Data Visualization and Forecasting of COVID-19, and Analysis of COVID-19 data with World Happiness Report

by

Sachin Mohan Sujir

A Project Report Submitted
in
Partial Fulfillment of the
Requirements for the Degree of
Master of Science
Supervised by

Prof. Jai Kang

School of Information

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

December 2020

The project “Data Visualization and Forecasting of COVID-19, and Analysis of COVID-19 data with World Happiness Report” by Sachin Mohan Sujir has been examined and approved by the following Examination Committee:

Prof. Jai Kang
Associate Professor
Project Committee Chair

Dr. Michael McQuaid
Senior Lecturer

Abstract

Data Visualization and Forecasting of COVID-19, and Analysis of COVID-19 data with World Happiness Report

Sachin Mohan Sujir

Supervising Professor: Prof. Jai Kang

Since the beginning of this year, the whole world has been facing a deadly disease caused by a coronavirus which is an acute respiratory syndrome coronavirus, said to be originating from Wuhan, China. This project focuses on COVID-19 (Corona Virus Disease 2019) based on the data from John Hopkins University and the data from The COVID Tracking Project. The report is also based on a dataset named World Happiness Report that ranks 155 countries by their happiness levels based on the GDP per capita, social support, healthy life expectancy, etc. The main idea of this project is to visualize COVID-19 data comparing different countries, with the highest number of cases, deaths, and recovered cases, based on their fatality rates, total infections, recovered cases over time, and also specifically comparing different states in the USA with the highest number of positive cases, deaths and recovery rate, and forecasting the number of confirmed cases, the number of deaths for a future date. The lack of prediction caused inadequate hospitalizations to the affected people. Forecasting was done using FbProphet, a time series forecasting tool developed by Facebook and Autoregressive Integrated Moving Average (ARIMA)-which is a form of regression analysis that measures the strength of one dependent variable with respect to other changing variables. The project also includes a discussion of the performance of the forecasting model in regards to COVID-19 data. The effectiveness of the models is evaluated based on the mean absolute error, root mean square error and r-squared metrics. The evaluations show that the ARIMA model is more effective for forecasting the pandemic. The forecasting results have the potential to assist governments to plan policies to contain the spread. Finally, the number of COVID cases and the recovery rates were compared with the GDP per capita of a country, social support by the people, and the healthy life expectancy of a country to find if a correlation exists between them and to visualize the findings if a correlation exists. By correlating COVID-19 data with the world happiness dataset, we could find how the socio-economic factors influence the rise/decrease in positive cases and recovery rates.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

The outburst of COVID-19 created a tough time across the world. The first case was reported in Wuhan, the capital of China's Hubei province, and was reported to the WHO (World health organization) on 31 December 2019 (Shereen et al., 2020). WHO confirmed it to be a global pandemic on the 11th of March and the infection rates have been rising exponentially ever since then (WHO, 2020b). The COVID-19 has been mutating in different regions and it has been acting differently in different places. This creates the need to find out the different impacts of COVID-19 in different regions. Health workers and the government aren't prepared for tomorrow. There might be a rise or fall in cases or an exponential rise in the number of positive cases in the future that might create a demand for more health equipment or workers. If there exists a rough estimate for the near future, the government could get a rough estimate and be well prepared in advance.

As we see the daily cases, there are too many countries suffering from COVID-19 even today. There are poor countries that have a low GDP and those with low GDPs might have poor testing equipment and this is going to rise the infections as a lot of people do not get tested. Other factors may include social support provided by the people of a country or the freedom given to the people. We know the infection spread through social contact, but the rise in cases and deaths may correlate with many other unknown factors. The COVID -19 acts differently in different regions, it is mutating differently making its predictions difficult. The uncertain nature of the virus makes the government and health care workers go unprepared for medical needs. The lack of prediction caused inadequate hospitalizations to the affected people. Per CDC, the increase in the number of cases is due to contact, droplets, and airborne transmission but, there could be other unknown factors contributing to the increase in the positive cases, increase in recovery, or the decrease in recovery rates. By correlating COVID-19 data with the world happiness dataset, we could find how the socio-economic factors influence the rise/decrease in positive cases and recovery rates.

1.2 Research Questions

- Is there a positive relationship between the number of deaths and the factors like a surge of positive cases and hospitalization rate at the time of a surge?
- How would the number of cases look in the near future, and is there a pattern in the increase of COVID-19 cases?
- Does the ARIMA model have better forecasting accuracy than the FbProphet model for pandemic forecasting?

- Countries with a great economy are on the top in high infection rate, are there any hidden factors that influence COVID-19 cases in a country?

1.3 Potential Benifits

The research has eyed the total number of cases, deaths, recoveries in different countries to gain better insight into the effects of COVID-19 across the globe. The research has also included a comparison of the effects of COVID-19 on different states of the USA as it is the current epicenter of the disease. Visualizing the data to gain better insight into the COVID-19 scenario. A lot of countries are out of ventilators and are currently having limited accommodation for patients suffering from COVID19 pneumonia (Elegant, 2020). Being able to accurately forecast when the outbreak would surge, which would significantly diminish the impact and curb the spread. Forecasting the number of cases will help health workers, and the government to be prepared for a rise in the number of cases. The comparison of the COVID-19 data with the world happiness report has given a better insight into the cause of too many positive cases and the recovery rates across the globe and would answer the research question. The ongoing COVID-19 pandemic has caused worldwide socioeconomic unrest, so, analyzing the socio-economic indicators would help governments to reform the socio-economic policies and take measures to reduce its spread. The socio-economic factors help governments to alter their policy accordingly and plan for the preventive steps needed such as public health messaging, raising awareness of citizens, and increasing the capacity of the health system

1.4 Dataset

1.4.1 John Hopkins University

Repository: Center for Systems Science and Engineering (JHU CSSE), updated on a daily basis. (Project, 2020)
Primary Source: WHO, WorldoMeters, COVID Tracking Project, etc, from January 21, 2020.
Data used in this project: Global Confirmed Time Series data, Global Recovered Time Series data, and Global Deaths Time Series data.
Time Period: 21st January - 2nd of November.

1.4.2 The COVID Tracking Project

Repository: The COVID Tracking Project, collected on a daily basis. (University, 2020)
Primary Source: Local or state public health authority data, from January 21, 2020.

Data used in this project: The total positive cases, the positive increase in cases each day, the total death cases, the death increase in each day, the increase in hospitalization, and the total recovered cases.
Time Period: 21st January - 2nd of November.

1.4.3 Kaggle

Repository: Kaggle (World Happiness Report- yearly publication by the United Nations that consists of a score given to a country based on the socio-economic indicators). (Kaggle, 2016)
Primary Source: Responders of a country which is based on a survey conducted by the UN.
Data used in this project: Gross Domestic Product per capita, Social Support, Healthy Life Expectancy, Freedom to make life choices, and Perceptions of corruption (socio-economic indicators) of 155 countries.
Time Period: 2019.

2 Literature Review

2.1 Prior Work

On 28 July 2020, we have 5.7 million active cases, 650K deaths, and about 1 million recovered cases of COVID-19 summing it to about 16 million cases worldwide making it one of the largest pandemics in the history of diseases ("COVID-19 CORONAVIRUS PANDEMIC," 2020). Comparing with the 2003 Severe Acute Respiratory Syndrome (SARS) outbreak, the death rate was 10%- 8098 cases and 774 deaths, and Middle East Respiratory Syndrome (MERS) had a death rate of about 34% between 2012 and 2019- 2494 cases and 858 deaths (CDC, 2020b), (WHO, 2020a). As we can see COVID-19 has a higher death rate than any other pandemic in recent times and is higher than SARS and MERS combined. The death rate for COVID-19 so far is 3.9% and the death rate is lesser compared to SARS and MERS. Even though the cases have been increasing exponentially over time, there has been a reduction in the number of deaths due to COVID-19. On 13 April 2019, the death rate was 6.17% is comparatively more than the death rate on 28 July 2020 (CDC, 2020a). However, the number of cases is still rising so the current death rate cannot be considered as it may be more or less than the death rate due to SERS and MERS when the pandemic ends.

The USA has the highest number of cases and casualties of the COVID19 and the recovery rate is maximum in China (Dixit, 2020). Even though COVID-19 started in China, the graphs show that China has dealt with the pandemic well. Graphs show that China had

followed very strict lockdown restrictions from the 23rd of January 2020 that kept the people out of exposure which thereby contained the disease in China (Dixit, 2020). As of April, the USA was on the top of the confirmed list followed by Italy and Spain (Khanam et al., 2020). But today, Italy and Spain have contained the confirmed cases and are having very few numbers of cases each day (Project, 2020). This is because Italy followed a lockdown from the 9th of March and Italy's confirmed cases graph had a rising curve because the lockdown was initiated when it was a little late (Dixit, 2020). It is quite clear from the graphs that men and elderly people are more prone to get the disease (Khanam et al., 2020). The fatality rates have been considerably lower than many other diseases like Swine Flu, Flu, etc in the past (BEGLEY, 2020). However, the death count has been high in the USA and many other major countries like the UK, Italy, and Mexico ("COVID-19 CORONAVIRUS PANDEMIC," 2020). Death count and death rate have two different understandings. The death count might be high when the number of cases is high but if the number of cases is low or moderate and if the death count is high then that means the death rate is high.

A forecast is an estimate about an unobserved event or a trend and its surrounding uncertainty, which is based upon previously observed data (Lutz et al., 2019). Epidemic prediction initially began in India with environmental data combining with health records. Risk maps were developed for diseases like leprosy, pneumonia and these risk maps forecasted the diseases 2-3 months in advance for the government to take actions (Rogers, 1925). One of the oldest methods for forecasting was Early Warning Systems (EWS) that predicted risk or modeled projections of outbreaks that were based on collected information from the affected regions for mitigation and response (Myers et al., 2000). EWS was based on 1. Survey of the disease, 2. Modeling the risk using historical and environmental data, and 3. Forecasting the potential risk using different predictive models and surveys (Myers et al., 2000). The prophet model by Facebook is a forecasting tool where non-linear trends are fit with yearly, weekly, and daily seasonality effects (Lyla, 2019). It determines the changes in trends by finding the change points from the data. FbProphet is a quantitative forecasting approach. FluSight is an evaluation technique for forecasting that is based on the probabilities of different outcomes (Lutz et al., 2019). It forecasts the probability of a peak, the number of cases for example, for a particular week. Prophet has been used in several other applications like sales, supply chain, etc. and its performance in the real-world has been quite accurate and the results were satisfactory(Zunic et al., 2020). According to (Luo et al., 2017) ARIMA model is used to capture patterns in time series with lesser computation efforts. Hence they have been adapted as a benchmark model to test the effectiveness of other forecasting models. This project uses the SARIMAX model to compare the forecasting performance with the FbProphet model.

2.2 Problem Statement

The lack of forecasting systems makes it hard for the healthcare systems to predict a surge and be prepared for the healthcare needs, also, it makes it hard for the government to handle the pandemic without understanding the hidden socio-economic factors that influence COVID-19 cases.

2.3 Proposed Solution

COVID-19 data is visualized using graphs and charts comparing different countries and focusing on COVID-19 data for the USA. The USA is specifically focused as it is the current epicenter of the COVID-19 with a huge number of populations being affected. The project has focused on forecasting the numbers globally and in the US. Forecasting is essential today as it can predict the impact of an epidemic which will allow tactical responses that can be implemented when the risk of the disease is high. Forecasting the number of cases would also be helpful for hospitals to be more prepared for a rise as the rapid increase in positive cases increases the number of deaths rapidly. Moreover, the health care sector can also get a clear picture of the numbers. The total number of cases differ from country to country and the factors may include- the quality of living in a country, social support by the people of a country, the freedom is given to the people (partial locked like the one in the USA); the number of testing may depend on the GDP (Gross Domestic Product) of a country; the number of people getting admitted to a hospital may depend on the GDP per capita; the death rate might depend on the healthy life expectancy. All these data are available in the world happiness report dataset and the analysis has determined a correlation exists between any of these aspects with the COVID-19 data and the factors are graphically visualized. For example, more number of positive cases for a poor country might be associated with the GDP per capita or the number of positive cases might be high for countries that have a low score on the social support provided by the people.

3 Methodology

The plan is comprised of data collection, pre-processing, visualizing COVID-19 worldwide data, visualizing COVID-19 USA state-wise data, forecasting the number of cases, deaths, and recovery, merging world happiness report data and finding a correlation and visualizing COVID-19 data with any of the aspects of happiness report data that has a positive correlation. This will be an iterative process with improvements in stages over a while.

3.1 Data Collection

COVID-19 data for both global and USA states were gathered from [11], [8] respectively, and the world happiness report data were gathered from [12]. The Covid-19 USA state-wise data contains- the date of the report, the state, positive cases, negative cases, hospitalized, recovered, and deaths over a period of time. The COVID-19 global data contains- country, the number of confirmed cases over time, total deaths over time, and the recovered over time. Finally, the world happiness report contains- country, social support, freedom, corruption, and the GDP per capita.

3.2 Data Pre-processing

This step includes cleaning and preparing the data for the visualization. This preparation of data included removing unwanted columns, noisy data, void data followed by converting the dates to a date format for better visualization. Later, the world happiness report was merged with global COVID-19 data.

3.3 Visualizing COVID-19 global data

Data is difficult to understand if it is in the form of a table and people who do not work with data would find it hard to understand it. So, Data Visualization is important to graphically represent and report data and explain it in layman's terms. So the COVID-19 data has been graphically represented in the form of graphs and charts for data like positive increase in cases, test result increase, hospitalized vs deaths, fatality rates, the positive test ratio for the tests conducted; over a period of time for different states in the USA and globally.

3.4 Forecasting the Number of Cases, Deaths, and Recovery

After visualizing different aspects of COVID-19 data, forecasting the number of cases, deaths, and recovery was done. The forecasting is for the near future like the week after the date we have in the data or a month after. This has been done using FbProphet and ARIMA models.

3.5 Merging world happiness report data and finding if a correlation exists

Merged and determined a correlation exists between the COVID-19 data and the world happiness report data to analyze what might cause an increase in the number of cases or the

factors the contribute to the recovery of cases.

3.6 Visualizing COVID-19 data with any of the aspects of happiness report data that has a correlation

A correlation exists, COVID-19 data has been visualized with the aspect that has a correlation.

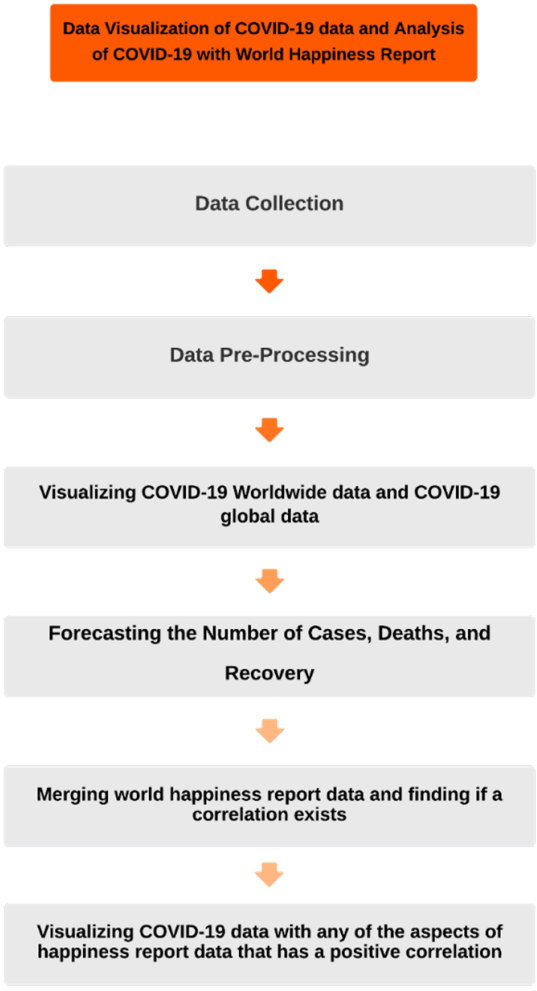

Figure 1. Flow of the project

4 Experiments and Results

4.1 Visualizing COVID-19 global data

In this section, the COVID-19 global data is visualized for the countries that currently have the highest positive cases, the highest deaths, and the highest recovered cases.

Table 1. Countries to be discussed

Aspect	Countries on top of the list and to be discussed
Positive Cases	USA, India, and Brazil
Deaths	USA, India, and Brazil
Recovered Cases	USA, India, and Brazil

4.1.1 Positive Cases

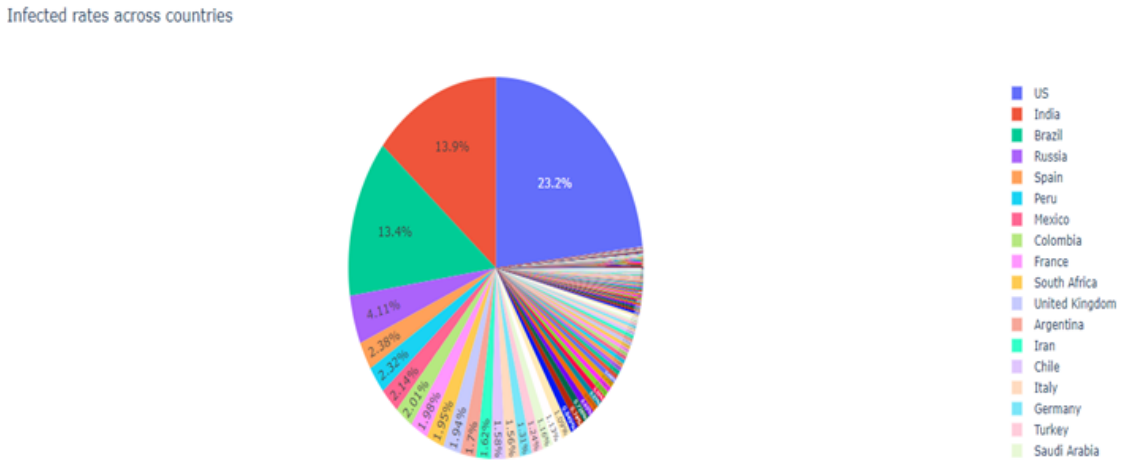

Figure 2. COVID-19 infection rates across countries

This section provides visual exploratory data analysis on COVID-19 global data using vertical bar charts and pie-charts. All the visualizations were done using Python. Firstly, the positive increase in the number of cases is visualized for India, the USA, and Brazil. The visualizations include the top countries affected by COVID-19 i.e., the USA and India, and

Brazil. The pie-chart shows the current standings of the COVID-19 infection rates across the globe. The United States, which is the current epicenter, tops with 23.2% followed by India with 13.9% and Brazil with 13.4%. However, the number of positive cases per day is decreasing in India and Brazil, which will be discussed in the forthcoming sections. So the top three countries would be discussed in this section.

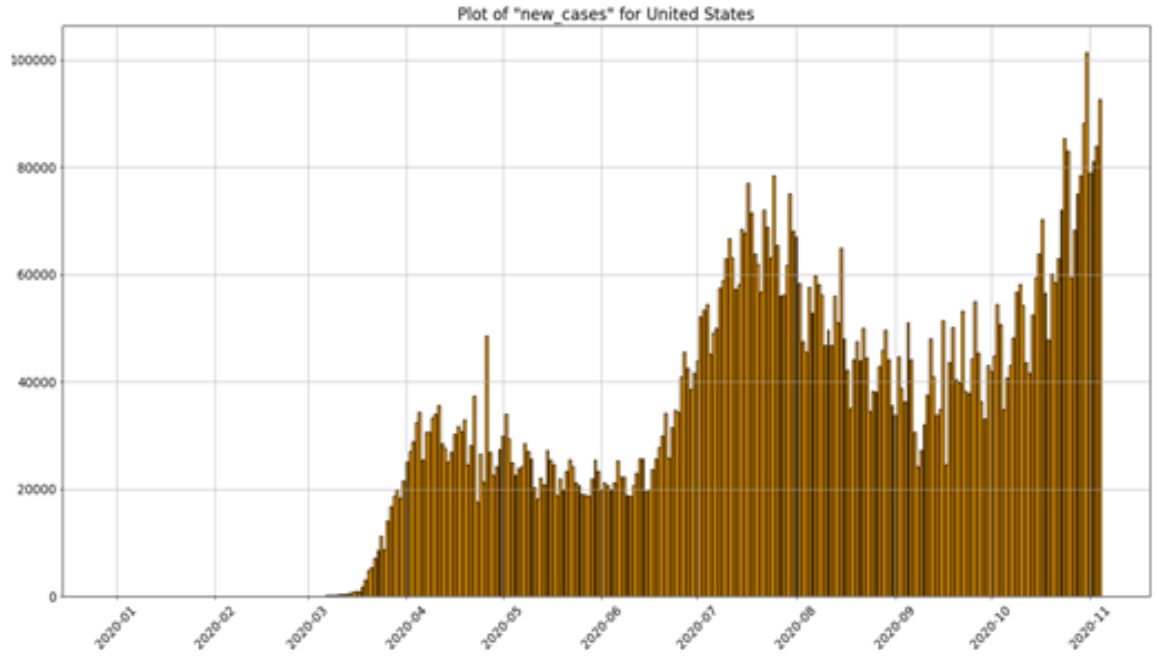

Figure 3. Infections over time in the USA

Based on the data used in the project, the first COVID-19 case in the USA was reported on 01/22/2020 and from figure 3 it is clear that the outbreak of the virus started to create an impact around March 2020. Since then, the infection spread rapidly, and the curve grew exponentially.

In India, the first case was reported on 01/30/2020 (based on the data) and from figure 4 it is understood that the outbreak started around May 2020. India was locked down from the 25th of March 2020 until the 31st of May 2020 (Wikipideia, 2020). From figure 4, it is clear that the number of infections was kept low until the lockdown period but since May 2020 the outbreak has been worse. Considering India's population, approx. 1.3 billion, the outbreak was forecasted at a much earlier date. But, because of the government's lockdown measures, the infections were averted to a certain extent.

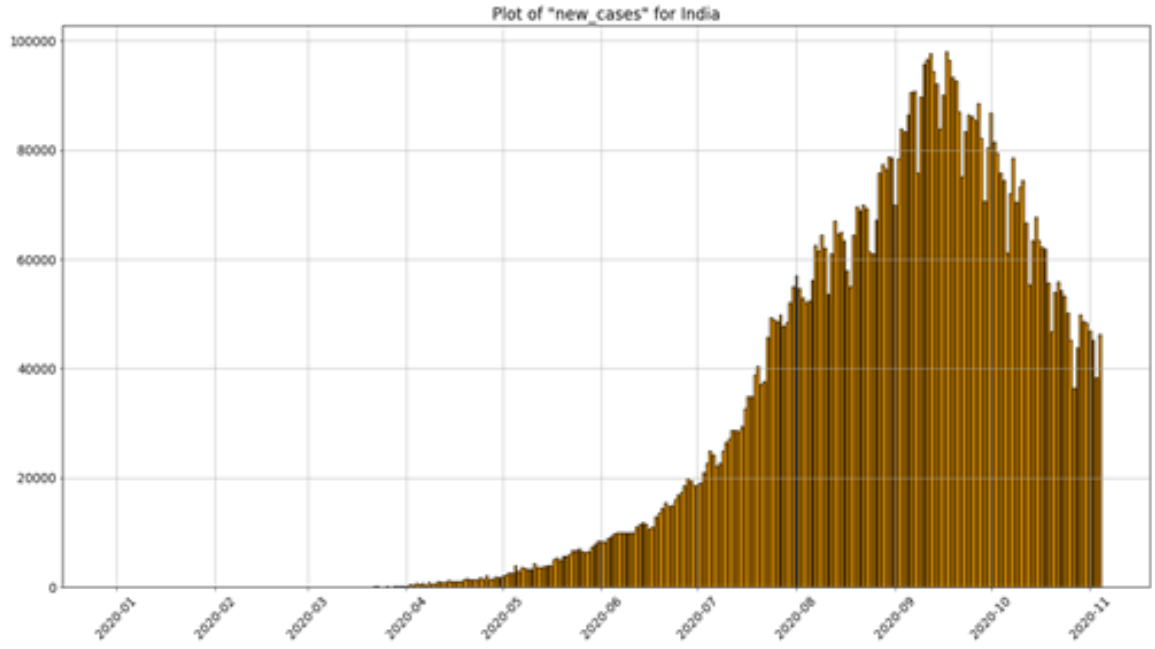

Figure 4. Infections over time in India

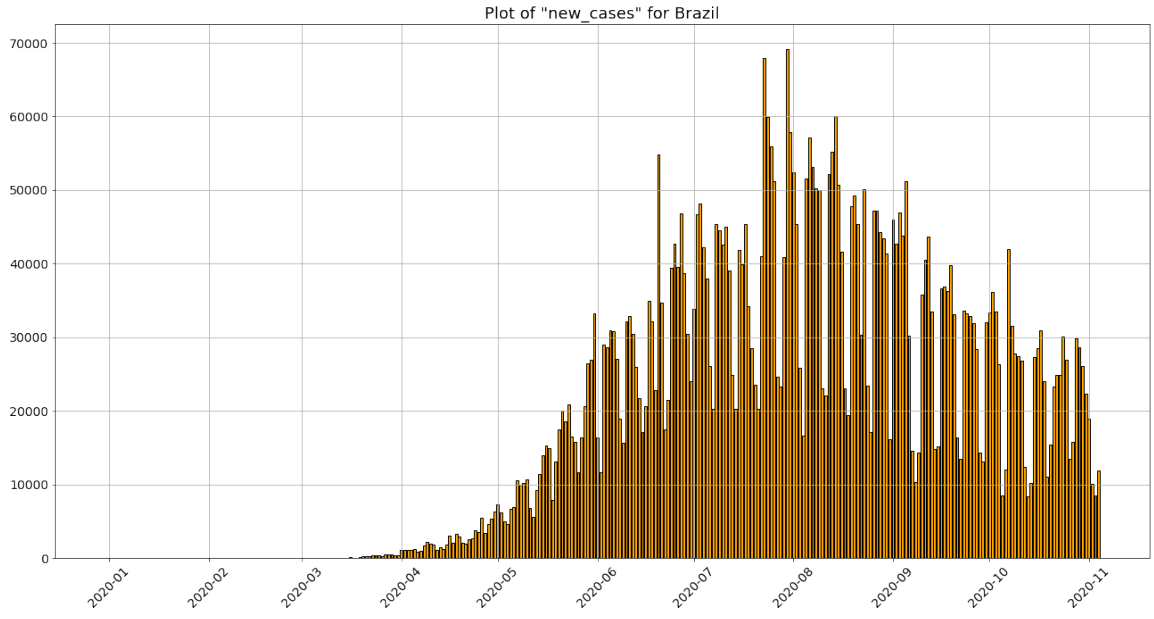

Figure 5. Infections over time in Brazil

From figure 5, it is clear that the infections started to rise in Brazil around June and it has been ballooning the caseload since then.

4.1.2 Mortality Cases

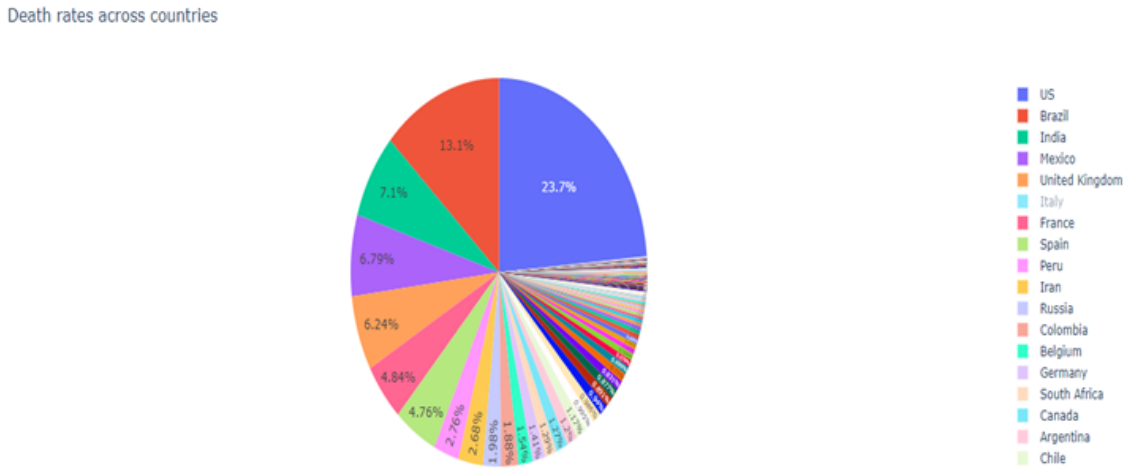

Figure 6. Death rates across the globe

The pie chart in figure 6 shows how low India is in death rates compared to the other majorly affected countries Brazil and the USA. The difference in percentage between the US and India is 16.6%.

As discussed in the previous sections, the death rate for COVID-19 is less than that of the death rate caused due to other coronaviruses (SARS, MERS, etc.) in the past. During a pandemic, the death rate is one of the major concerns as a lot of people die all across the globe. The USA, India, and Brazil are focused here as the three countries have the highest mortality cases. Figure 7 depicts the deaths that occurred due to COVID-19 over time. It does seem like the number of deaths per day has dropped but there are days where the number of deaths has been high.

Figure 8 depicts the number of deaths per day since the beginning of the outbreak in Brazil. Though the USA is the highest in the number of deaths, by looking at the figure of Brazil it looks like Brazil might surpass the USA in the number of deaths if it continues at the same rate. One of the positives for India is that the number of deaths per day is lying low compared to the population of the nation. India has the best recovery rate compared to any other country affected by the pandemic. Figure 9 shows how the number of deaths has occurred per day in India and the number is going down towards the end.

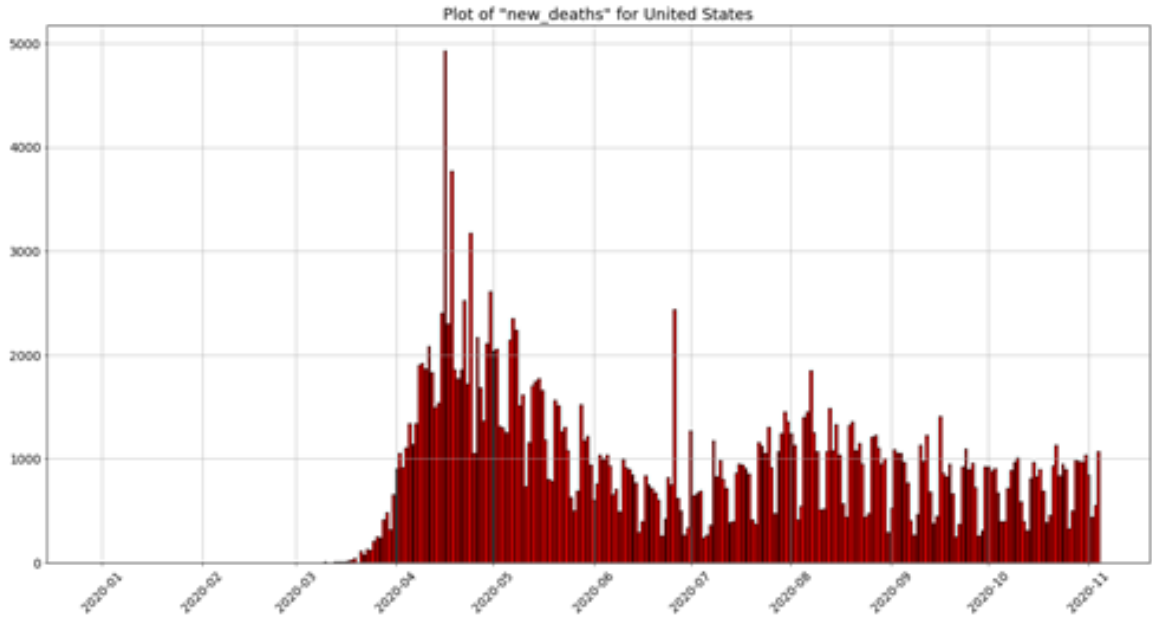

Figure 7. Deaths per day over time in the USA

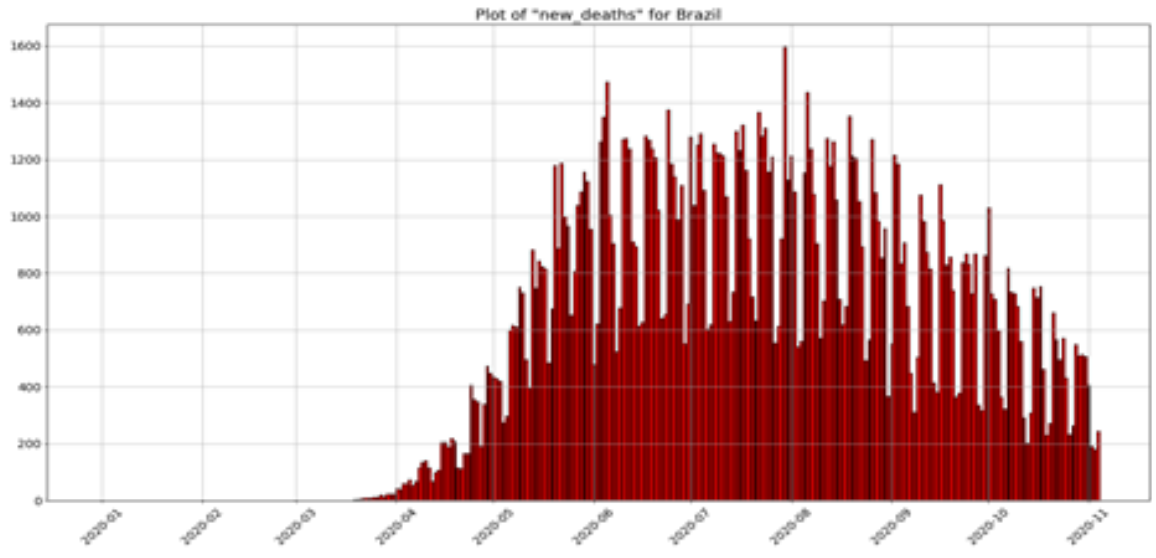

Figure 8. Deaths per day over time in Brazil

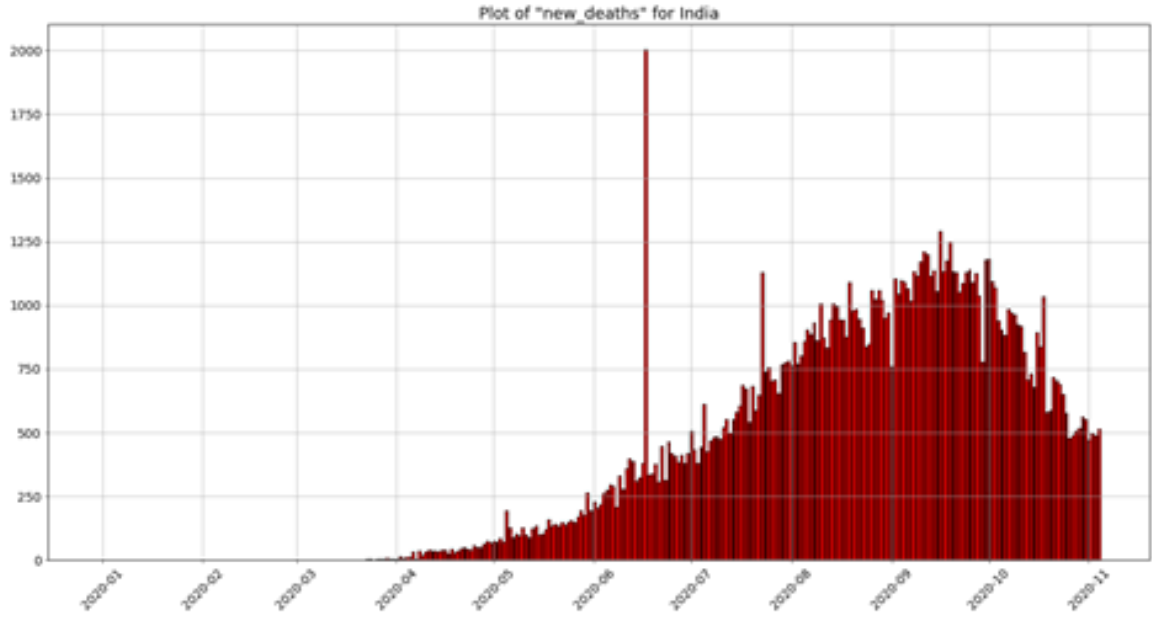

Figure 9. Deaths per day over time in India

4.1.3 Recovered Cases

One of the positives during a pandemic is the recovery rates. Even if there are a lot of infections, a higher recovery rate is always a positive in a negative situation. A higher recovery rate means the mortality rates are lower. Figure 10 shows the recovery rates across different countries. It is clear that the recovery rate of India is the highest with 17.9%. As stated, we can see that a lower death rate means a higher recovery rate. India had a lower death rate and has a higher recovery rate. Similarly, the US had a higher death rate and hence it has a lower recovery rate than that of India and Brazil.

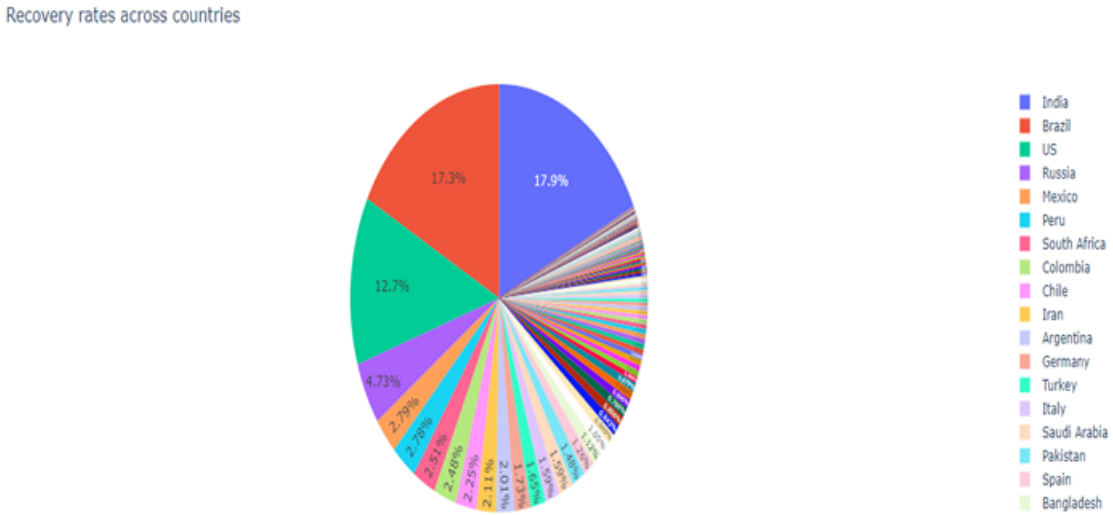

Figure 10. Recovery rates across the globe

Since this is an ongoing pandemic we may never know as the tables might turn. Other countries might surpass India in the recovery rate. For example, Italy had the worst impact of COVID-19, in terms of deaths, during April and May (Indolfi & Spaccarotella, 2020). But subsequently, as the USA, India, and Brazil surpassed the rates Italy went down the table. So, until the pandemic is over we wouldn't be able to tell which country had the best recovery rates or death rates and which country did well during the pandemic.

4.2 Visualizing COVID-19 USA state-wise data

Table 2. States in the US to be discussed

Aspect	States on top of the list and to be discussed
Positive Cases	NY, TX, CA, FL.
Deaths	NY, CA, TX
Recovered Cases	TX, NC, TN

This section provides visual exploratory data analysis on COVID-19 USA state-wise data using vertical bar charts. All the visualizations were done using Python. Firstly, the positive increase in the number of cases is visualized for the states in the US that are on the top of the table. The visualizations include the top states affected by COVID-19 including but not limited to New York, California, and Texas.

4.2.1 Positive Cases

According to the dataset used in this project, the first case in the USA was reported in Washington on 01/21/2020 and that was the first test conducted in the USA. Even though the first case was during the end of January 2020 the outbreak did not create a serious impact until March 2020.

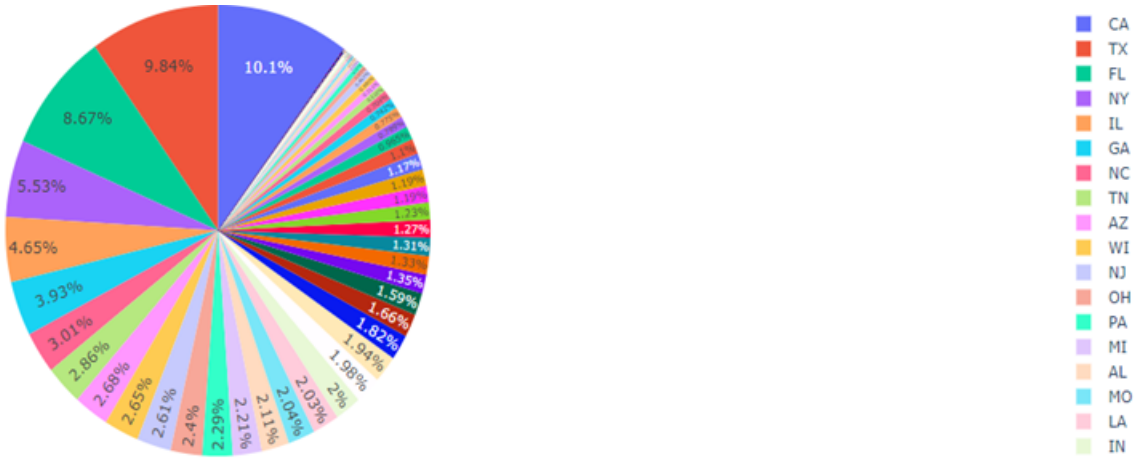

Figure 11. Total positive cases across the US

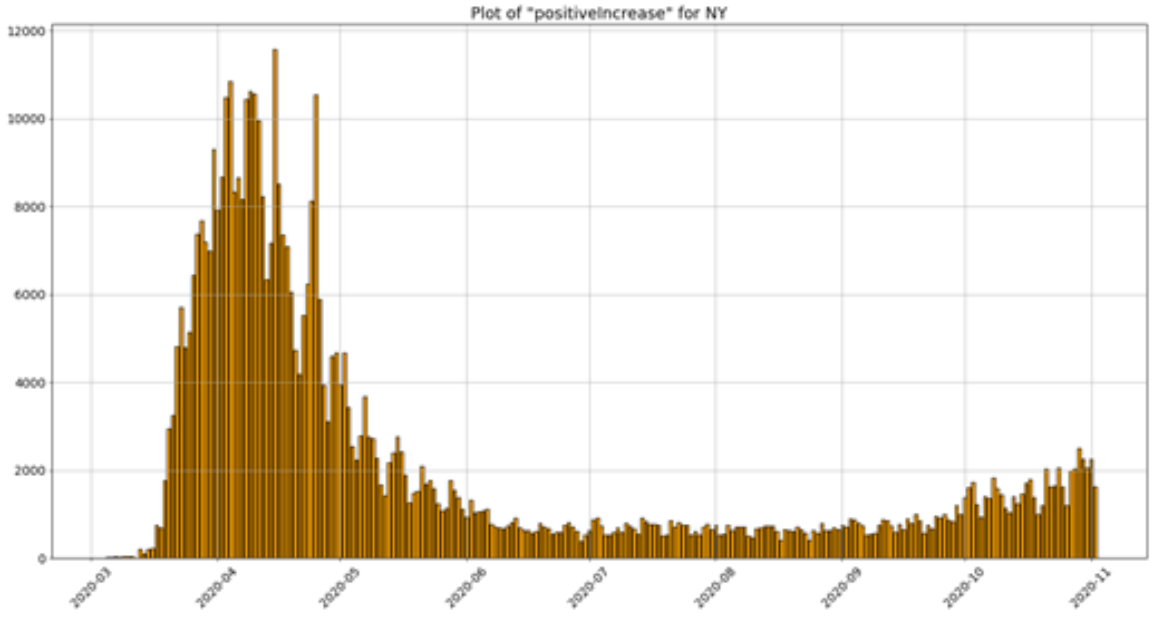

Figure 12. Positive Increase of infections in NY

NY was the first to get majorly affected by the virus in the USA. Figure 12 shows how the positive increase in infections has been each day in New York and it is clear that New York had a peak from March to May and there were government restrictions like working from home, schools and colleges were shut down, restaurants and pubs had limited access and the number of people in a closed area was also limited (State, 2020). All these restrictions and policies like wearing a mask flattened the curve up to a certain extent and as Figure 12 depicts, the number of cases came down during this period. On the other side of the USA, California had a slight rise in the number of cases per day which was when California shut down its economy to control the spread. But the economy was reopened in May 2020 which was when the number of cases started skyrocketing and California lost control of the spread (News, 2020). Figure 13 explains how the cases increased in California after the economy was completely opened in June 2020.

Figure 12 and Figure 13 depicts the tale of two different states (NY and CA) and they are completely two different stories. The peak is left-skewed for NY and California, it is more of a right-skewed peak. However, the case for Texas has been a little different. As seen in Figure 14, the numbers started skyrocketing towards the end of June 2020. Texas and California are still having a lot of positive cases each day. Based on the dataset used in the project California is currently the state with the highest number of COVID-19 cases in the USA. As discussed, the total number of cases for New York is lower than that of

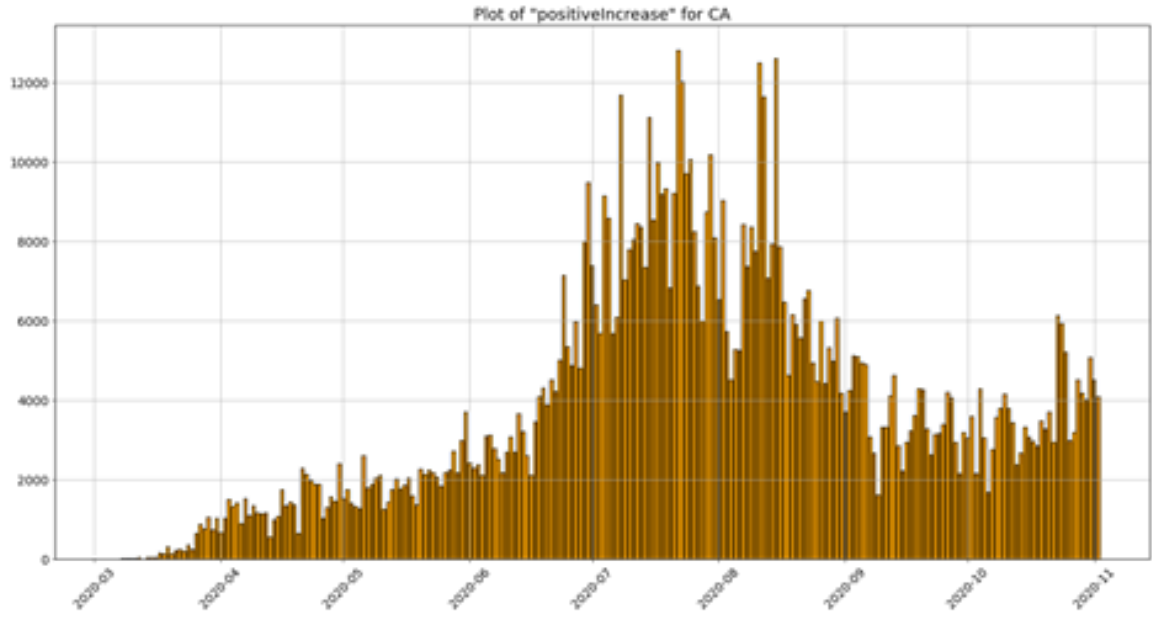

Figure 13. Positive Increase of infections in CA

California, and Texas. Figure 15 proves how NY has controlled the spread of the virus to a certain extent. During the first 50 days of the outbreak, NY was out of control, and CA,

TX, and FL were in control, and after 100 days of the outbreak NY flattened the curve to a certain extent but CA, TX, and FL had the total number of cases shoot up.

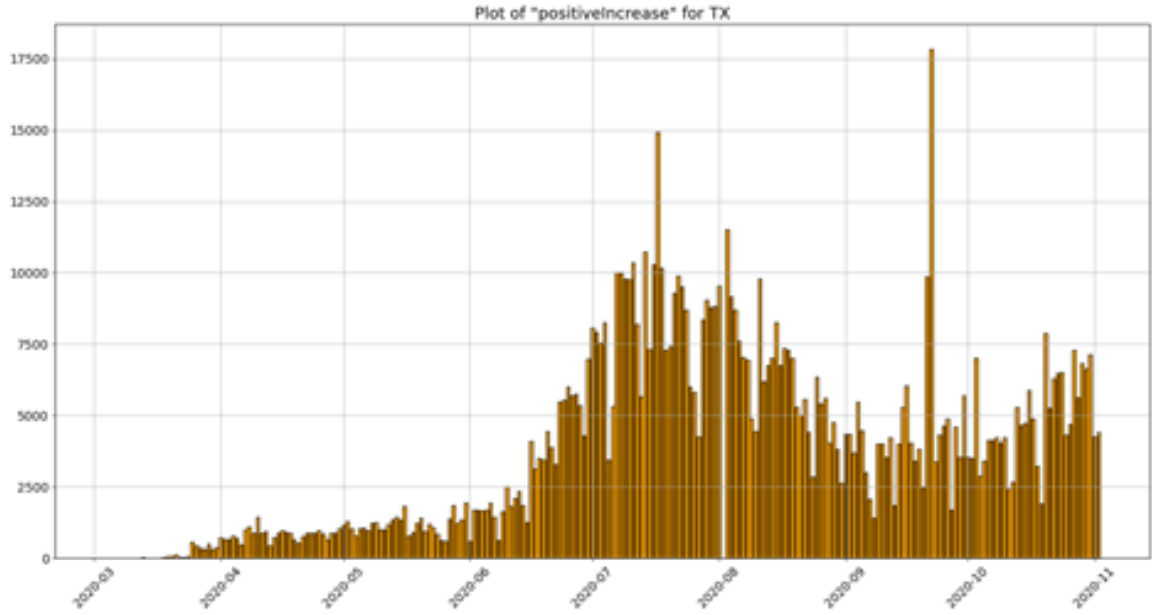

Figure 14. Positive Increase of infections in TX

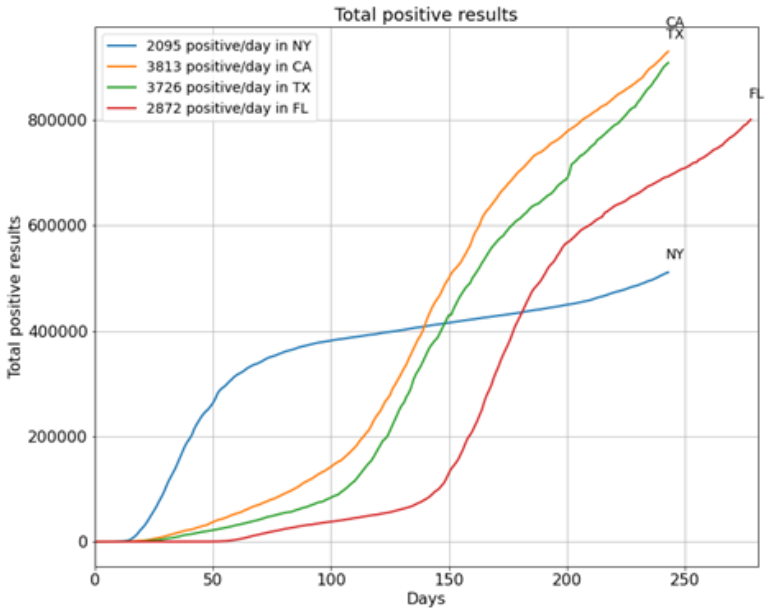

Figure 15. Total cases overtime for CA, TX, NY, and FL

Figure 16 visualizes the successive changes in new positive cases for the last 14 days, if the curve is under zero it means the state did better on that day, in CA, NY, and TX and it is clear that NY is having ups and downs every day but it is remaining low, at least lower than CA and TX at this moment. TX is quite unstable as it has too many cases a day and a lot less the other day.

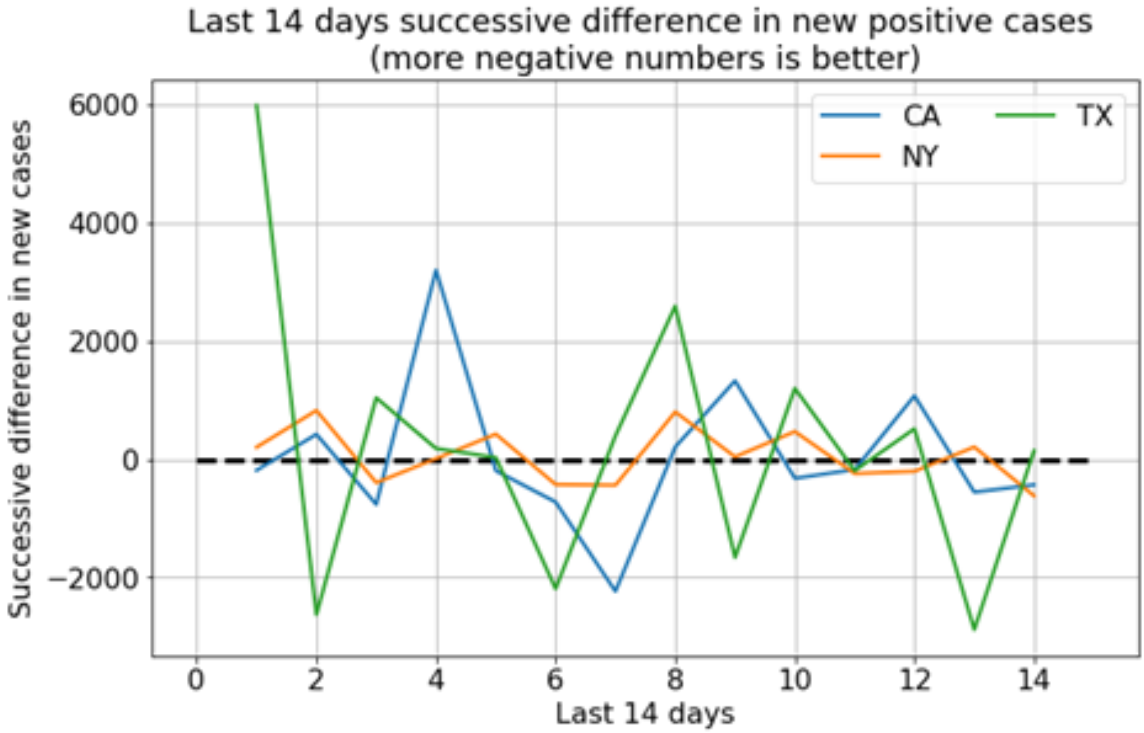

Figure 16. New cases last 14 days for CA, NY, and TX

4.2.2 Mortality Cases

As mentioned in the global COVID-19 visualization, the USA tops in the total number of deaths around the world. In this section, specific states that have the highest fatality rates are discussed.

Figure 17, the top states with the highest death cases are New York, Texas, and California, which is quite obvious as these are the top three states with the highest number of COVID-19 cases, with approx. 27% of the deaths in the USA.

Figure 18 visualizes the death increases per day in the state of NY and it depicts how the deaths were high between March and June 2020 and how the deaths have decreased over

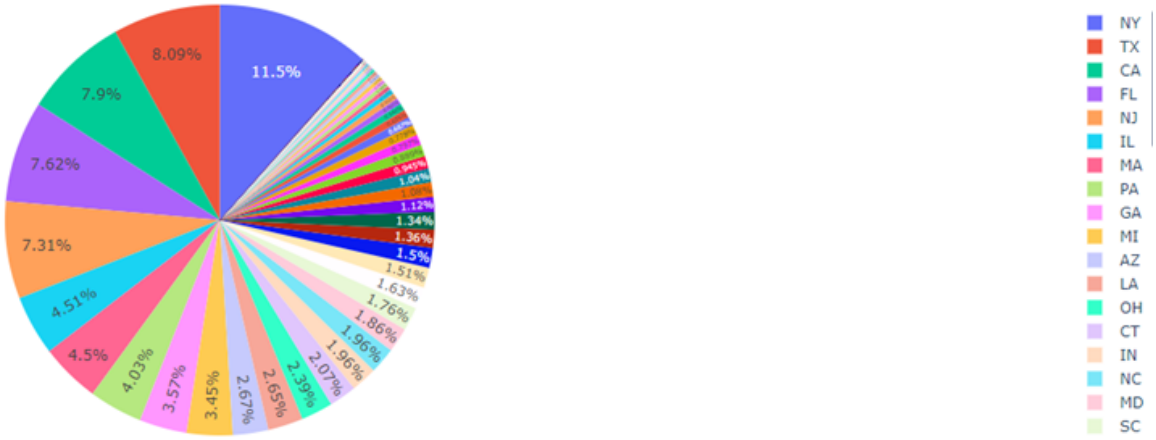

Figure 17. Total death cases across the US

time. So, NY is on the top because of the peak that occurred between March and June because the deaths have decreased drastically since mid of June 2020.

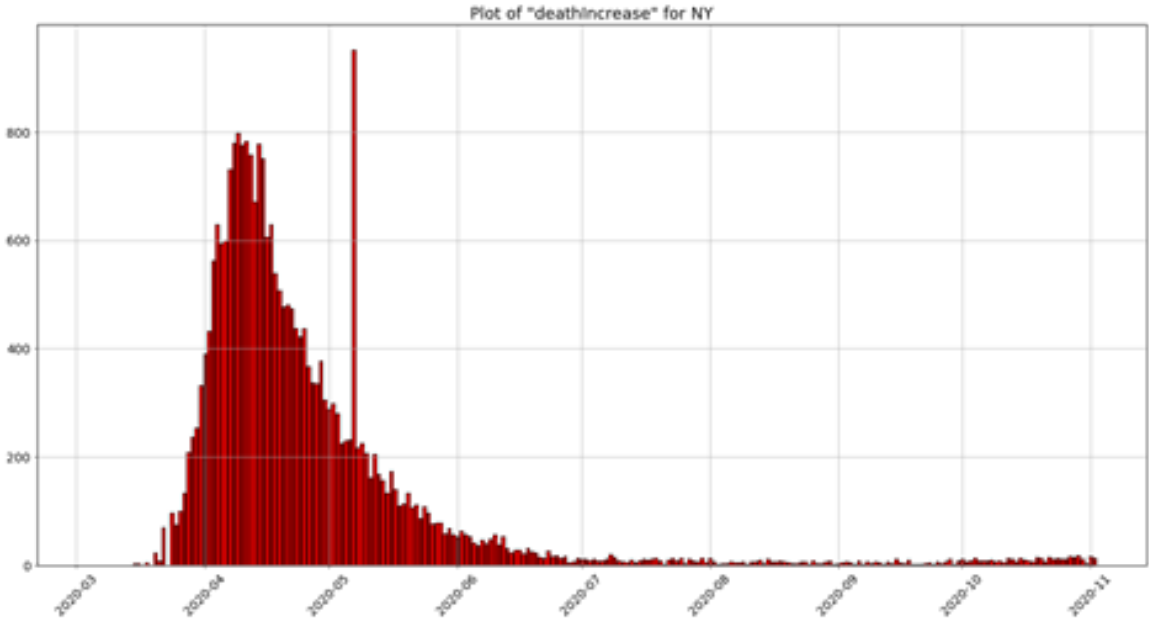

Figure 18. Death increase per day in NY

It does seem like New York wasn't prepared for the pandemic with enough hospital staff and equipment. Figure 19 visualizes how the hospitalization increase has also significantly increased the deaths. The slope looks positive and it justifies because there were unexpectedly too many cases in NY during the start of the outbreak in the USA. On the other side,

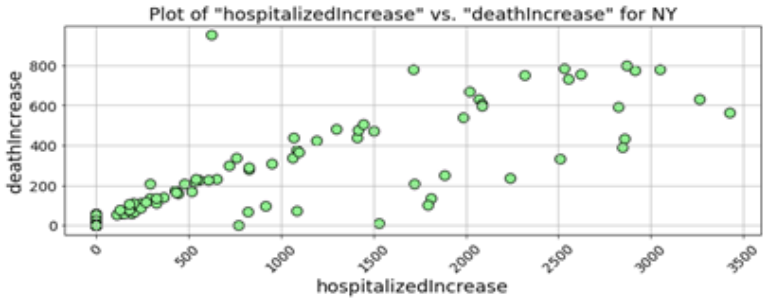

Figure 19. Death increase vs Hospitalization increase in NY

California hasn't been as bad as the US in the number of deaths per day. Figure 20 shows that the highest point in the number of deaths per day for California is approx. 250 but the highest point for NY was approx. 900 (based on figure 18). The positive for NY is that the number of deaths per day has decreased drastically and it is very low currently, but for California, it is still a lot higher.

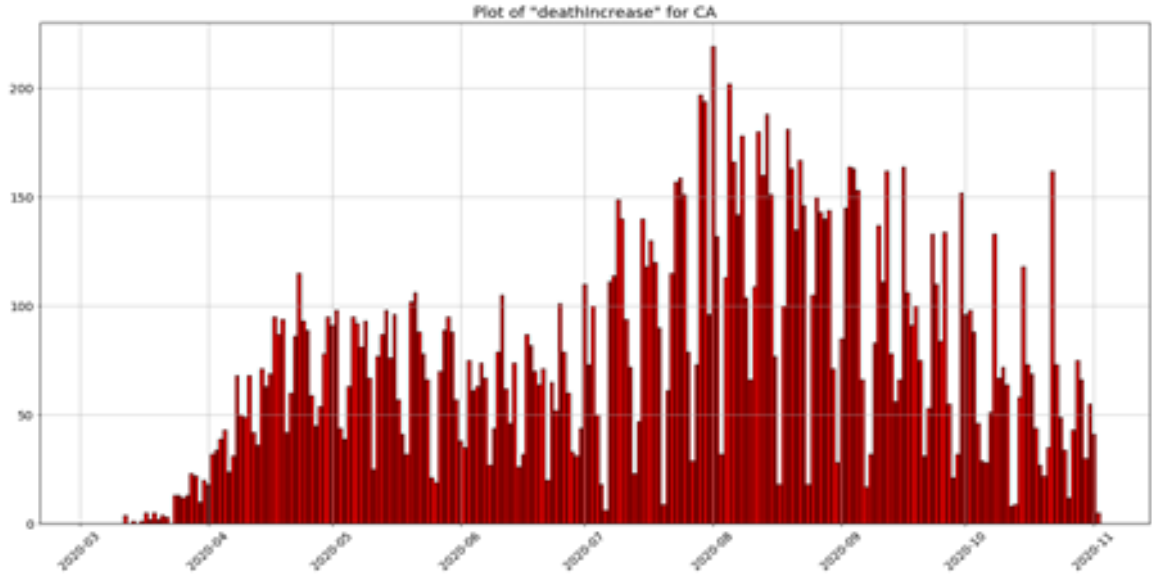

Figure 20. Death increase per day in CA

Texas has more deaths than CA currently which is also the state second in the total number of deaths in the USA. The peak of the outbreak started quite late in Texas around July 2020 and is seen in figure 14. Based on the visualizations of positive increase and the death increase per day it is also clear that the number of deaths is the highest during the peak of the outbreak for the states that are discussed here. For example, NY had its peak in the outbreak during April and May and the deaths per day were also high during that time. This

might be because too many were hospitalized during the period and NY state health care wasn't prepared. Figure 19 supports the claim that the increase in hospitalization (which is also an increase in the number of cases) significantly increased the deaths in NY. Figure 22 shows how increased cases have increased the deaths per day for the state of California and as mentioned this answers the claim that the deaths are highest during the peak of the outbreak.

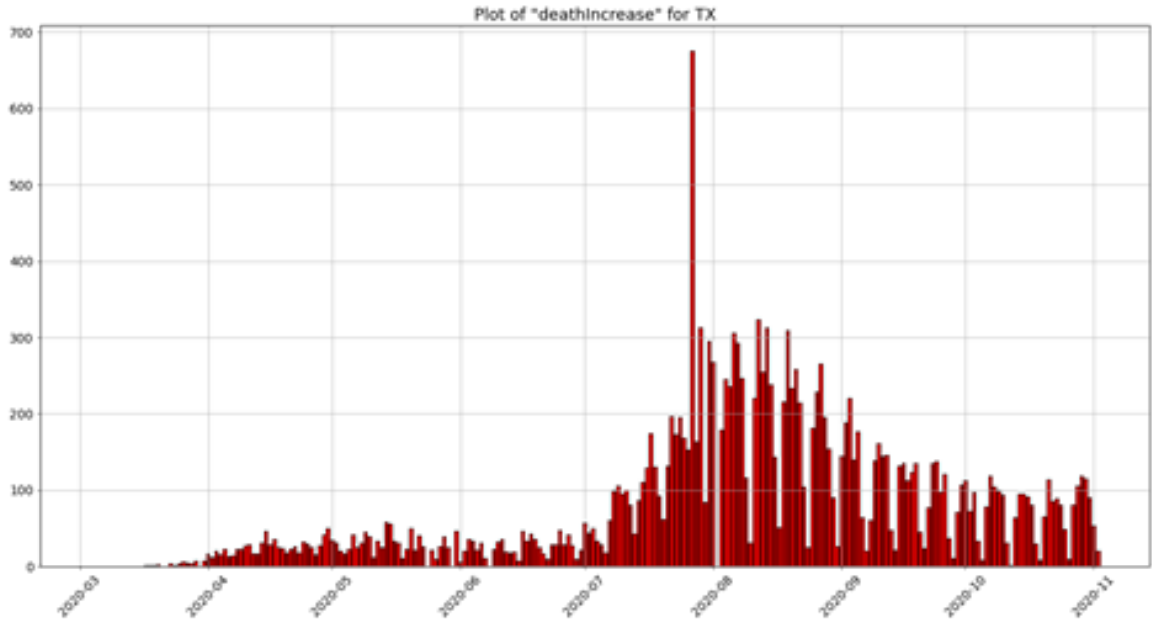

Figure 21. Death increase per day in TX

The fatality ratio is the total deaths that occurred to the positive cases of a region (a state in the US in this project). The death rates in the US were highest for NY but the fatality ratio is the highest for the state of New Jersey.

Figure 24 depicts the fatality ratio in volumetric visualizations across the different states in the USA. Connecticut, Massachusetts, and New Jersey have the highest fatality ratios which means that there are too many death cases for the number of positive cases that occurred in these states.

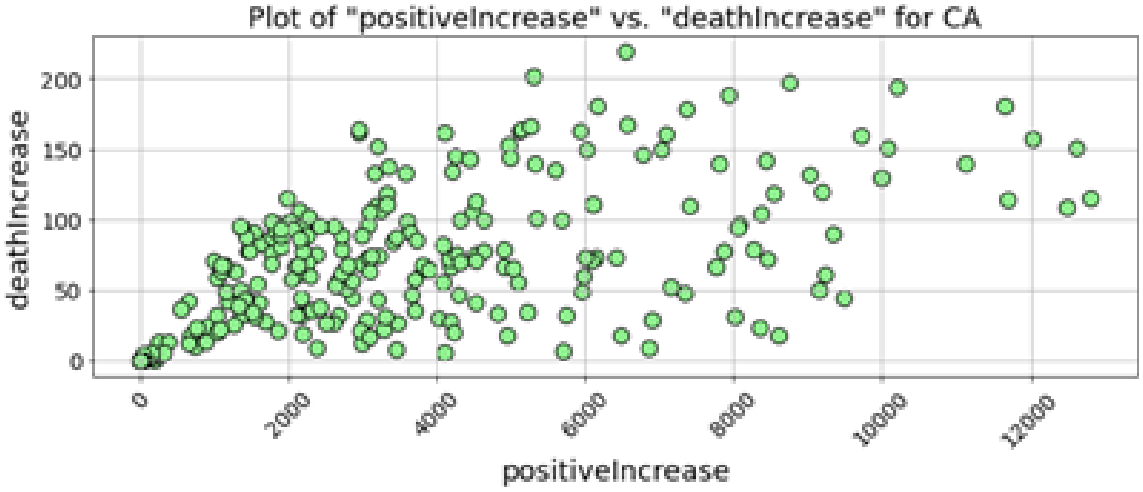

Figure 22. Positive increase vs death increase

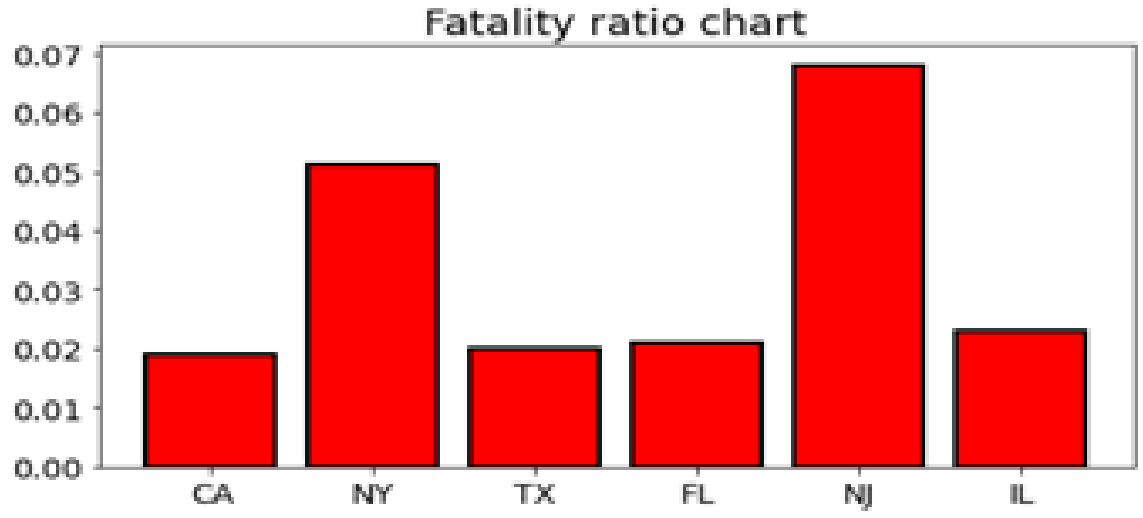

Figure 23. Fatality ratio for the USA

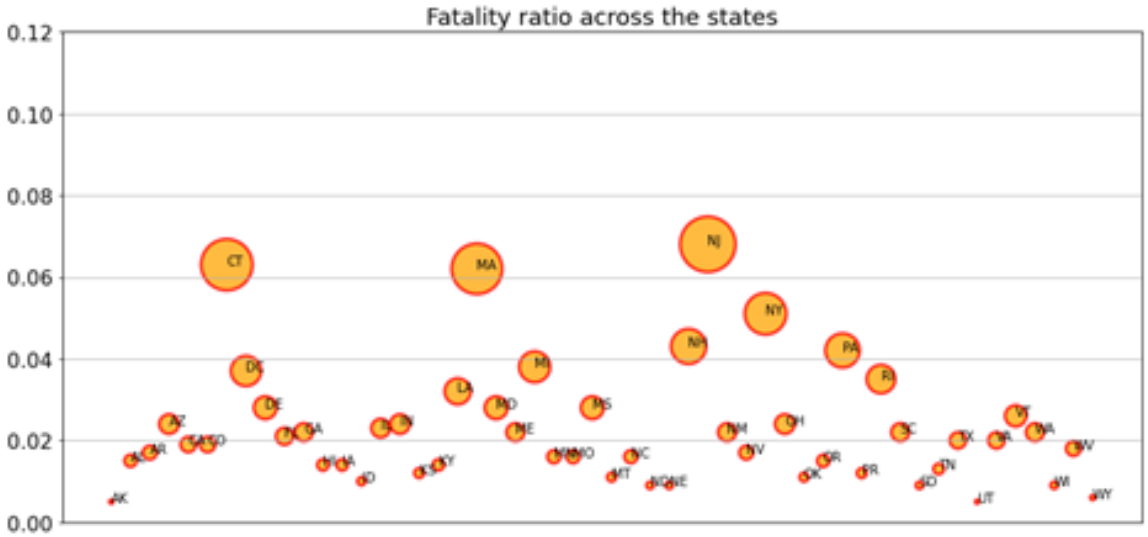

Figure 24. Fatality ratio (in volume) for the states of the USA

This section answers the first research question on how there is a rise in the number of deaths during the peak of the outbreak. We have also seen that the increase in hospitalization has an increase in the number of deaths. It can be inferred that during a peak of the outbreak, a rapid increase in the number of cases would increase the hospitalization number. This creates a shortage of ventilators, hospital beds, and equipment. This positive relationship could be because of the hospitals not being able to accommodate too many cases. This could be due to the lack of equipment and health workers during the unexpected rise in the number of cases. This could be the reason for the increase in the number of deaths during the peak. Figure 19 supports this claim of how the increase in the hospitalization rate causes an increase in deaths. There is a positive relationship between the number of deaths and the surge in positive cases. This creates the need for a forecasting system to foresee another spike that would enable hospitals to be prepared to provide sufficient hospitalization

4.2.3 Recovery Cases

In the USA, Texas, North Carolina and Tennesee have the best recovery rates making it approx. 34% of that of the USA. Since this is an ongoing pandemic these numbers are not going to be static and would change in the course of the future. The more the COVID testings are conducted the more it is easier to control the number of cases. This is true because the individuals who are tested positive could be quarantined and the individuals wouldn't spread the infections further. If the aren't many tests conducted, then the individuals with COVID wouldn't be tracked and the individuals would infect other people due to this. So during a pandemic testing is the most important step a government would have to take. Based on figure 26 that visualizes the total tests done per state, it can be seen that CA,

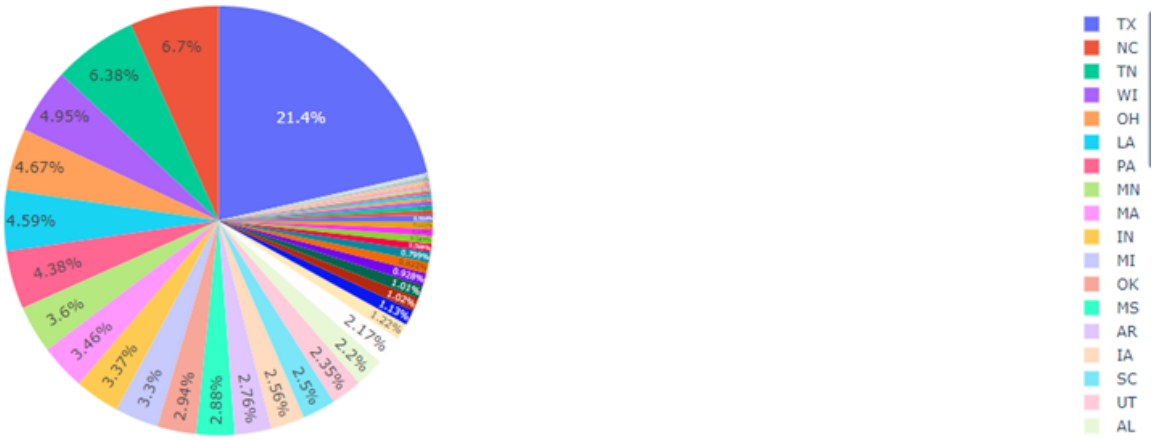

Figure 25. Recovery rates for the states of the USA

NY, and FL are doing a great job in testing and that might also be the reason NY was able to control the outbreak during the peak.

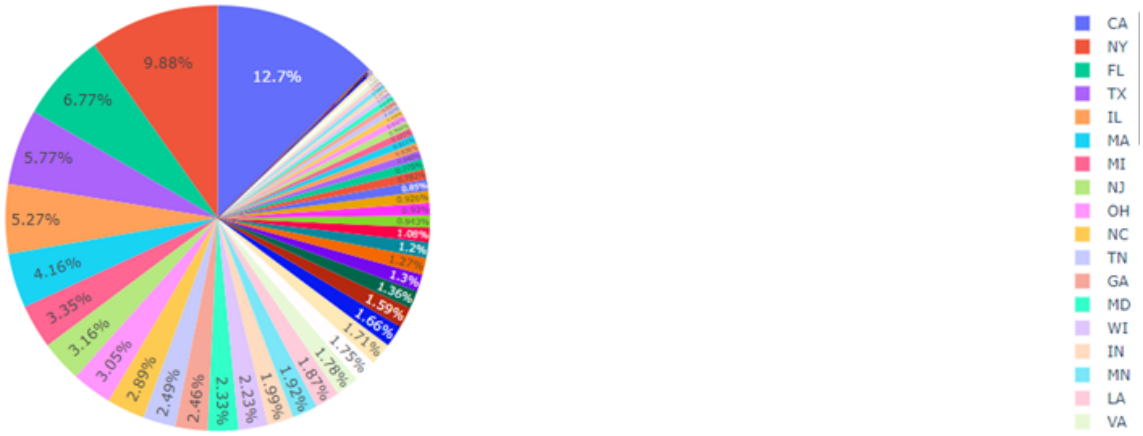

Figure 26. Total Tests conducted for the states of the USA

It was clear how the number of deaths increased during the peak of the outbreak. But, since this is an ongoing pandemic, the numbers aren't final. So we may have another peak anytime and this can be predicted using forecasting the numbers for the near future. If the numbers are forecasted, hospitals and health workers could be ready and the government could also see to it that if there is another peak in the number of infections, hospitals are ready with a sufficient amount of beds, and ventilators, and other equipment.

4.3 Time Series Forecasting Models

Forecasting is one of the most important concepts to implement during an ongoing process to get an idea of what the future has got. Especially, during the current pandemic situation, it is necessary as it is the COVID positive cases are quite unstable. Just when France, Italy, and the United Kingdom thought it had controlled the virus, it is now up for the second wave of the virus (based on the latest dataset). So forecasting tools are quite helpful in these types of situations, for example, a tool could forecast a sudden rise in the number of positive cases for a country in Europe and this could mean the neighboring countries could also be in trouble.

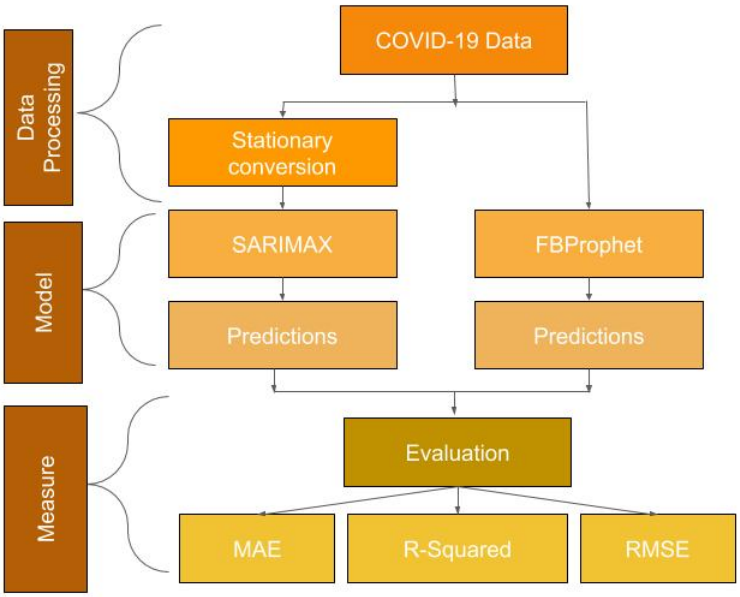

Figure 27. Flow chart for the time series forecasting section

In this project, the tool used to forecast is FbProphet which is a forecasting tool developed by Facebook Inc. and the tool has been used in a number of fields like the stock market, economy prediction, etc, and ARIMA tool which is an Autoregressive model based on a moving average and integration model.

4.3.1 FbProphet- A time-series forecasting tool

A time series is a sequence of data points plotted or tabulated in time order. It is a series of equally spaced points in time. Time series forecasting is a technique to use a model to

forecast future values with the previously observed values over time. A time series forecasting model takes into account- the trend, and seasonality. FbProphet is an additive model in which the predictor does not take a predetermined form and is modeled based on the information from the data, and is called a nonparametric regression model. It uses several non-linear and linear methods as components and time is the regressor. A nonparametric regression requires a good amount of sample sizes because the data must suit the needs of a model structure and the estimate (forecast). The additive model also uses a smoother- a smoothing technique that captures important patterns and trends in the data points and leaves out the noise or a sudden temporary rapid change in the data points. The noisy points are higher than the neighboring points are reduced and the points lower than the neighboring points are increased to create a smoother model. This will help in the COVID-19 case because when the trend has been going low there would be a rapid change the next day and back to normal the other day. This wouldn't affect the model to predict a higher value based on the sudden change. That is the reason the forecast has performed pretty well. The Prophet model is fit using Stan, a platform for creating data models and high-performance algorithms.

The tool requires two features to be inputted- the time-series dates and a value associated with that particular date, for example, the stock market price for a series of dates. In this project, the inputs would be the date and the time-series dates of the pandemic and the everyday value (total number of cases, the total number of people recovered, or the total number of people dead) associated with the date, and it is given in the format of 'ds' and 'y' where 'ds' is the input date in date or date-time format and 'y' is the actual value in numeric format.

The inputs given to the forecasting model Prophet() are- the confidence interval, which in this case is 0.96, and optional inputs like- the seasonality, holiday effects, and the growth of the curve- Linear or Logarithmic. The confidence interval here is 0.96 because this is an ongoing pandemic that is unstable and changing every day. So, there would be a larger room for the lower and upper limits for the prediction. The input given to the prediction of a future variable is the 'period', which is the number of dates in the future to forecast. The way the model works is by learning how the trends have changed over time and how the data has been going in recent times.

Once the forecast has been made, the data that is returned contains the 'yhat' value which is the actual number forecasted, the 'yhat_lower' which is the lower limit of the forecast, the 'yhat_upper' which is the upper limit of the forecast, and the 'ds' the dates the values are forecasted for. The 'yhat_lower' and the 'yhat_upper' value lower and higher than the forecasted value respectively.

	ds	y
85	2020-10-29	45023444
86	2020-10-30	45594203
87	2020-10-31	46070822
88	2020-11-01	46509809
89	2020-11-02	46959365

Figure 28. Sample input that would be given to the FbProphet model

	ds	yhat	yhat_lower	yhat_upper
115	2020-11-28	58565852.47	57181370.78	59897314.40
116	2020-11-29	58957622.97	57563798.75	60315365.58
117	2020-11-30	59402552.75	57926338.11	60854899.16
118	2020-12-01	59831475.36	58197998.52	61364682.37
119	2020-12-02	60307915.06	58601362.76	61911853.85

Figure 29. Sample output returned by the FbProphet model

4.3.2 Forecasting global time-series data using FbProphet

The pandemic has been around for almost a year now and it is still at its peak and just when it was getting better in a lot of countries it is now back with a second wave. But, countries like India, the USA, and Brazil is still in the worst situation and nothing much has changed. Figure 31 shows how the curve would move in the next 30 days and it nowhere close to getting flattened. It is seen that by the mid of December it would be more than 55 million global cases of COVID-19. Based on the current trend it is going higher and higher. The curve is the forecasted value pointing at the 'y' axis and the blue-shaded regions are the regions of 'yhat_lower', and the 'yhat_upper'. This means that the forecasted numbers might be lesser than the actual forecasted number up to the 'yhat_lower' number or it could be larger than the actual forecasted value up to the 'yhat_upper' value. The dotted points on the curve are the values that are forecasted for the data that is already present. This is similar to the technique used in a supervised machine learning model where the data is split into training and test data and the test data contains the label for the feature class and the accuracy of the model is tested using by checking if the test data was able to predict the accurate class present in data itself. Similarly, if the dotted points align with the curve, it means that the forecasted value and actual values are accurate. It is clear from Figure 31 that most of the dotted points are aligned with the curve and the model is doing great. There

are very few points out of the curve but still inside the shaded region.

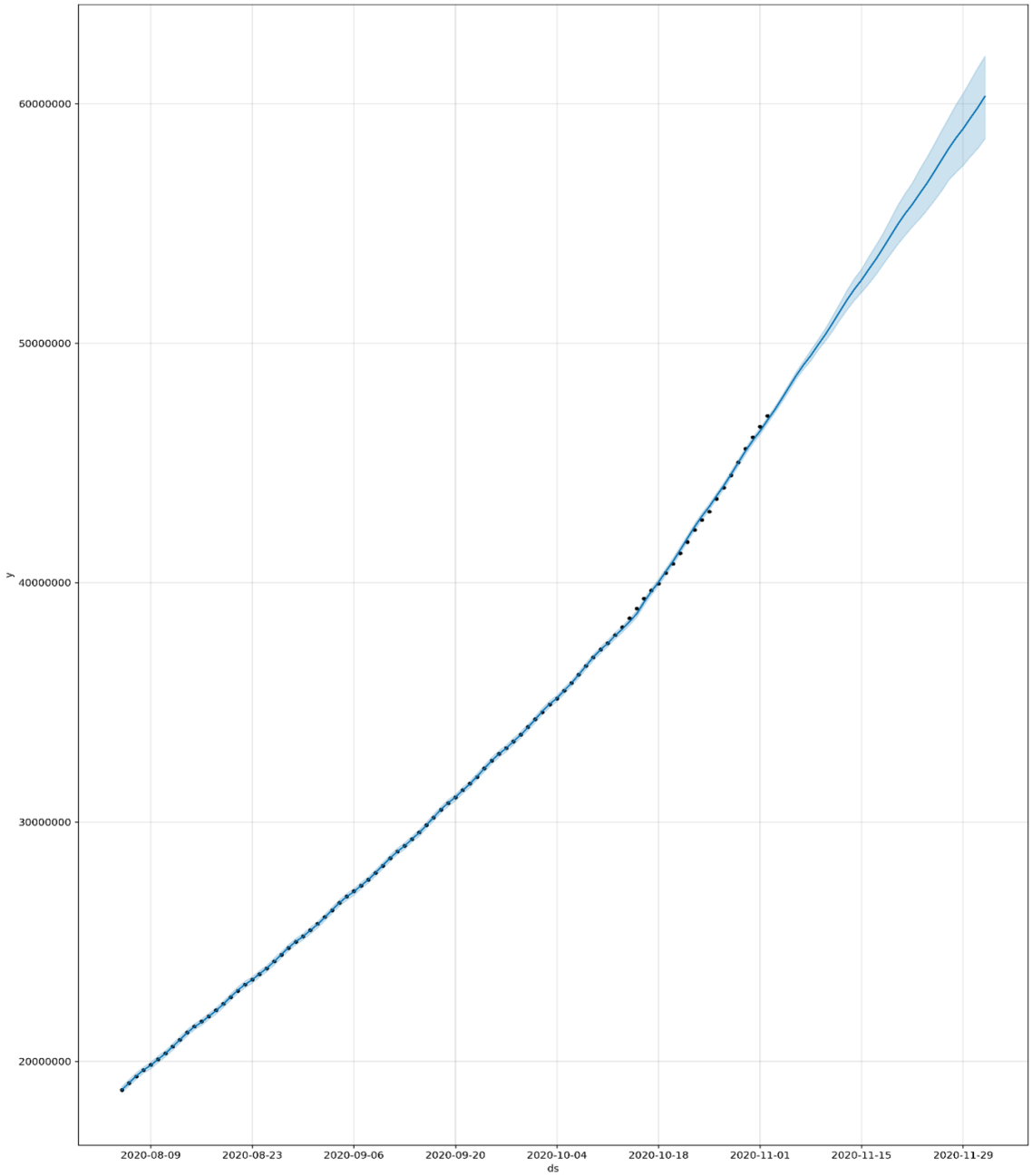

Figure 30. Forecasted Number of Cases globally

Figure 30 depicts the trend of the pandemic and also how the pandemic goes on a weekly basis. It does seem like the numbers go up during the end of a week and that might be because the virus doesn't show its effect immediately and it shows in a few days. So the

people get affected during the weekend due to the exposure outside and get positive later that week. It can also be inferred that a lot of people might get tested during the weekends and the positive test results turn up during the weekdays.

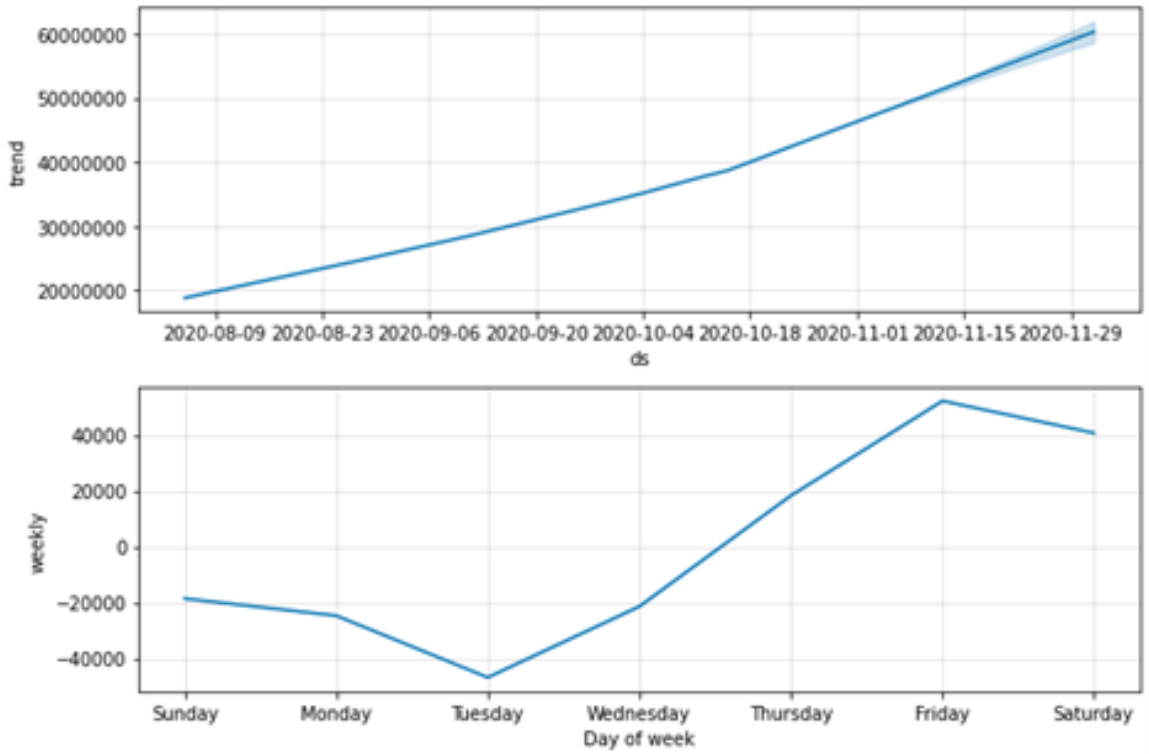

Figure 31. Trend of the pandemic for global data

Next up is the forecast for the number of deaths over the next thirty days and as figure 32 visualizes the numbers are still high as well and there doesn't seem to be a scope of a flat curve. As discussed earlier, there would be a higher number of deaths during the peak of an outbreak. Currently, the USA is facing another peak which might mean that the number of deaths might get high and even higher than the previous peak. The dotted points are mostly aligned with the curve and there does exist a margin of error and it is very small. The performance is discussed in the forthcoming section.

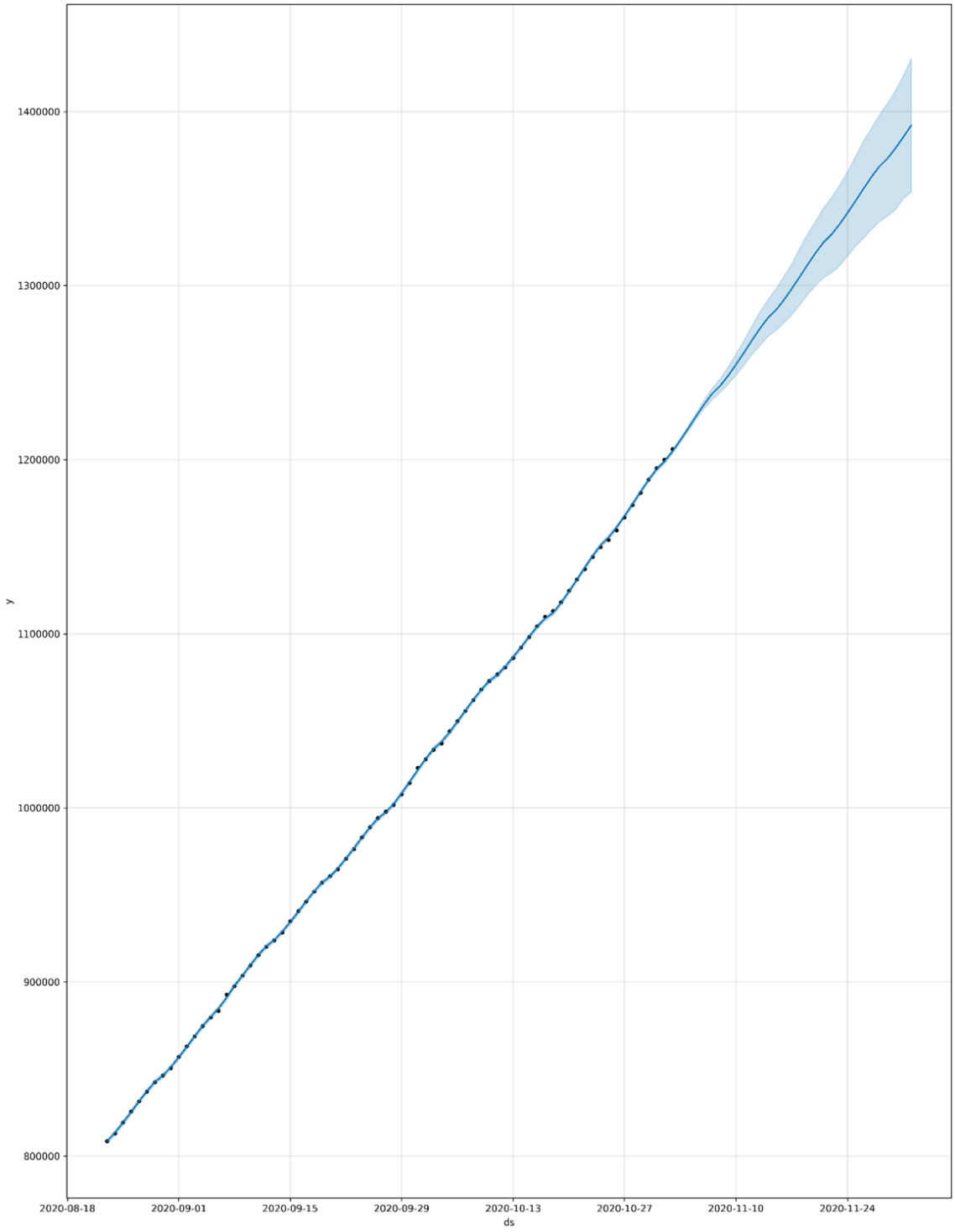

Figure 32. Forecasted Number of deaths globally

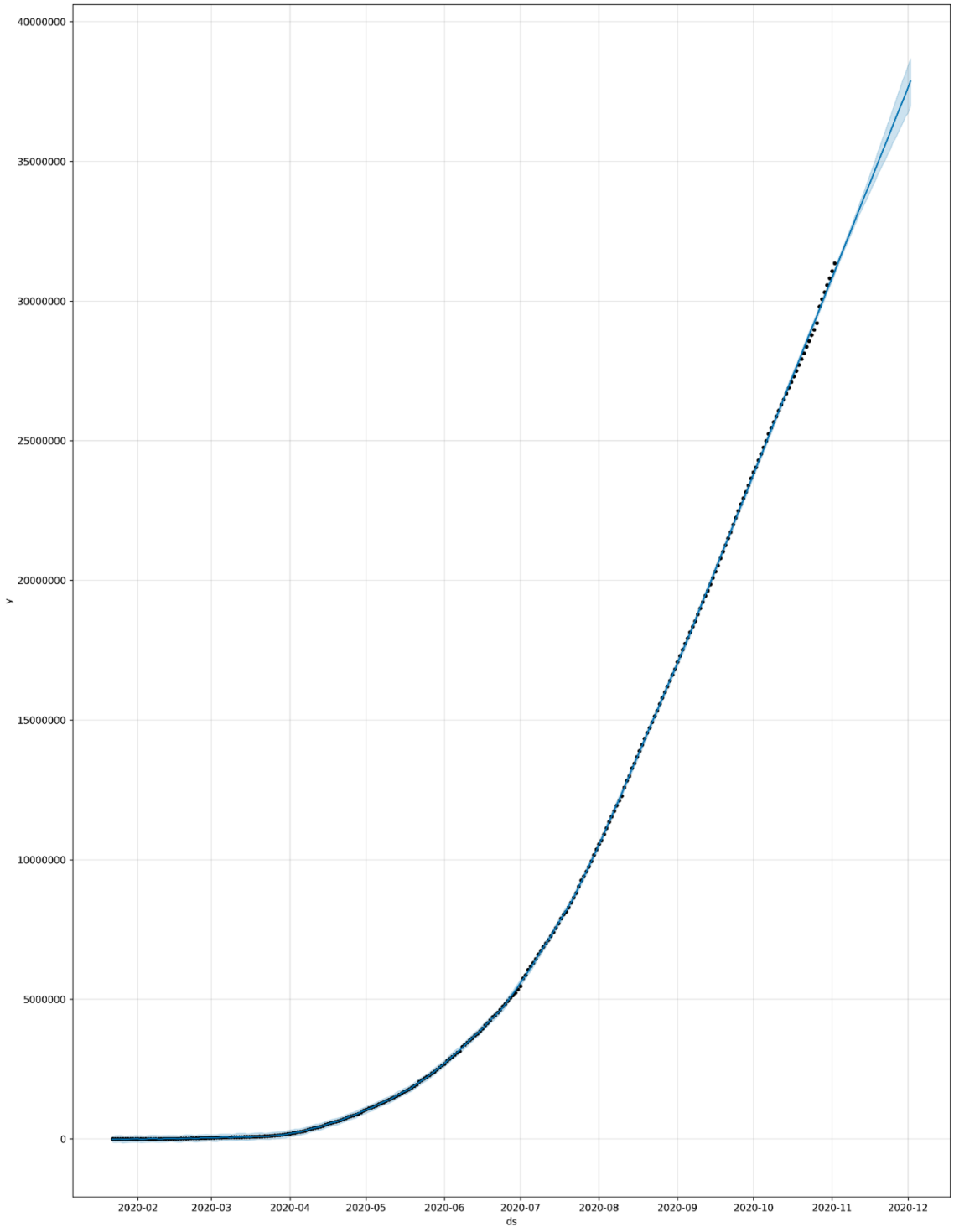

Figure 33. Forecasted number of recoveries globally

Figure 33 shows the forecasted curve for the number of recovery cases over the next thirty days and this is the only curve that needs to be as high as possible because the higher the curve is the lower the death rate would be. Again, the dotted points are well aligned within the curve but there are few points in the greater part of the curve that are outside which may be due to the rise and fall in the number of cases in recent times.

4.3.3 Forecasting the USA time-series data using FbProphet

In the earlier sections, the different aspects of the pandemic were discussed state-wise for the USA. This section discusses what the future would have for the USA in terms of the number of cases, number of deaths, and the number of recovered. Figure 35 depicts the curve moving further upward and there is no scope of the curve flattening. This is true because there has been a rapid increase in the number of daily cases in the USA in recent times. So it is going to be a rapid increase and it would be almost 11 million cases by the end of December which is rapid. This could be due to the election campaigns that have been happening for the past month. The rapid increase could be the result of the contact of too many people with the infected. So when a large group of people comes in contact with even one or two infected people there would be a huge increase in the cases, which might be the reason for the current rapidly growing trend.

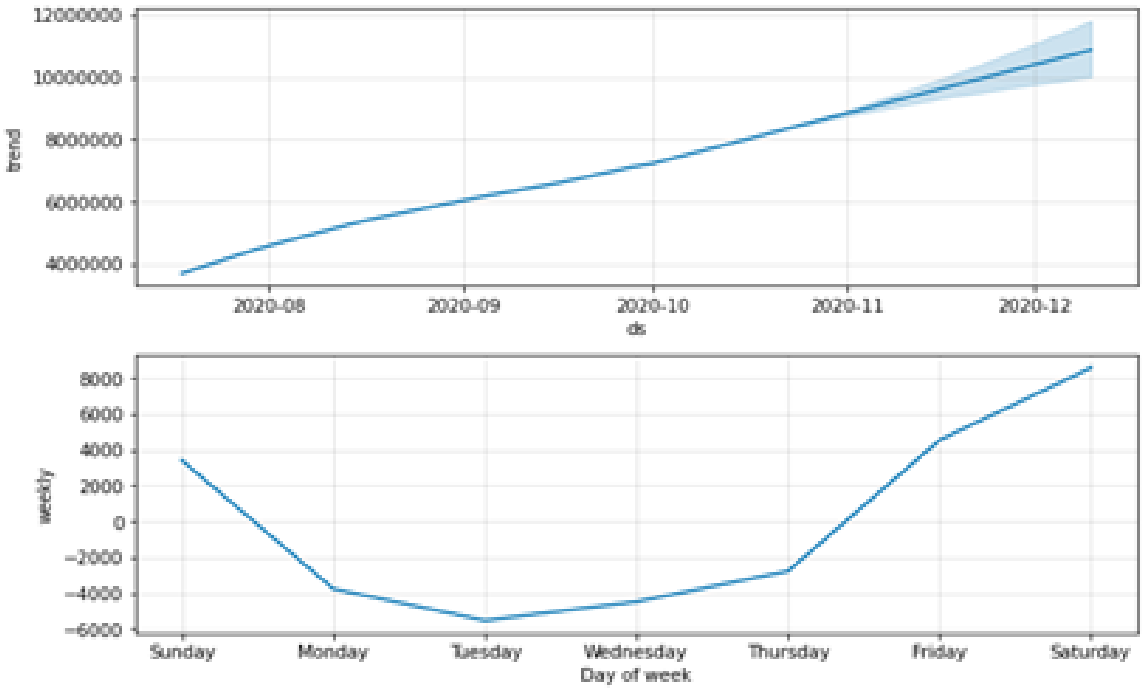

Figure 34. Trend of the number of cases in the USA

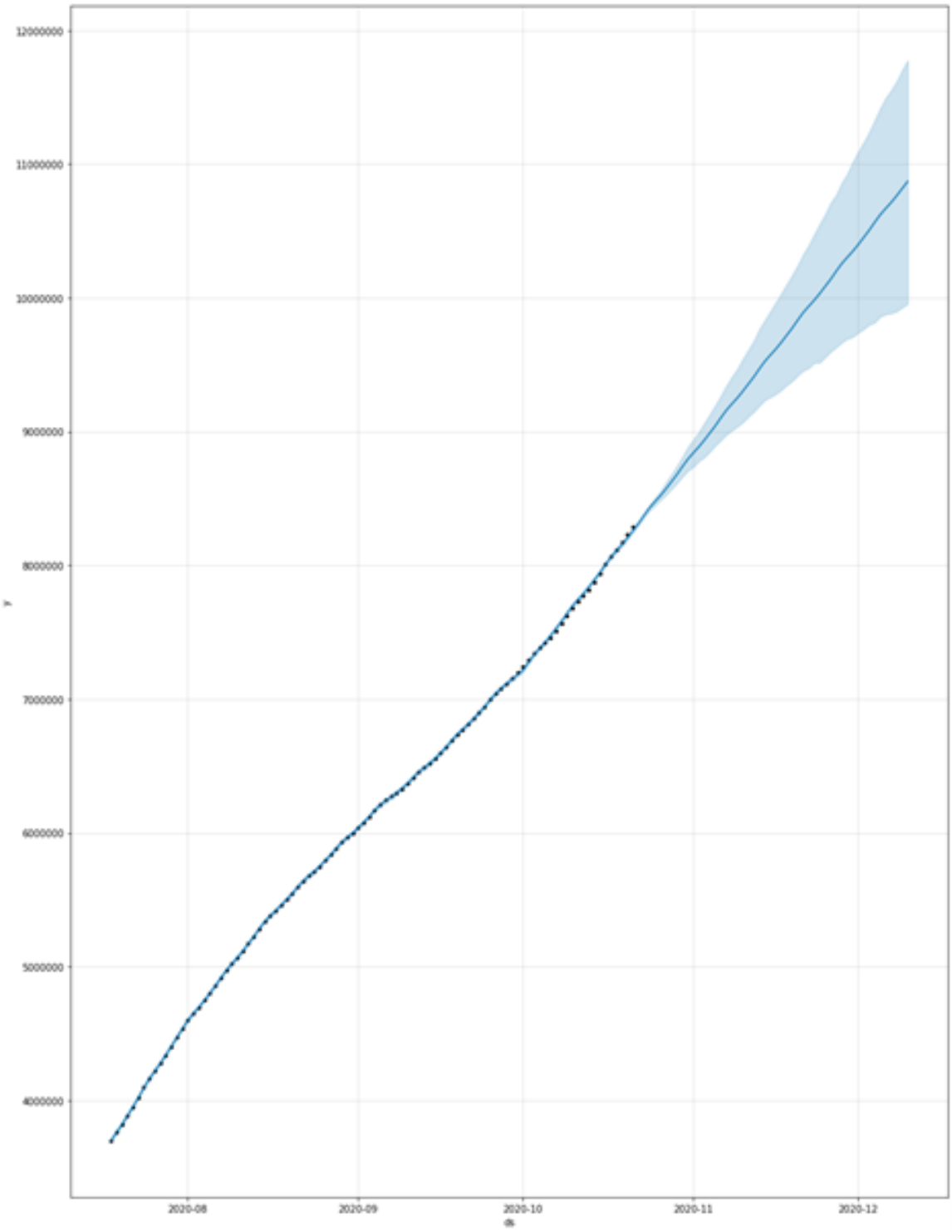

Figure 35. Forecasted number of cases in the USA

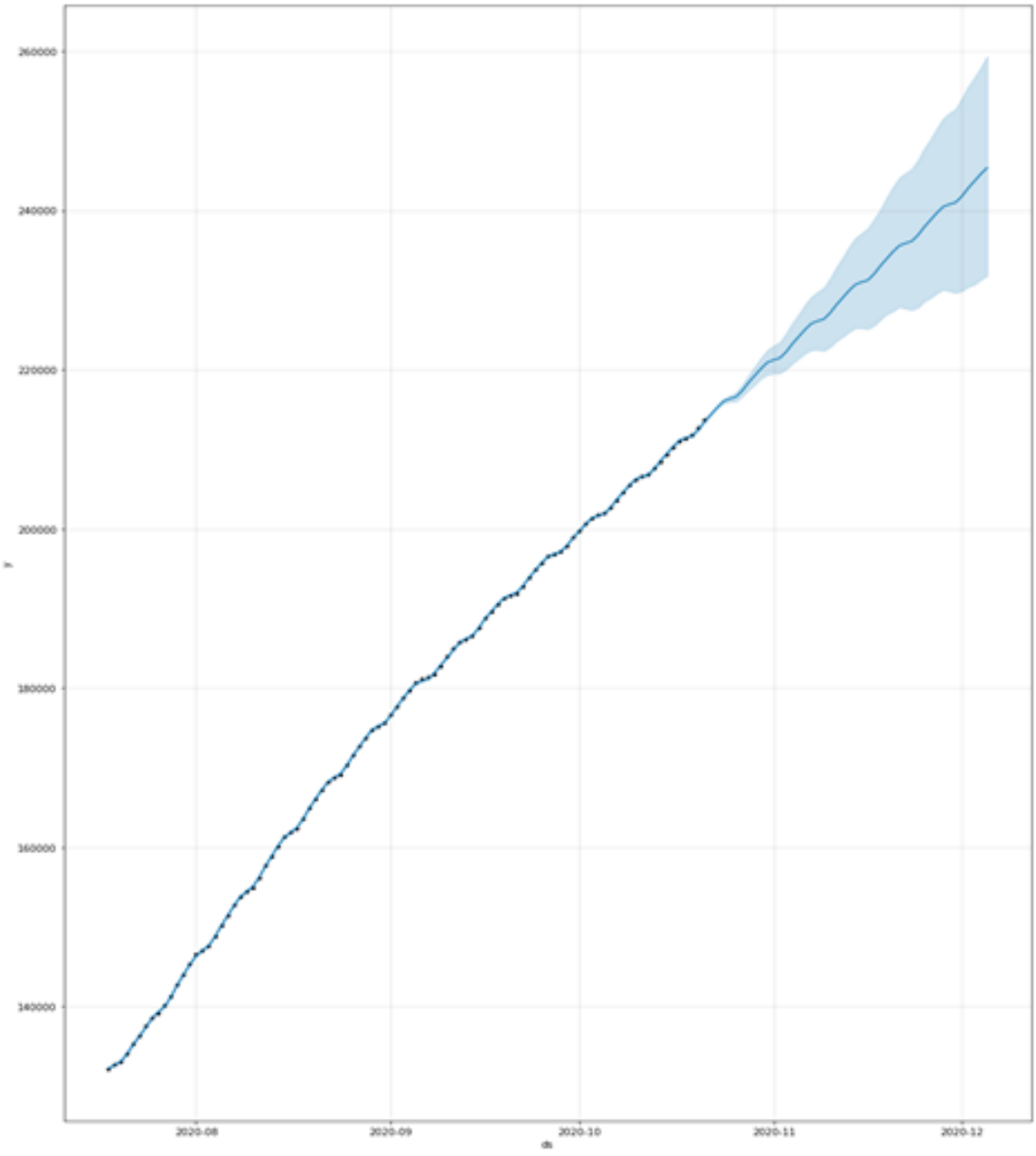

Figure 36. Forecasted number of deaths in the USA

Due to the instability in the number of daily cases in the USA (too high one week and pretty decent the other week), the forecasted curve for the number of deaths in the future is quite wiggly. Figure 36 shows how the number of deaths could cross around 250K by the end of December 2020. The USA is the country with the highest number of deaths and hence the recovery is also a little lower than the other top countries in COVID-19 cases. By the mid of

December 2020, there would be more than 4 million total recovered from the disease versus the 11 million predictions of the total number of cases. Figure 39 visualizes the forecast of daily positive cases in the USA over the next thirty days. Since the recent trend went from a low to a rapidly high number, the trend for the next few days would be a series of ups and downs. In the next 30 days, the peak it could go might be around 115K cases in a day. This might be true because the number of cases on 11/04/2020 was approx. 103K.

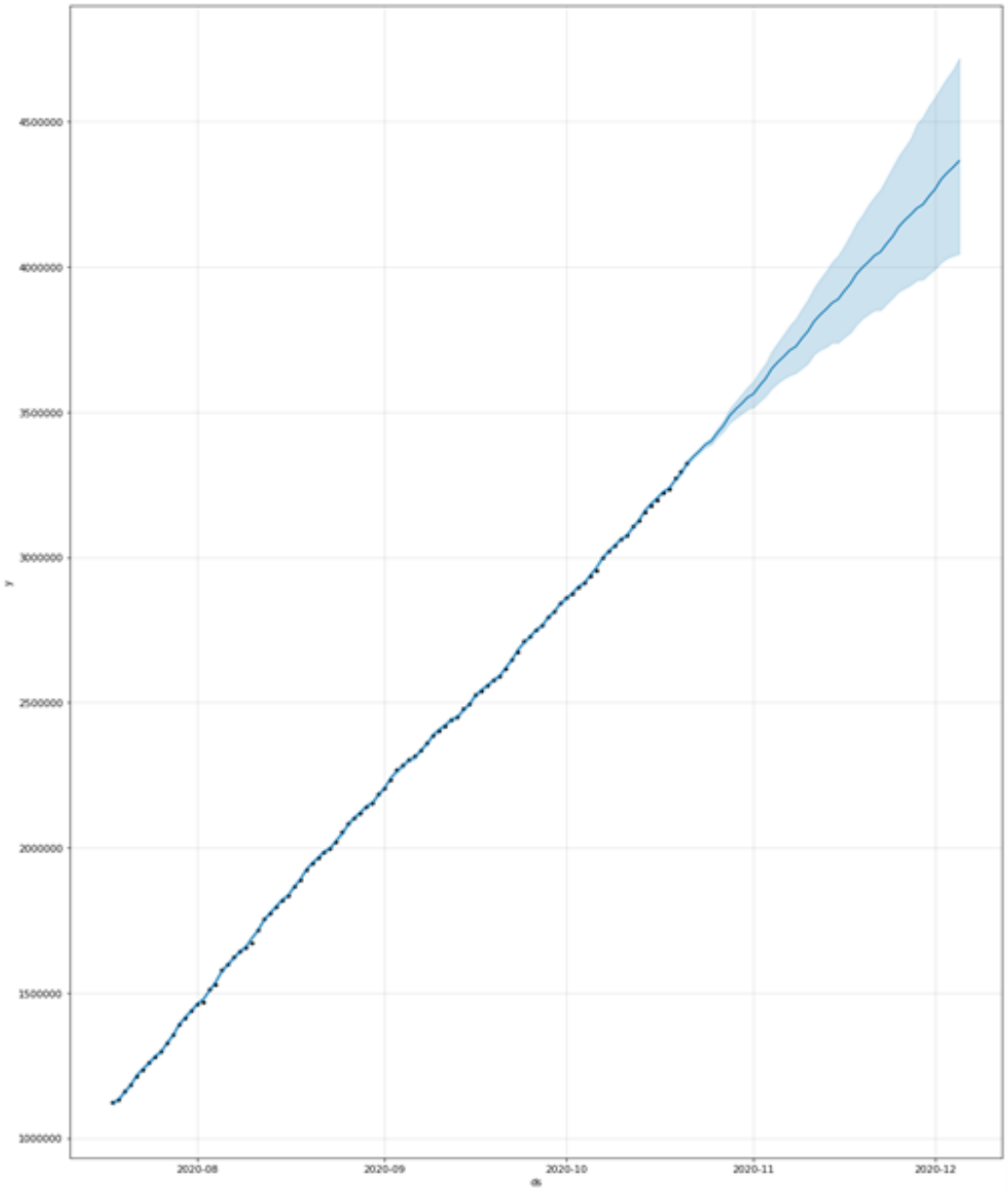

Figure 37. Forecasted number of recovery in the USA

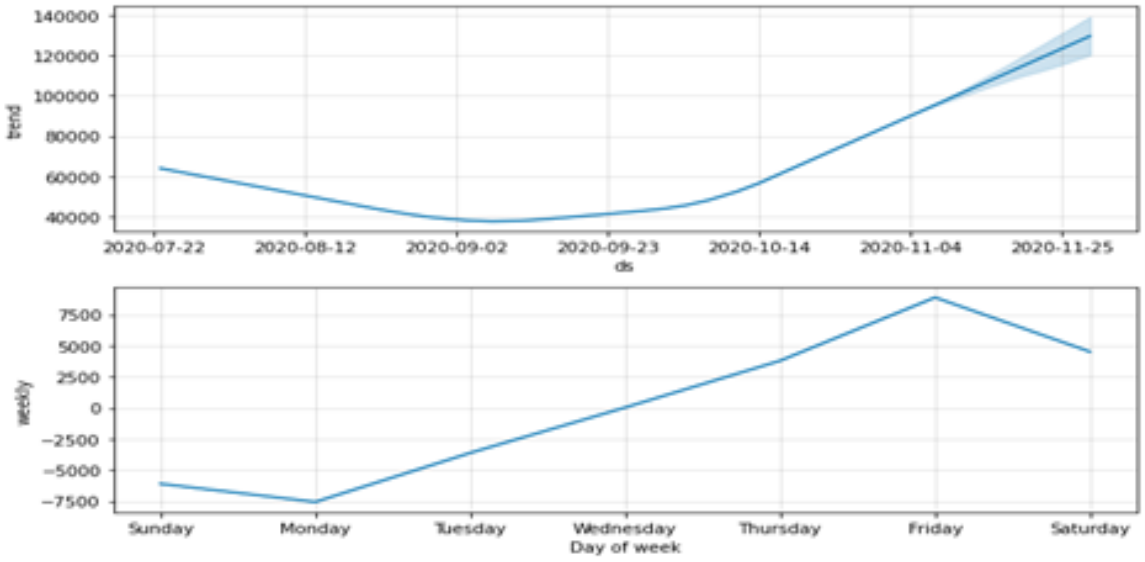

Figure 38. Trend of the daily increase in cases in the USA

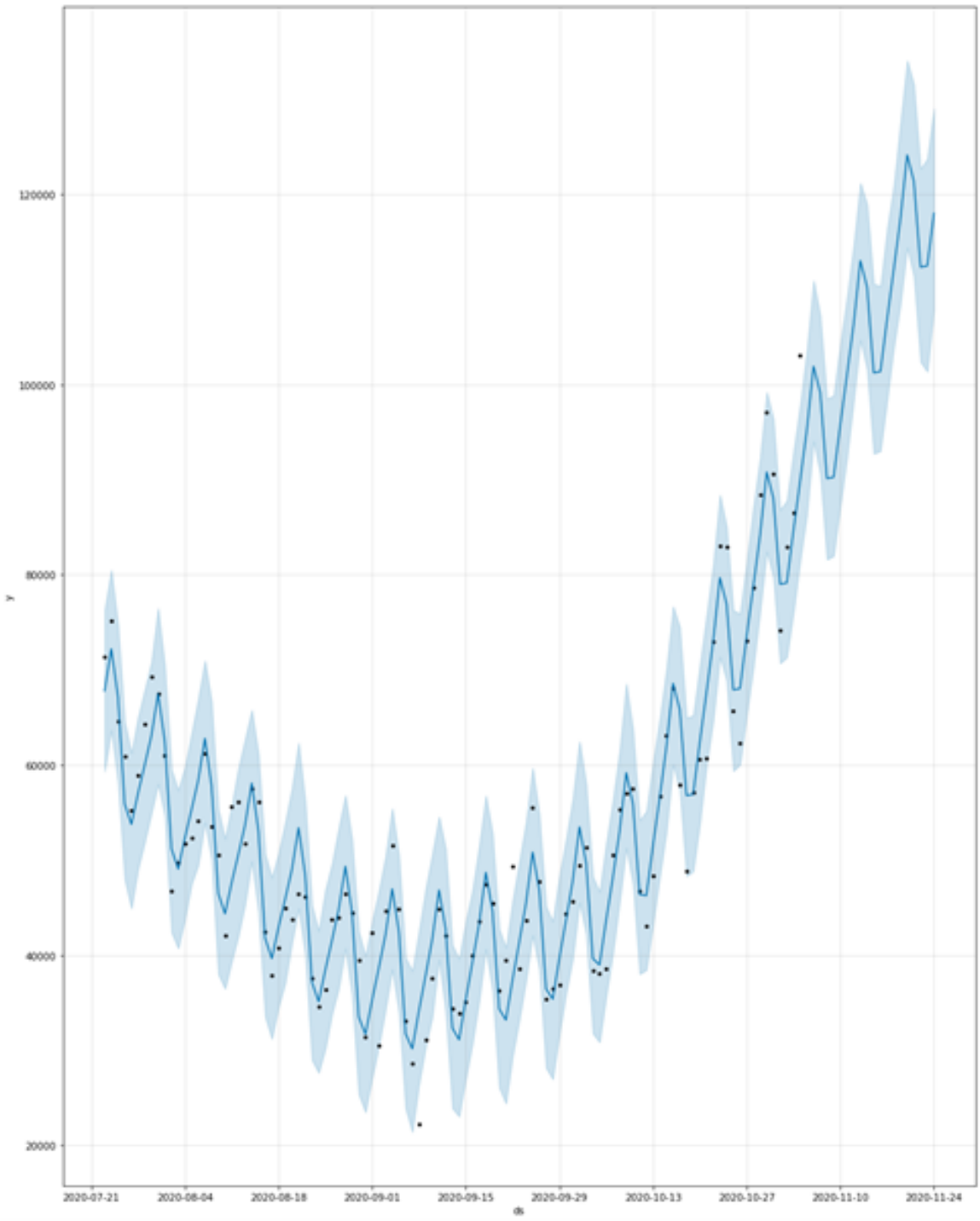

Figure 39. Forecasted Daily Cases in the USA

4.3.4 Autoregressive Integrated Moving Average- A time-series forecasting tool

ARIMA is an Autoregressive, Moving Average and Integration (I) model. The model is based on 'p' which is the order of autoregression, 'd' which is the order of differencing, 'q' which is the order of moving average (Wulff, 2017). ARIMA would fit only when the data mean and the standard deviation is constant. The differencing parameter d is a transformation variable to make the data mean and standard deviation constant(stationary data). The AR(p) model is given by.

$$z_t = \alpha + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + w_t \tag{1}$$

where zt1,zt2,ztp are the lags (historical values); 1,2,p are lag coefficients which are estimated values forecasted by the model; wt is the white noise.

$$\alpha = \left(1 - \sum_{i=1}^{p} \phi_i\right) \mu \tag{2}$$

where is mean of the process.

The MA(q) model is derived using the following:

$$z_t = \alpha + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q} \tag{3}$$

The three parameters p, q, and d are required for the forecasting.

"p" is the auto regressor which refers to a past value used to predict future value. "d" is the integration which is the amount of differencing and "q" is the moving average used to calculate intervals. To select the right p,d,q values we use the auto ARIMA function that would iterate through different p, d, and q values and come up with the best combination by achieving the best Auto Correlation Function. Autocorrelation is how the correlation between two data points changes as their separation changes. ACF is a function used to find a pattern in the data points.
The Auto ARIMA method uses AIC and BIC to find the best model for the given data points. The Akaike Information Criteria(AIC) and the Bayesian Information Criteria are used to measure the goodness of fit. The auto ARIMA method computes the model until the lowest AIC and BIC values have been achieved for a given set of data points.

4.3.5 Forecasting the USA time-series data using ARIMA

In the previous section, we saw forecasting using FbProphet. This section discusses the forecasting for USA state-wise data using the ARIMA model.

Table 3. ARIMA Model used

Data	ARIMA Model
USA-Daily Positive Increase	SARIMAX(3,1,3)
USA-Total Positive Cases	SARIMAX(3,2,2)
USA-Daily Death Increase	SARIMAX(3,0,1)
USA-Total Death Cases	SARIMAX(3,1,1)

Table 3 shows the different ARIMA models that were computed best for the respective data using the auto Arima function. The number represents the order, which are the values for p, d, and q discussed above.

Figure 40 visualizes the forecast of daily positive cases in the USA over the next thirty days. In the next 30 days, the surge could go might be around 120K cases in a day. The USA would see a surge that hasn't been seen before.

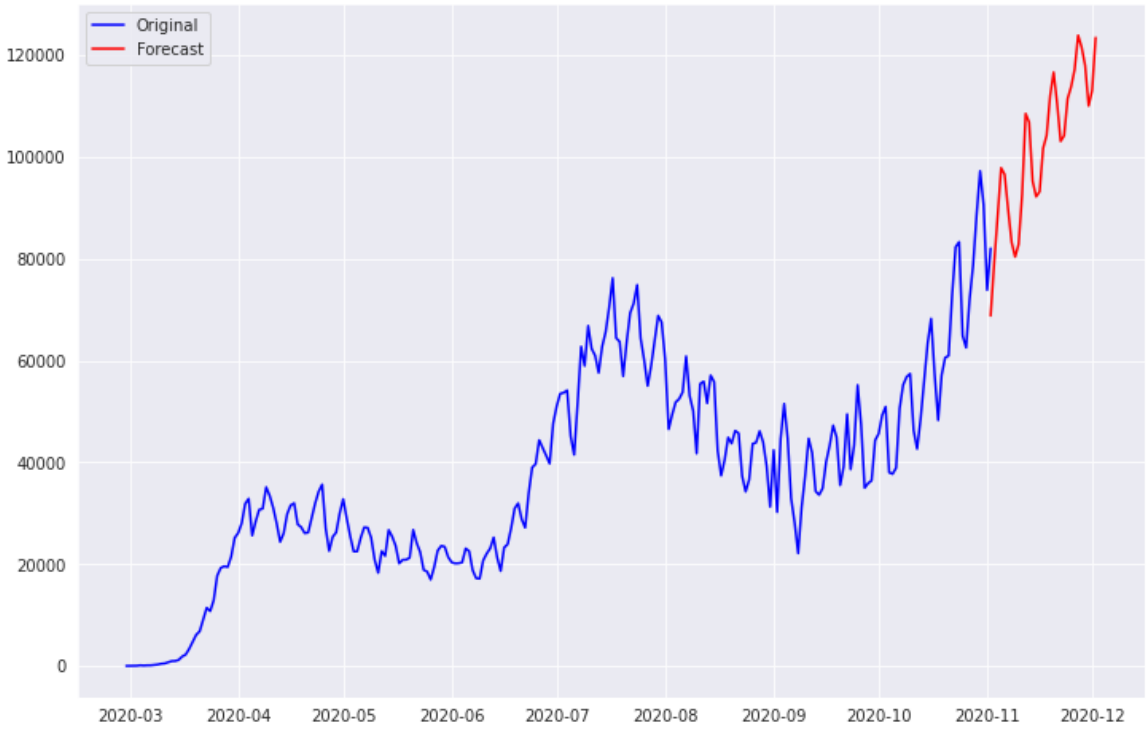

Figure 40. Forecasted daily positive increase in the USA

We know that a rapid increase in the number of cases would increase the number of deaths. So it would be necessary to forecast the death increase to get an idea of how the deaths might occur during this surge. Figure 41 shows that the forecasted number of deaths could get high in the next few days as we see a surge in the number of cases.

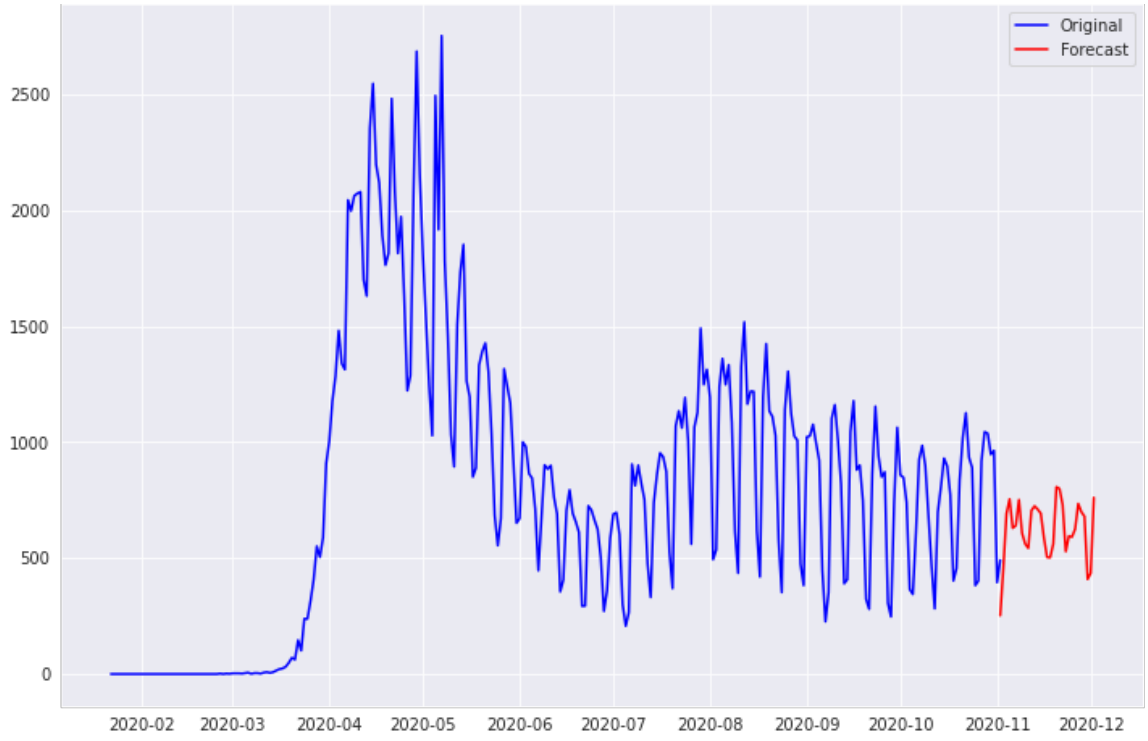

Figure 41. Forecasted daily increase in the number of deaths in the USA

Figure 42 depicts the curve moving further upward and there isn't a point where the curve might flatten. By December 2020, there might be around 12 million total COVID-19 cases in the USA.

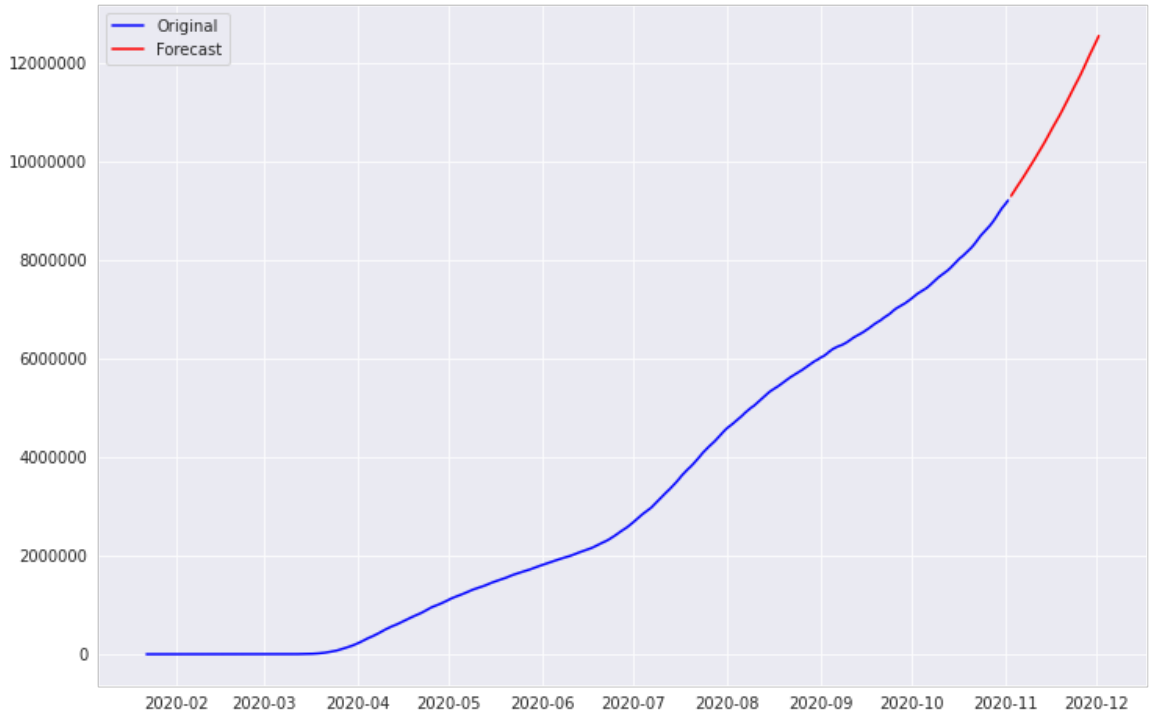

Figure 42. ARIMA Forecasted number of cases in the USA

Figure 43 shows how the number of deaths could reach around 250K by the end of December 2020. The USA is the country with the highest number of deaths and hence the recovery is also a little lower than the other top countries in COVID-19 cases.

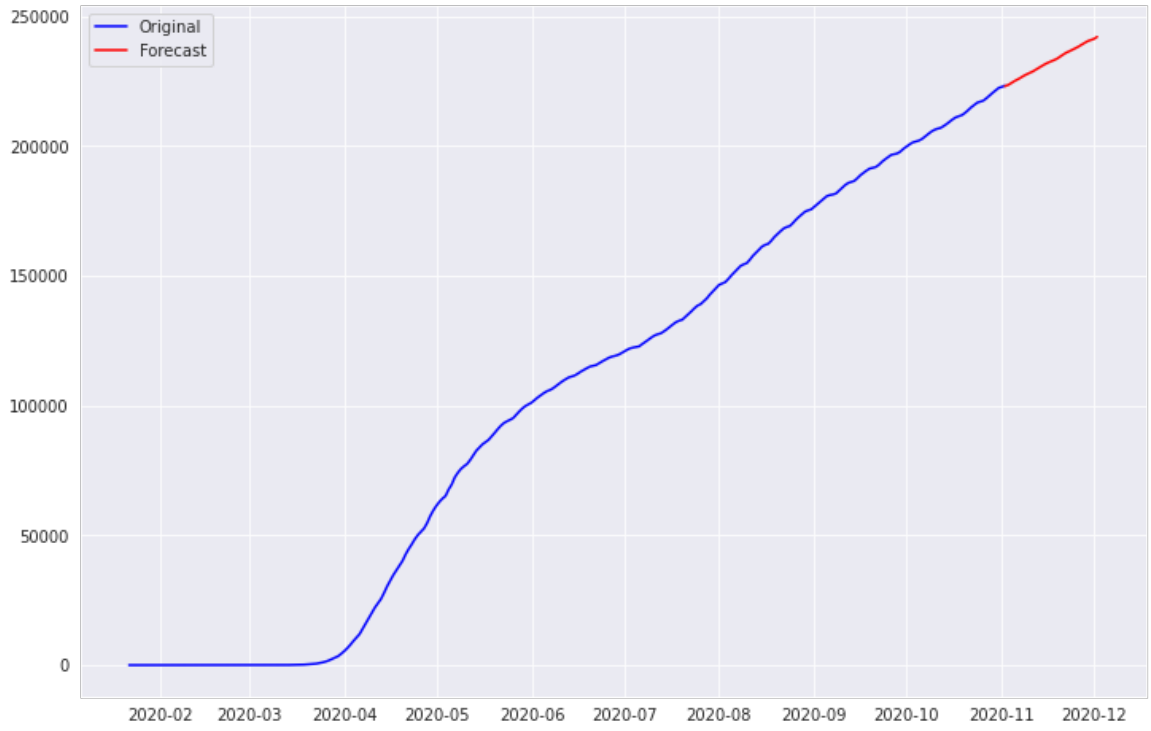

Figure 43. ARIMA Forecasted number of deaths in the USA

We have seen that both our models FbProphet and ARIMA have shown a forecast for a surge in the number of cases and deaths in the USA and the total positive cases have a steady increase in the USA.

4.3.6 Performance of the forecasting models

To evaluate the prediction models, we use the following statistical measures:

Mean Absolute Error (MAE): MAE is the measure of how a forecast might deviate from the actual value. It is the measure of error between the observed value and the forecasted value.

Table 4. Peformance Measures

Model Performance Evaluation				
Data	Model	MAE	RMSE	R-Squared
USA-Daily Positive Increase	SARIMAX(3,1,3)	4936.53	7449.68	0.96
USA-Daily Positive Increase	FbProphet	8946.01	11332.91	0.92
USA-Total Positive Cases	SARIMAX(3,2,2)	4328.53	6940.64	0.99
USA-Total Positive Cases	FbProphet	135154.51	235408.38	0.99
USA-Daily Death Increase	SARIMAX(3,0,1)	214.48	320.15	0.75
USA-Daily Death Increase	FbProphet	328.11	438.19	0.53
USA-Total Death Cases	SARIMAX(3,1,1)	214.48	320.15	0.99
USA-Total Death Cases	FbProphet	839.38	1502.99	0.99

Root Mean Square Error (RMSE): RMSE is the standard deviation of the prediction errors. It is the measure of how far the points deviate from the mean and how the residuals are spread out.

R-squared: It is the co-efficient of determination and it is the measure of how well-forecasted outcomes are replicated by the model, and it is based on the proportion of variation of outcomes explained by the model.

ARIMA vs FBProphet- USA daily cases: In this project, ARIMA and FBProphet have been used to predict future cases and deaths. ARIMA requires the data to be stationary for evaluation purposes. Dicky-fuller test was used to check the stationarity of the data and the data was not stationary. So the lag differencing approach was used to convert the data into a stationary form.

As we can see in figure 44, the ARIMA model very closely aligns with the observed value than the FBProphet model which is clearly inaccurate than the observed curve. It is clear that the ARIMA model SARIMAX(3,1,3) outperformed FBProphet. ARIMA has an RMSE of 7449.68 and FBProphet has an RMSE of 1132.91. The MAE for ARIMA is 4936.53 and for FBProphet it is 8946.01. From the results in table 4, we can clearly say that ARIMA has

outperformed the FBProphet model with respect in terms of error measures(lowest error value) i.e. MAE, RMSE for USA Daily cases data.

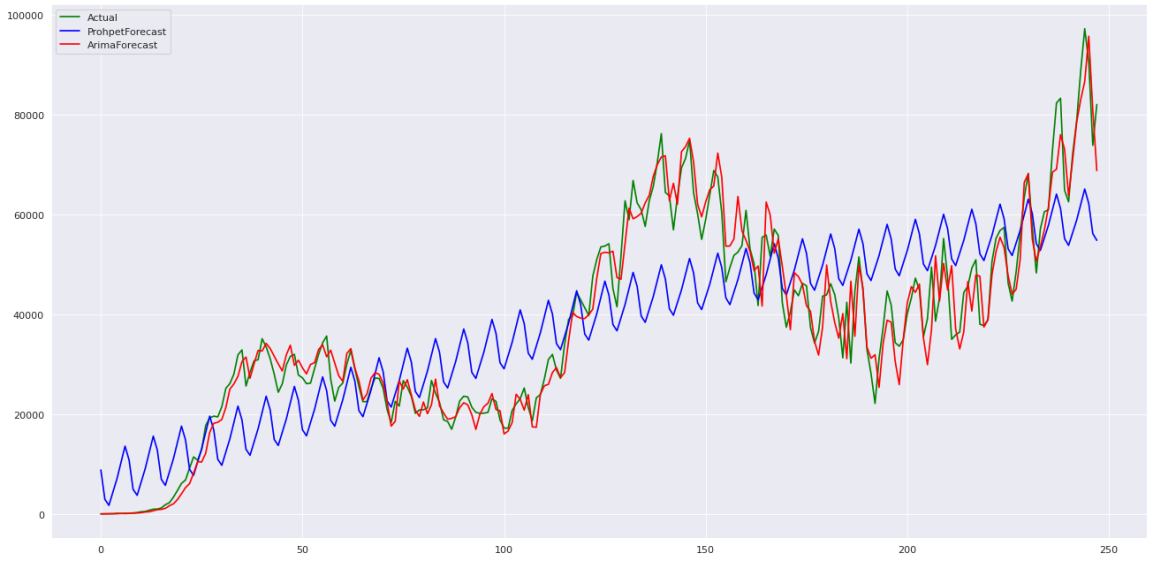

Figure 44. Daily Positive Increase in the USA- ARIMA Vs FBProphet Vs Obeserved value

ARIMA vs FBProphet- USA total positive cases: Here, in figure 45, we see that the two model-ARIMA and FBProphet closely align with the observed value but the FBProphet model has a slight deviation. So, It is clear that the ARIMA model SARIMAX(3,2,2) outperformed FBProphet in this case too. ARIMA has an RMSE of 69 0.64 and FBProphet has an RMSE of 235408.38. The MAE for ARIMA is 4328.53 and for FBProphet it is 135154.51. However, the r-squared score for both the models is 0.99 which means variance is explained by both the models. From the results in table 4, we can clearly say that ARIMA has outperformed the FBProphet model with respect in terms of error measures(lowest error value) i.e. MAE, RMSE for the USA total positive cases data.

ARIMA vs FBProphet- USA daily death cases: As we can see in figure 46, the ARIMA model very closely aligns with the observed value than the FBProphet model which is clearly inaccurate than the observed curve but aligns with the curve towards the end. It is clear that the ARIMA model SARIMAX(3,0,1) outperformed FBProphet. ARIMA has an RMSE of 320.15 and FBProphet has an RMSE of 438.19. The MAE for ARIMA is 214.48 and for FBProphet it is 328.01. From the results in table 4, we can clearly say that ARIMA has outperformed the FBProphet model with respect in terms of error measures(lowest error value) i.e. MAE, RMSE for USA Daily cases data. For this data, the r-squared value is quite low with 0.75 for ARIMA and 0.53 for FBProphet.

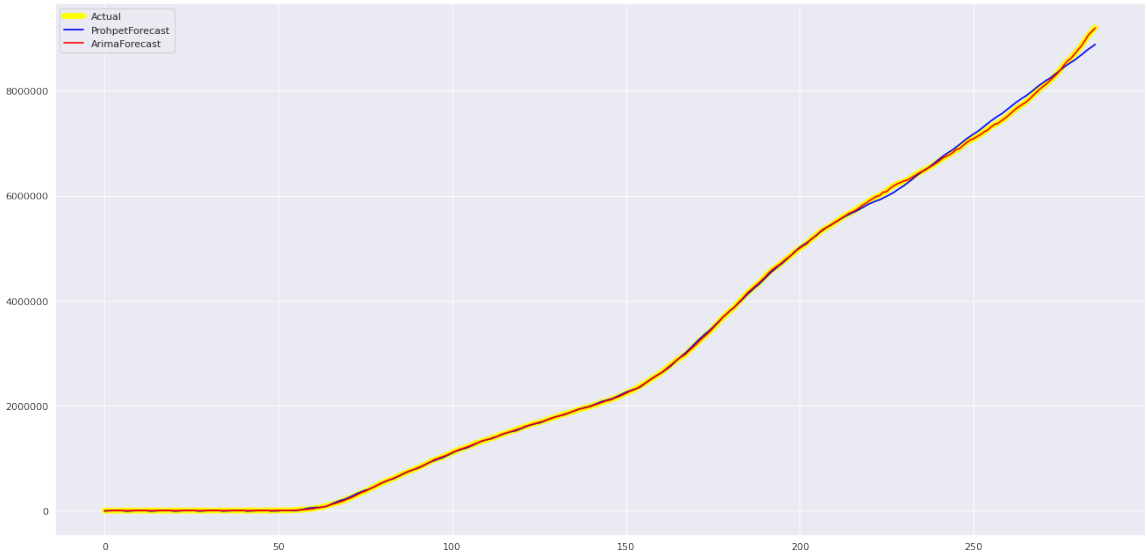

Figure 45. Total Positive Cases in the USA- ARIMA Vs FBProphet Vs Obeserved value

Figure 46. Total Daily Cases in the USA- ARIMA Vs FBProphet Vs Obeserved value

ARIMA vs FBProphet- USA total death cases: Here, in figure 47, we see that the two model-ARIMA and FBProphet almost perfectly align with the observed values. Of all the other data, this data has perfectly suited both models. ARIMA has an RMSE of 320.15 and FBProphet has an RMSE of 1502.99. The MAE for ARIMA is 214.48 and for FBProphet it is 839.388. However, the r-squared score for both the models is 0.99 which means variance is explained by both the models. Though both the models are a great fit with the observed

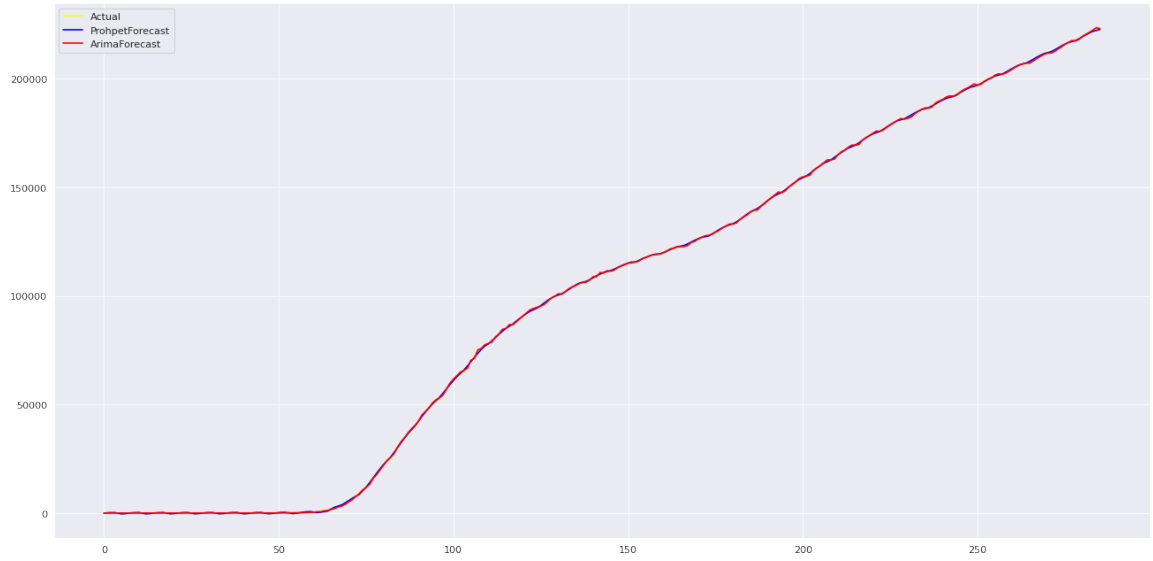

Figure 47. Total Death Cases in the USA- ARIMA Vs FBProphet Vs Obeserved value

values, from the results in table 4, we can clearly say that SARIMAX(3,1,1) has outperformed the FBProphet model in terms of error measures(lowest error value) i.e. MAE, RMSE for the USA total death cases data.

4.4 Visualizing COVID-19 data with any of the aspects of happiness report data that correlates

The final part of the project is merging the COVID-19 global data with the world happiness report to find a correlation between the aspects of the happiness report and the COVID-19 data. Upon merging the happiness report data with the COVID dataset, by joining using the country name common to both the world happiness report and the COVID data, the correlation was determined using the Pearson Correlation Technique. Correlation is a measure of the strength of a relationship between two quantitative dimensions. The correlation could be positive or negative depending upon the relationship. A positive relationship means that the two dimensions move in the same direction and in a negative correlation one increases and the other decreases or vice versa.

Perceptions of corruption of a country are the score given to a country based on the corruption present in the country. So a lower perception of corruption score means higher corruption in the country and a higher score means the country has lesser corruption cases. Figure 48 depicts a negative correlation between the positive cases around the globe and the corruption of a country. This means that the number of cases is high for the most corrupt countries (smaller corruption score). It is true because the countries get affected because

of corruption. Corruption in terms of COVID-19 may include corruption in health care, in the government, etc (International, 2020).

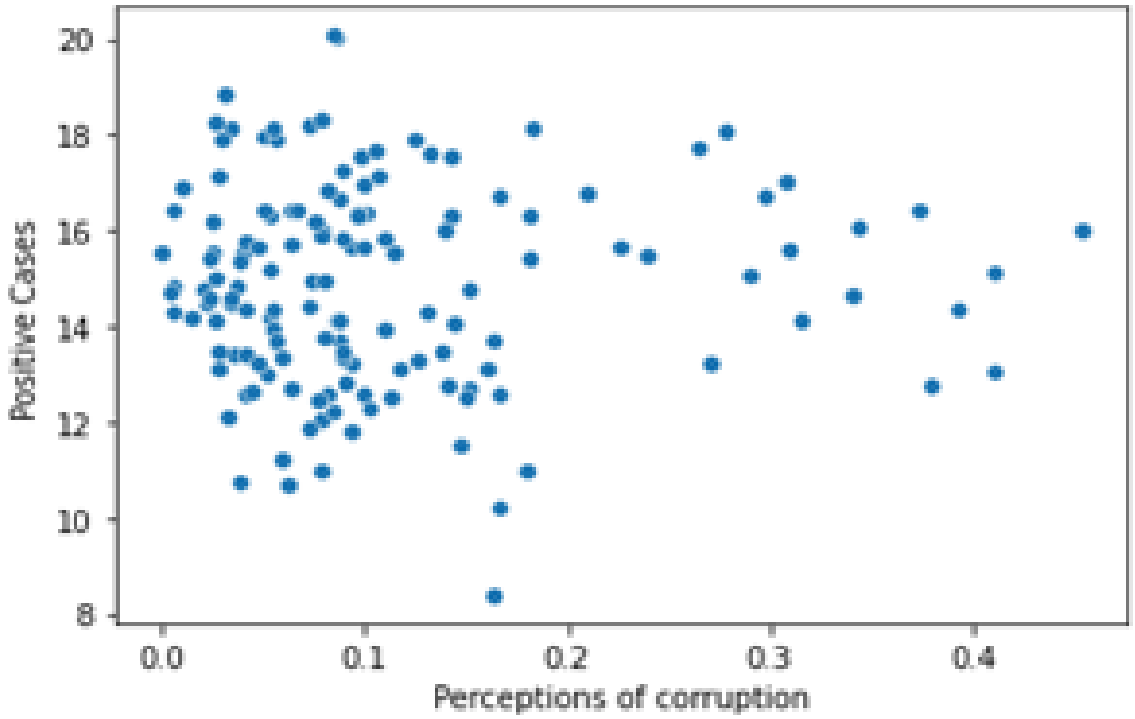

Figure 48. Positive Cases vs the perceptions of corruption

A lot of countries like India are having corruption like providing fake COVID-19 negative results to a person who is tested positive for the virus (Mahal, 2020). Corruptions like these are going to affect the positive increase in cases as the person with fake test results is free to move into the country and would affect many more. There is a positive correlation between the positive number of cases and the Gross Domestic Product per capita of a country. So as the GDP per capita increases, there are chances of more positive cases in that country. The USA, India, and Brazil with the highest number of COVID cases and with a high GDP per capita. This could mean that the GDP of the country is high so the number of testing centers could be more as the testing equipment are affordable for the country with a higher GDP and (Tauberg, 2020) supports this claim.

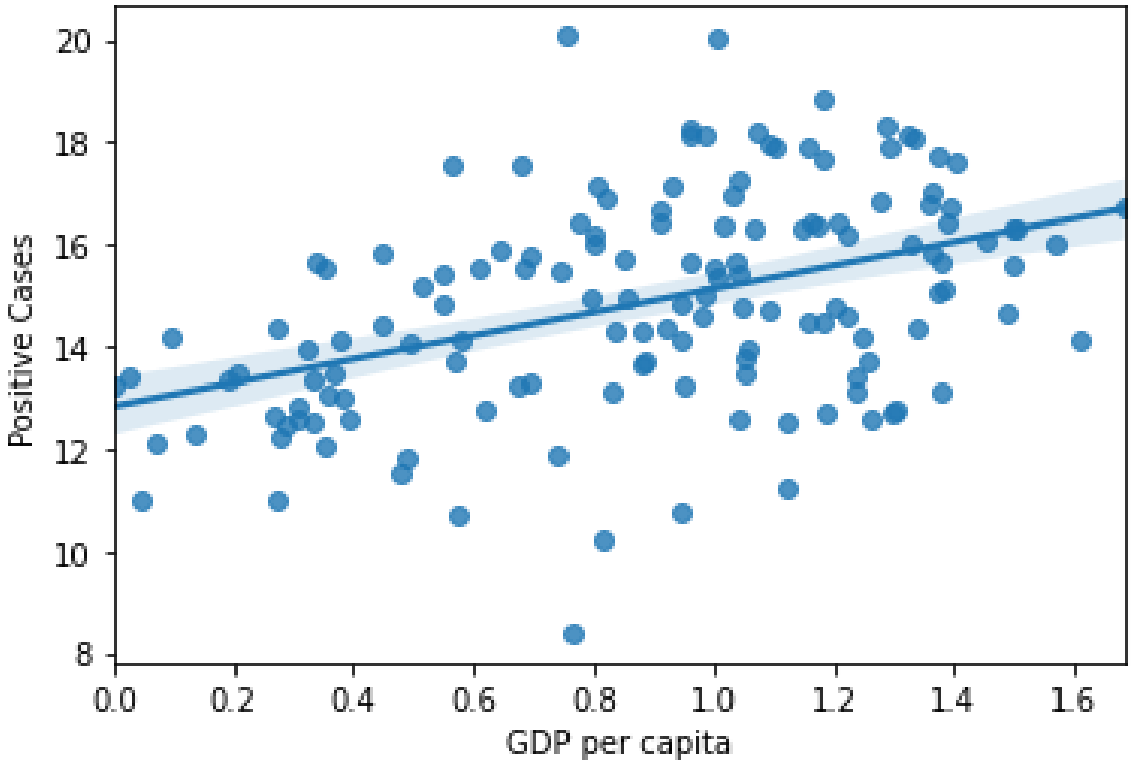

Figure 49. Positive Cases vs the GDP per capita

With the topic of GDP per capita, there is also a positive correlation between the recovery numbers and the GDP per capita of a country. As seen in figure 50, a positive correlation may be because a country with higher GDP per capita would have better sanitary systems, better hospitalization, and also the people would be able to afford a healthy living. For a country with a lower GDP, it would be difficult for the people to bear the medical expenses at a hospital.

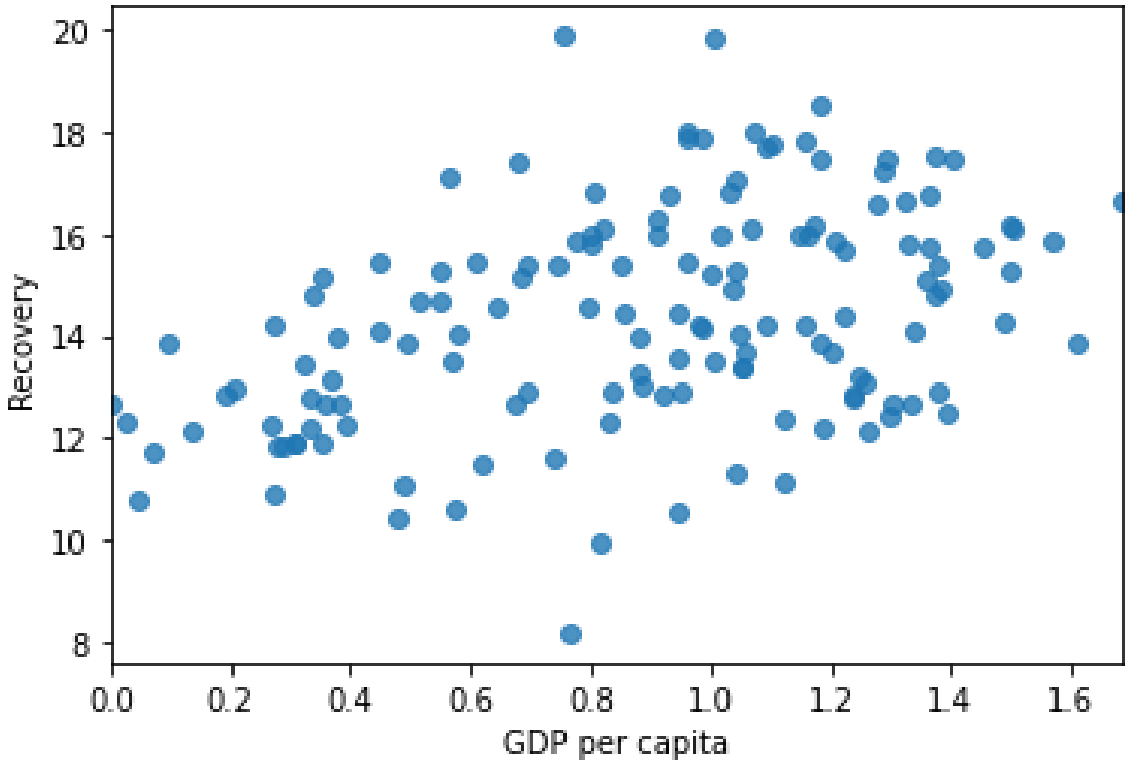

Figure 50. Recovery vs the GDP per capita

Also, a country with a higher life expectancy has higher recovery numbers as there exists a positive correlation between them based on Figure 51. This is maybe because of the obvious reasons as a healthy nation with healthy people would fight the virus easily. The healthy life expectancy factors might include good sanitation, a clean environment, etc, and again these would be better for a country with a better GDP.

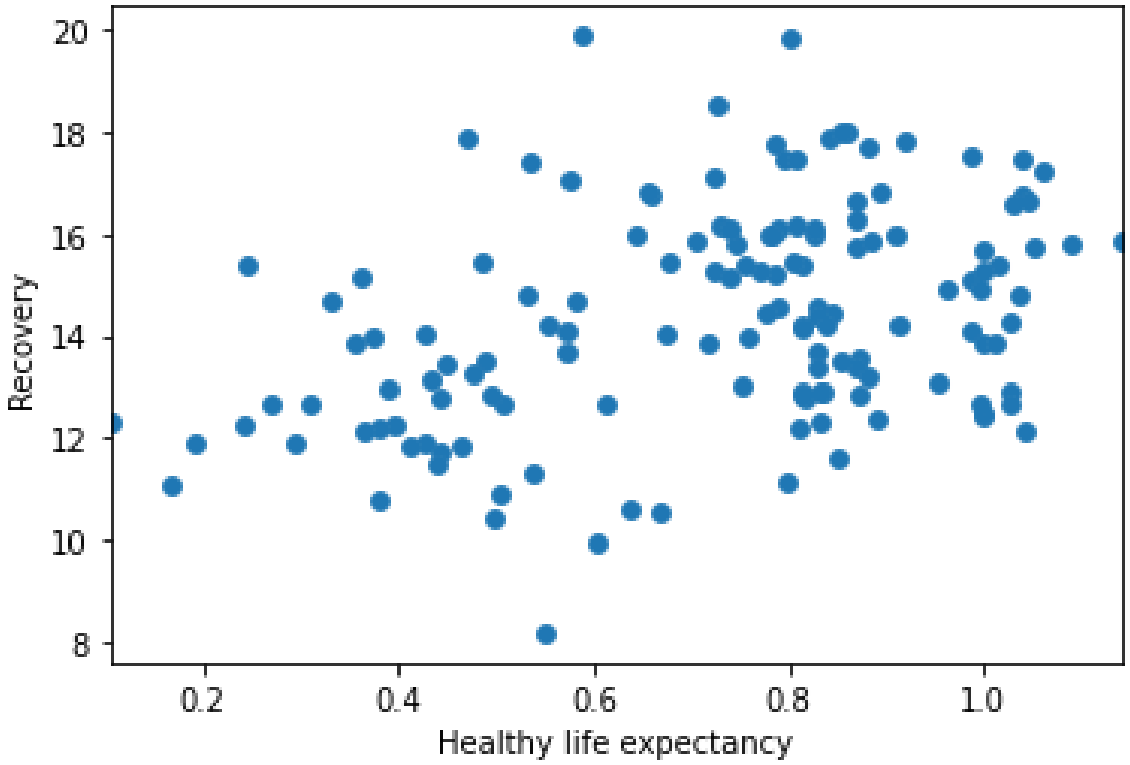

Figure 51. Recovery vs the Healthy life expectancy

Social support by the people would mean the cooperation and the support given by the people to the government of the people itself. Figure 52 depicts a positive correlation between social support by the people of a country and the recovery rate. During a pandemic, people must cooperate with the government and the policies of a government. For example, a study shows that wearing a mask could prevent the spread of the virus to a greater extent, although not completely. So, when people abide by the rules of the government, the spread could be avoided, and the recovery rate would be higher. As discussed in the earlier sections, the higher the peak in the outbreak, the lower the recovery rate would be and there would be a higher death rate.

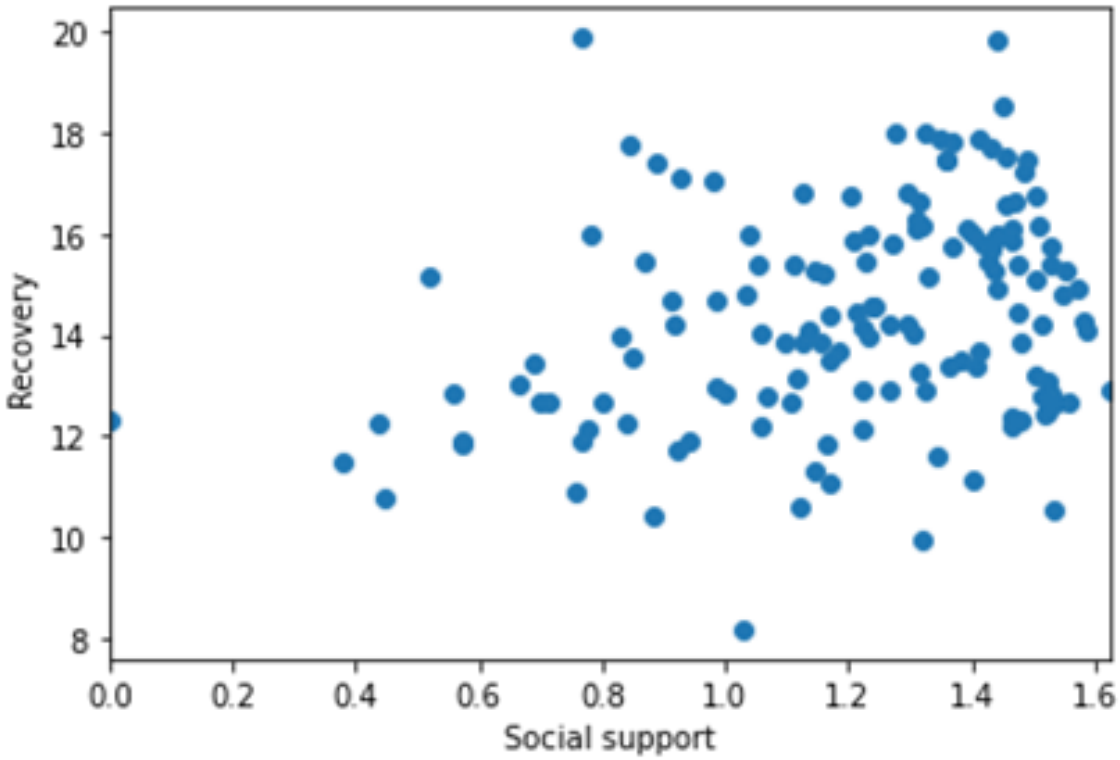

Figure 52. Recovery vs the Social support of the people

5 Conclusion

In this project, we have explored the COVID-19 data to get useful insights into the research questions mentioned in the Introduction. Based on the above experiments and visualizations, there have been more deaths during the peak of an outbreak. The USA is currently having a peak that has never been so high. If this peak of the outbreak goes out of control in the USA, there would be more deaths and the recovery rate would go down. We have also seen that the increase in hospitalization has an increase in the number of deaths. This answers the first research question that during a peak of the outbreak, a rapid increase in the number of cases would increase the hospitalization number. This creates a shortage of ventilators, hospital beds, and equipment and that causes an increase in the number of deaths during the peak. From the data visualizations, we infer that there is a positive relationship between the high number of deaths and high hospitalization rates during a spike of the outbreak.

We have done an analysis and prediction study of the disease using two popular forecasting models- ARIMA and FBProphet. Based on the forecast by the two models, there is going to be a rapid rise in the number of cases in the USA. This would also increase the number of deaths during the period. The increase is going to be rapid and it would be high that the USA has never seen before during this pandemic as there's going to be more than 100K cases a day. Similarly, the number of deaths forecasted is also going to be high and it would be around 1K a day and this is similar to the finding that the number of deaths increases rapidly during a surge. This answers the second research question and there is going to be a new peak in the USA which would also reflect on the global data. So, hospitals and health workers have to be even more prepared for this as we have seen how the hospitalization increase has also increased the death rate.

For all the data (USA Total Positive Cases, USA Daily Cases, USA Death Cases, and USA Total Death Cases), ARIMA has performed better compared to FBProphet on the scale of MAE, RSME error metrics, and also R-Squared measure. Out of all these models, the ARIMA model performed the best for the USA Total Death cases data with the lowest RSME, MAE, and the highest R-Squared value.

Finally, we saw that other factors could cause a rise in cases and reduce the recovery rate by correlating the COVID-19 data with the World Happiness Report. We saw how the number of positive cases for a country could increase with the corruption present in a country and we saw examples supporting the claim on how the COVID-19 results are bribed, the GDP per capita of a country could also increase the positive cases. On the other hand, the recovery rate also increases with the increase in the GDP per capita as a country with higher GDP per capita would have a higher GDP which would have better sanitary systems, hospitalization, and a healthy lifestyle. The recovery numbers also positively correlate with the healthy life

expectancy of a nation as there isn't a vaccine and a healthy body should fight the virus. Furthermore, social support by people is also one of the features that positively correlated with the recovery rate. This means that the people of a country must abide by the rules of the government and also help fellow beings. This also answers the final research question and we have discussed a few hidden factors associated with the rise in the number of cases and the recovery cases. In the future, we may have outbreaks, and during those outbreaks, the government must see to it that there is no corruption as the results tell us how corruption also plays a major role here. And, we as the people of the nation must abide by the rules and policies of the government to prevent the spread of the virus. Social support is most important during one of such pandemics. These apply to the current pandemic too as it is not over yet.

5.1 Limitations

- Since this an ongoing pandemic, we wouldn't be able to jump to a conclusion as we do not know what the future has got for us. The forecast is just a rough estimate of numbers for the future. To provide accurate numbers, it is best to forecast with an updated dataset for the future. The models used in this project have performed well but it limits our study to the effectiveness of the models, which can be further improved using an ensemble of multiple prediction models.

5.2 Challenges

While dealing with an ongoing pandemic it is difficult to conclude with an accurate statement. A lot of studies on COVID-19 that had concluded that the pandemic would be over by the end of this year haven't been accurate.

5.3 Future Work

- The forecasting results can be improved by taking various variables into account like health systems, hospital data, patient history, etc. using some deep learning and artificial intelligence techniques.
- More relevant datasets could be merged with the COVID-19 dataset and the unknown factors could be explored.

References

BEGLEY, S. (2020). Its difficult to grasp the projected deaths from covid-19. heres how they compare to other causes of death. https://www.statnews.com/2020/04/09/its-difficult-to-grasp-the-projected-deaths-from-covid-19-heres-how-they-compare-to-other-causes-of-death/

CDC. (2020a). Provisional death counts for coronavirus disease 2019 (covid-19). centers for disease control and prevention. https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm

CDC. (2020b). Severe acute respiratory syndrome (sars). centers for disease control and prevention. https://www.cdc.gov/sars/about/faq.html

Covid-19 coronavirus pandemic. (2020). https://www.worldometers.info/coronavirus/

Dixit, R. (2020). Analysis of covid-19 impact using data visualization. *International Journal for Research in Applied Science and Engineering Technology*, *8*. https://doi.org/10.22214/ijraset.2020.5132

Elegant, N. X. (2020). Ventilators are key to preventing coronavirus deathsbut does the world have enough of them? https://fortune.com/2020/03/17/coronavirus-ventilator-shortage/

Indolfi, C., & Spaccarotella, C. (2020). The outbreak of covid-19 in italy fighting the pandemic.

International, T. (2020). Citizens report covid-19 corruption. https://www.transparency.org/en/citizens-report-covid-19-corruption

Kaggle. (2016). World happiness report. https://www.kaggle.com/unsdsn/world-happiness

Khanam, F., Nowrin, I., & Mondal, M. R. H. (2020). Data visualization and analyzation of covid-19. *Journal of Scientific Research & Reports*, *26*. https://doi.org/10.9734/JSRR/2020/v26i330234

Luo, L., Luo, L., Zhang, X., & He, X. (2017). Hospital daily outpatient visits forecasting using a combinatorial model based on arima and ses models. *BMC Health Services Research*, *17*. https://doi.org/10.1186/s12913-017-2407-9

Lutz, C. S., Huynh, M. P., Schroeder, M., Anyatonwu, S., Dahlgren, F. S., Danyluk, G., Fernandez, D., Greene, S. K., Kipshidze, N., Liu, L., Mgbere, O., McHugh, L. A., Myers, J. F., Siniscalchi, A., Sullivan, A. D., West, N., Johansson, M. A., & Biggerstaff, M. (2019). Applying infectious disease forecasting to public health: A path forward using influenza forecasting examples. *BMC Public Health*, *19*. https://doi.org/https://doi.org/10.1186/s12889-019-7966-8

Lyla, Y. (2019). A quick start of time series forecasting with a practical example using fb prophet. https://towardsdatascience.com/a-quick-start-of-time-series-forecasting-with-a-practical-example-using-fb-prophet-31c4447a2274

Mahal, P. S. (2020). Agents taking bribe of 3,500 each from nris for early, fake covid test report in punjabs moga civil hospital. https://www.hindustantimes.com/cities/agents-taking-bribe-of-3-500-each-from-nris-for-early-fake-covid-test-report-in-punjab-s-moga-civil-hospital/story-MBWKRZp43mdvYlpZhQ234M.html

Myers, M., Rogers, D., Cox, J., Flahault, A., & Hay, S. (2000). Forecasting disease risk for increased epidemic preparedness in public health. *Adv Parasitol*, *47*. https://doi.org/10.1016/s0065-308x(00)47013-2

News, A. (2020). How california lost control over covid-19 despite early successes. https://abcnews.go.com/Health/california-lost-control-covid-19-early-successes/story?id=72008022

Project, T. C. T. (2020). The covid tracking project. https://covidtracking.com/

Rogers, L. (1925). Climate and disease incidence in india, with special reference to leprosy,phthisis,pneumonia and smallpox. *Journal of State Medicine*, *33*.

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). Covid-19 infection: Origin, transmission, and characteristics of human coronaviruses. https://doi.org/10.1016/j.jare.2020.03.005

State, N. Y. (2020). Continuing temporary suspension and modification of laws relating to the disaster emergency. https://www.governor.ny.gov/news/no-20214-continuing-temporary-suspension-and-modification-laws-relating-disaster-emergency

Tauberg, M. (2020). Comparing coronavirus economic and health impacts across nations. https://towardsdatascience.com/comparing-coronavirus-economic-and-health-impacts-across-nations-ef8e4ea52741

University, J. H. (2020). Novel coronavirus (covid-19) cases. https://github.com/CSSEGISandData/COVID-19

WHO. (2020a). Middle east respiratory syndrome coronavirus (mers-cov). world health organization. https://www.who.int/emergencies/mers-cov/en/

WHO. (2020b). Timeline of whos response to covid-19. https://www.who.int/news/item/29-06-2020-covidtimeline

Wikipideia. (2020). Covid-19 pandemic lockdown in india. https://en.wikipedia.org/wiki/COVID-19_pandemic_lockdown_in_India#:~:text=To%5C%20control%5C%20the%5C%20spread%5C%20of%5C%20coronavirus%5C%20outbreak%5C%

20in%5C%20India.&text=On%5C%2024%5C%20March%5C%202020%5C%2C%5C%20the,COVID%5C%2D19%5C%20pandemic%5C%20in%5C%20India

Wulff, S. S. (2017). Time series analysis: Forecasting and control, 5th edition. *Journal of Quality Technology*, *49*(4), 418–419. https://doi.org/10.1080/00224065.2017.11918006

Zunic, E., Korjeni, K., Hodi, K., & Donko, D. (2020). Application of facebook's prophet algorithm for successful sales forecasting based on real-world data. *International Journal of Computer Science and Information Technology*, *12*. https://doi.org/10.5121/ijcsit.2020.12203