COVID-19 Data Visualization, Growth Factor and Predictions

Mihir Patel

April 1, 2020

1.0 Introduction

1.1 Background

COVID-19 is a respiratory illness caused by a new virus. Symptoms include fever, coughing, sore throat, and shortness of breath. The virus can spread from person to person, but good hygiene can prevent infection. COVID-19 may not be fatal, but it spreads faster than other diseases, like a common cold. Every virus has a Basic Reproduction number (R0), which implies how many people will get the disease from an infected person. As per the initial research work, the R0 of COVID-19 is 2.7. Currently, the goal of all scientists around the world is to "Flatten the Curve." COVID-19 has an exponential growth rate around the world, which we will see in the notebook ahead. Flattening the curve typically implies that even if the number of confirmed cases is increasing, but the distribution of those cases should be over longer timestamp. To put it in simple words, if say suppose COVID-19 is going infect 100K people, then those many people should be infected in 1 year but not in a month. The sole reason to Flatten the Curve is to reduce the load on the Medical Systems to increase the focus of Research to find Medicine for the disease.

Every pandemic has four stages:

Stage 1: Confirmed Cases come from other countries

Stage 2: Local Transmission

Stage 3: Communities impacted with local transmission

Stage 4: Significant Transmission with no end in sight

Other more robust ways to tackle the disease are to cross-border shutdown, contact tracing, and quarantine people as necessary.

1.2 Problem Statement

The objective of this report is to study the COVID-19 outbreak with the help of some basic visualizations' techniques. Comparison of disease spread across the different states within the United States and perform predictions and Time Series forecasting in order to study the impact and spread of the COVID-19 in the coming days.

2.0 Data Acquisition and Cleaning

2.1 Loading data

Johns Hopkins University has made an excellent dashboard using the affected case data. Data is extracted from the google sheets associated with the dashboard. The data is available from January 22, 2020. Since the scope of this project is limited to the United States, the data will be extracted only from March 10 to March 31. The other data has been acquired from New York Time's GitHub repository; the data can be used for free of cost for educational purposes. However, author permission must be obtained for commercial use(s).

2.2 Column Description

The data contain several columns, including the latitude and longitude data, which will be used for geospatial analysis. The confirmed, death and recovered cases data will be used for statistical analysis and model training

1. Province_state - Province or state of the observation (Could be empty when missing)
2. Country_region - Country of observation
3. Last Update - Time in UTC at which the row is updated for the given province or country. (Not standardized and so please clean before using it)
4. Confirmed - Cumulative number of confirmed cases till that date
5. Deaths - Cumulative number of deaths till that date
6. Recovered - Cumulative number of recovered cases till that date
7. Lat – Latitude data
8. Long – Longitude data

3.0 Data Visualization and Analysis

The following outcome is recorded through basis analysis for quick. The total number of states include the cases from American Territories and Cruise Ships that arrived through various ports.

Total number of States with Disease Spread: ***63***

Total number of Confirmed Cases in the US: ***188172***

Total number of Recovered Cases in the US: ***7024***

Total number of Deaths Cases around in the US: ***3873***

Total number of Active Cases around in the US: ***177275***

Total number of Closed Cases around in the US: ***10897***

The approximate number of Confirmed Cases per Day in the US: ***8961.0***

The approximate number of Recovered Cases per Day in the US: ***334.0***

The approximate number of Death Cases per Day in the US: ***184.0***

The approximate number of Confirmed Cases per hour in the US: ***373.0***

The approximate number of Recovered Cases per hour in the US: ***14.0***

The approximate number of Death Cases per hour in the US: ***8.0***

The distribution plot in **Figure 1** shows the number of confirmed cases in the United States. The confirmed case distribution validates the exponential growth trend. The number of cases is rapidly increasing in last week. The number almost doubled in a week.

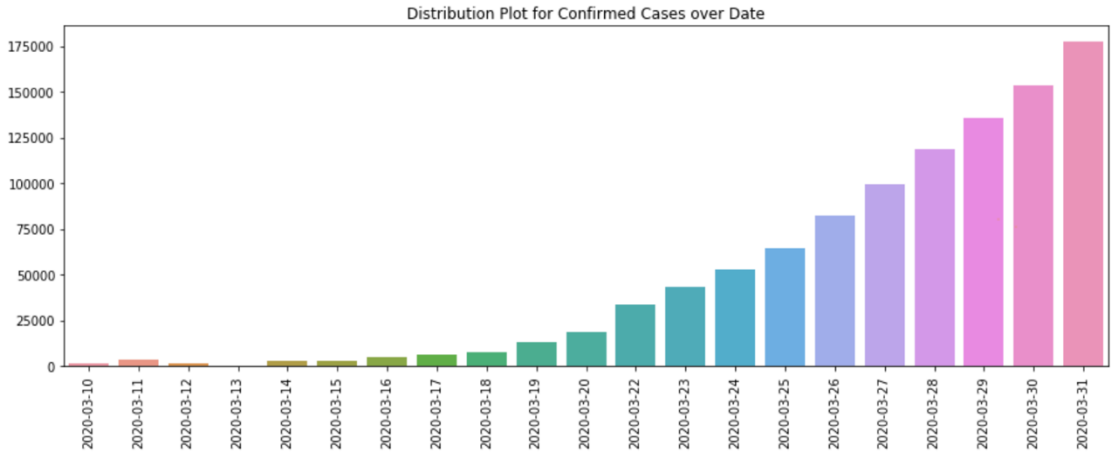

Figure 1 Distribution Plot for Confirmed Cases Over Date

Confirmed cases reached 200K as of April 1, 2020, and approximately 1800 people died within the US, and worldwide 52K people already died as of April 1.

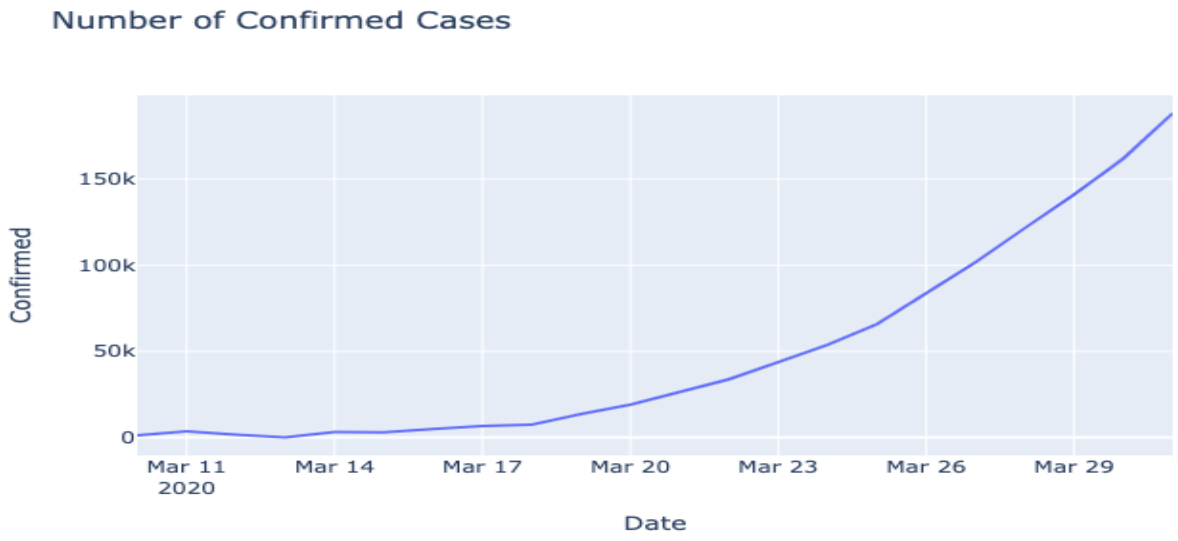

Figure 2 Confirmed Cases Progression

The closed cases have been calculated using the following formula

Closed Cases = Number of Recovered Cases + Number of Death Cases

An increase in the number of Closed Cases is an indication of Recovered or Death case numbers are drastically increasing in comparison to a number of Confirmed Cases. The following graph shows an increase in the number of closed cases. However, the exponential growth in closed cases is due to more recover cases and not necessarily due to death cases. We will differentiate this later when we look at the mortality rate

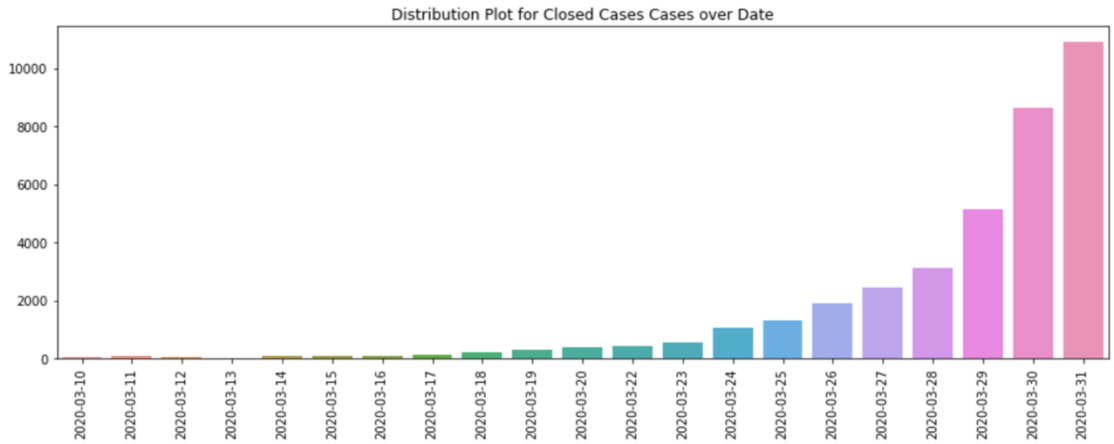

Figure 3 Distribution of Closed Cases Over Date

The active cases have been calculated using the following formula

$$Active\ Cases\ =\ Number\ of\ Confirmed\ Cases\ -\ Number\ of\ Recovered\ Cases$$
$$-\ Number\ of\ Death\ Cases$$

An increase in a number of active Cases is an indication of Recovered or Death cases number is dropping in comparison to a number of Confirmed Cases. We will look for the conclusive evidence for the same in this report later.

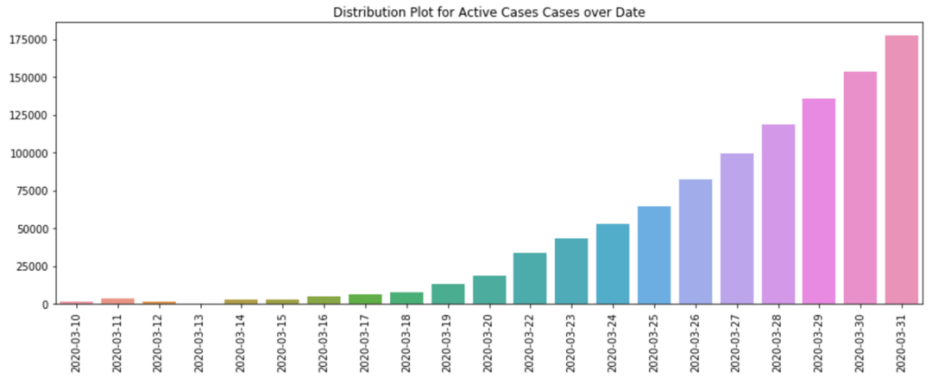

Figure 4 Distribution of Active Cases Over Date

Now, look at the increase in a number of cases altogether in **Figure 5**, as week progress in March-2020, the blue line indicate the 'Confirmed' cases which are increasing exponentially, the orange line shows the recovered cases and green line indicate an increase in a number of death cases.

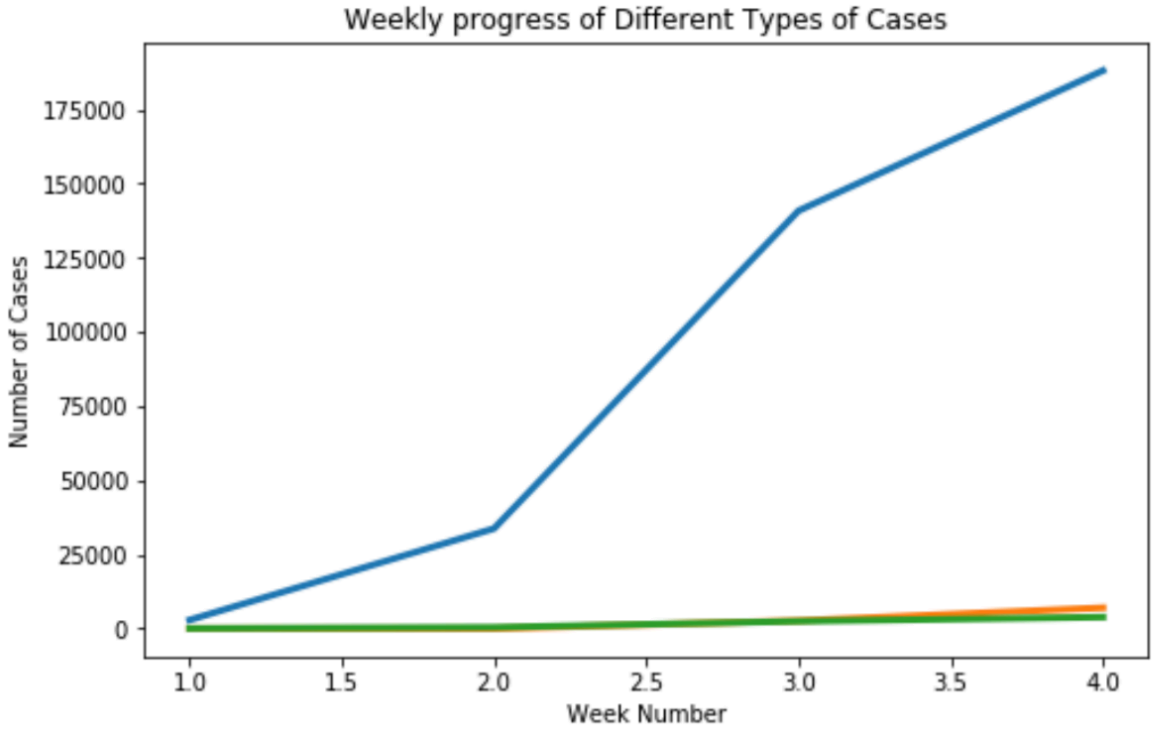

Figure 5 Weekly Progression of Different Types of Cases

Figure 6 shows an increase in confirmed and death cases for the last three weeks of March; it can be concluded that in the third week of March, the number of confirmed cases rose significantly.

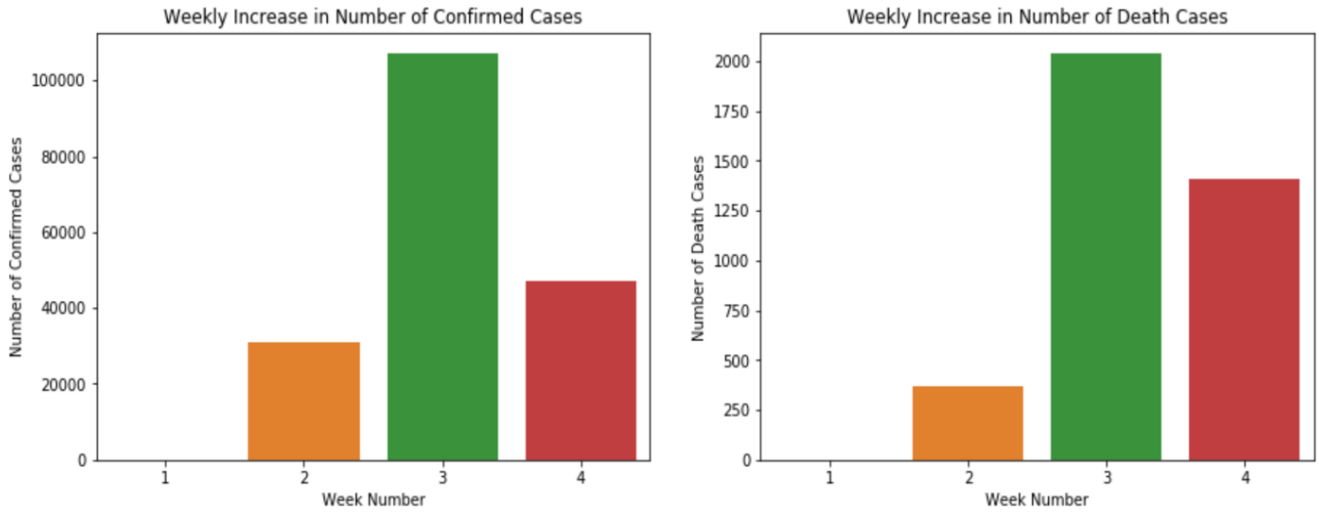

Figure 6 Weekly Increase in No. of Confirmed and Death Cases

The graph in **Figure 7** shows a daily exponential increase in the number of cases

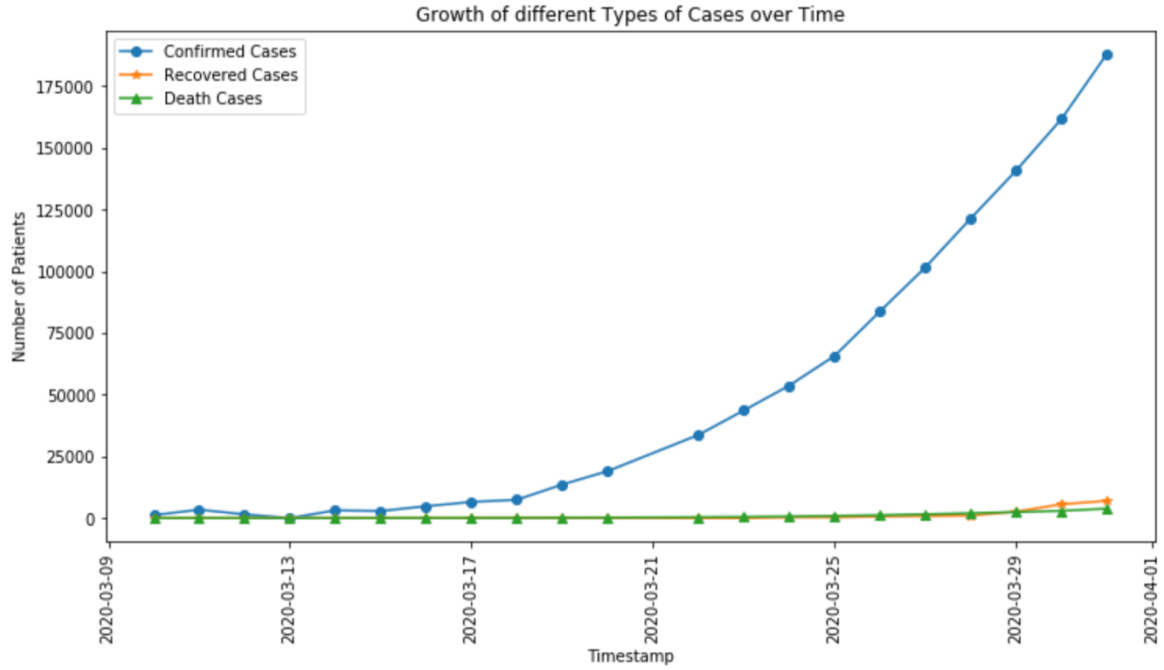

Figure 7 Growth of Different Types of Cases over Time

3. 1 Mortality and Recovery Rate

In the simplest form, mortality and recovery equation can be calculated as below

$$Mortality\ Rate = \left(\frac{No.of\ Death\ Cases}{No.Of\ Confirmed\ Cases}\right) * 100$$

$$Recovery\ Rate = \left(\frac{No.of\ Recovered\ Cases}{No.Of\ Confirmed\ Cases}\right) * 100$$

Figure 8 shows two graphs of mortality and recovery rate. The mortality rates remain low for COVID-19, which is good news. The second graph shows the recovery rate is also low, which is due to an increase in a testing capacity, which increases confirmed cases every day. Recovery rate will increase in upcoming weeks since strict social-distancing guidelines are already in place since March.

Mortality rate increment is pretty significant, along with the drastic drop in recovery rate falling even below the average recovery rate in the US. That's conclusive evidence of why the number of active cases are rising. Also, there is an increase in the number of closed cases as the mortality rate is a clear indication of an increase in the number of death cases

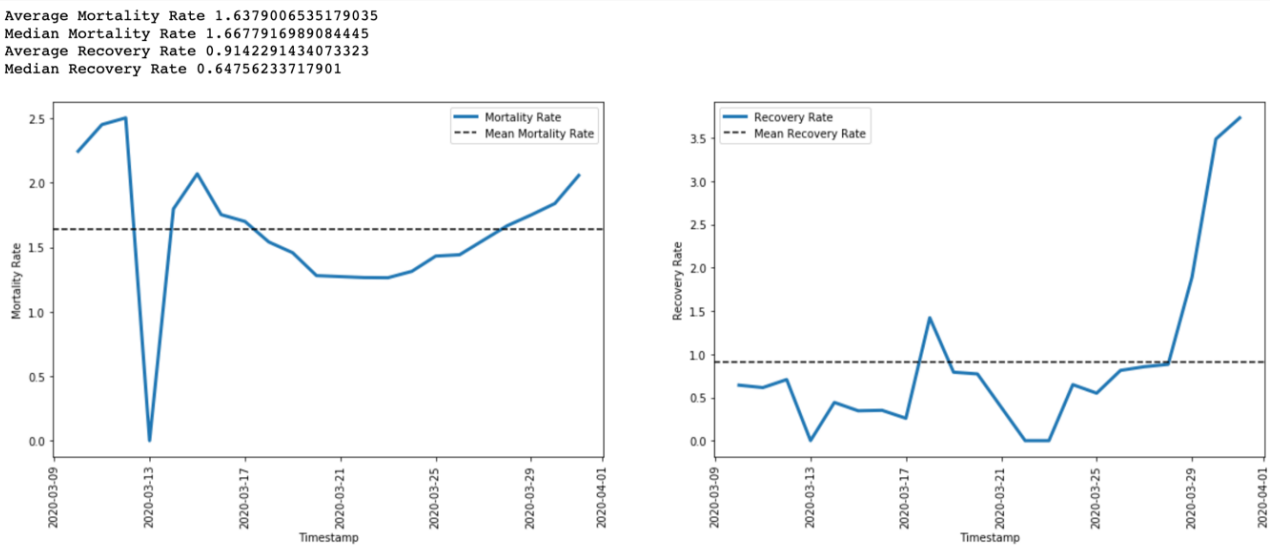

Figure 8 Recovery vs. Mortality Rate

3.2 Growth Factor

The 'Growth factor' is the factor by which a quantity multiplies itself over time. The following formula is used to calculate the growth factor

$$Growth\ Factor = \frac{Today's\ Total\ (Confirmed + Recovere + Deaths)Cases}{Yesterday's\ Total\ (Confirmed + Recovere + Deaths)Cases}$$

A growth factor above 1 indicates an increase in the corresponding case.

A growth factor above 1 but trending downward is a positive sign, whereas a growth factor consistently above 1 is the sign of exponential growth.

A growth factor constant at 1 indicates there is no change in any kind of cases.

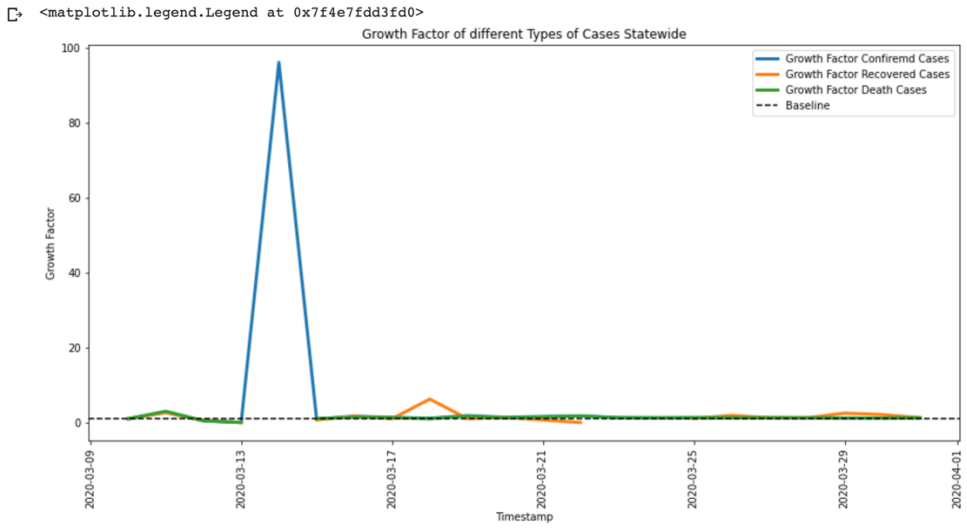

Figure 9 Growth Factor of different Type of Cases Statewide

3.3 Survival Probability

Survival Probability is the only graph that looks the most promising. Having an average survival probability of 98%+ across all states is a very good sign, but this is subject to change in upcoming weeks due to a sudden increase in a number of cases. The equation and other necessary information can be found in the notebook linked to this report.

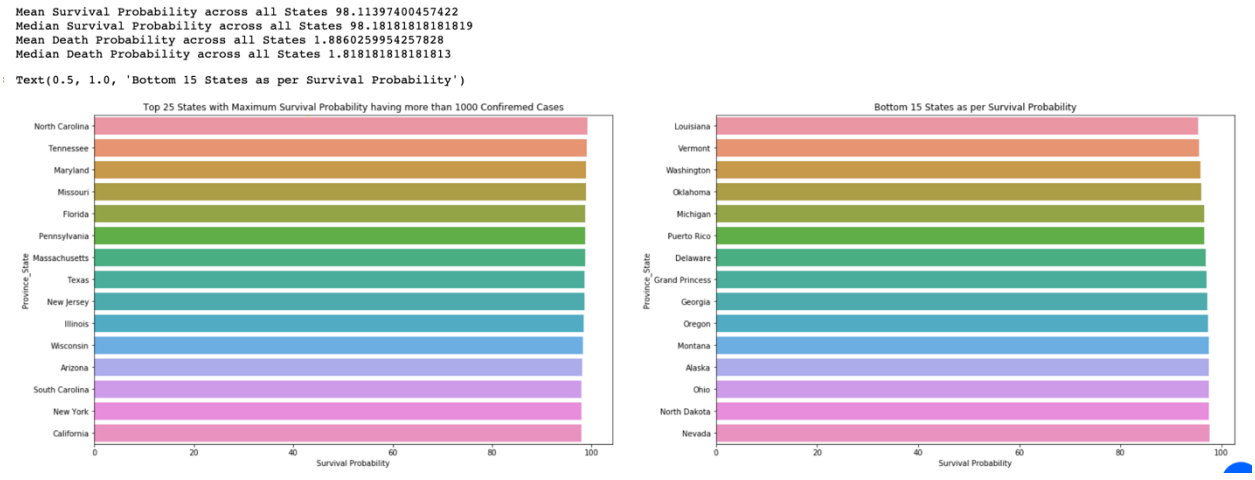

Figure 10 Top and Bottom States as per Survival Probability

3.4 Geospatial Analysis

COVID-19 geospatial analysis has been done at the county and state levels. Figure 11 provides numbers of cases in all categories (confirmed, death, and active) cases for all counties of the United States. The interactive graph can be accessed through the project notebook.

Figure 12 can be used to validate the number of confirmed cases at the state level; this interactive graph is also accessible through the project notebook.

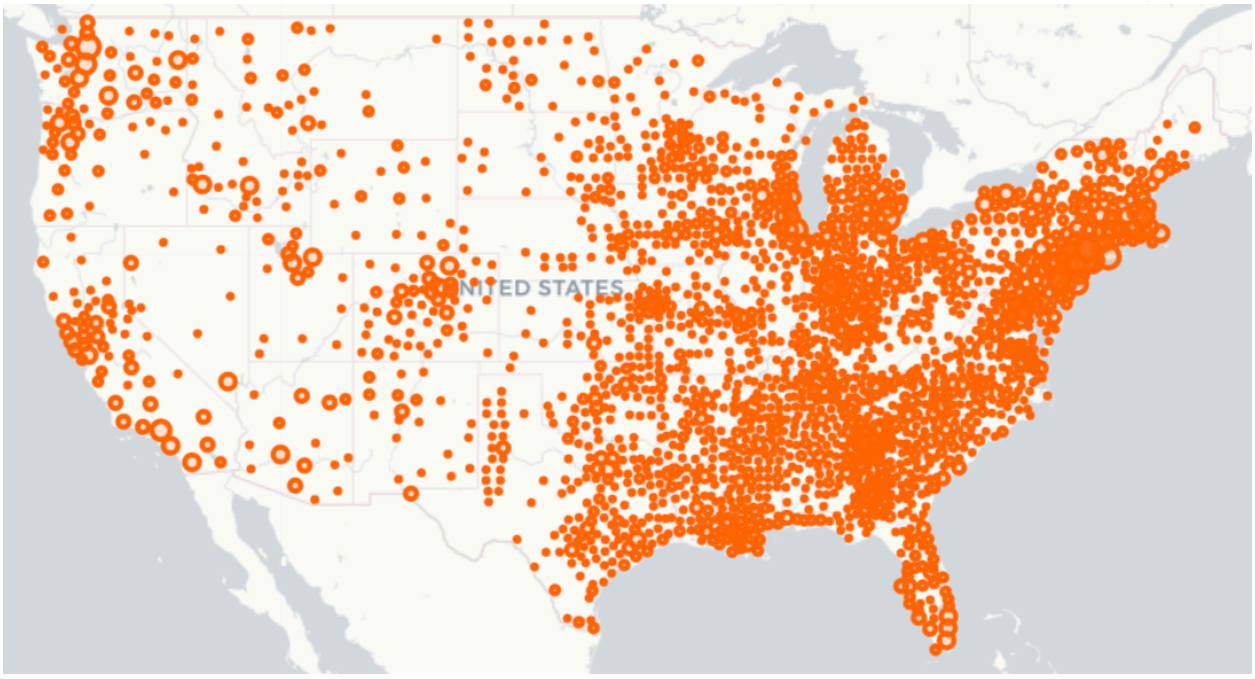

Figure 11 Counties Map with Disease Case Progression

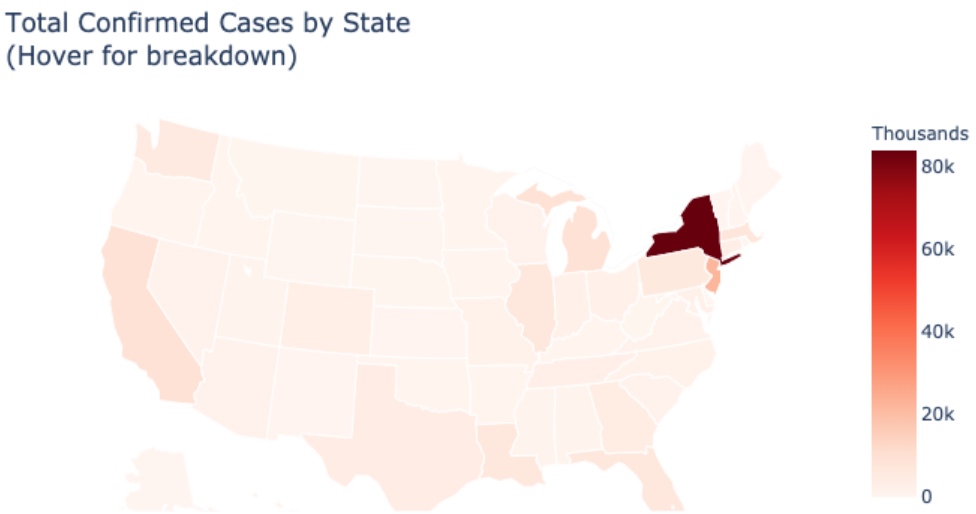

Figure 12 Confirmed Cases Map

4.0 Predictive Modeling

There are two types of models, regression, and classification, that can be used to predict the growth of COVID-19 spread across the country. The underlying algorithms are similar between regression and classification models, but the different audiences might prefer one over the other. This study is using a hybrid model since we are dealing with time-series data.

4.1 Regression models

I applied linear models (linear regression, Holt's linear model), Holt's winter Model, Support Vector Machines (SVM), and Facebook's prophet model for time-series predictions, using root mean squared error (RMSE) as the tuning and evaluation metric.

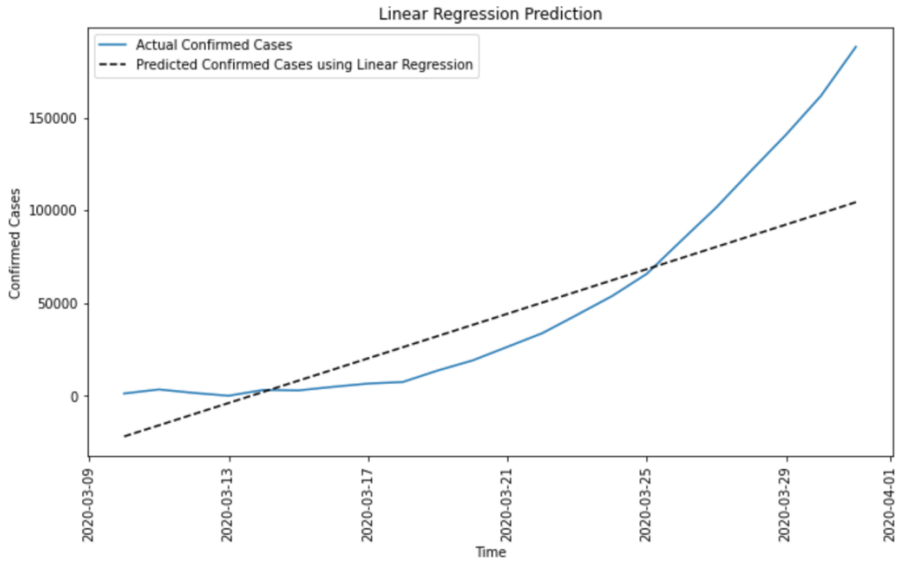

Figure 13 Actual vs. Liner Model Prediction

Since the COVID-19 growth is exponential, the simple linear regression is not a suitable choice. The support vector machine with a 5^{th} degree is more appropriate to predict the future probability with an 80% confidence interval.

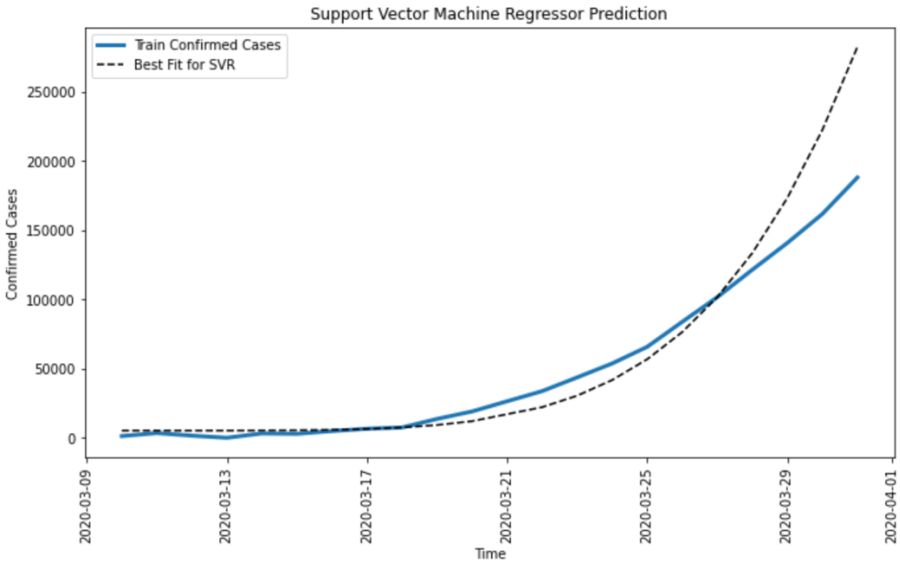

Figure 14 Actual vs. SVM Prediction

Holt's linear model is also not suitable since the relation is non-linear. However, the winter model does provide test data prediction with 87% confidence interval, but it underperforms while predicting for future dates, which is due to overfitting of data, this model can be useful but need more data for training and testing to make a reliable prediction.

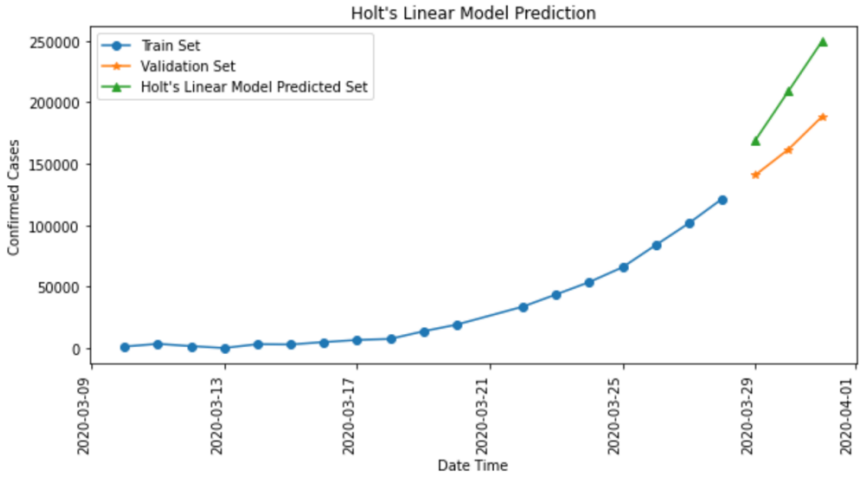

Figure 15 Actual vs. Holt's Linear Model Predictions

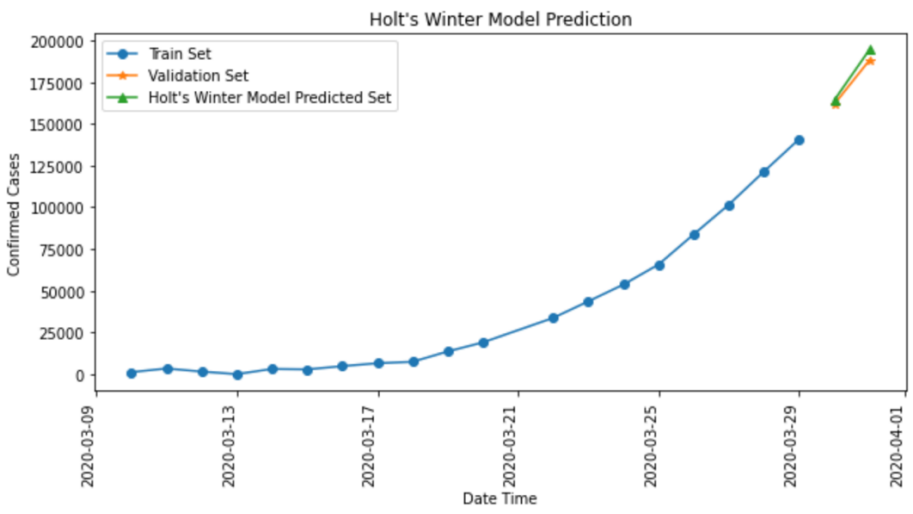

Figure 16 Actual vs. Holt's Winter Model Prediction

Facebook Prophet model also underperform due to lack of data; more data can be obtained in upcoming days for training and re-evaluation of this model.

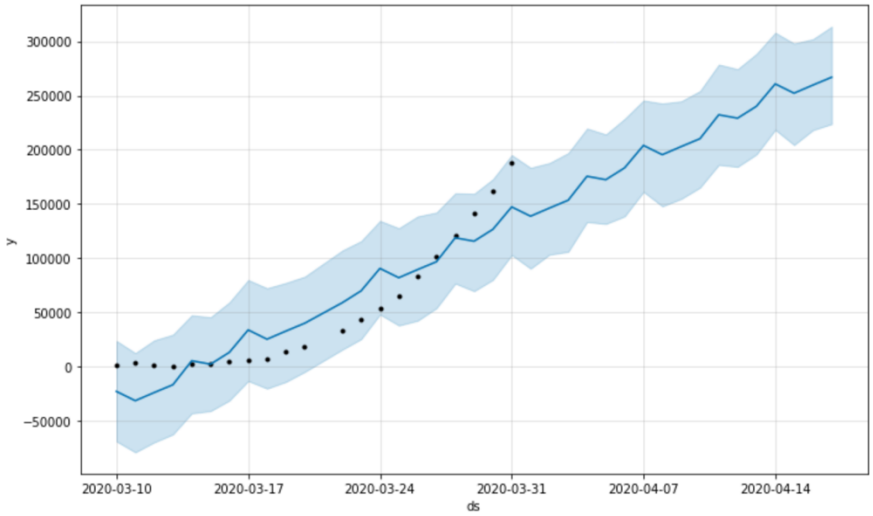

Figure 17 Actual vs. Facebook Prophet's Predictions

5.0 Discussion on performances of different models

Using the new approach of different sample weights, I built linear regression, SVM, Holt's linear model, Holt's winter model, and prophet model using weighted root mean squared error as the evaluation metric. For each model, hyperparameters were tuned using the same metric. SVM had the best performance among all models, which had ~40% less error, and prediction are correct with overall ~60% confidence bound for next week. However, more data is needed to retrain the model for better performance.

Date	Linear Regression Prediction	SVM Prediction	Holt's Linear Model Prediction	Holt's Winter Model Prediction	Prophet's Prediction	Prophet's Upper Bound
4/1/20	110394.25	355217.01	289484.19	225522.10	138833.65	181474.98

4/2/20	116408.09	442325.70	329586.96	263600.41	146292.53	191622.57
4/3/20	122421.94	545975.18	369689.73	305422.62	153527.18	195521.15
4/4/20	128435.78	668416.67	409792.50	355662.39	175597.33	217944.84
4/5/20	134449.63	812096.97	449895.2	410387.28	172453.10	219121.88

Table 1 Model Predicted Values

6.0 Conclusion

COVID-19 doesn't have a very high mortality rate, as we can see, which is the most positive take away. Also, the healthily growing Recovery Rate implies the disease is curable. The only matter of concern is the exponential growth rate of infection.

States like New York, New Jersey, and Michigan are facing some serious trouble in containing the disease showing how deadly the negligence can lead. The need for the hour is to perform COVID-19 pandemic controlling practices like Testing, Contact Tracing, and Quarantine with speed greater than the speed of disease spread at the state and federal level.

Stay Safe and follow Social -Distancing Practices!