UCD Certificate in Introductory Data Analytics

Cristina Vilaró Palacín

March/2021

This report has been prepared as part of the Certificate in Introductory Data Analytics course hosted by UCD.

The Github repository URL is https://github.com/cristinavip/UCDPA_Cristina-Vilaro-Palacin.git. Unfortunately, the original repository for this project was deleted, and therefore a new repository was created for the purpose of this course but it does not include all the commits performed throughout the project.

The information analysed displays COVID-19 data using some Python functions. The data used for the creation of this report was downloaded from the Kaggle site on the 13th of March 2021 as CSV files.

The datasets downloaded were part of the COVID-19 Global Dataset – version 31 (*worldometer_coronavirus_daily_data.csv* and *worldometer_coronavirus_summary_data.csv*) and the COVID-19 World Vaccination Progress – version 68 (*country_vaccinations.csv*). These datasheets are live and continually updated, therefore any data added to the dataset after that date will not be reflected in this report. The csv files were used as is, the only modification done was in the csv file *country_vaccinations.csv* the dates were reformatted to be able to transform them easily to a dataset.

Data has been analysed using PyCharm Community version 2020.3.3 launched via Anaconda 3 navigator. Python version 3.8 was used and the environment selected for the project was conda.

Pandas, numpy and matplotlib tools and libraries were downloaded in Pycharm and imported to the project. The CSV files mentioned above were also imported and merged into one unique dataframe based on the column 'country'. Duplicates based on all fields were dropped, although none were found. Na data was kept as is to not give false information. Moreover, blank cells were replaced by NaN.

The data will be presented in two blocks, based on the different CSV files imported (the first part of the data will belong to number of cases and deaths due to COVID19 and following will be data regarding vaccines). These CSV files contain a wide amount of data and not all will be reflected in this report.

Comparison of COVID19 confirmed cases and deaths per 1 million population per continent

For this figure, data was grouped by continents and the columns 'Number of Total Cases per 1 million Population' and 'Total Deaths per 1 million Population' values were summed. These columns were selected instead of the total numbers as it is a more suitable way to compare inter continents, as it accounts for the number of populations. Then the data 'Number of Total Cases per 1 million Population' and 'Total Deaths per 1 million Population' were sorted from higher to lower. The data was plotted using a bar chart representation:

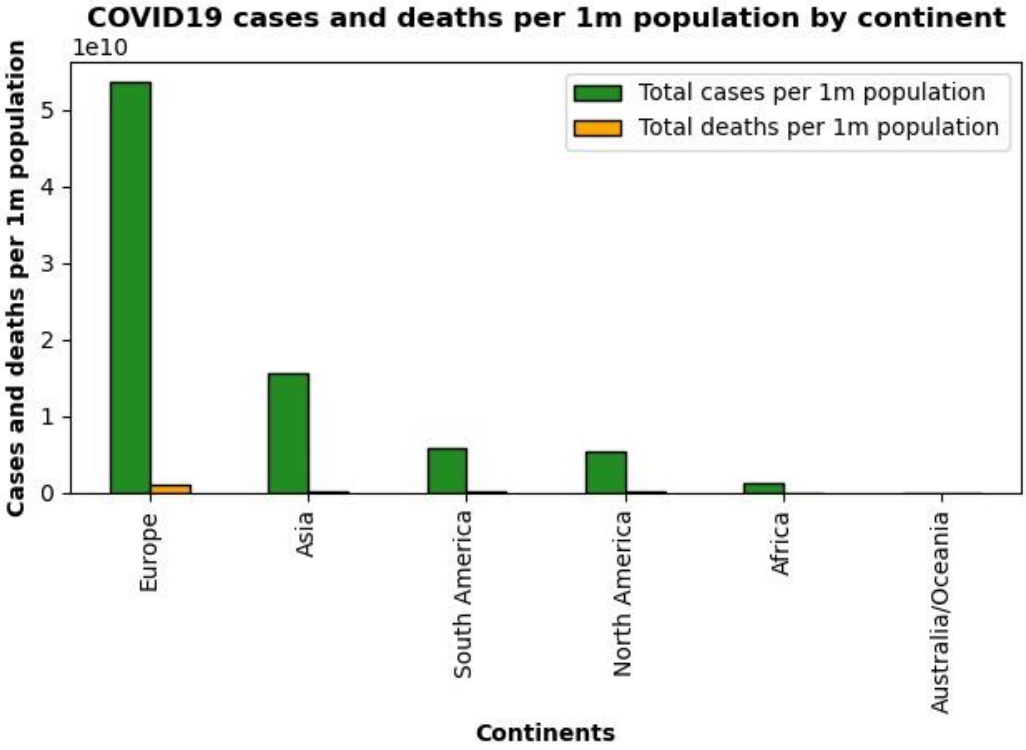

Figure 1 - COVID19 cases and deaths per 1 million population by continent

As we can see in the Figure 1, the continent with a higher number of cases per 1m population is Europe (over 50000 per 1 million population), followed by South America.

On the opposite side we find Australia/Oceania with a number of total cases of <250 (fairly visible on the chart). Regarding the number of deaths, Europe again is the continent with a higher number of deaths per one million population.

Comparison of COVID19 confirmed cases and deaths data and COVID19 confirmed cases and deaths per 1 million population in EU countries

For Figure 2, a list based on EU countries was created (instead of the label Europe), and only the rows in the listed country were selected (using isin). The data then was sorted from higher number in the column 'Total Confirmed' to lower, and plotted using a scatter graph. The 'Total Deaths' column data from these countries was added using a colour-bar. Grid lines were added to help reading the graph:

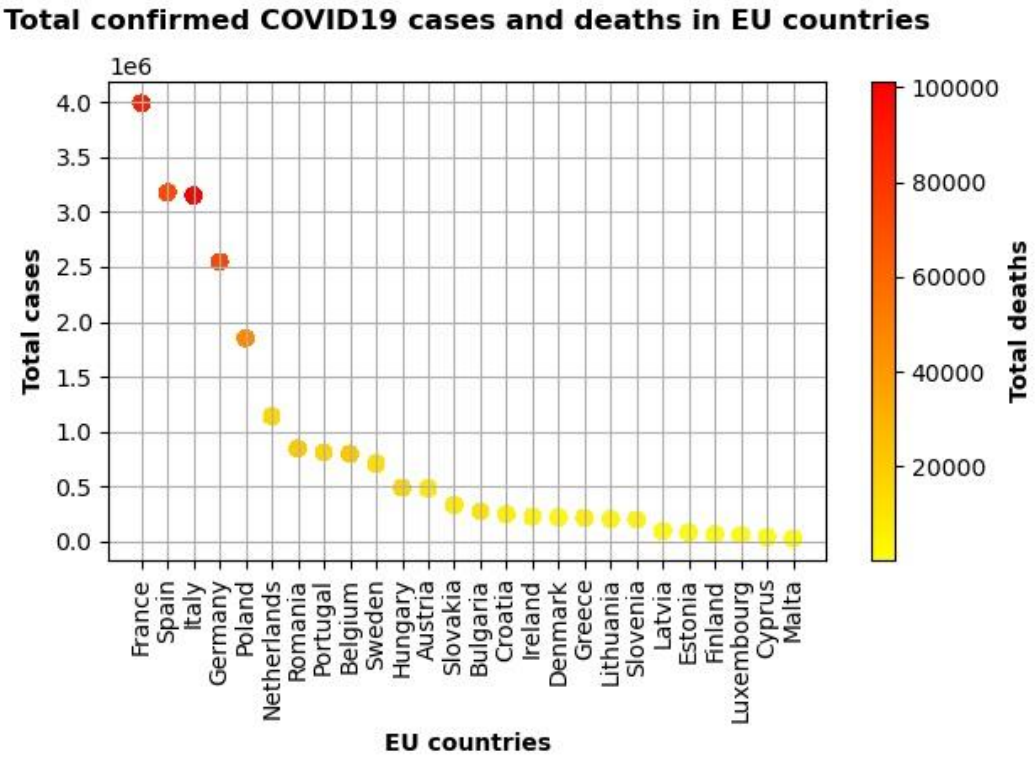

Figure 2 - Total confirmed COVID19 cases and deaths in EU countries

The graph above shows a right skewed shape. The EU countries with a higher number of total confirmed cases of covid19 are France leading the chart with around 4 million people, followed by Spain and Italy both with a similar number of confirmed cases, above 3 million confirmed cases. On the other side of the chart Latvia, Estonia, Finland, Luxemburg and Malta all have the lowest rate of cases, with a similar number between them with numbers way below the half million cases.

When comparing this data against the number of total deaths, it can be seen there is a direct relationship between the countries with a higher number of cases and higher number of deaths (red colour), the countries that fall in the middle of the graph (colour of the orange and dark yellow spectrum) and finally the countries with fewer number of cases have a faded yellow colour. Therefore, from this graph we can say there is a direct relationship between having a high number of total confirmed cases and total deaths.

Regarding number of COVID19 cases and deaths per 1m population graph (Figure 3), it followed the same rationale as Figure 2: the data only from the EU countries list was selected, the data based on the columns 'Total cases per 1m population' was selected from higher number to smaller, and data was plotted in a scatter plot with a colourbar reflecting the data from 'Total deaths per 1m population' column.

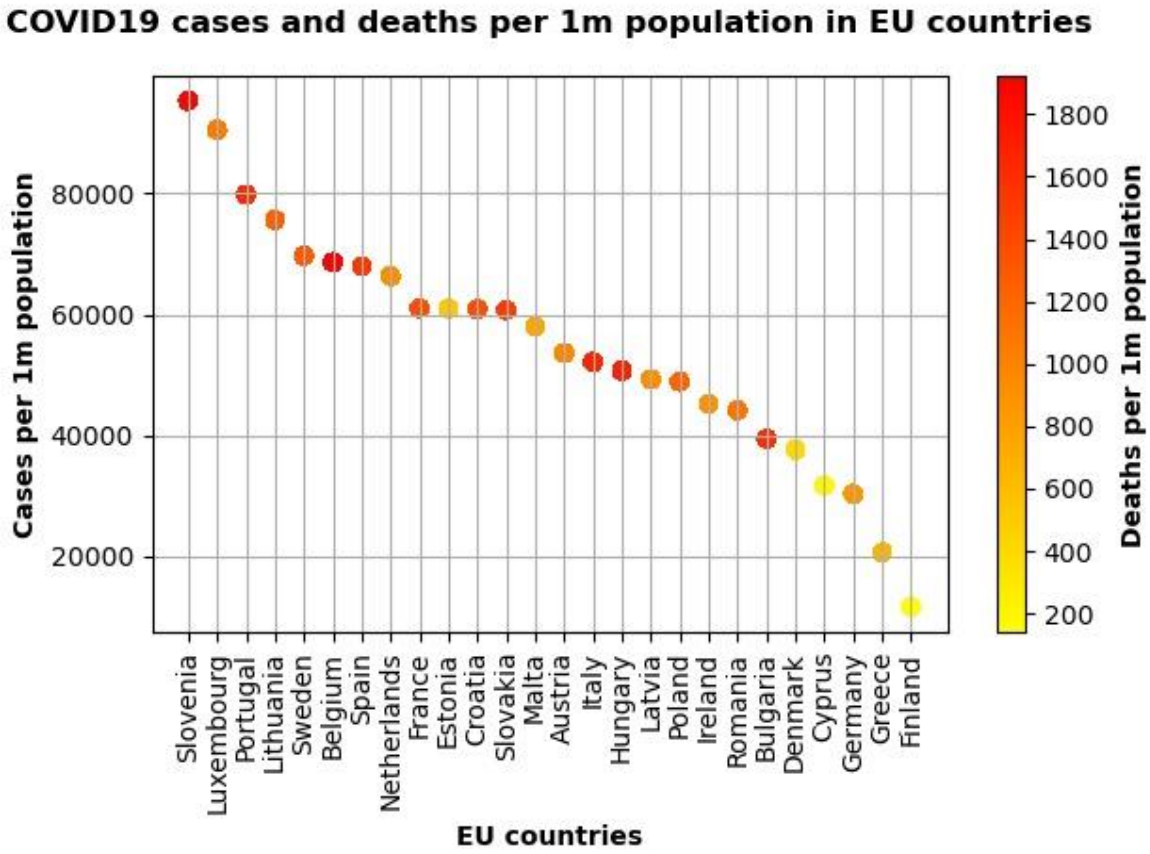

Figure 3 - COVID19 cases and deaths per 1 million population

The graph above shows a sigmoidal shape showing a normal distribution. The EU countries with a higher level of cases per 1m population in the EU are first Slovenia, Luxemburg and Portugal (above 80000 cases per 1 million population) while the countries with a lower level of cases per 1m population in the UE are Germany, Greece and Finland.

When comparting the data with the number of total deaths per 1 million population, overall a relationship can be seen between a higher number of total cases per 1 million population and total deaths per 1 million population (there is a higher concentration of red and dark orange on the side of higher number of cases and a higher concentration of weak orange and yellow on the side of lower number of cases), although the relationship is not as straightforward as it was in the figure reflecting the number of total confirmed cases and deaths.

What about Ireland? Cumulative number of COVID19 cases since start of the pandemic

As we can see in Figures 2 and 3, in both cases Ireland falls in the middle of the graph towards the side. For more specific data regarding the number of COVID19 cases the next figure was created indexing the column 'country' and then selecting only rows containing the word 'Ireland' using loc. The dates were changed from strings to datetime. The information was plotted using 'date_y and 'cumulative total cases' columns:

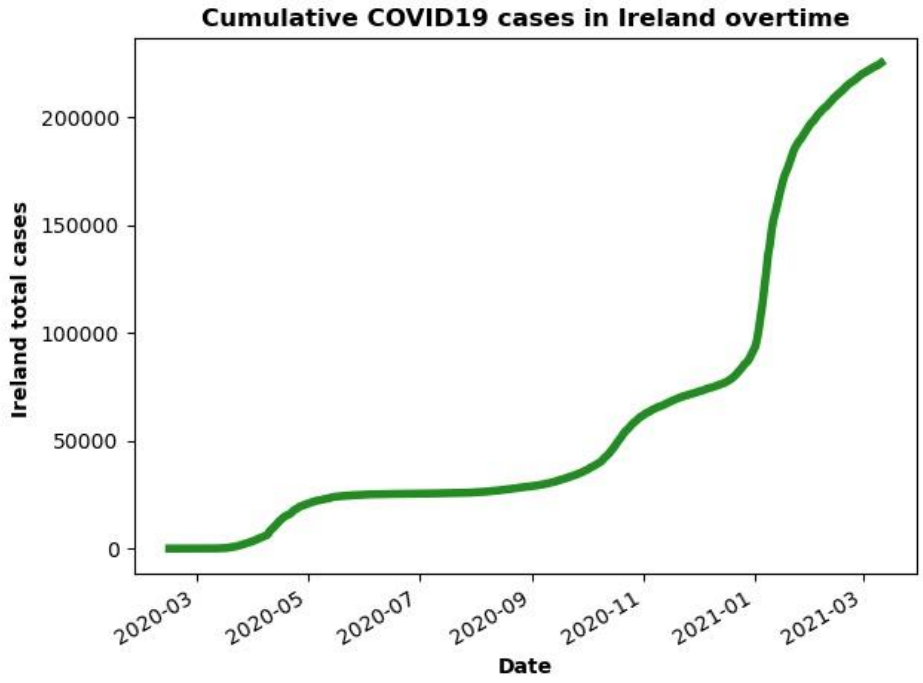

Figure 4 - Cumulative COVID19 cases in Ireland overtime

The data shows how there was a small increase at the start of the pandemic, especially after May 2020, plateauing after the first period of quarantine (from approximately June) increasing slightly after September, and continuously increasing during November.

A quick, sharp increase of cases it is seen at the end of December and throughout January, slowing down in February.

Comparison data of approved vaccines per number of countries and continent

Moving into vaccines data, Figure 5 wants to compare the different combinations of approved vaccines based on the number of countries where they are approved. The duplicates based on 'country' were dropped as keeping them would create a false reflection of reality – as such, only the first entry was kept. The 'vaccine' column was counted using value_count and plotted using a horizontal bar chart:

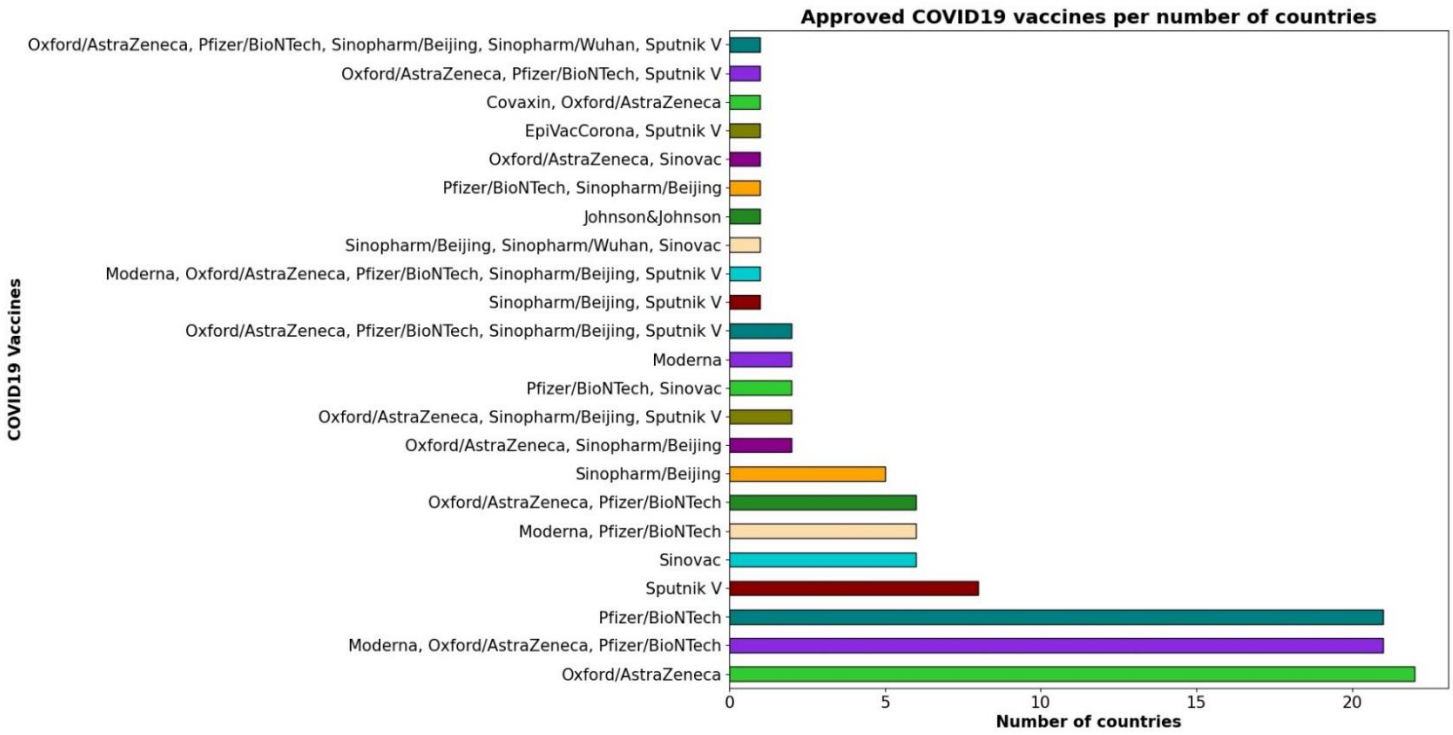

Figure 5 - Approved COVID19 vaccines per number of countries

From Figure 5 we can see the more common combination worldwide is Oxford/Astrazeneca, followed closely by Pfizer/BioNtech and Moderna-AstraZeneca-Pfizer/Biotech (this last combination reflects the ones approved in the European Union until when this excel datasheet was last updated in Keggle. This data does not reflect the population of each country but only the number of counties.

We can see a disclosure of this data based on continents in Figure 6 showing a proportion of which combination of vaccines are most and less common per continent. This data was grouped by 'continents' and 'vaccines' columns and plotted using a stacked bar chart:

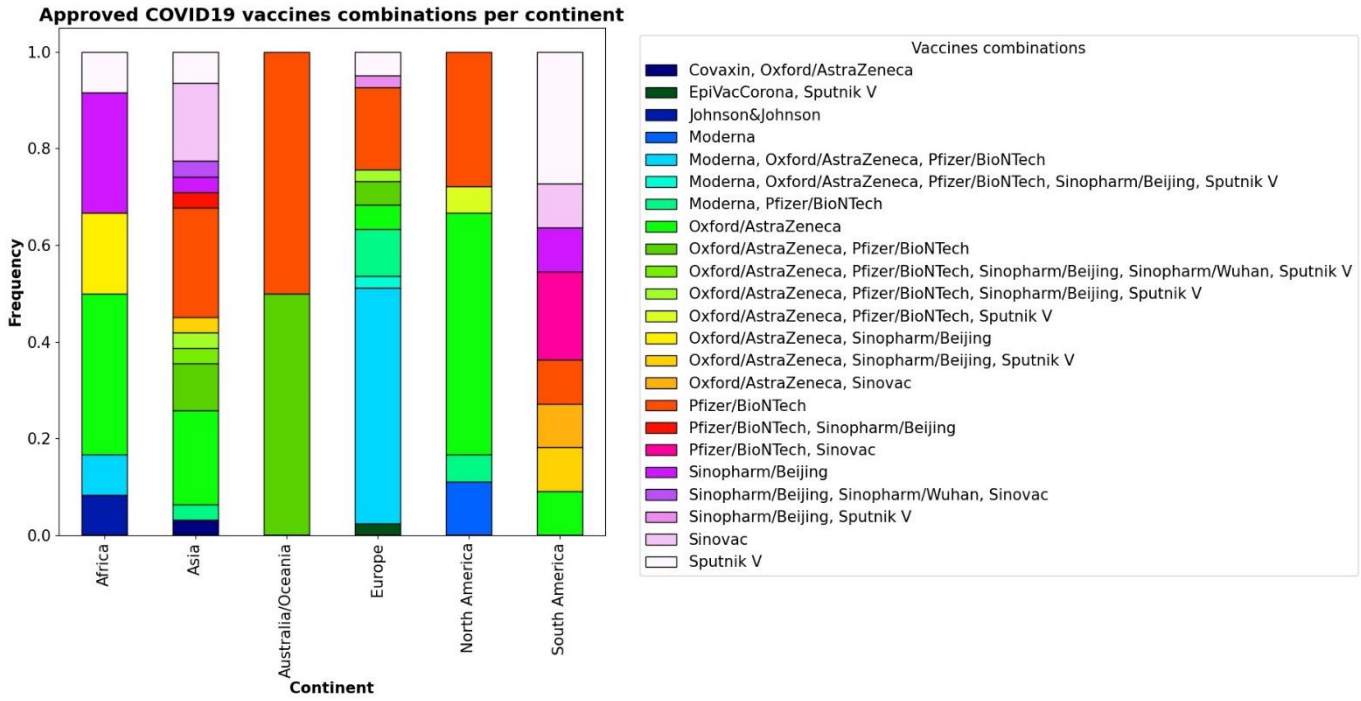

Figure 6 - Approved COVID19 vaccines combinations per continent

Figure 6 only reflects the frequency of an approved vaccine or series of approved vaccines per country in each continent, without accounting for the population of these countries. In Africa the vaccine used for more countries is Oxford/Astrazeneca, followed closely by the Sinopharm/Beijing vaccine.

In Asia Oxford/Astrazeneca and Pfizer/ BioNTech are the vaccines used for more countries in a similar percentage, while in Australia/Oceania the Pfizer/ BioNTech vaccine and the combination of Pfizer/Biontech and Oxford/Astrazeneca are equally used in the continent.

In Europe, the most used combination is Moderna, Oxford/Astrazenca and Pfizer/BioNTech vaccines. In North America the most used vaccine is the Oxford/Astrazeneca and finally in South America the most frequent vaccine is the Sputnik V.

Total vaccinations per country

First of all, the total number of vaccinations will be compared: the countries with a higher number of total vaccinations, only the latest entry (the most recent) of each country will be used, as it is deemed to reflect the latest cumulative data. Therefore, the data was sorted by vaccination date ('date_x' column), with the most recent entries at the top, and then duplicates based on country were dropped, keeping the first entry.

This data was indexed per column 'country' and sorted by the 'total vaccinations' column with the highest number at the top. The data then is sliced using iloc and only keeping the first 20 rows. Finally, the data is plotted using a bar chart. It presents as follows:

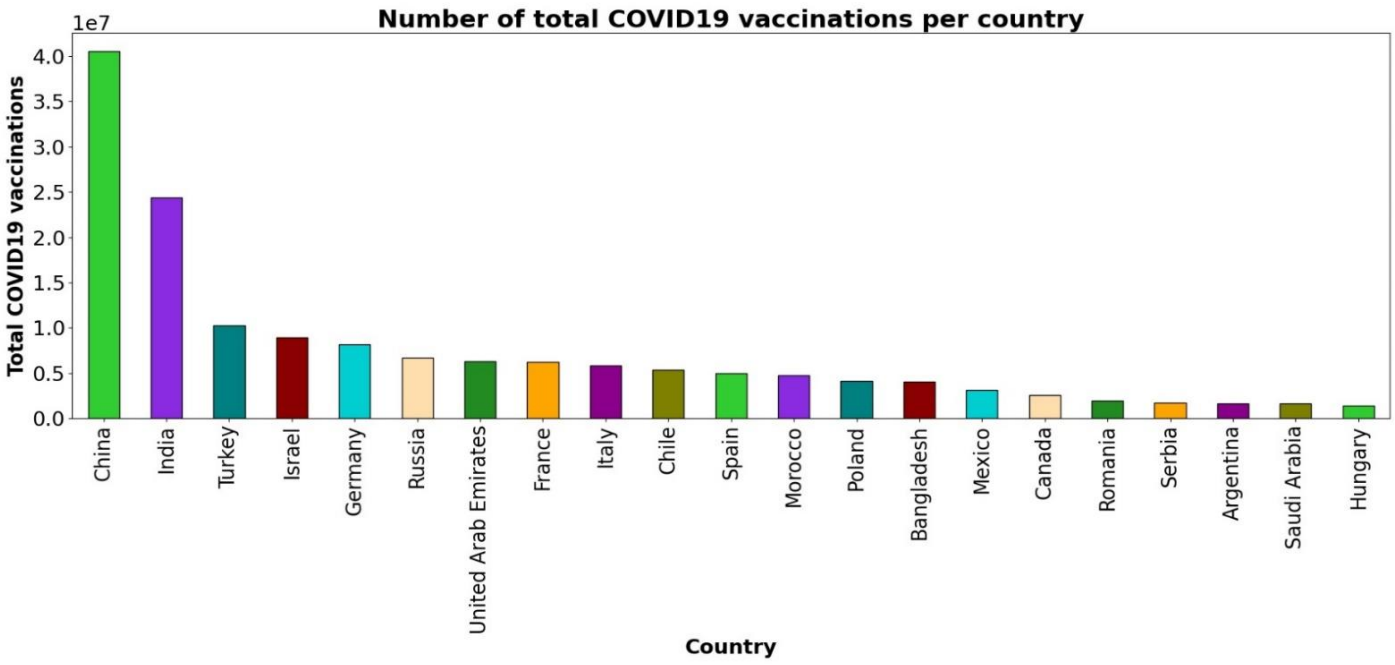

Figure 7 - Number of total COVID19 vaccinations per country

The chart above reflects the 20 countries with a higher number of vaccinations administered.

It is worth noting that the UK data is not reflected per se as the entries in the original csv file were entered as England, Wales, Scotland and Northern Ireland. Keeping that in mind, the countries with a higher total number of vaccines administered are China stretching far ahead, following by India, and further back Brazil, Turkey and Israel. The first EU country to appear on the chart is Germany in 5th position.

People vaccinated per hundred per country

To analyse the data per hundred the same rationale was used as above, but instead of using the 'total vaccination' column, the 'people vaccinated per hundred' column was used. The data was sorted by this column and iloc slicing was performed on the first 20 entries. The data then was plotted as a bar chart:

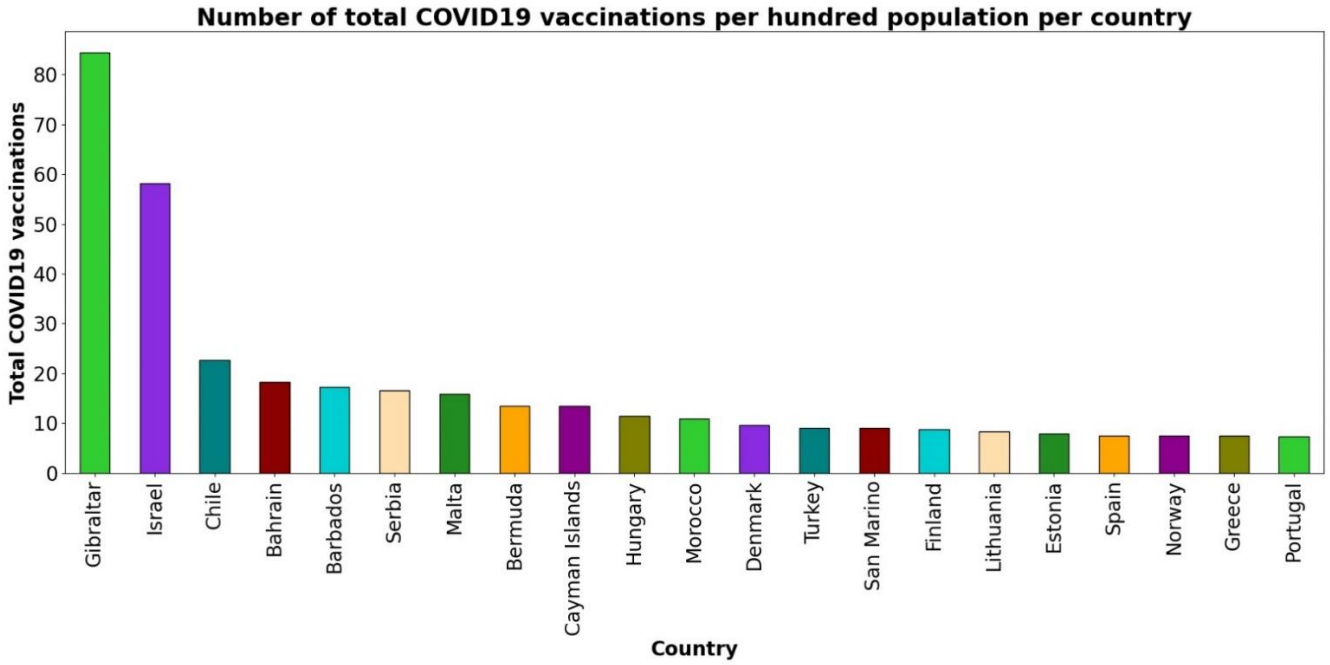

Figure 8 - Number of total COVID19 vaccinations per hundred population per country

The graph above shows how small countries are at the top of the graph (such as Gibraltar, with more than 80% people vaccinated and leading the chart, Seychelles, Cayman Islands, Anguilla and Bermuda, as examples).

This makes sense in a situation where production of the vaccine and its delivery have been the main setback in vaccination roll out.

The first big country to appear on the chart is Israel in third position, with almost 60 people out of 100 vaccinated. The first EU country to appear is Malta in the 11th position and with almost 20% of people vaccinated.

Daily vaccinations per million population in Ireland

Finally, looking at the numbers of vaccinations in Ireland, the next figure was created following the same approach as the figure regarding the cumulative number of cases in Ireland, but instead this time it shows data related to the number of daily vaccinations per million in Ireland. The data was indexed based on the 'country' column, and only the Ireland entries were selected using loc.

Columns 'vaccinations' and 'date_x' are plotted:

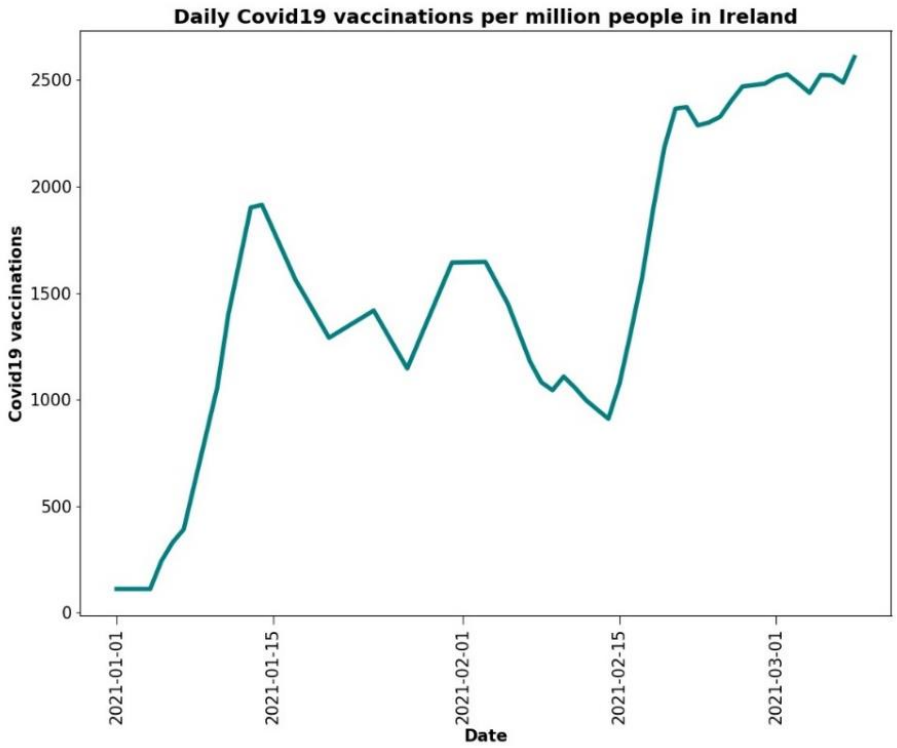

Figure 9 - Daily COVID19 vaccinations per million people in Ireland

As we can see in Figure 9, there was a fast start with the vaccinations process from the end of the December until mid-January, slowing down afterwards until mid-February (with a slight increase in the middle), and then rapidly increasing between mid-February to the beginning of March, reaching a plateau that extends until the most recent data added to the dataset.