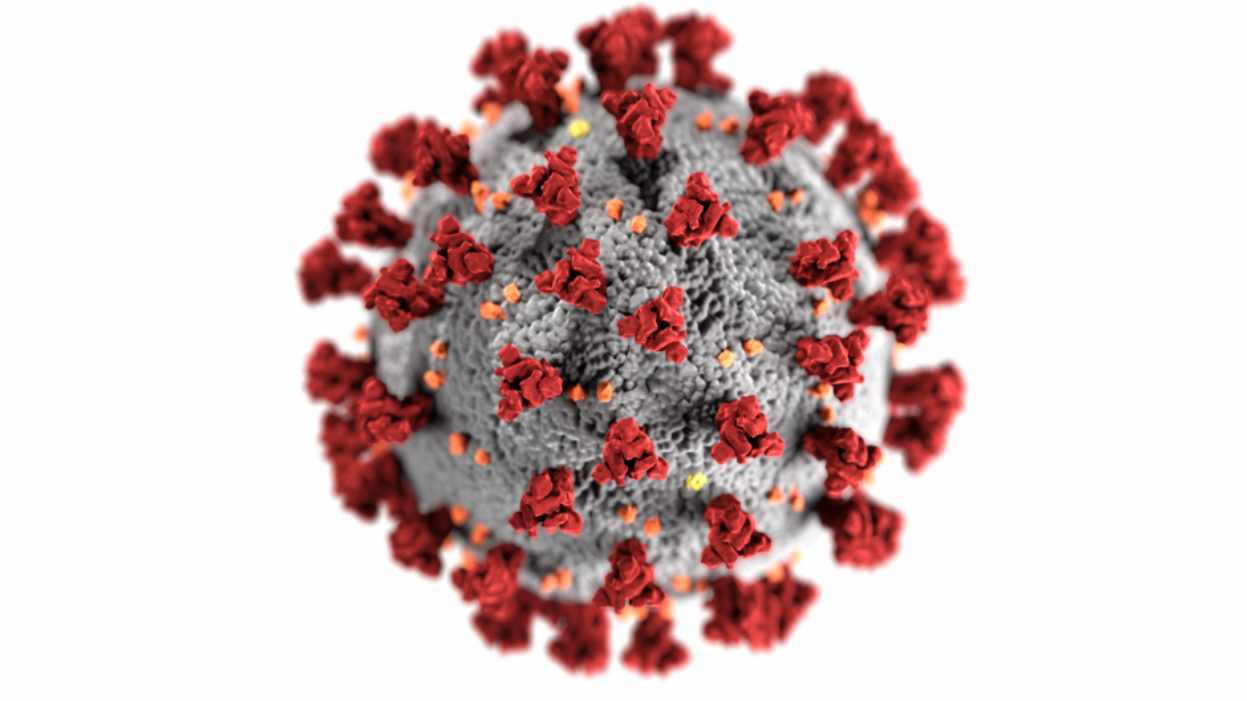
FIGHTING THE NOVEL CORONAVIRUS 2019: UTILIZING DATA SCIENCE

University of Illinois at Chicago



CONTENTS

1 CONTEXT, MOTIVATION AND RESEARCH QUESTION

COVID-19 is a virus that induces severe respiratory illness and is thought to have originated in Wuhan City, Hubei Province, China in December 2019. Since December 2019, the virus has spread rapidly on a global scale. The outbreak of COVID-19 caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Mainland China has been closely monitored by governments, researchers, and the public across the world. The rapid spread of positively diagnosed clustered cases affected hundreds of countries in a matter of weeks with hundreds of thousands — possibly even millions — of infections and deaths worldwide, and the impetus for a devastating wave of economic damage not seen since the 1918 Spanish influenza or the Great Depression.

Per the CDC, COVID-19 poses a serious threat; the virus is more contagious and lethal than the seasonal flu. The World Health Organization announced the COVID-19 outbreak a public health emergency of international concern on January 31st and on March 11th, 39 days later, it was classified as a pandemic. The first case of COVID-19 in the United States was reported on January 21st, 2020 in a man who traveled to China and began experiencing symptoms a few days after returning home to Seattle on January 15th, 2020. Per the CDC, while approximately 80% of cases involve mild symptoms, some individuals infected progress into severe pneumonia, multi-organ failure, and death. Research is still being conducted, but thus far, the risk of death for those infected with COVID-19 significantly increases for individuals above the age of 60 or for individuals with hypertension, heart disease, diabetes or lung disease.

The work of Kahneman and Tversky (1974) offer key insights on heuristics pertaining to making judgments when faced with uncertainty and highlight innate behavioral bias of human decision-making. Their work has aided the use of contemporary data science to construct comprehensive, effective processes for better decision-making; such models utilize substantial sets of data to produce useful information, drive performance, increase transparency, and especially, remove

irrational, instinctive behavioral biases in sports, investments, and even epidemics. Regarding the latter, some stress to just "follow the models" to avoid behavioral bias to which most expert epidemiologists fall victim and consequently, more people can die. Models may not be the only answer, but certainly, can offer reassurance when faced with a novel virus.

The goal of this project was to create a Python program to visualize the spread of the novel coronavirus 2019, perform network analysis tests at the country, state, and local level, and contribute insights to a terrifying, contemporary tragedy. This paper reports the results and describes the processes involved in completing the project.

2 NETWORK DATA, SUMMARY STATISTICS, AND VISUALIZATIONS

For this project, the inputs were gathered from data available in the GitHub data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL). For each observation, this data includes the date of observation in MM/DD/YYYY format, province or state of the observation when provided, country of observation, and cumulative number of confirmed cases, deaths, and recovered cases until that date. We collected data at different points in time.

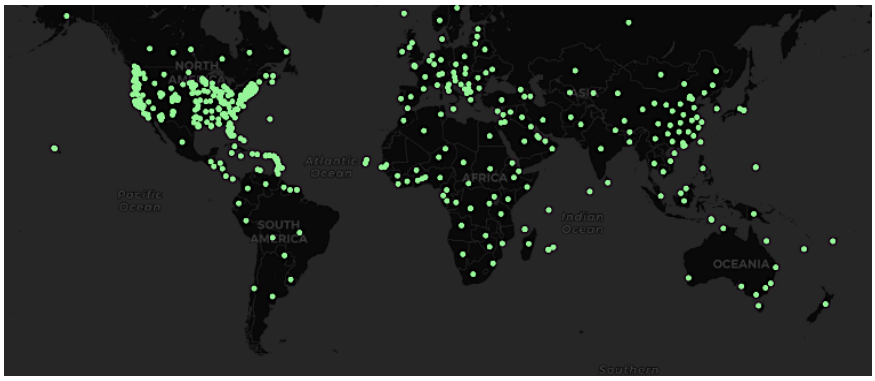
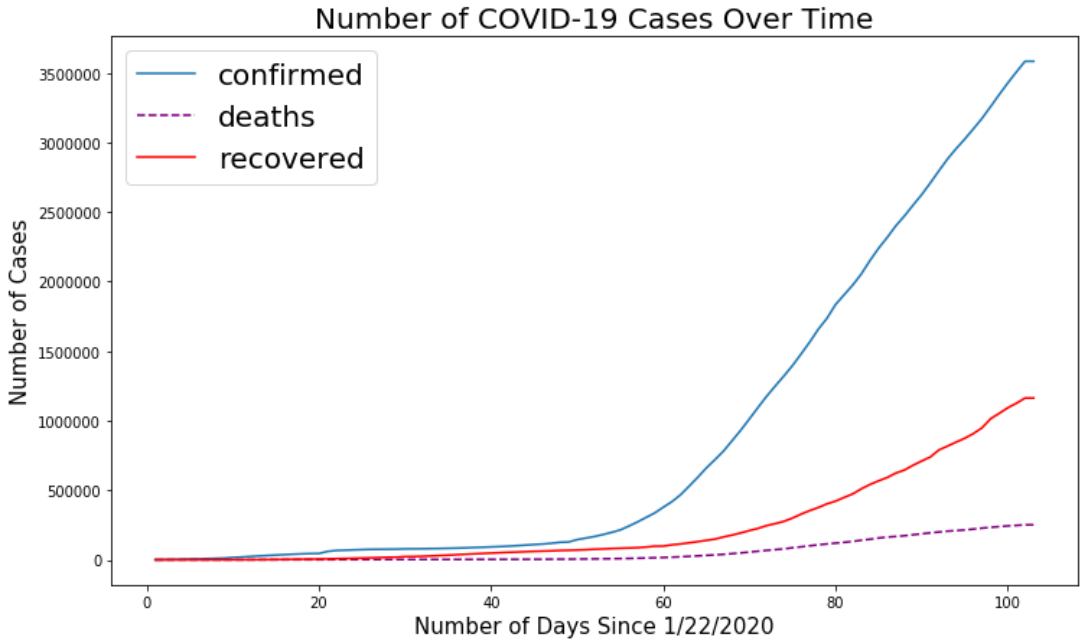


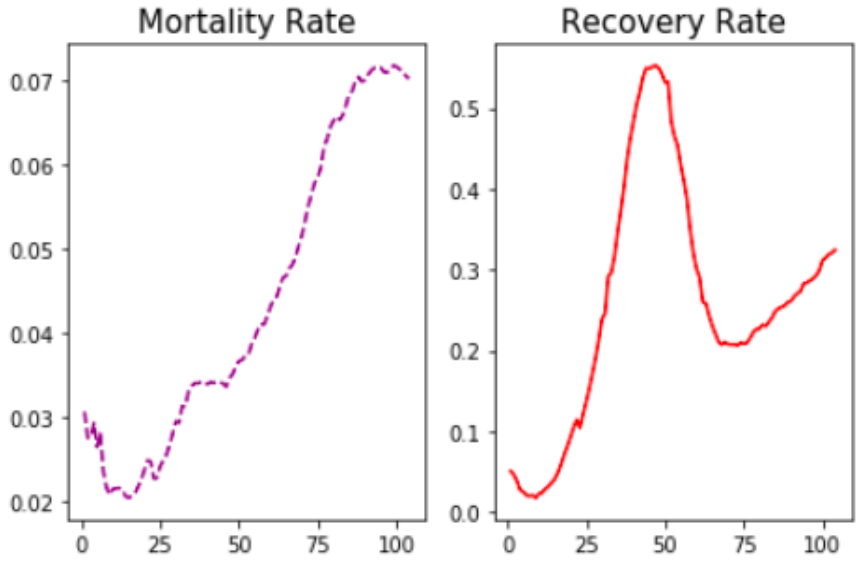
Figure 1: Countries with COVID-19 confirmed cases

The first thing studied was the spread of the COVID-19 virus. In order to do this, we started by looking at the data of confirmed, deaths, and recovered cases. This data was collected on March

3rd, 2020. At this point in time, the number of total confirmed cases worldwide was 336004, the number of total deaths was 14643, and the number of total recoveries was 98334. In the figure to the below we can see how the number of confirmed, deaths, and recoveries cases of COVID-19 increased day by day over the period starting on January 22nd, 2020 until May 4th,2020. From this picture, we can realize the number of cases since March 3rd until May 4th increases tremendously. More specifically, in this period of time, the number of confirmed cases increased by 3247051, number of deaths by 236894, and number of recoveries by 1064390.



```
Note: increases in COVID-19 cases lead to higher mortality rates
      and lower recovery rates
```

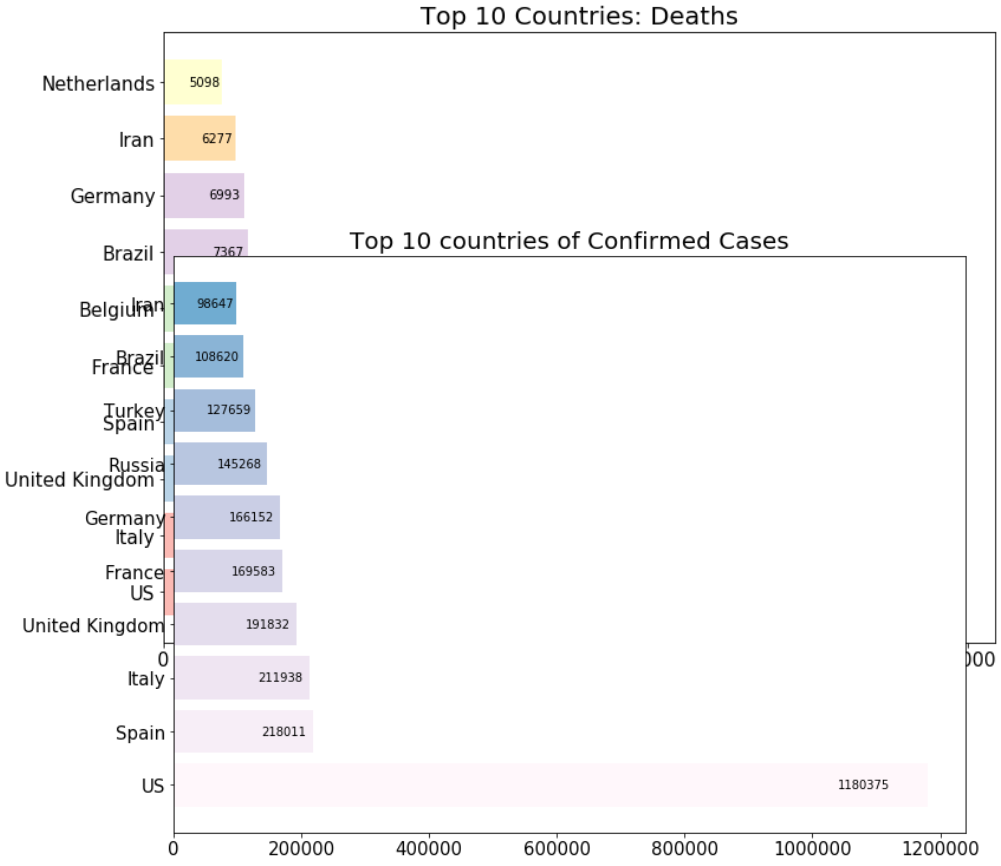


From this data, we also calculated the death and recovery rate. While the death rate as of March 3rd, 2020 of COVID-19 was 4.36% of all confirmed cases, the recovery rate was 29.27%. These rates were also calculated on May 4th, 2020 to see the differences. On this day, the mortality rate calculated was 7.02% of all confirmed cases and the recovery rate was 32.45%. Furthermore, we wanted to understand how the death and recovery rate has changed over the previously mentioned period of time. From the image below, we can see that as days passed, the mortality rate kept increasing while the recovery rate increased until around day 45, then declined. This is not surprising as the increase in confirmed COVID-19 cases lead to higher mortality rates and lower recovery rate.

Lastly, we thought it was important to understand how the COVID-19 was affecting countries individually. We plotted the 10 countries with the highest number of confirmed cases of COVID-19 and with the highest number of death cases. The country with the greatest number of cases confirmed was the United States with 1180375 cases. The second country with the most cases was Spain with 218011. The amount of cases in Spain helped us understand how fast the spread of the virus was in a

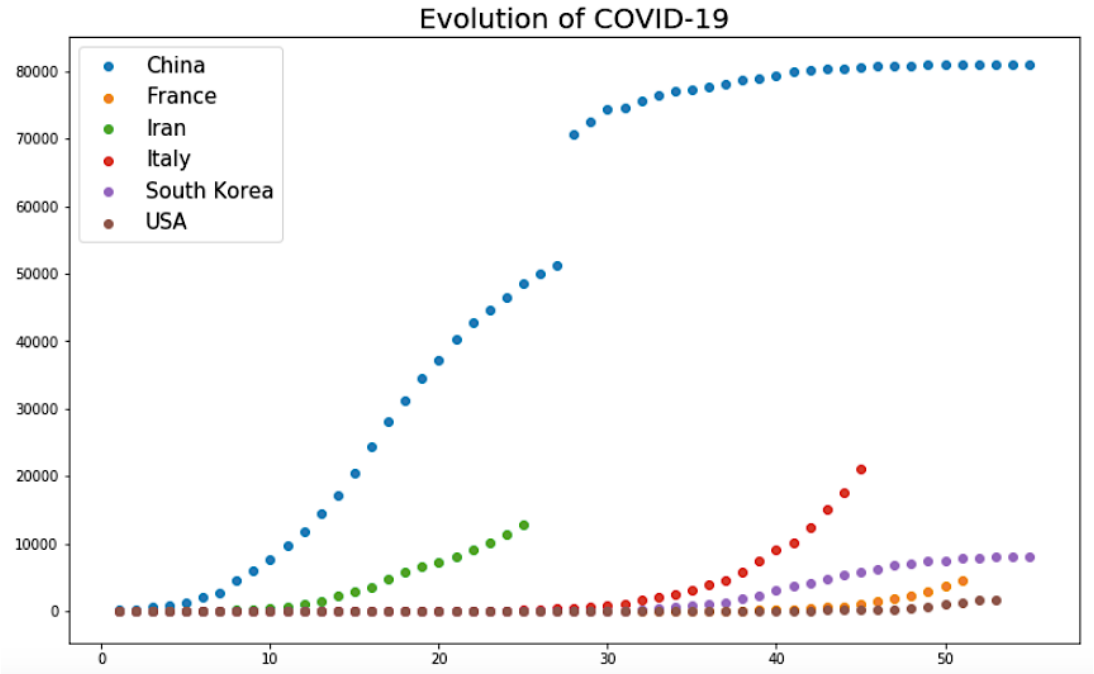
short period of time. It is interesting to see how the virus traveled that fast to another continent and how it reached that amount of cases. The countries that followed with the greatest number of cases were the U.K, France, Germany, Russia, Turkey, Brazil, and Iran, respectively. Ultimately, the country with the most deaths was the U.S with 68922, surpassing Italy with 29079 deaths.

As mentioned before, we wanted to know how fast the COVID-19 network spreads and its infection rate. More specifically, we wanted to know if the spreading of the COVID-19 pandemic follows an exponential growth model, meaning that the number of new infected cases follows an exponential function. An exponential growth occurs when the growth rate of a function is always proportional to the function's current size. Here, it means the number of cases constantly increases each day and it gets to big numbers fast. An initial



exponential growth of confirmed cases is generally expected for an uncontrolled outbreak, as observed during the 2009 Influenza A (H1N1) pandemic(de Picoli Junior et al, 2011).

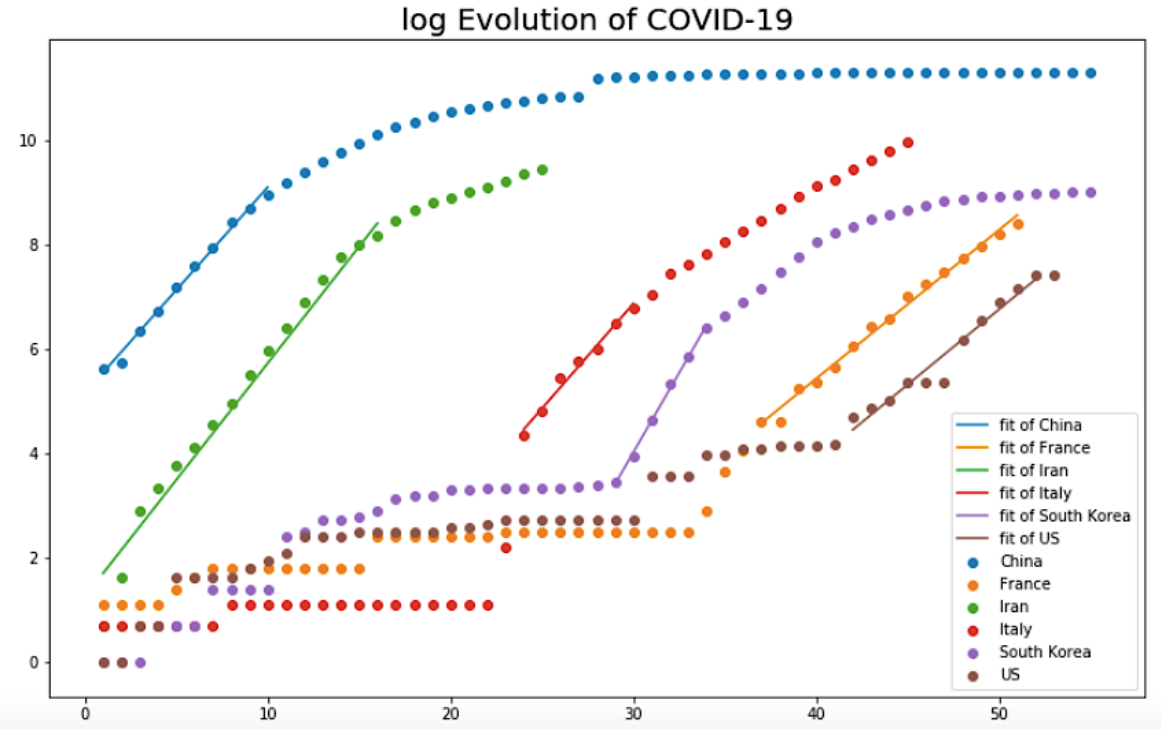
In order to do this, we looked at the data available from the following 6 countries: China, France, Iran, Italy, South Korea, and the U.S. From this data, we looked at the first days since the first discovery of contagion in each country besides China, which starts with 278 persons already infected. We mention that not all countries had 56 observations at the time the data was collected, but this data still works as we are trying to look at the percentage of increase of cases.



From this data, in order to see if the number of infected people is growing at an exponential rate, we calculated $N(t) = N_0e^{rt}$. Where N represents the number of infected people and depends on N_0, which is the starting number of infected. Lastly, e represents a natural number which is then raised to the r*t, where t is time. Thus, that is why it's called an exponential growth as the independent variable(t) is an exponent.

Then, to understand better these graphs, we turned these exponential functions into linear functions. To do this, we did the following for each six selected countries. First, we divided both sides

by N_0, which equals $N(t)/ N_0 = e^{rt}$. Then, to remove e, we took the natural logarithm of both sides of the previous formula to get $\ln(N(t) / N_0) = rt$, which leaves a linear function on the right side.



From this picture, we can see that only some parts of the data can be fitted linearly, which tends to be in the beginning. The parts of the plot that are linear, represent that that part is actually growing at an exponential rate. These results are not surprising since from previous research in epidemics we know that exponential growth of confirmed cases during pandemics is generally expected at the initial stage. Therefore, we can say that the COVID-19 spreads at an exponential rate at its beginning phase and doesn't continue to grow like this as otherwise, everyone would be infected fast. Lastly, from this figure we can observe the slopes of the lines seem relatively different between countries. This means the rates of COVID-19 infection rate from country to country differ.

We explored the different rates. From the plot corresponding to Iran and Italy, we see a linear fit is still possible to the latter part, which means the COVID-19 is still spreading exponentially in

these 2 countries. The beginning infection rates from the first 10 days since the first observation available of China is 0.39. Infection rate for France in the days 36-50 is 0.28. For Iran, the rate in the first 16 days was 0.44. Italy in the day 24 through day 30 since the first contagion had a rate of 0.40. The highest rate of infection observed in our data was in South Korea in the days 28-33, with 0.60, which lasted 6 days. Lastly, the rate of infection in the USA in the days 41-51 since first contagion was 0.28.

3 COMPARTMENTAL MODELS

Compartmental models rely on an arrangement of differential equations that can serve as mere representations of the different parts of epidemiological states of a population. They can serve as useful simulations for examining infectious diseases, such as the mechanisms by which they are transmitted, as tools to forecast the long and short path of an outbreak, and developing mitigation strategies for the epidemic (Abhay Shukla, 2020). In these models, a population is separated into distinct groups that possess the same features. A well-known compartmental model is the SIR model which possesses three compartments: **S**usceptible, **I**nfectious and **R**ecovered. This model can be modified to include **E**xposed and **D**eceased compartments such as in the SEIR and SEIRD forms. These models take on the following assumptions:

1. No individual can join the susceptible group – members of the population can only be infected once and it is the only way they can leave the susceptible group
2. All members of a population have equal chances of being infected and their age distribution is uniformly distributed between 0 and the life expectancy L
3. Homogeneous mixing of the population: the links a member has to the population are uniformly distributed

Moreover, these models incorporate the following parameters:

1. **β or infection rate**: parameter that exemplifies the contagiousness of a virus, and consequently, a critical factor in the spread of a disease from one individual to surrounding members of the population. This number of infected parties is linked to the number of connections per time unit (assumed static) and to the probability of disease transmission;
2. **γ or recovery/death rate**: the average ratio of the infected individuals which recover in the time unit
3. **σor incubation rate:** rate of latent individuals becoming infectious where the average duration of incubation is $1/\sigma$

Novel viruses and human behavior are dynamic and thus, difficult to anticipate. Consequently, SIR models and their related versions do not represent reality, but in fact, are just models that can provide useful information, at a high-level, that has the potential to save lives.

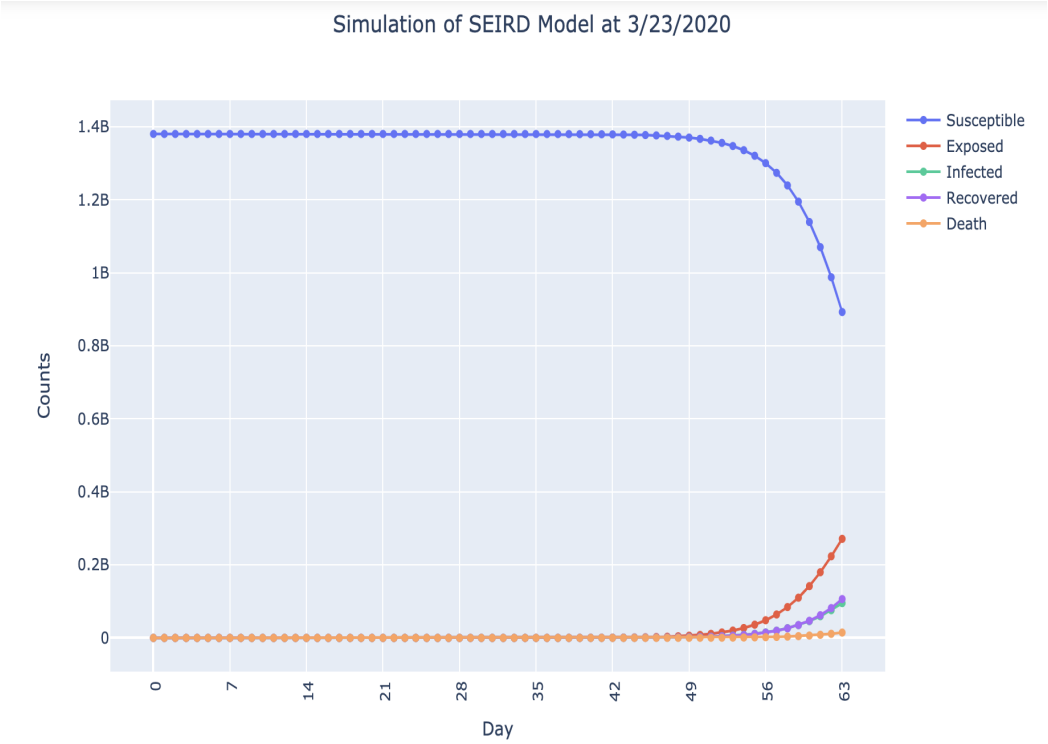
We were intrigued by the availability of data analytics tools that were available to create models of our own that could enable us to analyze COVID-19.

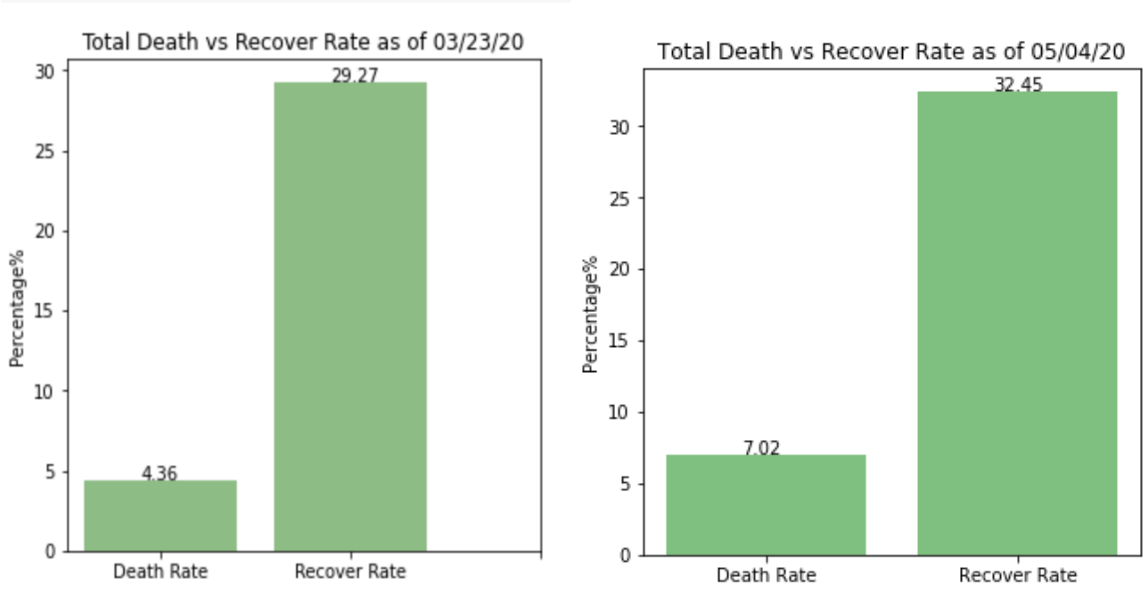
SIMULATION SEIRD MODEL

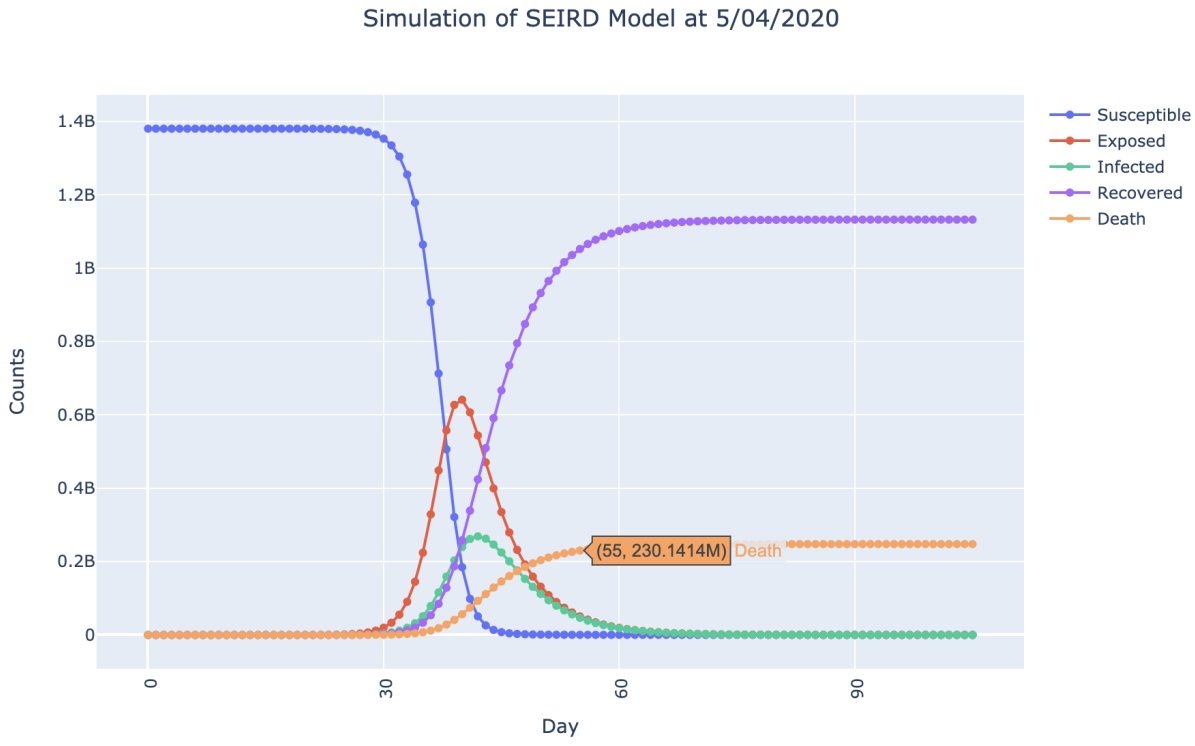
Equipped with data analysis tools and techniques acquired through our education, we assembled various SEIRD simulations to take a deeper look at the tragic pandemic that has reminded our deeply interconnected, massive world how small it can become.

Execution of the following simulation needed us to determine a system of Ordinary Differential Equation (ODE), solve it using a solver (Scipy.integrate package function **ODEINT**) and forecast for the time span we set and lastly, consolidate it with Jupyter Widgets to enable interactive modification of various parameters of the model and inspect its influence on the curves of SEIR and D. We present contrasting examples that depict the SEIRD models at March 23, 2020 and May 4th, 2020, respective of the disease transmission rates at the time. Furthermore, we have

provided corresponding bar plots demonstrating the relevant death and recovery rates at each respective date in COVID-19's progression.

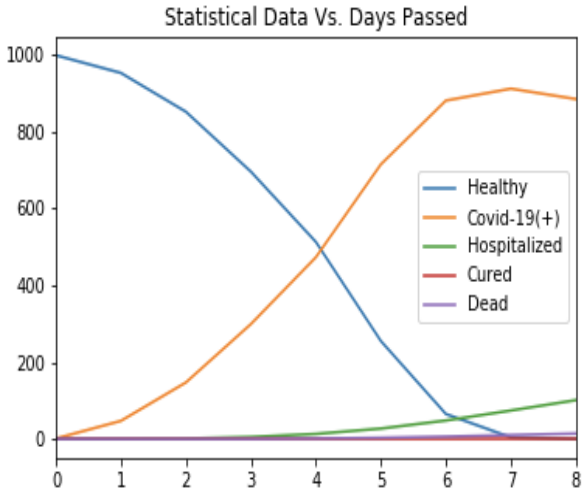






The juxtaposition is astounding. On March 23rd, 2020 the virus is just marking the beginning of its rapid, destructive, and lethal path that can be shown on May 4th, 2020; notably, the above figure shows "flattening of the curve."

Next, we wanted to dig even deeper to assist our understanding in epidemiology as it related to COVID-19 by incorporating an iterative loop, armed with the needed SEIRD parameters, including a distance and number of samples input to simulate the impact of social distancing on the mitigation of transmission. The following was more of a roughly random process, but nevertheless, offered another view on the mechanisms of the COVID-19.




```
------------------------------
Day: 12
- - - - - - - - - - - - - - - -
Healthy            7.0
Covid-19(+)      837.0
Hospitalized     129.0
Cured              8.0
Dead              19.0
Name: 12, dtype: float64
------------------------------
Day: 13
- - - - - - - - - - - - - - - -
Healthy            2.0
Covid-19(+)      798.0
Hospitalized     153.0
Cured             24.0
Dead              23.0
Name: 13, dtype: float64
------------------------------
```

4 CONCLUSION

After performing this network analysis and taking into consideration the different adjustments countries have taken to stop the spread of the virus, we find a significant relationship between the number of cases, deaths, and the type of mobility restrictions imposed by each country. For instance, we can see from the analysis, the U.S. and the U.K. have a lot of cases. Having the U.S. (top 1) and U.K.(top 4) makes sense as these countries opted for a medium level mobility restriction. On the other side, we can observe a high degree of relationship between the number of cases and the number of tourists a country receives each year. Spain and Italy being the most visited countries within Europe and also being the countries most affected within, validate our statement. Something important from this analysis and from our simulations is that after checking when a rate is increasing exponentially we

can conclude that many if not most of the countries have already passed this stage which hopefully indicates this virus won't be increasing at such rate.

BIBLIOGRAPHY

Abhay Shukla. (2020, April 11). Simulating Compartmental Models in Epidemiology using Python & Jupyter Widgets. Retrieved May 6, 2020, from Medium website: https://towardsdatascience.com/simulating-compartmental-models-in-epidemiology-using-python-jupyter-widgets-8d76bdaff5c2

Capitanelli, A. (2020, March 8). Modeling the spread of diseases - Towards Data Science. Retrieved May 6, 2020, from Medium website: https://towardsdatascience.com/modeling-the-spread-of-diseases-821fc728990f

CSSEGISandData. (2020, April 29). CSSEGISandData/COVID-19. Retrieved April 29, 2020, from GitHub website: https://github.com/CSSEGISandData/COVID-19

de Picoli Junior, Sergio et al. "Spreading patterns of the influenza A (H1N1) pandemic." *PloS one* vol. 6,3 e17823. 31 Mar. 2011, doi:10.1371/journal.pone.0017823

Exponential Model. (n.d.). Retrieved April 27, 2020, from https://web.ma.utexas.edu/users/davis/375/popecol/lec5/exp.html

Exponential growth. (n.d.). Retrieved April 27, 2020, from https://www.tau.ac.il/~tsirel/dump/Static/knowino.org/wiki/Exponential_growth.html