Predicting Recovery from Covid-19 for Different Countries Based on models from Currently Top-Pandemic Ridden Countries

Daniel Cleary
Dept. of Computer Science
Stevens Institute of Technology
dcleary@stevens.edu

Parth Parab
Dept. of Computer Science
Stevens Institute of Technology
pparab@stevens.edu

Saniya Attar
Dept. of Biomedical Engineering
Stevens Institute of Technology
sattar1@stevens.edu

***Abstract*—This project is designed to predict recovering trajectory for various countries based on the first world countries that were very highly infected with the first wave of COVID-19(Coronavirus) which are China, South Korea and United States of America.**

***Index Terms*—coronavirus, machine learning, polyfit, algorithms.**

I. INTRODUCTION

This report consists completed status of the project and the limitations faced, with detailed analysis of steps taken.

II. PROJECT DISCUSSION

A] ABOUT THE VIRUS

COVID-19 or coronavirus is a newly discovered infectious virus that is highly contagious. People infected with the virus will experience mild to moderate respiratory illnesses and can recover without any medically necessary interventions. Older populations suffering with pre-existing comorbidities like Cardiovascular Disease, Diabetes, Chronic Respiratory Diseases, and cancer are more likely to develop serious illness. At this time, there are no specific form of vaccines or treatments for COVID-19, but there are several clinical trials that are looking for potential treatments and World Health Organization(WHO) will be continuing to update this information as the clinical findings advance further [1].

B] WORLD CRISIS

More than 2 million infected cases have been currently reported around the world with approximately 500,000 recoveries and estimated more than 100,000 fatalities. The reported cases are increasing by the rate of 4% every day. United States is currently the country with the highest number of cases (~700,000) in the world, followed by China and Italy. China being the initial country affected by the virus, has also been the one with highest recovery rate with approximately 80,000 total cases and more than 70,000 recovered [2]. South Korea has implemented strict quarantine laws, which includes prosecution of people who are suspected but will not cooperate to get tested for the virus. United States has currently only implemented quarantine laws for the state of New York [3].

C] ABOUT THE PROJECT

Our main goal to design this project was to apply machine learning algorithms in order to roughly estimate or predict the recovery rate for different countries based on the recovery strategies of the above mentioned countries that have been most heavily affected since the start of the pandemic. This would help us analyze how the recovery plans in different countries can be modified to “flatten the curve” and avoid lesser deaths and more patients recovered over a certain period.

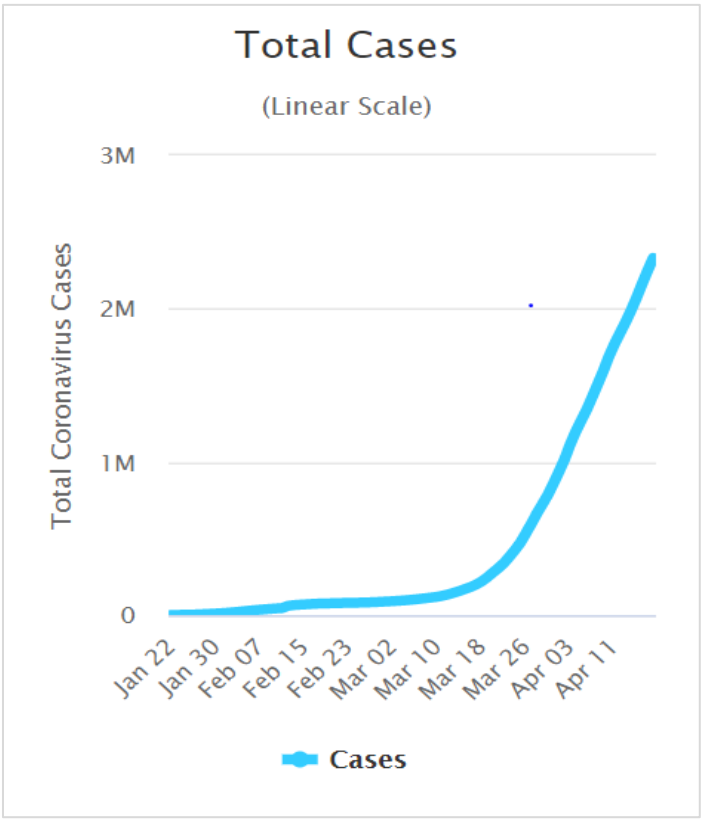


Fig 1: Total number of Coronavirus Cases around the world from the months of January to April [2].

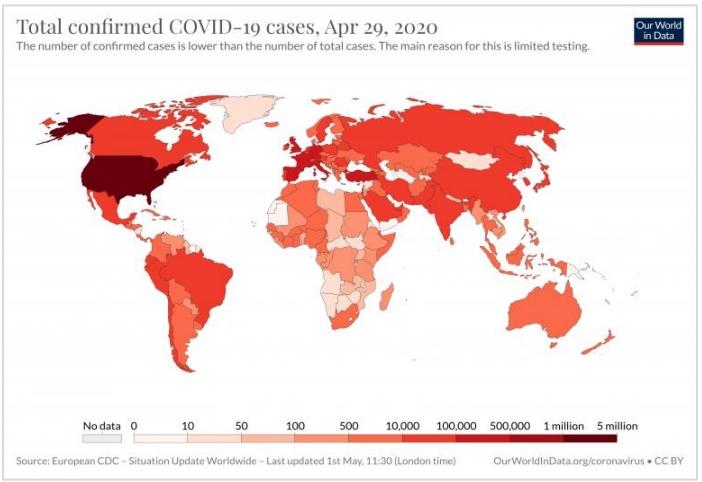


Fig. 2: Color map representation for the spread of coronavirus [8].

III. METHODS

A] LIBRARIES AND DATASETS

We will be using Google Colab for the purpose of this project since it would allow all the partners to work on the code simultaneously, utilizing Google's Cloud servers and leverage Google hardware. We can write, execute Python in our browser.

We have utilized the data from Github repository since it contains the most recently updated CSV files with cases reported from all around the world and is cleanly filed for each country separately.

The data first began to be reported in the repositories from Washington State in January 2020 [7]. New York Times started by releasing a series of data files with cumulative counts of cases around USA and then at state and county level overtime [7]. Github combined this time series data from local governments and health departments and attempted to provide a case wise report of the ongoing outbreak. The data was made public in response to requests researchers, scientists and government officials who would like to access the data to better understand the nature of the outbreak.

We are importing Pandas and NumPy libraries for data management. Pandas provides easy to use structures and data analysis tools, in-memory 2D table object called Dataframe which resembles a spreadsheet with columns and row labels. NumPy library provides objects for multi-dimensional arrays and computing functionalities that are designed for high-level mathematical functions and scientific computation. We also imported Beautiful Soup library for scrapping the data form GitHub, it extracts all the URLs found within a page's tag [4]. Requests Library is basically used to fetch the data and be loaded into BeautifulSoup which then extracts the data from the now stored URLs. We also imported the Scikit Learn Metrics which implements functions assessing prediction errors for specific purposes such as regression metrics. In future we will use this to calculate MSE and regression plots between various countries and the training dataset which we have set to the above mentioned three countries that will be part of our prediction analysis.

B] EQUATIONS

We used Polynomial Regression fit to avoid underfitting or overfitting of the data, because it would increase the complexity of the model and make it fit better with lesser noise. This was used to compare the population affected of three countries (US, South Korea, and China) and to project the rate of infection for 250 days for the aforementioned countries. Mean Absolute Error equation was used to measure the difference between two continuous variables and is best used with regression models which is part of our project [5]. MAE would also give us the median target value instead of MSE (Mean Squared Error) which would where optimal prediction is the mean [6]. We used MAE to find the error between of each of the three countries in our training data as test dataset against each other in addition to Italy and France.

C] PRE-PROCESSING

Pre-processing consisted of taking each csv file from the github and removing all but the five countries used for testing. Each district labelled for the country was compressed into a single row with columns consisting of confirmed cases, deaths, and recovered cases. The metric for number of active cases was then calculated and appended to the list, as well as the percentage of the total population for each country.

Prior to modeling the three selected countries each dataset was compressed to only contain the confirmed cases per day. Then for each of the three selected countries, the dataset was split into a 70/30 for training and testing sets. The training set was then used to make a regression model for a given order. The MSE was then calculated for the model in comparison to the training set then the testing set. This was repeated for orders 0 to 10. Whichever order produced the least error in terms of the testing set was selected, with the first few orders ignored as the training set would need time to become properly adjusted. The resulting MSE values were plotted per country, as well as the equation chosen given the most accurate model order.

After determining the equations for each model, the mean absolute error was calculated between the remaining two countries and each model, with each of these values being stored to a data frame. Whichever country produced the lowest mean absolute error was considered to be the closest model, with the model choice being noted and stored.

D] EDA (EXPLORATORY DATA ANALYSIS)

JHS GitHub has multiple names for the same country/region. We used a dictionary to re-index it to a standard value. The data also does not show the amount of population affected. This was calculated by subtracting the total population with the active cases at the time. All data values are dynamic and change as new data is scraped.

The end dataset is a list of 5 countries CDRA (Confirmed Deaths, Recovered and Active cases) and population affected with each row representing the date. The first record being of 01/22/2020 to the current date.

E] STEPS

1) We first code for scraping CSV file names from GitHub using BeautifulSoup and Requests.

Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed	Deaths
South Carolina	US	2020-05-05 02:32:34	34.22333378	-82.46170658	33	0
Louisiana	US	2020-05-05 02:32:34	30.295064899999996	-92.41419698	134	10
Virginia	US	2020-05-05 02:32:34	37.76707161	-75.63234615	425	6
Idaho	US	2020-05-05 02:32:34	43.4526575	-116.24155159999998	710	17
Iowa	US	2020-05-05 02:32:34	41.33075609	-94.47105874	1	0
Kentucky	US	2020-05-05 02:32:34	37.10459774	-85.28129668	81	13
Missouri	US	2020-05-05 02:32:34	40.19058551	-92.60078167	12	0
Oklahoma	US	2020-05-05 02:32:34	35.88494195	-94.65859267	65	3
Colorado	US	2020-05-05 02:32:34	39.87432092	-104.3362578	1815	68
Idaho	US	2020-05-05 02:32:34	44.89333571	-116.4545247	3	0
Illinois	US	2020-05-05 02:32:34	39.98815591	-91.18786813	40	1
Indiana	US	2020-05-05 02:32:34	40.7457653	-84.93671406	8	1
Mississippi	US	2020-05-05 02:32:34	31.47669768	-91.35326037	149	9
Nebraska	US	2020-05-05 02:32:34	40.52449420000001	-98.50117804	204	3

Above in the data is the original data that we got from the mentioned repositories.

2) We create a final data dataframe with population of each country. We store value for processing, each row would be the date starting at 1/22/2020 and each column is the country sorted by value remaps.

```
[ ] #Code for scraping CSV file names from Github
    links=[]
    if requests.get("https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports").status_code==200:
      divs = BeautifulSoup(requests.get("https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports").content, 'htm
      for idx, div in enumerate(divs):
        url_href=div.select("td.content span.css-truncate.css-truncate-target a.js-navigation-open ")
        if url_href!=[]:
          links.append("https://raw.githubusercontent.com"+url_href[0].get('href')).replace('blob/',''))
    links = links[1:-1]
    links

['https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-22-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-23-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-24-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-25-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-26-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-27-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-28-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-29-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-30-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-31-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/02-01-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/02-02-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/02-03-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/02-04-2020.csv',
 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/02-05-2020.csv',
```

Above is the initialized code we used to scrape the data from the mentioned data repositories.

3) We then create for loop that would remove unwanted columns, the only columns we would need are 'Country/Region', 'Confirmed', 'Deaths' and 'Recovered' and 'Amount of Population' was simultaneously calculated based on the data we scraped.

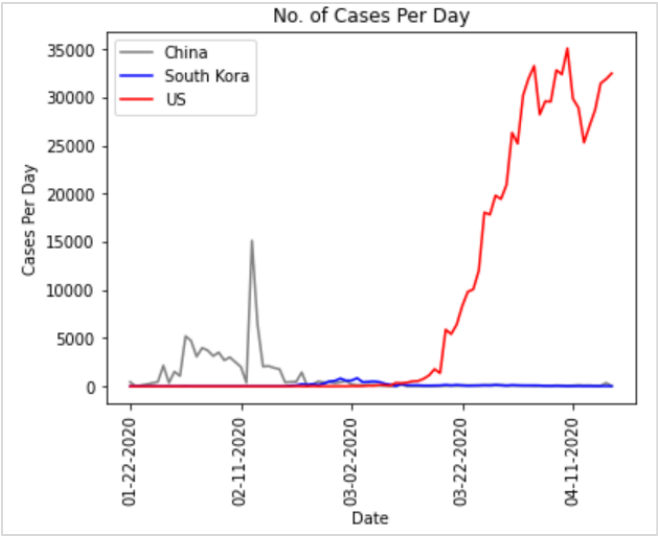
```
finaldata.head(20)
```

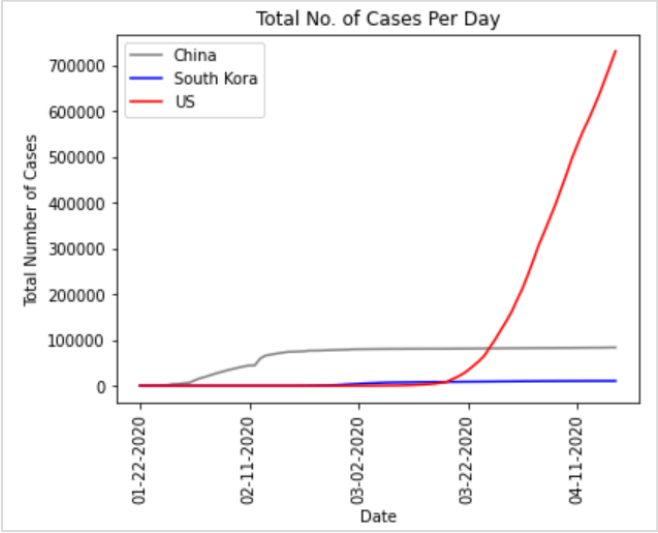
	Country/Region	Confirmed	Deaths	Recovered	Amount of Population
01-22-2020	China	444.0	17.0	28.0	2.878788e-07
01-22-2020	South Korea	0.0	0.0	0.0	0.000000e+00
01-22-2020	Italy	0.0	0.0	0.0	0.000000e+00
01-22-2020	US	0.0	0.0	0.0	0.000000e+00
01-22-2020	France	0.0	0.0	0.0	0.000000e+00
01-23-2020	China	444.0	17.0	28.0	2.878788e-07
01-23-2020	South Korea	0.0	0.0	0.0	0.000000e+00
01-23-2020	Italy	0.0	0.0	0.0	0.000000e+00
01-23-2020	US	0.0	0.0	0.0	0.000000e+00
01-23-2020	France	0.0	0.0	0.0	0.000000e+00
01-24-2020	China	549.0	24.0	31.0	3.564214e-07
01-24-2020	South Korea	0.0	0.0	0.0	0.000000e+00
01-24-2020	Italy	0.0	0.0	0.0	0.000000e+00
01-24-2020	US	0.0	0.0	0.0	0.000000e+00
01-24-2020	France	0.0	0.0	0.0	0.000000e+00
01-25-2020	China	761.0	40.0	32.0	4.971140e-07
01-25-2020	South Korea	0.0	0.0	0.0	0.000000e+00
01-25-2020	Italy	0.0	0.0	0.0	0.000000e+00
01-25-2020	US	0.0	0.0	0.0	0.000000e+00
01-25-2020	France	0.0	0.0	0.0	0.000000e+00

4) We calculate adjusted value for each country and then store it to main list and then calculate percent population and store it in the dataframe.

5) We then start Exploratory Data Analysis (EDA) by preparing dataframes for graphs for the countries US, China and South Korea.

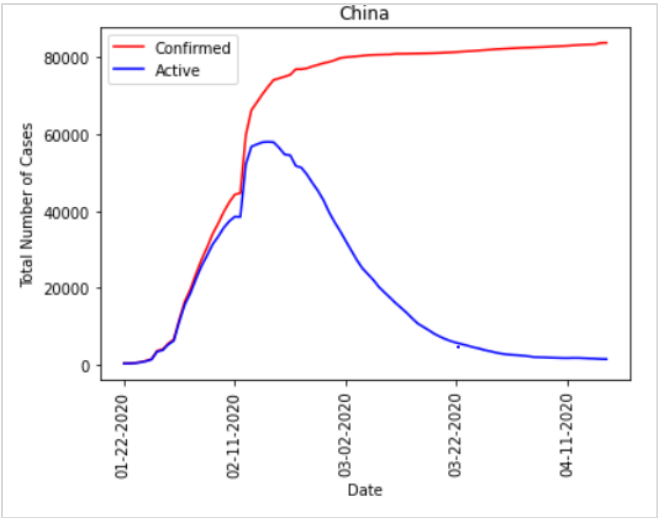
6) We then create graphs for case counts.

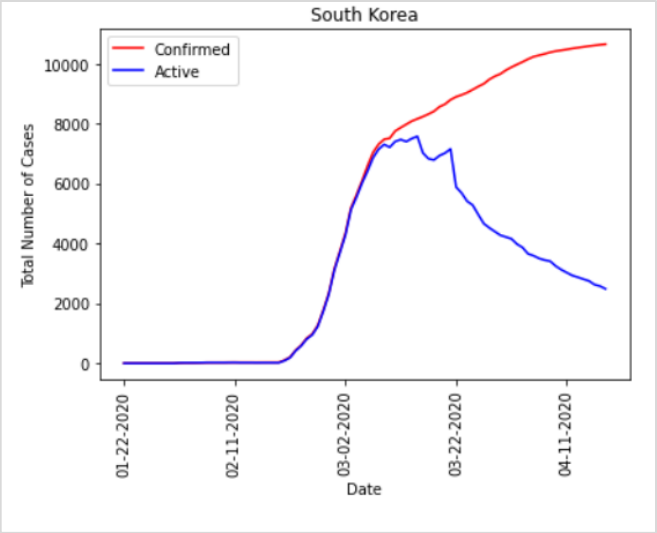


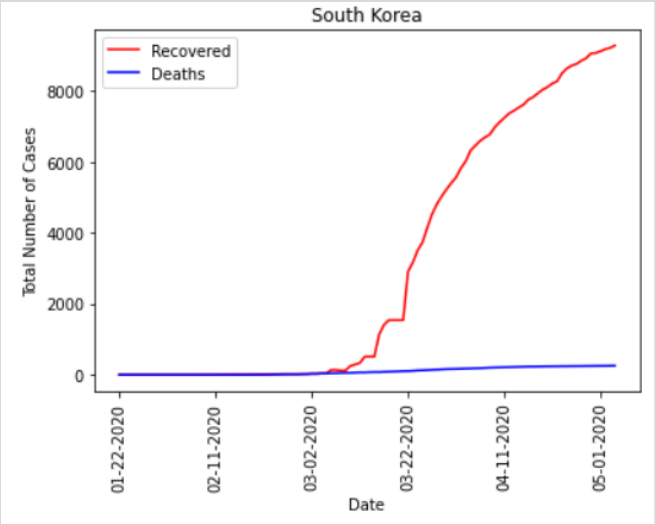


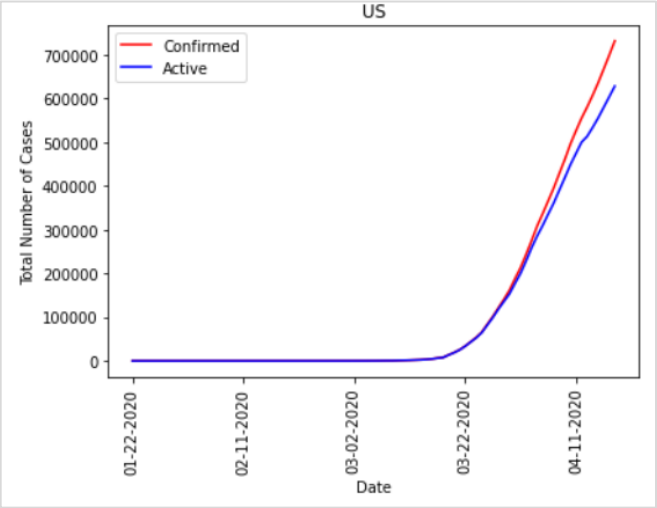
The above two graphs show us the range of cases on a daily and cumulative basis, here we see in the sample space, in the initial days China and South Korea being affected the most but with strong preventive measures decreasing overtime as compared to US which is dormant at first but shoots up overtime and not flattening to date.

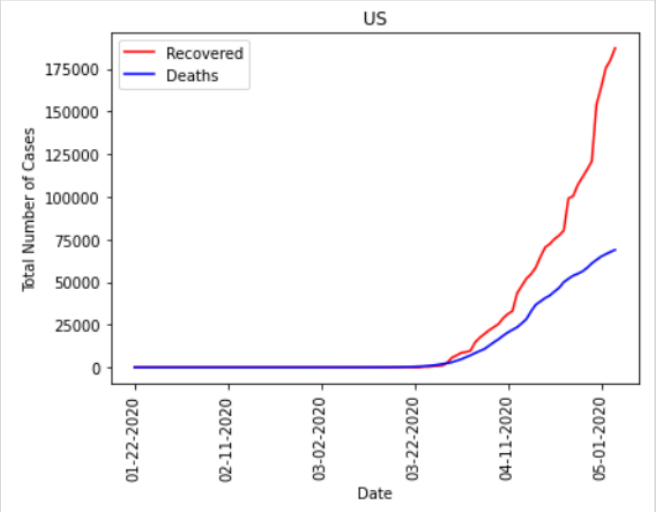
7) We then create line graphs for Confirmed vs. Active Cases.







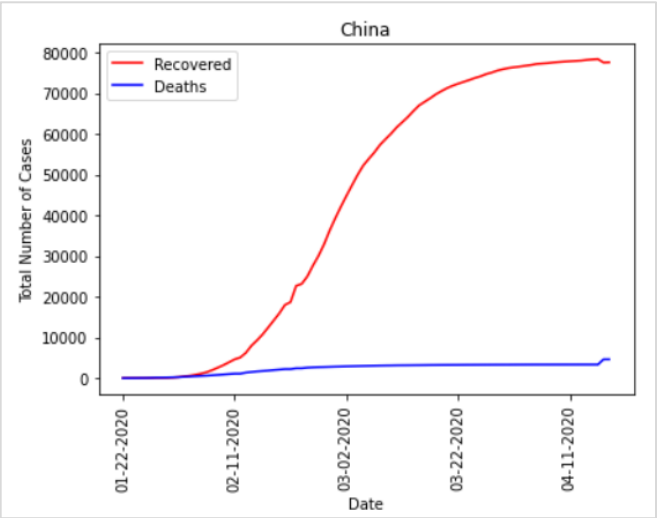




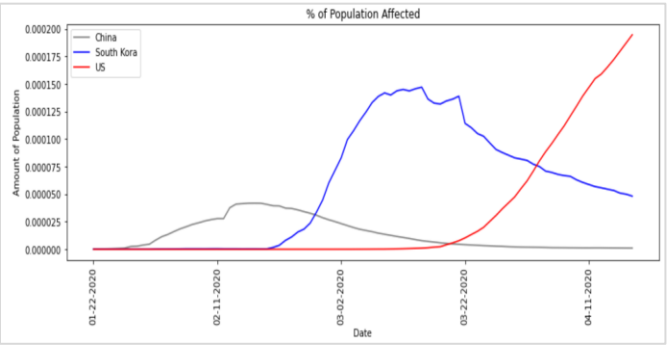
The above graphs show China and South Korea having more gap between confirmed and active cases overtime meaning that the total active cases are far less than the total confirmed cases which not being the case with US as the lines almost converge noting the active and confirmed cases are almost at the same count giving keeping the healthcare industry at risk.

The above graphs show us that that the South Korea and China having less deaths and more recoveries meaning that more and more people are healing as compared to dying which isn't the case with the US at the initial days but is slowly diverging as stronger stay and home measures are imposed and followed.

8) Then we plot line graphs for Recovered vs. Deaths

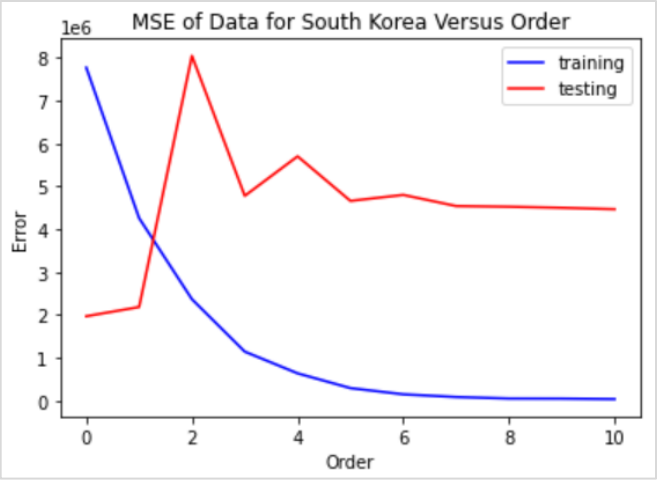


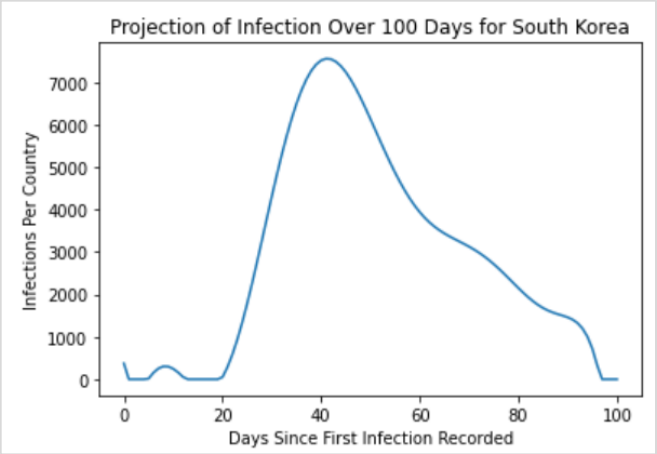
9) Comparing the percentage of population affected in all three countries

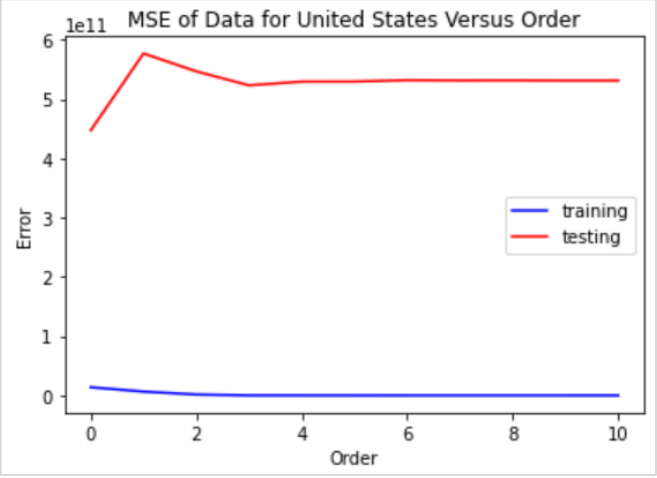


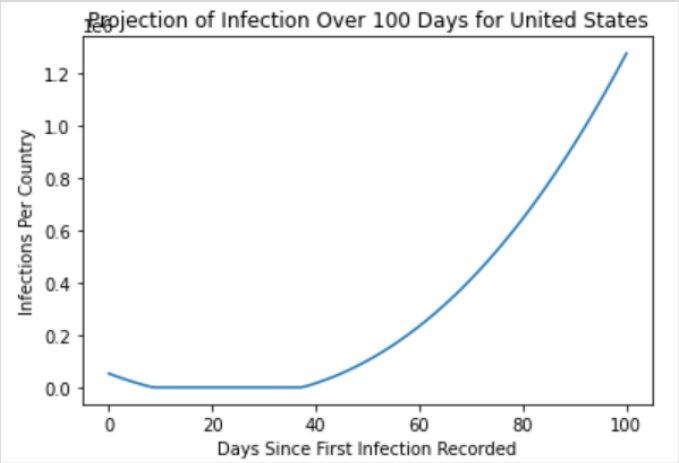
The above graph shows us the percentage of population affected. This graph is relative because the amount of population and spread of the virus also varies. But this graph shows that South Korea and China were most affected during the first and mid of the peak whereas United States increased dramatically in the later stage.

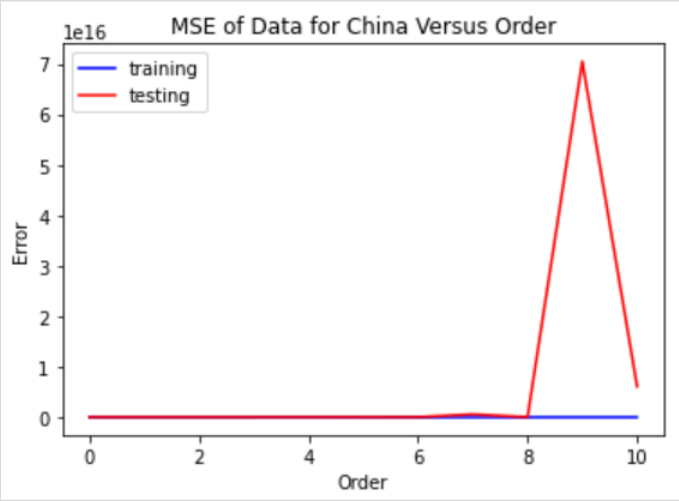
10) We create model equations for South Korea, US and China and generate plots for error values. Plot projection of the two models for the first 50 days after infection appearance. The best order for South Korea is determined to be 10, best order for US is 3 and the best order for China is 3.

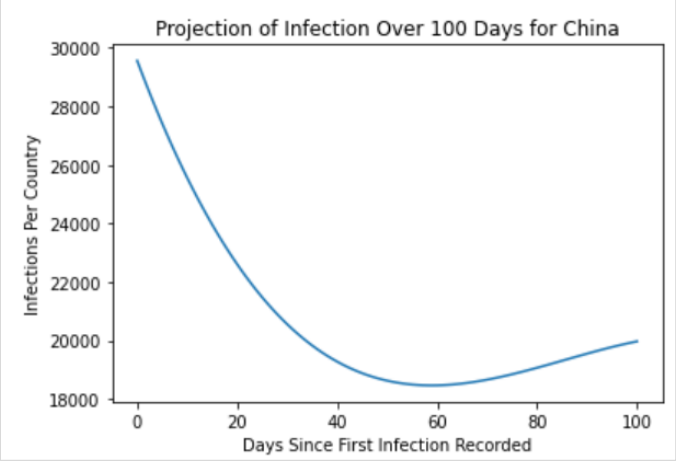












11) Lastly, we calculated the MAE for the three countries against themselves and two of the other countries Italy and France. We stored the data in frame for readability.

IV. RESULTS

We used MAE to calculate and to compare other country datasets (Italy and France) to each of the model and found that most of the models are closer to South Korea and US and China only matched themselves.

	MAE for SK Model	MAE for US Model	MAE for China Model	Closest Model
China	1.648436e+08	2.216293e+06	14382.396325	China
Italy	4.402087e+04	1.981143e+05	44620.883068	South Korea
France	3.345292e+04	2.091153e+05	36125.079431	South Korea
United States	2.273706e+05	3.678000e+04	229620.851238	United States
South Korea	1.637073e+02	2.420274e+05	17740.845599	South Korea

Fig. 3: MAE for other countries vs our model countries.

Italy and France are mostly closest to South Korea model which tells us that they are having almost the same rate of recovery as Korea. If they keep following the same quarantine measures, with the same aggressiveness then these countries are well on their way to have more recovered patients and lesser active cases.

V. LIMITATIONS

If we extended our prediction analysis past 150 days, the model begins to 'crack' or not plot well enough and that could be due to the inconsistencies in the data. This limitation can be overcome as new data becomes available over time; the error rate will decrease to give more accurate results

Code designed for this project is dynamic enough to handle any future changes that are made to the csv(s). It would be interesting to see how the MSE decreases over time and if more countries are added, how would they be conceived on the model and if they are well on their paths of recovery from the pandemic as well.

VI. FUTURE DIRECTIONS

Currently all the results can tell us how close a given country is to it. What this means is that a country that is almost identical to a model will be given the same assessment as one that is further from it but not closer to the other models. Going forward we would like to expand the metric to include relative distance between models. This would be able to give more context to the assessment.

Currently only France and Italy are referenced in the results. This was because initially they had a large infected population meaning they could be used without much concern. With countries showing more total infections the current dataset can be expanded to include them.

After making a quantitative analysis of how close a country is to a given model an analysis should be made dealing with current measures being taken to prevent infection spread. This would involve a comparison of between a given country and the model it closest resembles. From there any changes could be suggested in order to move a country away from a bad model and closer to a good model based on the comparison.

V. REFERENCES

[1] "Coronavirus," *World Health Organization*. [Online]. Available: https://www.who.int/health-topics/coronavirus#tab=tab_1. [Accessed: 20-Apr-2020].

[2] "Coronavirus Cases:" Worldometer. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 20-Apr-2020].

[3] P. Duddu and Praveen, "South Korea Coronavirus (COVID-19) outbreak, measures and impact," Pharmaceutical Technology, 02-Apr-2020. [Online]. Available: https://www.pharmaceutical-technology.com/features/coronavirus-affected-countries-south-korea-covid-19-outbreak-measures-impact/. [Accessed: 20-Apr-2020].

[4] "Beautiful Soup Documentation," Beautiful Soup Documentation - Beautiful Soup 4.9.0 documentation. [Online].

[5]M, "Mean Absolute Error ~ MAE [Machine Learning (ML)]," Medium, 22-Feb-2018. [Online]. Available: https://medium.com/@ewuramaminka/mean-absolute-error-mae-machine-learning-ml-b9b4afc63077. [Accessed: 20-Apr-2020].

[6] "Regression loss metrics on the Peltarion Platform," Peltarion.com. [Online]. Available: https://peltarion.com/knowledge-center/documentation/evaluation-view/regression-loss-metrics. [Accessed: 20-Apr-2020].

[7] Nytimes, "nytimes/covid-19-data," GitHub, 04-May-2020. [Online]. Available: https://github.com/nytimes/covid-19-data. [Accessed: 06-May-2020].

[8] Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020, March 04). Coronavirus Pandemic (COVID-19) - Statistics and Research. Retrieved May 11, 2020, from https://ourworldindata.org/coronavirus.