Milestone Report: Analyzing COVID-19 Trends on a Global Scale

Saksham Chawla, Ishaan Lubana, Emilio Rivera
Virginia Tech CS Department, CS 4824
PIDs: schawla2, ishaan, emilio532

Abstract

The goal of this project is to visualize the impact of COVID-19 on nations around the world and successfully identify which countries were the most and least impacted based on the number of confirmed and death cases of COVID-19.

1 Methodology

1.1 Data Source

For our research project, we decided to use data on the Coronavirus made available by the World Health Organization Coronavirus (COVID-19) Dashboard. The WHO began collecting data on December 31st 2019 and the data set was last updated on October 22nd 2021. This data set provides us with the date, country code, country, WHO region, new cases, cumulative cases, new deaths, and cumulative deaths on a country by country basis. There are over 150 thousand data points collected within this data set(WHO).

Not only are we provided with a csv file containing daily COVID-19 statistics, but the WHO also provides us with another csv file containing cumulative statistics. Within the second file, we are provided with categories such as cumulative cases per 100000 population, cumulative cases in total, cumulative deaths in total, cumulative deaths per 100000 population, etc. These categories are aligned with 237 different countries, providing us with the most relevant and latest information amongst these different countries.

1.2 Approach to Data Analysis

So, first we assessed the type and amount of data the data set contained and the kind of the things we could hope to assess from it. We discovered that since the data set had a lot of data samples (approximately 156 thousand data points) and the data was broken down from country to country, we could meet our initial proposal goal of creating a dashboard that broke down the severity and impact of COVID-19 on a country by country basis.

The first thing we did was analyze the type of data we were dealing with for each column. We then organized the data by country and by cumulative cases and deaths.

2 Visualizing Impact on a Global Scale

2.1 Data Source

After getting a good idea of the kind of data we had, we then began to create our country-by-country dashboard. We began by first organizing confirmed and death cases by the number of occurrences

over different time periods (over the last month, 6 months, year, ever). Below is an example of the kind of visual we came up with to show this.

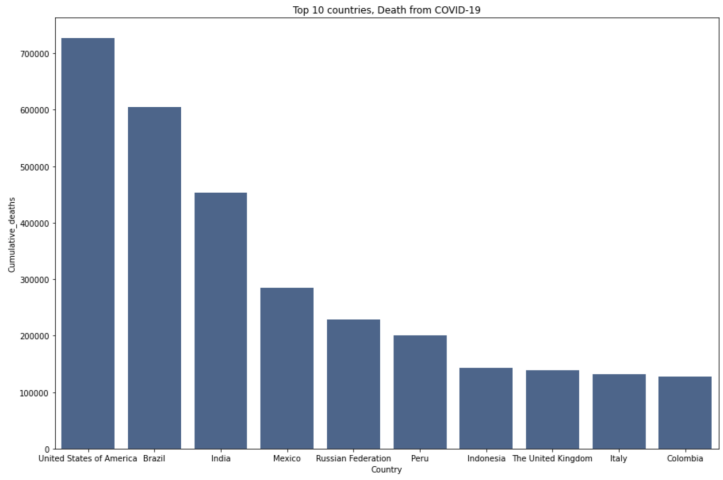

Figure 1: Top 10 Countries with the Most Cumulative Deaths over the last 22 Months

We noticed that the top three most impacted countries (US, India, and Brazil) remained as the most affected countries for most of the study, while the other 7 of the top 10 switched positions (WHO). This suggests that these countries had, in general, a poorer response to the COVID-19 pandemic and should have taken better action to protect their citizens. This shows the top 10 countries with the most confirmed and death cases over the duration of the study (approximately 22 months of data).

3 K-Means Clustering and Regression

3.1 Categorizing Countries for the Most Effective Approach for Medical Assistance

We then wanted to categorize the countries that have been affected by COVID-19 in different groups to get a better understanding of which countries need the most help. For this, we wanted to use a simple but effective unsupervised algorithm: K-means clustering (Sharma). The goal of this algorithm is to group data with some defined similar characteristics together to discover patterns. For this, we look for a fixed number of clusters which are defined as a collection of data points aggregated together because of certain similarities (Garbade). We start with a group of randomly selected centroids and then do repeated calculations to optimize the position of the centroids (Garbade). A centroid is defined as the imaginary or real location representing the center of the cluster (Garbade). This seems to be a clear correlation between the K-means cluster value for degree 3 and degree 4.

Country	Cumulative_cases	Cumulative_deaths	Clusters
Philippines	2740111	41237	0
Ukraine	2725385	63003	0
Malaysia	2413592	28234	0
Netherlands	2064729	18280	0
Iraq	2042117	22875	0
...	...	...	...
Indonesia	4238594	143153	3
Mexico	3767758	285347	3
Poland	2961923	76359	3
South Africa	2918366	88835	3
Peru	2192205	199945	3

Table 1: Breakdown of country data on cumulative cases into 4 Clusters

3.2 K-Means Clustering Implementation

This whole process stops when the centroids have stabilized or the defined number of interactions has been achieved (Sharma). For this part, “pandas”, “numpy”, “matplotlib.pyplot” and “KMeans” from the “sklearn.cluster” package were used. After a group consensus, we decided to divide the countries into 4 different categories to assess which needed the most medical assistance. Our four categories are as followed: "Crisis", "Urgent", "Risky", and "Manageable". We believe that resources towards addressing the COVID-19 pandemic should be prioritized towards countries that have been worst affected by the virus.

3.3 Linear and Polynomial Regression Algorithm Implementation

We implemented various regression algorithms to see if we could find a model that accurately represents the cumulative cases and cumulative deaths for our data set. In addition, we wanted a quantitative measure (as well as a visual one) that could more accurately show how accurate or inaccurate a model is. For this purpose, we also calculated the Root Mean Square Error (RMSE) for each algorithm. The RMSE is an error metric that represents the square root of the result of the Mean Square Error (MSE) function (Moody). For further context, MSE is a function that determines the difference between the actual and model predicted value of a variable, thereby indicating its accuracy (Moody).

We decided to begin with a simple linear regression model using the cumulative cases data over the course of the entire study. Linear regression is a predictive analytic technique that uses historical made to predict a chosen output variable (Hadar). Even before implementation, we were fairly sure that this model would be inaccurate since COVID trends were obviously getting worse and worse over the course of the last year. As you can see below, our initial hypothesis regarding the accuracy of this model was fairly accurate.

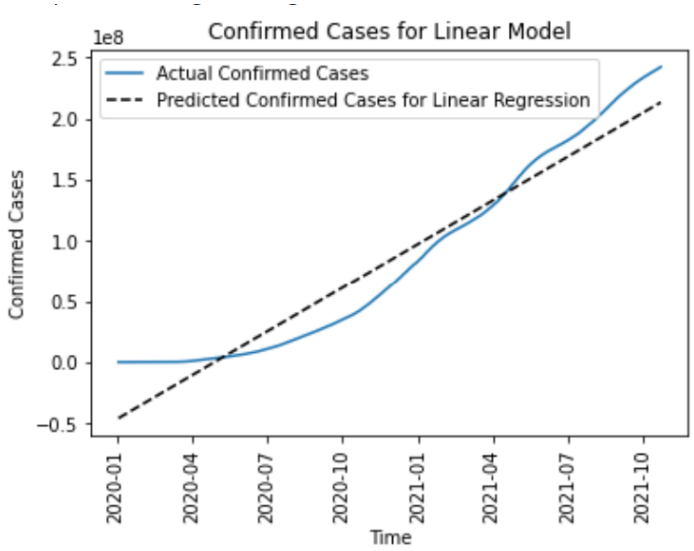

Figure 2: Linear Regression Model on Cumulative Cases Data over 22 Months of Study

We then implemented two separate polynomial regression models to see how much we could improve upon the linear regression model we began with. Polynomial Regression is a form of regression analysis in which the relationship between the independent variables and dependent variables are modeled with regard to a polynomial of a pre-selected degree (Abhigyan). We decided to use 2 and 5 as the assigned degrees of our polynomial regression models. Other than the change in degree, there was no other change in the implementation of these models. Which made the whole implementation and testing process must faster.

As one can, the polynomial regression models are much more accurate on the data than the linear regression model. This makes sense since the COVID-19 pandemic worsened over time. In addition, we can see that the fifth degree polynomial model is much closer to the actual data than the 2nd degree polynomial model. This also makes sense because a polynomial to the fifth degree affords more accuracy than the 2nd degree model, for better or for worse (in fear of over fitting).

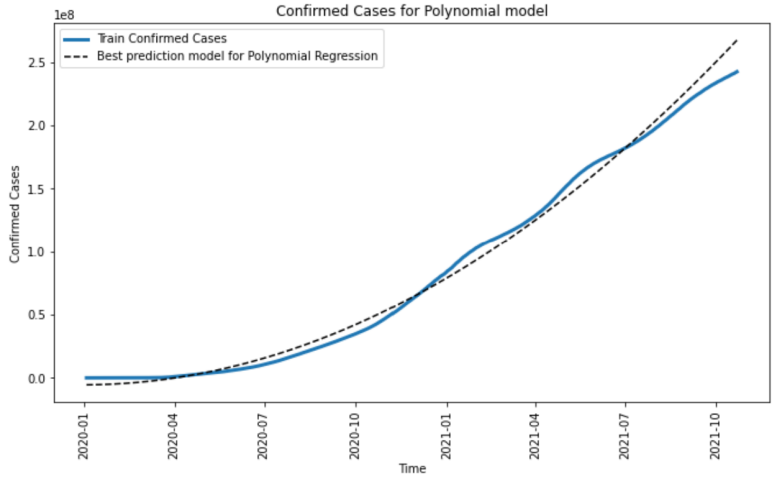

Figure 3: Polynomial Regression Model (Deg=2) on Cumulative Cases Data over 22 Months of Study

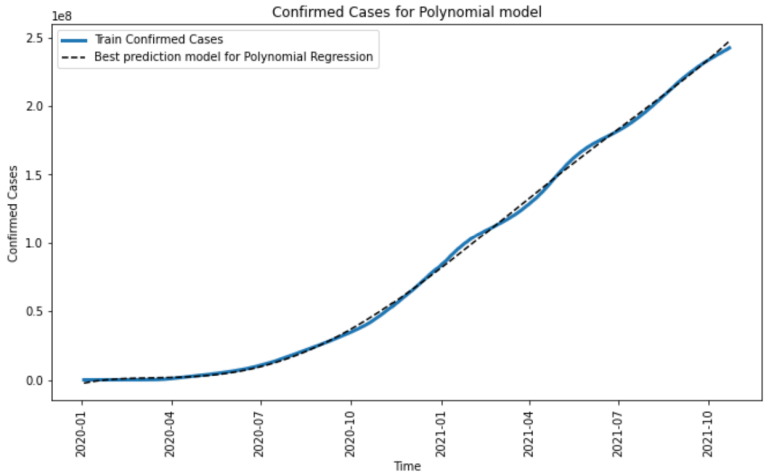

Figure 4: Polynomial Regression Model (Deg=5) on Cumulative Cases Data over 22 Months of Study

In addition, we attempted to implement a logistic regression model to the same data set as well in order to see if it could serve as a better alternative to the polynomial regression models. Logistic regression is a parametric classification model where instead of fitting a straight line to our data (like in linear regression), we fit an S-shaped curve (a sigmoid) to the data set. Over the last week, we've worked on tuning the model but have still struggled to come up with a model that's even close to the accuracy we achieved with the polynomial regression models. Logically, this does make sense since the data won't fit well to an S-shaped curve given that the COVID-19 infection rate was, for about 8 months, an exponential curve.

4 Contributions

Every Wednesday evening, our group would usually meet over Zoom where we would discuss outcomes, data collection, coding, algorithms, and upcoming challenges. Usually, the group would spend time working in a simultaneous manner via the Zoom screen sharing tool so that all members are on the same page. In terms of what we did so far, Ishaan and Saksham were responsible for deriving the idea and our means of research. Ishaan helped research and find an appropriate algorithm that would be useful for our project. Likewise, Saksham was responsible for finding the primary source of data to be collected from.

All three members made contributions to the code we have so far, and all three members made contributions to the Milestone report. So far, the work has been evenly distributed and we have all managed our schedules accordingly so that we can be efficient and on the same page.

Broader Impact

We want our project to help those that don't have much information on COVID to get an idea of its impact on a global scale. Over the last year, we've seen many awareness campaigns about the impact and spread of the Coronavirus and we were interested in creating a simplified version of that. The positive impact of this work will be to inform and educate the general public about the impact of coronavirus in their country as well as other countries around the world.

References

[1] Garbade, Michael J. "Understanding K-Means Clustering in Machine Learning." Towards Data Science, 12 Sept. 2018, https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

[2] Sharma, Pulkit. "The Most Comprehensive Guide to K-Means Clustering You'Ll Ever Need." Analytics Vidhya, 19 Aug. 2019, https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/.

[3] "Sklearn.cluster.kmeans." Scikit-Learn, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[4] World Health Organization, World Health Organization, https://covid19.who.int/info/.

[5] Hadar, Yonatan. "Introduction to Linear Regression in Python - Towards Data Science." TowardsDataScience, 5 Feb. 2019, towardsdatascience.com/introduction-to-linear-regression-in-python-c12a072bedf0.

[6] Moody, James. "What Does RMSE Really Mean? - Towards Data Science." Medium, 28 May 2021, towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e.