**Examining the Efficacy of the Covid-19 Vaccine**

Term Project

07/31/2022

DSC 630-T301

The Algorhythms: Andrew Jordan & Jahedur Rahman

**Introduction**

Early in the year 2020 the world was struck by the rampant spread of the Covid-19 virus, a deadly pathogen that required an unprecedented global response to curtail. Chief among the world’s focus was the creation of a vaccination against Covid-19, a goal that was achieved-with initial distribution following shortly after-in 2021. As this pandemic has had a global effect, most countries have tracked and reported data regarding many aspects of Covid-19 response including hospitalizations, infection rate, severity, mortality, testing, and vaccination rate. With this amount of data, from a global sample size, potentially predicting the long-term effects Covid-19 vaccination will have on the global Covid-19 response is not only possible, but it has become a social responsibility to do so.

**The Data**

Our dataset comes from a combined Johns Hopkins University (JHU) Covid-19 dataset and global data reports compiled by *Our World in Data*. Together these two sets of data provide insight into each country’s case numbers, deaths, hospitalizations, tests, and vaccination rate at an ongoing pace from January 1, 2020 through today. Each of the above variables listed are further broken down within the dataset by continent, new and total reports, and reports per capita. As there is no lack of data available with which to work there appears to be little cause to split the data into training and test data, however the initial analysis may prove otherwise. A link to the dataset is provided below.

**Modeling**

There are two steps we must take to ensure we properly evaluate the data, exploratory data analysis and regression modeling. Through visualization and other Exploratory data analysis techniques we expect to identify any potential relationships and trends occurring within the data. In particular, we anticipate our visualizations will provide a great deal of insight into Covid-19 infection, mortality, hospitalization, and vaccination rate over time. This phase of modeling will require far less time and evaluation than the next, but is a necessary step that will provide the framework by which the next modeling phase will be guided.

To determine the effect the Covid-19 vaccine has had on the global pandemic response we will use a multiple linear regression model to evaluate several relationships, most notably between Covid-19 vaccination rate and reported case data variables. Though the initial Exploratory data analysis may provide unexpected potential relationships to examine more deeply, we expect the initial analysis to guide the focus of our regression modeling to Covid-19 vaccinations’ impact on the reported global case data.

**Evaluating Results**

Examining the predictive model’s results will be done so through the examination and comparison of R2 and p-values, as well as through the creation of visualizations to compare actual versus expected results. Comparing actual versus predicted results will be done so by means of a scatter plot, a simple and efficient tool that can be created to display results both at a singular point in time and throughout the entirety of the pandemic.

**Risks**

By not splitting the dataset into training and test data we run the risk of overfitting, compromising the accuracy of our model’s results. There is also risk contained in including too many variables, potentially creating an indication of statistical significance that does not actually exist.

**Contingency Plan**

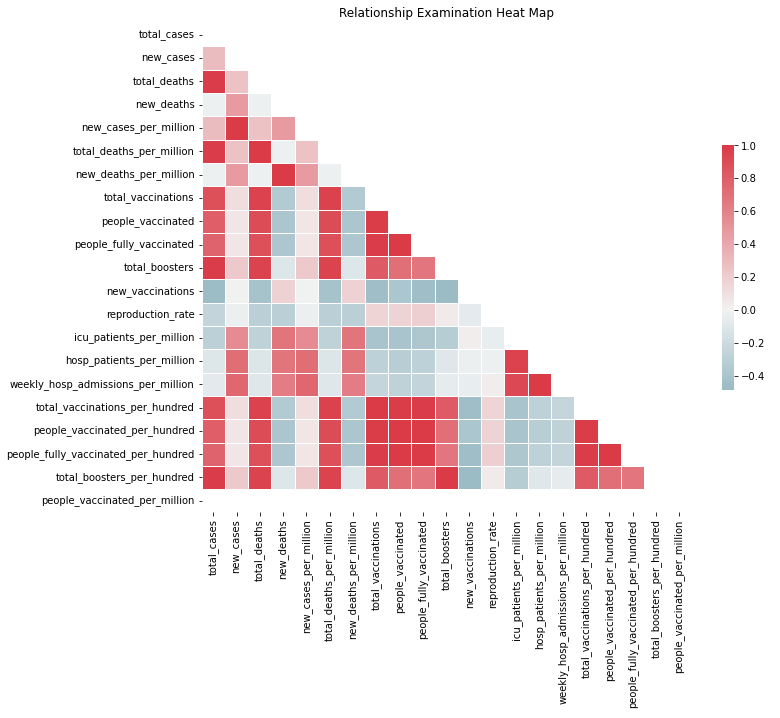
The two portions of our contingency plan exist to counter the two biggest risks posed by our proposal. If we determine we have compromised the accuracy of our model through overfitting, we will split our data into training and test data before re-running our regression model. The second portion of our contingency plan relies upon risk mitigation through the evaluation of our initial exploratory data analysis to identify and include only the most relevant variables. Failure to do so will require a re-evaluation of the variables included and starting the modeling process over.

**Preliminary Analysis**

Our main Covid-19 dataset consists of Covid-19 data from about 241 countries. Each country has about 800 to 900 records of Covid-19 data. In total there are 194,032 rows of data. In addition, there are 67 columns/variables in the data set. It is highly likely that not all the columns will be used. Some columns we will most likely use are new\_cases, new\_deaths, reproduction\_rate, weekly\_icu\_admissions, and weekly\_hosp\_admissions. With the data we will use, we will be able to answer the questions that we want answered.

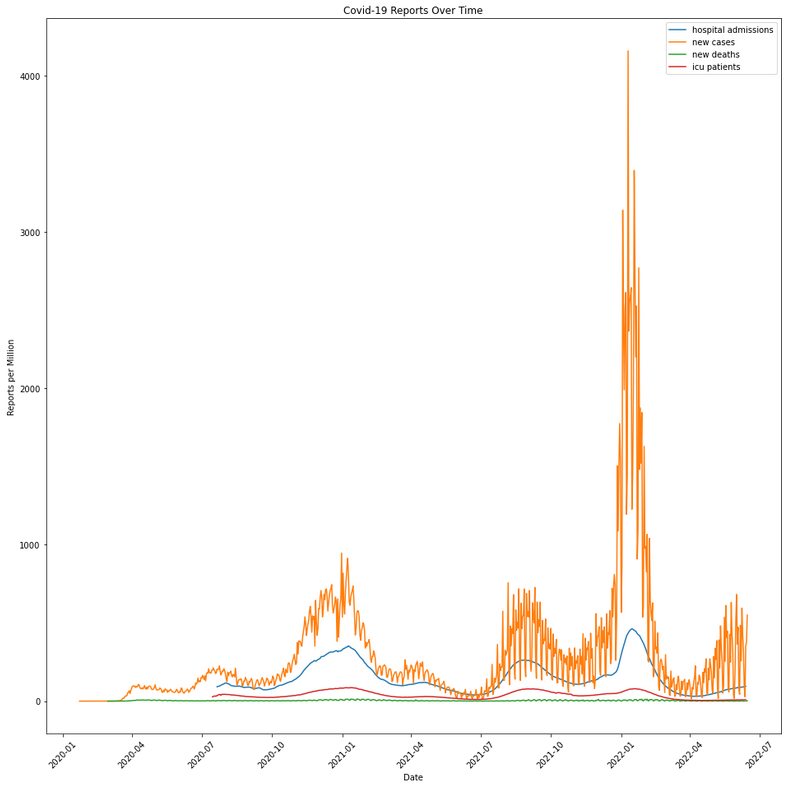
For our initial findings, we used a heatmap.

Figure 1



Doing this helped us narrow down the variables we will be using. In addition, using scatterplots and/or line charts will help in comparing multiple variables.

Figure 2



In this chart we can observe that these variables are related to each other.

The data that we are using will be adjusted/transformed. Since there are so many columns in this dataset, some of the columns will not be necessary. In addition, there are many NaN values. Right now, we have decided to fill those NaN values with zeroes. However, we may decide to change this decision depending on future results. Additionally, picking which countries in the dataset that will be used is something we are thinking about throughout the project. For now, we are only using United States Covid-19 data. The results we observed when using only United States Covid-19 data is that there was a drop in hospital admissions after vaccination rollout, a spike as new variants arose, and then a drop again when boosters rolled out. We have also observed that new deaths drop when vaccinations spike, and vice versa.

For our model we first chose to use a multiple linear regression model to evaluate several relationships, most notably between Covid-19 vaccination rate and reported case data variables. However, after plotting different variable and comparing the relationships between them, the data isn’t very linear for many of the comparisons/relationships. So now we will be using Random Forest Classifier.

So far, we still believe our original expectations seem reasonable. The results we have observed so far are favorable and will help in answering the questions.

**Data Preparation and Modeling**

Our first step in preparing the data for analysis was to curate the data with which we will be running the model. Selecting only data from within the United States, we identified the most relevant features and they were extracted and entitled “covid\_subset”. Missing data remained, and as the missing data only represented the absence of data at that point all null values were filled with zero.

To ensure best results from the model a new feature was created entitled “change”. This new feature was to act as the target variable, tracking when there were positive changes to weekly hospital admissions per million with a 0 for no increase from the previous week and a 1 indicating an increase from the previous week.

Using a 70/30 split the “covid\_subset” data was split into training and test data with the “change” feature as the target. The data was fit to the Random Forest Model, and unlike our early attempts with a Linear Regression model we saw excellent results with a returning accuracy score of approximately 95%.

To further verify the validity of our results a Confusion Matrix Heat Map was generated, as seen below in figure 3. The accuracy of our model was again evident as there were only 4 instances in which change was predicted and did not occur, and only 13 instances in which change occurred but no change was predicted.

Figure 3

Chart, treemap chart

Description automatically generated

**Interpretation**

Our initial results appear very clear: hospitalization rate per million showed a significant relationship with vaccination rate. As the vaccination rate increased, the exponential rate at which hospitalizations had been increasing first slowed and then began to decrease. When observing the data, hospitalization rates after vaccination rollouts began only showed significant week-to-week increases as new Covid-19 variants arose, but again tailed off following further vaccination booster rollouts.

**Initial Recommendation**

Though ideally the data could be separated by variant type, the real-world logistics of tracking exactly what type of Covid-19 each individual has had is unrealistic. Without that any recommendation is reliant upon the results shown by our model, which again indicate a strong relationship between vaccination and lower rate of hospitalization. Our conclusion is in line with the medical community’s recommendation: being vaccinated correlates with a lower chance of hospitalization from all Covid-19 variants seen thus far.

**References**

https://github.com/owid/covid-19-data/tree/master/public/data

*CDC Museum COVID-19 Timeline*. (2022, January 5). Centers for Disease Control and Prevention. https://www.cdc.gov/museum/timeline/covid19.html

Assistant Secretary for Public Affairs (ASPA). (2022, August 4). *COVID-19 Vaccines*. HHS.Gov. https://www.hhs.gov/coronavirus/covid-19-vaccines/index.html