With the unexpected introduction of a new strain of respiratory diseases from the Coronavirus family, the world is witnessing an influx of cases. In this report, I will be analyzing the COVID-19 data set provided by Johns Hopkins University Center for Systems Science and Engineering (CSSE) by training a machine learning model to predict future cases given the current reported cases. This report will specifically investigate cases within the US and its comparisons with other countries leading in case numbers.

The machine learning method used to model the data is a regression model. Depending on the trend of the data, either a linear or polynomial regression model was used to fit the data. The sklearn library was used to train these models and predict future targets. The data used in this analysis is the total confirmed cases for a country of interest each day between 01/22/2020 to 03/21/2020. This data is split into 10% testing data and 90% training data. The features and targets used to train and test the model are the known total confirmed cases on a given day and the predicted total confirmed cases after some x amount of days, where x is an integer representing the number of days in the future to predict cases for.

The following lines of best fit were acquired after training models to predict total cases after one day and three days on the training data. Fits were acquired based on confidence score (>95%) and expected accuracy of predictions. The x axis represents the features, the total known cases on a given day, and the y axis represents the targets, the predicted cases after one or three days. The scatter points plotted are the true values of the training features and targets.

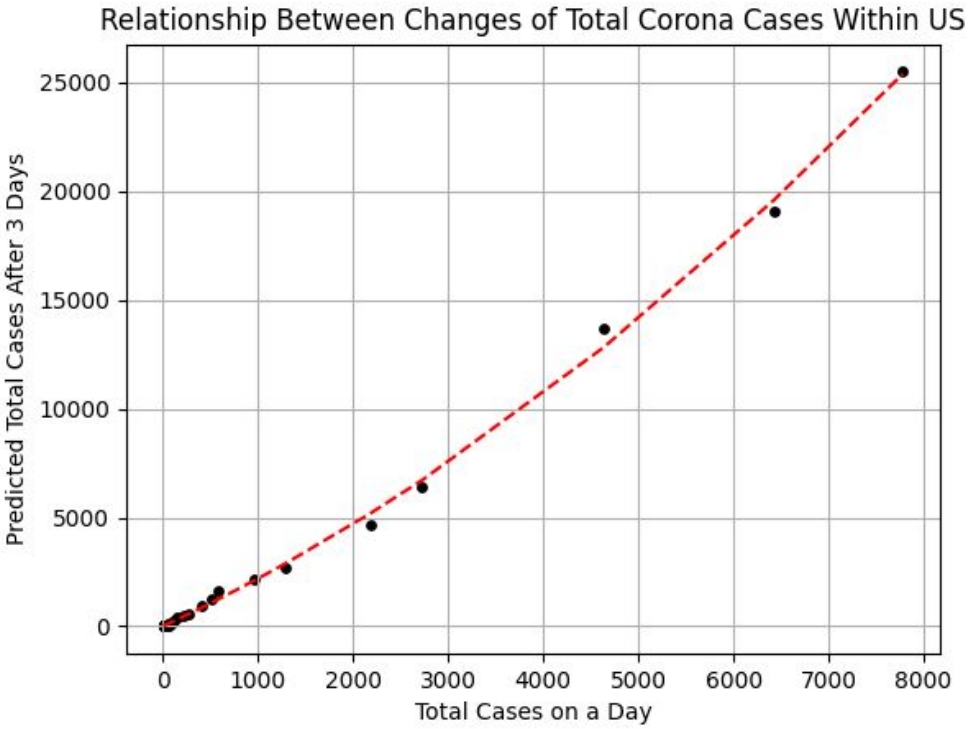

Figure 1

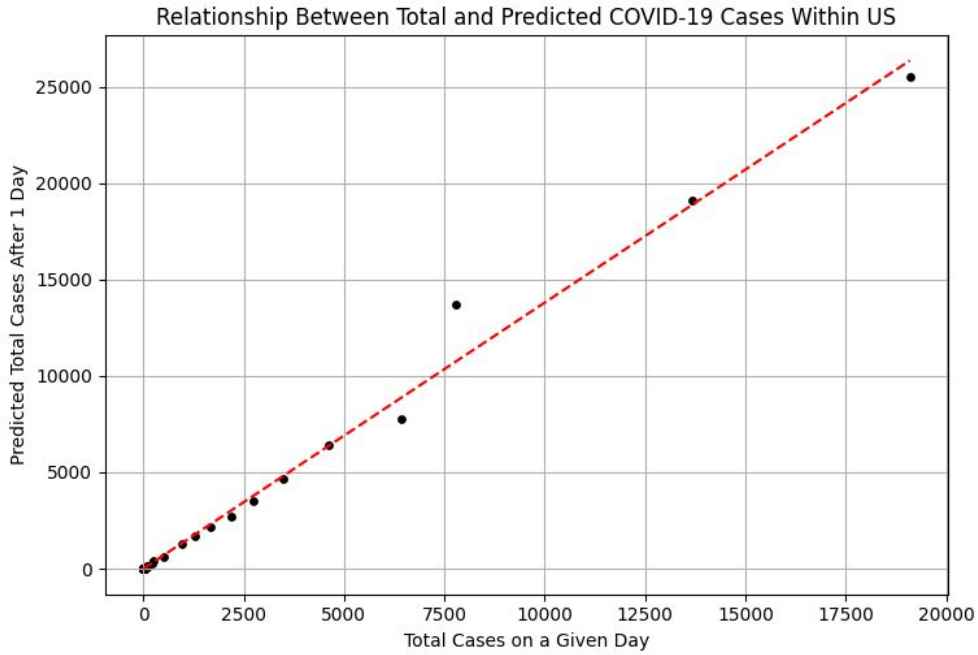

Figure 2

The figures above showcase different fits for the represented data. When training a model to predict future cases after one day, a linear regression model was used to fit the data, while a polynomial regression model was used to fit the data when training the model to predict future cases after three days. Polynomial regression was attempted with the one day model, but overfitting occurred and the confidence score from the test data was 80%, which is below the threshold of acceptable scores (95%).

A reason that both models produced different trends may be due to the loss of data when increasing the number of days to predict for. In the algorithm, the target values assigned to the training features are essentially the same dataset, but shifted downwards according to the number of days to predict for. For example, if there are ten case values in a feature set and we want to predict three days into the future, the target assigned to the first feature value would be the fourth feature value, since that fourth feature value is the value that is three days ahead of the first feature value. This means that only the first seven feature values would have targets assigned to them, and thus the remaining three values at the end of the feature set would get discarded from the training and test set. As the number of days to predict for lowers, the more the dataset is preserved. Although three values may not seem as a significant loss in data, the dataset is relatively small and thus is easily influenced by the modification of data. By the figures above, it can be seen that out of all the values within the dataset, only the last 10 values show a rise in the number of cases within the US. Therefore, it can be inferred that the shape of the trajectory is heavily influenced by these trailing values, and thus removing them from the training set may result in an inaccurate representation.

In order to preserve the integrity of the dataset, the linear regression model used to generate figure 2 is used to perform further analyses. After training the model, more predictions were made on this trained model by iterating ten times and using the last predicted value as the new feature value to predict the next day's case numbers. This method of acquiring predictions allows for generating large amounts of predictions without limiting the training or testing dataset. The following figure displays the trajectory of cases within the US and its predictions for the next 11 days:

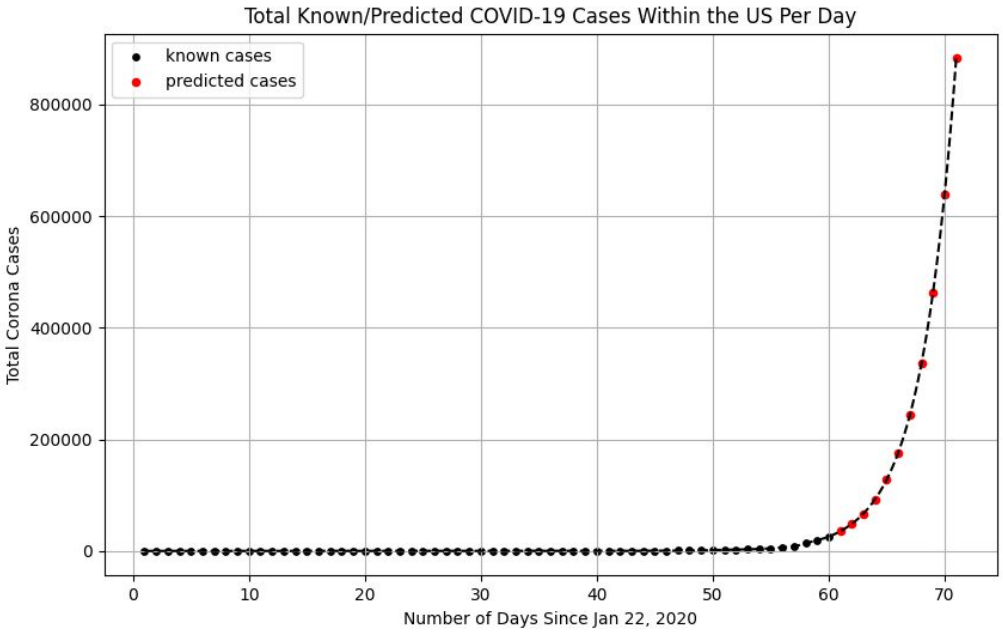

Figure 3

Figure 3 shows an exponential curve and estimates that after 11 days, total cases within the US will rise from approximately 25,000 to 900,000. The rate at which the case numbers increase were modeled as well, as shown below:

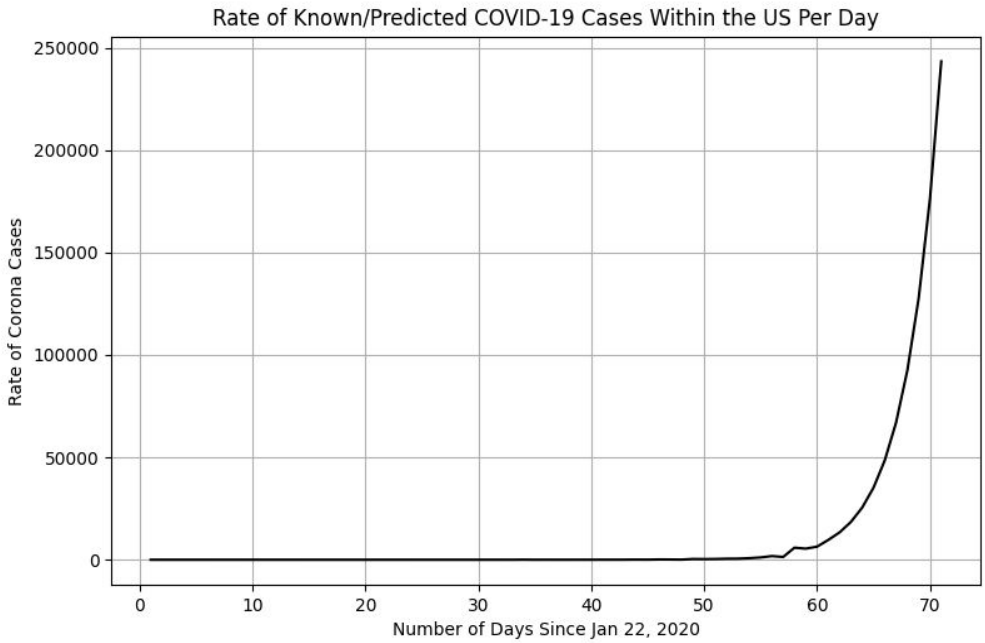

Figure 4

Similar to the predicted total cases, the rates of total cases between each day shows exponential growth, reaching a maximum of 250,000 cases per day. To gauge how this data compares with other countries with high case numbers, case predictions and rates were modeled for Italy and China, as shown below:

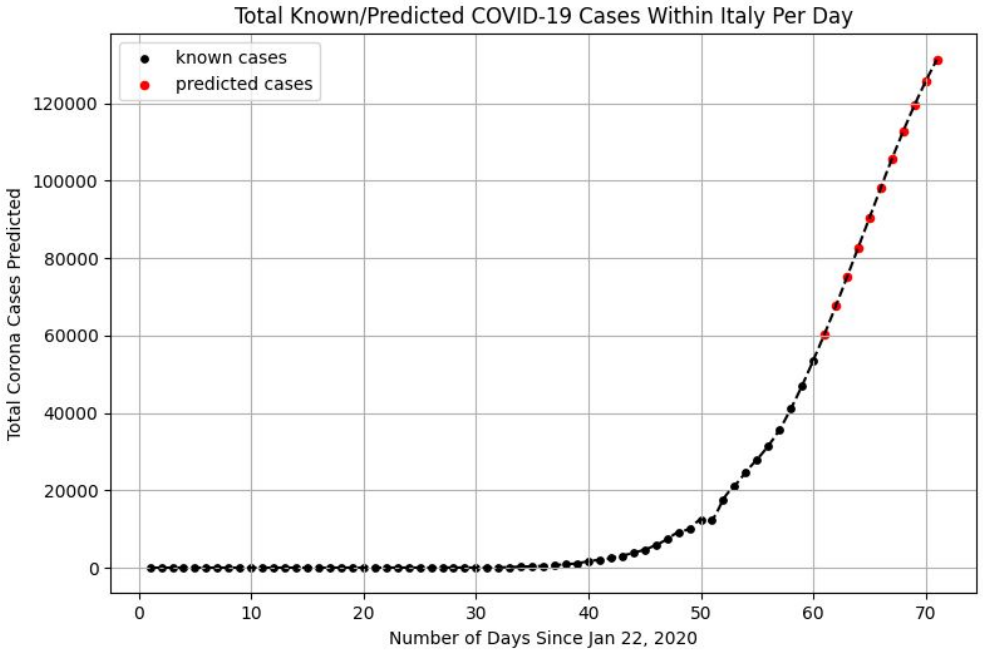

Figure 5

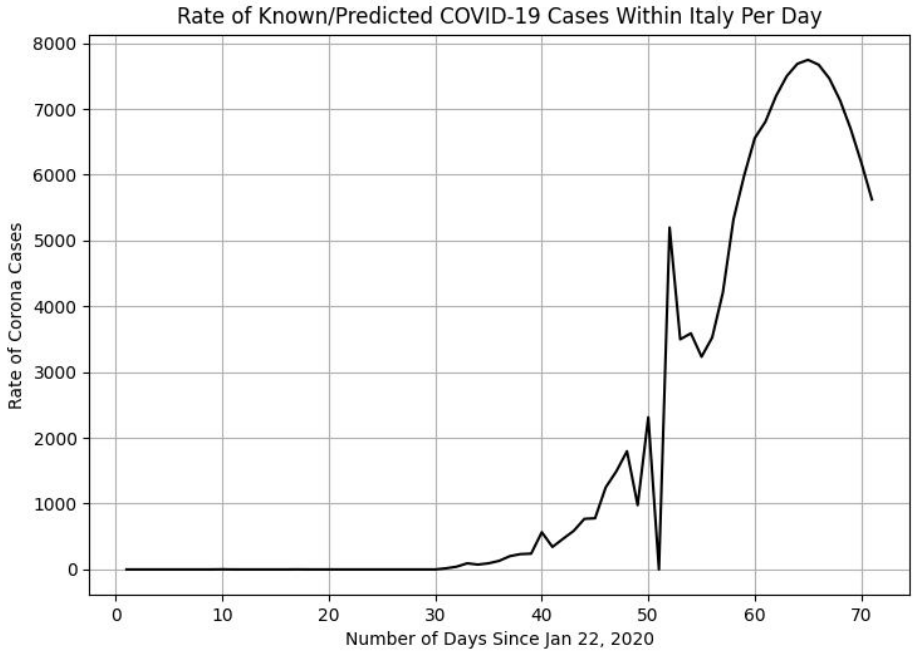

Figure 6

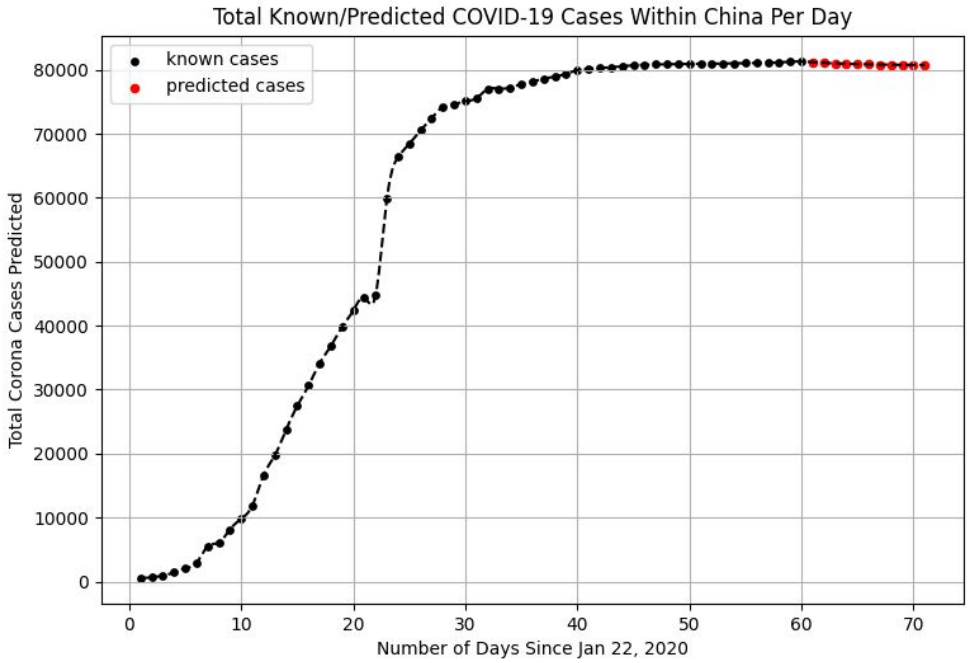

Figure 7

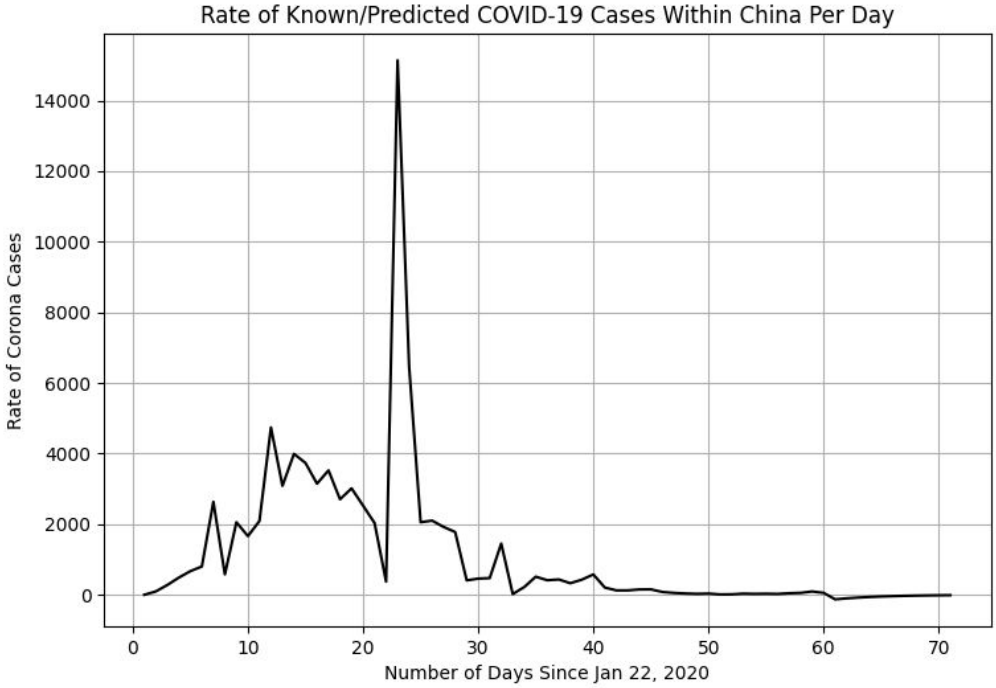

Figure 8

The trajectory of cases for Italy follows an exponential curve similar to the US. However, the curve starts to slightly level off on the final predicted day. This leveling off may either indicate a peak, which may be followed by a downfall, or may be temporary as seen with the data represented in China's model (Figure 7). China's trajectory is comprehensive and meaningful because this is where the virus originated from and thus may serve as an indicator for countries with rising cases such as the US and Italy. The first portion of China's trajectory shows an exponential curve and an eventual leveling off, as in the case with Italy. However, before it can fully level off, there is an abrupt rise in the number of cases, followed by another period of sustained leveling off. This trajectory is interesting for two reasons: first, the sharp rise in cases after cases had begun to start decreasing in rate is unexpected. This sudden rise is the highest rate of cases in China's trajectory, and may be the result of increased testing or changes in contamination control, such as no longer enforcing quarantine among civilians. The second reason why this trajectory is interesting is because China has the biggest population in the world, yet the maximum number of cases before leveling off is merely 80,000. The US is not nearly as populous as China, yet its predicted number of cases exceeds 800,000 with no indicator of leveling off. Although the time it takes for symptoms to show and cases to be identified may slow down the trajectory, China's recently documented cases and its predicted cases shows continued leveling off and a slight decrease in case numbers, which may be an indicator of a halt in tests performed or some form of underreporting.

It is worth noting that the predicted results for all countries may differ from actual documented case numbers because it is not possible to estimate the true number of people who contracted the virus within a day due to a myriad of reasons. These reasons may be that testing resources are under a limited supply, symptoms may take anywhere from a few days to weeks to show up, and people may recover without the need of getting tested. Therefore, it may be worth viewing this data as the estimated number of infections spread within a day based on previous case numbers.