

|  |
| --- |
| Analysis of Novel Coronavirus (COVID-19) data for United States  ATDM |
|  |
| July 19  St Clair College of Applied Arts and Technology  Authored by: Data Wizard- Group 01 |



Contents

[Project- Submission 4](#_Toc77627391)

[Proposal 4](#_Toc77627392)

[Descriptive Statistics and Exploratory Data Analysis (EDA) 6](#_Toc77627393)

[Data Cleaning and Transformation 10](#_Toc77627394)

[Transformation using Python- 10](#_Toc77627395)

[Data Analysis 12](#_Toc77627396)

[Key patterns and visualization- 12](#_Toc77627397)

[Business Questions 28](#_Toc77627398)

Project Guide

Yingge Wang ([YWANG@stclaircollege.ca](mailto:YWANG@stclaircollege.ca))

Project Group Members

|  |  |  |
| --- | --- | --- |
| Student Name | Student ID | Student Email |
| Harsh Parmar | 779093 | [hp97@myscc.ca](mailto:hp97@myscc.ca) |
| Nishil Patel | 782837 | [np85@myscc.ca](mailto:np85@myscc.ca) |
| Raman Keshari | 783376 | [rk409@myscc.ca](mailto:rk409@myscc.ca) |
| Shreyas Mahendra | 770049 | [sm208@myscc.ca](mailto:sm208@myscc.ca) |
| Priyanka Harsukhbhai Ghetiya | 779258 | [pg48@myscc.ca](mailto:pg48@myscc.ca) |

## **Section Number- 002**

## **Academic Integrity**

We, Harsh Parmar, Nishil Patel, Raman Keshari, Shreyas Mahendra and Priyanka Harsukhbhai Ghetiya, hereby state that we have not communicated with or gained information in any way from any person or resource that would violate the College’s academic integrity policies, and that all work presented is our own. In addition, we also agree not to share our work in any way, before or after submission, that would violate the College’s academic integrity policies.

Analysis of Novel Coronavirus (COVID-19) data for United States

# Start Date- 1st June 2021 End Date- 28th July 2021

# Objectives- Analysis of the dataset using EDA and visualization

# This project study aims to describe the health outcomes of people diagnosed with COVID-19 in the US, over time and in relation to the characteristics of the virus, by combining COVID-19 information from different states, hospital, general practice and death registry data.

## **Background**

The novel coronavirus disease, named COVID-19 on 11 February 2020, is caused by SARS-CoV-2 virus. The outbreak was declared a Public Health Emergency of International Concern on 30 January 2020. While the number of confirmed cases worldwide and in the US is reported daily, detailed data on the outcomes of people who test positive for SARS-CoV-2, and predictors of outcomes, are still scarce.

Outcomes are likely to vary with context, including according to extensiveness of surveillance and testing, health systems functioning and population characteristics. Evidence to date has come primarily from various states of the US that are further along in the pandemic.

**Goal**

The aim of the project is to provide a list of priority areas for work and health research, that addresses evidence gaps and emerging evidence needs within the context of global pandemics generally and COVID19 specifically.

# Requirements

# The dataset is present on GitHub, which has been downloaded and loaded for our data analysis. Following are the technological requirement and tools used for this project.

|  |
| --- |
| Technological Requirements |
| Python 3.9 |
| Anaconda- Jupyter Notebook |
| Tableau |
| Excel |
| GitHub |
| MS office |

# Project- Submission

DataSet- USA daily state reports for Novel Coronavirus (COVID-19)

Source- <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data>

GitHub Link- https://github.com/ShreyasM6/ATDM-Project-Repo

## Proposal

We will be working with the publicly available dataset from JHU CSSE’s COVID-19 data repository. The data repository consists of different datasets related to covid-19.

What is the problem you are solving?

Data in this dataset is a collection of the covid-19 cases in the US, in which various locations are taken into consideration. The aim of the project is to provide data analysis of covid-19 pandemic, various cases have been studied like most affected areas due to this pandemic. Study of data from various states is combined to show the growth of cases and recovery. Through this project we are trying to achieve a sustainable solution and understanding of the behavioral trend of the crisis and, a step towards helping people to understand the spread and predict the raise in cases for the US.

The data analysis will provide a complete understanding of the critical information which will help address the business question that we are trying to answer. EDA and visualization are the most prominent way to perform analysis and obtain the required or desired details.

**Who benefits from this project?**

By the analysis of the dataset the results help any Government or national body to examine the current situation and establish clear protocols to help the stop of the virus. The measures collected in the dataset will provide an accurate statistic to take critical actions in order to contain the spread of the virus. The analysis of the tends in the dataset will be beneficial to have a better idea in the behavioral patterns of the cases and help to mitigate the spike. With the help of the results determined by the analysis, the local bodies can help to suppress any serious conditions that may arise, as the result would give us clear directions of what precautionary measures to be implemented. Hence by carrying out any kind of analytics to provide insights about the dataset will help a number of different sectors and avoid similar situation in future.

Analysis of the dataset using EDA and visualizing, will provide a way to convey information for any desired audience who has no prior knowledge about the dataset. The analysis using EDA and visualization can be taken advantage by the Government or national body to squeeze important details as much as possible.

**Why is the project important?**

The pandemic has already taken grip over life of people. Since the start of the pandemic, some countries are facing problem of ever-increasing cases. Through the data analysis of cases one can analyze how the US is doing in terms of controlling the pandemic. Analyzing data leads to adapt the prevention model that are doing great in terms of lowering the cases, Predictions can be made with the dataset available to the individual/country/organizations, thus helping to decide how far we are able to control the pandemic or up to what extent we should guide preventive measures.

Analysis is more important than ever during these unprecedented times. As a society, we've seen how important even basic line graphs, bar charts, and heat maps are to understanding the spread of the virus. We've heard a lot about various models in terms of predicting deaths associated with COVID-19. Many people want to see the data and understand the facts in this rapidly changing environment.

What are the business questions you are looking to answer or objectives you are looking to achieve from this project?

Business questions-

Which state was the most affected from covid-19?

Which state has the highest deaths from covid-19?

When did the COVID-19 cases had the highest peak?

When did the COVID-19 cases had the highest death rate?

When did the COVID-19 cases started to drop?

When did the death rate started to drop?

With the EDA and visualization performed, did you observe any pattern in the dataset?

Top five states confirmed and deaths statistics in the US?

Did you observe any unusual behavior of the dataset?

Why do you think EDA and visualization important?

Objectives-

With the help of the data we can estimate the cases and requirements that are essential for the treatment of patients effected by COVID 19. Improve on distribution and control the tools that are necessary in this critical times for example:- If a country is suffering from high amount of covid cases then the essentials are more required there and with the project we can understand how much tools are required for the stability of the situation.

The main objective is to describe the health outcomes of people diagnosed with COVID-19 in the US, over time and in relation to the characteristics of the virus, by combining COVID-19 information from different states, hospital, general practice and death registry data.

**Describe the data at a high level, explain the data collection process, source of data, etc**

The dataset represents the detailed characteristics of the 2019 Novel Coronavirus, which provides information about the US and provinces with the number of confirmed, deaths and recovered cases. This list includes a complete list of all sources ever used in the data set since January 2020. Some sources listed here (e.g. ECDC, US CDC, BNO News) are not currently relied upon as a source of data. The data is collected from multiple sources such as World Health Organization (WHO) and different US data sources at the state (Admin1) or county/city (Admin2) level.

The dataset which is used for analysis has categorical, continuous, and discrete data, the dataset has different data type like int, float, and object.

**What insights does this data give and how can it be used in future?**

The analyzed data provides information of different states in the United states which consists of the confirmed, death and recovered cases based on a timeframe and the location. This information can be used by the state government to analyze the current situation and help to suppress the severity.

The dataset provides the insights of the day-to-day measures of increase and decrease of the cases, deaths, and recovery for the US. This kind of data is very useful in the current situation as it gives the government, the Ngo’s, and other medical and social organization to analyze the covid affected areas for a better decision making.

This data is also future proof as it shows the pattern of transmission of the covid 19 and this kind of dataset can help the countries more resilient to any kind of pandemic in future.

What type of problems are solved with this approach?

With analytics, we are trying to acknowledge the general public about the pandemic and providing solutions to overcome this critical situation. In this project we are approaching with an open mindset and providing solution for general public rather than businesses.

Any problem related to COVID-19 or information can be fetched by the following preformed analysis, the insights gained by this can hugely incorporated in any real word model to address the needs of the situation in any given time period.

Can the performed analysis be implemented to any real word scenario?

Since the considered dataset is live data recorded for the US from 1st January 2020 to 13th July 2021, the performed analysis is a best fit for any real word scenario related to COVID-19. With the help of visuals provided below, decisions can be taken faster and act on the situation quicker than expected.

## Descriptive Statistics and Exploratory Data Analysis (EDA)

Step 1: Confirm the data is correctly loaded

Before loading the dataset to a dataframe, we need to verify and make sure if the dataset has consistent column names across all the files.

Importing the dataset can be achieved with the pandas packages in python, following is the Code to import a single file to dataframe-

import pandas as pd

covid=pd.read\_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse\_covid\_19\_data/csse\_covid\_19\_daily\_reports/01-01-2021.csv")

Step 2: Describe the data

Describe the data: The dataframe has 18 columns which gives information about the Province name and timeframe, the number of Confirmed cases, death toll and Recovered cases. It also provides statistical information which are derived columns (calculated columns) like Active cases, Incident\_Rate and Case\_Fatality\_Ratio which will be helpful to perform more EDA on the dataset.

Summary statistics: Summary statistics can be performed in python using dataframe.describe() function. The describe function will provide the measures to describe the dataframe like- count, mean, std, min, max and much more.

Breakdown of variables (numeric, ordinal, categorical)

|  |  |
| --- | --- |
| Variable Type | Column Names |
| Categorical | Province\_State, Country\_Region, ISO3 |
| Numerical | Lat, Long, Confirmed, Deaths, Recovered, Active, Incident\_Rate, Total\_Test\_Results, People\_Hospitalized,Case\_Fatality\_Ratio, UID, Testing\_Rate, Hospitalization\_Rate |
| Date | Last\_Update |
| Ordinal | FIPS |

Step 3: Check the validity of data

Define schema: With the help of df.dtype function from the pandas package, we can determine the schema of the dataframe. Most of the variables are objects, float, or integer types.

|  |  |
| --- | --- |
| Variable Type | Column Names |
| Object | Province\_State, Country\_Region, ISO3 |
| Integer | Confirmed, Deaths, Recovered, Active, FIPS, Total\_Test\_Results, People\_Hospitalized, UID |
| Float | Last\_Update, Lat, Long, Incident\_Rate, Case\_Fatality\_Ratio, Testing\_Rate, Hospitalization\_Rate |
| Date | Last\_Update |

Understand the data: The primary information from the dataset is the daily cases of COVID-19 from US. Which provides statistics about the number of confirmed, death and recovered based on the province and timeframe.

Step 4:  Answer the following questions:

1. Does the data include missing, incomplete, or invalid records?

Yes, our dataset has missing and incomplete records. The missing and incomplete data is handled based on the requirement of the analysis to be performed. In our cases we are replacing the missing values with 0.

2. Does your data include outliers?

Yes, our dataset has outliers. Since the recorded data is live data, we can not eliminate the outliers as they also play an important role in providing the critical information.

3. Is the data segmented into groups?

Yes, the dataset is divided in groups based on the province, and timeframe. Grouped data will help us to provide information based on states with number of confirmed/death/recovered, and month/year.

4. Is the data imbalanced (a large number of the records represent a majority class and very few records represent the minority class)?

No, the dataset is not largely concentrated or representing a majority class. The data is equally divided for all the stated in the US.

5. Are some data elements highly correlated with each other?

Yes, the dataset has correlation with each other, as few columns are dependent and helps us to determine the trend of another column and also derive a calculated column.

6. How was the data collected?

The data is collected from multiple sources such as World Health Organization (WHO), European Centre for Disease Prevention and Control (ECDC) and different US data sources at the state or county/city level. The data was loaded using pandas library into a dataframe from github.

7. What are the inclusion criteria for your data?

Inclusion criteria are defined on the key target feature which includes analysis on demographic, and geographic characteristics.

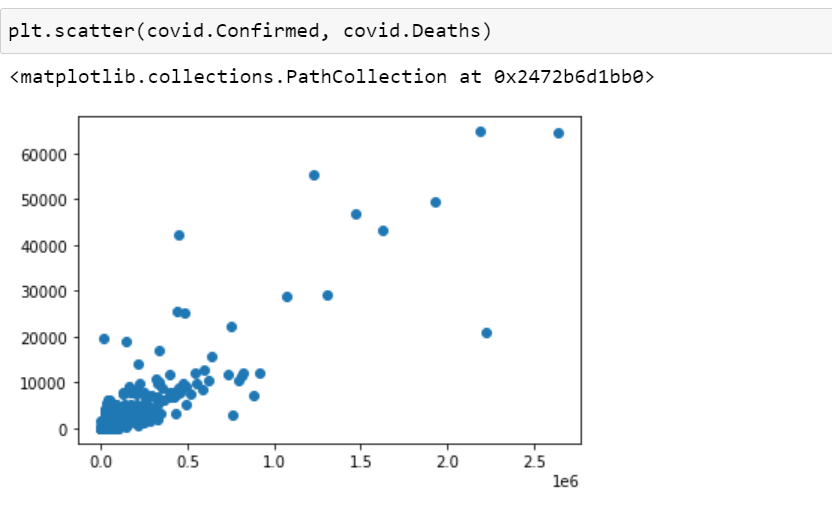
8. Can you generate preliminary visualizations for individual features?

Using the dataset we can induce visualization to check which province has the highest number of deaths or highest number of confirmed cases. By the visualization created we can have more data insights about the dataset.

Step 5: Use visualization to understand and explore, but not to explain

By using Matplotlib package we can produce visualization to understand the behavior of the dataset.

The following plot is a scatter plot displaying the relation between the number of confirmed and deaths. As we can see the number of confirmed and deaths are directly proportional.



## Data Cleaning and Transformation

The COVID-19 data files considered for EDA is from 1st January 2020 to 13th July 2021, about 460 files close to 27 thousand rows and 20+ columns.

### Transformation using Python-

1. Load all the required libraries for analysis-

Python provides a wide range of libraries for analysis which must be imported prior performing our analysis. In our EDA process we are importing the following libraries.

1)numpy package for arithmetic operation.

2)pandas package for dataframe handling.

3)OS package for interacting with the operating system.

4)glob package for retrieve files/pathnames.

5)date package to work with date and time.

1. Set the path from where the CSV files are picked-

Using the glob package we can set the path where the COVID-19 data files are present to load them into a single dataframe. Also list all the files present in the directory using the OS package.

1. Create a dataframe-

Using the pandas package, create an empty dataframe. The created empty dataframe will be used to load all the CSV files dynamically in a single run.

1. Load the files into a single DataFrame-

Loading multiple CSV files into a dataframe is achieved in a single piece of code, which iterates between all the files and append the data to the dataframe. The for-loop operator helps us to iterate each file and load all the files to a temporary dataframe present in the directory. The append operator consolidates all the files loaded in the temporary dataframe to the newly created dataframe. To avoid ambiguity while performing EDA- we have created a new column which holds the filename for each row, which allows us an ability to trace back for verification and validity.

1. Retrieve the information of the loaded dataframe-

After loading the data into the dataframe, we need to check the dimension of the dataframe and the data type of each columns, which helps us to understand the loaded data and perform any modifications if required. The .info operator retrieves all the information of a dataframe describing the categorical and continuous variables.

1. Find the sum of NaN values for each column-

It is important to know the number of NaN values present in the dataframe for each column before performing EDA. As the result will determine the way to handle the NaN values. In general, any real time dataset will contain many NaN values which requires us to eliminate handle them in an efficient manner which would not effect the original dataset.

1. Replace the NaN values-

As part of EDA, the key aspect is the way of handling the NaN values. Part of this solution is to drop the NaN values, however dropping the NaN values in our dataset will wipe out half of the data, resulting in losing valuable insights from the dataset. Although dropping NaN values might be a solution in some cases but replacing them with 0 is the best practice.

1. Data Transformation-

Substring the filename column to retrieve only the date information and exclude the file type. This transformation will help to create new columns and perform analysis based on a timeframe. It is very important to have a column which represent the timeframe of the data, which will let us to represent the data in a specific period of time.

1. Creating new columns from the existing columns-

Derive columns from already existing columns which provide information about the timestamp (Month, Year).

1. Rename the columns for readability-

Renaming the columns will help us to understand the column information better.

1. Write the transformed data into a CSV file-

To avoid performing transformation repeatedly, we are writing the transformed data from the dataframe to a CSV file.

*All the mentioned above EDA and transformation is performed using python on Jupyter Notebook, the notebook has the required code along with headings and comments which has been uploaded on GitHub.*

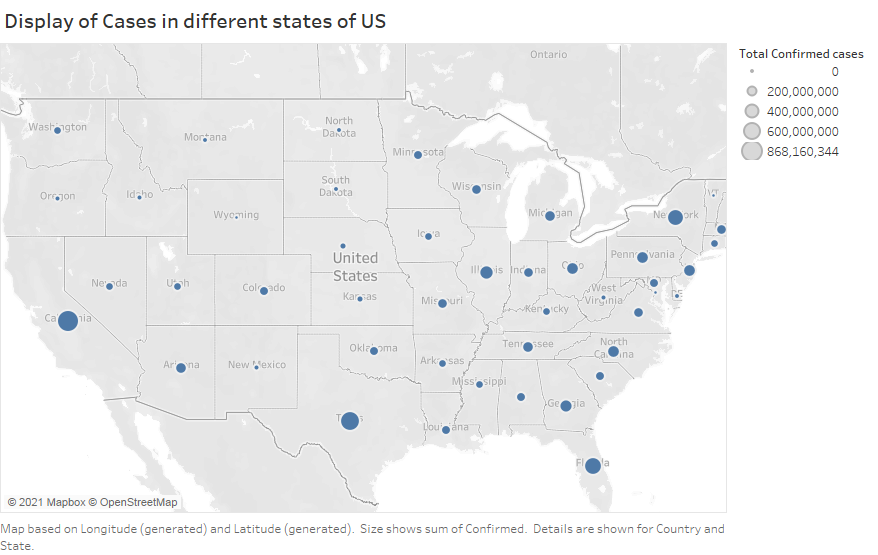
## Data Analysis

The COVID-19 dataset is visualized in Tableau after transformation, to identify the key patterns and to answer the business questions.

### Key patterns and visualization-

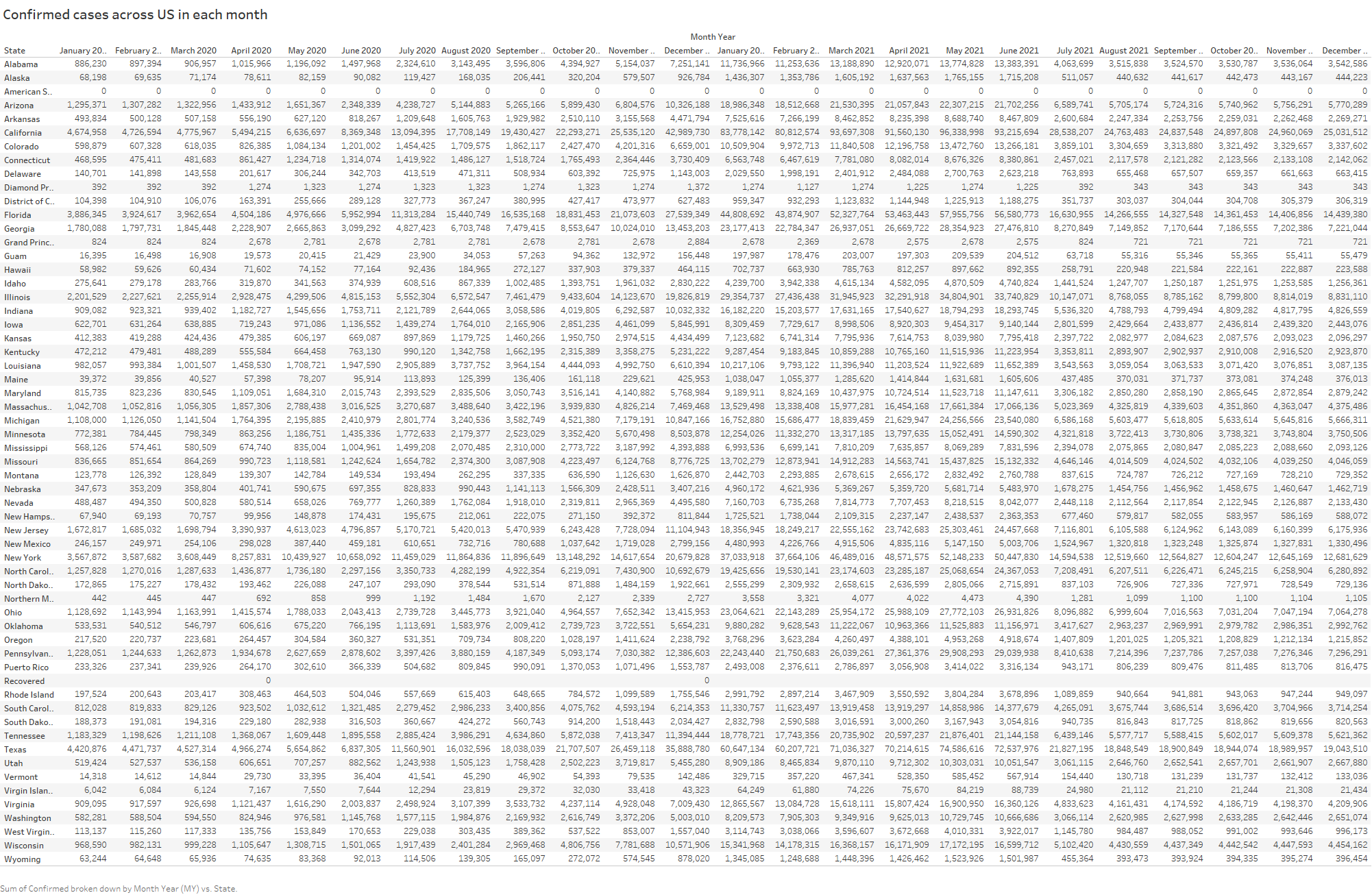
Load the transformed data to Tableau for visualization

*Confirmed cases in US*



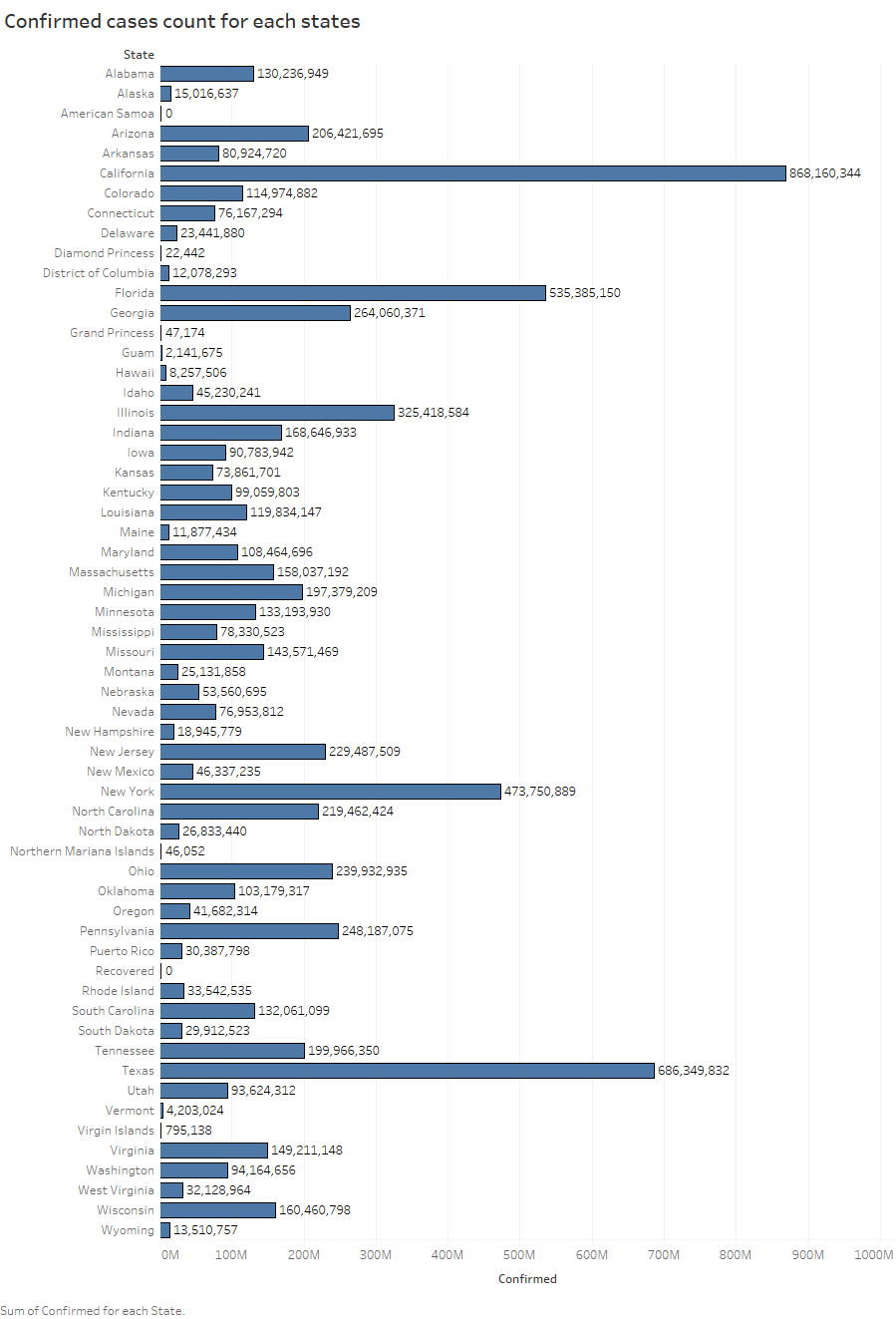
The above geographical distribution of the cases in US displays the spread of COVID-19 virus. The larger the size of the circle, larger is the intensity.

Statistical information about the confirmed cases in all the states of the US for each month starting from January 2020 to July 2021

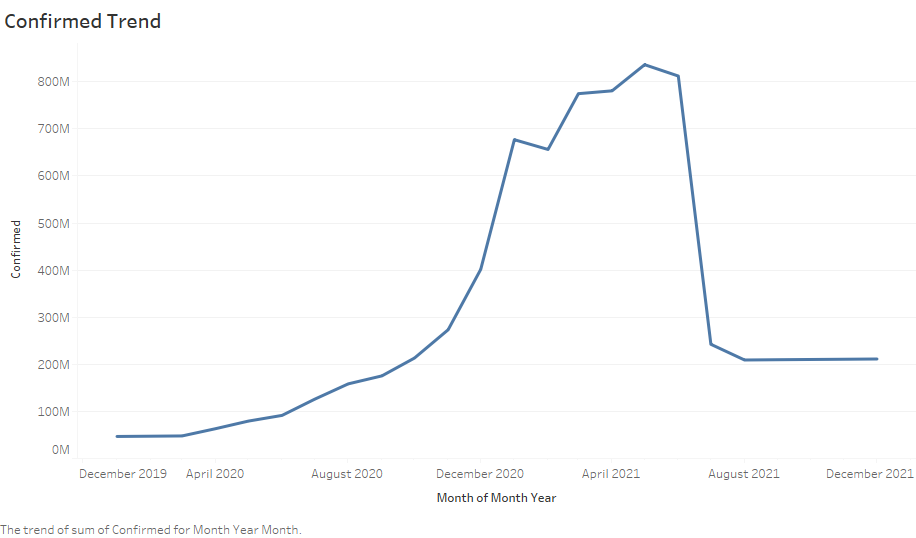


While delivering the results found during statistics, the report has to be precise and concise, with the help of the above tabular representation of the data for the total confirmed cases for each state in a month, provides complete information in a single report. Anybody looking at this report can acquire knowledge about the number of cases confirmed in the US, in addition this report provides specific information with states and timeframe.

Bar Plot representation of the confirmed cases in US



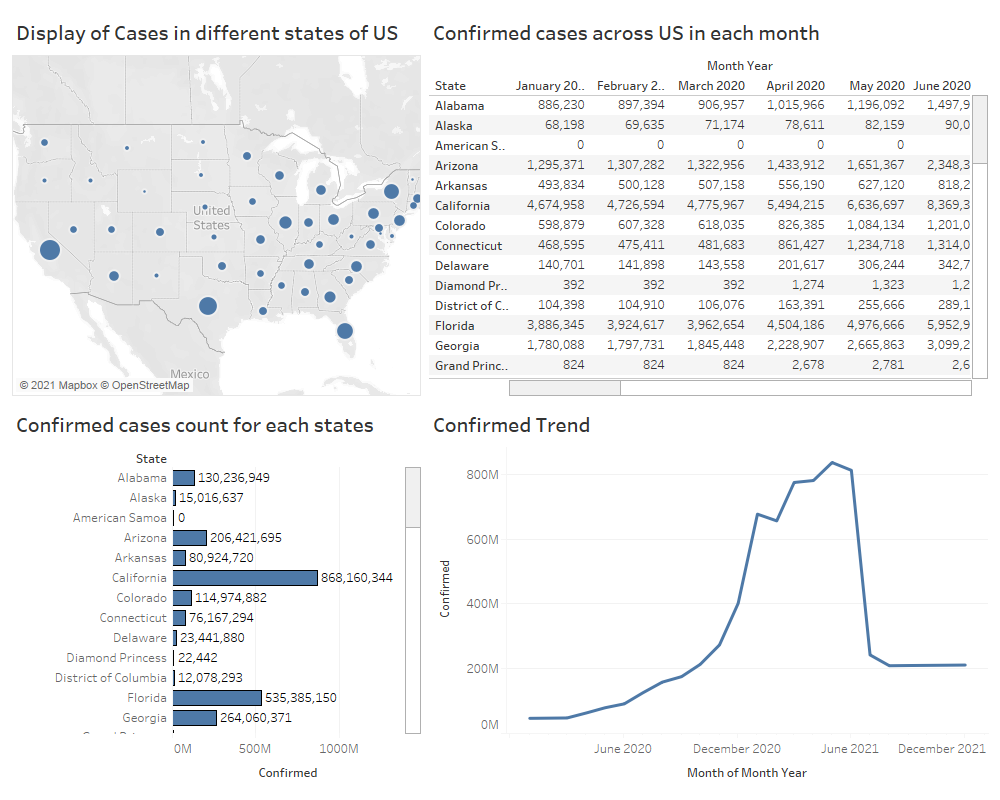
Line plot representing the trend of confirmed cases over the timeframe



The above line graph displays the information of the rise and fall of the COVID-19 cases in US.

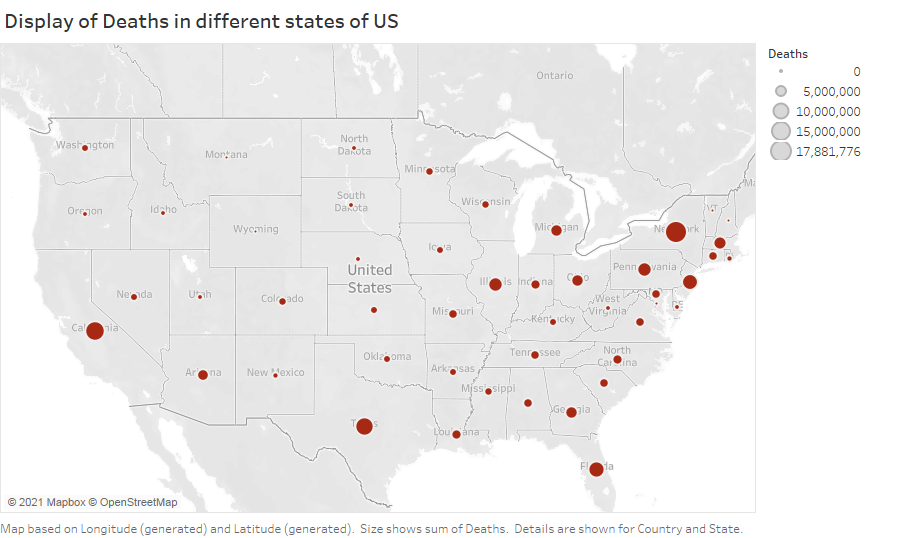
Line graph helps us understand the key insights of the cases confirmed in the given timeframe, information like when was the spike and decrease in cases can be fetched without investigating the actual numbers.

Dashboard for Confirmed cases



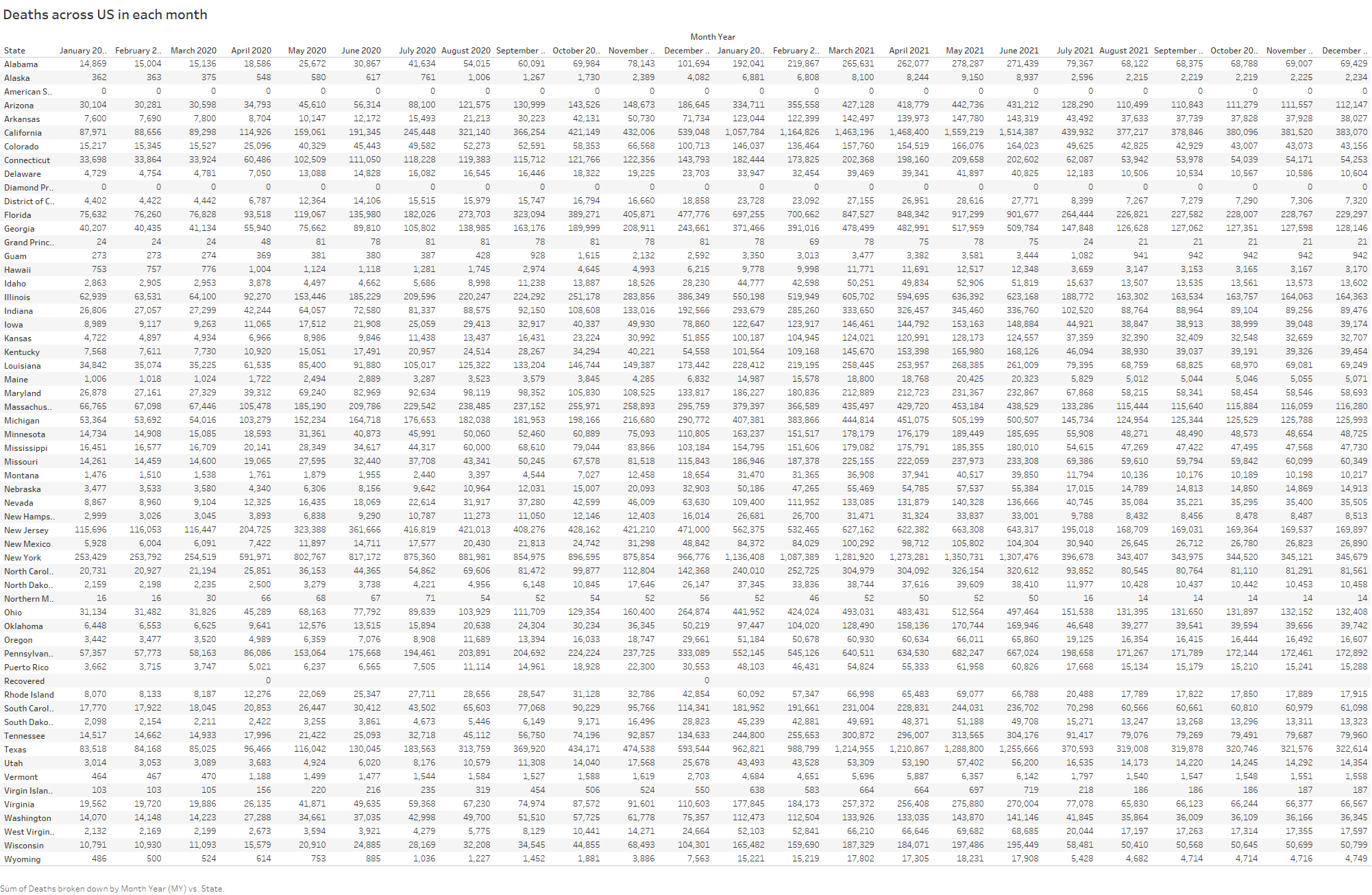
Dashboard helps us to consolidate many reports into a single frame, providing multiple aspect of information at a time. Dashboards can provide us quick information hazel free, in the above dashboard we can fetch any can kind of information with respect to the confirmed cases for any state in the US.

*Death cases in US*

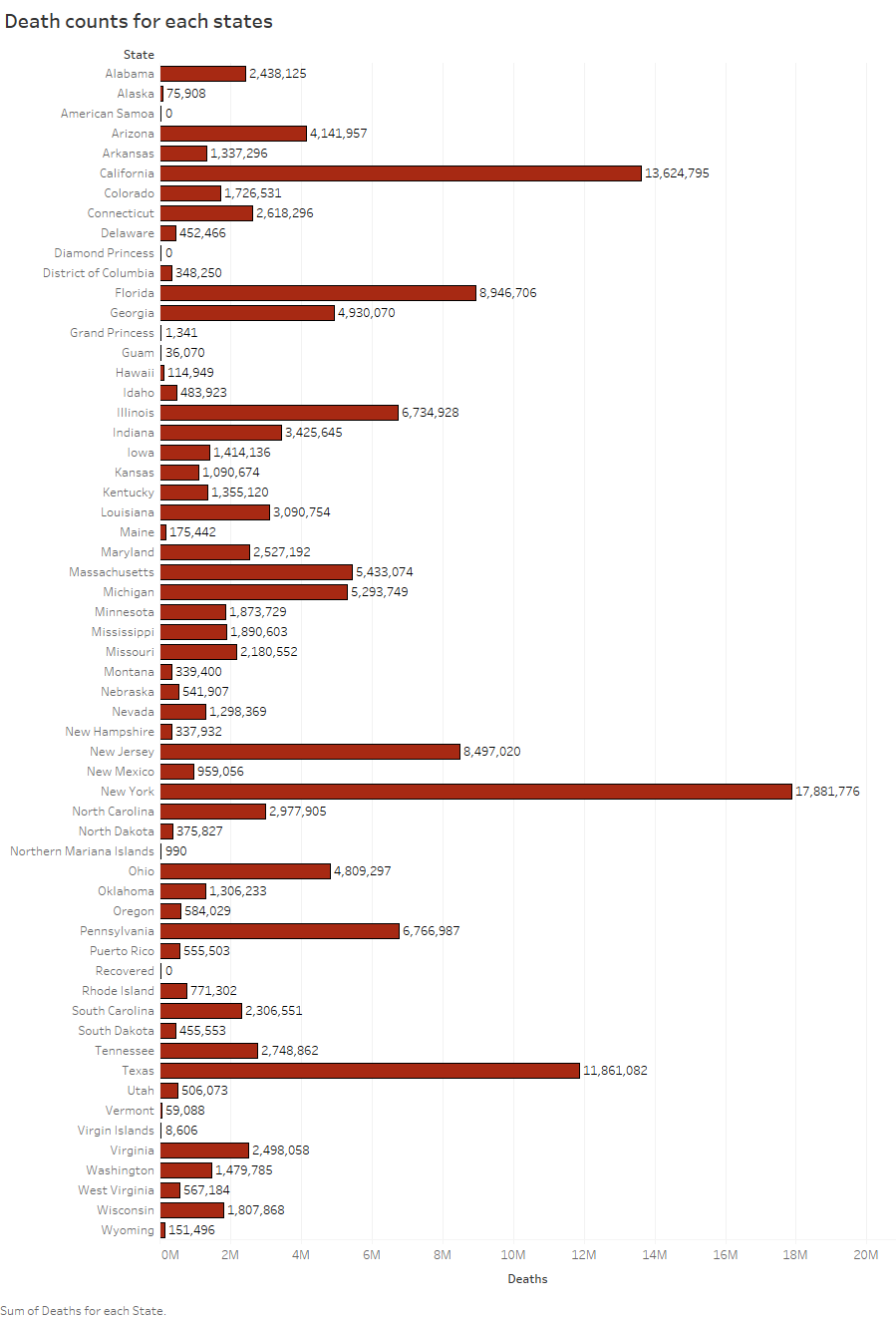


The above geographical distribution of deaths in US displays the severity of COVID-19 virus. The larger the size of the circle, larger is the intensity.

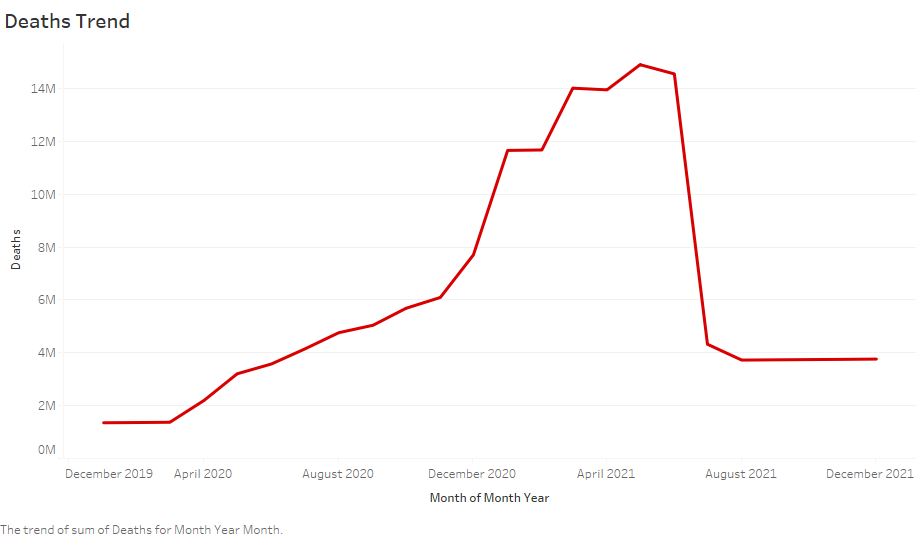
Statistical information about the death cases in all the states of the US for each month starting from January 2020 to July 2021



The above tabular representation of the data for the total death cases for each state in a month, provides complete information in a single report. Having a look at this report can acquire knowledge about the number of deaths confirmed in the US, in addition this report provides specific information with states and timeframe.

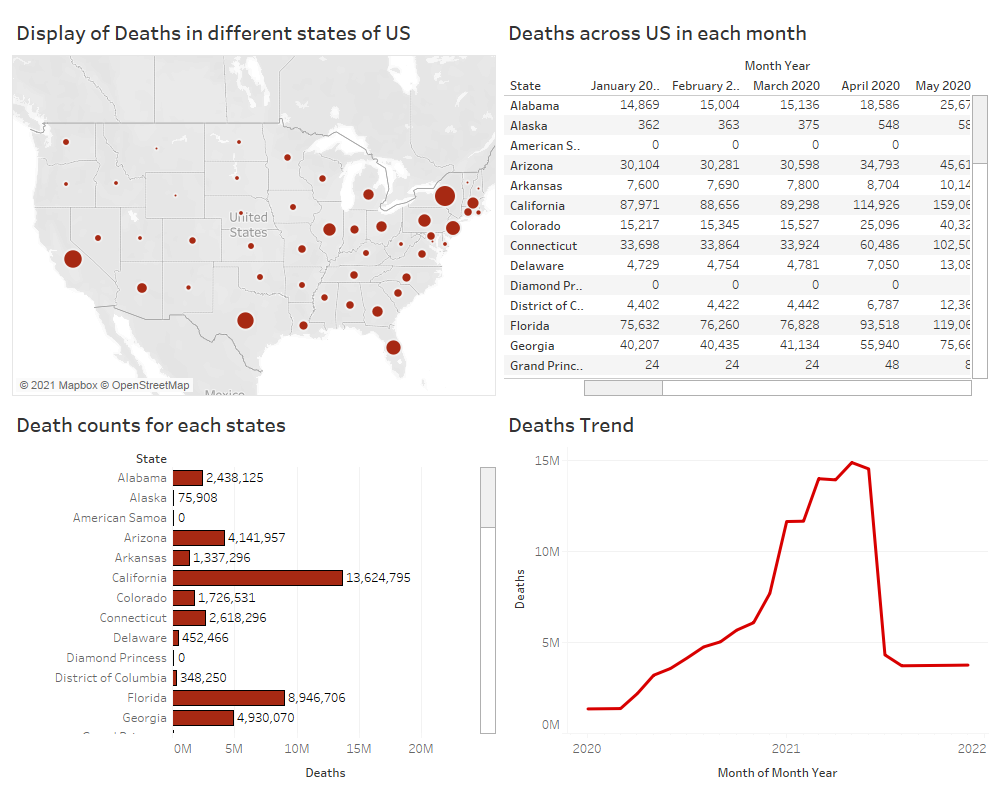
Bar Plot representation of the death cases in US

Line plot representing the trend of death cases over the timeframe



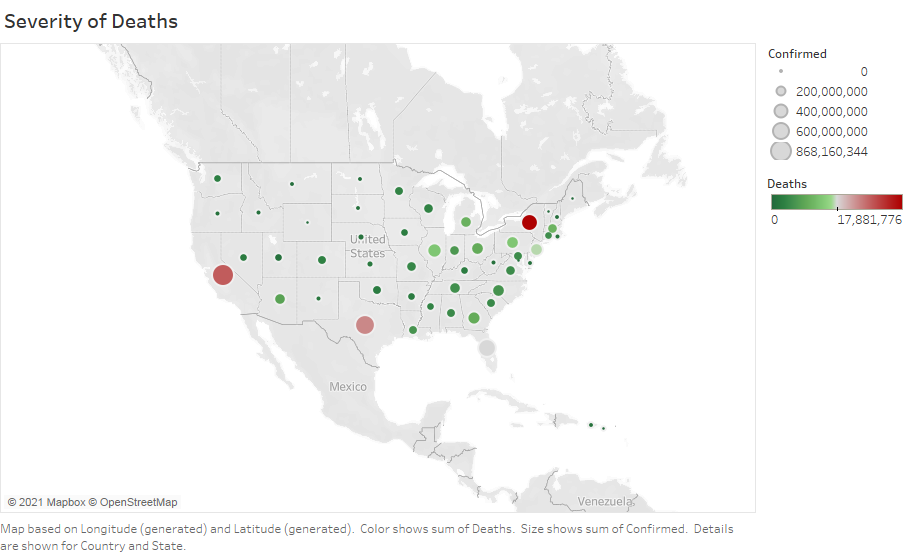
The above line graph displays the information of the rise and fall of the COVID-19 deaths in US.

Dashboard for Death cases



The above dashboard providing wide range of information for the death cases in the US.

Severity of Deaths in the US



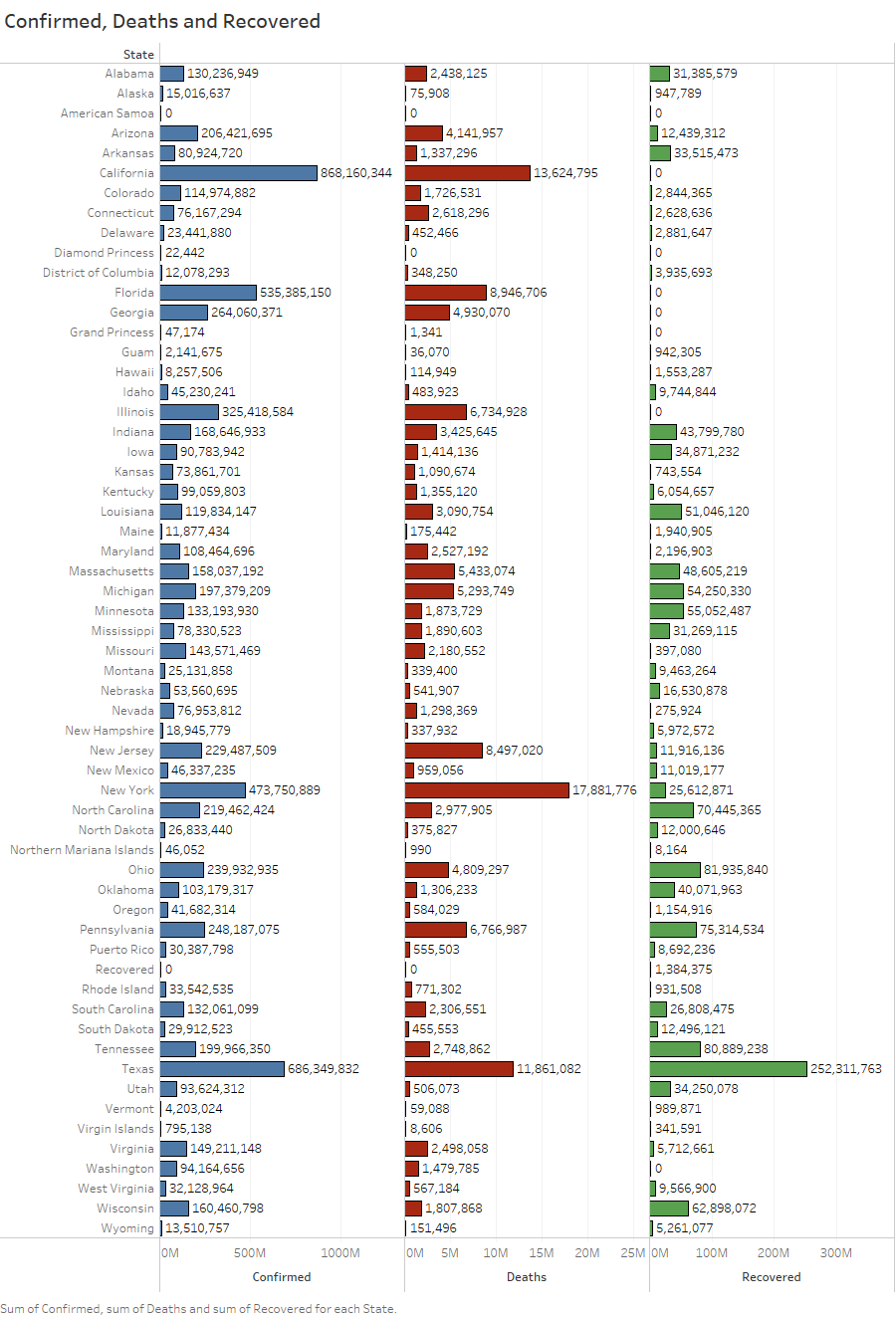
The above geographical distribution of the deaths in US displays the severity of COVID-19 virus. The darker and larger the size of the circle, larger is the intensity.

Quick statistics of COVID-19 dataset

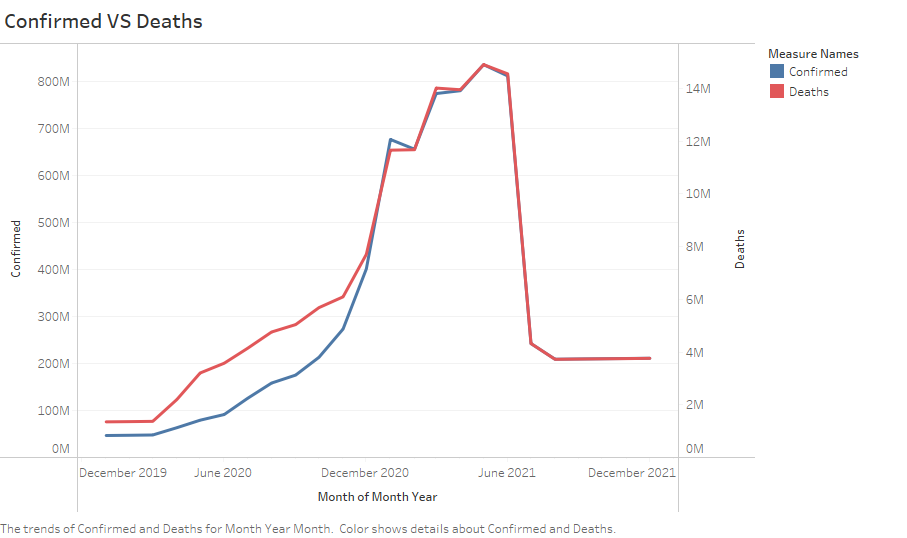


The above table displays the information of confirmed, deaths and recovered cases for each month in the US.

Bar plot for Confirmed, Death and Recovered cases



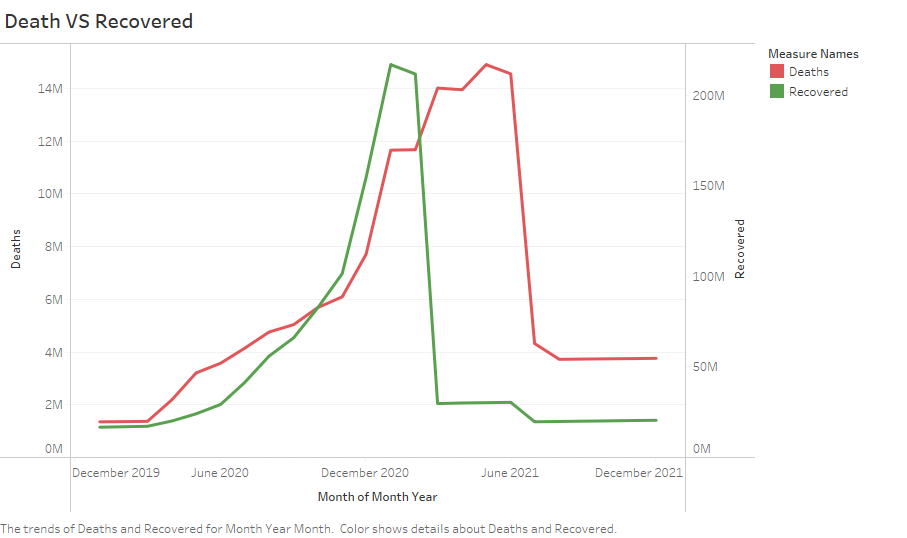
Confirmed VS Deaths trend



The above line graph displays the relation between the confirmed and death cases. We can see that as the confirmed cases increased the death cases also increase.

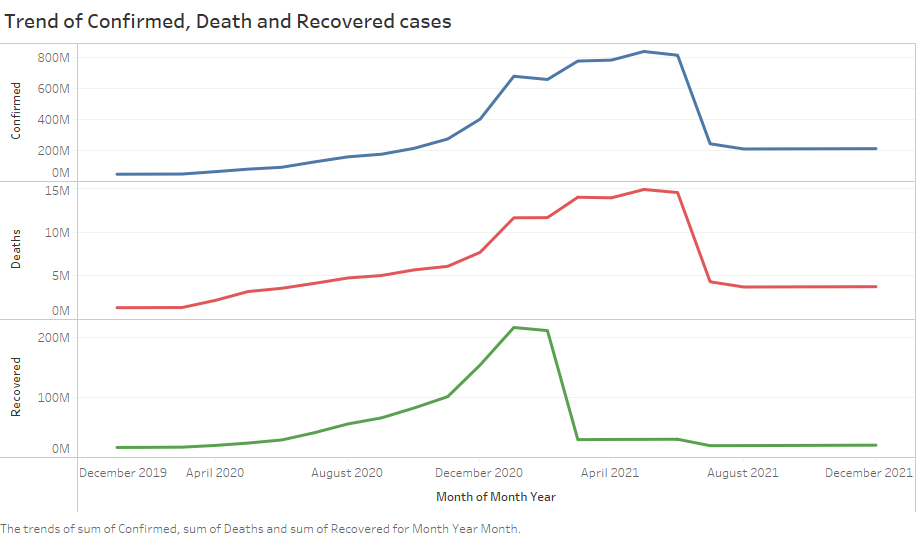
NOTE- The Y axis intervals are different for confirmed and deaths, confirmed cases intervals are displayed on the left, and deaths on the right.

Death VS Recovered

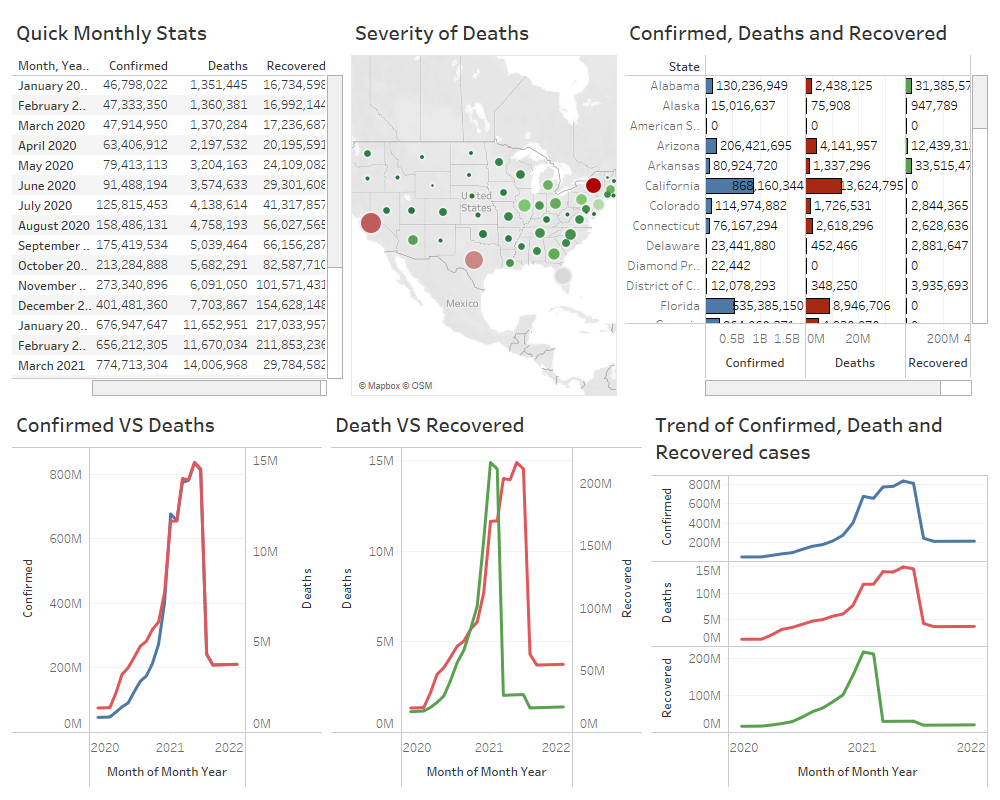


NOTE- The Y axis intervals are not the same for both the measures.

Trends of different measures



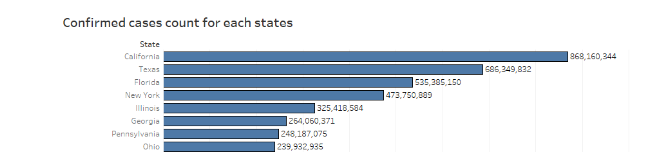
Overall insights of the COVID-19 dataset



The above dashboard provides the overall insights and the behavior of the dataset.

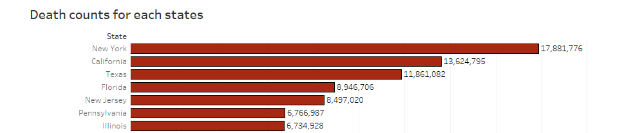
### Business Questions

Which state was the most affected from covid-19?



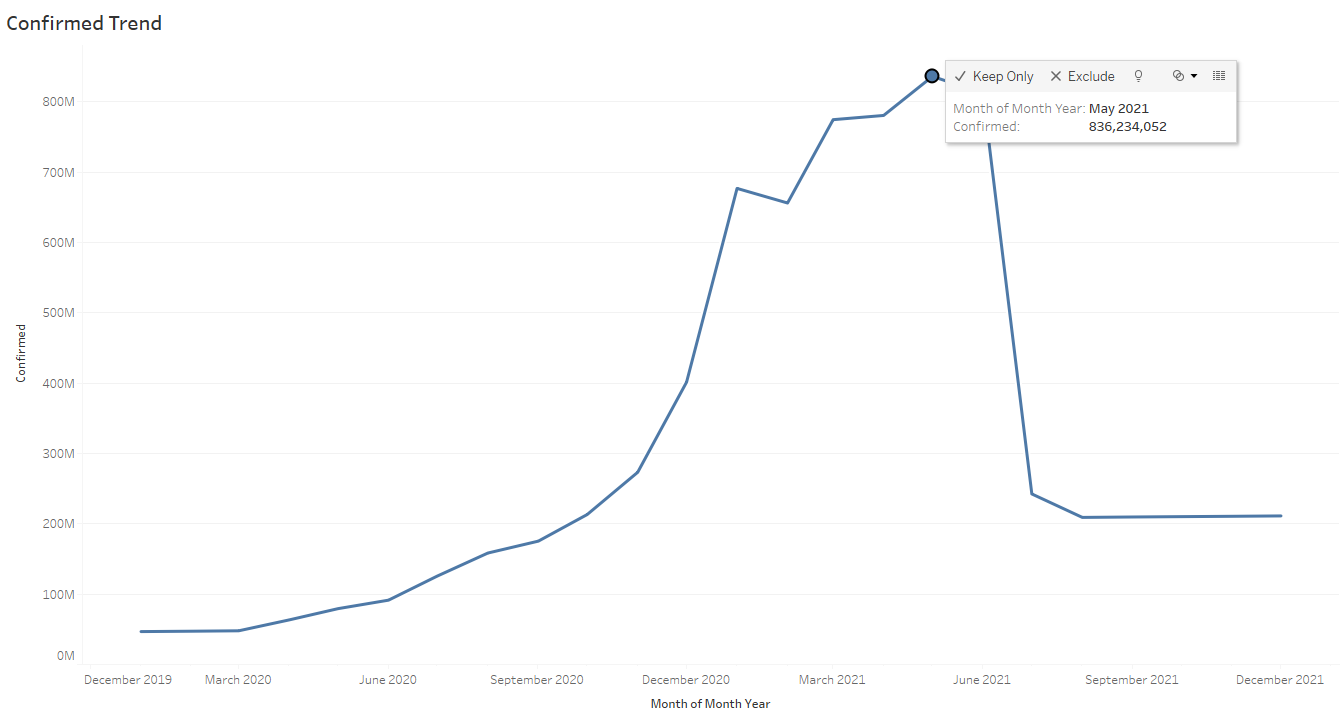
As displayed in the above visual, California state is the most affected from COVID-19

Which state has the highest deaths from covid-19?



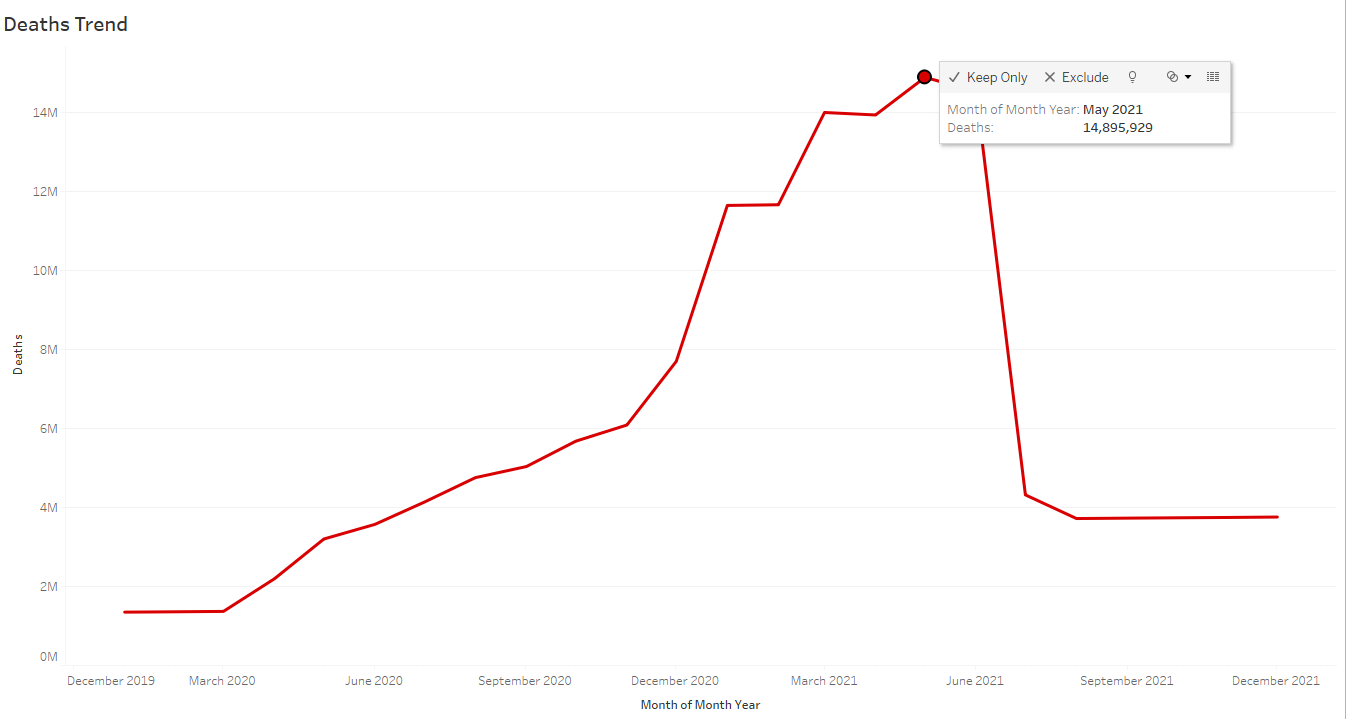
As indicated in the above visual, New York is the state with most deaths.

When did the COVID-19 cases had the highest peak?



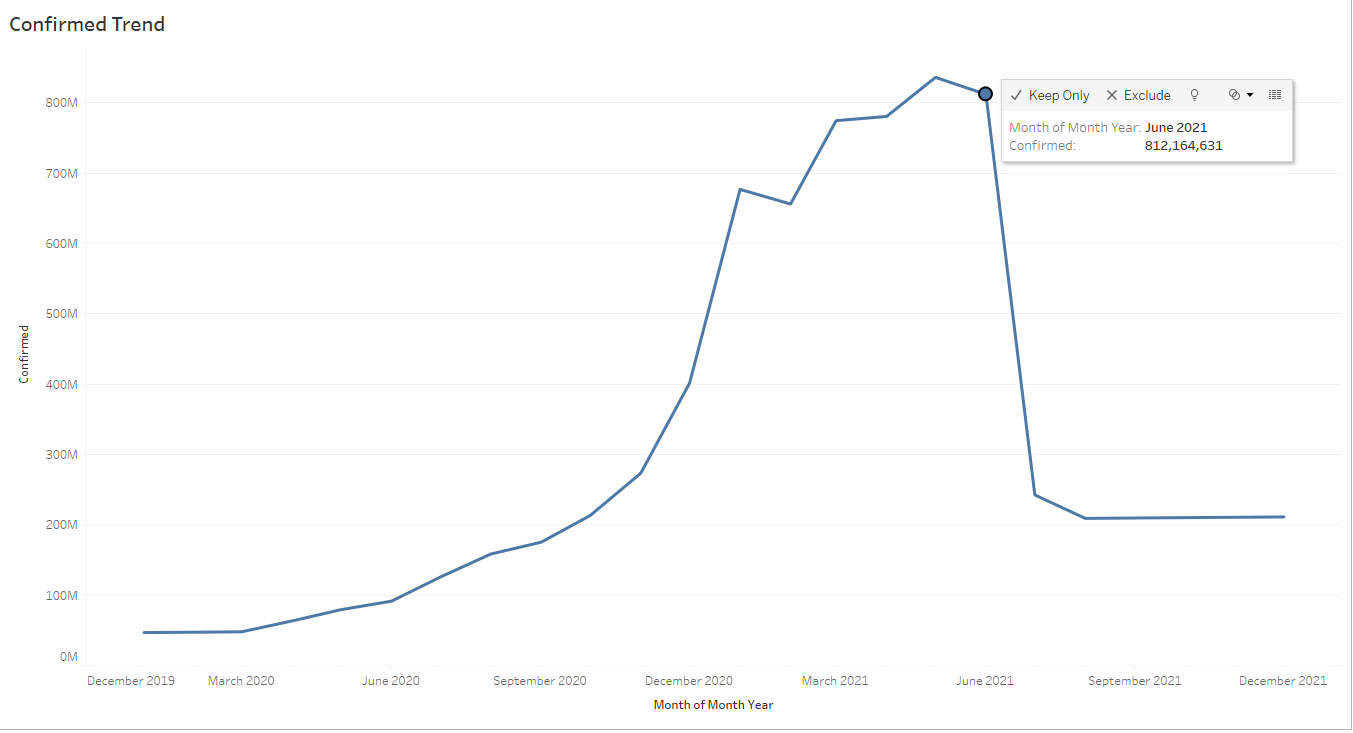
May 2021 was the month which had the highest peak of COVID-19 cases, with 836,234,052 cases.

When did the COVID-19 cases had the highest death rate?



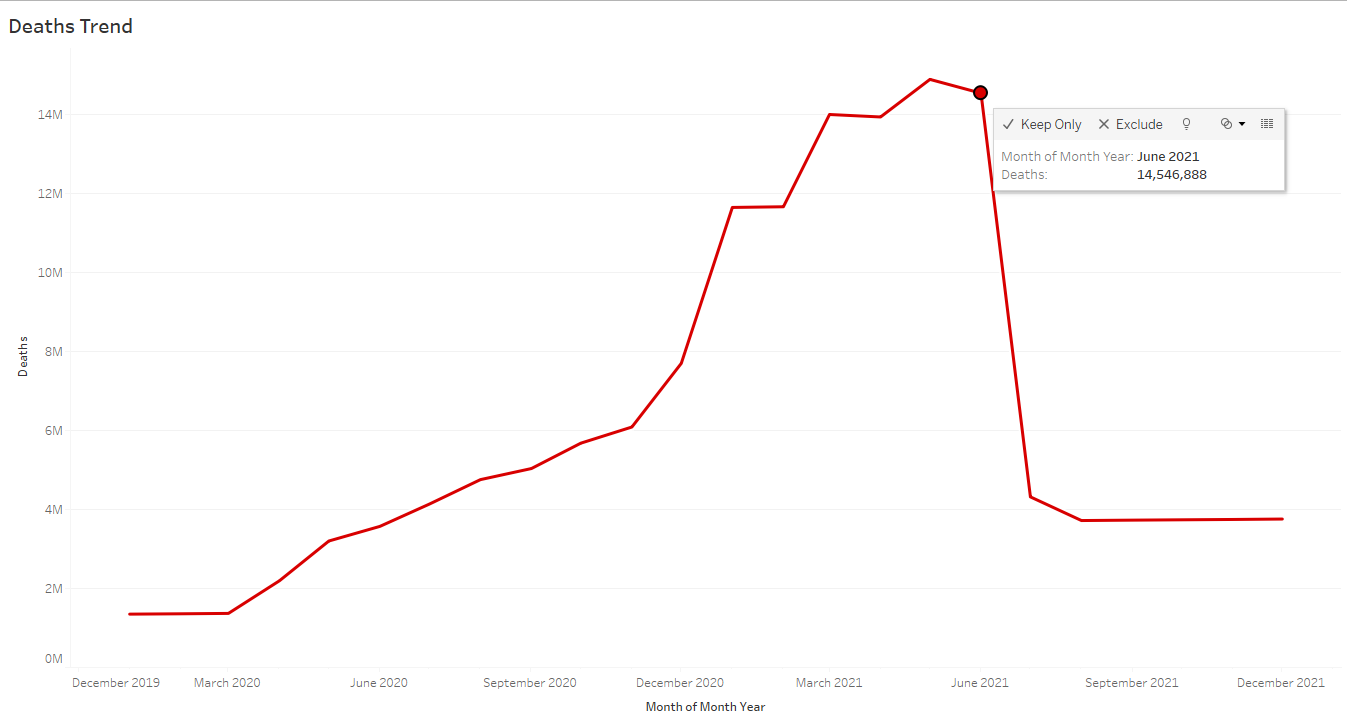
May 2021 was the month with highest death cases of COVID-19, with deaths of 14,895,929.

When did the COVID-19 cases started to drop?



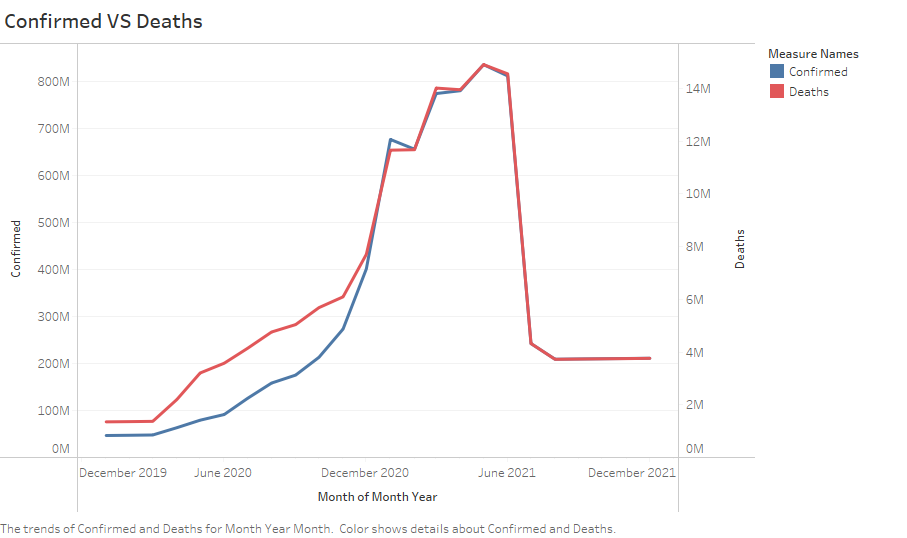
During the month June 2021, there is a huge drop of COVID-19 cases.

When did the death rate started to drop?



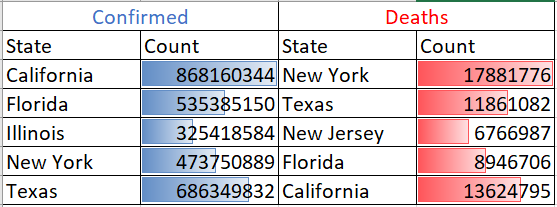
During the month of June 2021, the death rate dropped drastically.

With the EDA and visualization performed, did you observe any pattern in the dataset?



In the above trend, we can see there is a direct relationship between the confirmed and deaths column, as the number of confirmed cases increased, the death counts also increased.

Top five states confirmed and deaths statistics in the US?



Did you observe any unusual behavior of the dataset?

Yes, there are unusual behaviors, During the month of June 2021 the number of confirmed and death cases dropped thoroughly. The drop was caused by the massive vaccination drives conducted by the national governments to the general public to tackle the virus.

The number of confirmed and death cases had a linear raise until October 2020, however there is an immediate uplift of the cases until April 2021. This was due to, the US government had enforced strict lockdown all over the country until September 2020, after the relaxation of the rules the virus spread was escalated.

Why do you think EDA and visualization is important?

After performing all the above analysis using EDA and visualization, both helps us to provide accurate insights about the data before making any assumptions. EDA and visualization provide us a better understanding of the variables and the relationships between them, it helps to understand the data better to measure its impact on the business and communicates the insight visually to internal and external audiences. By performing EDA and visualization on our dataset provides precise information and helped us to answer all our business question effortlessly. EDA and visualization are two important aspect of data analytics when trying to get insights on a dataset.

*All the mentioned above screenshots and visualization is performed on Tableau the required code/file (.twb) has been uploaded on GitHub.*

Thank you,

Data Wizard Team