Miranda Chavez 2266537, Umar Hussain 1627677, Shiv Vyas 2230744

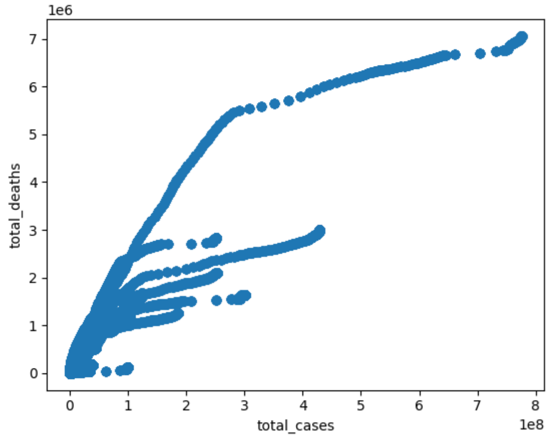
The Pandemic Prevention Analysis – Final Report

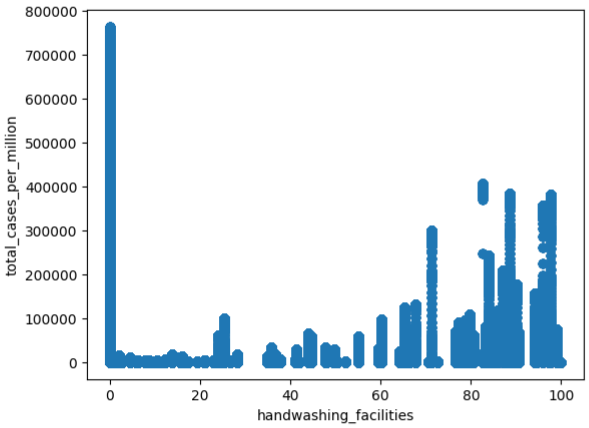
The first task involves preprocessing the ‘owid_covid_data.csv’ data. The first step in preprocessing involved standardizing the data. First, a copy of the data was made, then sorted by date. Since the data is ordered by the ‘iso_code’ feature first and ‘date’ second, the trends in data would follow a seasonal or cyclic pattern, rather than one consistent trend. If a scaler were to be applied to the data before reordering, it could cause some issues with the standardization process. After the data was reordered, a MinMaxScaler was applied to the data. This is since the data is enumerable, there are no negative values. The pattern of the data is also not normal. Using StandardScaler would turn some of the values into negative values and reform the data to a normal shape.

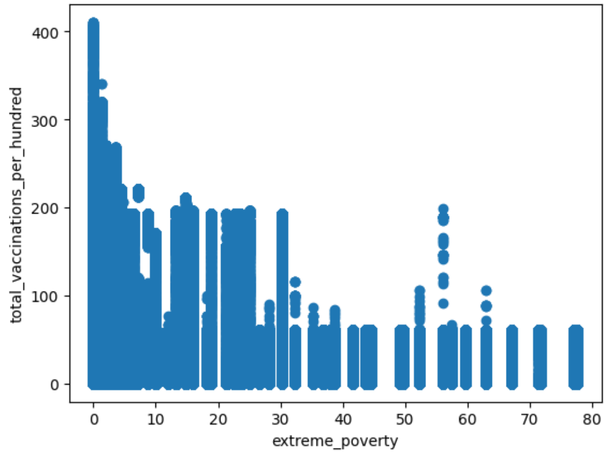
The next part of the first task involved plugging in missing values. First, the standardized data was ordered back to its original order. Then, these missing values were plugged in by iterating through each country. This is since each country may have its own trends and due to the use of the continent sub-data frames and interpolation. If we used these tools over the entire dataframe, trends from one country’s data may bleed into other country’s data. A sub-data frame is made of each country. First, values that may be present in the sub-data frame belonging to a country’s continent is plugged in to the country’s data frame where possible. After, any additional missing values are interpolated to connect any gaps in data. Afterwards, the number 0 is plugged in for any more missing values.

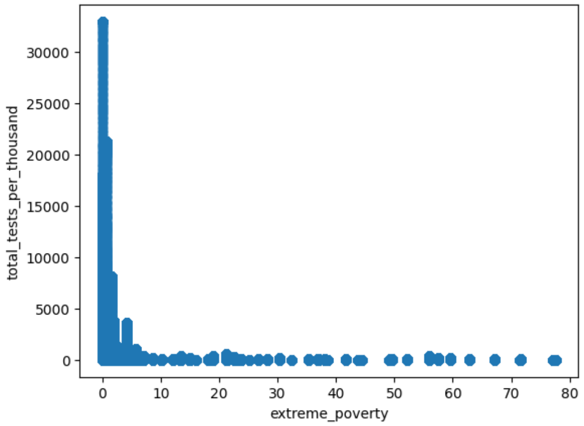
The second task involves finding correlations between the different features given in the ‘owid_covid_data.csv’ dataset. Three correlation tables, as given with the three csvs, were made.

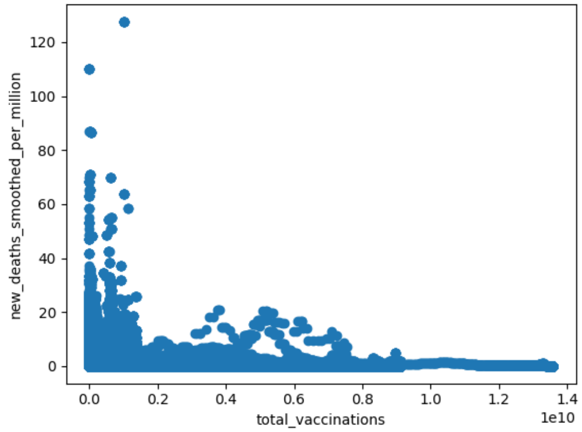
One for the correlation between each and every feature, another for the highest correlation between each intersecting feature, and another for the highest inverse correlation between each intersecting feature. These tables provided some links between different features, but many were quite separated. Some correlations were to be expected. For example, features ‘total_cases’ and ‘total_deaths’ are almost directly related, meaning an increase in one feature would indicate an increase in the other. The features ‘handwashing_facilities’ and ‘total_cases_per_million’ also had a fairly significant inverse relationship with each other, which could indicate the handwashing facilities’ effectiveness in combatting COVID. The features ‘extreme_poverty’ and ‘total_vaccinations_per_hundred,’ and ‘extreme_poverty’ and ‘total_tests_per_thousand’ also had inverse relationships, which could point to higher rates of extreme poverty indicating a lack of access to resources that could’ve prevented the spread of COVID. There was also some inverse relationship between ‘total_vaccinations’ and ‘new_deaths_smoothed_per_million,’ which could indicate the effectiveness of the vaccine, but it is important to note that the correlation value between the two is quite weak.



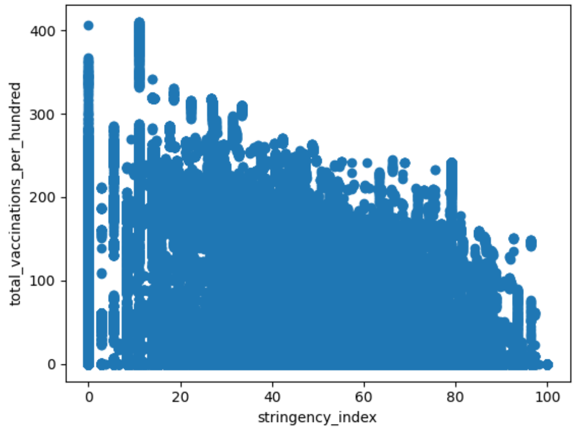


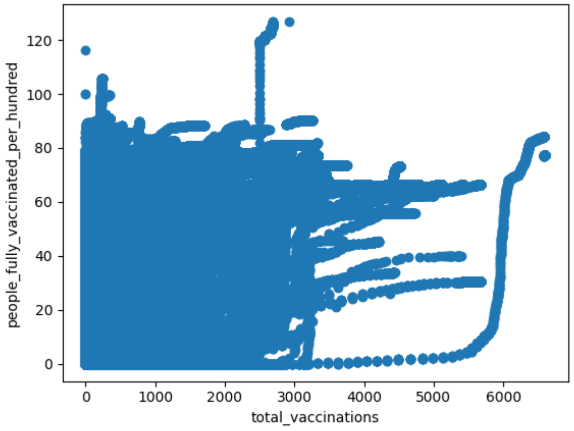


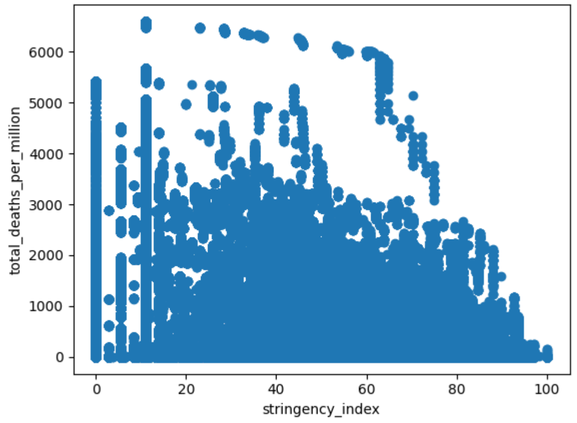




Some were more surprising. ‘stringency_index’ and ‘total_vaccinations_per_hundred’ had a large inverse correlation. On the surface, it wouldn’t make much sense, but considering the definitions of these features, these higher inverse correlations don’t necessarily mean they actually influence each other. The ‘stringency_index’ feature varies over time, meaning some countries have more strict regulations for preventing covid at some times than at others. Both of these features increased in response to the COVID outbreak. This means that the increased COVID cases caused an increase in the ‘stringency_index’ and the ‘total_vaccinations_per_hundred’ feature. The same can be said for the relationship between the ‘stringency_index’ feature and ‘total_deaths_per_million’ feature. Due to the timings of vaccinations and the speed in the change of a country’s response to COVID, there could very well mean that there is an inverse relation between the two features, but not necessarily because one caused the other. The same is also true for certain features, like ‘total_deaths_per_million’ and ‘people_fully_vaccinated_per_hundred’ having some positive correlation.

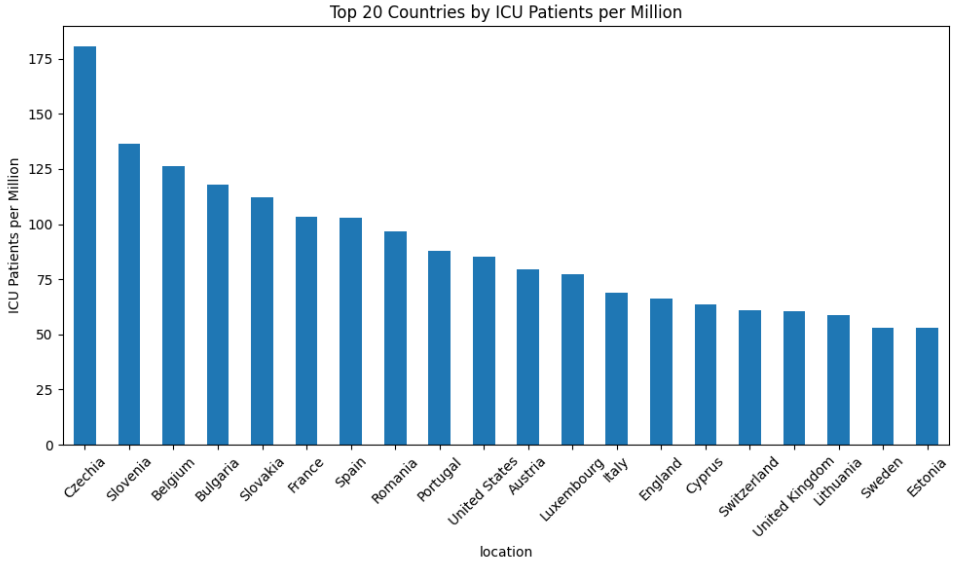






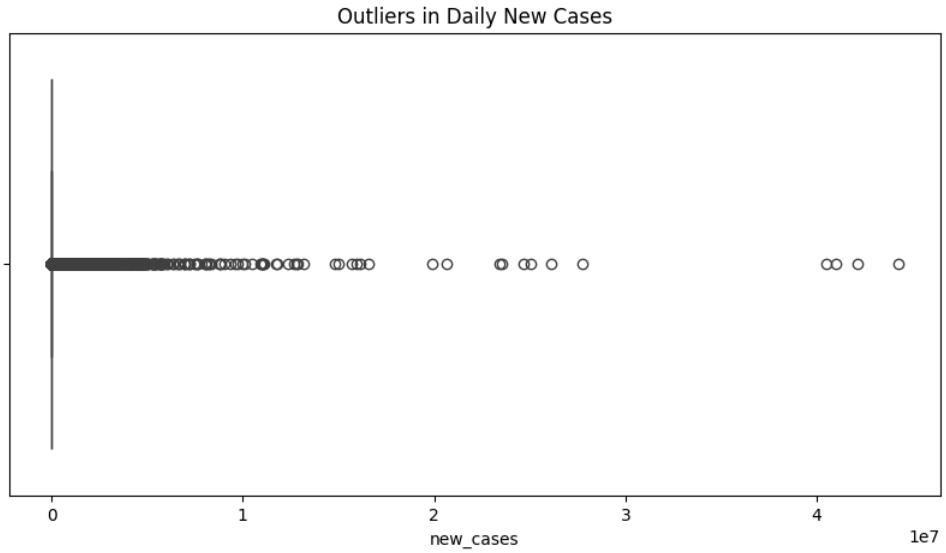
Overall, while there may be some significant correlations between certain features, it isn’t as straightforward as pointing to one feature and claiming it as the sole cause or solution to COVID. Like with the correlations between ‘total_deaths_per_million’ and ‘people_fully_vaccinated_per_hundred,’ there are still some aspects not explicitly stated in the data that must be considered. But even so, it does go to show that certain features, like increased access and use of COVID vaccinations can indeed lead to a decreased rate of COVID cases. Of course, these statements are more nuanced that what’s on the surface of this data.

The next task was to find out the severity of COVID in each country. The key feature was icu_patients, which is the number of COVID patients in intensive care. This was done by first converting the dates to datetime format, then filtering the rows with valid ICU data such as location, total_cases, total_deaths, and population. The next objective was to group by the country and get the max values of ICU and hospital patients, then sorting the ICU patients per million. Finally, a plot was made to find the top twenty countries with the highest ICU use per million. The top five countries were Czechia, Slovenia, Belgium, Bulgaria, and Slovakia. According to these results it seems that COVID was the most severe in central Europe. This could be due to GDP, dense populations, and even the health infrastructures of this part of the world. Analyzing severity through ICU patient data has several key benefits. First, it can help pinpoint periods when health systems were most overwhelmed, which often align with major COVID-19 waves such as Delta or Omicron. These insights are important for evaluating public health policies and hospital preparation.



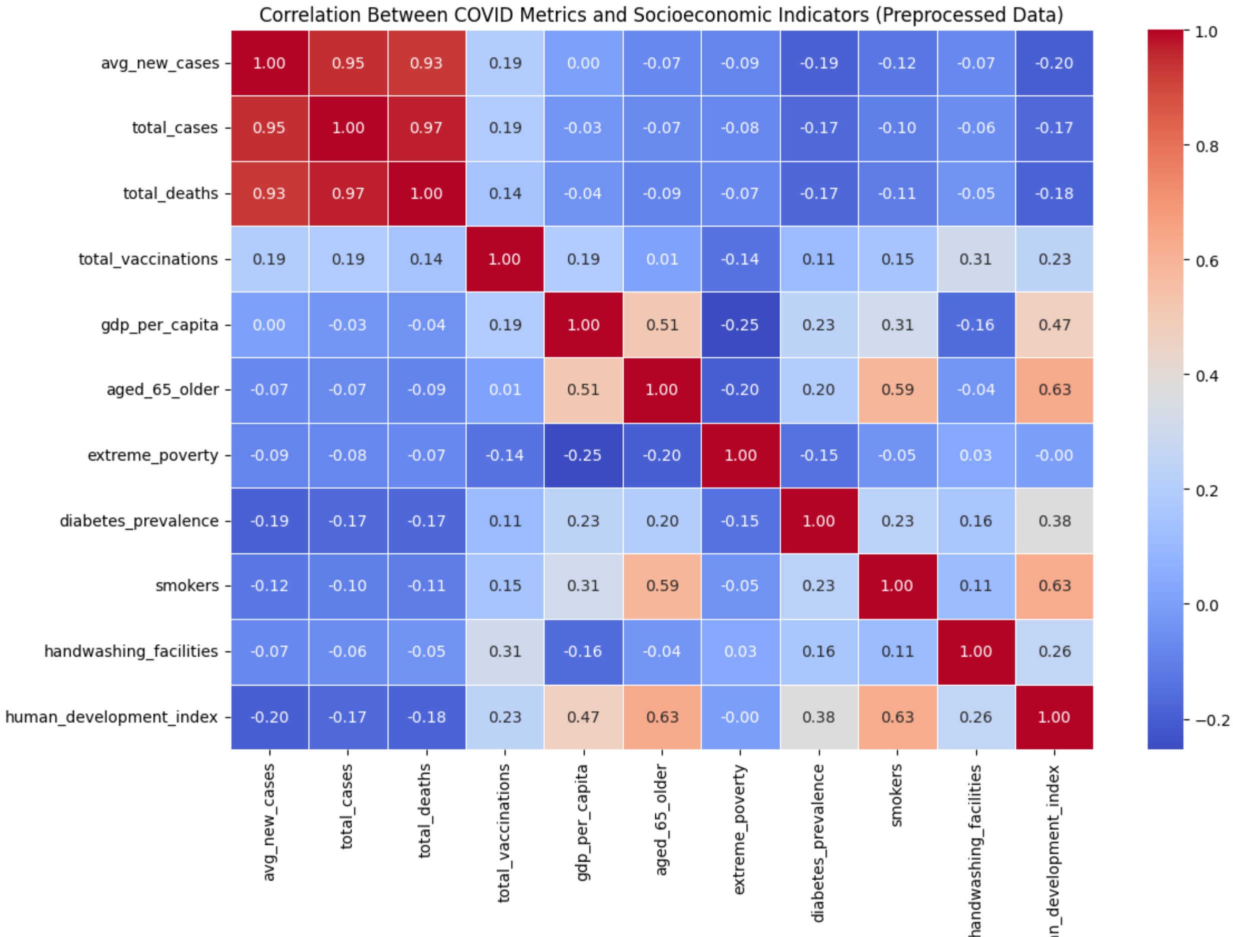
The next task that was done was to find the outliers and interpret what they can represent. This was done by creating a box plot using new_cases to find the outliers in daily new cases.

Outliers can inform global health strategy by revealing where responses were most or least effective. Furthermore, this type of analysis can act as an early warning system for future surges, providing public health officials with the opportunity to take swift action. They can represent healthcare emergencies.



The last task involved searching for relationships between active rising cases of COVID globally and variables that weren't/aren't direct responses/actions taken to COVID's infection, severity, and lethality. For this task, we mainly used 'total_cases', 'new_cases', and 'total deaths' (representing the COVID cases side) to directly compare with 'gdp_per_capita', 'aged_65_older', 'extreme_poverty', 'diabetes_prevalence' (% of population affected with diabetes), 'smokers', 'handwashing_facilities', and 'human_development_index' (representing the indirect affects to COVID spread). Lastly there's another variable within my task that I wanted to include as I wanted to see the direct correlation between the COVID cases side and its direct correlation between indirect affects to COVID spread side. This was the 'total_vaccinations' variable. It is obvious that some variables would not have a clear correlation to COVID cases and related deaths. Therefore, we can get a variable that does have an impact on

COVID cases and related and see whether or not the other variables have an indirect effect on the spread and lethality of the COVID virus. With the variables defined we can now discuss the actual data and relationships. We use coefficients of each variable in order to identify the relationship between each variable. As is the case, 1 represents a positive relationship between 2 variables, -1 represents a negative relationship between 2 variables, and 0 represents no clear relationship between the variables. One thing to note is the number of variables that can affect a single variable an analyst may focus on. Since the nature of COVID-19 and the ensuing pandemic was global affecting every facet of life, many variables have many different effects on other variables that may be in focus. This is important to note because a greater number of variables that affect a single focus variable will decrease the strength of the relationship between the focus variable and any other variable the analyst would want to compare too. Therefore, it shouldn't be too shocking that many of these correlation values aren't strong considering the scale of the pandemic. However, with all that said here's the graph.



We use a heat map as it’s the best way to identify the relationship between two variables using correlation coefficients. A colder graph represents a negative relationship, and a warmer graph represents a positive relationship. It’s hard to identify the relationships through only color therefore we have the values as well between each variable. I'll only be talking about the correlation between variables that seem relevant. For example, let’s start off with ‘age_65_older’ and ‘total_deaths’. It’s fairly obvious that an increase in total deaths (related to the COVID virus) would mean a decrease in the population of this aged 65 and older. Older people in general were greatly affected by the virus and its lethality. Starting off though let’s look at the GDP and the human development index. We can see that there is almost no correlation between the GDP and cases/deaths which would be surprising, however strictly economy would have no effect on the spread of the virus. It’s how the government the money within their economy that would affect the virus, which we see in the human development index. As the index percentage increases, the rate of cases and deaths decreases. This would make sense as a higher index represents the overall wellness and wellbeing of a population within the country, which takes into account health and safety. As such, the human development index has a great effect on the total cases and deaths, therefore this is one aspect that governments should look at to improve to prevent spread. We mentioned GDP before as not being correlated to total cases and deaths. However, GDP does have some correlation to total vaccinations. If we look at the total vaccinations, we can see that it has some correlation to cases and deaths as well. GDP having a positive correlation to vaccines makes sense as a higher GDP relates to a higher production of medication and transportation. However, the positive correlation between vaccines and cases is confusing. One thing to note is the way you can look at the data. Just because vaccines and cases are positively correlated doesn't mean that vaccines cause cases to rise. In fact, the increased number of cases may mean that there would be an increased number of vaccines shipped or produced within that country. As such, we see the indirect correlation that GDP has on cases and deaths. These are just a few examples that I have shown and there are a decent more that are relevant to the topic. So, let's move on to what we can learn from this. We can start by looking at the ways in which we ship vaccines and medication to different areas as we can see from the map that variables such as 'extreme poverty' and 'gdp_per_capita' are affected by or affect the 'total_vaccinations' that occur within a country. We should also look into those that are of old age or have health problems from outside sources such as 'age_65_older', 'diabetes_prevalence', and 'smokers' as they are somewhat affected by the virus for total cases and deaths. There are other solutions that can be mentioned such as the effects that GDP can have on facilities, which can reduce spread but for the most part immediate action such as vaccine production and transport. As for prevention, we should focus on reducing disease and bad habits that may impact the spread of future pandemics and focus on the health and safety of the citizens within the country (human development index).

Overall, with the sheer amount of data and different variables affecting each other, it can be difficult to point to certain variables as definitive sources or solutions to combating the spread of COVID. However, there are some relationships of note that can point to more effective strategies world leaders could use in the event of another pandemic. The number of vaccinations administered, for example, shows an inverse relationship with the number of COVID cases and deaths. This could point to the effectiveness of vaccines. Unfortunately, with the 'stringency_index' variable, it was difficult to fully determine if having stricter rules about preventing COVID would help combat the virus. Since the stringency index rises in response to increasing cases of COVID, it is easy to draw inverse relationships between the index and number of deaths. However, is still some nuance to the relationship, due to the nature of the stringency index. Despite knowing this, however, it is still difficult to determine the effectiveness of a higher stringency index either way. We should also take prevention and reactive steps in order to make sure the spread of any future disease is limited. One step to take is to focus on the general safety, health, and wellbeing of the overall population within the country. There is a chance that in the future we will have to deal with another pandemic and with that in mind, we need to analyze what happened in 2019 and 2020 in order to take better measures in ensuring that we protect the populous and those around us.