COVID-19 CASE STUDY

IST 718 – Big Data Analytics

Sunday 7:30 pm

Michael Armesto - Patrick Aslakson - Jordan Spector

September 24, 2021

Introduction

Big data analytics is the use of advanced analytic techniques against very large, diverse big data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes. Big Data Analysis is a process of extracting and discovering patterns in large data sets involving methods that involve an amalgamation of machine learning, statistics, and database systems. Data Mining is a process performed by a data scientist where raw data is manipulated into more useful information. Data mining is the process of discovering interesting and useful patterns and relationships in large volumes of data.

Such useful patterns could be used in the evaluation of a dataset for the COVID-19 Pandemic that is currently rampant all over the globe. In this case, several datasets are utilized in the evaluation of death rates and other indicators caused by the virus in order to make recommendations for an efficacy for vaccinations and other mandates such as mask wearing.

There are many techniques or models used for data mining at the data scientists disposal.

Specifications

The COVID-19 pandemic continues to be a significant problem for the world population as new variants of the virus emerge from different countries. At the time of this research paper, more than 201 million cases have been recorded worldwide, 4.27 million deaths reported.

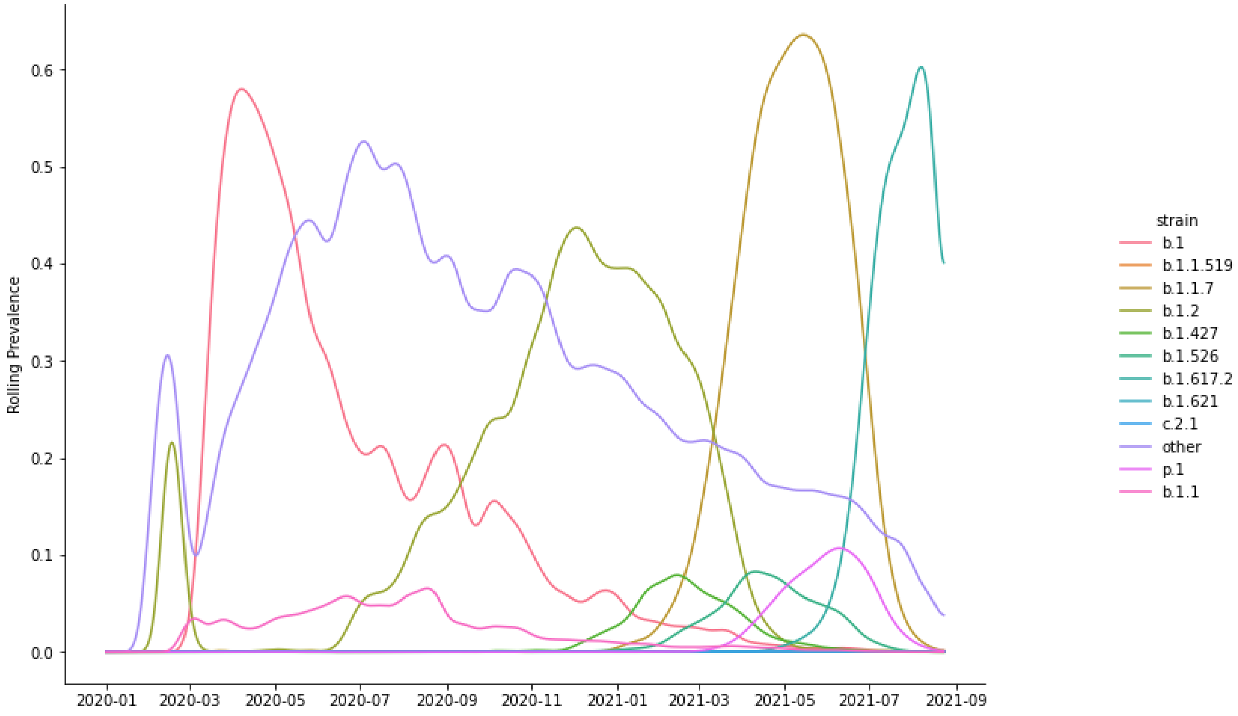
Mitigation measures continue to vary per country with no centralized guidance. There is a lack of clarity around government-imposed restrictions and how strictly those restrictions should be enforced. Variations in mitigation strategies lead to variations in successful prevention methods. A larger number of cases increases the probability of new variants emerging in the population. These variants threaten vaccine efficacy and natural immunity. This will inevitably require the necessity for renewing restrictions on travel, work, vaccination mandates, and greater strain on the economies of the nations hit hardest by COVID. Therefore, our null hypothesis states countries who have instituted less stringent mitigation tactics and lower vaccination rates will have increased death rates on a per-capita basis.

The datasets for this analysis are the Our World in Data (OWID) dataset. The dataset consists of 62 columns and more than 112,000 rows of data made up of string, time-series, and numeric data. The pertinent data in the dataset includes COVID total cases, new cases, total deaths, new deaths, and etc. The next dataset includes a metric describing a country's stringency on COVID-19 intervention on every day between January 2020 and today, totaling 109,055 rows. This metric is described further below.

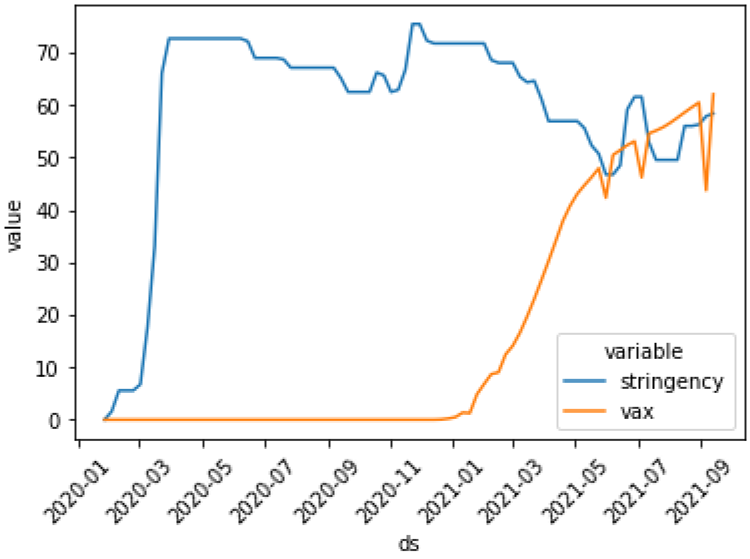
Observations

Several datasets were considered in the course of the evaluation of the project including mask data, vaccine manufacturer data, and variants data all proved to be inconsistent and difficult to combine with the OWID dataset to provide great confidence in the models. The time-series data was the most difficult to work with in that it was the most difficult to manipulate and build into a meaningful working dataframe. The most prevalent issue with the COVID data is that there is a large amount of data. Two of our datasets are so large that they must be broken down into smaller incremental pieces, by country is the current method.

The dataset on COVID-19 variants was found to be useful only in exploratory analysis. While it provided prevalence data on a daily basis, the number of countries included were limited, and inconsistencies were found within the data, including many missing datapoints and rapidly changing data. More concerning, is that while the data was uploaded by a Kaggle user, an academic source of origin could not be identified, and the methodology of its creation remains unknown. Though an attempt was made to incorporate variant data into further analysis, mostly focused on variants of concern such as the highly contagious Delta variant, the data was not found to be correlative with spread, or severe cases. It's very possible that because the outbreak of concerning variants such as the Delta and Lambda coincided with the spread of COVID-19 vaccines, the effects of the variants have become muddled and therefore appear insignificant. The following plot depicts the prevalence (as a 5-day rolling average) of several COVID-19 variants over time in the United States. Here, it can be seen that the b.1.1.7 strain and the b.1.617.2 strain (commonly known as the UK and Delta strains, respectively) became more prevalent in the United States during 2021, confirming researchers' fears that these strains could be more contagious than previous forms of COVID-19.



Stringency Policies include the following areas of concern on a scale typically from 0 to 5. The specific policy and response categories are coded as follows: School closures, Workplace closures, Cancel public events, Restrictions on gatherings, Close public transport, Public information campaigns, Stay at home, Restrictions on internal movement, International travel controls, Testing policy, Contract tracing, Face coverings, and Vaccination policy. The stringency index utilized in this analysis is an aggregate composite score using these metrics. This index was created by researchers from the University of Oxford, and is described in further detail here. The following graph demonstrates the overall Stringency Index of the United States over time, as well as the vaccination rate, expressed here as the running total of vaccinated people out of every 100 residents.

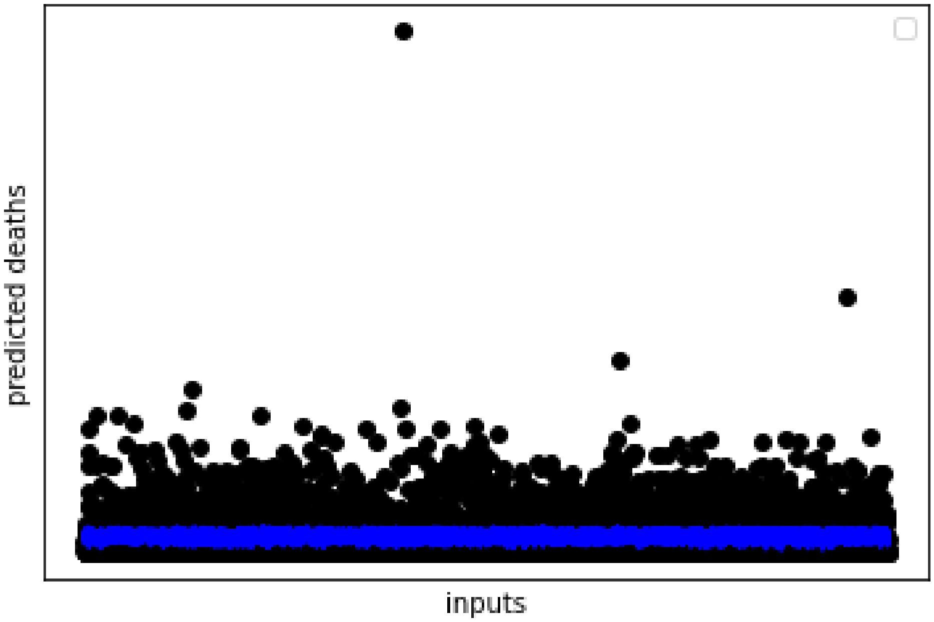


Analysis & Models

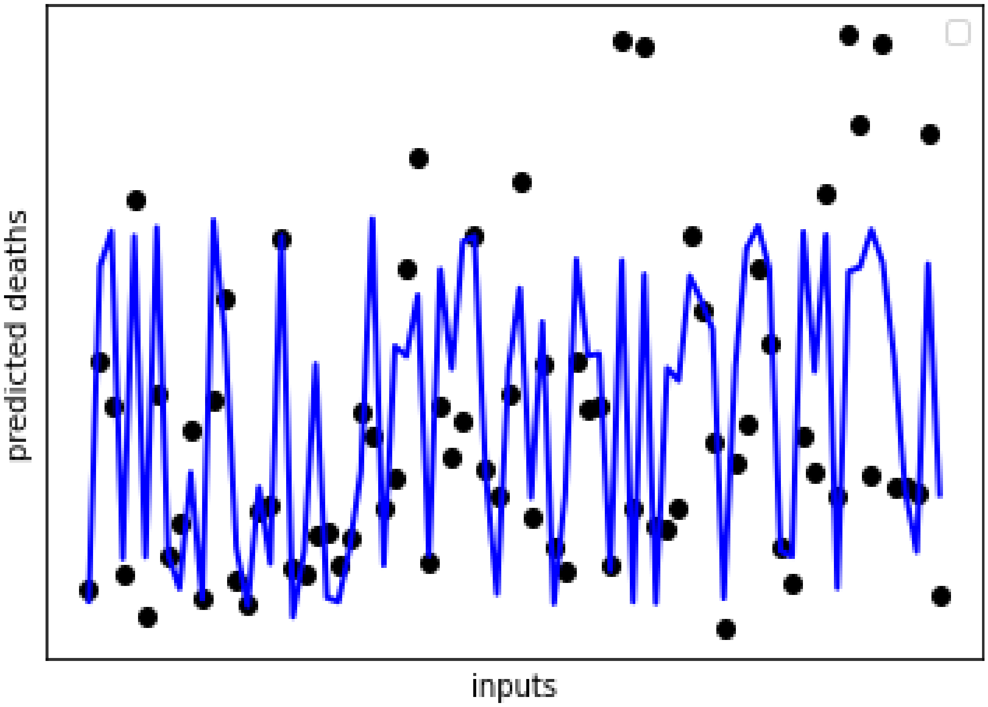
Models

Linear Regression

The first model utilized in the analysis is the Linear Regression model from SciKit Learn. We can see from the Linear Regression of the World data that the inputs stringency_index, facial_coverings, and people_fully_vaccinated_per_hundred are represented by the black scatter point and the predicted deaths are the blue line.   We repeated this with up to eight different variables with little or no success in increasing this accuracy of the model. The statistics for the World model are: Coefficients: [0.32951289 0.04169519 -0.02196575] , Mean squared error: 26.86, Coefficient of determination: 0.03.

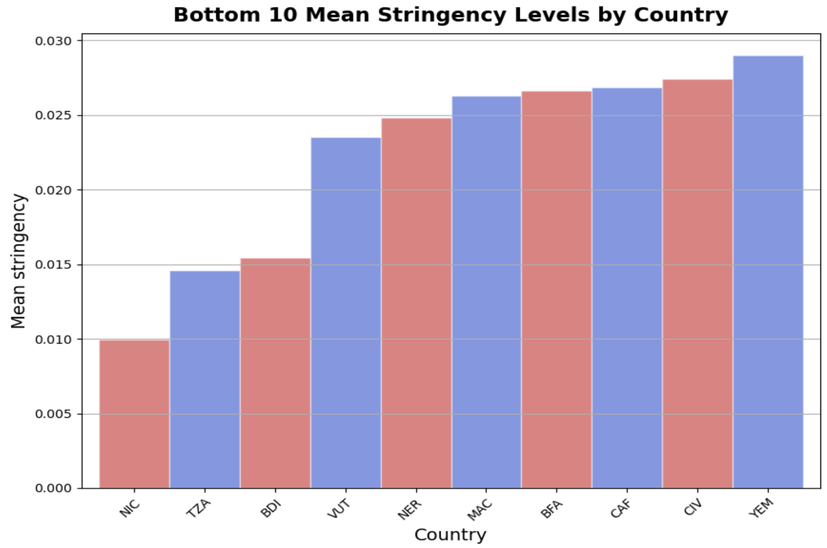


Next, we evaluated the Linear Regression and isolated the data to the United States. This increased the accuracy of the model by over 50% and reduced the mean squared error significantly. Yet in evaluating the coefficients for the model, they proved to be counterintuitive in that the coefficient for the facial_coverings variable is positively correlated and lends to the notion that facial covering lead to increased death rates. Inherently, we know this to be a false positive. The statistics for the US model are: Coefficients: [2.34443187 -0.10282321 -0.13333016], Mean squared error: 4.57, Coefficient of determination: 0.51.

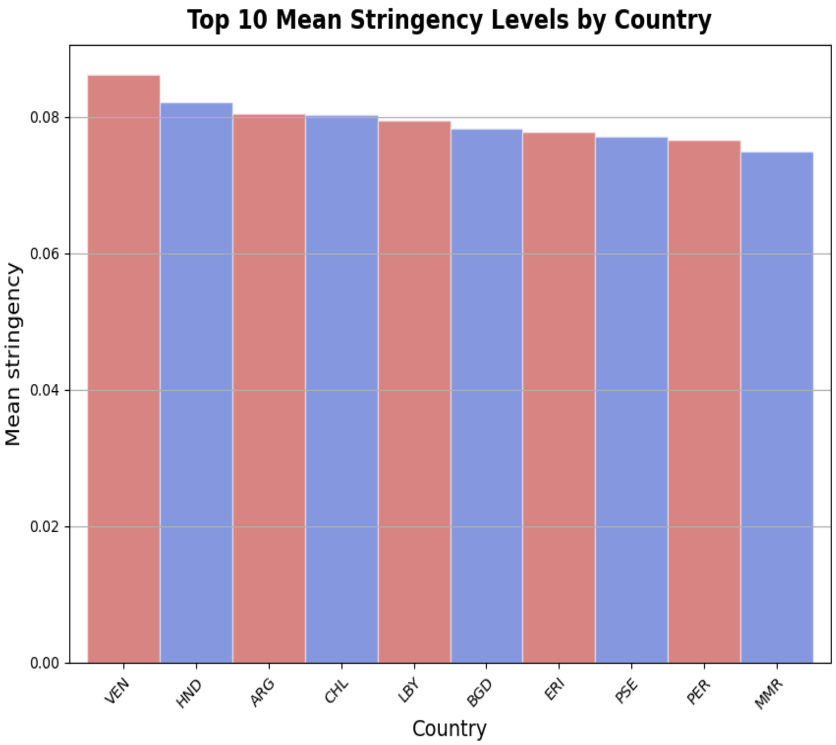


In reviewing the stringency policies of the various countries, they reveal that there is a

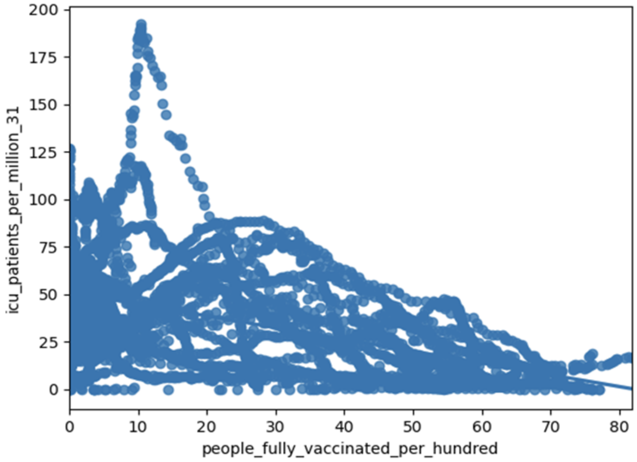
-.33 correlation between stringency and new cases per million lagged and -.28 between stringency and new deaths per million for top 5 most stringent countries. Shown here:



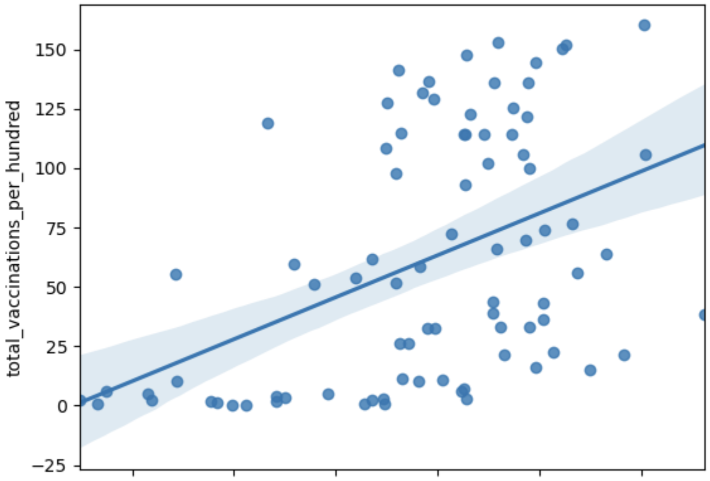
Further, we look at the top 10 stringency levels and the correlation of -.06 between stringency and new cases per million lagged and .02 for new deaths per million for bottom 5 most stringent countries. Shown here:



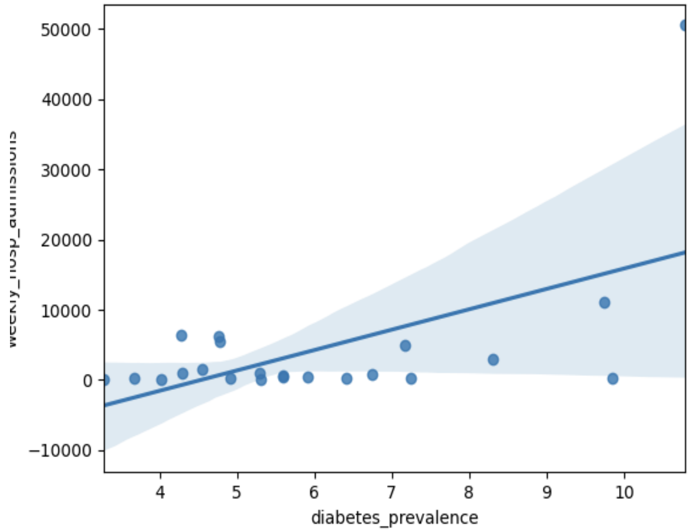
To continue a more in depth look into the data, we now turned to looking at the number of ICU patients versu fully vaccinated people. The R-value is R = -.47, which translates to as more people were fully vaccinated ICU patients per million declined, considering a two month lag.



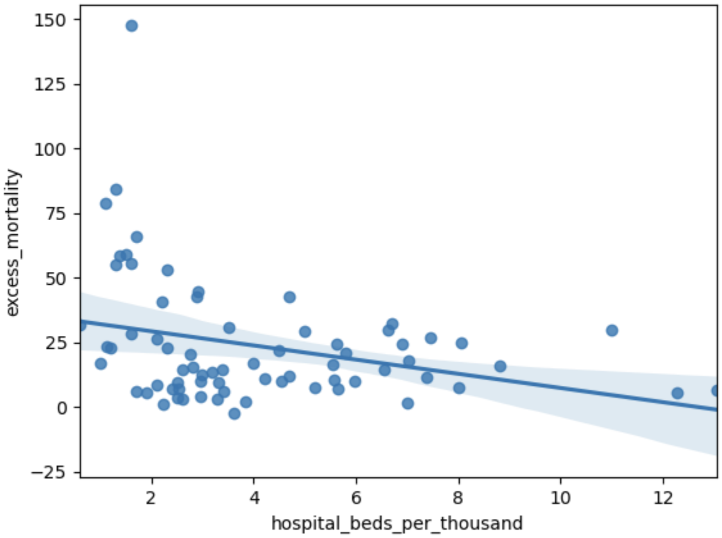
People fully vaccinated per hundred are highly correlated with less ICU beds per million on a two month lag. This supports the idea that vaccinations are effective at preventing severe illness from COVID.



As the stringency index goes up, so too do total vaccinations per hundred. As countries make vaccines available to local populations and more accessible, it shows people are willing to get the shot more often than not.



Among the countries with diabetes data, there was a positive correlation (R = .56) between higher diabetes prevalence on average and weekly hospital admissions on average. This lends to the theory that people with pre-existing conditions are more susceptible to severe illness.



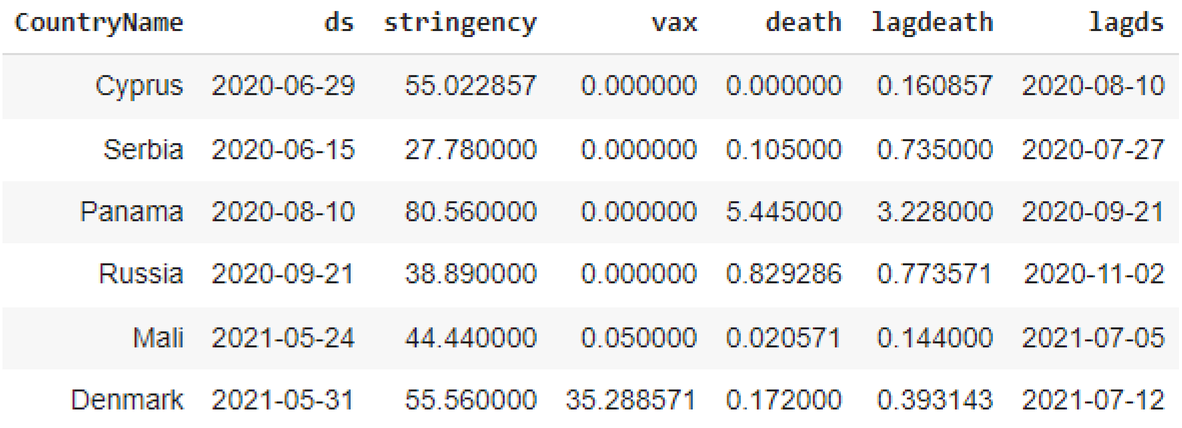
With an R value of -.30, more hospital beds per thousand led to less excess mortality throughout different countries. This indicates that countries who had less beds may have been overloaded by COVID surges, resulting in an inability to treat other illnesses with priority.

Facebook Prophet

In order to create a forecast from time series data, Facebook's Prophet library was used. Prophet was selected because of its ease of use, and built-in features such as yearly, weekly, and daily seasonality. Deaths per million was selected to be the target variable, as exploratory analysis found it to be more stable and more predictable than the rather volatile case rates. Due to national factors such as differences in testing accessibility, motivation/ability to be tested, and more, simply relying on the number of daily positive tests seemed likely to be tenuous at best. While rather morbid, working with death data is, in theory, more concrete, as a death is more likely to be documented and reported. Similarly, death rate is more likely to be affected by vaccination rates, as research has found COVID-19 vaccines to be more effective in preventing fatal cases than cases themselves.

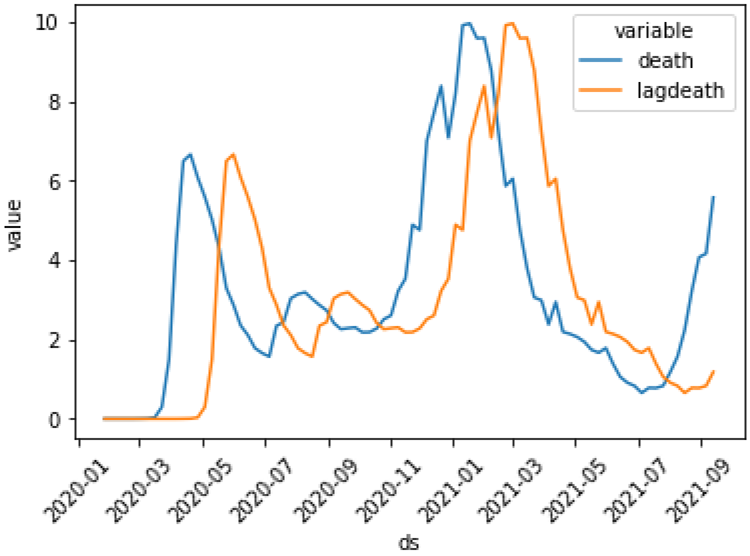
To forecast deaths per million by country, two datasets were used. The first was the previously described Our World In Data dataset. This provided country, date, deaths per million (daily), and people vaccinated per hundred residents (running total). Deaths were analyzed on a daily basis to see how external factors affected a country on a particular day, while a running total of vaccinations was used because the total number of people vaccinated should affect the spread of disease, rather than the number vaccinated on a particular day. Different countries had different start dates in provided data, but generally the date range analyzed was mid-January 2020 to September 18th, 2021. The second dataset provided country, date, and the previously described Stringency Index. This was merged in by country and date. To account for volatility and daily seasonality in the data, weekly averages were taken for all metrics. This way, if residents in an area tend to receive more vaccinations on Fridays, for instance, those temporary peaks would not affect the analysis.

Next, a lagged dataset was derived for the deaths per million residents. In theory, a country's policy changes, such as mask mandates and restaurant closures, may only provide a noticeable impact in case rates and death rates weeks after implementation. Similarly, vaccinated individuals may take weeks to reach peak immunity levels. To test this, a second death rate metric was created by taking the original death rate, and "lagging" it by adding 6 weeks to the date, then matching it back in. This way, a day, say June 15th, could be measured by its current stringency, current vaccination numbers, and the deaths that would occur 6 weeks later.

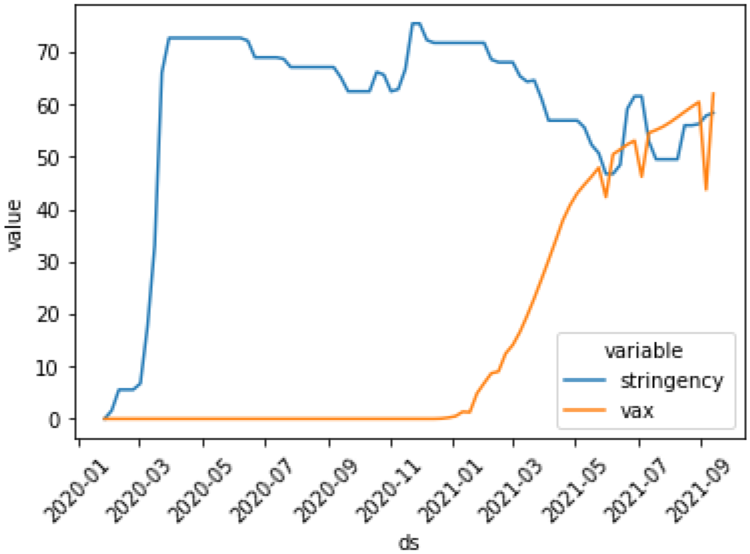


CountryName	ds	stringency	vax	death	lagdeath	lagds
Cyprus	2020-06-29	55.022857	0.000000	0.000000	0.160857	2020-08-10
Serbia	2020-06-15	27.780000	0.000000	0.105000	0.735000	2020-07-27
Panama	2020-08-10	80.560000	0.000000	5.445000	3.228000	2020-09-21
Russia	2020-09-21	38.890000	0.000000	0.829286	0.773571	2020-11-02
Mali	2021-05-24	44.440000	0.050000	0.020571	0.144000	2021-07-05
Denmark	2021-05-31	55.560000	35.288571	0.172000	0.393143	2021-07-12

The graph below illustrates the death rate and lagged death rate in the United States.



The following graph illustrates the changes in the stringency index and the number of vaccinated people per hundred residents in the United States. Note that the increased vaccination rate beginning in 2021 pairs with a drop in the death rate seen in the graph above. The anomalous drops in vaccination rate seen below may be a result of the weekly aggregation, or flaws in the original dataset.



Next, the forecasting models were created. All models were trained using data from January 2020 through June 2021, and tested using data from July 2021 through September 18, 2021. Models were created individually for each country, and eight varieties of models were created based on combinations of lagged or unlagged deaths, and whether or not stringency and vaccinations were added as regressors. Models were evaluated using mean absolute error (MAE), the average difference between the predicted and actual values.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

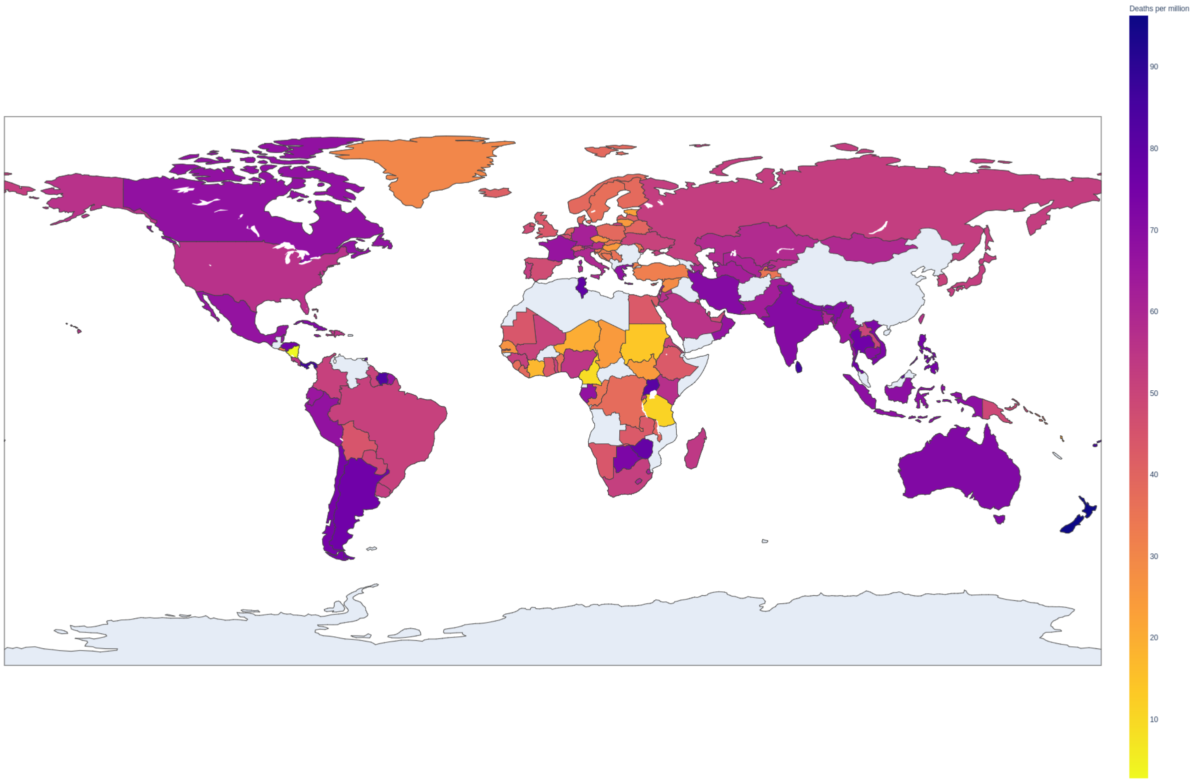
Results for each variation of the model are depicted in the table below.

Model	**Death type**	**Regressors**	**MAE**
1	Unlagged	None	2.181
2	Unlagged	Stringency, vaccinations	2.007
3	Unlagged	Vaccinations	2.079
4	Unlagged	Stringency	1.777
5	Lagged	None	2.209
6	Lagged	Stringency, vaccinations	2.091
7	Lagged	Vaccinations	2.155
8	Lagged	Stringency	1.982

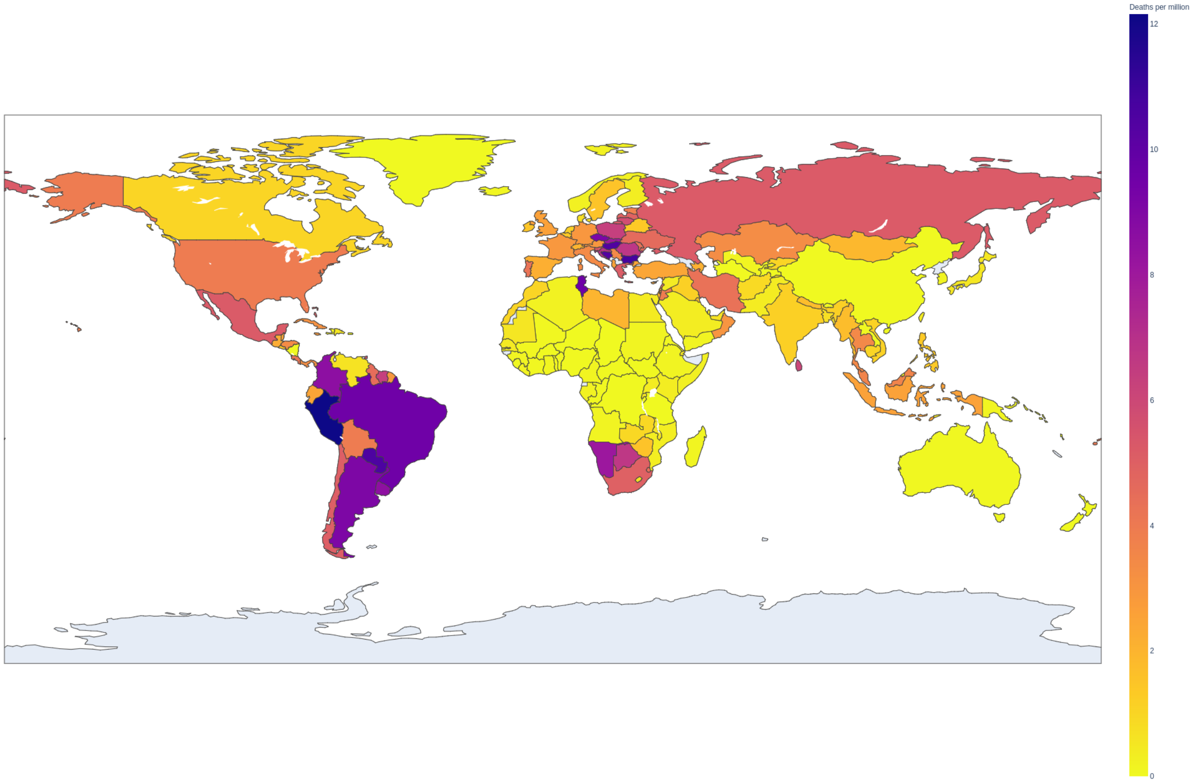
Looking through the results table, a few patterns become clear. Unfortunately, the lagged death data did not prove more predictable, with all lagged models performing slightly worse than their unlagged peers. The regressors added to the predictive power of the models. Stringency proved to be a more helpful regressor than vaccinations. In fact, while vaccinations improved model performance over the base models, it detracted from the success of the models that included stringency. The overall best combination was unlagged deaths predicted using only stringency as a regressor.

After testing the various models, a new model was created using the entire timeframe to train, and forecast into the future. It was found that stringency could be predicted with a high degree of accuracy. Knowing that this number was not as subject to change as death rate, stringency was forecasted into the future 20 weeks beyond the final date of available data for each country. This stringency data was then used as a regressor to predict the future death rate.

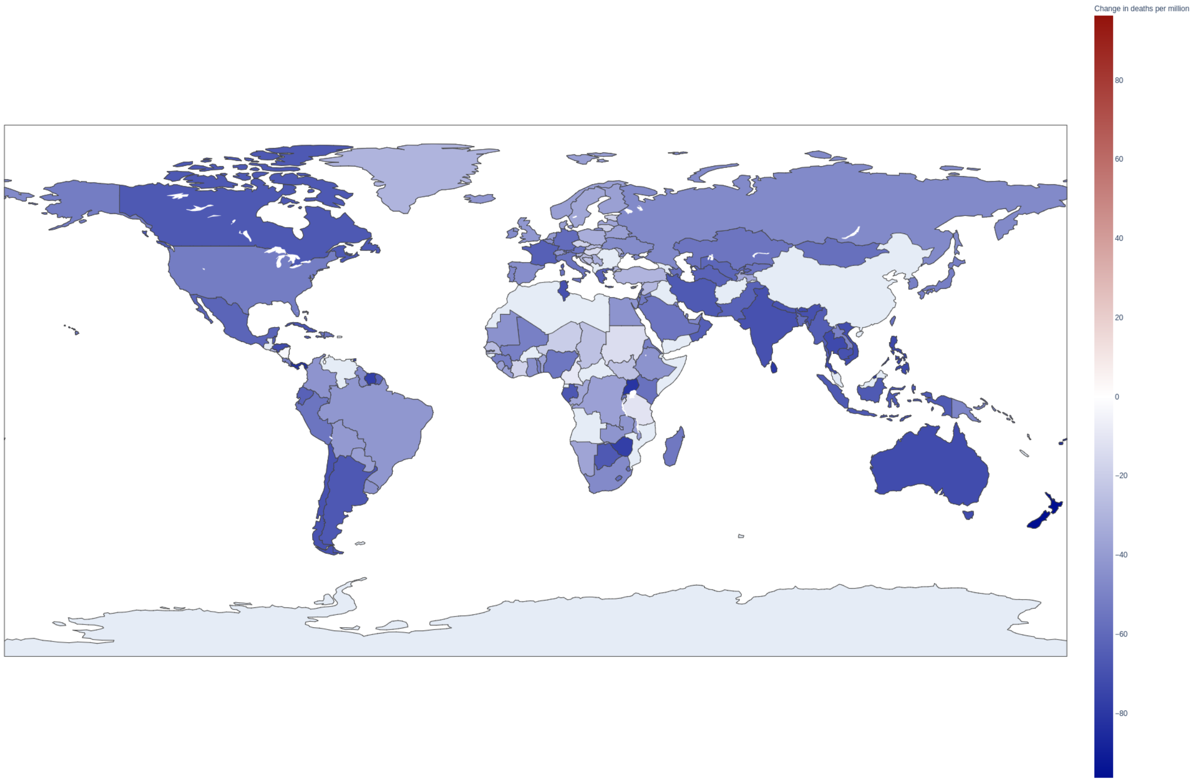
The image below depicts the rate of death per million residents in each of the 185 countries that had data available for the week of August 30, 2021.



The next image depicts the predicted rate of death per million residents on the week of November 14th, 2021. For most countries, this was 4 weeks after the last available data in the dataset. Note that some countries that did not have data available on August 30 are included here, as their previous data was forecasted up to this point. Also note that the scale here is much smaller, with the maximum rate being 12 per million, rather than 90 per million as in the previous image.



Fortunately, a large overall decrease is seen. The image below depicts the change in death rate between those two time periods. Note that every country with data made available is predicted to have a decrease in death rate.



Recommendation

We are by no means doctors or health care professionals, however based on our findings in the above research, we believe it would be in a person's best interest to consider getting vaccinated and continue to be cautious in public. Both of these measures will help to curb the possibility of contracting COVID-19. Though the vaccination data was not found to be the most helpful indicator for predictive analysis, exploratory analysis revealed a clear pattern with increased vaccinations and reduced deaths. Further, governments should consider distributing vaccines as widely as possible, and maintaining high levels of stringency. Data on COVID-19 variants was not found to be helpful for in-depth analysis, though this could be reevaluated if another, more comprehensive dataset is made available. Although much data is available on the COVID-19 pandemic, from a country-level perspective, the timeframe is rather limited. The rapid spread of variants and the development of COVID-19 vaccines muddle the not-quite two year span, and make the most important factors difficult to identify. In the same vein, a country-level analysis may be too broad, considering the diversity of different regions, particularly in large countries like the United States. This analysis may serve best as a method of monitoring a short-term forecast in COVID-19 deaths, and as a predecessor for future reevaluation, when more data is available on vaccinated and unvaccinated populations, and with respect to the various strains of COVID-19.

Resources

1. https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv
2. https://masks4all.co/what-countries-require-masks-in-public/
3. https://covidtracker.bsg.ox.ac.uk/
4. https://www.kaggle.com/ruchi798/covid19-variants-and-prevalence