COVID-19 Pandemic Forecasting Using LSTM

1. Introduction

This project is for the Capstone Project in the IBM Advanced Data Science Specialization on Coursera. It aims to utilize all Data Science Concepts learned in the IBM Advanced Data Science Professional Course. Within the project, the COVID-19 trend will be explored using, analyzed and forecasted using Machine Learning tools such as LSTM. The code will be programmed in python using the Jupyter notebook environment, which can be found on my GitHub repository at https://github.com/wnich/COVID-19-Pandemic-Forecasting-Using-LSTM.git.

1.1 Background

The COVID-19 (coronavirus disease 2019) is an infectious disease that is caused by the SARS-CoV-2 virus and was first reported in China in late 2019, as the cause of a new type of pneumonia [1]. The COVID-19 is infectious and quickly transmitted through close human-to-human contact. It has spread throughout the world with the number of infected cases and deaths rising, which becomes a pandemic. Even though the use of Covid-19 vaccines led to a decrease in mortality rate, mutations of the virus can cause new COVID-19 surges.

As known, variants are a common occurrence in viruses. They are constantly changing through mutations, and new variants of a virus are expected to occur over time. Most of the SARS-CoV-2 virus mutations do not make a big impact on mankind. However, there are still those that people need to be concerned about. Therefore, to minimize the number of infected cases, many countries followed procedures that included getting vaccination, quarantine, online schools and businesses, and bans on travel [2, 3].

1.2 Business Problem and Target Audience

The COVID-19 pandemic has fundamentally altered the worldwide landscape and dramatically changed people’s lives and behaviors. Border closures, rampant hoarding, self-isolation policies, and an economic recession have all resulted from a high reproduction rate and an increased risk of complications. Thus, it is crucial to determine the spread rate to help governments to plan public health and develop policies and strategies to minimize the risk of transmission COVID-19. It also allows companies to have ideas of the demand so that they can act accordingly and plan their way out. To determine the COVID-19 spread rate is to accurately predict the number of cases at a given time.

1.3 Objective

The primary objective of the project is to perform exploratory data analysis and to aid in the accurate forecasting of confirmed COVID-19 cases by using Long Short-Term Memory (LSTM) architecture, a deep learning technique for building the time-series forecasting model.

Deep learning techniques have been used for a prediction of the COVID-19 and feeding it into demand forecasting. They have recently been successfully applied to a variety of difficult prediction problems encountered in the real world, such as time-series forecasting. Since the COVID-19 pandemic has shifted into a global pandemic, real-time data examinations are required to provide the population with a strong course of action to combat the infection. LSTM is one of the most effective and widely used deep learning approaches. It has been applied by several studies to forecast COVID-19 cases. LSTM is a type of Recurrent Neural Network (RNN) that can better handle long-term memory. An unit is typically made up of a cell, an input gate, an output gate, and a forget gate. The cell retains values over arbitrary time intervals, and the three gates control the flow of sequence pattern information into and out of the cell [4].

2. Data Collection

Our World in Data has a complete COVID-19 dataset, which is updated daily throughout the duration of the COVID-19 pandemic. It includes the following data:

- Vaccinations -- Official data collated by the Our World in Data
- Tests & positivity -- Official data collated by the Our World in Data
- Hospital & ICU -- Official data collated by the Our World in Data
- Confirmed cases -- JHU CSSE COVID-19 Data
- Confirmed deaths -- JHU CSSE COVID-19 Data
- Reproduction rate -- Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C
- Policy responses -- Oxford COVID-19 Government Response Tracker
- Other variables of interest -- International organizations (UN, World Bank, OECD, IHMEv)

Sources: https://github.com/owid/COVID-19-19-data

3. METHODOLOGY

3.1 Data Preparation and Cleaning

3.1.1 Extracting data from the Github repository and create dataframes

Create dataframes of global COVID-19 cases, vaccination and variant information with the data extracted from the Github repository. Then, cleaning the data to be ready to be used for analysis and time-series forecasting.

3.2 Exploratory Data Analysis

Exploratory data analysis refers to the crucial process of conducting preliminary investigations on data in order to uncover patterns, spot anomalies, test hypotheses, and understand keys using summary statistics and graphical representations.

After cleaning the data, the next step is to visualize better pictures of global COVID-19 trends:

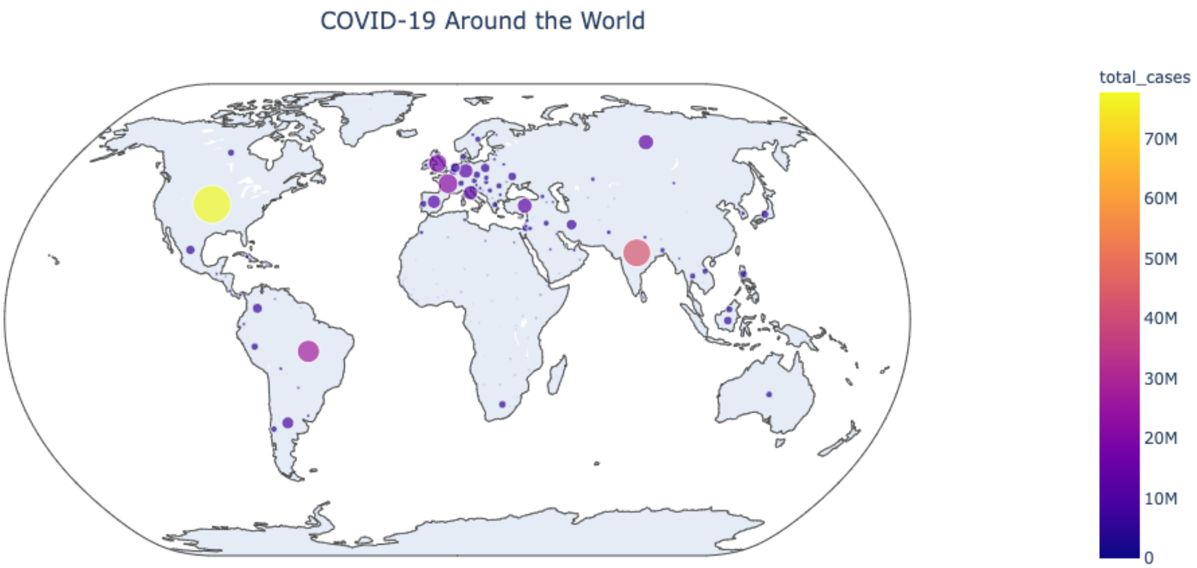


Figure 1: Global COVID-19 Infection Rate
The darker coloration indicates the increased concentration of COVID-19 cases.

Next, let's explore the trendlines of the global total COVID-19 confirmed cases, total deaths, new confirmed cases and new deaths since the pandemic started to provide useful insights.

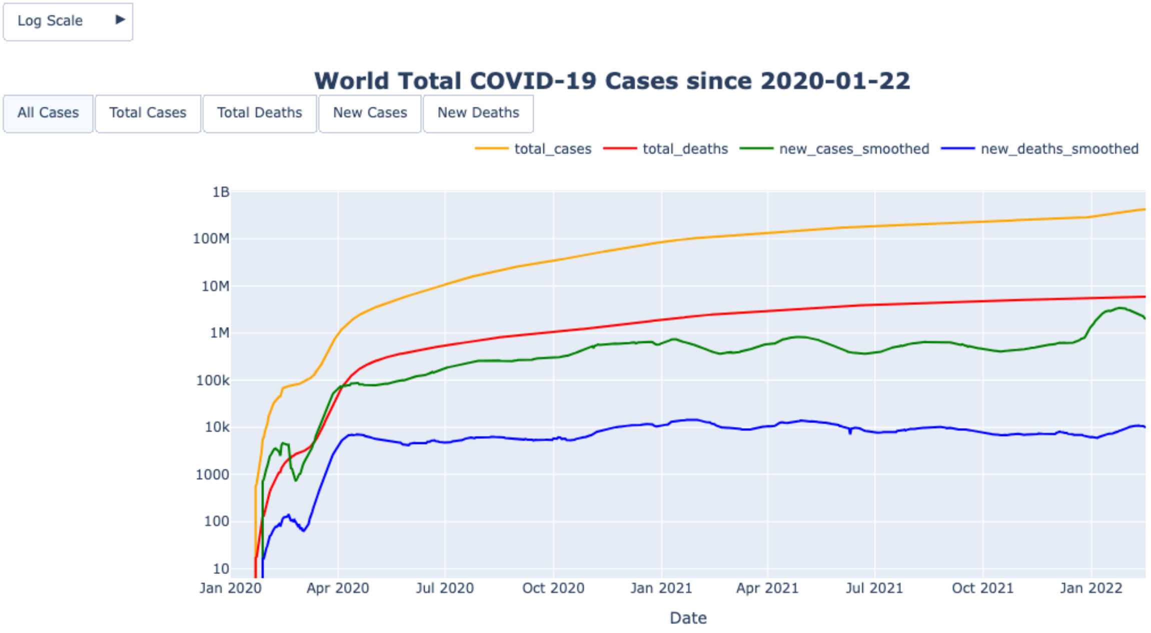


Figure 2: Log Scale of the Global COVID-19 Cases

As shown in the figure above, the global total confirmed case trend is slightly increasing but the trend in new confirmed cases has been decreasing parabolically since the end of 2021. While the total deaths trend is fairly steady.

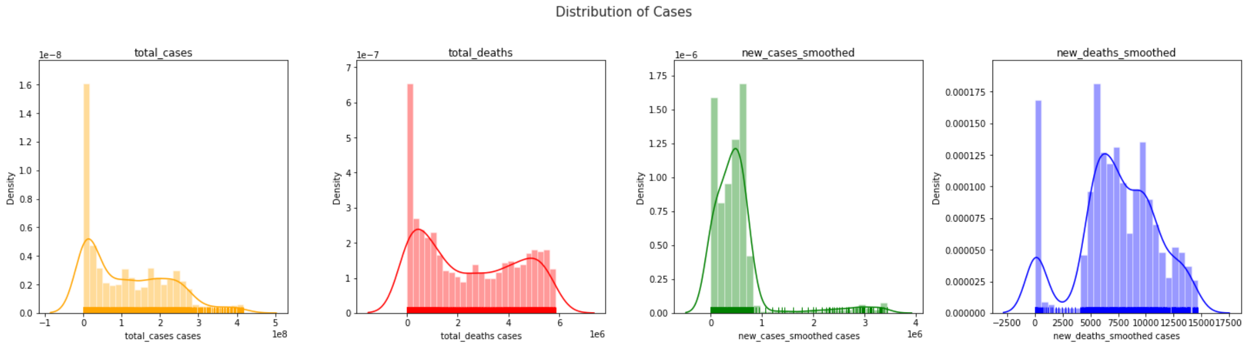


Figure 3: Distribution of Global COVID-19 Cases: Total Confirmed Cases, Total Deaths, New Confirmed Cases, and New Deaths

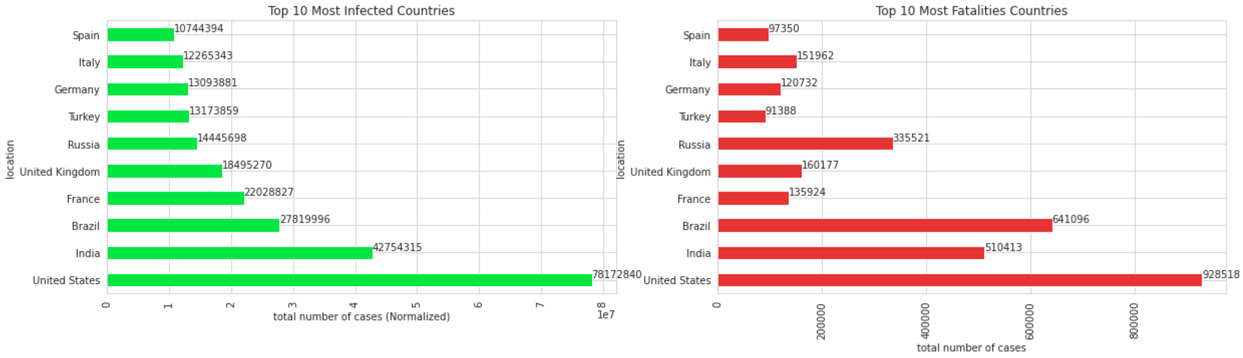


Figure 4: Top 10 Countries with Highest Number of COVID-19 Cases and Total Deaths

The above figure shows the numbers of confirmed cases and deaths of the top 10 most affected countries where COVID-19 has spread. We can see that the top most infected countries also have high mortality rates. The United States has been the hardest hit. There have been 77,707,349 confirmed cases. There have been 919,255 deaths in all (updated on 02/12/22).

Let's have a look at the spread of the following variants of concern: 'Alpha', 'Beta', 'Delta', and 'Omicron'.

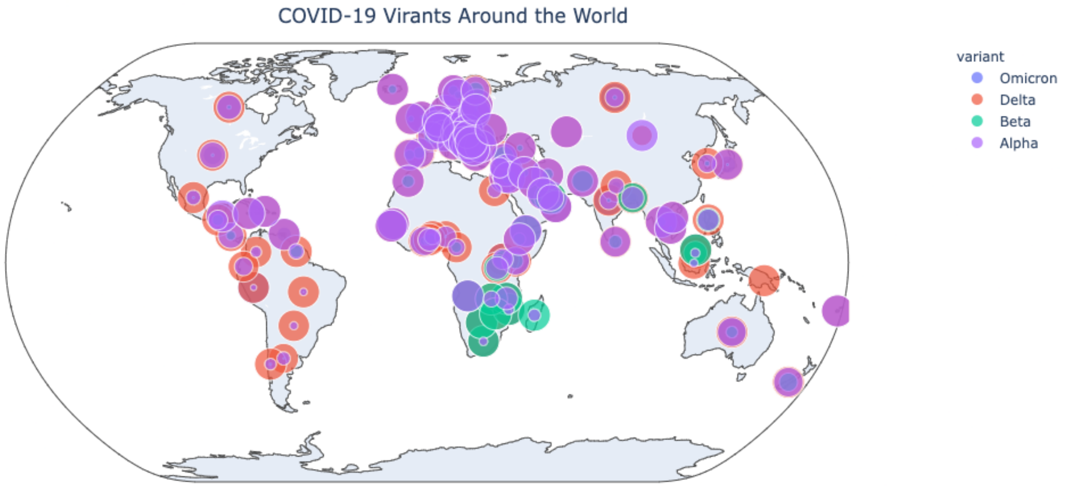


Figure 5: Amount of COVID-19 Cases per Variants of Concern Around the Globe

The table and plot below display the top 5 countries affected by each variant of concern the most. The latest COVID-19 variant, Omicron, has been detected in the UK the most. While the most infected variant in the United States is Delta, followed by Alpha.
Note that the dataset is a little old so the numbers might have changed by now.

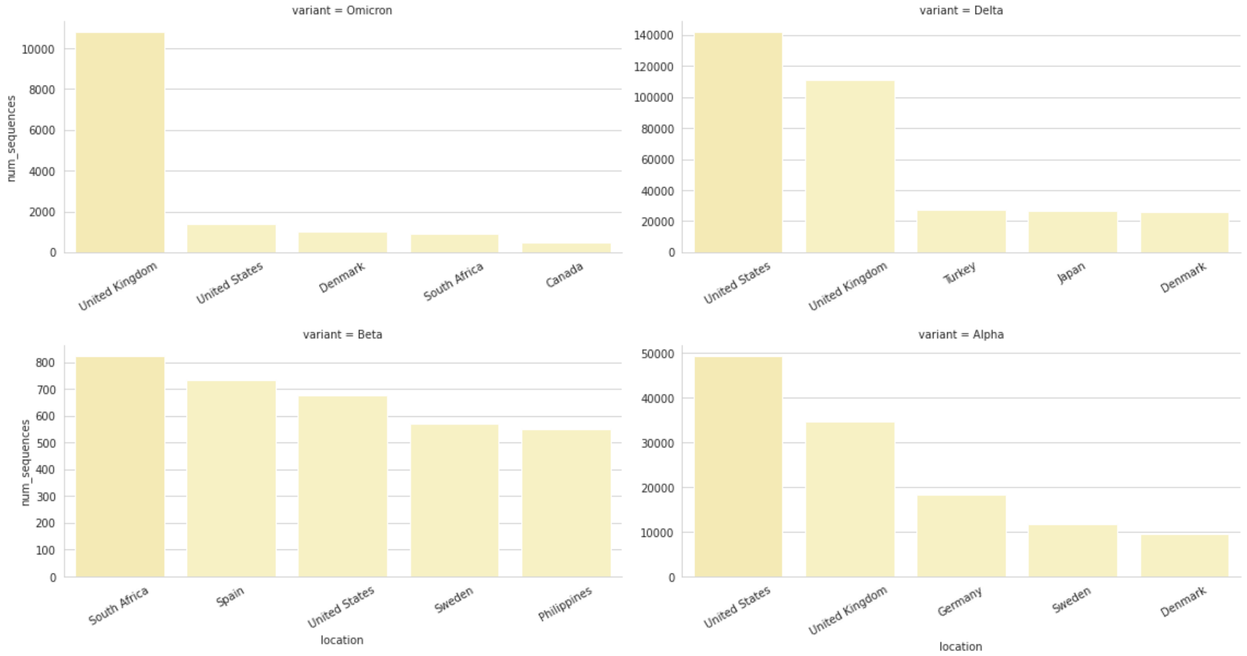


Figure 6: Top 5 Countries Affected by each Variant of Concern

Before driving into the analysis, let’s look at COVID-19 vaccination.

The following choropleth map is the total vaccinations around the world, per country, which shows the total number of COVID-19 vaccination doses administered. It is important to note that this is a single dose and may not equal the entire number of people vaccinated, depending on the specific dose regimen (e.g. people receive multiple doses).

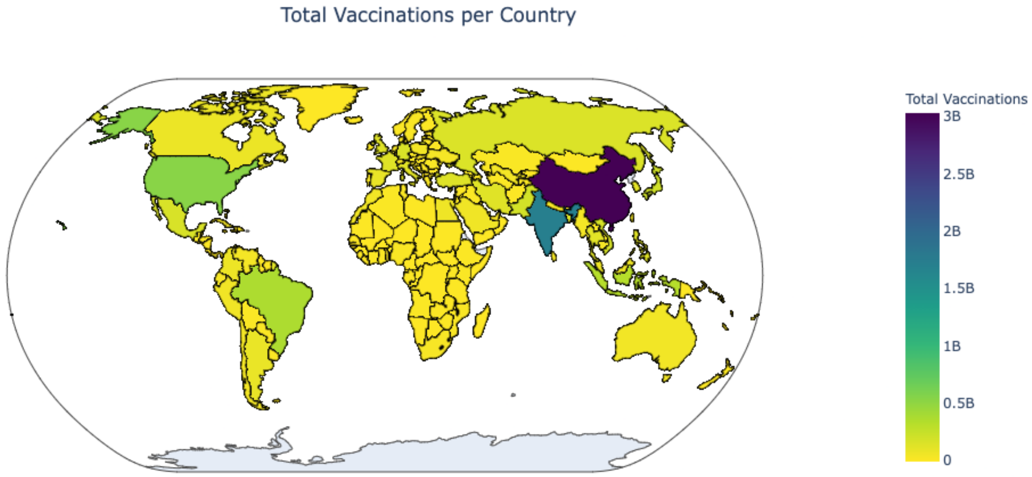


Figure 7: the Total Number of COVID-19 Vaccination Doses Administered Around the Globe

The map depicts which COVID-19 vaccine brands are accessible globally and the number of countries that use them. AstraZeneca vaccine is the most widely used around the world, followed by Pfizer and Moderna, respectively.

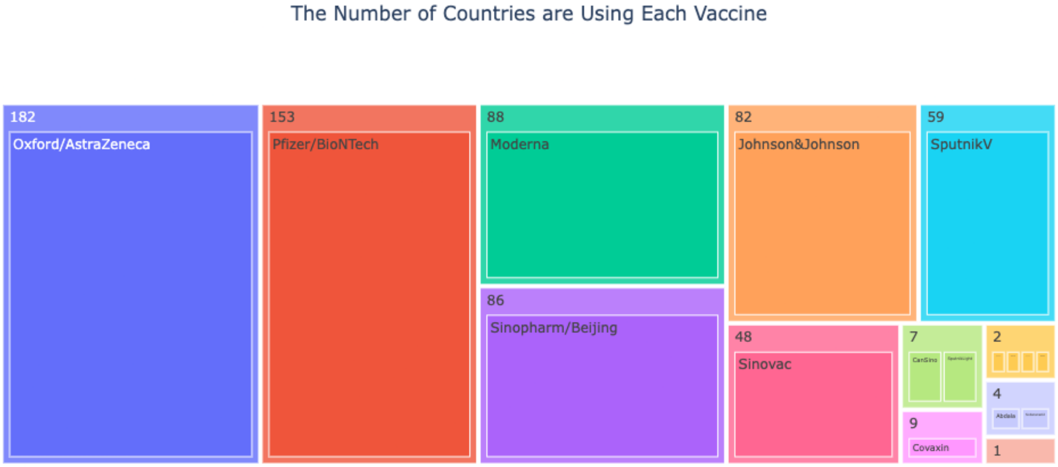


Figure 8: Number of countries where each vaccine is being used

Displayed in Figure 9 is a plot of vaccine combinations against COVID-19 that are preferred presently in the world. The most used combinations are the following, respectively:

- CanSino, Sinopharm/Beijing, Sinopharm/Wuhan, Sinovac, and ZF2001– China;
- Covaxin, Oxford/AstraZeneca, and Sputnik V – India;
- Johnson & Johnson, Moderna, and Pfizer/BioNTech – U.S.;
- Johnson & Johnson, Oxford/AstraZeneca, Pfizer/BioNTech, and Sinovac – Brazil;
- Johnson & Johnson, Moderna, Novavax, Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing, and Sinovac – Indonesia.

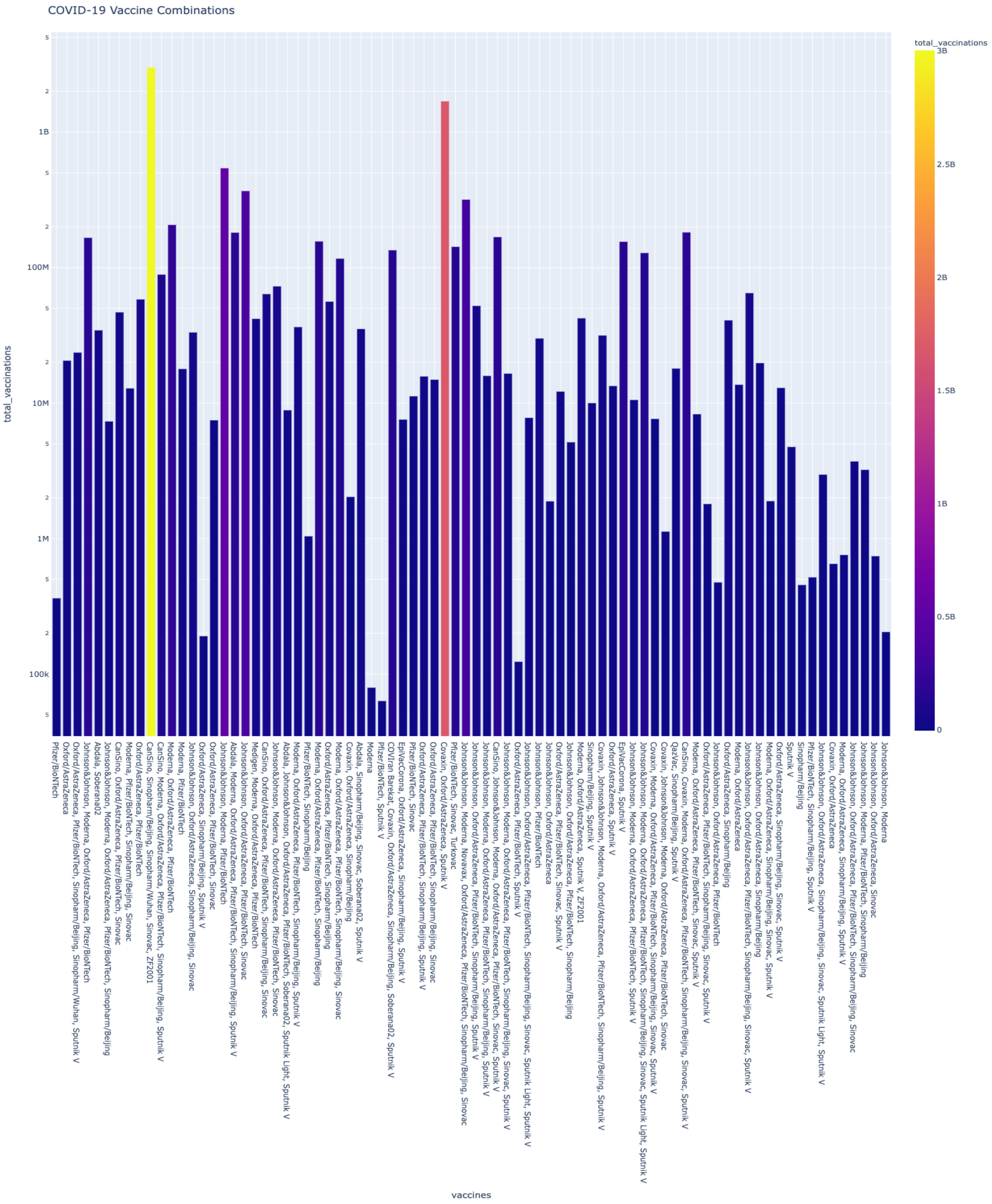


Figure 9: COVID-19 Vaccine Combinations

The following chart shows the top 10 countries with the percentage of fully vaccinated population. Fully vaccination entails receiving all of the doses specified by the vaccination regimen, which is a two-dose vaccine, and this information is only available for countries that provide the number of doses delivered in the first and second administrations. Gibraltar's percentage of fully vaccinated population exceeds 100%. It is due to vaccination of non-residents.

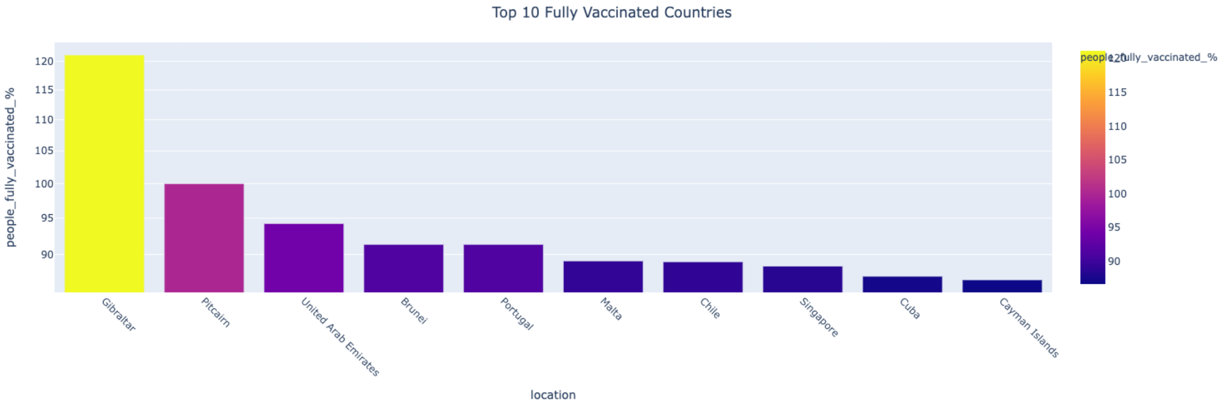


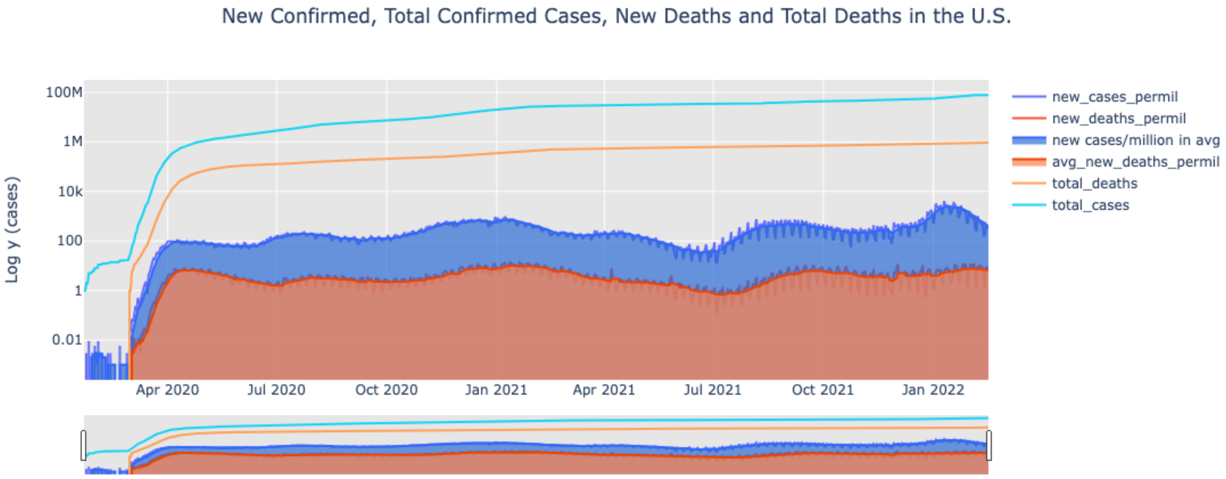
Figure 10: Top 10 Fully Vaccinated Countries

Select some countries in the following conditions to compare breakdown of cases:

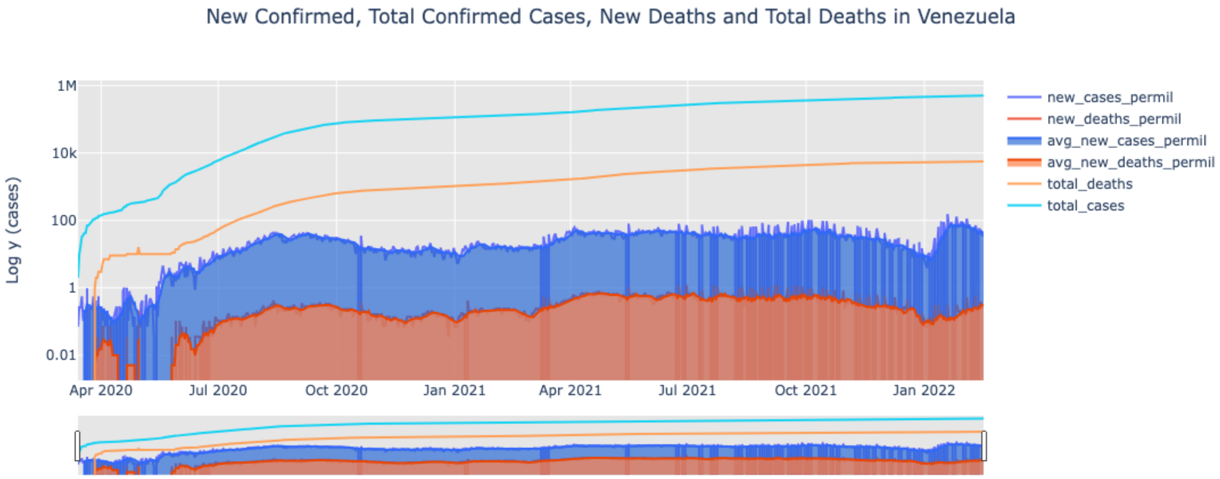
- Mainly uses mRNA vaccines – U.S.;
- Mainly uses viral vector and/or inactivated vaccines – Venezuela;
- Mix and match vaccine types and classified as a high-risk area – Germany;
- Lowest vaccinated population percentage – Haiti.

First, let's look at trendlines of new confirmed, total confirmed cases, new deaths and total deaths with respect to time of each country individually.

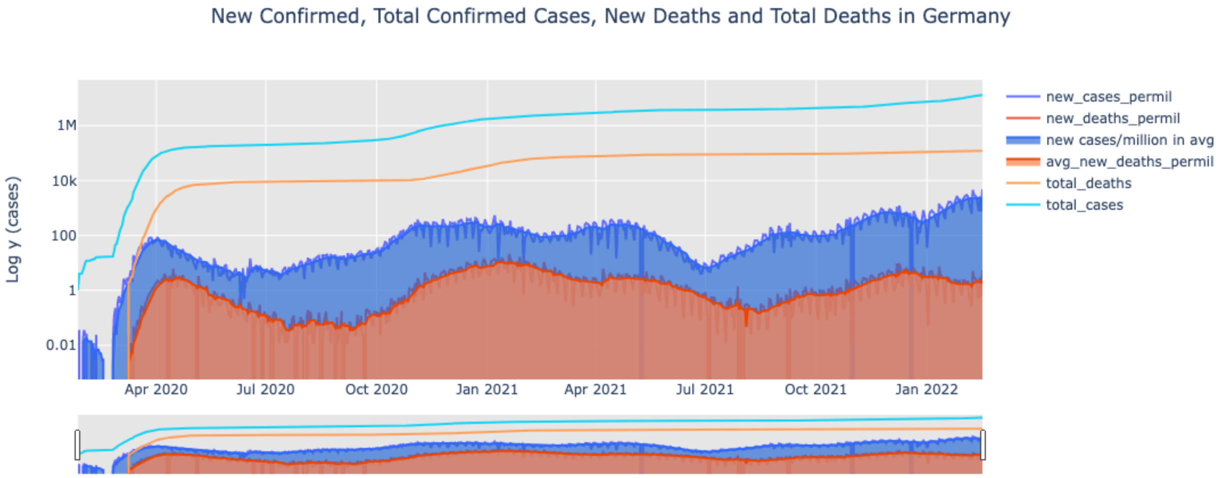
United State



Venezuela:



Germany:



Haiti:

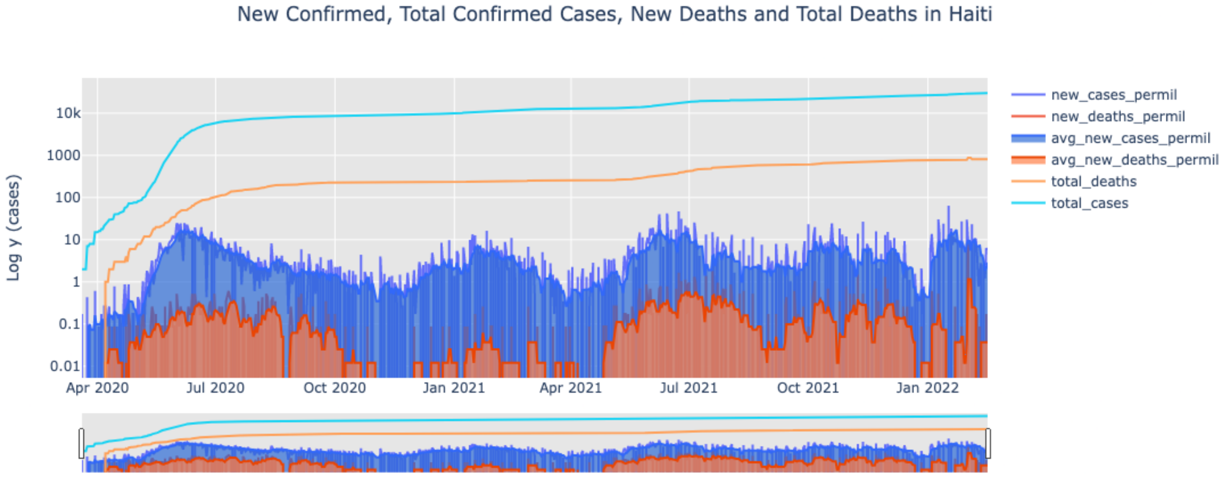


Figure 11-14: Cases and Trends of COVID-19 in the United States, Venezuela, Germany and Haiti, since the pandemic started

To get better pictures of each case in these countries, let’s compare each case in the countries.

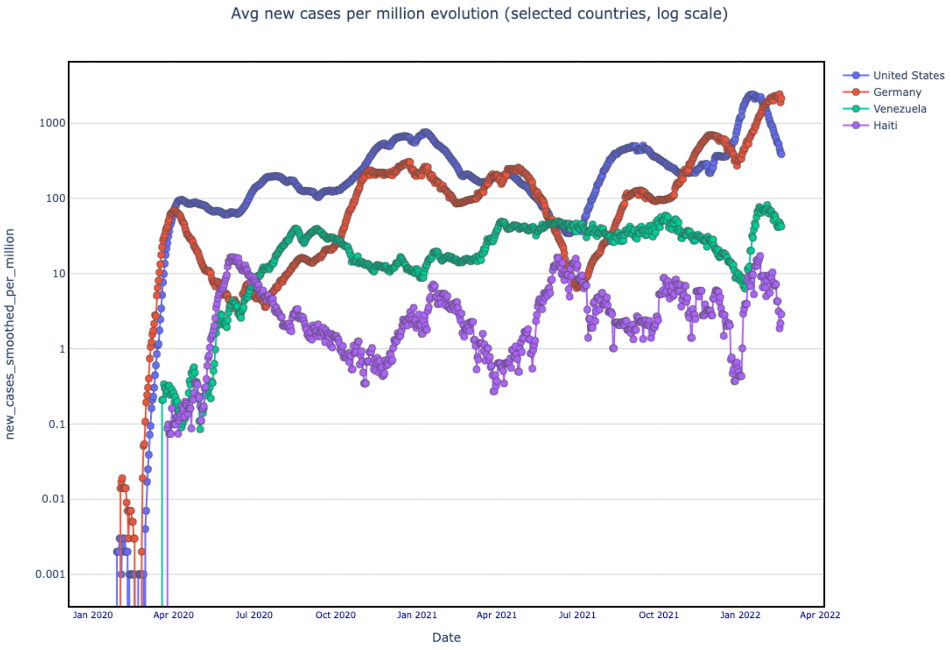


Figure 15: Comparison of the New Cases per Million (averaged over a rolling 7-day window) in the United States, Venezuela, Germany and Haiti, since the pandemic started

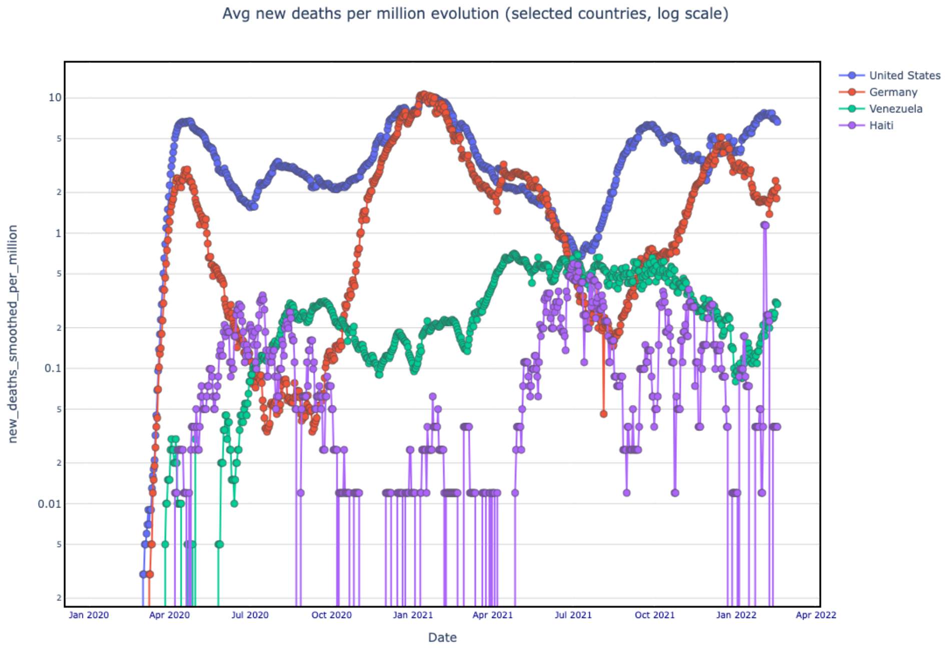


Figure 16: Comparison of the New Deaths per Million (averaged over a rolling 7-day window) in the United States, Venezuela, Germany and Haiti, since the pandemic started

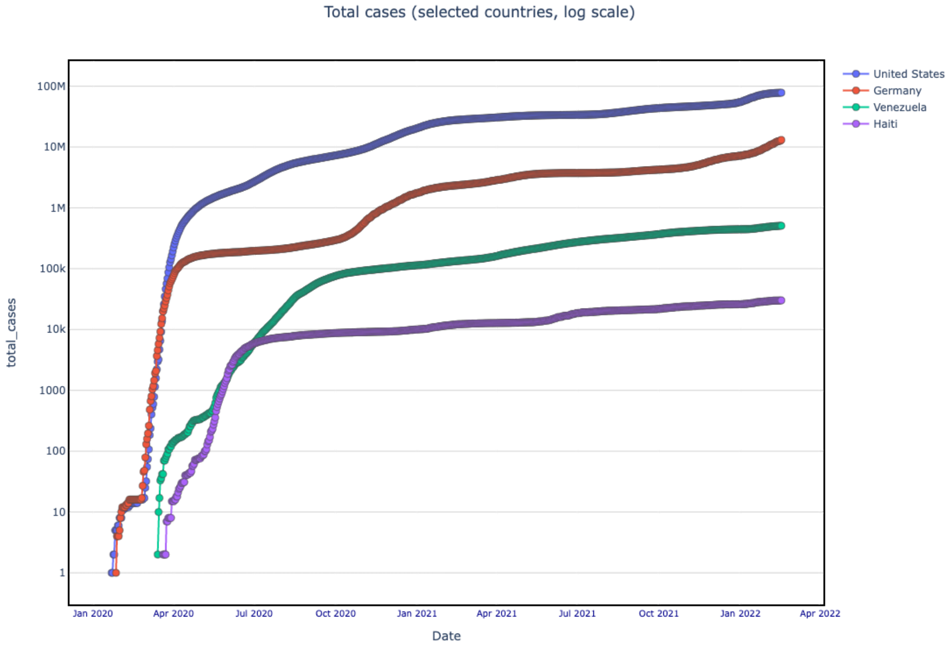


Figure 17: Comparison of the Number of Confirmed Cases in the United States, Venezuela, Germany and Haiti, since the pandemic started

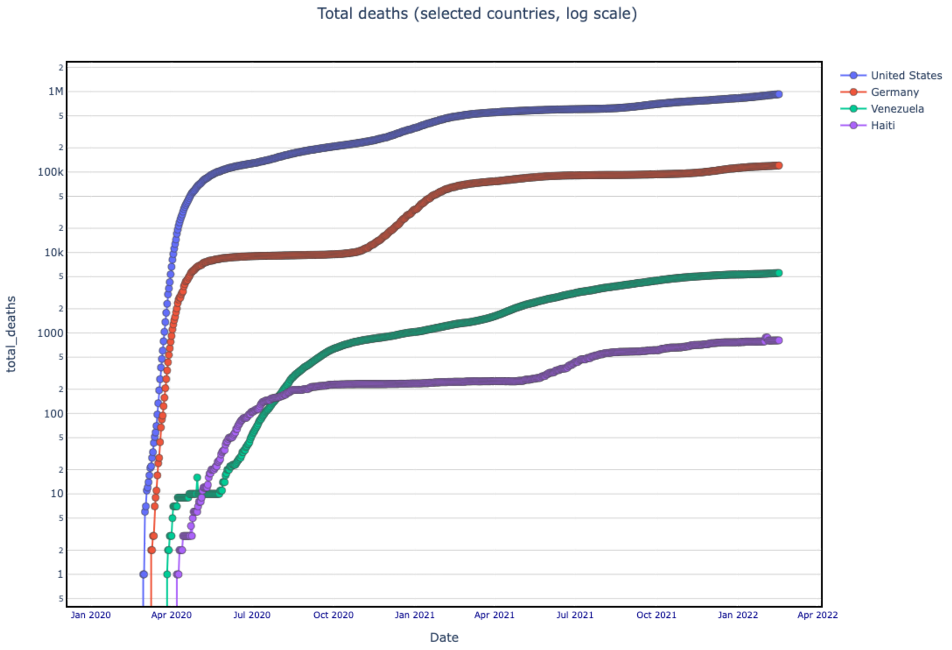


Figure 18: Comparison of the Number of Deaths in the United States, Venezuela, Germany and Haiti, since the pandemic started

As Omicron surged, the number of new cases per million people in each nation peaked in January 2022. However, the number of new deaths in the United States, Venezuela, and Germany are lower than it was before COVID-19 vaccine was introduced. In the month of January 2022, Haiti, one of the countries with the lowest vaccination rates in the world, had the highest number of deaths.

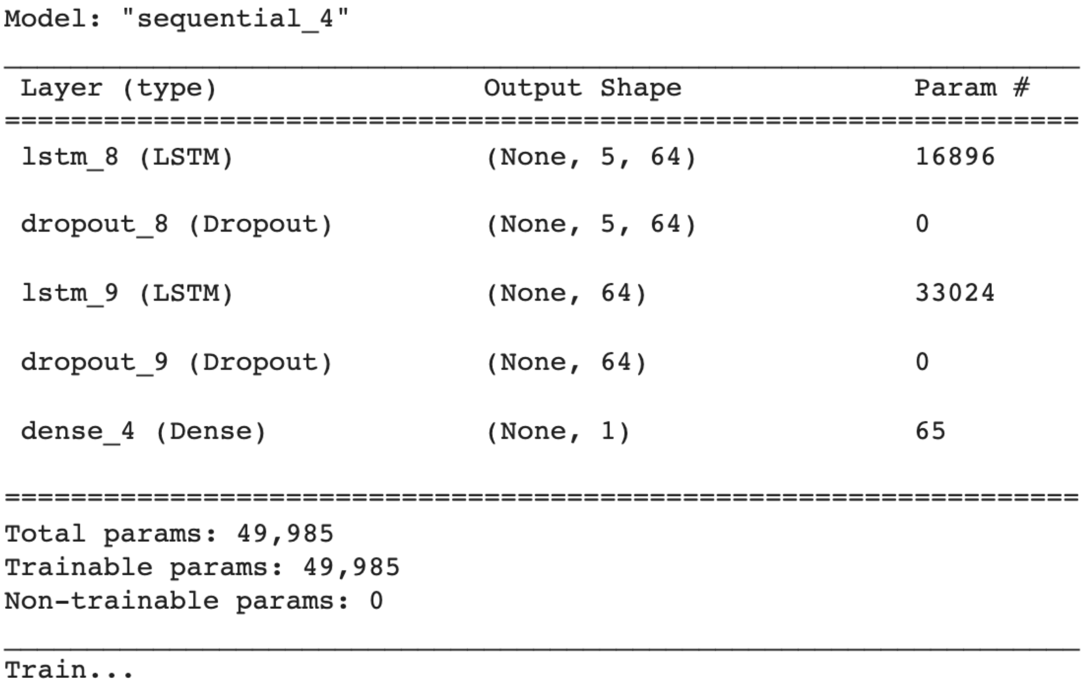
3.3 Data Analysis

After data preparation, the next step is to build a model to forecast total COVID-19 confirmed cases and deaths by using deep learning technique, LSTM, which is a type of recurrent neural network, but the memory performance is better than that of traditional recurrent neural networks.

3.3.1 Model Architecture

This section will highlight the actual model architecture that was utilized in the project to forecast the trends of COVID-19 cases of the 4 countries.

To forecast the COVID-19 numbers, we built a model with the help of LSTM architecture. This model was implemented using Python’s Tensorflow-Keras API. The LSTM network consists of 2 LSTM layers, 2 dropout layers, and 1 dense layer. The first LSTM layer is the input layer with 64 hidden neurons. The activation function is rectified linear (relu), which is widely used. “return_sequences=True” is an important parameter while using multiple LSTM layers as it enables the output of the previous LSTM layer to be used as an input to the next LSTM layer. If it is not set to true, the next LSTM layer will not get the input. The next layer is a dropout layer to improve validation or testing accuracy to prevent overfitting. Next, another LSTM layer with 64 cells and followed by a dense layer as the output layer. Compiling the model using Adam optimizer, which is for handling sparse gradients and noisy problems. A summary of which is shown in figure 19. The model takes in the historical total confirmed COVID-19 cases and trains on individual countries, making case predictions.



```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_8 (LSTM)               (None, 5, 64)             16896

 dropout_8 (Dropout)         (None, 5, 64)             0

 lstm_9 (LSTM)               (None, 64)                33024

 dropout_9 (Dropout)         (None, 64)                0

 dense_4 (Dense)             (None, 1)                 65

=================================================================
Total params: 49,985
Trainable params: 49,985
Non-trainable params: 0
_________________________________________________________________
Train...
```

Figure 19: LSTM Model Summary

3.3.2 Training and Testing

With the common goal of making the most accurate COVID-19 spread predictions within the 4 countries, there were several important aspects to be considered within the training process. Since the number of COVID-19 cases gets rather large over time, the model's calculation during training may be very slow.To fix this, we rescaled the data using sklearn's MinMaxScaler before splitting it into training and testing data. We split the X and Y so that X contains cases for a given number of previous days (time step) and Y has the readings for the next day. This way the model will be trained to predict the number of cases on a certain day based on the trend in the number of cases within the previous number of days. The prediction time frame for a model used 80% of the data as its training set, while the remaining 20% was used as the prediction test set. This results in an approximately 16-day prediction period for the test data.

After splitting the dataset, the LSTM model was then trained using the train and test data was used to validate it. Model.fit() was used for this purpose. An essential aspect of the training process for each model is the hyperparameter tuning. A hyperparameter is any parameter typically in the form of a variable whose value is used in machine learning to influence the learning process. For the LSTM model, the main hyperparameters of focus were the following:

- Time steps = 5;
- Activation functions = relu;
- Optimizer = Adam;
- Learning rate = 0.001 (and reduced by using ReduceLROnPlateau);
- Loss function = MSE;
- Epochs = 100;
- Batch size = 64.

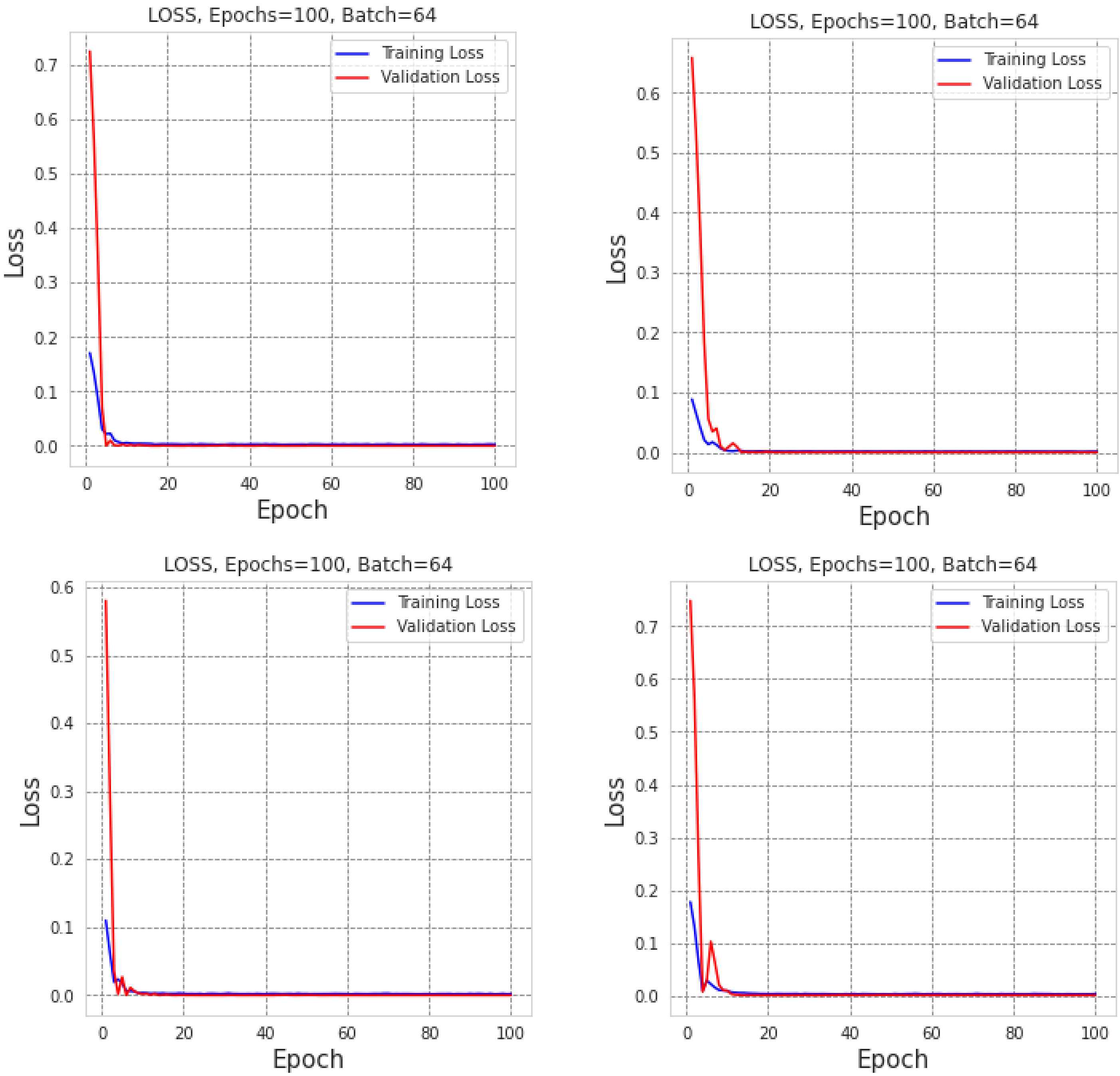


Figure 20: LSTM Model Training Loss Curve of Total Confirmed Cases in Each Country

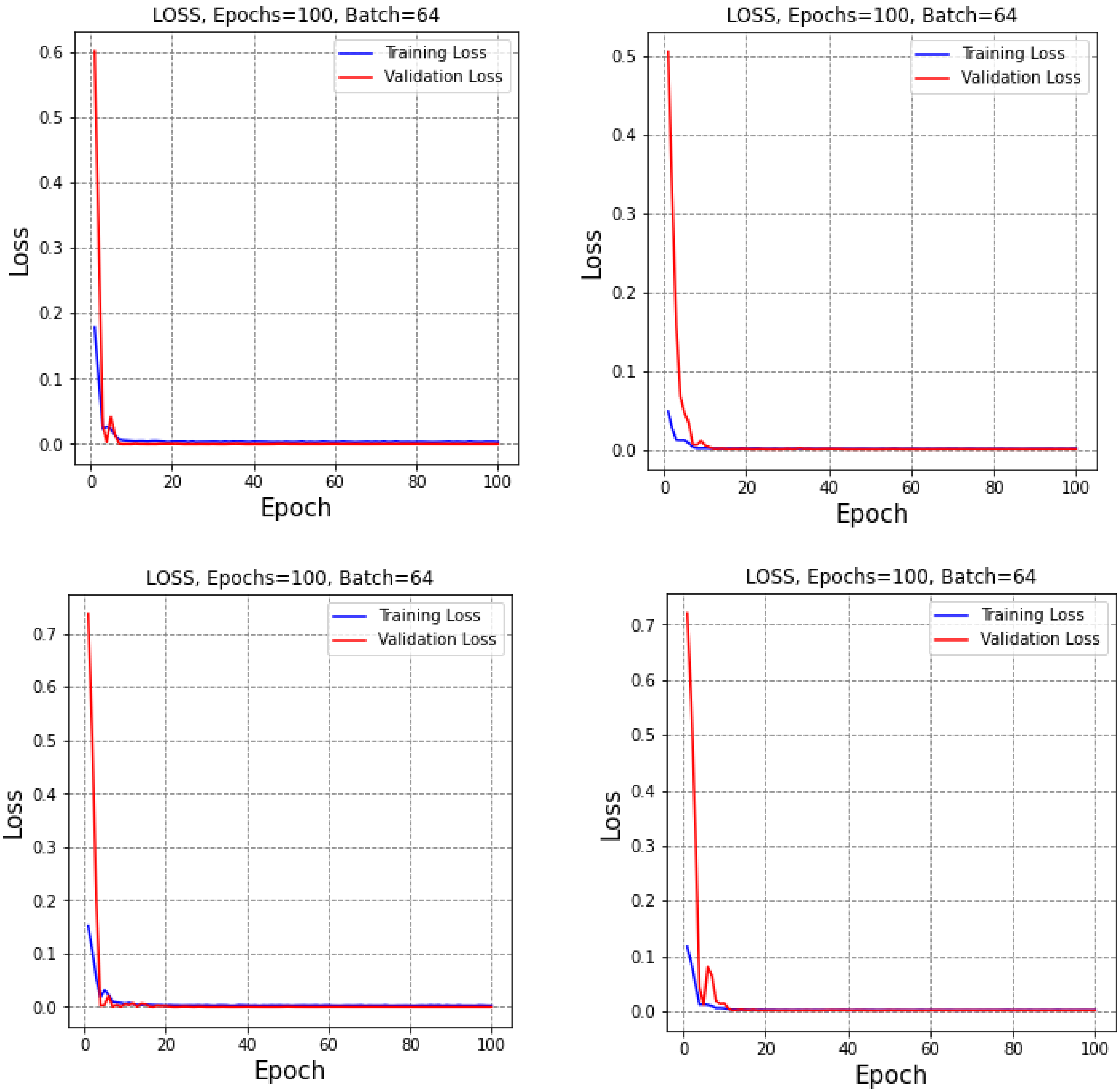


Figure 21: LSTM Model Training Loss Curve of Total Deaths in Each Country

Before getting into the actual prediction, let’s evaluate the model using built-in API, “model.evaluate()” to generate evaluation scores.

The train and test scores of each country from the model evaluation are the following:

- **U.S.**

Total confirmed cases:

```
19/19 [==============================] - 0s 4ms/step - loss: 1.7268e-04
Train score: 0.00017267858493141830
5/5 [==============================] - 0s 5ms/step - loss: 0.0079
Test score: 0.007869580760598183
```

Total Deaths:

```
19/19 [==============================] - 0s 3ms/step - loss: 2.3852e-04
Train score: 0.0002385158877586946
5/5 [==============================] - 0s 4ms/step - loss: 4.2780e-04
Test score: 0.0004277975531294942
```

- **Venezuela**

Total confirmed cases:

```
18/18 [==============================] - 0s 4ms/step - loss: 3.2359e-04
Train score: 0.00032359102624468505
5/5 [==============================] - 0s 4ms/step - loss: 8.3401e-04
Test score: 0.0008340147905983031
```

Total Deaths:

```
18/18 [==============================] - 0s 3ms/step - loss: 2.1986e-04
Train score: 0.00021985553030390292
5/5 [==============================] - 0s 3ms/step - loss: 0.0010
Test score: 0.0010150948073714972
```

- **Germany**

Total confirmed cases:

```
19/19 [==============================] - 0s 4ms/step - loss: 7.1212e-05
Train score: 7.121173985069618e-05
5/5 [==============================] - 0s 5ms/step - loss: 0.2832
Test score: 0.2832081615924835
```

Total Deaths:

```
19/19 [==============================] - 0s 3ms/step - loss: 1.9241e-04
Train score: 0.00019241093832533807
5/5 [==============================] - 0s 6ms/step - loss: 8.9625e-04
Test score: 0.0008962524589151144
```

- **Haiti**

Total confirmed cases:

```
18/18 [==============================] - 0s 4ms/step - loss: 0.0010
Train score: 0.0010093741584569216
5/5 [==============================] - 0s 5ms/step - loss: 0.0062
Test score: 0.006190353538841009
```

Total Deaths:

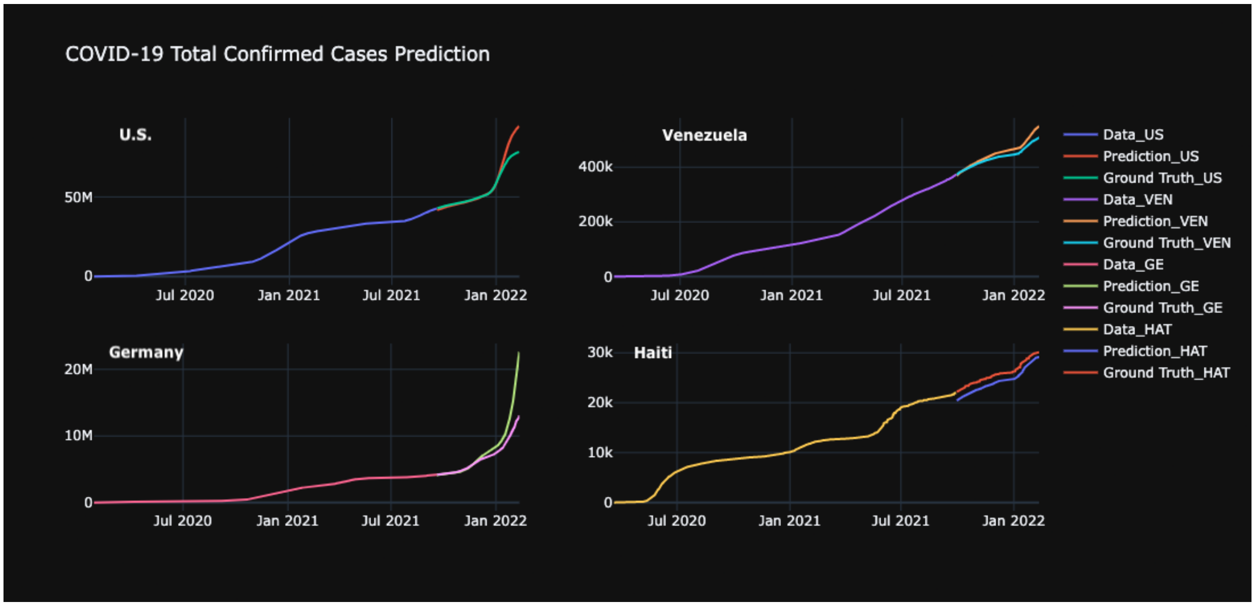
```
18/18 [==============================] - 0s 3ms/step - loss: 2.5719e-04
```

```
Train score: 0.0002571873483248055
5/5 [==============================] - 0s 4ms/step - loss: 0.0211
Test score: 0.021069351583719254
```

3.3.3 Predictions

After building the LSTM model, we will predict values of total confirmed cases and deaths. In order to see the prediction and see the accuracy, first, we will predict the output of the X_test data. This is the output that we get from the test data. To accurately plot the values we need to bring the prediction and y_test data back to the original bounds of the data. Thus, we can plot graphs between the actual cases compared to predicted cases to see the overall accuracy of the model.

By using the COVID-19 confirmed case and death counts from January 2020 to September 2021 as training data, the model predicts the transmission of the disease on the test data from September 2021 to February 2022, shown in Figure 22-23, where the predicted values are represented in a red line on the United States, an orange line on Venezuela, a light green line on Germany and a purple line on Haiti graphs. While the green, blue, pink and red lines in the graphs of the United States , Venezuela, Germany and Haiti, respectively, represent the actual values. The LSTM model's predictions tend to slightly overestimate the actual total confirmed cases of every country, except Haiti where it tends to overestimate the actual values. The predictions on the number of deaths in Germany and Haiti are slightly off; they are underestimated, however, they still follow the trends. These reveal an accurate prediction ability from the model.



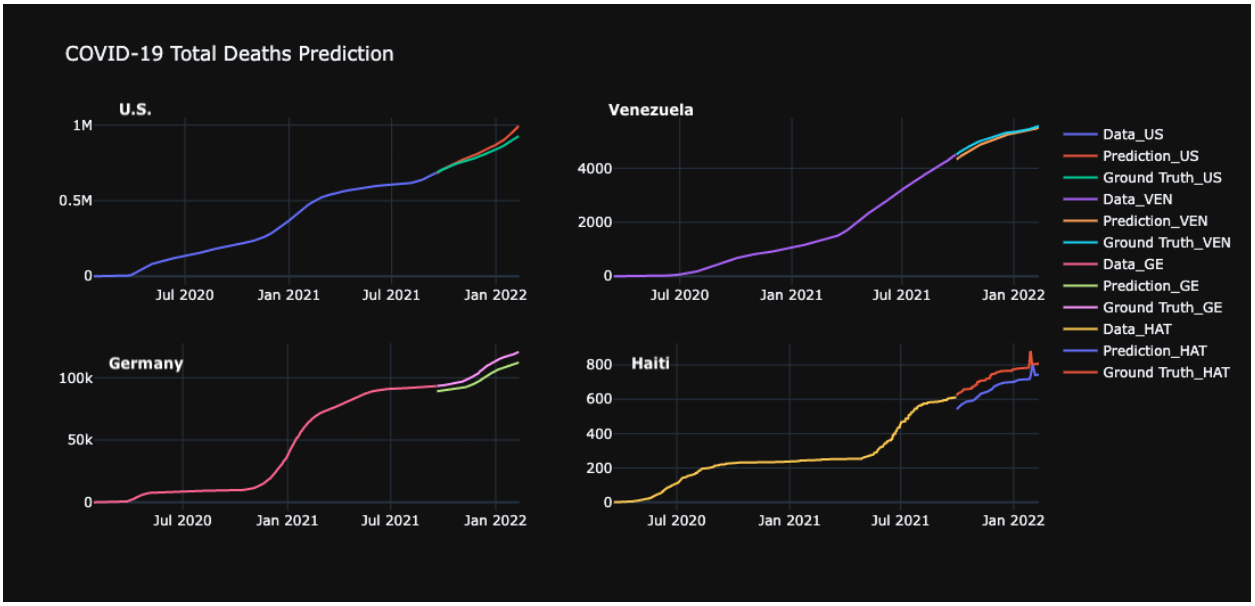
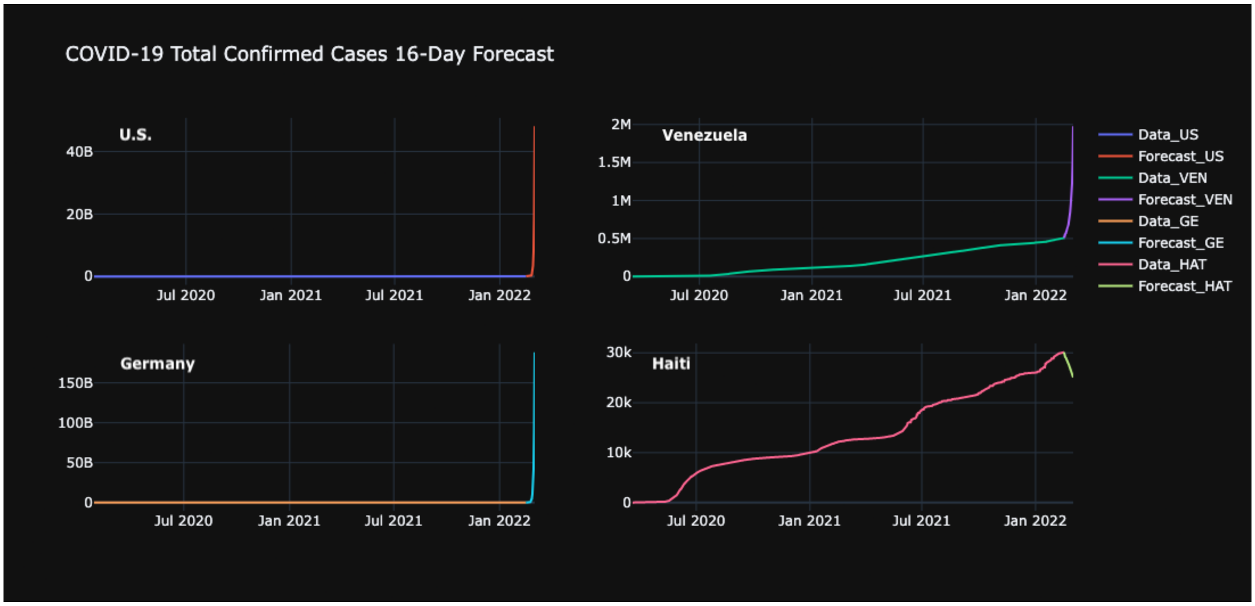


Figure 22-23: LSTM Predictions of COVID-19 Total Confirmed Cases and Deaths over the United States, Venezuela, Germany and Haiti

3.3.4. Results

Using the COVID-19 case counts from Our World Covid Data, we were able to apply the data to our model to make 16-day forecasts on total confirmed cases and deaths.
The model forecasted that the total confirmed cases in every country will drastically increase over the next 16 days (starting from February 16, 2022), except Haiti that will decrease or have an unknown tendency. In the United States, the number of deaths will rise. While Venezuela's trend is uncertain or predicted to remain stable. The number of deaths in Germany and Haiti are forecast to likely decrease.



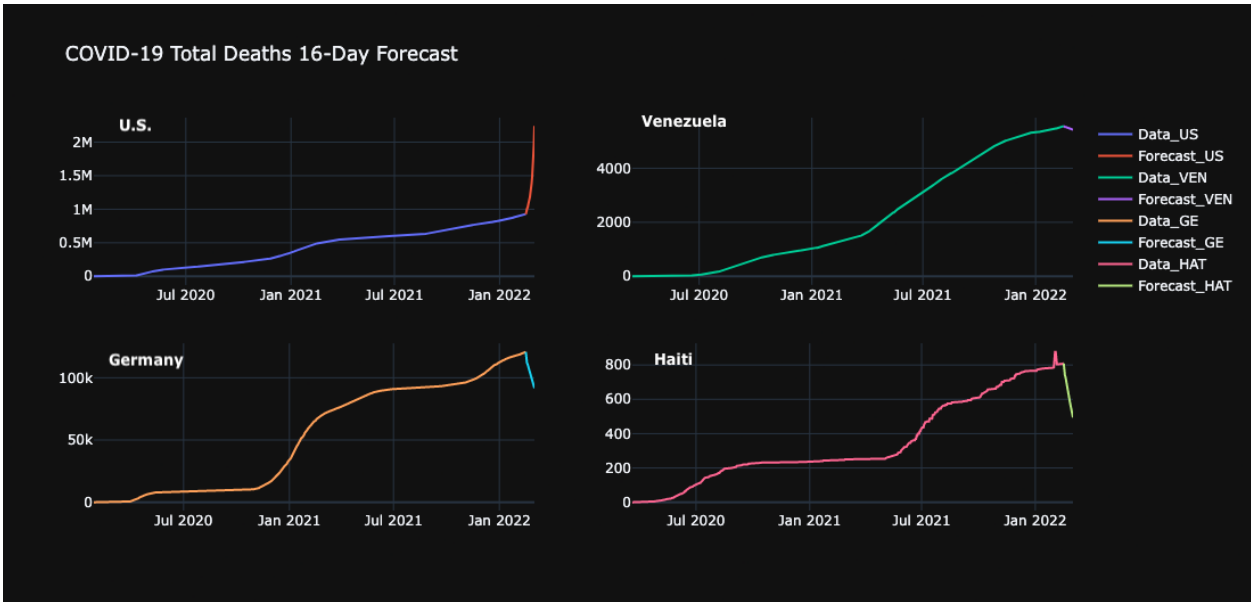


Figure 24-25: LSTM 16-Day Forecasts of COVID-19 Confirmed Cases and Deaths in the United States, Venezuela, Germany and Haiti

4. Conclusion and Discussion

By performing exploratory data analysis (EDA), we can learn more about the datasets while also extracting valuable information from them and determining whether or not there are any flaws. For the predictions, we chose RNN for time-series prediction over other models because LSTMs are better at discovering complicated pattern logics from data by memorizing what's valuable and rejecting what isn't. The LSTM model, which is used for forecasting, featured a very simple neural network structure. Moreover, despite being by far the most advanced model, Machine learning usually benefits from having several features to train on in order to encourage effective learning and develop the most robust models feasible.

The COVID-19 pandemic has posed an unprecedented threat to economics and healthcare systems all across the world. We propose one deep learning model to explore this challenge using COVID-19 case-count data: Long Short-Term Memory (LSTM). This model can forecast the spread of COVID-19 across countries in advance. Though the deep learning model produced unsatisfactory results, we anticipate that with additional improvements to our model, we will be able to construct a better foundation of assessment and preparation for future pandemics.

Bibliography

[1] World Health Organization (WHO), “Coronavirus disease (COVID-19) pandemic,” 2020, https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-COVID-19-19/novel-coronavirus-2019-ncov. View at: Google Scholar

[2] M. Ozaslan, M. Safdar, I. H. Kilic, and R. A. Khailan, “Practical measures to prevent COVID-19: a mini-review,” Journal of Biological Sciences, vol. 20, no. 2, 2020. View at: Publisher Site | Google Scholar

[3] M. U. G. Kraemer, C.-H. Yang, B. Gutierrez et al., “The effect of human mobility and control measures on the COVID-19 epidemic in China,” Science, vol. 368, no. 6490, pp. 493–497, 2020. View at: Publisher Site | Google Scholar

[4] Hochreiter, Sepp and Jurgen Schmidhuber (1997). “Long Short-Term Mem- ory”. In: Neural Computation. URL: https://www.bioinf.jku.at/publications/ older/2604.pdf.