An Analysis of the Worldwide Progress of COVID-19 Vaccinations using Apache Spark

Charlotte Dalenbrook

February 21, 2021

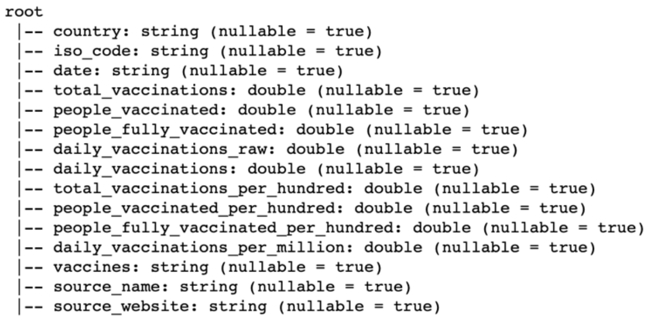
Spark, Term 2, MBD

IE School of Human Sciences and Technology

Background/Scenario Description

The COVID-19 pandemic has had unprecedented consequences on all aspects of human life for over a year. From losses in livelihoods to closures of entire economies, the pandemic will have long lasting effects that exceed the enormous loss of lives alone. One reason for hope in recent months has been the fast and effective development of vaccines by several different entities around the world. These vaccines, if applied on enough people, could create herd immunity among the population, and eradicate the disease. Herd immunity usually occurs when between 70-90% of the population is immune. Although vaccines were developed at an unparalleled speed, the implementation of vaccination schemes and delivery of vaccinations has been anything but ideal. In many countries the organization of vaccinations has been very slow, and supply issues have led to further delays. Gaining insights into the vaccination progress around the world could help quell doubts about when heard immunity will achieved.

The dataset chosen for this report consists of data about the worldwide vaccination progress per country per day, since vaccinations have started. The columns in the dataset (all columns shown in Figure 1) consist of a country and date, as well as information relating to how many vaccinations have been administered, and finally which brand/type of vaccination was been used. Although this dataset is currently still quite small, it will expand greatly over the coming months and years and setting up an analysis with Pyspark like the one described throughout this report, ensures that, even as the size significantly increases in the future, the analysis of vaccination progress can be done in an organized and time efficient manner. Later on in the analysis, the data was also enriched with a dataset about the population by country in 2020, in order to gain more insights into the number of vaccinations performed per total population in a country. The datasets can be found at the following links:



```
root
 |-- country: string (nullable = true)
 |-- iso_code: string (nullable = true)
 |-- date: string (nullable = true)
 |-- total_vaccinations: double (nullable = true)
 |-- people_vaccinated: double (nullable = true)
 |-- people_fully_vaccinated: double (nullable = true)
 |-- daily_vaccinations_raw: double (nullable = true)
 |-- daily_vaccinations: double (nullable = true)
 |-- total_vaccinations_per_hundred: double (nullable = true)
 |-- people_vaccinated_per_hundred: double (nullable = true)
 |-- people_fully_vaccinated_per_hundred: double (nullable = true)
 |-- daily_vaccinations_per_million: double (nullable = true)
 |-- vaccines: string (nullable = true)
 |-- source_name: string (nullable = true)
 |-- source_website: string (nullable = true)
```

Figure 1 Global Vaccination Progress Data Schema

Dataset	Link
COVID-19 Vaccination Progress	https://www.kaggle.com/gpreda/covid-world-vaccination-progress
Population by Country 2020	https://www.kaggle.com/tanuprabhu/population-by-country-2020

Goal of the Analysis

By analyzing the dataset described above, the following business questions surrounding the global vaccination progress can be answered:

1. In which countries has the vaccination program the most effective so far? And which countries are performing worst in vaccinating their population?
2. What is the per country and worldwide progress of the vaccinations?
3. Which vaccine type has been used the most?

4. At the current rate when will 70% of the population be immune? I.e., when will herd immunity be achieved?

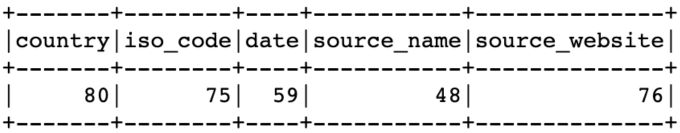
In order to properly answer these questions, the dataset must first be explored and cleaned. As mentioned previously, the dataset consists of several columns of information on the vaccination progress per country, per date since the beginning of vaccinations. The dataset's columns can be split into 3 main categories as follows:

- **Information related columns:**
    - *Country:* the country for which the information of the row is
    - *iso_code:* the iso code of the country
    - *date:* the date of the information in the row
    - *source_name:* the name of the source of the data in the row
    - *source_website:* the URL of the source of the data in the row
- **Vaccination Count related columns:**
    - *total_vaccinations:* the total number of vaccinations that have been administered in the country
    - *people_vaccinated:* the total number of people that have received 1 or more vaccines
    - *people_fully_vaccinated:* the total number of people fully vaccinated
    - *daily_vaccinations_raw:* the total vaccinations that day
    - *daily_vaccinations:* the total vaccinations that day (unclear)
    - *total_vaccinations_per_hundred:* the vaccinations administered per 100 people
    - *people_vaccinated_per_hundred:* the people vaccinated per 100 people
    - *people_fully_vaccinated_per_hundred:* the fully vaccinated people per 100 people
    - *daily_vaccinations_per_million:* the daily vaccinations per million people on the given day
- **Type related columns:**
    - *vaccines:* the brand(s)/type(s) of vaccinations being administered in the country

Exploration/Cleaning of Information Related Columns

It is important for the information related columns not to contain nulls, since the date and country column act as a unique identifier for each piece of information in the dataset. Thankfully, out of the information related columns, only the ISO code value contains nulls. Since the country column does not contain nulls, this is not an issue, and therefore the ISO column doesn't need to have any values imputed.

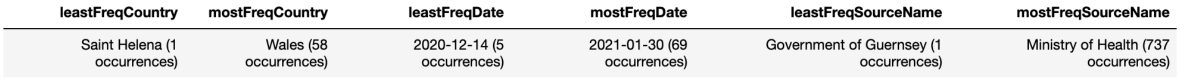
By finding the maximum and minimum of the data column we can see the date range for which the data is valid. The earliest record in the dataset was on the 13.12.2020, while the latest record was on the 09.02.2021, meaning the dataset spans a range of 59 days. This number of days is also reflected when exploring the number of unique values in each of the column, for which the date value is 59, as shown in figure 2.



```
+-------+--------+----+-----------+--------------+
|country|iso_code|date|source_name|source_website|
+-------+--------+----+-----------+--------------+
|     80|      75|  59|         48|            76|
+-------+--------+----+-----------+--------------+
```

Figure 2 Number of Unique Values per Information Related Column in the Dataset

Another important thing to note about the data is that there are only 80 unique countries, meaning not all 195 countries are represented in the dataset. This is most likely due to the fact that not all countries have begun vaccinating at this point in time. As for the frequency of occurrences of each of the columns, the country, date and source name were the only ones it made sense to analyze. We can see from figure 3 that some countries have information on almost all dates in the range, while others only have one set of information on one day. Finally, on some dates (30.01.2020), many countries have information about vaccinations taking place that date, while the 14.12.2020 only has 5 corresponding rows of information in the dataset.

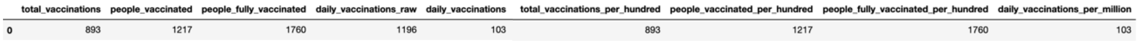


leastFreqCountry	mostFreqCountry	leastFreqDate	mostFreqDate	leastFreqSourceName	mostFreqSourceName
Saint Helena (1 occurrences)	Wales (58 occurrences)	2020-12-14 (5 occurrences)	2021-01-30 (69 occurrences)	Government of Guernsey (1 occurrences)	Ministry of Health (737 occurrences)

Figure 3 Most and Least Frequent of Occurrences within the Dataset Information Columns

Exploration/Cleaning of Vaccination Count Related Columns

The dataset has many different columns concerning the number of vaccinations administered on each date. For these numerical columns it makes sense to create a summary of all of the values in the columns. From this summary we can already gain insights such as the maximum number of vaccinations that have been administered in any country (which is 43,206,190), as well as the counts within each column, which varies significantly. This variation in the counts hints to the fact that there are a lot of nulls in the data, which is confirmed when checking the null counts of each column (figure 4)



	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_hundred	people_vaccinated_per_hundred	people_fully_vaccinated_per_hundred	daily_vaccinations_per_million
0	893	1217	1760	1196	103	893	1217	1760	103

Figure 4 Null Counts of Vaccination Related Columns

For the purpose of this analysis and the business questions to answer, the most important columns are the total vaccination column, the number of people fully vaccinated column, the daily vaccinations column and the people fully vaccinated per 100 people column. Using these columns, as well as some of the information columns and the data from the population dataset, all of the business questions can be answered. When deciding to need to impute the missing values in the dataset, these are therefore the columns that were focused on.

The logic behind imputing the nulls within the columns is as follows:

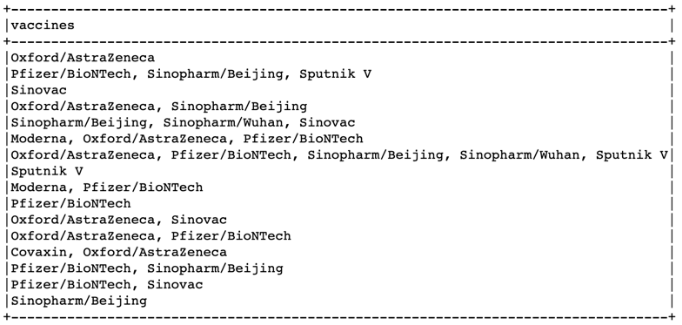
- For each country, organize the rows with respect to the date column and add a counter starting with the earliest date and increasing in order

- If the value of the first date on which the country appears in the data set is null, impute it with a 0
- For any row after this date, impute the null with the most recent non-null value in the row above it (i.e., if the day before 100 people were fully vaccinated, we assume on that day also 100 people were fully vaccinated)

After applying these imputations on the nulls and replacing nulls with 0's in the daily vaccinations, all other columns related to vaccination progress were dropped, since they contained too many nulls and could be recalculated later on using the now processed columns, as well as the population dataset.

Exploration/Cleaning of the Vaccine Column

The vaccine column contains information about the vaccine(s) that each country is vaccinating their citizens with. In some countries this is only one type of vaccine, while other countries use several brands/types. Figure 5 shows the unique types of combinations that exist in the vaccine column.



```
+----------------------------------------------------------------------------------+
|vaccines                                                                          |
+----------------------------------------------------------------------------------+
|Oxford/AstraZeneca                                                                |
|Pfizer/BioNTech, Sinopharm/Beijing, Sputnik V                                     |
|Sinovac                                                                           |
|Oxford/AstraZeneca, Sinopharm/Beijing                                             |
|Sinopharm/Beijing, Sinopharm/Wuhan, Sinovac                                       |
|Moderna, Oxford/AstraZeneca, Pfizer/BioNTech                                      |
|Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing, Sinopharm/Wuhan, Sputnik V|
|Sputnik V                                                                         |
|Moderna, Pfizer/BioNTech                                                          |
|Pfizer/BioNTech                                                                   |
|Oxford/AstraZeneca, Sinovac                                                       |
|Oxford/AstraZeneca, Pfizer/BioNTech                                               |
|Covaxin, Oxford/AstraZeneca                                                       |
|Pfizer/BioNTech, Sinopharm/Beijing                                                |
|Pfizer/BioNTech, Sinovac                                                          |
|Sinopharm/Beijing                                                                 |
+----------------------------------------------------------------------------------+
```

Figure 5 Unique Combinations of Vaccines in the Vaccine Column

Upon further basic analysis the least frequently used vaccine or vaccine combination is the "Sinopharm/Beijing" combination, which only occurs a total of 7 times in the dataset, while the most frequent vaccine combination is "Moderna, Oxford/AstraZeneca, Pfizer/BioNTech", which occurs a total of 703 times in the dataset. It is also important to mention that the vaccine column has no nulls.

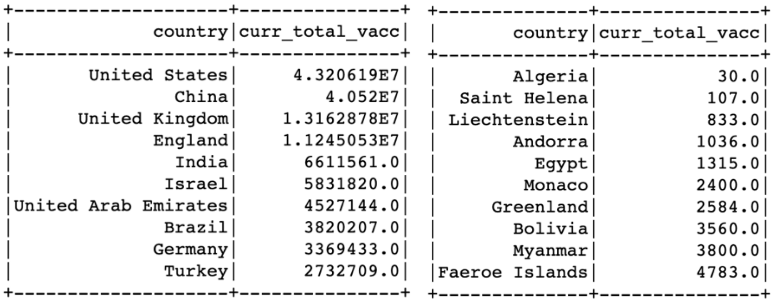
Analysis of the Data

After exploring and cleaning the data, it can now be used, with some more manipulation, to create an analysis which will answer the business questions mentioned earlier. Again, although the dataset is currently still very small, it is expanding rapidly every day, with a potential of 71,175 new lines if every country continues to record their daily progress (even more with autonomous regions, etc.). This is why making use of Spark in order to analyze it makes sense, as the same code can be re-run over the coming months and years in order to analyze and gain insights about vaccination progress once again.

Question 1: In which countries has the vaccination program been the most effective so far? And which countries are performing worst in vaccinating their population?

In order to answer this question, two approaches can be taken. The total vaccinations can show how efficiently a country has been able to administer vaccination overall, however this does not take into account population size of a country, and how much of the population of a country has been fully immunized (both vaccines necessary have been administered). Therefore, the total vaccination amounts, the number of people fully vaccinated, and the number of people vaccinated per 100 people will be analyzed.

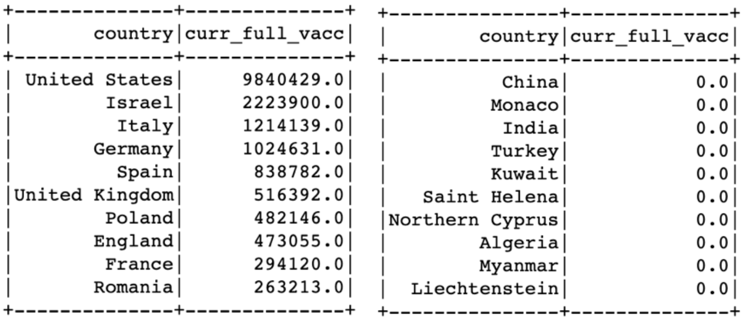
In order to find which countries have administered the highest number of vaccinations, we can simply find the maximum value of the total vaccination column per country. This will allow us to sort these values in descending order and create tables with the countries with the highest total vaccinations and those with the lowest. Figure 6 shows the results of the total vaccinations column. As we can see, the United States, China and the UK (which includes England) are leading the number of total vaccinations administered, while Algeria, Saint Helena and Lichtenstein have the lowest total number of vaccinations administered.



```
+--------------------+---------------+  +--------------+---------------+
|             country|curr_total_vacc|  |       country|curr_total_vacc|
+--------------------+---------------+  +--------------+---------------+
|       United States|     4.320619E7|  |       Algeria|           30.0|
|               China|        4.052E7|  |  Saint Helena|          107.0|
|      United Kingdom|    1.3162878E7|  | Liechtenstein|          833.0|
|             England|    1.1245053E7|  |       Andorra|         1036.0|
|               India|      6611561.0|  |         Egypt|         1315.0|
|              Israel|      5831820.0|  |        Monaco|         2400.0|
|United Arab Emirates|      4527144.0|  |     Greenland|         2584.0|
|              Brazil|      3820207.0|  |       Bolivia|         3560.0|
|             Germany|      3369433.0|  |       Myanmar|         3800.0|
|              Turkey|      2732709.0|  |Faeroe Islands|         4783.0|
+--------------------+---------------+  +--------------+---------------+
```

Figure 6 Top 10 and Lowest 10 Countries According to Number of Total Vaccinations Administered

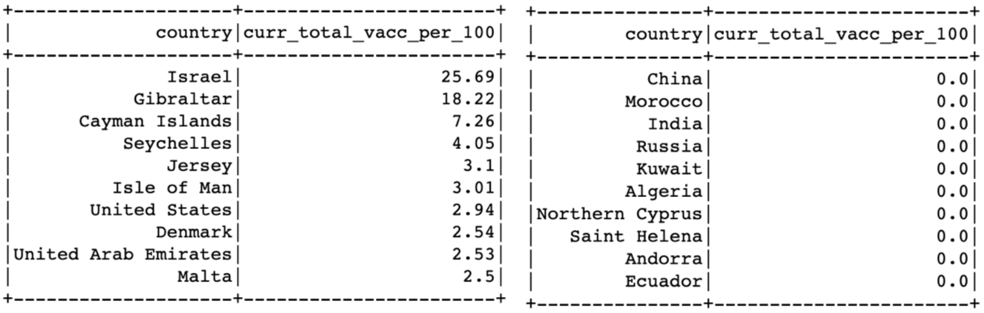
Figure 7 shows the top 10 and lowest 10 countries according to the number of people that have been fully vaccinated in each country. We can see here that missing information regarding the people that have been fully vaccinated is a problem, since a place like China which has administered over 40,000,000 vaccines and has the second highest number of total vaccines overall, has 0 people fully vaccinated. We cannot say that the data is entirely accurate, but from the dataset we can see that overall, the countries that are doing the best in fully vaccinating their population are the United States, Israel and Italy, with many countries in Europe following closely behind. The countries performing worst according to this measure include China, Monaco, India and more, which all currently have 0 people vaccinated (who all probably have many missing values for the fully vaccinated column).



```
+--------------+--------------+  +---------------+--------------+
|       country|curr_full_vacc|  |        country|curr_full_vacc|
+--------------+--------------+  +---------------+--------------+
| United States|     9840429.0|  |          China|           0.0|
|        Israel|     2223900.0|  |         Monaco|           0.0|
|         Italy|     1214139.0|  |          India|           0.0|
|       Germany|     1024631.0|  |         Turkey|           0.0|
|         Spain|      838782.0|  |         Kuwait|           0.0|
|United Kingdom|      516392.0|  |   Saint Helena|           0.0|
|        Poland|      482146.0|  |Northern Cyprus|           0.0|
|       England|      473055.0|  |        Algeria|           0.0|
|        France|      294120.0|  |        Myanmar|           0.0|
|       Romania|      263213.0|  |  Liechtenstein|           0.0|
+--------------+--------------+  +---------------+--------------+
```

Figure 7 Top 10 and Lowest 10 Countries According to the Number of People Fully Vaccinated

The final measure we can use to answer the first business question is finding the countries with the highest number of people fully vaccinated per 100 people and those with the lowest. This indicator takes into account the size of the population of each country, which, in most cases, also indicates the GDP and resources the country can spend on vaccines. Measuring the progress per 100 people in the country is therefore a fairer way of comparing the progress between countries. Again, in the case of the number of people fully vaccinated per 100 people, there were a lot of null values before imputing 0's or previous values instead. This means that many countries may have 0 people fully vaccinated in the dataset, however this does not reflect the real-life situation. In figure 8 we can see that Israel, Gibraltar and the Cayman Islands seem to be doing the best in fully vaccinating their populations, since they have the top values for the number of people vaccinated per 100 people. For the top 10 lowest counts, all values are 0. Countries like China and India are in the lowest 10, which is most likely simply due to missing information, since these two countries have some of the highest number of vaccinations being distributed, as we saw in figure 6.



```
+--------------------+-----------------------+  +---------------+-----------------------+
|             country|curr_total_vacc_per_100|  |        country|curr_total_vacc_per_100|
+--------------------+-----------------------+  +---------------+-----------------------+
|              Israel|                  25.69|  |          China|                    0.0|
|           Gibraltar|                  18.22|  |        Morocco|                    0.0|
|      Cayman Islands|                   7.26|  |          India|                    0.0|
|          Seychelles|                   4.05|  |         Russia|                    0.0|
|              Jersey|                    3.1|  |         Kuwait|                    0.0|
|         Isle of Man|                   3.01|  |        Algeria|                    0.0|
|       United States|                   2.94|  |Northern Cyprus|                    0.0|
|             Denmark|                   2.54|  |   Saint Helena|                    0.0|
|United Arab Emirates|                   2.53|  |        Andorra|                    0.0|
|               Malta|                    2.5|  |        Ecuador|                    0.0|
+--------------------+-----------------------+  +---------------+-----------------------+
```

Figure 8 Top 10 Highest and Lowest Countries According to the Number of People Fully Vaccinated per 100

Overall, there are different measures that can be taken into account to decide which countries are doing best and worst with respect to vaccinating their population. The measure that is most likely the fairest in deciding this is the number of people that have been vaccinated per 100 people in the population. This takes into account only the people that are fully vaccinated and therefore as immune as they can be, and also takes into account of total population of the country and hereby how well a country is doing with respect to achieving herd immunity.

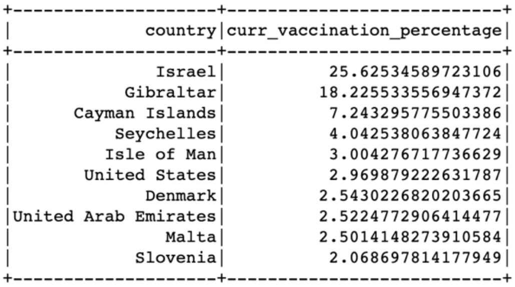
To answer the question, Israel is by far the most advanced, as they have vaccination 25 out of all 100 people in the population. This means they have reached over one third of what is necessary to build heard immunity (70/100 people) within their population. As for the countries that are doing worse, it is a better indicator to look at the number of total vaccinations administered, as information regarding the people fully vaccinated is missing in many countries. With this indicator, Algeria is clearly doing the worst overall, as they have only administered a total of 30 vaccinations throughout their country in the last 59 days.

Question 2: What is the per country and worldwide progress of the vaccinations?

The second business question requires us to once again look at the number of vaccinations that have been administered, but this time focusing on the amount administered with respect to the total population of a country and the world population.

First, we have to find the sum of people that have been fully vaccinated in all countries. To do this we find the largest statistic for the fully vaccinated column for each country (since this will also be the most recent; you cannot have the number of people fully vaccinated decrease from one day to the next, only increase). Next, we can divide this sum by the total world population, which currently lies around 7.8 and billion people, and multiply that by 100. The result of this calculation is the percentage of the worldwide population that is fully immune due to vaccinations and amounts to 0.2498% of the population.

As for the progress per country, we can do the same calculation, but of course dividing by the country's population instead of the worldwide population. The results of this are show in figure 9. As we can see, Israel, Gibraltar and the Cayman Islands are performing the best with respect to vaccinating their population. This coincides with the data from figure 8, which finds the highest full vaccinations per 100 people. The fact that the numbers are so close are a good sign that the data in our dataset is accurately calculated.



```
+--------------------+---------------------------+
|             country|curr_vaccination_percentage|
+--------------------+---------------------------+
|              Israel|          25.62534589723106|
|           Gibraltar|         18.225533556947372|
|      Cayman Islands|          7.243295775503386|
|          Seychelles|          4.042538063847724|
|         Isle of Man|          3.004276717736629|
|       United States|          2.969879222631787|
|             Denmark|         2.5430226820203665|
|United Arab Emirates|         2.5224772906414477|
|               Malta|         2.5014148273910584|
|            Slovenia|          2.068697814177949|
+--------------------+---------------------------+
```

Figure 9 The Top 10 Countries with the Highest % of their Population Fully Vaccinated

From the information we have extracted we can now answer question 2. The worldwide progress of vaccinations up until 09.02.2021 was that 19,487,189 had been fully vaccinated, making up around 0.25% of the worldwide population. On a per country basis, the countries shown in figure 9 have made the most progress in vaccinating their population. Overall, however we can say that progress is quite slow, as countries have vaccinated as low as 2% of their population but have managed to land in the top 10 countries.

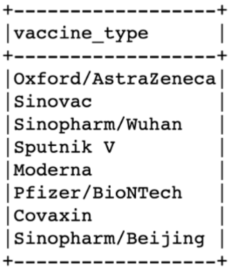
Question 3: Which vaccine type has been used the most?

In order to answer this question, the vaccine column must be manipulated. In its raw form, the column contains the names of all the vaccination types/brands which have been administered in a country. In order to discover which type has been used the most we can start by looking at the number of occurrences for the vaccination combinations. The result of the top 3 combinations of these calculations is below:

1. Moderna, Oxford/AstraZeneca, Pfizer/BioNTech: 703
2. Pfizer/BioNTech: 627

3. Oxford/AstraZeneca, Pfizer/BioNTech: 1139

Clearly the Pfizer/BioNTech, Oxford/AstraZeneca and Moderna vaccinations have been used the most, however other than the Pfizer one, these are all combinations of vaccinations and not individual ones. In order to find the possible occurrences of each one, we can count the number of times each individual vaccination was found in the dataset. Figure 10 shows a list of all the individual vaccination types/brands in the dataset.



```
+------------------+
|vaccine_type      |
+------------------+
|Oxford/AstraZeneca|
|Sinovac           |
|Sinopharm/Wuhan   |
|Sputnik V         |
|Moderna           |
|Pfizer/BioNTech   |
|Covaxin           |
|Sinopharm/Beijing |
+------------------+
```

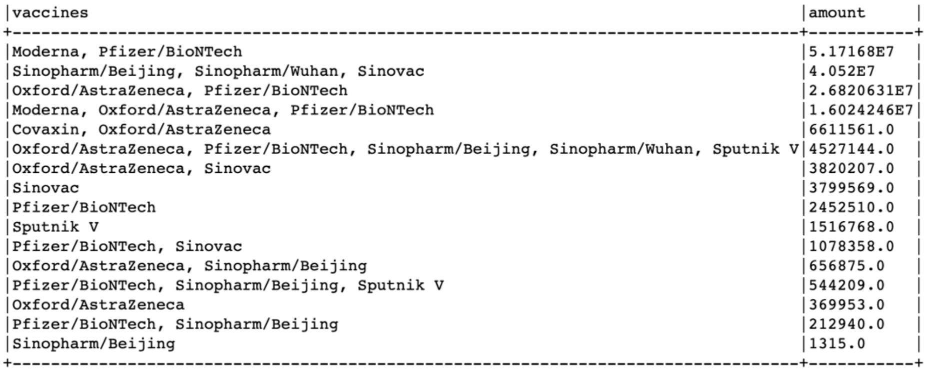
Figure 10 Vaccination Types in Dataset

The following are the results when occurrences of each individual vaccine are counted in the dataset:

1. Pfizer/BioNTech : 2263 occurences
2. Oxford/AstraZeneca: 1278 occurences
3. Moderna: 1063 occurences

This dataset does not have any information on how many of the total vaccinations each type of vaccine has accounted for, however it is clear that in both the highest occurrences of groups and the highest occurrences of each individual vaccine, the Pfizer/BioNTech, Oxford/AstraZeneca and Moderna are clearly leaders.

Finally, we can calculate the total number of vaccinations that have been administered of each combination of vaccines, in order to answer exactly which vaccination type(s) has been administered most often. As show in figure 11, the Moderna, Pfizer/BioNTech, and Oxford/AstraZeneca vaccinations seem to have a high number of total doses administered, as you would expect, but in this case the Sinopharm/Beijing, Sinopharm/Wuhan and Sinovac vaccinations are also very high on the list. These were previously probably not as obvious as they are mainly used in China, and therefore have few row occurrences, but since the population of China is large, they also have a high total amount of vaccinations administered.



```
+----------------------------------------------------------------------------------+-----------+
|vaccines                                                                          |amount     |
+----------------------------------------------------------------------------------+-----------+
|Moderna, Pfizer/BioNTech                                                          |5.17168E7  |
|Sinopharm/Beijing, Sinopharm/Wuhan, Sinovac                                       |4.052E7    |
|Oxford/AstraZeneca, Pfizer/BioNTech                                               |2.6820631E7|
|Moderna, Oxford/AstraZeneca, Pfizer/BioNTech                                      |1.6024246E7|
|Covaxin, Oxford/AstraZeneca                                                       |6611561.0  |
|Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing, Sinopharm/Wuhan, Sputnik V|4527144.0  |
|Oxford/AstraZeneca, Sinovac                                                       |3820207.0  |
|Sinovac                                                                           |3799569.0  |
|Pfizer/BioNTech                                                                   |2452510.0  |
|Sputnik V                                                                         |1516768.0  |
|Pfizer/BioNTech, Sinovac                                                          |1078358.0  |
|Oxford/AstraZeneca, Sinopharm/Beijing                                             |656875.0   |
|Pfizer/BioNTech, Sinopharm/Beijing, Sputnik V                                     |544209.0   |
|Oxford/AstraZeneca                                                                |369953.0   |
|Pfizer/BioNTech, Sinopharm/Beijing                                                |212940.0   |
|Sinopharm/Beijing                                                                 |1315.0     |
+----------------------------------------------------------------------------------+-----------+
```

Figure 11 The Total Number of Vaccinations Administered by Group of Vaccination Types

To conclude the findings and answer question 3, we can say that according to the number of countries in which a vaccination is administered, the Pfizer/BioNTech, Oxford/AstraZeneca and Moderna are clearly leaders. On the other hand, if we want to measure according to the total number of doses administered, the previous 3 vaccinations are still high on the list, but the Sinopharm/Beijing, Sinopharm/Wuhan and Sinovac vaccination also join the top spots.

Question 4: If we continued at the current rate, when would her immunity be achieved?

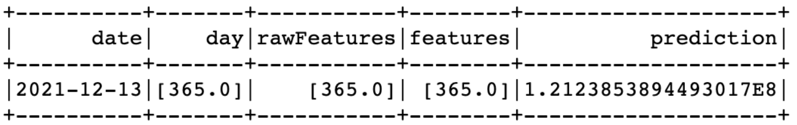
The final business question is a little more difficult to answer than the rest. In order to find how long it would take for herd immunity to be achieved we need to analyze how many people have been vaccinated over the past, and what the trend looks like in the future. The first, over-simplified way to do this is by simply looking at the previous 59 days for which we have information, and assuming that all future days will have the same amount of vaccination progress. Doing this is a simple calculation like the one shown below:

$$\frac{0.7}{x} = \frac{0.0025}{59} \rightarrow\ x = 16{,}520\ days = 45.26\ years$$

Of course, this way of calculating how long it will take to vaccinate 70% of the population is way over simplified, and doesn’t take into account issues such as how some countries haven’t started their vaccination schemes, the supply problems with some vaccinations, and many more factors that will cause the amount of vaccinations being administered to increase significantly in the future. In order to properly estimate when we will achieve herd immunity, we need to analyze the trend of the number of people who have been fully vaccinated vs. the days.

An algorithm that can be used to calculate a continuous variable based on other variables, in our case only the date, is Gradient Boosted Tree regression. It doesn’t really make sense to apply an ML algorithm to this data, however the GBTRegressor from the Pyspark ML library simplifies the use of this algorithm significantly, and it was a chance at improvement to the prediction. After preparing the data accordingly and executing all the steps necessary, the GBT model is able to predict a test set from the dataset with a RMSE of $2.05367*10^6$. This RMSE is very high, which is in part due to the large numbers of vaccinations, but also due to the fact that the model is simply not good. The model relies solely on the day value, which is the date transformed into a count from day one of the dataset, in order to predict the number of people totally vaccinated, which in no context would make more a good amount of features for prediction with a tree model.

From the model that was fit, although not ideal, we can predict how many people are vaccinated in a year from now, and with that the percentage of immunity that the world will have at that time. The prediction generated for the 13.12.2021 by the GBT model is that around 121 million people will be vaccinated at that time, as shown in figure 12. This would mean that 1.55% of the population will be vaccinated in December of 2021. As before, this would mean very slow progress, and years would be needed to vaccinate the people necessary for the world to be completely immune to COVID-19.



```
+----------+-------+-----------+--------+--------------------+
|      date|    day|rawFeatures|features|          prediction|
+----------+-------+-----------+--------+--------------------+
|2021-12-13|[365.0]|    [365.0]| [365.0]|1.2123853894493017E8|
+----------+-------+-----------+--------+--------------------+
```

Figure 12 The Prediction for the Number of People Vaccinated on 13.12.2021, 1 Year After Beginning of Dataset

All this being said, it does not make much sense to apply an algorithm to this data, since there are very few features that help with prediction, and in the data is only for the first 2 months of vaccinations, a time in which many countries were still struggling to properly organize vaccinations, let alone start vaccinating their population. Therefore, although the outlook from the data is not very bright, once more data becomes available over the coming months, the supply of vaccinations becomes more stable, and countries become more efficient in administering the vaccine, the prediction of the time required will reduce significantly.

Conclusion

Overall, after an in-depth analysis of the dataset, we can say that the progress of vaccinating the global population against the COVID-19 virus is not much. Overall, 19,487,189 people have been fully vaccinated in the 59 days that the dataset entails, making up around 0.25% of the world population. The countries that have done the best at getting their population fully vaccinated include Israel, Gibraltar and the Cayman Islands. Many other countries on the other hand, have not even begun vaccinating their citizens. It can also be concluded that the three main vaccinations that are being used around the world are the Moderna, Oxford/AstraZeneca, and Pfizer/BioNTech vaccines. These vaccines are being used in many countries and have also been administered very often. Finally, we can say that if vaccination rates continue as they have been, only 1.55% of the population will be vaccinated in a year from now, much less than is necessary in order to reach the percentage of the population required for herd immunity and be able to return to normal life.

In conclusion, the progress of COVID-19 vaccinations around the world has been nowhere near necessary in order to reach heard immunity solely through vaccinating within the next year or even few years. It can be said from the analysis within this report that governments worldwide need to work much harder to vaccinate a significantly higher number of people daily in order to make sure that their populations become fully vaccinated as soon as possible. The world is relying on the vaccinations to bring back normality, however their implementation is much slower than it needs to be to prove effective.

In the future, the code used for the analysis can be applied again in order to measure progress, but with more data. As said before, the dataset is currently very small, and Spark is not necessary in order for the analysis to be fast and efficient, however the dataset will be growing rapidly, as more countries, vaccines and data becomes available. When it does grow over the coming years, having implemented everything in Spark will make sure that the analysis is fast and smooths. Some additional steps that could be taken to improve the predictions (question 4) specifically, would be to enrich the dataset with more features such as the available healthcare workers to administer vaccinations, the budget for vaccinations in each country, the importance a country's government places in vaccinating, and more. All this information could contribute to the speed at which people are being vaccinated and will therefore also help in predicting more accurately how soon we reach herd immunity.