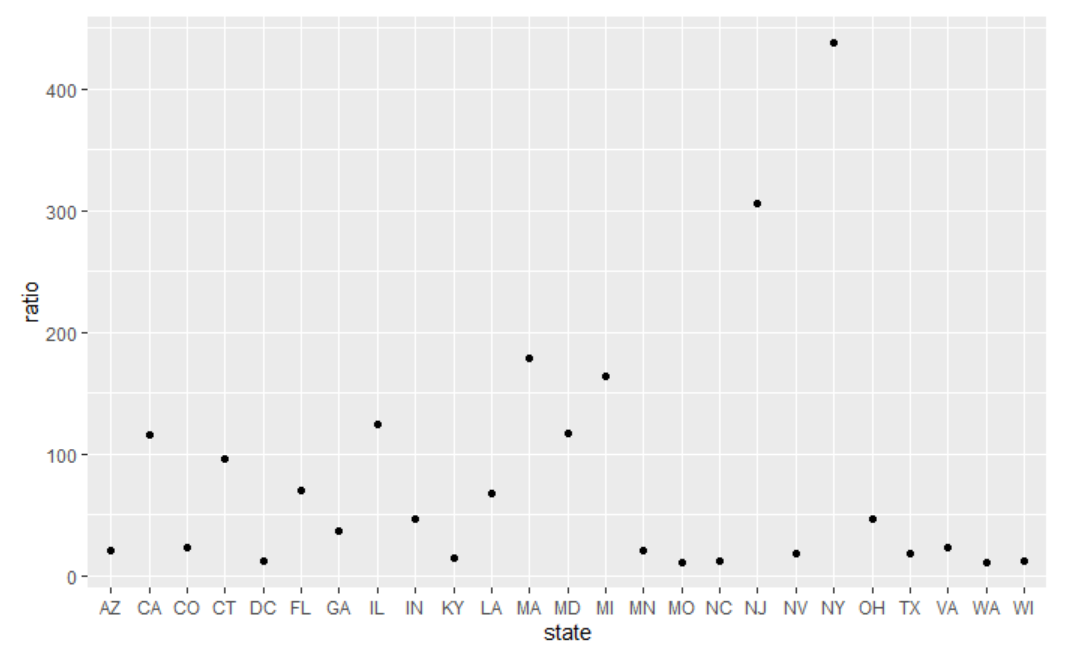
Below graph shows the graphical representation of the score of the states that will get out of this pandemic. Highest score implies fastest recovery and from the result we can see that Vermont(VT) has the highest possibility to get out of this pandemic and is safe compared to other states.



**Figure:** Graphical representation of higher possibility to recover fastest

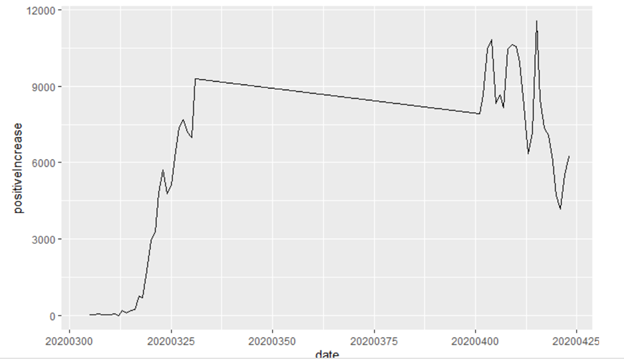
* **Identifying which state in the US has been hit hardest.**

The below graph shows the graphical representation of the score of the states that has been hit hardest and will be the last to get out of this pandemic. Highest score implies least recovery and And from the result, we can see that New York(NY) is in a critical situation and New Jersey(NJ) is in the second position.

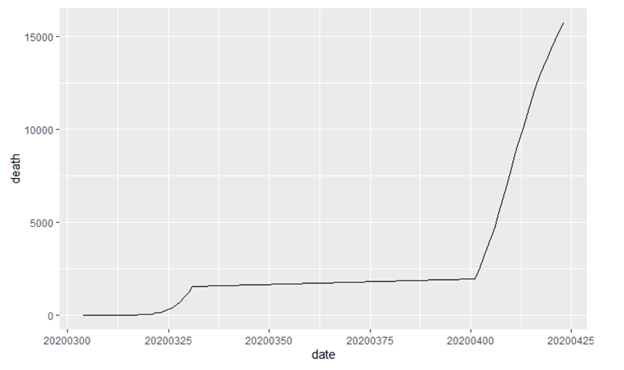
****

**Figure:-** Graphical representation of states that are in a critical situation

* **Trends of positive cases in USA**



* **Trends of deaths in USA**



**Future recommendations:**

* Digital platforms or apps to keep citizens informed, enable public participation, and/or offer open data; Digital tools to enable public participation.
* The government should take prior measures as the pandemic is with spurt increase every day otherwise the situation in socio-economic problems may lead to riots and chaos by poor and ill people.
* Controlling rumors to avoid panic among the public.