Vaxelence: Statistical Analysis of the Global Impact for Individuals Fully Vaccinated Against COVID-19

Nathan Dennis, Tanner Huck, Andy Wen

Problem Statement and Background

COVID-19 was a disease which caused a global pandemic originating in 2019, formally declared a pandemic by the World Health Organization March 11, 2020[1]. This pandemic has caused millions of cases and deaths from the disease worldwide, especially during the beginnings of COVID-19 from 2020 to 2021, where there was no widely available vaccine. Because of this there was much worry about the virus with both its high mortality rate and how quickly it spread, causing a worldwide shutdown[2]. Many businesses were closing or having employees work remotely, schools cancelled in person classes, and mask-mandates became a very common theme for indoor settings with many people[1]. Since this pandemic began however, many vaccines have been created to counteract the spread and morality rate of the disease. These vaccines have caused a subsequent decline in the number of cases and deaths over time. While the disease is still present and people are getting it, the rate of which people obtain the disease is much slowed[1].

For this project, we now want to look into the vaccine itself and the effect it has worldwide. We wanted to find data from the beginning of the pandemic until the present day to see what trends and correlations there are with levels of vaccinations and factors such as deaths and cases from COVID-19, since these are the primary concerns about the disease. We discovered that Our World in Data has COVID-19 data since the beginning of the pandemic with an open dataset on Github which we used[3]. This dataset contains data across every country and every day of the pandemic, monitoring factors such as cases, deaths, and eventually vaccinations when those became available. We were able to find a cleaned dataset in Kaggle[4], which combined the data for every country across all days in the dataset. We used their Python code to alter the Our World in Data github dataset, where each country has their own unique COVID-19 data. There were many other variables in the data, such as life expectancy and average age, that we decided to not use, but could be interesting to inspect in the future.

This problem has affected everyone in the world. The impact the COVID-19 pandemic has had on the world since 2019 has been a major factor in many peoples lives. Whether that be simply getting the disease and becoming sick, or losing others because of the disease, or losing your job because of the shutdowns administered due to the pandemic, everyone has in some way experiences the downsides of this pandemic. This prompted for people to care about when a vaccine or cure would be widely available, where a vaccine was created and quickly administered to counteract it[6]. In the case of "better" solutions, there is no specific better solution for this pandemic. The disease since the vaccines have been administered has slowed down significantly, so there wouldn't be much implications if any other solutions were created other than a potential cure.

For past research, there is no shortage of research on how the vaccines have impacted the pandemic. We are actually using results we discovered in a project last year in STAT 342, where we analyzed the efficacy of the BNT162b2 COVID-19 vaccine. We found a very high efficacy, 90%+ during testing trials, concluding the vaccine was highly effective in preventing the spread of COVID-19[5]. Medical Journals found similar results for other vaccines, with 95% efficacy for Pfizer, 94% for Moderna, and 92% for Sputnik V[6]. We also have found visualizations that can back up our results, which we will state and reference inside the results section, including a map from the New York Times[7] for COVID-19 vaccinations worldwide.

Our overall problem statement is: How has the COVID-19 vaccine impacted the global population?

Questions:

Question 1: How do COVID-19 vaccines, cases, and deaths from COVID-19 vary across the world?

-We plan to make a world map for this question with different levels of color in regions for all 3 variables.

Question 2: Is there a relationship between number of fully vaccinated individuals with deaths from COVID-19?

-After observing the world map, we were interested in if there was a relationship between these variables. Specifically, we want to see if an increase in the proportion of fully vaccinated individuals leads to an increase in the proportion of deaths from COVID-19. We will use data visualization and regression modelling to observe any relationships.

Question 3: What region has the lowest proportion of people fully vaccinated against COVID-19?

-We are interested in this question due to observing in the world map and wondering what continent/region has lower vaccination rates. We say continent/region since "Oceania" is considered a region in our analysis, which isn't a continent.

We decided to investigate these questions, which are trimmed from our initial proposed questions to narrow the questions down as suggested.

Data and Methods

As mentioned previously, we have a dataset for every countries COVID-19 data taken from Our World in Data. The variables we were interested in this data involve vaccinations, deaths, and cases from COVID-19. We are particularly interested in the scaled variables for analysis. The exact variables we are interested in represent deaths per million, cases per million and fully vaccinated individuals per hundred from COVID-19, which will help with analysis so we don't have to deal with differences in population. There are many other characteristics in the data, including many variables we do not plan to use in analysis, such as GDP Per capita and life expectancy. This dataset was cleaned through a Kaggle notebook, where every column from the original Our World in Data dataset was grouped by location and averaged out, as the original dataset had data for almost every day.

It is noteworthy that even though this is a cleaned dataset, there are still missing values in the data. Some countries don't report their data or the data is inaccurate, so in this cases we cannot assess the data for these countries and have to work with the data we have. In regards to data preparation/featurization, we did not create any new variables nor changed any of the variables other than removing N/A values. We took the code from Kaggle to clean the dataset, then when observing the data we concluded it was fit for analysis outside of the missing values. The Python code from the Kaggle notebook was our method of data preparation.

We used data visualization to investigate every research question, such as scatter plots and box plots, with conclusions based off these visualizations. Different conclusions can be made depending on the visualization. Scatter plots are helpful to observe trends between two numeric variables with dots on a graph, which is why we will use this investigate relationships between variables. Box plots are useful to observe the spread of a numeric variable based on different levels of a categorical variables. We are most interested in the median and spreads of the boxplots we create to observe differences in each level of the variable we are using. Also, we plan to use a bar plot to plot the average of the vaccinations across different continents/regions. This is a quick method to observe which countries have the highest, or lowest averages of a given variable. The world map is also another form of data visualization we will use, since it was the obvious choice to observe variation across countries for different variables using a choropleth map with different colors.

To complement these visualizations, we wanted to get some numerical answers rather than results just looking at a graph. To do this, for the question where we want to assess relationships between variables we have used single linear regression model. Single linear regression involves one response and one predictor variable, both continuous, we are interested in seeing how the predictor affects the response and change together,

which can be achieved using single linear regression. We also calculated a correlation coefficient to measure the strength of this relationship.

Tools

We used programming languages to create our data visualizations, prepare the data for analysis, perform statistical testing, and create regression models. We first used Python to clean the data similar to the Kaggle dataset, using the code provided in the notebook. We have used R to create data visualizations, using ggplot2, then create the linear regression models. For the world map visualization, we initially planned to use an Observable Notebook with Vega-Lite to create an interactive world map, but had much trouble doing this. Instead, we reverted to using Python to create this world map.

We used these tools for many reasons, the primary one being due to familarity. We have all used R and Python through our coursework and felt comfortable using those skills on this project. We felt like our knowledge of data visualization packages/libaries in both Python or R would be sufficient to make the data visualizations for this project. In R we used ggplot2 to make each scatter plot, box plot, and bar graph. We also used R for our machine learning model and to calculate the correlation coefficient, as we knew how to perform single linear regression in R using the "lm()" function. All the output is a reflection of what R outputted, other than the world maps which were pasted in from Python.

Another reason we used these tools is due to how widely used they are. Many Data Scientists use R and Python for analysis, which we wanted to gain more experience with for a project. The methods we chose with data visualization and machine learning can easily be done using both R and Python as mentioned above, which is why many Data Scientists and us use these tools.

The ML model worked well and the features we chose to compare, vaccinations per hundred and deaths per million, worked for our model. We had no trouble fitting the model and found a significant result between the two variables, which we expand on in our results. We did not have any key variables being insignificant, maybe because we only investigated the relationship between two variables. Had we included more variables in our ML model, maybe we would find insignificant variables. Either way, deaths per million and vaccinations were significant features and worked well for our model.

Results:

Question 1: How do COVID-19 vaccines, cases, and deaths from COVID-19 vary across the world?

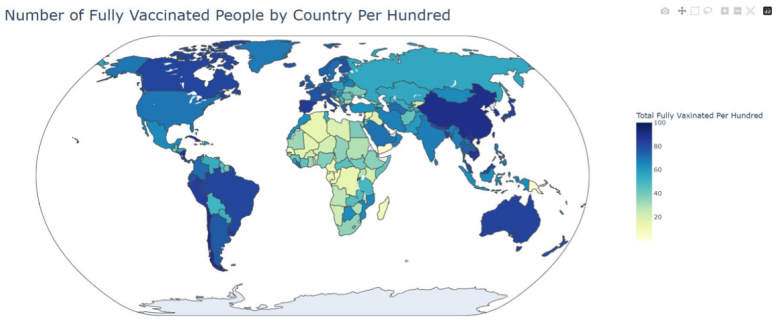


Figure 1: Image for COVID-19 Vaccines Across the World

This world map displays the number of people fully vaccinated by country per hundred across the world. We can see that there are many countries with darker colors, representing very high vaccination levels. This was slightly surprising, but when doing outside research we actually found that many countries have high

vaccination levels that we were unaware of. This was also very similar to a world map created by the New York Times[7] on the same variables, backing up our results and showing high levels of overall vaccinations.

An unfortunate observation is that Africa has significantly lower levels of vaccinations in their population. This may be due to availability of vaccines in Africa, there would be less providers and less opportunities to get vaccinated, backed up by research proving Africa got significantly less vaccines administered[8]. This prompted our final question, where we wanted to investigate if Africa really did have the smallest number of fully vaccinated individuals in their population and how this number varies across the world.

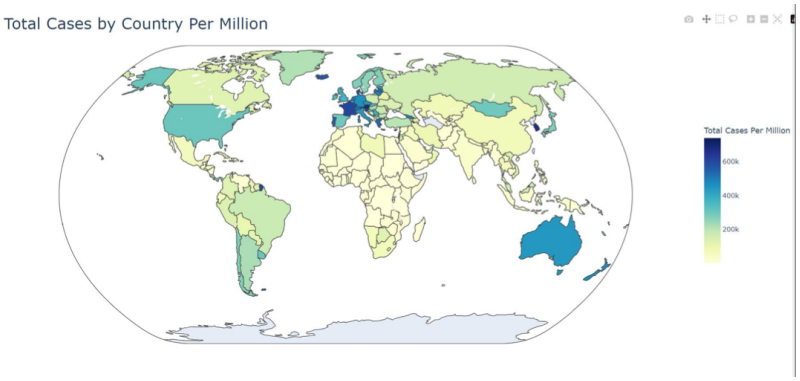


Figure 2: Image for COVID-19 Cases Across the World

This world map displays the number of COVID-19 cases by country per million across the world. It is very interesting to see that many countries actually have a low amount of cases per million, as we expected many more countries to have a higher number of cases. We expected this since we assumed since the beginning of COVID-19, many countries had large outbreaks of the disease. However, we didn't account for this data taking into account the history of COVID-19, until the present day. Since the many outbreaks vaccines have been created to counteract the disease, hence lowering the average cases in many countries.

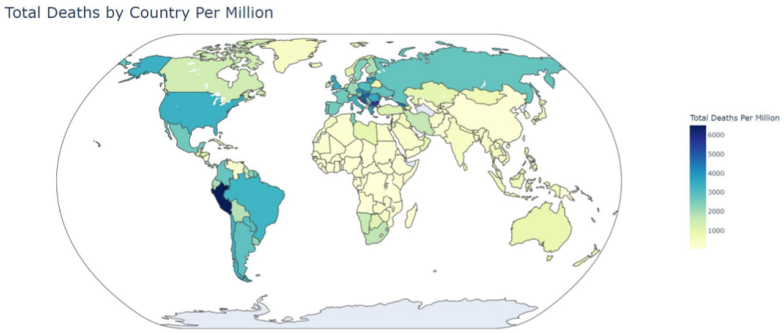


Figure 3: Image for COVID-19 Deaths Across the World

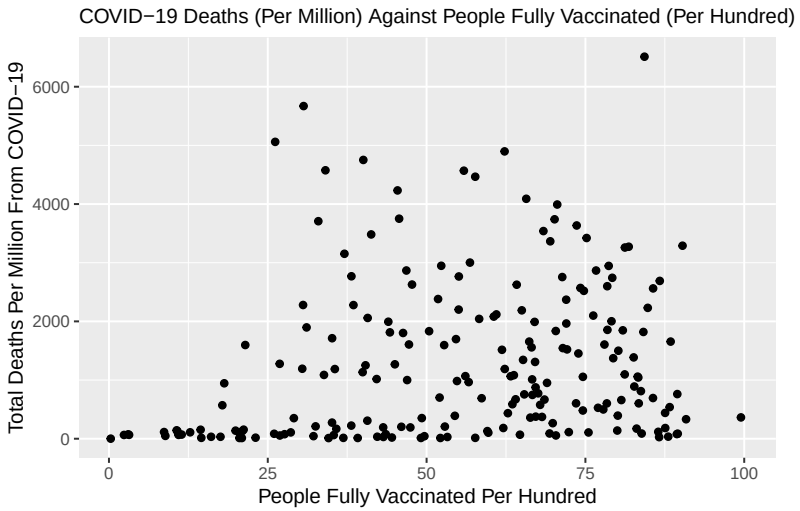
The final world map displays the number of COVID-19 deaths by country per million across the world, which seem to be moderate worldwide. We can see that Europe, North America, and South America have generally high deaths per million. These results were not very surprising, as we expected these regions to be higher in total deaths. Africa, Australia and some parts of Asia have relatively low deaths per million. This is similar to an Our World in Data world map created for deaths per million across the world[9]. We see very similar results, with Africa having the lowest deaths per million based off the map.

Overall, these world maps help us analyze how COVID-19 has varied across the world, allowing us to conduct subsequent analysis and ask more questions based on these visualizations. After looking at these maps we wondered if there was any relationships between number of fully vaccinated individuals with deaths from

COVID-19. If we look at the first 2 maps for vaccines and deaths, we see that there may be a positive correlation between the variables that we wish to investigate. It is noteworthy that there was some missing data for these variables, but they were automatically omitted in Python when creating the maps.

Question 2: Is there a relationship between number of fully vaccinated individuals with deaths from COVID-19?

To begin, we want to observe a relationship between two numeric variables, which means a scatter plot and/or linear model would be useful. First, we created a scatter plot to observe any trends by looking at the points for both variables. In this question we had to remove some data since there were missing values for either people fully vaccinated per hundred or deaths per million.



Looking at the scatter plot, we see no obvious correlation and relationship between the variables. One observation made is that as the number of people fully vaccinated per hundred increases, the total deaths per million also slightly increases, implying a weak, positive linear relationship. The data is very scattered and no definite conclusion can be made looking at this data. It is interesting to note that most of the data is towards the bottom half of the scatter plot, with few points closer towards the top of the plot.

To back up this visualization, the correlation can be calculated between these 2 variables we are interested in. This will help to identify the strength of a linear relationship between two variables. R is used to do this calculation, with this output:

```
## [1] 0.195294
```

The correlation coefficient was calculated to be 0.195294, which indicates there is a weak, positive relationship in the data. This is the same result from the scatter plot, but with a numeric answer. It is now reasonable to wonder the exact relationship between the variables, as the number of people fully vaccinated per hundred increase how much does deaths per million increase? We can use a simple linear regression model to observe the exact increase between these variables. Since the only variable we are interested in is the slope, that is the variable we will look at alongside the corresponding p-value. This output is created from fitting a linear model in R, then extracting the coefficients and p-values associated. The first line is the model intercept and slope, then the second line is the p-values for both

```
##        (Intercept) deaths_per_million
##        51.19511310         0.00339561
```

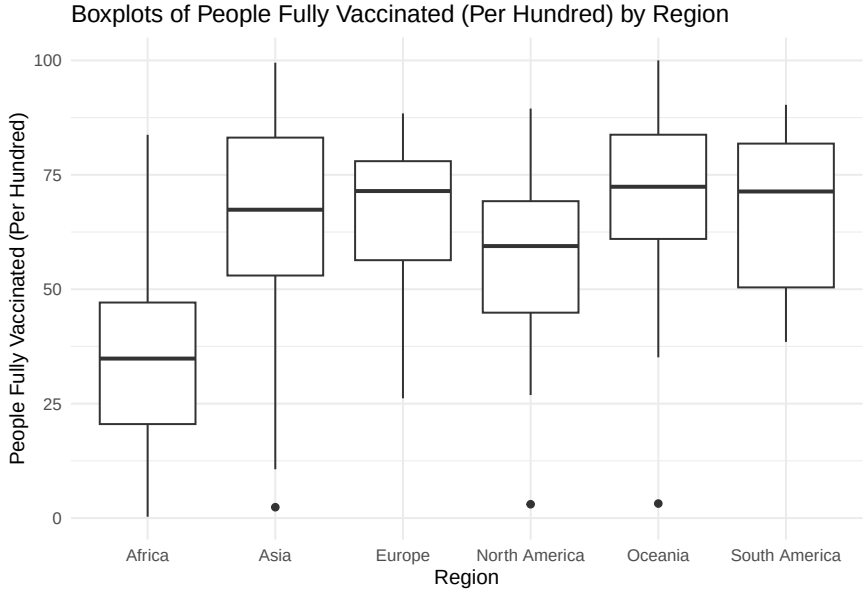
```
##        (Intercept) deaths_per_million
##       4.226662e-58       5.122599e-03
```

Looking at the simple linear regression model output, we can see that the estimate of the slope for deaths_per_million is about 0.003396, with a p-value of about 0.00512. This means that as the number of people fully vaccinated per hundred increases by 1, we would expect the number of deaths per million in a country to increase by 0.003396. With a p-value of 0.00512 in the linear model, we can conclude that this estimate is significant and is non-zero. This backs up the previous results with more numeric answers, as we see the exact relationship between the number of fully vaccinated people per hundred and the number of deaths per million.

With all of these results, it was discovered that there is a relationship between the fully vaccinated individuals and the deaths per million in a country. There is a weak, positive linear relationship between the variables which is backed up with data visualization and numeric results. These results were somewhat surprising, as our group figured that the relationship would be negative, and more fully vaccinated individuals would imply lower deaths. However, we didn't take into account that this is data since the beginning of the pandemic, many deaths were occurring before the vaccines were even administered. These results are more geared towards the relationship since the start of COVID-19, where these results could imply that regions with a higher number of fully vaccinated individuals per hundred had many deaths before vaccines were administered, which is why more people got vaccinated.

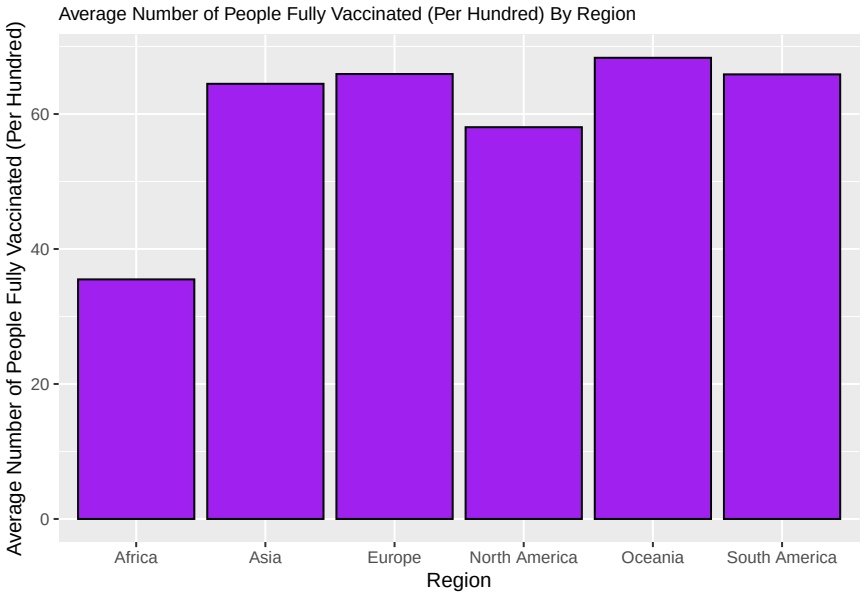
Question 3: What region has the lowest proportion of people fully vaccinated against COVID-19?

After viewing the world map, we were interested in what Regions, which we consider to Continents and Oceania, had a small number of fully vaccinated individuals. We saw some discrepancies in Africa compared to the rest of the world. Now, to observe the spread and median between these variables a boxplot can be created.



Using the boxplot we can observe the spread of the number of people fully vaccinated per hundred in each continent and other statistics such as the median. As we expected from the world map, Africa has by far the lowest median number of people fully vaccinated per hundred, as seen by the median line near the middle of each boxplot. The spread of number of people fully vaccinated per hundred is relatively the same for every continent, there are no major differences in the spread. The medians are also all about the same, other than Africa, suggesting that every continent but Africa has around the same number of people fully vaccinated

per hundred on average. To calculate the exact average, a bar graph graphing the means could answer this question.



Looking at the bar graph, again we can see that Africa has by far the lowest number of people fully vaccinated per hundred on average across the continent. This again backs up the conclusion made from the world map visualization, which is what we expected. Every other region has nearly the same average, with North America being slightly lower but every region having a high average.

Overall, these visualizations answer how the number of fully vaccinated individuals vary by region, scaled to include per hundred to not deal with population issues. Africa had by far the smallest number of people fully vaccinated per hundred, which we expected from looking at the world map visualization. Every other Continent and Oceania was relatively similar, with a high number of people fully vaccinated per hundred on average. This was also expected, as we did not see much difference in the world map for the other regions. We also expected this from past research done as well, where it was discovered that there were low rates of COVID-19 vaccinations across Africa which created a cause for concern[8]. Of the 9.5 billion doses of the vaccine administered at the time of the paper, only 3.4% went to Africa.

TL;DR

This project analyzed how the COVID-19 vaccine has impacted the world, as through global vaccination maps and data visualization we observed high fully vaccinated rates across the world with the exception of lower vaccination rates in Africa. Further analysis of world maps also revealed a relatively low number of cases per million and a moderate number of deaths per million worldwide. Subsequent analysis provided a surprising weak, positive relationship between people fully vaccinated per hundred and deaths per million from COVID-19.

Team Contribution

Percentage Breakdown: Nathan Dennis (33.33%), Tanner Huck (33.33%), Andy Wen (33.33%). All equal.

We gave each other an equal proportion of the total contribution since we felt like everyone contributed equally. To be specific about who worked on what, Tanner was in charge of the first question and created the world maps in Python. He had more familiarity with creating the maps in Python than Nathan or

Andy, so he took charge of that. Nathan and Andy worked on the other 2 questions, Nathan question 1 and Andy question 2, which focused on writing code in R to make the visualizations and numeric results. We all did analysis together after visualizations were created, as we felt it would be more natural to have everyone contribute to the analysis rather than doing it alone. We all also worked on the introduction to our project, such as the problem statement and background, data and methods, and tools. We also all tried to find references since the beginning of the project, which we worked on together. Overall, we all worked on our separate questions to write code but came together for analysis, introduction, and any other aspects in the project.

We are all satisfied with the distribution.

Project Reflections

The project overall went well. We found no issues running our R code or creating the maps in Python, creating the visualizations went very well as we knew from past knowledge how to create them. The analysis went very well as we formulated conclusions as a group based on the results we created. We also slowly worked on the other portions of the project together, finding outside resources that we references to help our overall report. After doing this, creating the overall project report went very well since we planned everything out beforehand.

There were some issues that came up while doing the project. First, we originally planned to make the world maps using an Observable Notebook, which we learned in a class titled CSE 412. We had a lot of trouble making the Vega-Lite code run in Observable, so we decided to switch over to Python. Also, the data was not exactly perfect, which was expected. There were many missing values for several reasons, one of which could be that some Countries don't report all their data. The data also could be inaccurate, as not all COVID-19 data will be accurate, but we could do little to counteract this issue as Countries may report false data. Another issue was the scaling of the data. We understood before doing the project we could not analyze variables such as total_deaths, as the population of a region would affect our results. Countries with a higher population may have more deaths than Countries with a lower population, but the overall deaths if you scale tells a different story. To counter this issue, all our analysis was done using scaled data.

Through this project we also learned many new lessons. First and maybe the most important, we learned how to better work as a team. On this project we worked very well together each step of the way and had a plan to finish the project. We were all on the same page throughout the process which made the project flow much better. We also learned how tough it can be to work with real world data. This data was far from cleaned, we were lucky enough to find a mostly cleaned dataset on Kaggle, or else cleaning the data would have taken a lot of time. We also learned that what we might expect before analysis won't always be true, which is why we had some surprising results. It is important to not assume anything before conducting analysis in many cases, as that may influence how you interpret the results. In general we also made sure to be mindful of ethical components in this project, such as understanding the global disparities in vaccine availability, which is important when conducting data analysis.

We have a github link posted in the bottom of our references for all our code done for this project and the Rmd for this report.

References

[1] CDC Museum Covid-19 Timeline." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 15 Mar. 2023, www.cdc.gov/museum/timeline/covid19.html

[2] Dessie, Z.G., Zewotir, T. Mortality-related risk factors of COVID-19: a systematic review and meta-analysis of 42 studies and 423,117 patients. BMC Infect Dis 21, 855 (2021). https://doi.org/10.1186/s12879-021-06536-3

[3] https://github.com/owid/covid-19-data

[4] https://www.kaggle.com/code/abmsayem/impact-of-covid-19-from-cases-to-vaccines

[5] https://github.com/NathanDennis1/STAT342-Final/blob/main/STAT342_Project.pdf

[6] Chirico, Francesco, et al. "Safety & Effectiveness of COVID-19 Vaccines: A Narrative Review." The Indian Journal of Medical Research, U.S. National Library of Medicine, Jan. 2022, www.ncbi.nlm.nih.gov/pmc/articles/PMC9552389/#:~:text=In%20clinical%20trials%2C%20three%20vaccines,of%20COVID%2D19%20infection10.

[7] The New York Times. (2021). COVID-19 Vaccinations Tracker. The New York Times. https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html

[8] Al-Kassim Hassan M, Adam Bala A, Jatau AI. Low rate of COVID-19 vaccination in Africa: a cause for concern. Ther Adv Vaccines Immunother. 2022 Mar 24;10:25151355221088159. doi: 10.1177/25151355221088159. PMID: 35355936; PMCID: PMC8958672.

[9] https://ourworldindata.org/grapher/total-covid-cases-deaths-per-million?tab=map&time=2023-05-08

Github Report https://github.com/NathanDennis1/SOC225-Project