Members: Abebe Amare(ama100), Eugene Kim (yk171)

Part 1: Introduction and Research Questions

On April 26th, 2020 the worldwide COVID-19 death toll passed 205,000(CNN) . The world is fighting against this global pandemic in which the day-to-day lives of millions are heavily impacted throughout the world population. Although the World Health Organization WHO carefully analyzed the spread of the virus, they were unable to declare it as a global threat at its earliest stage of outbreak due to several unexpected and confounding variables. This has cost the world several thousands of deaths and millions of positive cases worldwide. Many civilians are questioning when all things will go back to normal. Some questions our group had specifically about this virus are: How can we use statistical data on global pandemics to predict and prevent the spread of viruses prior to global outbreak? Which are the countries with the leading amount of cases? What key estimations need to be made in order to declare a virus as a potential global threat?

After our group discussed these research questions we decided to gain insight on preventive procedures and predictions. We planned to design an AI-based predictive model that would help detect global pandemics at their earliest stage.Our project concentrates on two areas of research: building a predictive model that could assist in detecting pandemics at their earliest stage and using current data to extract a SIR model for COVID-19. We used data about the current COVID-19 outbreak to build a follow-up SIR model to visualize and better comprehend the nature of global pandemics. SIR is a model that is organized by Susceptible(S), Infected(I) and Recovered(R) individuals. These will be used as dependent variables to measure rates of change. The SIR model will allow us to visualize the nature of the spread of the pandemic and better anticipate impending cases. Since working from the prototype, we developed a predictive model as well as an SIR model.

Part 2: Results

Using the analyze.ipynb and write_csv.ipynb in our git repo, we generated the following graphs to illustrate how fast the number of cases is increasing in the affected countries ranked in the top seven according to (https://www.worldometers.info/coronavirus/) along with the total daily number of cases worldwide. Figure1 shows the line graph for the number of cases in the highly affected area and the number of cases is increasing exponentially overall. In addition to that, the log graph that is plotted in Figure2 shows the rate at which the number of cases is increasing exponentially for example in the US. We categorized countries based on whether they have an increasing rate or decreasing rate as seen in Figure1 and Figure2. This indicates that they have exponential growth of cases, for example, the line for the US (brown line) in Figure 2. This can be used to test our model. We also plotted the line graph for the total number of cases in the world in Figure 3. As we can see from the graph the number of cases was increasing exponentially and recently started to decline. The significance of this trend of the data will help us to come up with an AI system that will detect pandemics at their earliest stage. These were the conclusions made upon the prototype.

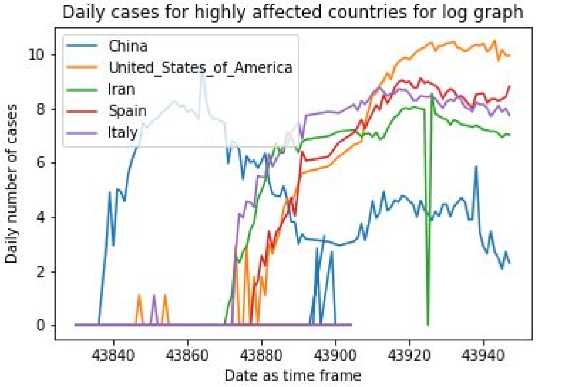


Figure1: The line graph for the top daily cases for the world as of April 24.

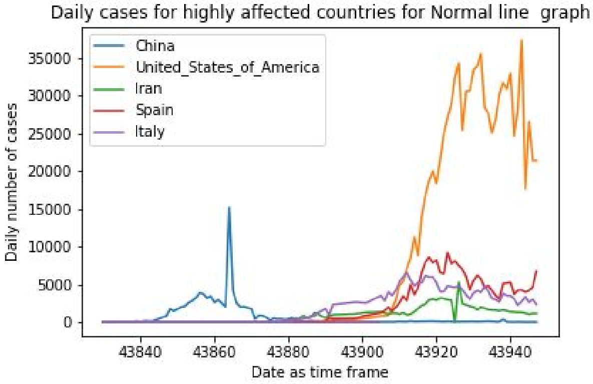


Figure2: The log graph for the number of cases in highly affected countries as of April 24.

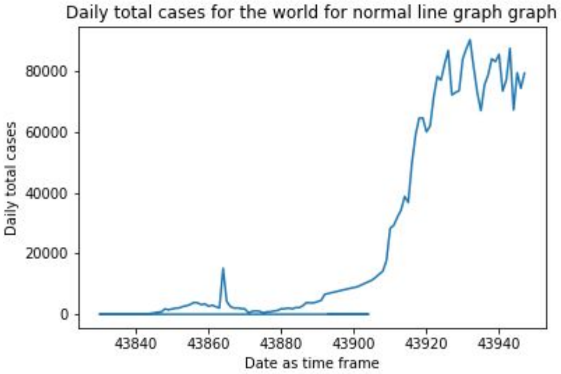


Figure3: The line graph for the daily number of cases in the world.

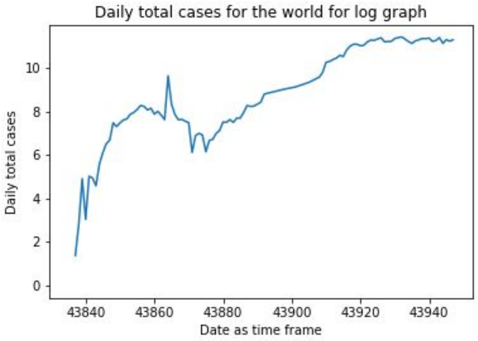
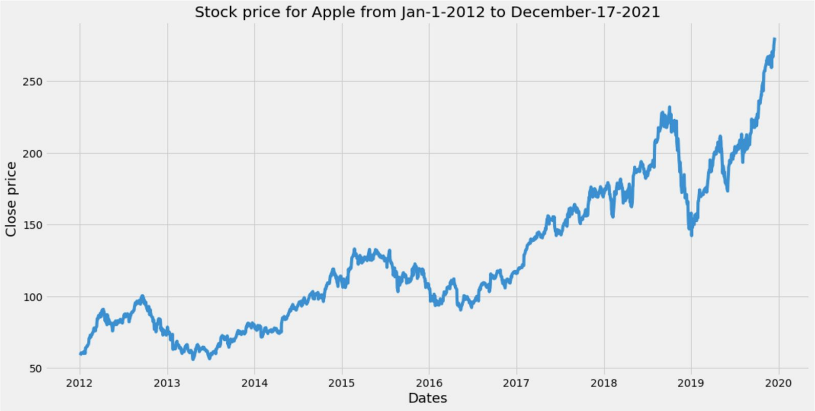
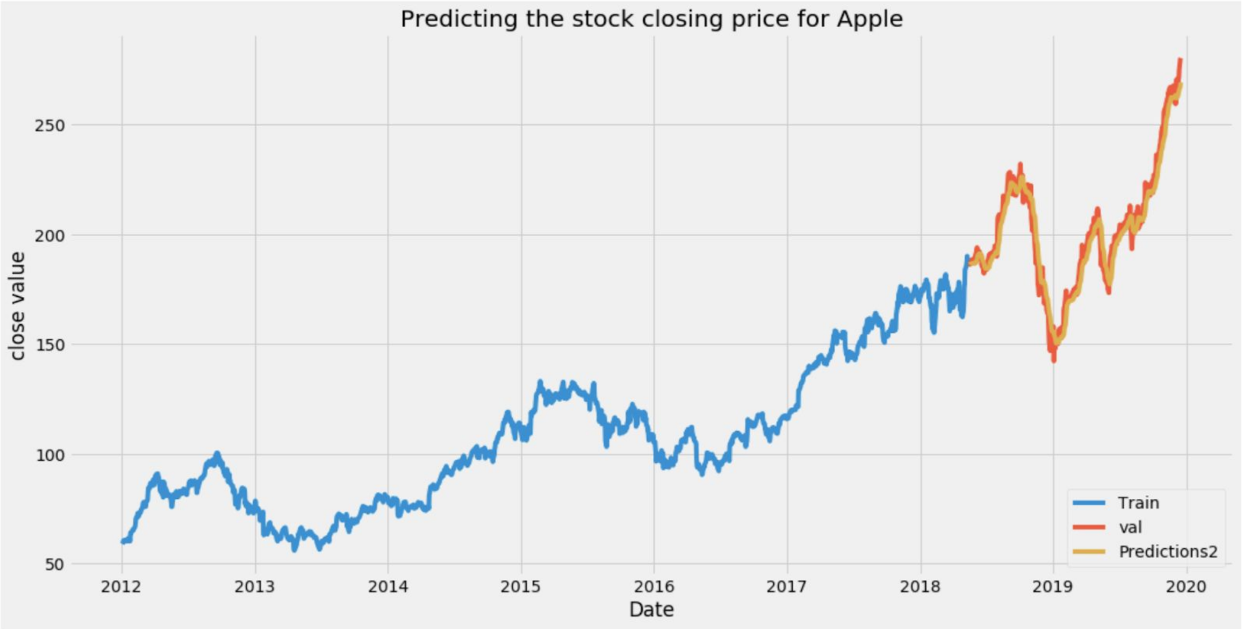


Figure 4: The log graph for the daily number of cases in the world.

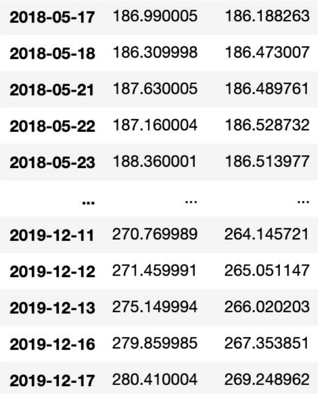
We decided to test our predictive model on a large data set to ensure that our model was competent and compatible for large data. The graph to the right displays the Stock prices data for Apple from 2012 to 2020.



Using 80% of this as training data, we built a model to predict the remaining 20% as shown below.

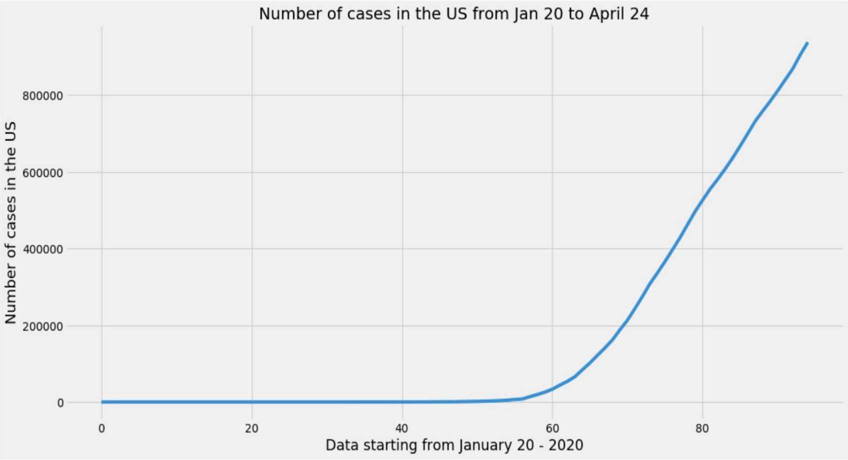


The prediction(yellow) closely follows the actual values of the 20% that came from the actual data values(red). This showed us that our data was able to correctly predict the 2019-2020 data based off of the data from 2012-2018. The table to the left shows the date, actual value and predicted value respectively for a couple of dates near mid-2018 and then end of 2019. As you can see in the table most of the values were off by less than 2 in 2018 and less than 15 at the end of 2019 which is much further ahead. These are proportionately significant predictions and so we decided to further use our model on COVID-19 data in the US.

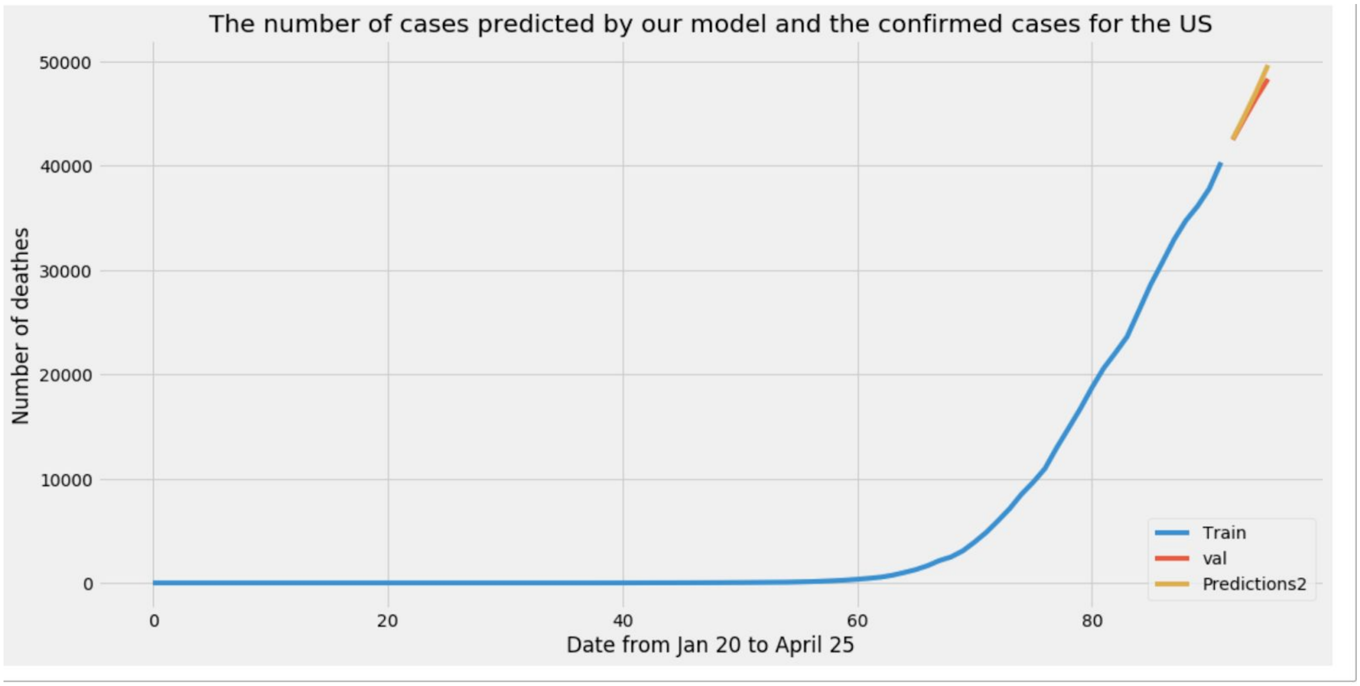


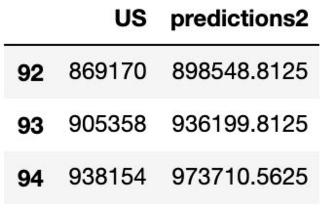
2018-05-17	186.990005	186.188263
2018-05-18	186.309998	186.473007
2018-05-21	187.630005	186.489761
2018-05-22	187.160004	186.528732
2018-05-23	188.360001	186.513977
...	...	...
2019-12-11	270.769989	264.145721
2019-12-12	271.459991	265.051147
2019-12-13	275.149994	266.020203
2019-12-16	279.859985	267.353851
2019-12-17	280.410004	269.248962

The graph to the right displays the number of cases in the US over a period of 94 days. The number of cases remained close to zero until around day 55 in which the number of cases steadily increased linearly.



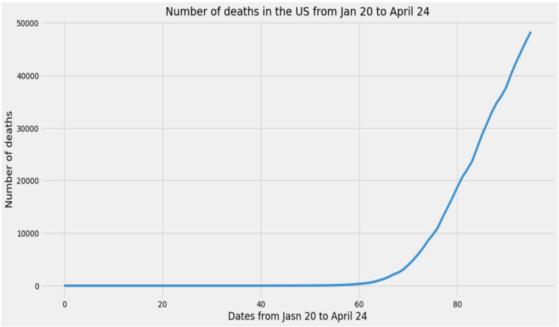
The graph below shows the number of cases predicted by our model compared to confirmed cases in the US. 91 of 94 days(97%) were used as training data to predict the remaining 3%.

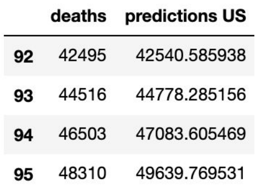




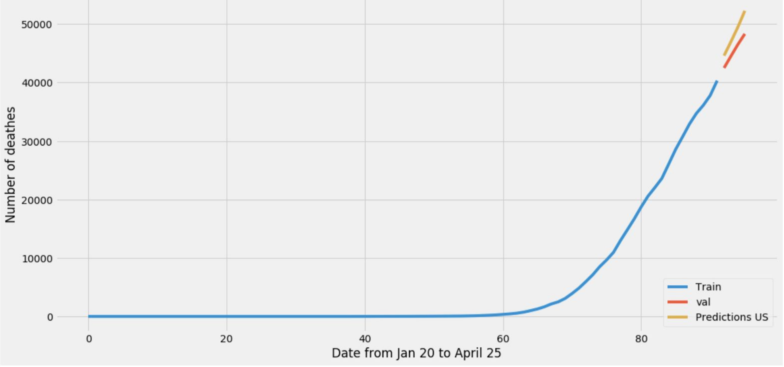
	US	predictions2
92	869170	898548.8125
93	905358	936199.8125
94	938154	973710.5625

The table to the left shows the day number, the actual value of cases and the predicted value for that day respectively. Corresponding to the graph above, the table shows the predicted values of cases for the last 3 days. If we examine the table on the left, we can see that the prediction for day 92 of the US has a difference of 29,378 cases which is just 3.3% of the actual value of cases. We repeated this procedure for the deaths in the US and obtained statistically significant results.

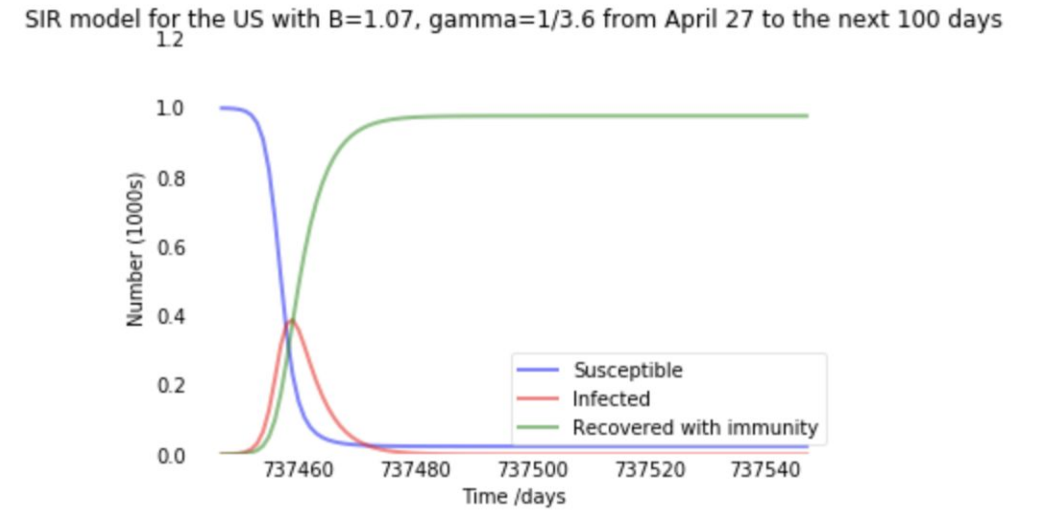




	deaths	predictions US
92	42495	42540.585938
93	44516	44778.285156
94	46503	47083.605469
95	48310	49639.769531



The predictive model we generated and the raw data we collected helped us to devise a SIR model that will characterize the spread of the disease. The SIR model uses three differential equations to come up with an equation that quantifies the number of infections and recoveries as a function of time. B is the number of people that a single person can infect in a day, and gamma is the number of days that person takes to recover from the disease. The significance of the SIR model is that it can be used to detect a pandemic at its earliest stage. Below is our SIR model for the US. We used B=1.07 people/day and gamma=1/(3.6) days and generated the



following graph by solving the above differential equation and our model predicts that the rate at which the disease is spreading is flatting and will decline in the next 10 days as shown in the following graph. The code used to generate this graph can be found in our git repo.

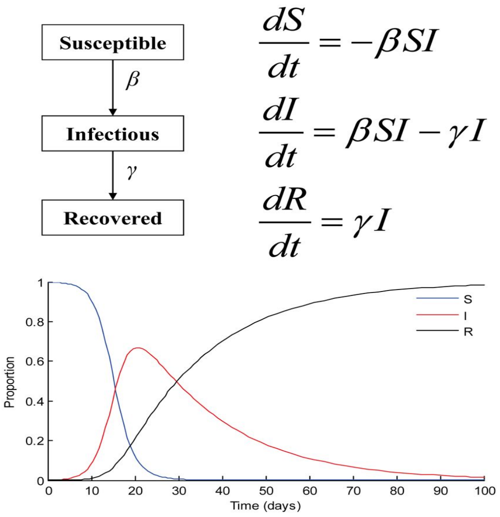
The significance of our model is that it can be used to predict the spread of any disease if the above numbers are experimentally determined. Our code is robust enough to predict the general behaviour of the spread of a disease given the number of infected people, susceptible population , B and gamma. According to our model a disease is a pandemic if the log graph of the infected number of people has linear or exponential growth rate. Our model predicted that Covid-19 should have been declared as a pandemic before January 25th. We used the data for china and the log value of the number of infected people shows a linear increase and then an exponential increase.

Part - 3: Methods

We used a data source from the European Center for Disease Prevention and Control. The CSV file contains data on the geographic distribution of COVID-19 cases around the world alongside. Each row contains the name of the country, the number of new cases per day, their date, the number of deaths that day and the population size. We then went through a process of data scraping to find the raw data for the number of corona cases in the world. We collected most of the data from the following websites:
https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide
https://www.worldometers.info/coronavirus/
https://coronavirus.jhu.edu/data/new-cases.
The data was not organized well as it was writing the name of a country in a column with a number of cases and deaths for the past 97 days. The problem with this data is that the name of a country is written more than it needs to be and it is not convenient to extract a piece of information from the csv file. We created an xls file write script and we extracted and organized the data.

Once we collected the data, we worked on cleaning the data to eliminate unprecedented values. We then created two excel files(cases2.xls and deaths2.xls), one that contained the number of cases and one with the number of deaths and then cleaned any leftover data. We formatted the data through a tabular form using python script(write_csv2.ipynb) in .xlsx file. The code we wrote and used can be accessed in the Git repo. Once we organized the data into .xlsx file, we used the (analyze2.ipynb) file in the git repo to access the dataset which contains the number of deaths/cases for each country. We utilized a built-in function to convert .xlsx/.xls files to pandas DataFrame and then converted it to a pandas DataFrame and extracted the data by using functions we defined.

For the predictive model(model_US_death.ipynb and model_US_cases.ipynb), we imported packages like sklearn, keras and matplotlib and counted the number of rows and columns for the US data we obtained. We built a line graph to visualize the number of deaths to see how the data is changing. Then we created a new dataframe with only the column specified by the deaths/cases and converted it to a numpy. We created training data of 95% of the actual data and we called it data2. We split the data into x_train and y_train and then converted it into a numpy array to build the LSTM model. We used Long short-term Memory (LSTM) which is an artificial recurrent neural network architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections and it can be used with a myriad of data types like video and image processing. We scaled the data we collected to be between 0 and 1. Then we built a model that uses 30 LSTM sequences, 15 Dense sequences and one dense output. We compiled and optimized the model to reduce the mean square error. This allowed us to have an accurate x_test and y_test. We then obtained the models predicted values and calculated the real mean square error. This resulted with a graph of actual values then predicted values. The comments in the code confirm the details of this method.



$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Then for the SIR model, we built it to characterize the spread of communicable diseases in society and predict the demographic of the spread of the disease. This model uses the rate at which the number of susceptible people is changing, the rate at which the infected people recover and the rate of infections in the community. The above graph for the SIR summarizes the differential equation that should be solved to characterize the spread of a disease in a SIR model for corona. We generate SIR model to predict the ending time of the pandemic and B and gamma values used to generate the model for the USA comes from a reaserch project posted on https://ibmathsresources.com/2020/01/28/modelling-the-spread-of-wuhan-flu-coronavirus/. We used the values calculated for the British to generate the SIR model for the US. Since the US and the United Kingdom have similar health facilities and economic status, we thought it would not make a huge difference in the model. We used the odeint of the ***scipy*** to solve the differential equations and matplotlib to plot the resulting relation between number of infected, recovered and susceptible.

Part 4: Discussion

After analyzing our initial graphs from our prototype, it was clear that the rate at which COVID-19 was spreading was drastically different amongst countries all over the world. Some of the countries with the most cases were the US, China, Italy, Spain and Iran, The rate at which it was growing was clearly exponential and spread much faster than other countries. In order to figure how the curve was flattened for particular countries, we built a predictive model. With training data and test data, we were successfully able to predict the rate at which the cases would increase. This is beneficial to data analysts that already have large datasets. Since we used a majority of the data as training data, we weren't able to predict possible values far in the future. Some questions we had for future research are: How much of a time interval is needed in order to predict further into the future? What is the effect of the advancement of the health system in the spread of the disease? What is the effect of population density?

Some of the challenges we encountered while we were doing were the lack of big data to test our model and weak internet connections to process the data in general. Our predictive model predicts well even if it is only provided with a small testing data, but our SIR model for the US was generated by assuming the parameters for the differential equations come from the UK, and our model deviates from the actual data. Our model would be reliable and to the point if the parameters used to generate the model are experimentally determined for specific countries. As we examined data that was further from our training data in time, the accuracy decreased. Our SIR mode shows that the current spread of the disease reaches its peak in the next 10 days and the curve will decline at a recognizable rate. Our model can be used to predict how long a pandemic will last, when to declare as a pandemic and how long it would take for the pandemic to be over.

Appendix

- Code-base:https://bitbucket.org/aabebe1966/cs216_final/src/master/
- Used-resources:
    - https://ibmathsresources.com/2020/01/28/modelling-the-spread-of-wuhan-flu-coronavirus/.
    - https://www.worldometers.info/coronavirus/
    - https://www.worldometers.info/coronavirus/
    - https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases?force_layout=desktop
    - https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/