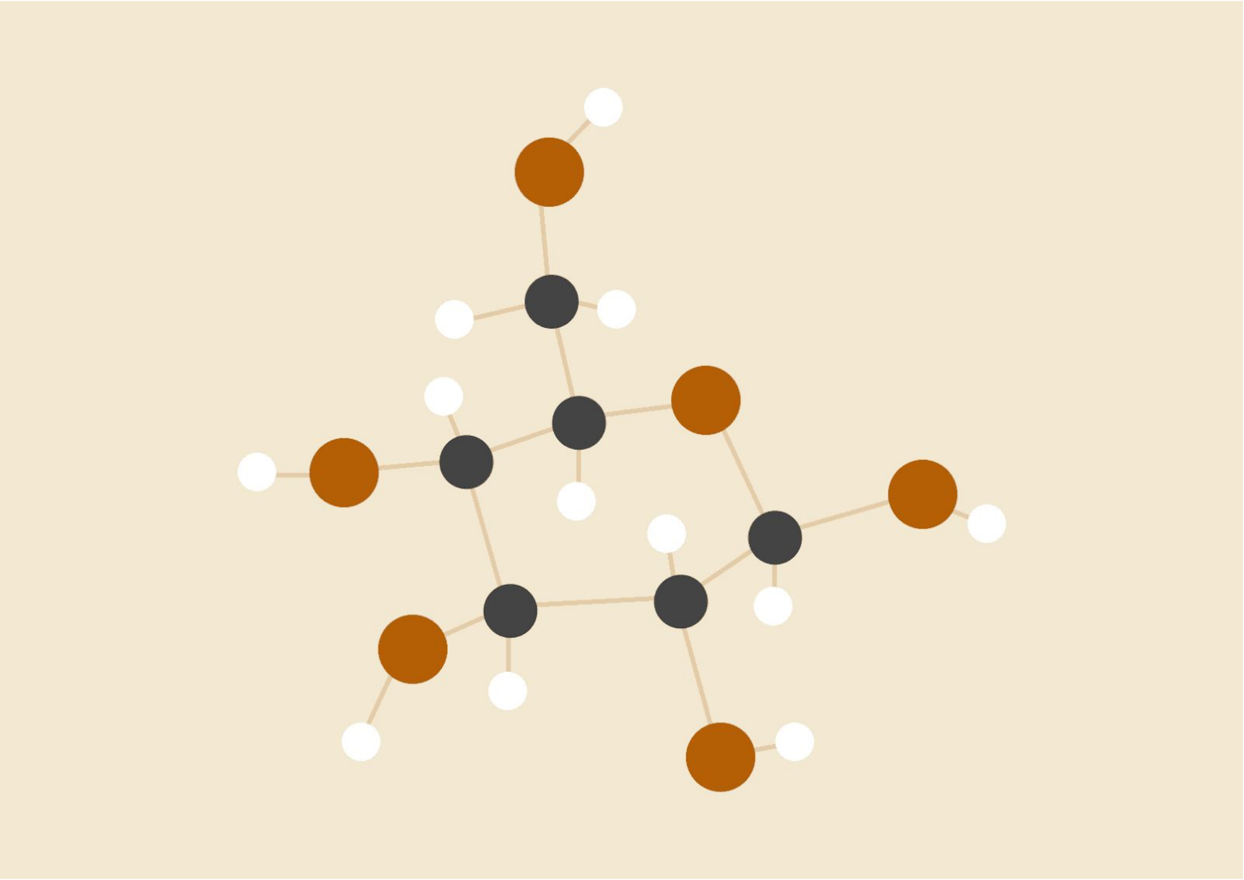
COVID 19 DATA ANALYSIS Final Report



VEDANT MAJMUDAR,

SANKET SHAH,

TANMAY SARIN

CS 418 Data Science Final Project

Introduction

The project attempted to analyze COVID 19 data. The data was acquired from the GitHub repository of John Hopkins University. The project focused on using the time-series data, which stored the data showing how many people contracted the virus, how many resulted in deaths, and how many were able to recover. It specifically focused on providing a global view of how the different countries of the world tackled this unprecedented situation and then gave a closed look to the United States. The later part of the project takes a closer look at the management of each state in the US. It provided a visual representation of the severity of the situation on a global scale. It focused on analyzing how a country tackled its own situation instead of providing a comparison. The results show that countries like New Zealand have been able to conquer COVID-19, and countries like Australia are handling the situation remarkably. However, countries like the United States, in particular, show no control over the disease as the reported confirmed cases and deaths seem to increase exponentially every day. The project also tried to point out how various states in the US handled the situation of COVID-19. The goal was to point out the states which needed better management. To achieve this goal, the team used linear regression prediction models of various states and was able to determine that states like Indiana needed better management, and states like Maryland and New York have shown remarkable improvement in their management recently. With these goals, we will be able to understand the growth of COVID 19 across the globe and have a better understanding of how much the virus has already spread.

Body

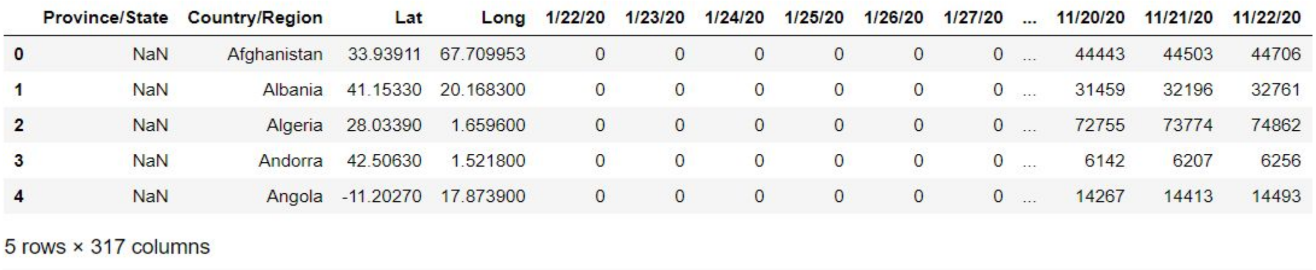
Data

Data Source

We got our Data from the John Hopkins University GitHub repository. We used the time series data, which is updated daily with details about the number of confirmed, recovered, or deaths that occurred due to the pandemic. The data is collected from various individual sources and gets compiled together much like the World Health Organization (WHO) database. Thus it is reliable data, which is very important for any data science project. The repository can be accessed by clicking here.

Data Description

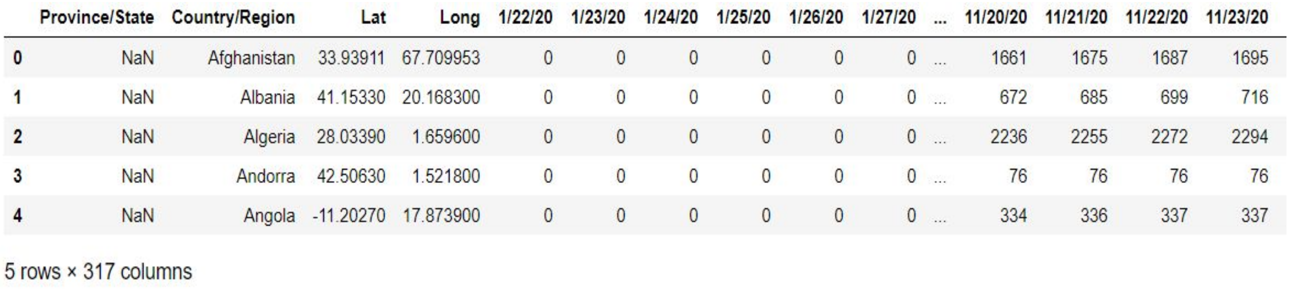
We have two different sets of data. One of them is about the countries. This data consists of a daily number of the confirmed, from January 22nd, 2020, and contains information of about 272 different countries and provinces. The other dataset contains information very similar to the previous one, but about the different States of US. The COVID positive date starters from January 21st, 2020. This also consists of the name of the city but we are not making use of it.



	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	11/20/20	11/21/20	11/22/20
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	44443	44503	44706
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	31459	32196	32761
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	72755	73774	74862
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	6142	6207	6256
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	14267	14413	14493

5 rows × 317 columns

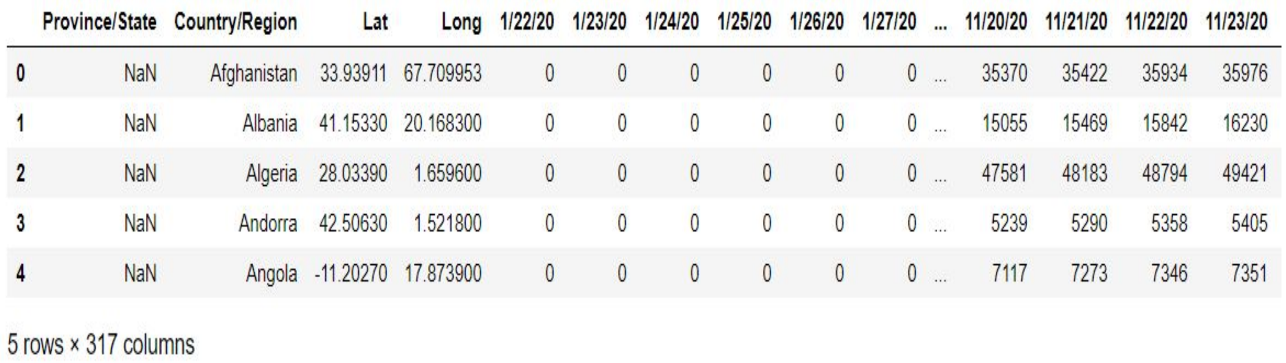
Confirmed Cases data snapshot by countries



	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	11/20/20	11/21/20	11/22/20	11/23/20
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	1661	1675	1687	1695
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	672	685	699	716
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	2236	2255	2272	2294
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	76	76	76	76
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	334	336	337	337

5 rows × 317 columns

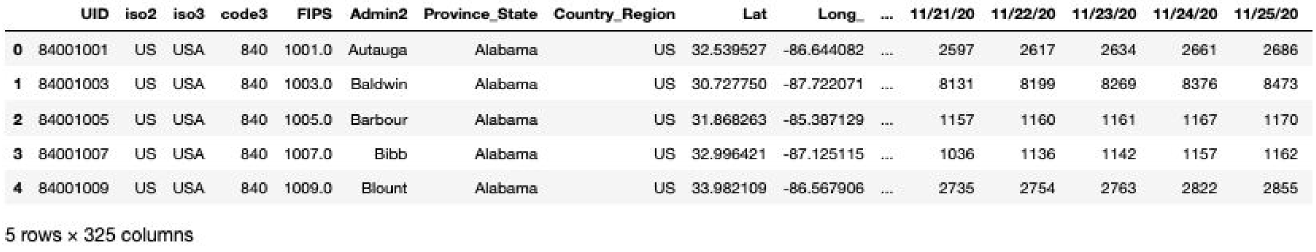
Death data snapshots by countries



	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	11/20/20	11/21/20	11/22/20	11/23/20
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	35370	35422	35934	35976
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	15055	15469	15842	16230
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	47581	48183	48794	49421
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	5239	5290	5358	5405
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	7117	7273	7346	7351

5 rows × 317 columns

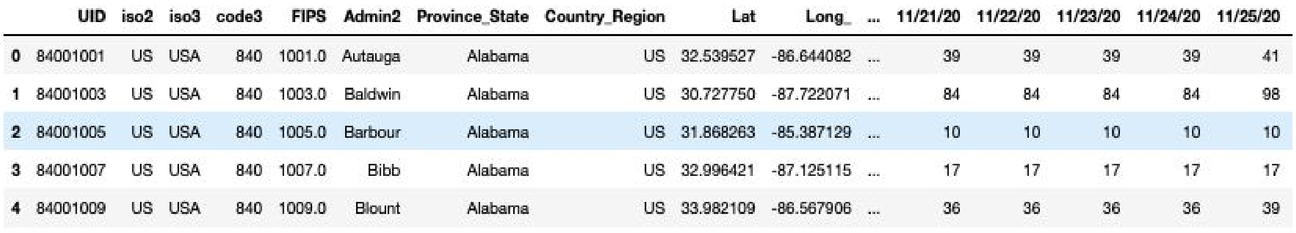
Recovered Cases by countries



	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	...	11/21/20	11/22/20	11/23/20	11/24/20	11/25/20
0	84001001	US	USA	840	1001.0	Autauga	Alabama	US	32.539527	-86.644082	...	2597	2617	2634	2661	2686
1	84001003	US	USA	840	1003.0	Baldwin	Alabama	US	30.727750	-87.722071	...	8131	8199	8269	8376	8473
2	84001005	US	USA	840	1005.0	Barbour	Alabama	US	31.868263	-85.387129	...	1157	1160	1161	1167	1170
3	84001007	US	USA	840	1007.0	Bibb	Alabama	US	32.996421	-87.125115	...	1036	1136	1142	1157	1162
4	84001009	US	USA	840	1009.0	Blount	Alabama	US	33.982109	-86.567906	...	2735	2754	2763	2822	2855

5 rows × 325 columns

Confirmed Cases based on the United States



	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	...	11/21/20	11/22/20	11/23/20	11/24/20	11/25/20
0	84001001	US	USA	840	1001.0	Autauga	Alabama	US	32.539527	-86.644082	...	39	39	39	39	41
1	84001003	US	USA	840	1003.0	Baldwin	Alabama	US	30.727750	-87.722071	...	84	84	84	84	98
2	84001005	US	USA	840	1005.0	Barbour	Alabama	US	31.868263	-85.387129	...	10	10	10	10	10
3	84001007	US	USA	840	1007.0	Bibb	Alabama	US	32.996421	-87.125115	...	17	17	17	17	17
4	84001009	US	USA	840	1009.0	Blount	Alabama	US	33.982109	-86.567906	...	36	36	36	36	39

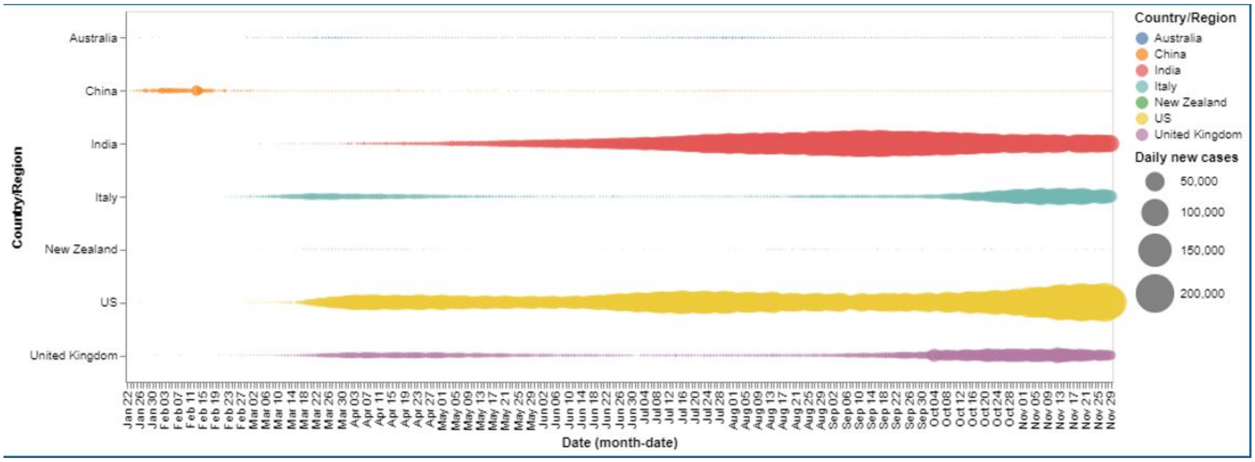
Deaths based on the United States

Analysis

Countries:

The country time series data acquired from the John Hopkins Github repository, contained some serious outliers like the data of cases from the ships that were out in the ocean at the time when pandemic affected the world. It also contained data of provinces of some of the countries which were unable to provide a concise estimate initially. Some of the data also contained empty values for either confirmed, recovered or deaths. All these had to be handled in order to make the data workable. The data was thus cleaned. The data of the ships was removed from the project. An approximate estimation was generated and the data of the provinces of various countries was reduced to a single data entry row. All the empty data values were handled.

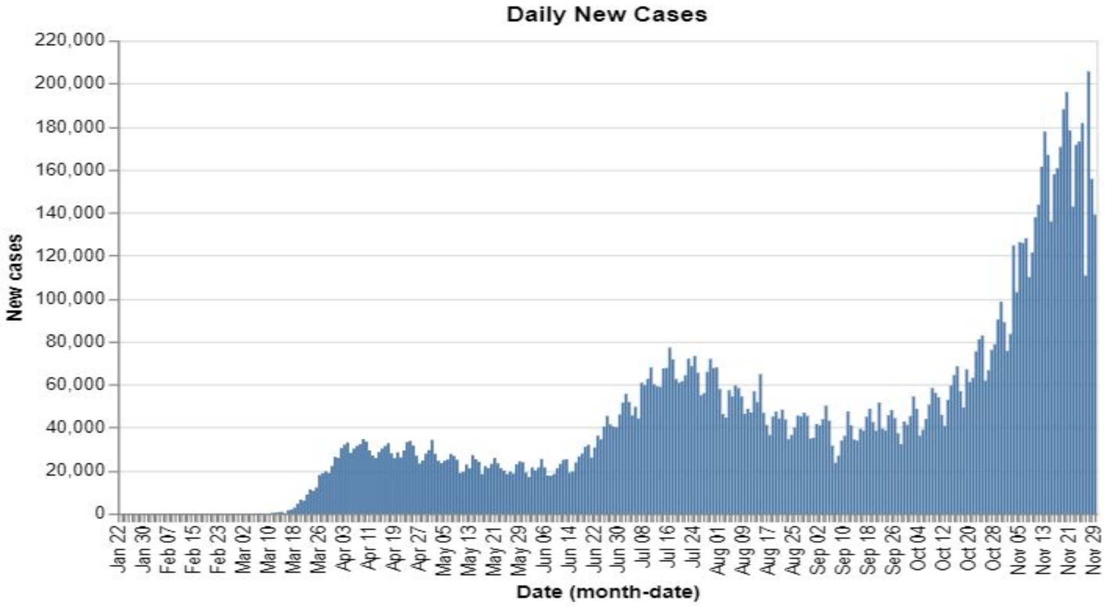
The project provides a visual representation of the daily number of reported cases in various countries, some of the major ones being New Zealand, India, United States, United Kingdom, Italy and Australia. From the visualization, it was evident that New Zealand by far, has been the most successful country in the world to experfully handle the situation of COVID-19, as the daily reported cases decreased to a pleasing 0, with Australia being at a close second place. On the other hand, countries like the United States had seen a huge explosion in the daily reported cases, with numbers reaching a staggering high of more than 200,000 daily cases. The visualization also showed that countries like India, which showed remarkable control during the beginning of 2020 were able to tactfully handle the situation initially, but this control broke down in the middle. However, with the decline in their everyday cases, it is evident that the country is making a comeback and is beginning to handle the COVID-19 situation in a better way.



United States:

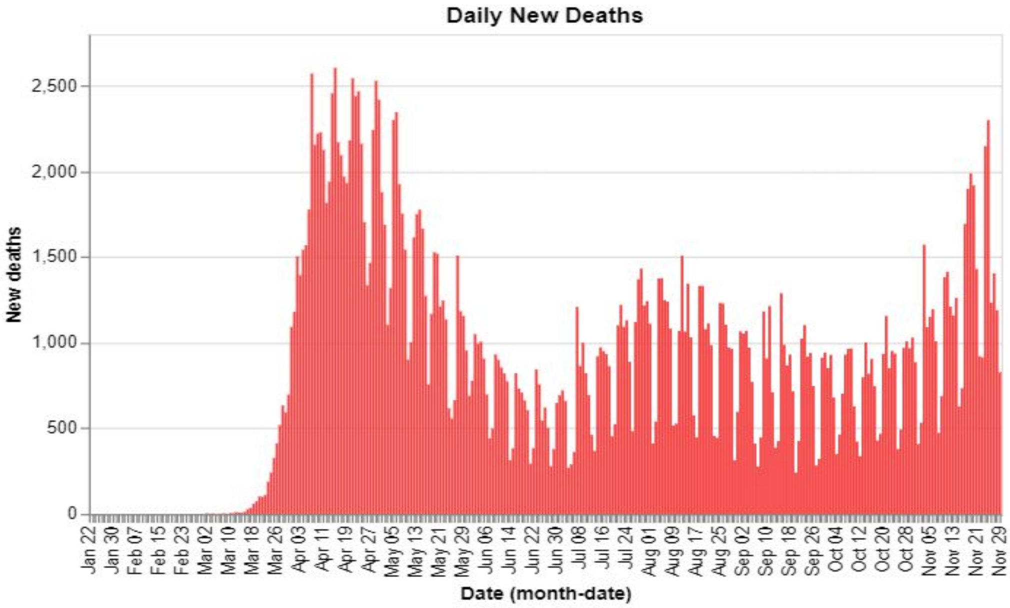
In regards to the United States, the data again had some outliers, like having some unknown values, empty values, negative values, etc which needed to be cleaned in order to work with the data.

After cleaning the data, the next step was to find the confirmed cases on the daily basis that were reported in the United States. It was essential to find the confirmed cases since that gives the idea of how the country is doing in regards to handling the situation and if the country needs to take some majors regarding it or not.

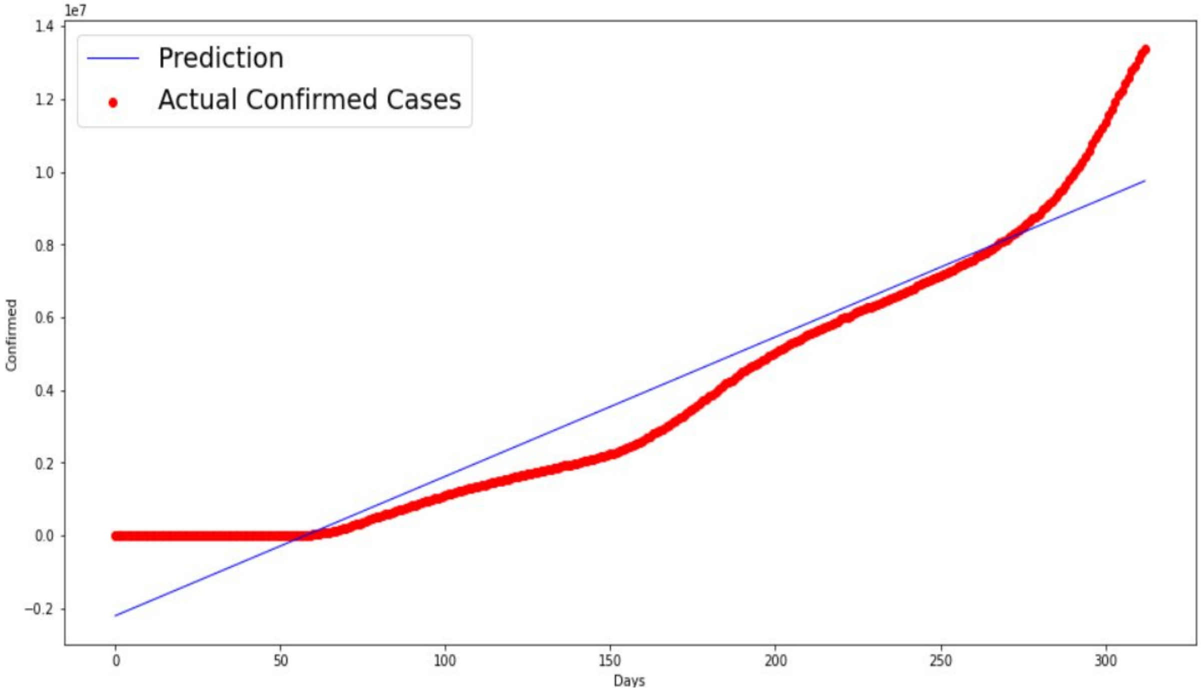


The United States initially didn't have many cases but after a while the cases grew exponentially which shows that the decisions that were taken by the government were bad or didn't work out as they thought. One can argue that the lockdown majors were taken out before they should have. The scenario could have been different.

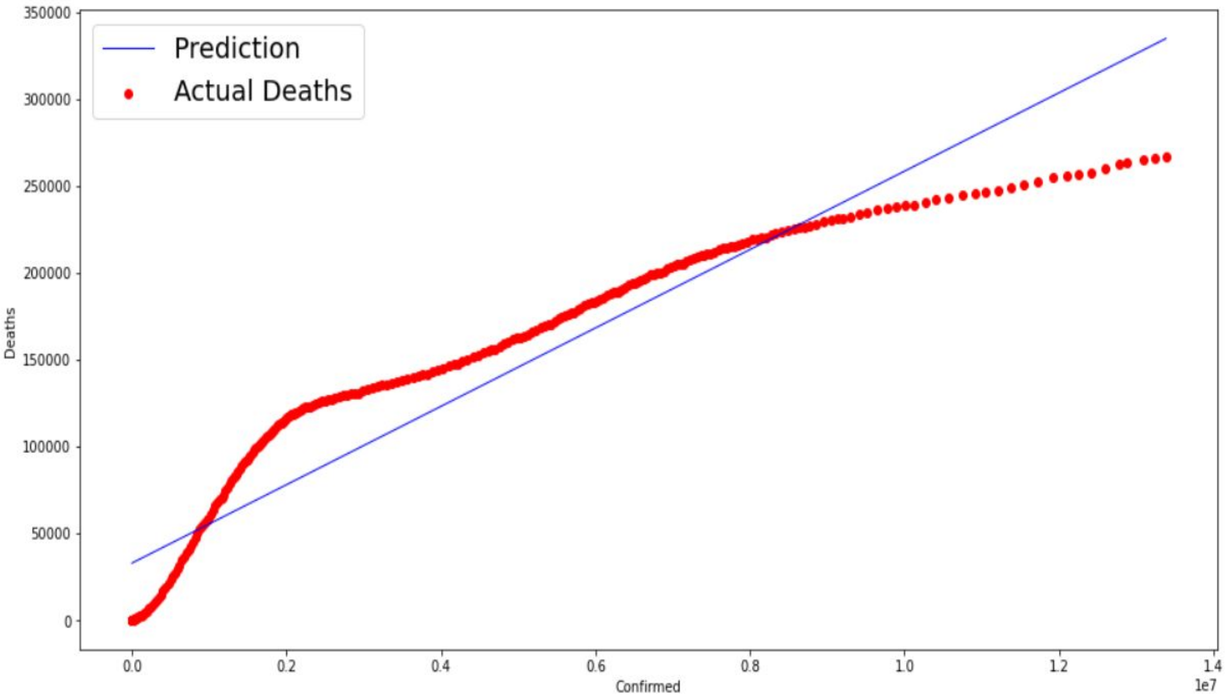
Next comes the finding of the deaths on the daily basis that were reported in the United States. This is another important part since by this data we can tell if the people in the US face COVID-19 more severely than other countries or not. Like for example, the percentage of the deaths based on the confirmed cases are the lowest in India unlike the United States.



The Data for the United States is then separated from all the other data to perform the linear regression modeling on it. There are two different kinds of modeling being performed. The first one consists of the number of days vs the number of confirmed cases, and the second one consist of the number of deaths vs number of confirmed cases The graphs are as below:



The graph above shows the linear regression modeling. The blue line shows the prediction for the confirmed cases whereas the red line shows the actual data. According to the graph, the cases are growing slowly, for the initial 260-270 days, after which they grow exponentially. The x-axis consists of the number of days, whereas the y-axis consists of the number of confirmed cases on that day, starting from january 22nd. It can be seen that there are no signs of number case decreasing, which implies that the management for tackling COVID-19 has not been reliable.

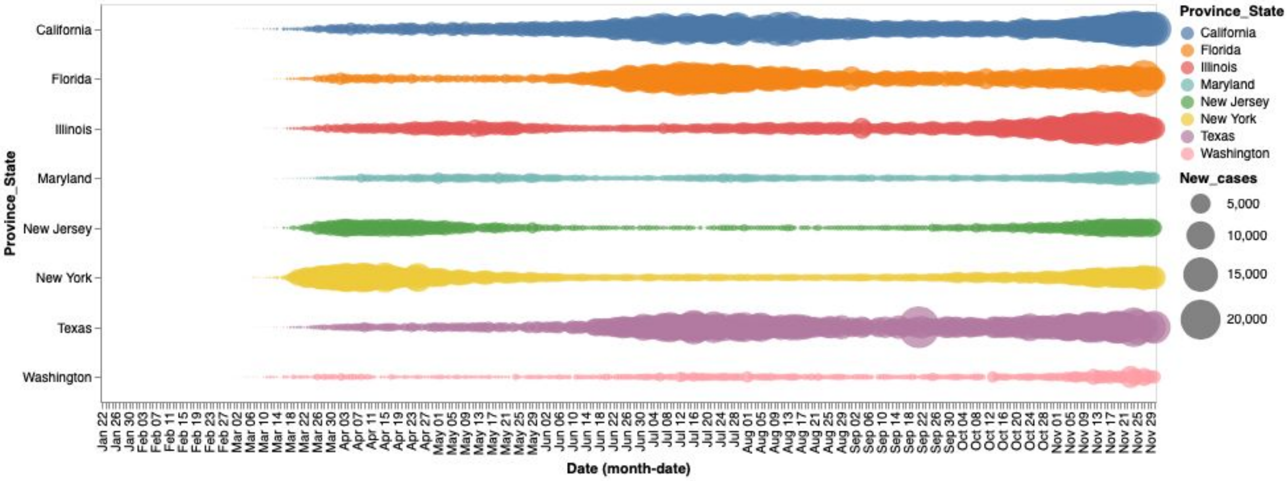


The graph above shows the linear regression modeling. The blue line shows the prediction for the confirmed deaths, whereas the red line shows the actual data. According to the graph, the number of deaths were higher in the beginning, but later on, when the number of cases was higher the number of deaths was under the predicted deaths. The x-axis consists of the number of confirmed cases, whereas the y-axis consists of the number of confirmed deaths due to COVID 19, starting from january 22nd. The decline in the deaths implies that although the daily cases continued to increase, the medical prowess in tackling the situation also increased, which resulted in considerably less deaths than the expected prediction.

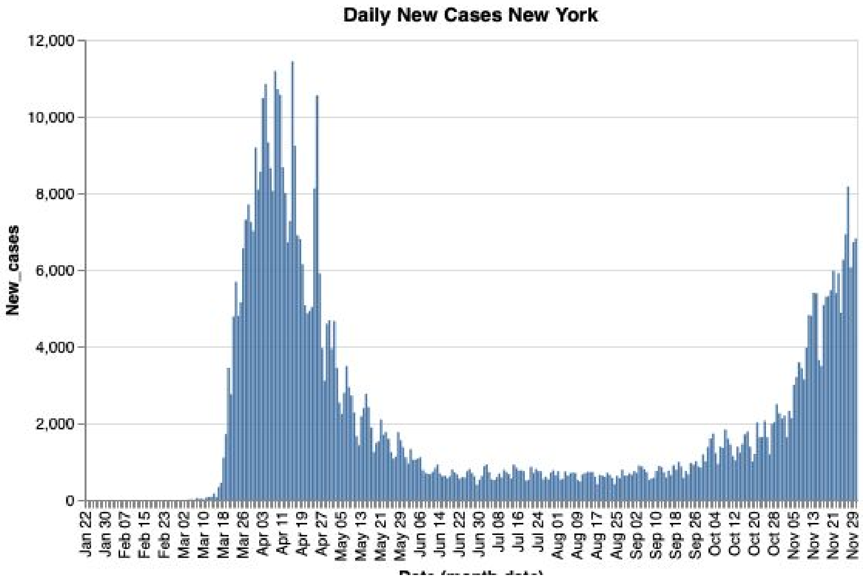
States within United States:

After focusing on the United States in general, the next task was to find the severity of the situation among the states within the United States. First step again was to clean the data so that the finding can be accurate.

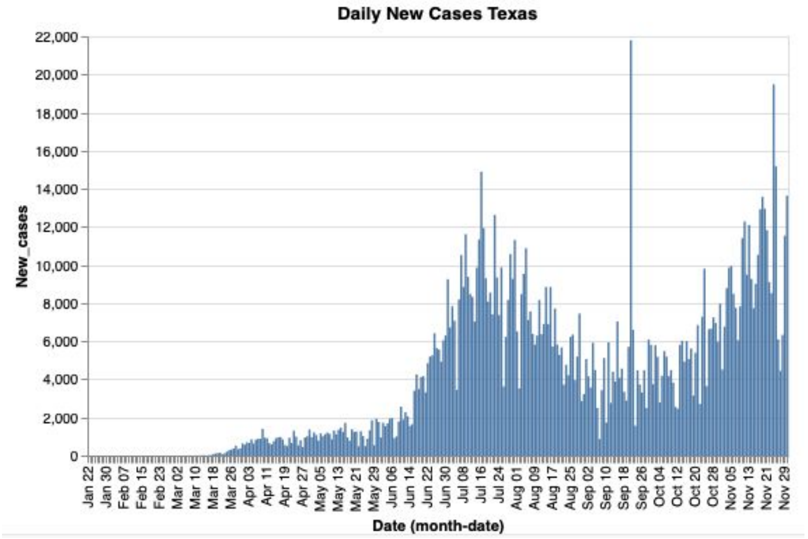
After cleaning the data, the task was again to find the daily new cases and the daily deaths of different US states. Getting this data gives more insight into the states and can know if the state needs to take more precautions or not.



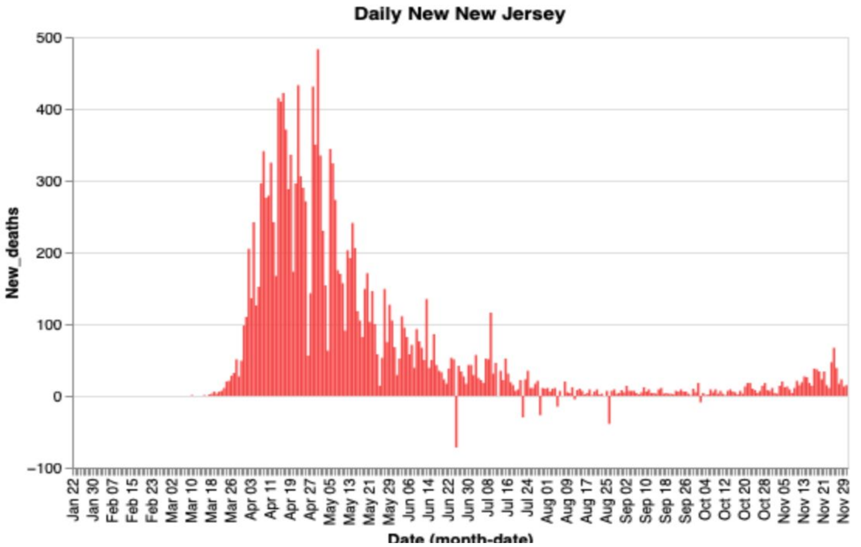
The team worked on comparing the cases and deaths between all the states and found out that some cases are doing better than others. And some states need to severely take precautions to bring the severity down.



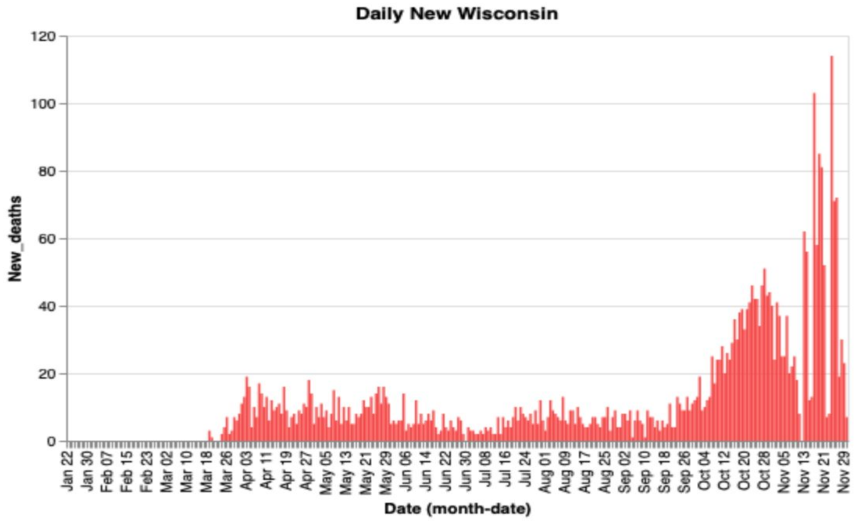
Cases in New York were very high in the initial days but the state understood that and took the necessary precautions that needed to be taken which brought the cases down.



Cases in Texas were under control in the initial days but the state didn’t take the matter seriously which eventually brought the cases up by a huge number.

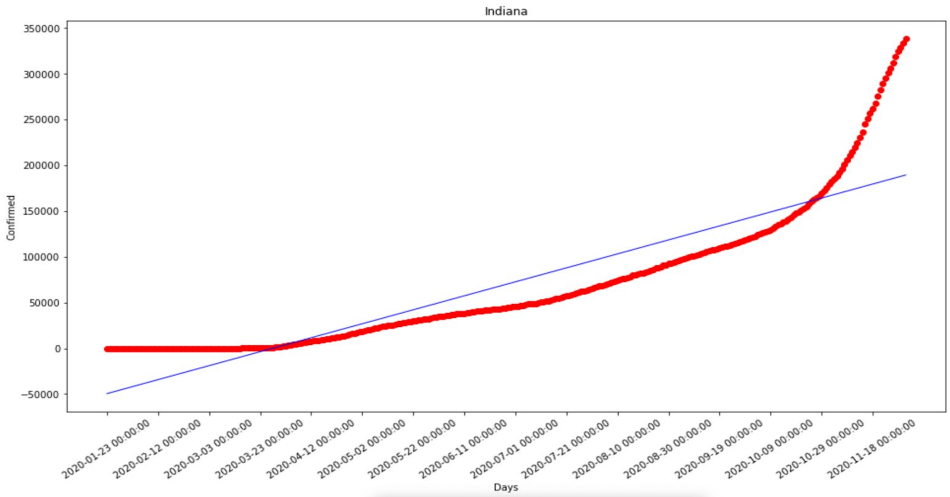


Daily deaths in New Jersey

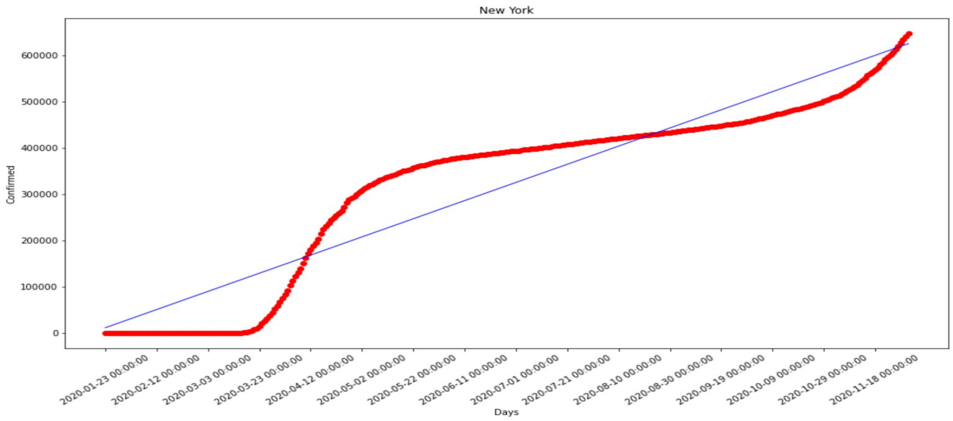


Daily deaths in Wisconsin

As performed above we are also performing the predictions modeling for all the US states, some of the unique ones are shown and explained below.

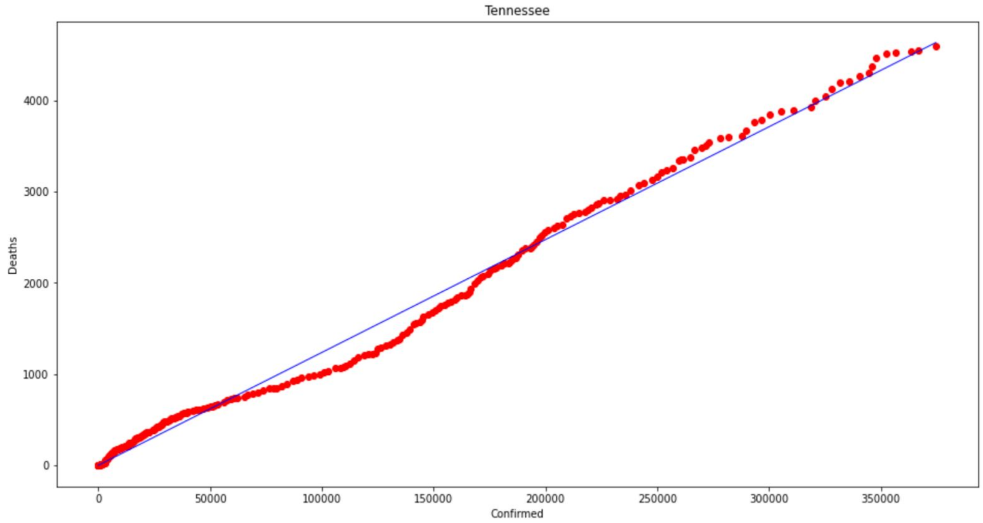


Indiana Confirmed Cases

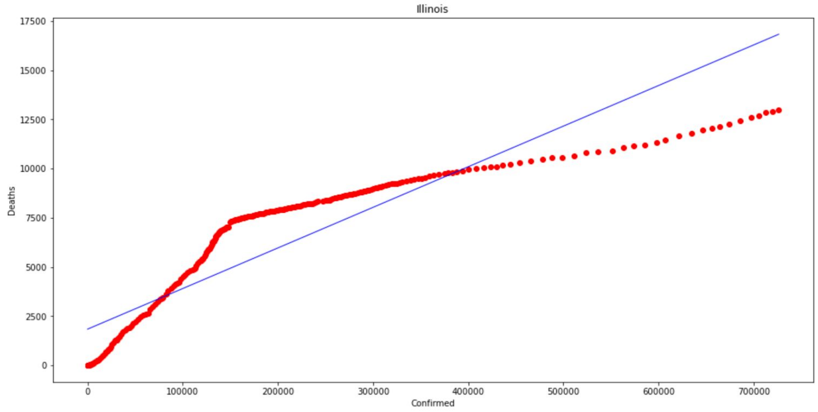


New York Confirmed Cases

Above are two linear regression models for the state of New York and Indiana. This modeling is for the number of confirmed cases per day vs the number of days. The blue line shows the prediction, whereas the red line shows the actual data. From the comparison we can see that NY had a caviar outbreak of COVID-19 around june, which was brought under control by septamber . Whereas in Indiana it had it well and controlled until September and the outbreak started getting out of control. The factors like the more international traveling in the state of NY then in the IN, may have caused this.



Tennessee total deaths



Illinois total deaths

Above are two linear regression models for the state of Tennessee and Illinois. This modeling is for the number of confirmed deaths per day vs the number of confirmed cases per day . The blue line shows the prediction, whereas the red line shows the actual data. From the comparison we can see that TN has more number of deaths as predicted, whereas in IL the number of deaths have been decreasing drastically. With this, we can determine that the mitigation steps taken by the TN state are not as reliable as taken by the state of IL.

CONCLUSION

To conclude, the team was successfully able to accomplish what we initially planned to do. We were able to get insights into how the countries are tackling the situation. We were also able to point out countries that need better management and determine how effectively a country is tackling the situation. For the United States in General, we were able to generate Linear Regression prediction models, which enables us to gauge the performance of each state in handling the situation. From our result, it was evident that -

- Although we have made a lot of progress in tackling the situation of COVID-19, it is still not enough as the number of new cases rise daily.
- The prediction model shows that the performance of most of the states in the US has been poor.
- Some states like New York have recovered nearing the end of 2020.
- States like Indiana have worsened as the measures implemented by the government have weakened.
- On a global perspective, New Zealand should be declared as a role model for everyone, being the only country to have completely and successfully defeated Coronavirus.