Today we are going to explore together the dataset provided by the UK Government about COVID-19 cases and vaccinations. The goal of the Government is to increase vaccination rates through marketing campaigns to promote the vaccine. I have carried out research leveraging Python to help inform the Government about the next steps.

We notice interesting aspects about the data provided which need further investigation. For example, for inland UK we have data just for Moffat in Scotland in both data frames, which also present values equal to zero for the recovered and hospitalised last entries. With print(cov.info()) I notice that Deaths, Cases, Recovered, and Hospitalised are missing 2 rows of data for Bermuda which I replaced at row 875 and 876 on 21st and 22nd September 2020.

With print(vac.describe()) I notice that I get the minimum, Q1, and Q2 as 0 for Vaccinated, First, and Second doses. This could happen because, considering the change over time for vac, we see that the vaccinations took place well after cases started (e.g. in Anguilla they started from 11/01/2021).

I also noticed wrong entries such as that of Saint Helena, Ascension and Tristan de Cunha is under Subregion Name “Northern Europe”.

A precious insight I gathered is that the data for each State come from a single geographical point. Also, we have the value counts for each State set at 632. This hinted me to a direction which we will return on later.

The value counts for cov[‘Lat’] are only 632, and vac[‘First Dose’] gives me a length of only 3206 and row zero displays a value of 4284.The Dataframe has the columns names as default index.

Here are some initial insights I have gathered about vaccinations: I calculated the IQR, lower and upper limit, and the calculated range for Cases, Deaths, Recovered, and Hospitalised. These are statistical parameters that help us in understanding the data and their distribution for drawing a clearer picture.

How has the number of cases changed over time? In cov we first notice that we start with hospitalisations and then the other values raise too (see the case of Anguilla). Bermuda records the first case in 19/03/2021 and every other country presents its own peculiarities. Similarly, we note a change over time for the vaccinated individuals.

We still need to assess the quality of the data that has been provided. There appear to be misleading data, for example in Gibraltar as well as other States. First, deaths by COVID-19 start and then cases start been recorded too as if there wasn’t an established method to record or check positive cases.

Deaths, Cases, Recovered, and Hospitalisations are cumulative and there are more Hospitalisations than cases from row 4050 to 4167, after that Hospitalisations reduce abruptly, then raise again. The Recovered values stop abruptly as we shall see later.

A technique I have adopted, is that of breaking down in smaller section the DataFrame and wrangled the data.

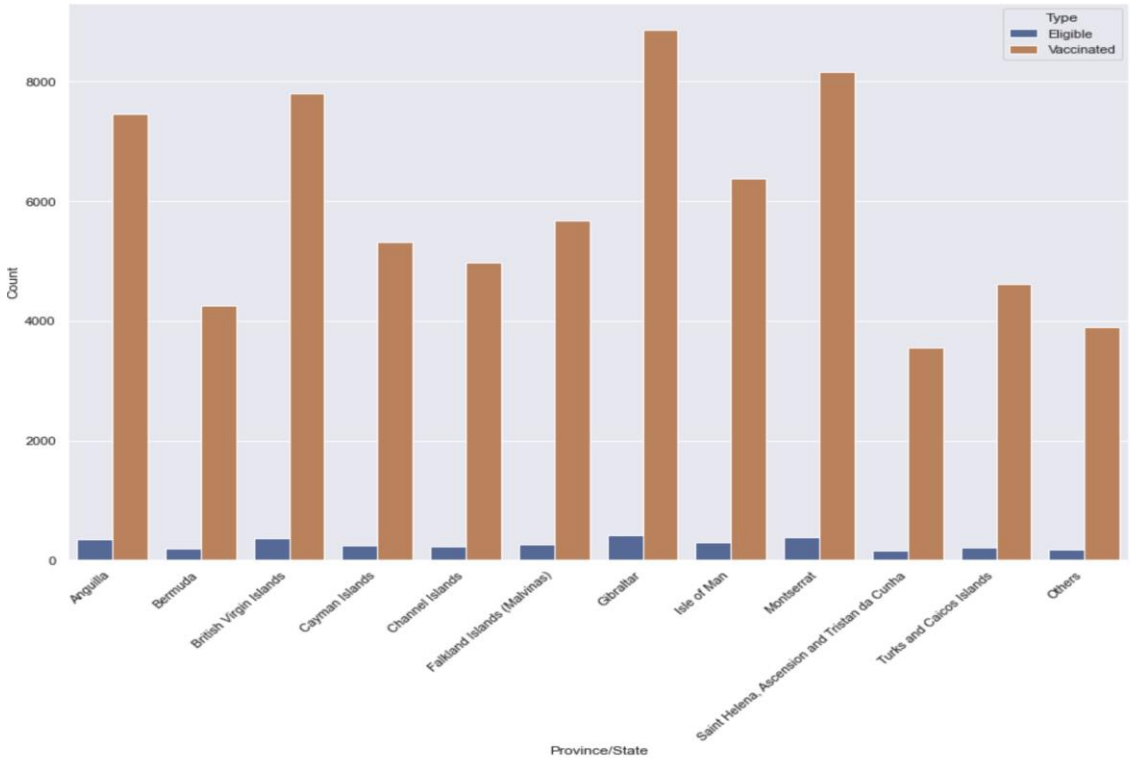
I believe that the data provided is a sample of a population but is in no way representative: by employing this data we are going to get wrong results. Sampling is a technique adopted in analysing data but in this case, we have the tools to know precisely the details of the population. The above looks like a clustered sample, where randomly selected subgroups (clusters) from the population are employed. A cluster could be geographical, such as a single town in a Province/State but again the numbers aren't adding up and especially the category "Others" has a strong impact the whole dataset.

I have a solution for you and will show where better data can be gathered through public repositories, such as the WHO or ONS.

Given this clarification, I want to show you under a technical point of view how I would approach the matter and to do this I will employ the dataset provided as an example to answer to the questions and suggest successful solutions.

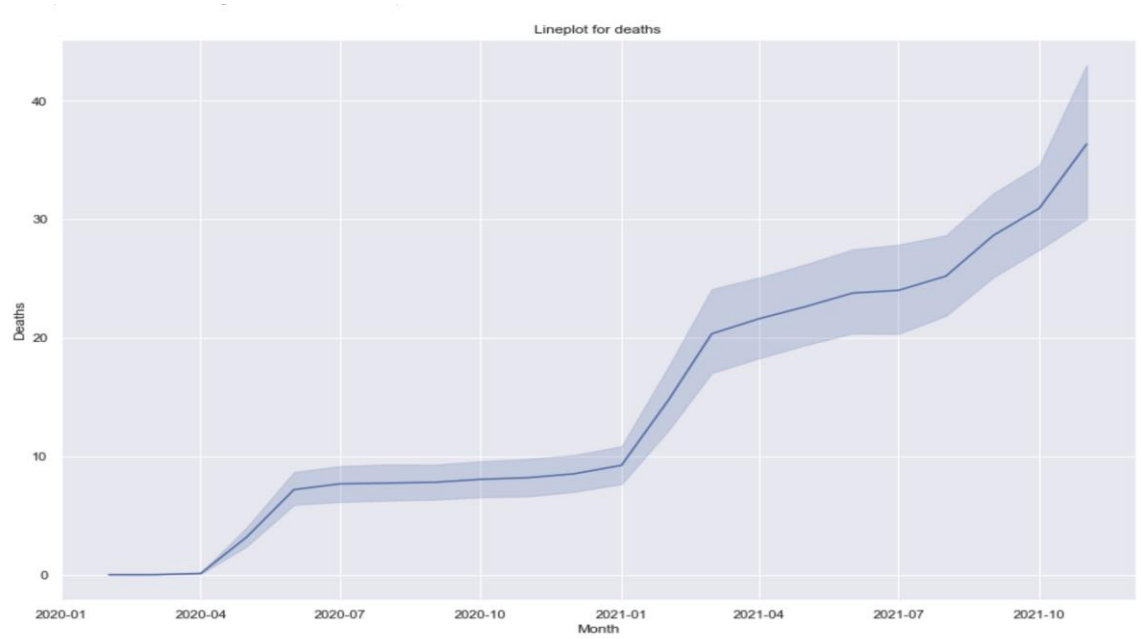
Considering the vaccinations rates in the provinces, to inform in the decision making for the marketing campaign, I have determined if there are differences over time between the first and second dose application. Also, which Province/State has the highest number of individuals who have received a first dose but not a second dose? It is Gibraltar. We are going to explore these data further in detail through visualisations I made with Python.

First, we have a barplot that allows us to compare first and second doses through the States:



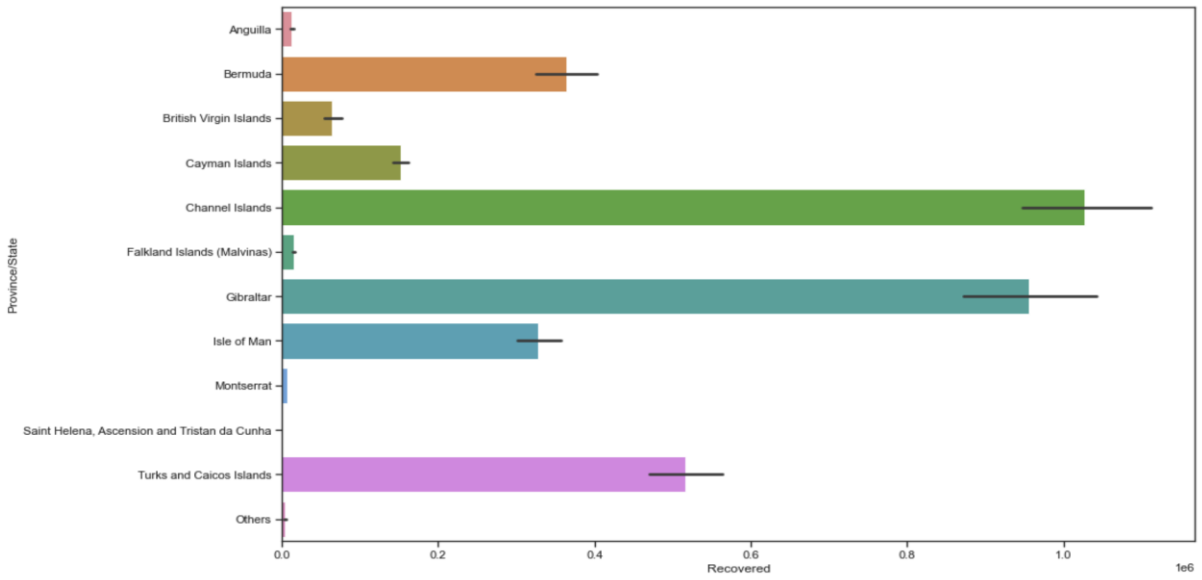
We see at a first glance the percentage of the first dose to fully vaccinated individuals and spot the countries where you need to concentrate the marketing campaign.

As mentioned earlier, the "Others" group causes the data to skew therefore I removed it before creating the following visualisation about the total deaths:

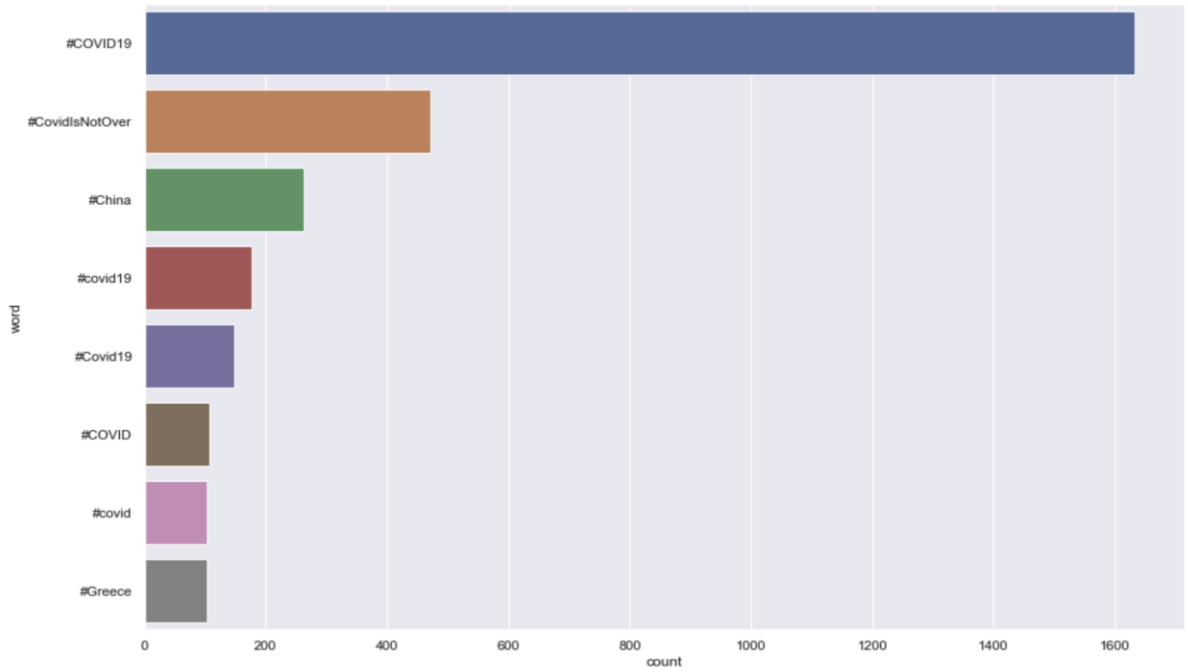


Here I converted the Date into Months to have a more granular view of the data. We are dealing with cumulative cases and notice a jump after January 2021.

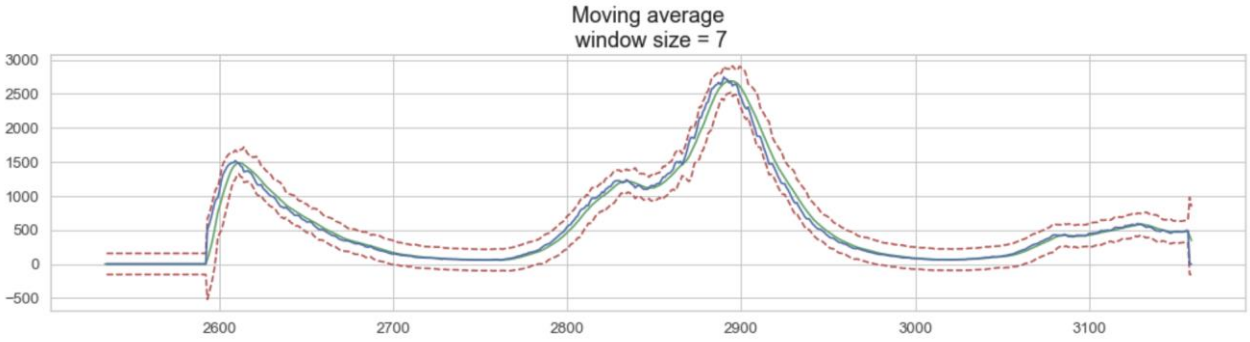
The region that has had the most recoveries and has this been consistent over time has been the Channell Islands therefore you shouldn't concentrate the campaign there.



I have scraped the Twitter data provided and conveyed the information to see the most popular hashtags:



Finally, I performed a time-series analysis on the Channell Islands region to forecast the hospitalisation rates so that hospitals can be prepared for upcoming surges.



Thanks to the last 3 days results for the biggest difference between the daily value and the rolling 7-day mean we can perform a predictive analysis and be prepared to face emergencies.

Additional questions:

Quantitative and qualitative (or categorical) data are both employed in analytics: dimensions (such as region, product category, order priority, etc.) contain qualitative or categorical data, while measures contain quantitative data (e.g., sales volumes, population numbers, shipping costs etc.). The employment of both data is necessary to provide business results as we need to deal with discreet and continuous values, for example continuous data is measured (not counted).

A word about data ethics and continuous improvement: both need to be taken inot account to avoid bias or at least we can exercise damage control. Every change we make is going to have an impact on people's life therefore it is mandatory to look at the bigger picture and question the very basic points we are exploring, by applying critical thinking on a personal and organisational level while also nurturing continuous improvement.

Lastly, to solve the data quality issue it is sufficient to write 2 lines of code to import, read, and filter the data:

```
covid_who = pd.read_csv('https://covid19.who.int/WHO-COVID-19-global-data.csv').

 vaccination_who = pd.read_csv('https://covid19.who.int/who-data/vaccination-data.csv').
```