Predictive Analysis on Global COVID-19 Dataset

Matt Dannheisser & Henry Luong

Motivation

COVID-19 has impacted most of our lives at one point or another within the past year. The world has been faced with a pandemic and companies are rapidly mass-producing vaccines to combat this virus. We hope that the vaccines will eventually stop the spread, but it is difficult to know how long this will take and what will be the rate at which the spread is curbed over the near future. Our goal is to address these concerns by developing a model that utilizes each countries' data and its input features. By creating these forecast models, it will let users determine the current state of the world and the most likely direction that it is heading.

Overview of Methodology

Our analysis predicts how the future spread of COVID-19 will be affected by the supply of the vaccines and the willingness of individuals to take the vaccines. To predict the supply of the vaccines, we modeled a scenario provided by the World Health Organization's Fair Allocation Framework (file in supporting documentation.) The WHO's FAF breaks the distribution of vaccines into two stages. The first stage promises that each country will be distributed vaccines for 20% of its respective population so that at-risk individuals and front-line health care workers can be vaccinated first. The second stage promises equal distribution of the vaccine based upon available supply. The FAF did not mention when the stages would begin, but we begin stage one on December 22nd, 2020, and stage two on April 1st, 2021. Following this framework, rapid supply by manufacturers across the world aim to have enough dosages for every person by the end of 2021. However, this does not mean that everyone will be willing to take a vaccine. We modeled demand using a report by the IPSOS Group that samples the varying willingness levels for a country's citizens to take a vaccine and bins those figures into representative probabilities which are then weighted against the population data resulting in demand predictions by country. For example, when presented the statement: "*If a vaccine for COVID-19 were available, I would get it*" subjects that responded "strongly agree" were assigned a 95% likelihood of getting the vaccine while those who answered "strongly disagreed" was assigned a 1% chance of getting the vaccine.

The final part of our analysis utilizes the daily new case data from January 1st, 2019 to February 2nd, 2021 inside or a SARMA function in order to forecast daily cases that demonstrate the trend and seasonality. Each country's daily new cases projections are then assigned weights corresponding to the number of susceptible people (those who have not received a vaccine or have not contracted the virus.)

Hypothesis

Our hypothesis before we began was that there was a negative correlation between the supply rate of vaccine and the spread of COVID-19 whereas the vaccines started increasing, we will see the future spread start to decrease.

To fully realize this scenario, we had to identify a multidimensional dataset that can provide us with the most up-to-date and accurate data for each country. Hannah Ritchie of ourworldindata.org was able to curate this data provided by the John Hopkins University's Center for Systems Science and Engineering team. All their data is updated daily and includes data on confirmed cases, deaths, and testing. Our

model will provide predictions by country using daily contagion data, Ipsos Group S.A. research reporting a country's willingness to take a vaccine, and population data.

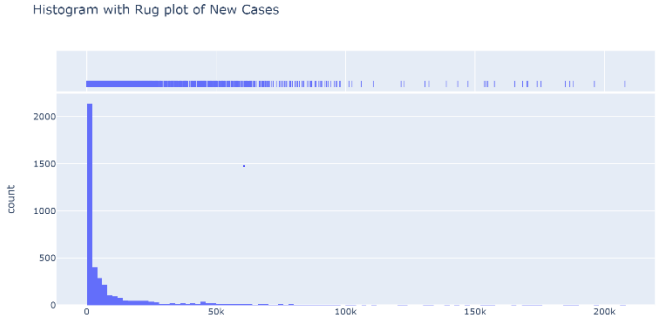
Data Sources

	COVID-19 Spread	Global Attitude on Vaccine	Country Population
Description	Daily new case of COVID by country. This will confirm total and daily cases all provided by CSSE at John Hopkins University.	Sampled likelihood of citizens to take COVID-19 vaccine based on survey research by the IPSOS Group.	The total population for each country based on the latest United Nations Population Division Estimates
Size	10,655 Kb	680 Kb	Small HTML table
Location	Our World in Data	PDF provided in folder	Population by Country (2020) - Worldometer
Format	CSV	PDF	HTML
Acess Method	Direct Access	Manually entered into pd.DataFrame	API
Variables Used	country, date, total cases, new daily cases	countries, varying levels of willingness to take vaccine	country, population

- This analysis utilizes data pulled on 12/4/2020 with datapoint dates ranging from 12/31/2019 - 11/29/2020. All data points past this range represent projected figures.
- Only countries that are included in the Global Attitude on Vaccine study were included.

Data Exploration

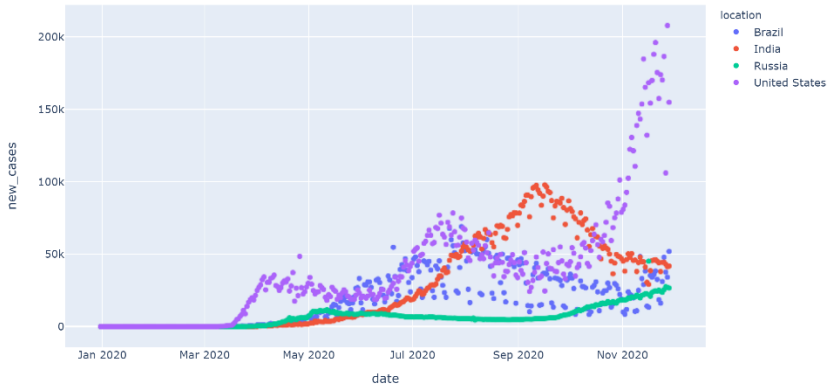
We went into exploration in the search of anything that stood out of the ordinary. Through conversions of comma-separated values and web scraping of portable document formats into data frames via the pandas' library, we were able to detect any noise within the dataset. Using the describe method shows a five-number summary that revealed negative values within the new_cases feature. We speculated that this might have been due to human error reporting. Furthermore, these values only accounted for .1% of the dataset, and so we decided to keep them



due to its' negligible effects. We also had to keep in mind that the data shows the total global data as well and so when we run our model, we need to remove it in case of misinterpretation.

A histogram of each feature lets us determine the shape and continuous spread of the data. We also want to confirm that the distribution for the number of new cases followed a right tail distribution since usually the number of cases for each country start out small with unexpected increases being the outlier.

A time-series analysis via scatter plot and line plot of each feature allows us to spot any outliers within the dataset. Overall, we were able to spot only a couple of points that were interesting, to say the least. For example, China reported the number of new cases to be as high as 15,000 on February 13th, only to fall about 75% the following day. We attributed this reporting as unusual but not uncommon because COVID-19 originated from Wuhan, China. Similarly, other countries have the same trend when reporting high numbers, but we decided to leave them in since we did not want to reduce our training data and cause overfitting in our predicted analysis. The scatter plots were also useful for observing trends and seasonality in the dataset at a high level. We deemed that a combination of an autoregression and seasonality analysis could be useful for making projections.



In our COVID-19 dataset, we have a certain attribute called reproduction rate that will allow us to measure how effective our collective behaviors, such as mask-wearing and social distancing, are in slowing the growth of the virus. If the rate is above 1, the virus spread is increasing, if it is below 1, that means that the virus is spreading slowly. We can look at the data of reproduction rate and see that it looks fairly normally distributed with some outliers. We believe that the reproduction rate offers significant value in helping reduce the spread of COVID-19 but cannot verify without additional research and data as the pandemic continues.

Data Manipulation

Data manipulation occurred in three key steps:

1. **Assigning the demand for the vaccine**

First, a dictionary is created with the willingness bins ranging from “strongly agree” to “strongly disagree” comprising the keys and the respective probabilities that individuals in that bin will get the vaccine comprising the values. Then the Global Attitudes data frame is merged with the population data frame. This allows for the country’s demand to be assigned in a for loop that checks the bins and assigns weighted probability times the population for each bin. The demand for each country is the sum of these weighted figures. The result is:

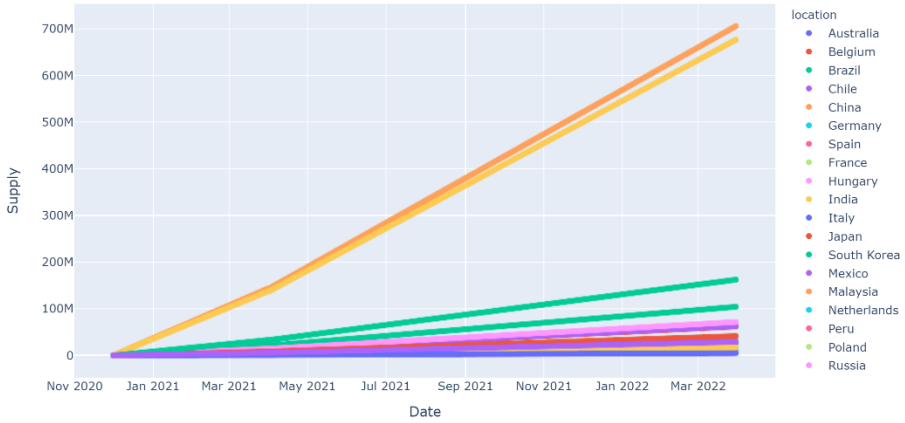
	location	population	Vaccine Demand
0	Argentina	45195774	29,354,655.0
1	Australia	25499884	18,997,414.0
2	Belgium	11589623	6,655,920.0
3	Brazil	212559417	164,606,013.0
4	Canada	37742154	24,800,369.0
5	Chile	19116201	11,270,912.0
6	China	1439323776	1,035,017,727.0
7	France	65273511	31,004,918.0
8	Germany	83783942	47,698,198.0
9	Hungary	9660351	4,243,792.0

2. **Building forecasts for the supply of the vaccine based upon FAF hypothetical**

The daily supply of the vaccine is determined by assigning projection rates under Stage 1 and Stage 2 of the FAF. The first stage is assigned a production rate of 20% of the population divided by the length of Stage 1 which has been set to 90 days. This supply figure is assigned to each country under Stage 1. Then in Stage 2, we determined that the supply chain would experience a 30% boost as more vaccines receive governmental approvals. These daily values are assigned to each country under Stage 2. Then a net total is applied to each day giving the total supply per country each day under COVID. This resulting data frame is merged with the Supply data frame on the country aka location column. The result is:

Date	location	population	Supply	Vaccine Demand
2020-12-01	Argentina	45195774	0.0	29,354,655.0
2020-12-02	Argentina	45195774	37,351.9	29,354,655.0
2020-12-03	Argentina	45195774	74,703.8	29,354,655.0
2020-12-04	Argentina	45195774	112,055.6	29,354,655.0
2020-12-05	Argentina	45195774	149,407.5	29,354,655.0

The projected supply and demand shown in the figure below represent the FAF. It demonstrates a scenario where every country acts in the best interest of the global population rather than in the best interest of the individual country.



3. Forecasting the daily cases without the vaccine and then adjusting those cases for the reducing susceptible population

First, the COVID data frame is pivoted so that the dates are on the index, countries are on the columns and daily new cases are the values. The countries are looped through and a SARMA model is fitted to each country's new cases. A SARMA model was chosen from other time series analysis as it allows for the trend and seasonality to be demonstrated. The SARMA function is given parameters p=30, q=5, and trend="t". This means that the autoregression used a lag of 30 days, the moving average used a 5-day window, and then there is a linear time trend. These estimators were chosen for simplicity, but later extensions of this analysis will include estimator optimization.

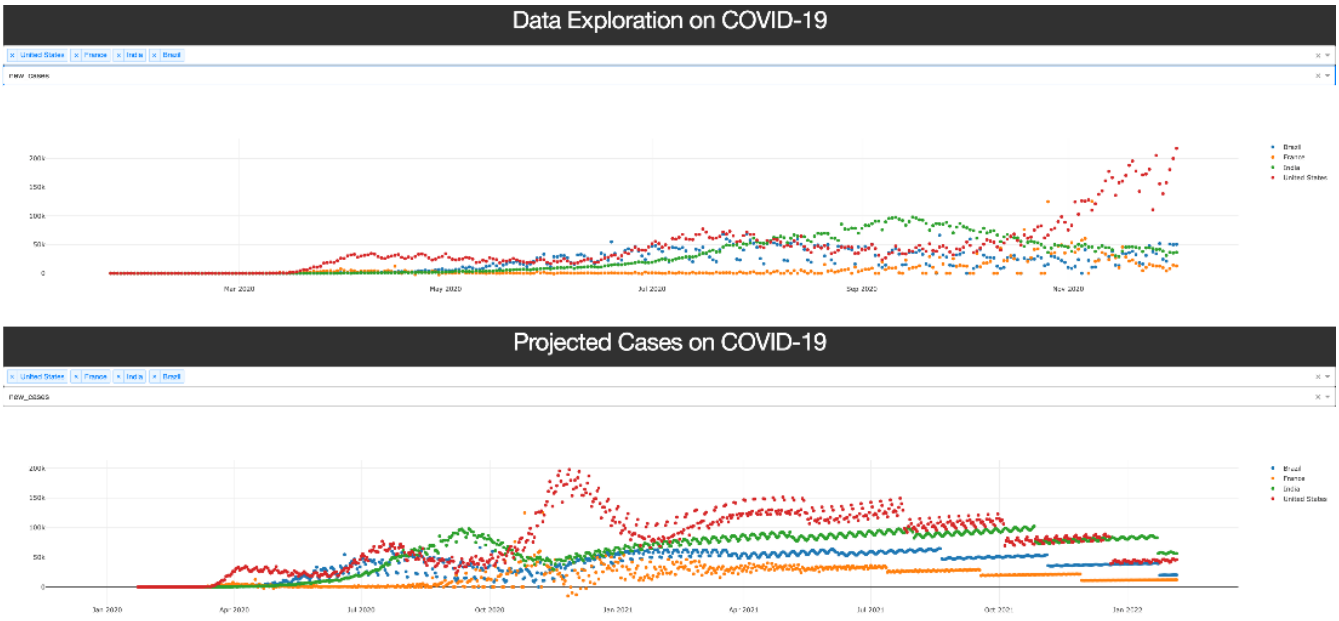
Once the daily predictions pre-vaccine has been determined, the data is now ready to be combined with the vaccine supply and demand data. The vaccine supply and demand is used to determine the total number of people that have the vaccine and are no longer susceptible (assuming a 100% efficacy.) The minimum of the two figures is used. This is because the supply starts to lower and then eventually catches up to the demand in many countries. Although the supply might overtake the demand, the minimum is the only one that is used so that countries who are averse to taking the vaccine will experience worse results. This daily figure will be added to the total cases creating the total immune figure. This figure is then divided by the population and giving the percent immune figure. This percent immune is then used as a weight on the daily cases based upon the weighting figure to the right. For example, if a country has reached 70% immunity by January 1, 2022, then the daily cases will be reduced by .90%.

	Immunity	Affect on Daily Cases
0	0.0	1.0
1	0.1	1.0
2	0.2	0.9
3	0.3	0.8
4	0.4	0.6
5	0.5	0.4
6	0.6	0.2
7	0.7	0.1
8	0.8	0.0
9	0.9	0.0

Once we were able to obtain the predicted data and the future projected total and new cases for each country, we decided that to let it speak for itself. While staying objective but allowing for some room for user interactivity, we built a dashboard through Plotly that features both the current number of cases as well as the future spread of COVID-19. Dash is rendered in the web browser and allows users to be able to view the scatter plot and pick which feature they would like to view and compare with each country as they see fit. We felt that the dash should be as simple as possible to avoid any cluttering while still be visually appealing.

Once the dashboard has been created, users are able to interactively control which countries to view via the multi-dropdown component. With the countries selected, users are able to view the many features that the graph has to offer (total_cases, new_cases, stringency_index, and reproduction_rate. The top plot views the current cases beginning January 2020-current. The bottom plot allows viewers to view the forecast or future projections of total and new cases per country.

An issue that we had hoped to accomplish when incorporating the dashboard was to reduce the number of countries viewed in one scatterplot. With too many countries, we run the risk of not being able to analyze the trends of each country due to our perception. What the dash dropdown widget allows us to do is let us, as viewers, select which country we would want to compare the total or new cases and see if the spread is being reduced or not.



Conclusion/Summarization

It should also be noted that we must make assumptions that the COVID-19 dataset is the most accurate and precise measurement of data that we are able to use. Recognizing and reporting the true number of individuals who are infected with COVID-19 in the world is essential in not only predicting the future spread but also understanding the disease. The true number of infections in the data is many times

greater than the number of reported cases simply due to the fact that more than half of the infection rates are not detected or reported.

After looking at our model and the projection for the number of total and new cases in each case, we can hypothetically confer that infection rates for each country will reach a stationary phase as supply exceeds the demands. As national immunity increases, viral infection rates will start to decelerate causing a steady rate as shown in the projected cases. In countries with large demand, we do expect projected cases to continue to rise a lot longer until supply has reached an inflection point that exceeds this vaccine demand pressure. For example, it can be seen in the graph that large countries such as the United States' constant rate of infection begin to slow down around January 2022 compared to smaller countries such as France, which seems to slow down around July-October 2021. Furthermore, our assumptions would require additional research since our model only depended on the vaccine's supply and demand. If we were to include additional factors such as the stringency index, reproduction rates, etc. into our model, we do foresee a significant change in the projections of COVID-19.

Statement of Work

Both team members fulfilled their responsibilities laid out in the Project Proposal and both played essential roles in delivering this report. Specifically, Henry was pivotal in conducting the data exploration, building the dash interactive visualization. Matt was responsible for the data manipulation and building the SARMA model for making forecasts.