PROJECT

COVID-19 ANALYSIS

Gowtam Potluri

COURSE ID:DS5020-34143

Abstract

In late December 2019, a previous unidentified coronavirus, currently named as the Coronavirus Disease-2019(COVID-19)[1], emerged from Wuhan, China, and resulted in a formidable outbreak in many cities in China and expanded globally. This report shares views on how the virus spread globally and has an adverse impact on some countries. It sheds some light on proactive measures taken by some countries which helped them curtail the spread. It also discusses a mathematical expression for the model and provides a method to curve fit the data to give insights into the future. To achieve this, Python has been used to process, read and visualize the data. By examining patterns in the data, it highlights how preventive measures would reduce the spread of the disease.

Introduction

From Ebola in West Africa to Zika in South America to MERS in the Middle East, dangerous outbreaks are on the rise around the world. The number of new diseases per decade has increased nearly fourfold over the past 60 years, and since 1996, the number of outbreaks per year has more than tripled[2]. Each of these outbreaks has an impact on its origin country and also globally. For Example, Consider the outbreak of MERS in Saudi Arabia in 2012. Since 2012, MERS has been reported in 27 countries including the United States, Republic of Korea, and Italy[3]. At the end of November 2019, a total of 2494 laboratory-confirmed cases of MERS were reported. Almost after a decade of the outbreak, there is still no specific treatment or a vaccine available for the disease. This is the case with almost all other outbreaks. In case if they have a vaccine, Developing and globally distributing it will likely take 4-6 months (4 months to produce first doses of vaccine, and 6 months to produce enough to give to a large number of people), even while mathematical models demonstrate that virus could spread globally within 6 months[4]. So, it's always better to enforce a framework that provides guidelines on preventive measures in such events of an outbreak. This report analyzes the current data of COVID-19 to provide an insight into how the virus spread from the origin country globally and has an adverse impact worldwide. It highlights how some countries have compelled preventive measures like contact tracing and have significantly reduced the spread of the virus. It also describes an exponential model that could give possible estimations for the near future on the spread of the virus.

Data Acquisition

The data used in this report to process, read and visualize patterns that established some facts on how proactive measures taken at the right time could curtail the spread of the virus is obtained from the GitHub[5] layer of Johns Hopkins University operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Johns Hopkins ensures the integrity of the data by acquiring it from reliable sources like World Health Organization, the U.S. Centers for Disease Control and Prevention, the European Center for Disease Prevention and Control, the

National Health Commission of the People's Republic of China, 1point3acres, Worldometers.info, BNO, state and national government health departments, local media reports, and the DXY, one of the world's largest online communities for physicians, health care professionals, pharmacies and facilities[6]. The Johns Hopkins COVID-19 data is shared in the public domain in the format of a CSV file. This file contains some of the important information about the virus with labels such as Date, Country/Region, Province/State, Lat, Long, Confirmed, Recovered, and Deaths. The information is cataloged day-wise for each country when the first case is recorded. For a few countries, the information is also available for multiple regions. To better visualize the data, It has been processed at multiple levels. First, different regions of a country are brought to a single time-series. For this pandas and pandasql modules were used.

```
def process_data():
    query1="SELECT Date,Country,SUM(Confirmed) as Confirmed,SUM(Deaths) as
Deaths,Lat,Long FROM df WHERE Province NOT NULL GROUP BY Date,Country ORDER BY
Country"
    query2="SELECT Date,Country,Confirmed,Deaths,Lat,Long FROM df WHERE
Province IS NULL"
    query3="SELECT Date,Country,SUM(Confirmed) as Confirmed,SUM(Deaths) as
Deaths,Lat,Long FROM processed_dataset GROUP BY Date,Country ORDER BY Country"
    dataset_with_provience=sql.sqldf(query1)
    dataset_without_proviences=sql.sqldf(query2)

processed_dataset=dataset_with_provience.append(dataset_without_proviences,ig
nore_index=True )

    return sql.sqldf(query3)
```

As shown in the above snippet, the query1, and query2 are used to fetch the data for countries with and without province. The processed_dataset is the dataframe which appends the previous data obtained. Now, query3 is run on the processed_dataset which gives the total number of Confirmed Cases and Deaths per day for all countries bringing them into a single time-series. For performing some analysis only the top 10 countries in terms of the number of Confirmed cases and Deaths are obtained.

```
def get_top10_countries():
    processed_dataset=process_data()
    top10_countries_by_confirmed=(sql.sqldf("select
Date,Country,Max(Confirmed) as Confirmed,Lat,Long from processed_dataset GROUP
BY Country ORDER BY Confirmed DESC LIMIT 10",locals()))
    top10_countries_by_death=(sql.sqldf("select Date,Country,Max(Deaths) as
Deaths,Lat,Long from processed_dataset GROUP BY Country ORDER BY Deaths DESC
LIMIT 10",locals()))

    return
top10_countries_by_confirmed,top10_countries_by_death,processed_dataset
```

Also for the ease of processing, some of the labels in the original dataset like Country/Region and Province/State are changed to Country and Province respectively.

Data Analysis

We live in an interconnected and increasingly globalize world thanks to international jet travel, people and the disease they are carrying can be in any city on the planet in a matter of hours. To understand why and how the COVID-19 spreads its key to look at the timeline of it. The reports of the confirmed cases were increasing at the time of the Chinese new year, the time when many people travel across the country and the world. This is one of the reasons for the spread of COVID-19.

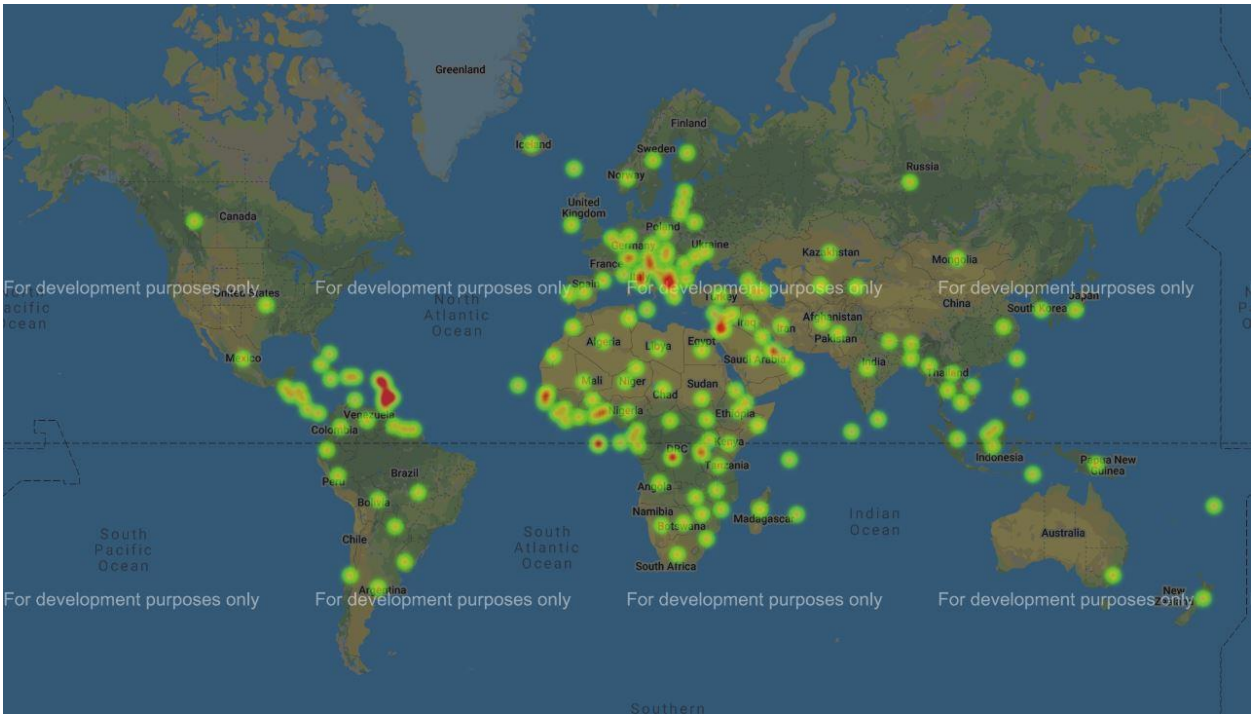

Figure 1: Heat Map of the COVID-19 across the world (NOTE: Data of provinces for few countries are not available.)

With the COVID-19 spread globally, It has impacted most of the countries with some taking the highest toll. To better understand which countries are most affected, The box plot has been used to analyze the global data of COVID-19 and recognize the countries which have significantly large numbers in terms of confirmed cases and deaths.

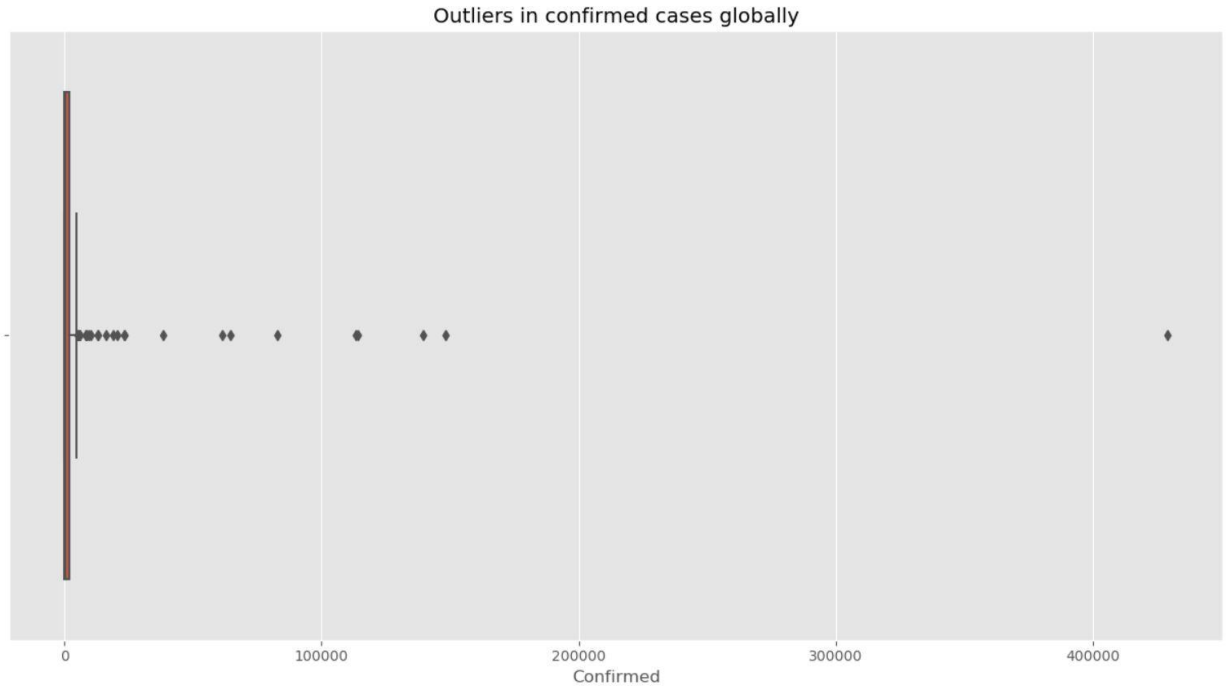

Figure 2: The above box plot shows the outlier, in this case, the US with the highest number of confirmed cases(>400000) as of 8th April 2020

From the above box-plot, we can interpret how the US has the highest number of confirmed cases with 429052 recorded as of 8th April 2020. Similarly, the box-plot of the number of deaths reported globally displays how Italy, unfortunately, has the highest reported cases of deaths with 17669 cases.

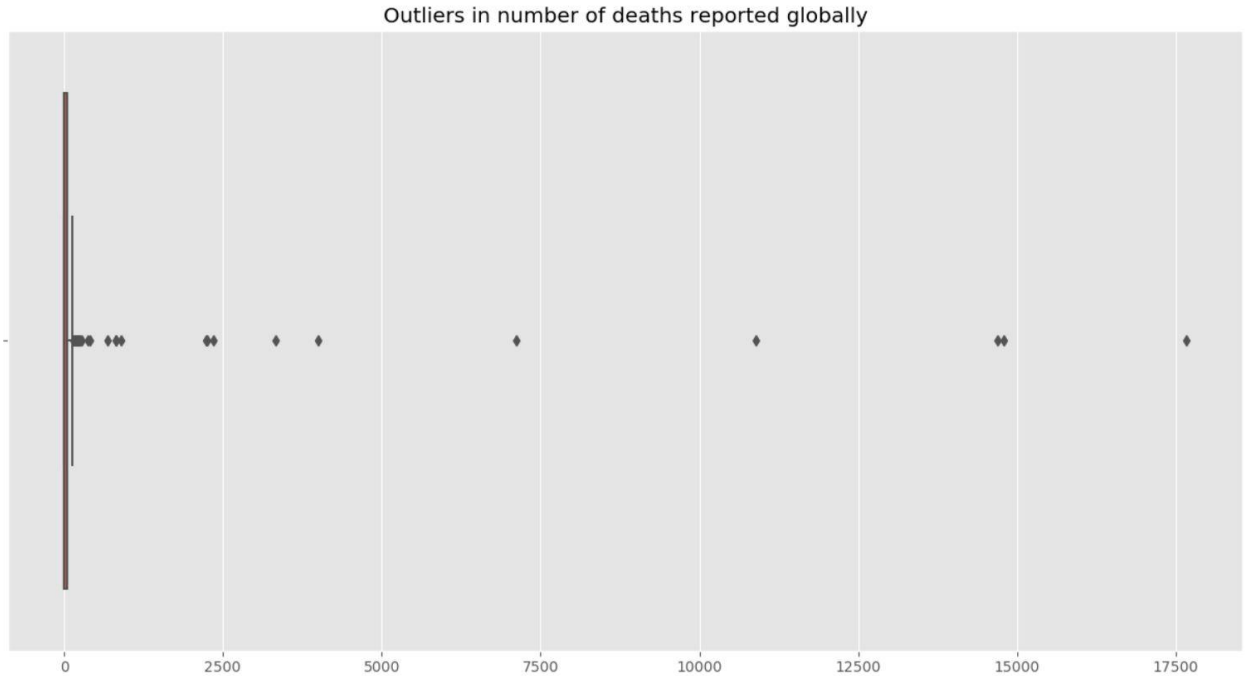

Figure 3: The above box-plot shows the outlier, in this case, Italy with the highest number of reported deaths(>17500) as of 8th April 2020

The reasons for such high numbers in these countries are discussed in the coming sections. Deaths from the COVID-19 in Italy are increasing from the first case reported, with the country reporting 919 deaths in a single day on March 27 — the biggest single-day death toll reported in any country since the start of the outbreak. Italy has the second-highest old population in the world[7]. It is one of the factors that is affecting the country's death rate.

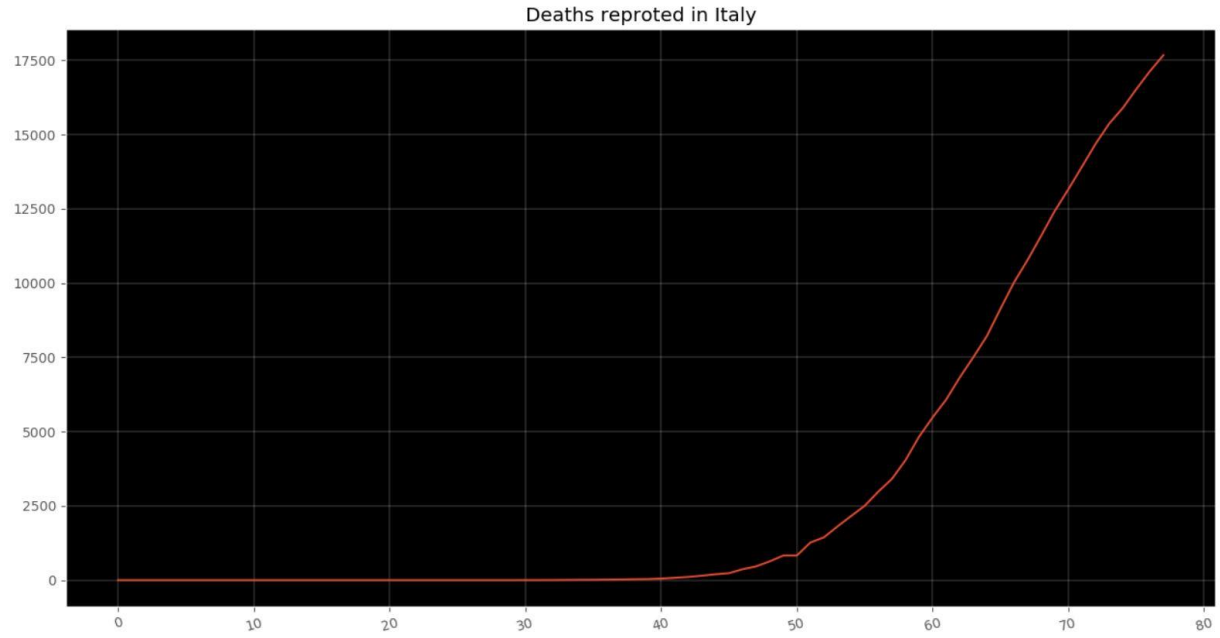

Figure 4: The above time-series graph represent the number of deaths reported in Italy as of 8th April 2020

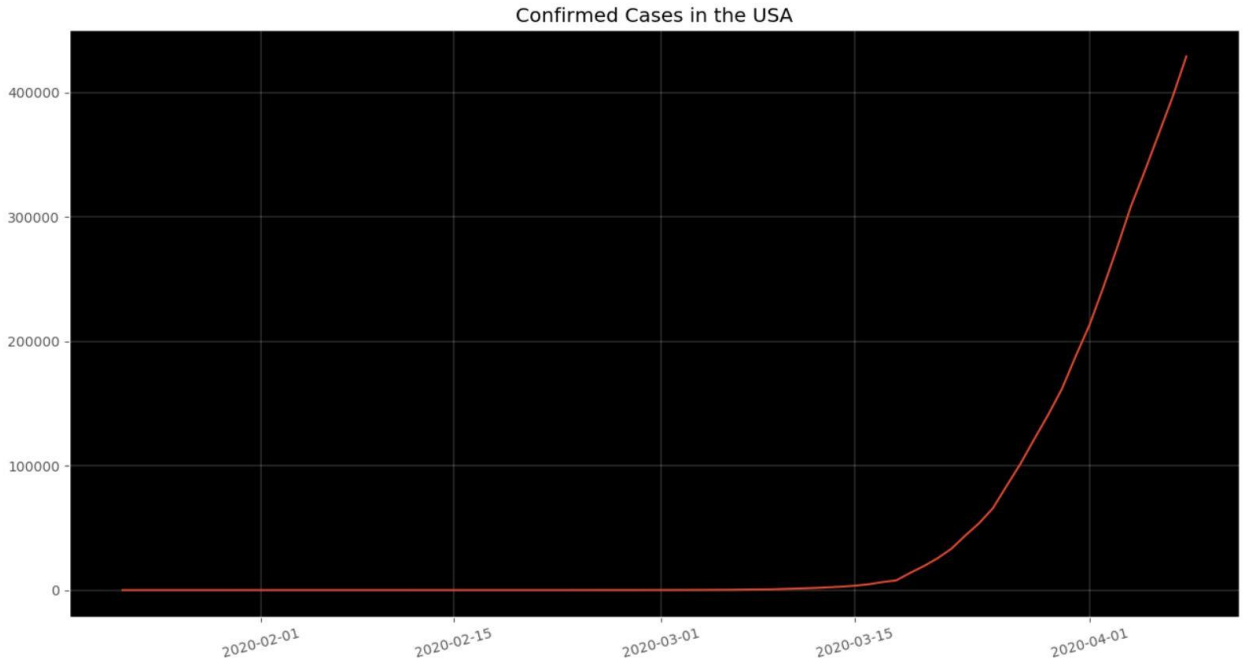

Figure 5: The above time-series graph represent the number of confirmed cases in the USA as of 8th April 2020

The US has leapfrogged the rest of the countries in terms of the number of confirmed cases reported. This was mainly because when the US was reporting very few cases, things were already getting bad under the radar. Due to the mismanagement at the beginning of the spread of COVID-19 in the US and implementing proactive measures like social distancing and lockdown only after a month of first reported case[8] might be the reasons for the highest number of confirmed cases in the world when compared to other countries which have already escaped the exponential growth. Further, we shall discuss how different countries that were affected at the same time as the US, escaped the exponential growth with the implementation of strict preventive measures.

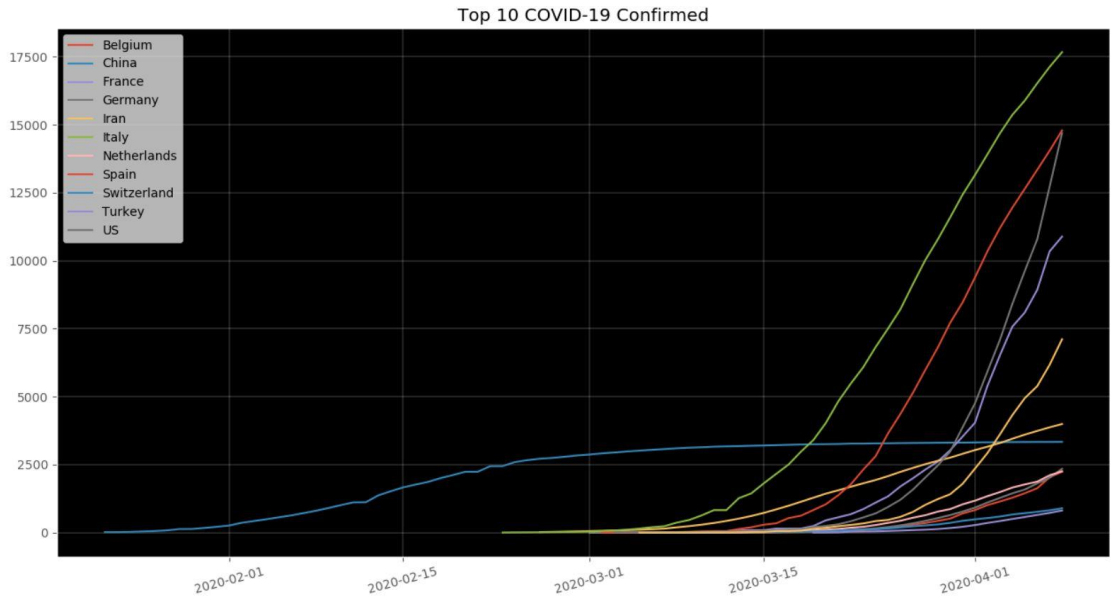

Figure 6: Time-series of top 10 countries in terms of confirmed cases after their 100th case is reported

The above time-series graph helps us understand how some countries like China and, Iran have significantly reduced their new cases. It is also important to understand how South Korea, which has reported its first case on the same day as the US has markedly contained the spread of the virus. To discern this let us go back in time, starting in late February 2020, South Korea was reporting a sharp increase in COVID-19 cases. With over 5000 infected, they were registering some of the highest numbers of confirmed cases in the world but then something changed. While cases in most other countries continued to rise, Korea's numbers started leveling off. In the below visualization, comparing the number of confirmed cases between the US and South Korea(in log scale), look at how the curve starts to bend. It indicated that Korea managed to contain the spread of the virus early on and they were able to do it because they'd learned a lesson from a few years ago, when they fought a different coronavirus outbreak called MERS[9]. The lessons all came into play when the next outbreak took hold in the country. As of February 17th, 2020, There were only 30 confirmed cases of COVID -19. Despite the low numbers, health authorities had already started working with biotech companies to develop testing kits for the COVID-19 and soon, they had thousands of test kits ready to go. By February 29,2020 the number of cases increased dramatically by increasing to 3000. The health authorities had already equipped hospitals with COVID-19 test

kits and when a patient is tested positive they have implemented a contact tracing system. The government had tracked every person the patient might have come in contact with and tested them as well[10]. Many of them were tested positive and are then isolated. The government teaming up with local health authorities have set up more than 600 testing locations that screen as many as 20000 people per day. With this system in place, when anyone tests positive, the government can test and trace their contacts to continue to break the transmission chains of the COVID-19 on large scales. By examining the trends in key features, indicates how mismanagement at the beginning of the spread of COVID-19 in the US like not enforcing a nationwide lockdown, and making test kits available had such a huge impact on the increase of the number of cases.

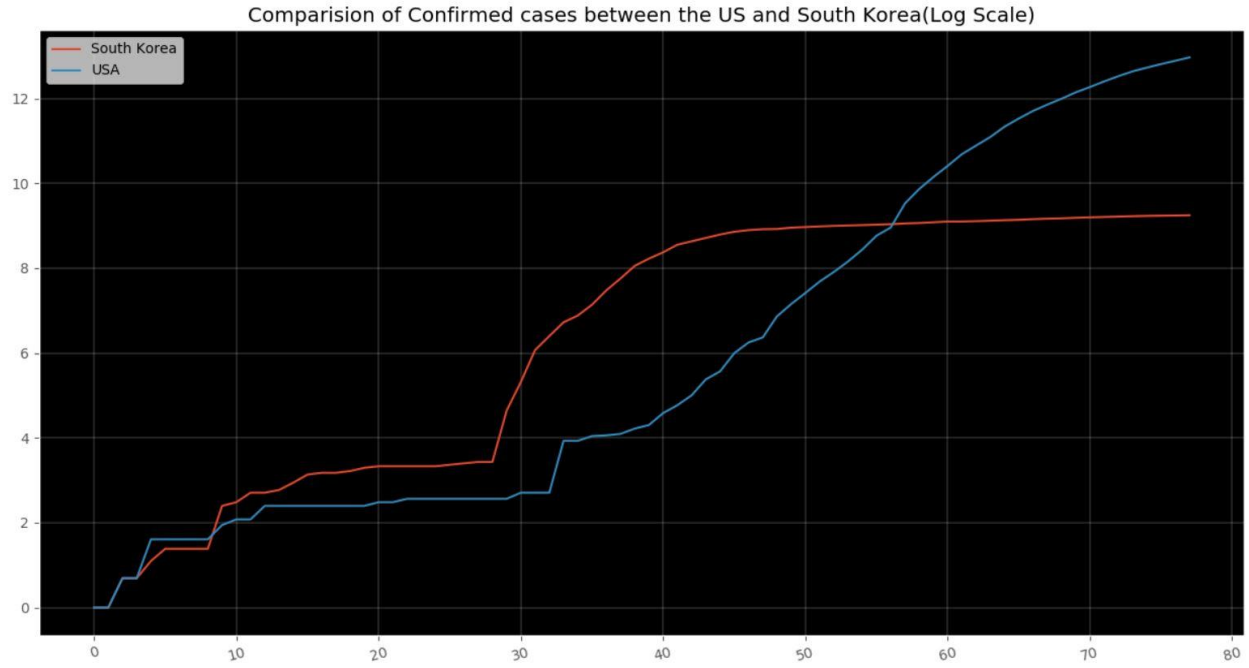

Figure 7: Time Series comparison of the confirmed cases between the US and South Korea(Y-axis- cases reported in log scale, X-axis – time in the number of days

To better understand how social distancing can help curb the spread of the virus, a model has been developed which can give insights into the near future on the possible number of increase in the cases. Since it is an exponential model, testing the predicted numbers against a country that has implemented social distancing at the early stage of the spread of the virus can help us understand how preventive measures can largely restrict the spread of the virus. The Exponential Growth model is considered for modeling the COVID-19 outbreak as epidemiologists have studied that such type of outbreaks in the first period of an epidemic follows Exponential Growth. The formula for exponential growth is $y(t) = a * exp(k * (t - t_0))$ where a is the number of cases at the beginning of the outbreak. y(t) is the number of cases at any given time t-t0 and k is the growth factor. However, when processing the data, it only has the number of cases per day and not the growth factor. The best method to find the growth factor from empirical daily observations is to use a statistical model called Linear Regression. Linear regression allows us to estimate the best values for a and b in the following formula, $y = a + b * x$. So, we have to rewrite the formula in

the form that has the shape of the Linear Regression. Logarithms allow us to rewrite the function in the correct form:

$$Y = c + m * x$$

$$c \Rightarrow \log(a) \qquad m \Rightarrow \log(k) \qquad x \Rightarrow \log(t)$$

*** For simplicity, the time is considered as the number of days since the first case is reported**

We now use the log of the Confirmed cases instead of the Confirmed cases. So, in python, the log transformation is applied to the Confirmed column as shown below.

```
qrr['logInfo']=np.log(qrr['Confirmed'])
qrr['Time'] = np.arange(len(qrr))
```

Now, the numpy modules polyfit function is used to curve fit the data which returned the coefficient values for the equation. Let's convert the equation back by applying exponential.

$$a=0.1594$$
$$k=-2.232$$

$$\log(a)= 1.1728$$
$$\log(k) = 0.09$$

$$y(t) = 1.17*0.09^t$$

So, for any given t(day) the model can predict the possible number of confirmed cases. However, there are some things to be considered. It's important to note that the predictions shown below are only an example to show how statistics can be used in epidemiology. In real life different models would be tested besides exponential growth which is not done for this model This Linear model is only the best estimate of exponential growth function, it has a certain error margin. The exponential growth function is not necessarily the perfect representation of the epidemic. Now to test the model, India's data for COVID-19 is considered. Below we can see the curve fit for the data. The red dots are the actual data of the Confirmed cases in India and the blue line is the best fit for the model.

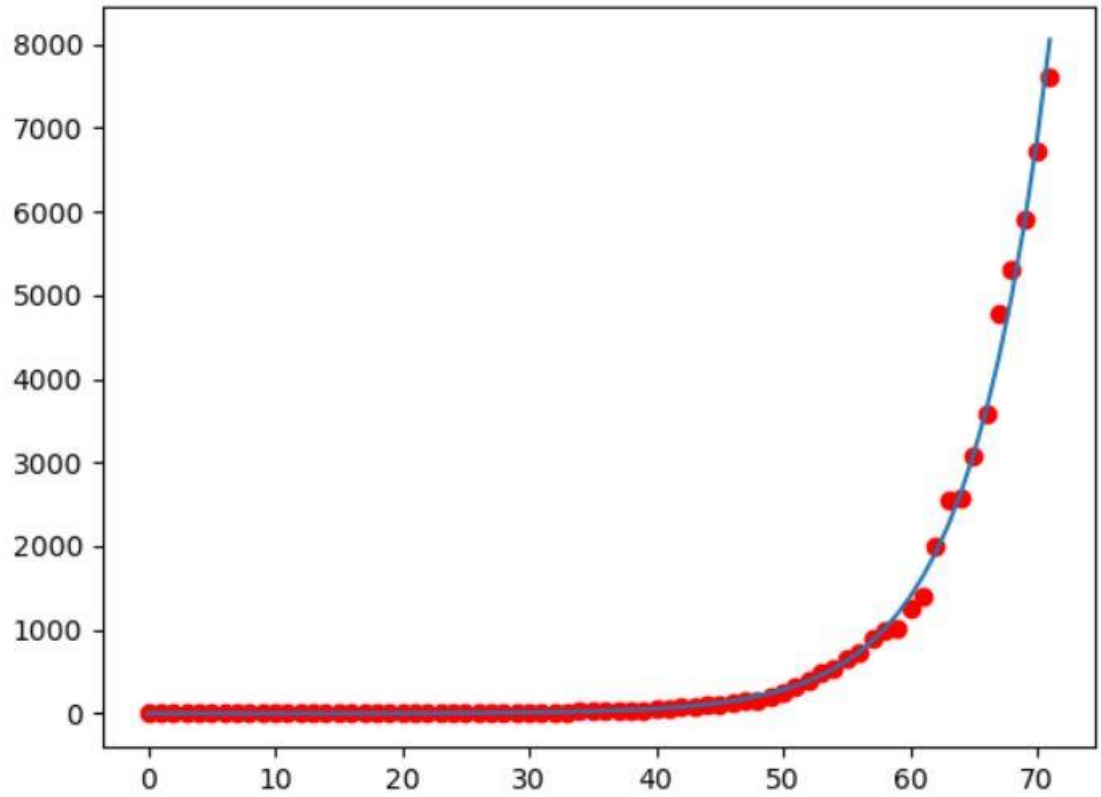

Figure 8: Curve Fitting of the COVID-19 cases in India

```
C:\Users\gowta\Desktop\study\python\project>python tt.py
    Unnamed: 0                 Date Country  Confirmed  Deaths   Lat  Long   logInfo  Time  Predictions
0         6248  2020-01-30 00:00:00   India        1.0     0.0  21.0  78.0  0.000000     0     0.097944
1         6249  2020-01-31 00:00:00   India        1.0     0.0  21.0  78.0  0.000000     1     0.114869
2         6250  2020-02-01 00:00:00   India        1.0     0.0  21.0  78.0  0.000000     2     0.134719
3         6251  2020-02-02 00:00:00   India        2.0     0.0  21.0  78.0  0.693147     3     0.158000
4         6252  2020-02-03 00:00:00   India        3.0     0.0  21.0  78.0  1.098612     4     0.185303
..         ...                  ...     ...        ...     ...   ...   ...       ...   ...          ...
67        6315  2020-04-06 00:00:00   India     4778.0   136.0  21.0  78.0  8.471777    67  4257.610213
68        6316  2020-04-07 00:00:00   India     5311.0   150.0  21.0  78.0  8.577535    68  4993.356972
69        6317  2020-04-08 00:00:00   India     5916.0   178.0  21.0  78.0  8.685416    69  5856.246251
70        6318  2020-04-09 00:00:00   India     6725.0   226.0  21.0  78.0  8.813587    70  6868.249224
71        6319  2020-04-10 00:00:00   India     7598.0   246.0  21.0  78.0  8.935640    71  8055.133850

[72 rows x 10 columns]
9447.120979273603
```

The above image shows the accuracy of the model. As we can see how close the predictions column match with the confirmed cases column.

Conclusion

In this report, we have discussed how easy it is in modern times for the virus to become an outbreak and create a global pandemic. It also emphasizes on how proactive measures enforced at the right time could curtail the spread of the virus significantly. With the comparison between the USA and

South Korea, it was distinctly shown how South Korea with its quick response for social distancing and other preventive measures was able to curb the spread of the virus while the USA reporting its first case on the same day as South Korea is recording the highest number of cases in the world. So, as the saying "Prevention is better than cure." In such events of a pandemic, it is important to enforce strict measures at the right time to help reduce the spread of the virus.

---

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it
[2] https://www.who.int/csr/don/archive/year/en/
[3] https://www.who.int/csr/disease/coronavirus_infections/faq/en/
[4] https://www.historyofvaccines.org/index.php/content/articles/vaccines-pandemic-threats
[5] https://github.com/CSSEGISandData/COVID-19
[6] https://coronavirus.jhu.edu/map-faq
[7] https://www.prb.org/countries-with-the-oldest-populations/
[8] https://www.aljazeera.com/news/2020/03/emergencies-closures-states-handling-coronavirus-200317213356419.html
[9] https://www.who.int/westernpacific/emergencies/2015-mers-outbreak
[10] https://www.who.int/csr/disease/coronavirus_infections/technical-guidance-contact/en/