Project Report
COVID-19 World Wide Visualizations and Predictions

Aakanksha Duggal, Harshal Savla, Priyanka Debnath
duggal.aa@husky.neu.edu
savla.h@husky.neu.edu
debnath.p@husky.neu.edu

Abstract

Late 2019 witnessed the birth of the current pandemic, CoronaVirus (COVID-19). It started from China, but has rapidly spread ever since then it has been multiplying rapidly with space and time. Almost 220 countries have been affected and don't seem to stop anytime soon. By the looks of it, it might look like a medical problem, but as a Data Scientist, what is the best that can be done? Overcoming this pandemic is the need of the hour, for which important decisions need to be made after understanding the underlying data. The aim of this project is to provide some insights, visualizations and predictions that can allow the government to take appropriate measures in dealing with this situation. The results show the sequential increase in infected numbers of people and deaths. However, the expectation is to find a way to flatten the curve.

Introduction

Problem Statement

The barrage of stats surrounding the CoronaVirus pandemic can be overwhelming and difficult to parse. Every day, public health officials release frightening new numbers, and it's hard to understand if any of the draconian measures being taken have any chance of success. And, actually, it's apparently still hard for some media consumers, who may be getting flawed and misleading information from various corners of the internet, to even understand why such measures are not a total over-reaction.

We are trying to visualize these enormous chunks of data that we have to correctly understand how this lethal virus spreads, whether factors like demographics are at all involved in determining which areas around the world are seeing growing numbers and to see a comprehensive analysis of these numbers helping infer any trends about the spread, if at all.

The intent is to investigate the curve through visualization and make cogent predictions about what is next.

Importance of the Solution

These visualizations and predictions help us understand the logic as to why social-distancing seems to be working, which is extremely important since it might be our most potent ammunition against this invisible enemy virus.

News outlets are good sources for specific growth trends and day-by-day increases, particularly when there's local context for how a region is responding. But if we want to track the raw mathematical progress of the pandemic, we may need something more specific. There are plenty of raw figures to look at — tests executed, confirmed cases, hospitalizations, deaths — and each one can be tracked over time or against total population numbers. If we look at the numbers right, we can get a sense of how well a particular region is doing at containing an outbreak. But that takes the right graph and the right perspective on exactly what the numbers mean.

The math involved in these visualizations is significant since it says that the spread can be slowed, public health professionals say, if people practice "social distancing" by avoiding public spaces and generally limiting their movement.

What is already known

The original cases in humans appeared around the city of Wuhan, in central Hubei province in China, where scientists identified this 2019-nCoV.

- In November 2015, studies predict that this virus could be a potential threat to humans by contracting the bat SARS-like CoV SHC014.

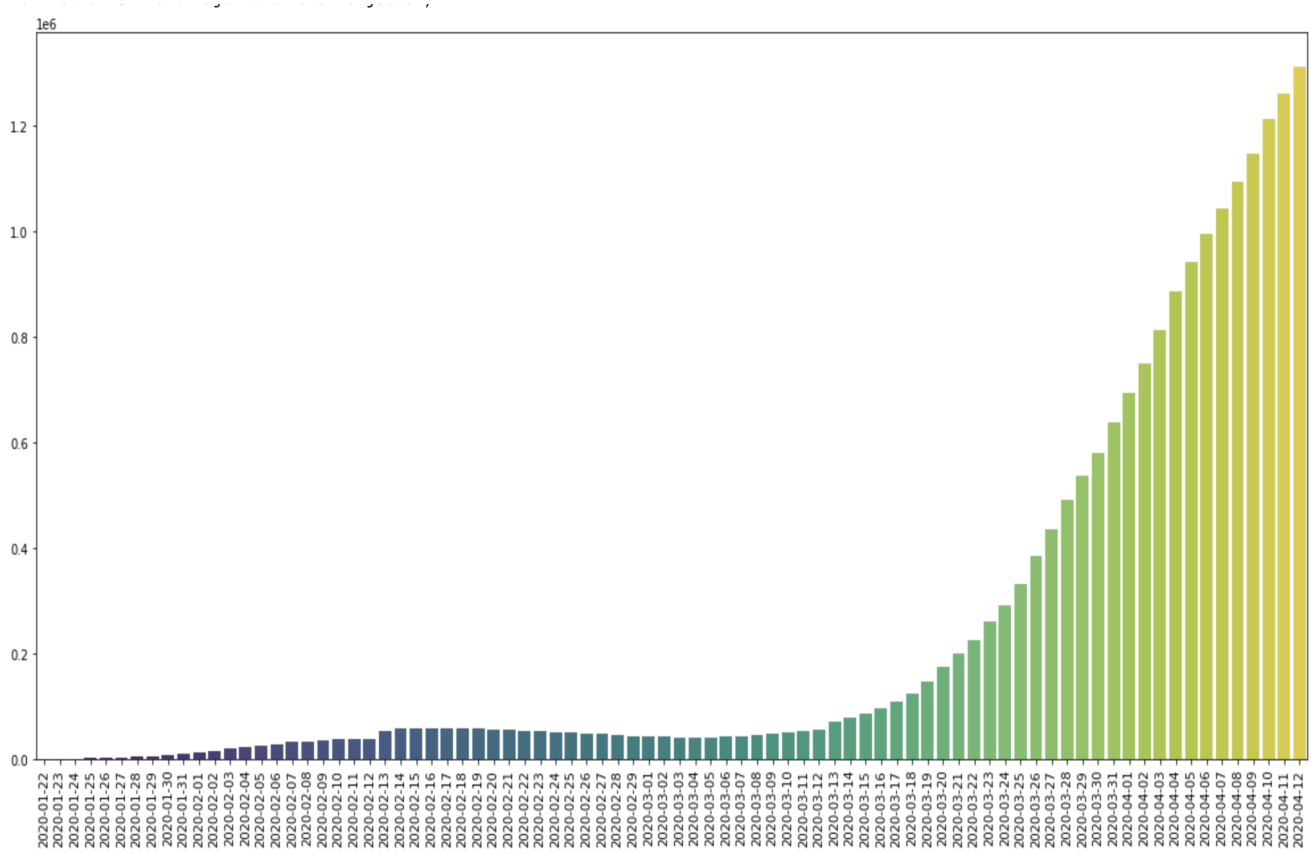


Figure 1: Growth of CoronaVirus

- Late November 2019, origin of COVID-19 is suspected to be the bat crossover to humans.

- On January 9th, sequencing data for the 2019-nCoV virus were released. It is discovered as a novel CoronaVirus by deep sequencing and etiological investigations by at least 5 independent laboratories of China.

- A number of cases have also been confirmed in Thailand, Japan, Taiwan, South Korea, USA, and Europe over the next few months. Although it was believed to be originally transmitted from animals to humans, human to human infections are prominent. On January 24th, there were 830 reported cases and twenty six confirmed deaths from 2019-nCoV, which could be far from the real numbers. On January 26th, models suggest there could be 30,000 - 200,000 humans with 2019-nCoV.

- On February 21st, COVID-19 is thought to be primarily transmitted from one person to another via respiratory droplets, and no evidence of airborne transmission. Asymptomatic carrier transmission has been suspected.

- On March 3rd, WHO published a 40-page report of the WHO-China Joint Mission on COVID-19, with detailed findings on the virus and the disease, assessment of current situation, and recommendations for global communities.

- On March 11th, COVID-19 was characterized by WHO as a pandemic. Social Distancing has been the general response to control the disease. It used to slow the spread of the disease in order to maintain critical care capacity for the sickest of patients. Social distancing is hard to implement and a common question is how much and for how long to implement it.

Because of the frequency with which newer data becomes available, they may not reflect the exact numbers reported, state and local government organizations or the news media. Numbers may also

fluctuate as agencies update their own data.
These are extensively evolving circumstances and everyday, we learn a little bit more about this silent killer.

Methodology

COVID-19 is changing life as we know it across the world. According to the WHO, there are more than 1.8 million confirmed cases of people with COVID-19 and more than 117,000 people have died from the disease—a death toll that has far surpassed that of the severe acute respiratory syndrome (SARS) epidemic that occurred in 2002 and 2003. We're trying to gain some insight on how the virus spread and how it has affected different countries.

Using a dataset acquired from Johns Hopkins University, we intend to demonstrate certain aspects like the rate of growth of number of cases, mortality rate, perform country wise analysis and compare the differences, use supervised machine learning algorithms like linear regression and SVM (Support Vector Machine) to perform predictions in order to study the impact of COVID-19 in the coming days.

Code

Data cleaning and Preprocessing

- Removing irrelevant columns

```
data.drop([“SNo”],1,inplace=True)
```

- Changing to Datetime

```
data[“ObservationDate"]=
pd.to_datetime(data["ObservationDate"])
```

- Grouping the data according to the Observation Date

```
date_wise_data = data
.groupby(["ObservationDate"])
.agg({"Confirmed":’sum’,"Recovered":’sum’,
"Deaths":’sum’})
```

Visualization

Calculating Growth Factor

- Growth factor is the factor by which a quantity multiplies itself over time.
- The formula used is: Formula: Every day's new (Confirmed,Recovered,Deaths) / new (Confirmed,Recovered,Deaths) on the previous day.
- A growth factor above 1 indicates an increase in corresponding cases.
- A growth factor above 1 but trending downward is a positive sign, whereas a growth factor constantly above 1 is the sign of exponential growth.
- A growth factor constant at 1 indicates there is no change in any kind of case.

```
daily_increase_confirmed=[]
daily_increase_recovered=[]
daily_increase_deaths=[]
for i in range(date_wise_data.shape[0]-1):
    daily_increase_confirmed.append(((date_wise_data["Confirmed"].iloc[i+1]/date_wise_data["Confirmed"].iloc[i])))
    daily_increase_recovered.append(((date_wise_data["Recovered"].iloc[i+1]/date_wise_data["Recovered"].iloc[i])))
    daily_increase_deaths.append(((date_wise_data["Deaths"].iloc[i+1]/date_wise_data["Deaths"].iloc[i])))
daily_increase_confirmed.insert(0,1)
daily_increase_recovered.insert(0,1)
daily_increase_deaths.insert(0,1)
```

Code : Calculating Growth Factor

Country-wise analysis

- Here we're trying to depict the comparison of top 15 countries in terms of number of confirmed cases and number of deaths.
- Countries like France, Spain, USA, China, Italy, Mexico, UK, Turkey, Germany, Thailand which have a high number of tourists or a high number of International Students are the most affected countries because of COVID-19.

```
country_wise_data=data[data["ObservationDate"]==data["ObservationDate"]
                       .max()].groupby(["Country/Region"])
                       .agg({"Confirmed":'sum',"Recovered":'sum',"Deaths":'sum'})
                       .sort_values(["Confirmed"],ascending=False)
country_wise_data["Mortality"]=(country_wise_data["Deaths"]/country_wise_data["Confirmed"])*100
country_wise_data["Recovery"]=(country_wise_data["Recovered"]/country_wise_data["Confirmed"])*100

fig, (ax1, ax2) = plt.subplots(1, 2,figsize=(27,10))
top15_confirmed =country_wise_data.sort_values(["Confirmed"],ascending=False).head(15)
top15_deaths = country_wise_data.sort_values(["Deaths"],ascending=False).head(15)
#Top15 for confirmed
sns.barplot(x=top15_confirmed["Confirmed"],y=top15_confirmed.index,ax=ax1, palette="viridis")
ax1.set_title("Top 15 countries as per Number of Confirmed Cases")
#Top15 for deaths
sns.barplot(x=top15_deaths["Deaths"],y=top15_deaths.index,ax=ax2,palette="viridis")
ax2.set_title("Top 15 countries as per Number of Deaths")
```

Code : Countrywise Analysis

Predictions

LinearRegression

- In the snippet, the data is split into train and validation sets.

```
date_wise_data["Days Since"]=date_wise_data.index-date_wise_data.index[0]
date_wise_data["Days Since"]=date_wise_data["Days Since"].dt.days

train_ml=date_wise_data.iloc[:int(date_wise_data.shape[0]*0.90)]
valid_ml=date_wise_data.iloc[int(date_wise_data.shape[0]*0.90):]
model_scores=[]

lin_reg=LinearRegression(normalize=True)

lin_reg.fit(np.array(train_ml["Days Since"]).reshape(-1,1),np.array(train_ml["Confirmed"]).reshape(-1,1))
```

Code : Linear Regression

- The training data is used to train the linear regression model, and is validated against the validation set.

```
prediction_valid_linreg=lin_reg.predict(np.array(valid_ml["Days Since"]).reshape(-1,1))

model_scores.append(np.sqrt(mean_squared_error(valid_ml["Confirmed"],prediction_valid_linreg)))
print("Root Mean Square Error for Linear Regression: ",np.sqrt(mean_squared_error(valid_ml["Confirmed"],prediction_valid_linreg)))
```

Code : Linear Regression Prediction

Support Vector Machine

- Use the same data that was split for linear regression and Predict the results.

```
#Intializing SVR Model and with hyperparameters for GridSearchCV
svm=SVR(C=1,degree=6,kernel='poly',epsilon=0.01)

#Performing GridSearchCV to find the Best Estimator
svm.fit(np.array(train_ml["Days Since"]).reshape(-1,1),np.array(train_ml["Confirmed"]).reshape(-1,1))
```

Code : Support Vector Machine

```
prediction_valid_svm=svm.predict(np.array(valid_ml["Days Since"]).reshape(-1,1))

model_scores.append(np.sqrt(mean_squared_error(valid_ml["Confirmed"],prediction_valid_svm)))
print("Root Mean Square Error for Support Vectore Machine: ",np.sqrt(mean_squared_error(valid_ml["Confirmed"],prediction_valid_svm)))
```

Code : Support Vector Machine Prediction

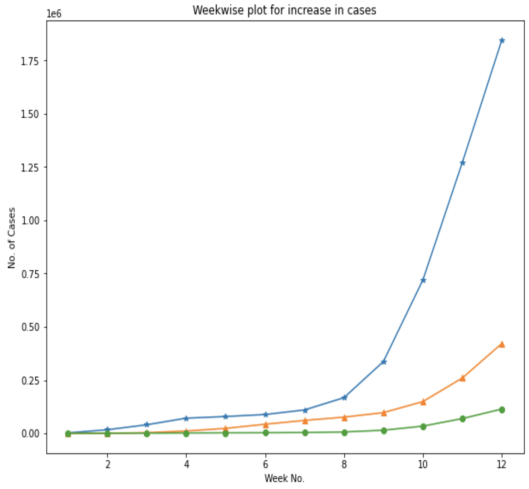
Results

Since the motive behind this project was to find some insights, visualizations and predictions in order to help community take appropriate measures. Following are some major observations and takeaways from the project :

Week-wise Analysis for Confirmed cases, Recovered cases and Deaths.

- Confirmed cases increase rapidly after Week 5.
- Growth rate of the Recovered case is observed better after Week 6.
- Deaths have seen a significant increase after Week 10.

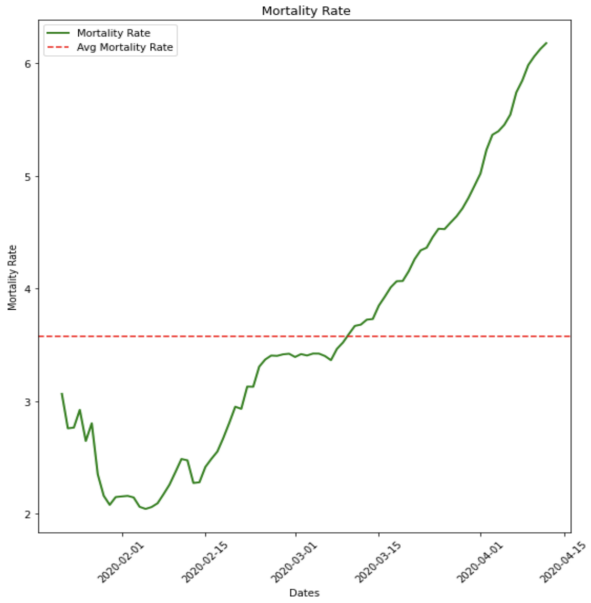
```
Blue : Week wise confirmed cases
Orange : Week wise recovered cases
Green : Week wise deaths
```



Graph 1: Week-wise Increase in Confirmed cases, Recovered cases and Deaths

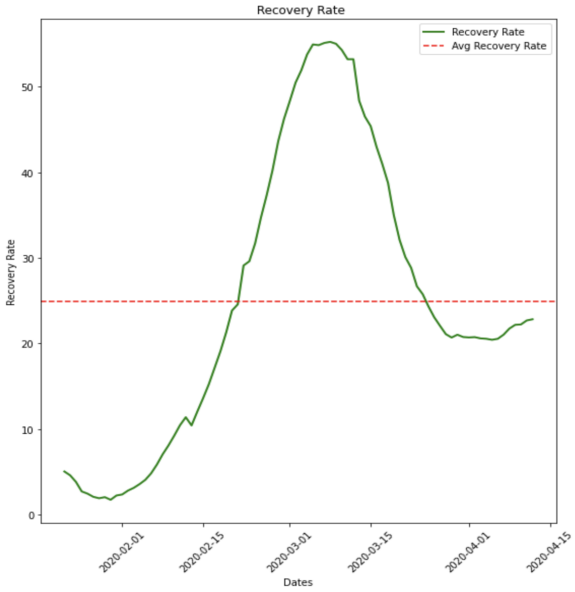
Mortality and Recovery Rate

- Mortality rates rise beginning of April. They increase so rapidly, crossing the Average Mortality rate in April first week.



Graph 2: Mortality Rate

- Recovery rate looks like a bell curve and achieves a peak around early March.
- With increase in cases the recovery has not been that fast.
- It has been slower with the increase in number of infected cases.

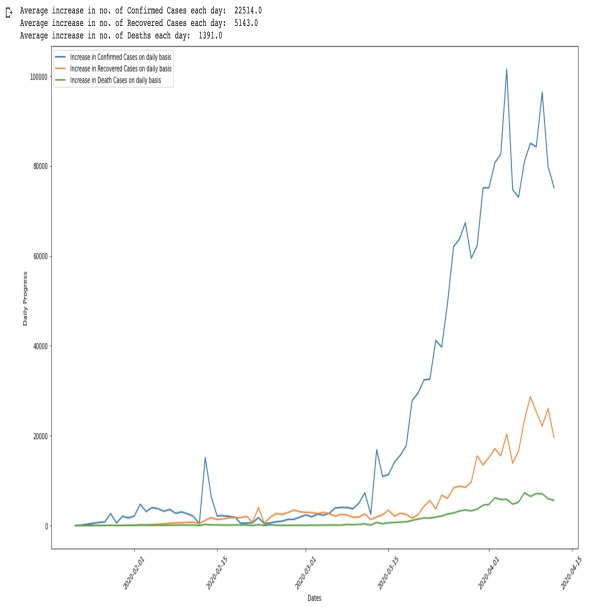


Graph 3: Recovery Rate

Daily increase in Confirmed, Recovered and Death Cases

- Daily increase in Confirmed Cases is more erratic and intense than daily increase in Recovered cases, which says how serious this pandemic is.

- Deaths seems to be a comparatively more stable curve

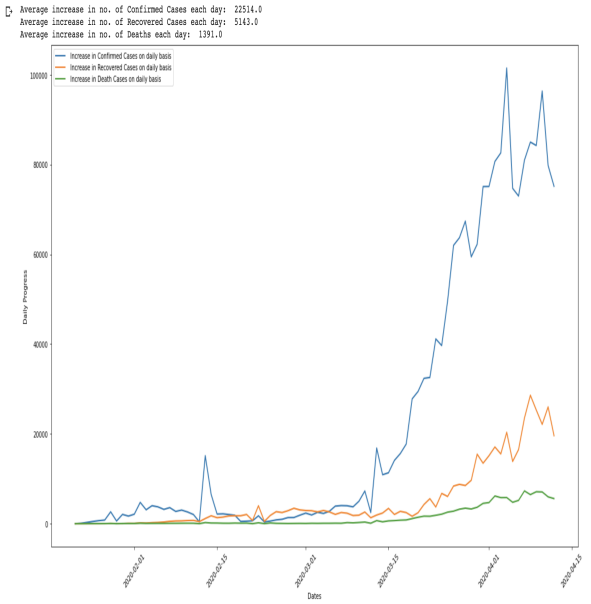


Graph 4: Daily increase in Confirmed, Recovered and Death Cases

Growth Factor for Confirmed, Recovered and Death Cases

- Growth Factor for all the different kind of Cases were pretty high and unpredictable around the beginning of Feb which is when this spread had started to take its shape towards a global pandemic.

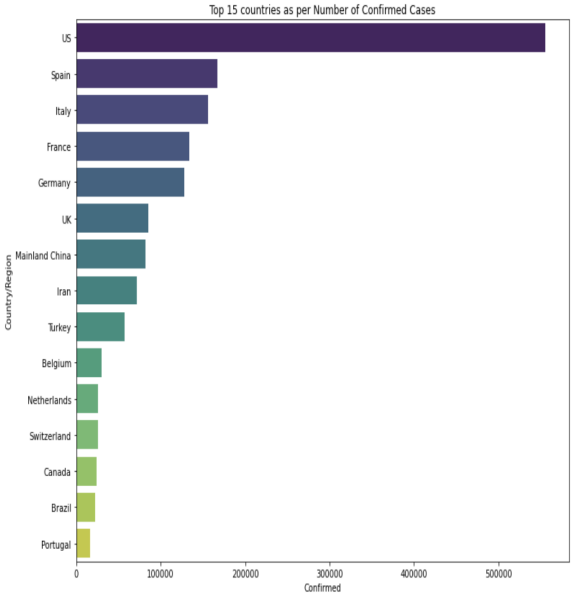
- It has since stabilized since the past 2 months.



Graph 5: Growth Factor for Confirmed, Recovered and Death Cases

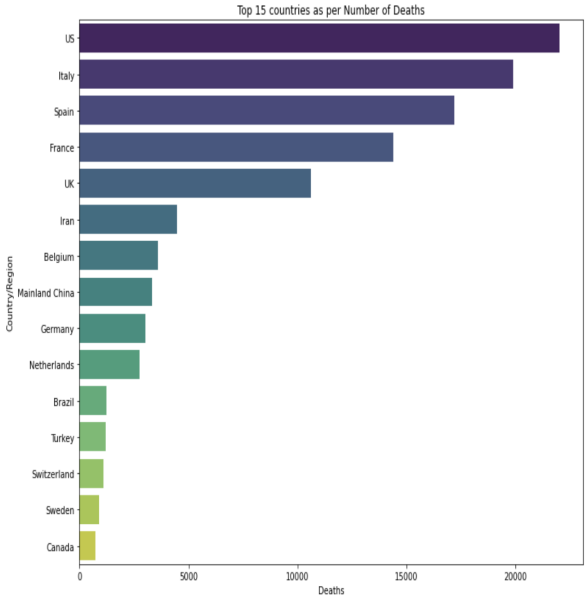
Countrywide Analysis

- US infections are by far the worst, as we see both the highest number of confirmed cases vs the deaths.



Graph 6: Top 15 Countries with Most Confirmed Cases

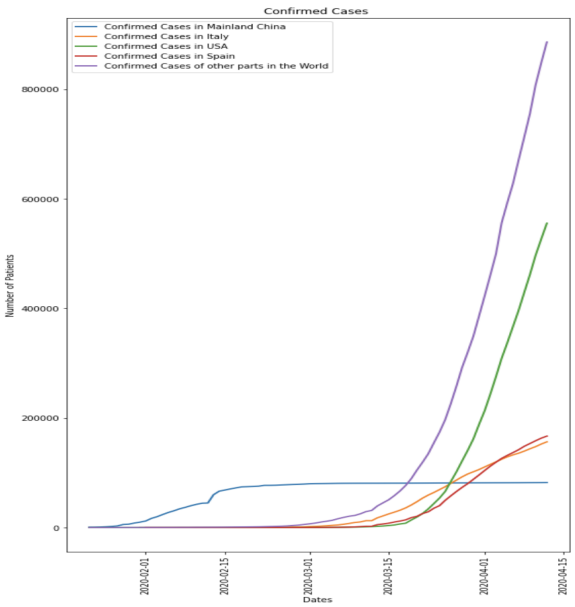
- Fewer number of deaths compared to the number of confirmed cases, the ratio is not as outweighed for US vs other countries as is the confirmed cases number.



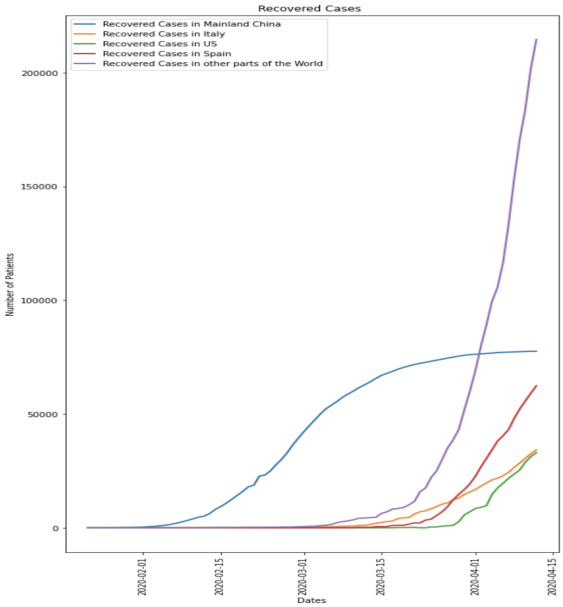
Graph 7: Top 15 Countries with Most Deaths

Comparison - China, Italy, US and Rest of the world

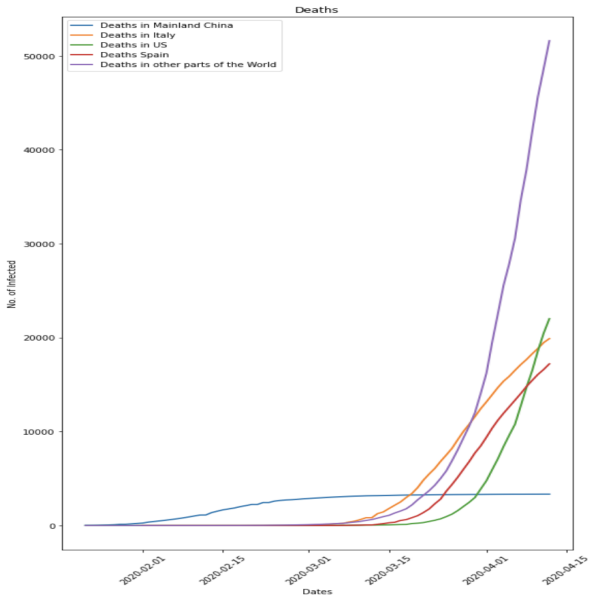
We can see that China has managed to flatten the curve, however, the rest of the world is experiencing a spike in the total number of cases and number of deaths.



Graph 8: Comparative Study of Confirmed Cases : China, Italy, US and Rest of the world



Graph 9: Comparative Study of Recovered Cases : China, Italy, US and Rest of the world

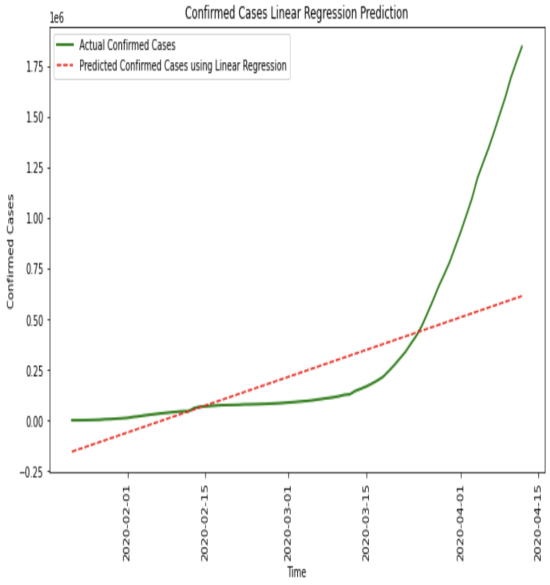


Graph 10: Comparative Study of Deaths : China, Italy, US and Rest of the world

Predictions using Linear Regression and Support Vector Machine

Linear Regression :

- The predictions of the Linear Regression model are completely off.
- Root Mean Square Error for Linear Regression: 959965.38 .

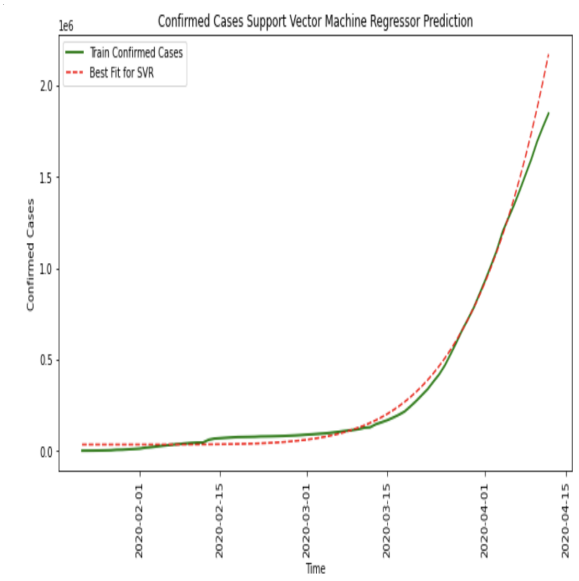


Graph 11: Linear Regression Results

SVM:

- SVM handles non-linear data more efficiently and this is evident in the plot.

- Root Mean Square Error for Support Vector Machine: 161859.97552267514 .



Graph 12: Support Vector Machine Results

Comparative Study of Linear Regression and SVM Results :

- Linear regression is good when the relation between covariates and the response variables is linear.
- SVM on the other hand handles non-linear data efficiently. This is because it transforms non-linear data into linear data and then draws a hyperplane.

	Dates	Linear Regression Prediction	SVM Prediction
0	2020-04-13	623747.629952	2334663.456037
1	2020-04-14	633243.319049	2508201.263523
2	2020-04-15	642739.008145	2692514.054271
3	2020-04-16	652234.697242	2888130.525870
4	2020-04-17	661730.386338	3095598.598038

Graph 13: Comparative Study of Linear Regression and SVM Results

Conclusion

COVID-19 does not have a very high mortality rate as seen above and is one of the most positive conclusion. Also, steady increase in Recovery Rate implies the disease is not chronic and can be cured. The most concerning thing is the exponential increase of the confirmed cases.
Countries like USA, Spain, United Kingdom,and Italy are facing major issues in containing the disease. The need of the hour is to perform COVID-19 pandemic controlling practices like staying at home/Quarantine, practice Social distancing, Testing and Contact Tracing with a speed greater than the speed of disease spread at all levels.

Future Work

This COVID-19 study is very extensible. This means that there could be a number of things we could include as part of the future scope for this analysis report when we can get hold of additional data. Some of these are outlined below :

- Time Series Forecasting for Virus Spread with different models to understand how this virus will spread over time given what we have data for.
- Time Series Forecasting for Deaths with different models to understand how many deaths would occur over time given what we have data for.
- Flattening curves drawn for Cases vs Date since when Social Distancing officially is in effect in a country/state which show that Social Distance is working.
- Stock prices comparative study to see how this global pandemic has attributed to the recessive economic market across the world.
- Comprehensive Analysis about how with the growth of the virus, various industries like the airline, travel and restaurant industry have taken the direct hit.
- Trends to see impact on Countries' GDP with time while everyone struggle to cope with this virus.
- Analysis of how this outbreak directly relates to extremely high unemployment requests files in some of the world's biggest economies.

References

[1] Must see CoronaVirus data Visualizations

[2] The best graphs and data for tracking the CoronaVirus Pandemic

[3] Scientific Facts for Understanding Curing the COVID-19 Pandemic

[4] What is a Support Vector Machine, and Why Would I Use it?

[5] Linear Regression — Detailed View

[6] RMSE: Root Mean Square Error