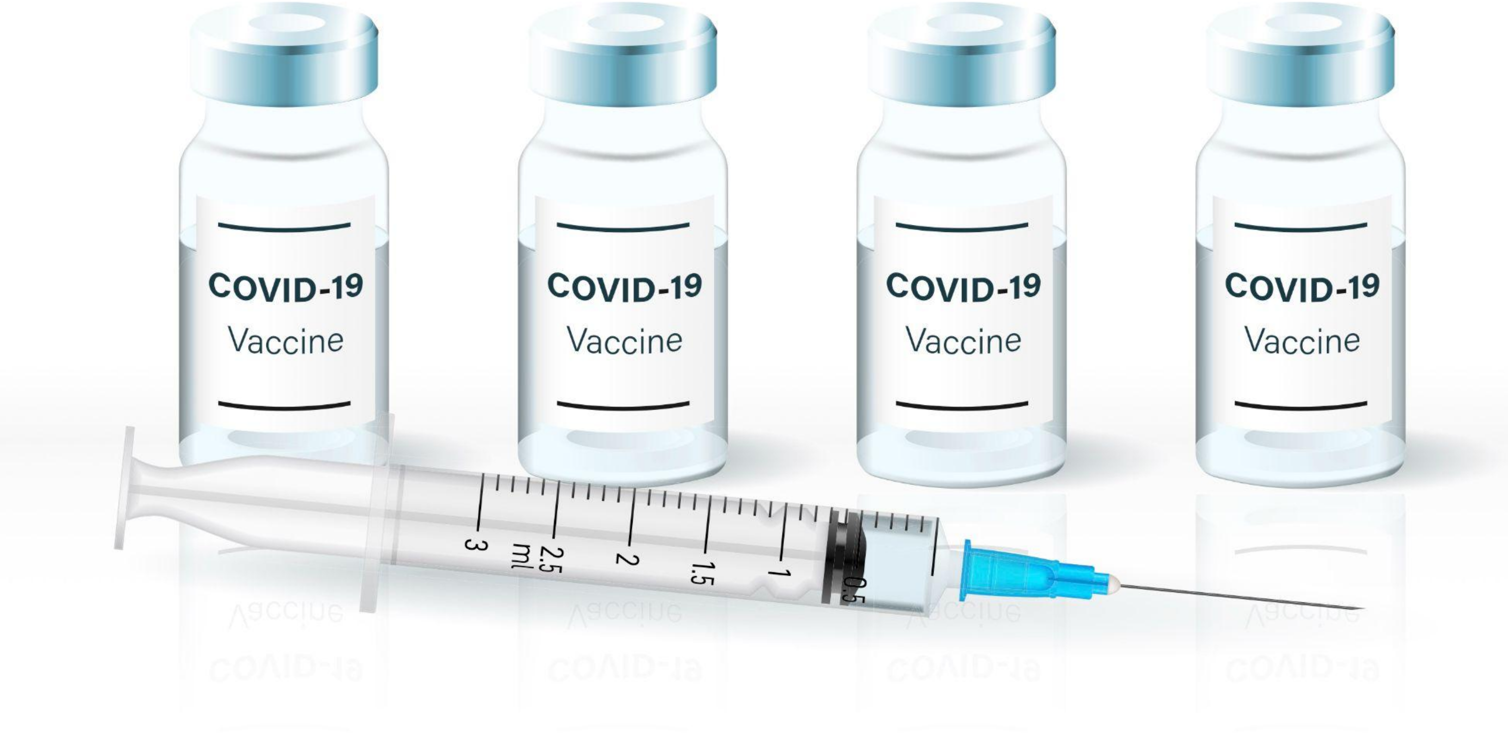


Covid-19 vaccination patterns and trends

Data analysis to inform the campaign to increase vaccinations

Report by Nikki Stopford

Background and context

Problem statement

The UK government wants to launch a campaign to increase the number of fully vaccinated people - these are people who have received both their first and second doses of the vaccine. This report aims to provide insight to help focus the UK government's marketing efforts to increase Covid-19 vaccination rates.

Questions

To effectively inform this work, the UK government team would like to understand the following:

1. What are the total vaccinations received by a region and how trends have developed over time?
2. Where should marketing efforts be focused based on identifying:
    a. Areas where there is the biggest variation between those people receiving their first dose but not their second dose.
    b. Areas where there have been the biggest volumes of recoveries so these can be avoided in initial marketing campaigns.
    c. Whether deaths have been increasing across all regions over time or if it looks like peaks have been reached.
3. What trends can be identified using unstructured Twitter data - specifically data with #coronavirus and #vaccinated hashtags?
4. Which regions have received a peak in hospitalisation numbers and which haven't yet reached this peak?

Approach

Python was used to analyse the data and the analysis was done using Jupyter Notebook - a free, open-source tool conducting analysis, and storing code and notes. This report, the Jupyter Notebook and the data can be accessed via the Github repository set up for this project.

GitHub is a software tool used to manage the source code for analysis. It is a stable and well-supported Version Control System (VCS); making it a reliable place storing all the data and the code for the analysis to enable others in the team to effectively collaborate on this project and help us track changes.

Data exploration and cleaning

I started by importing and loading the Covid-19 'cases' and 'vaccinated' .csv data files into Jupyter Notebook and started exploring the data by viewing the basic shape and data types and viewing the head and tail of the data within the DataFrames.

Based on the early observations, I identified that the 'Date' column in both data files was recorded as an object data type so I imported the DataTime module to manipulate the Date class.

There are missing values in two rows of the data in the 'cases' data file - these are both in the Bermuda Province/State. Where values are missing across four columns (deaths, cases, recovered and hospitalised). There is no missing data in the 'vaccinated' file.

The minimum and maximum dates in the two datasets are aligned - the first record of data collection is 22 January 2020 and the last record is 14 October 2021.

Data analysis

More in-depth exploration of the Gibraltar region was conducted. Data for Gibraltar 'cases' and 'vaccinated' was filtered. The 'case' data was subset to view only data for deaths, cases, recovered and hospitalised. The 'vaccinated' data were subset to view only data for the fully vaccinated and first and second dose data.

The data was aggregated and descriptive statistics were generated, and the minimum and maximum dates were viewed. The observations from this in-depth exploration are shown in the Results section for question 1.

As the next step for analysis, it would be useful to plot the distribution of the daily data by region for the volume of cases, deaths, recoveries, hospitalisations and vaccinations. This will better help us visualise trends and patterns in the data.

It would also be useful to look at the data in terms of cumulative cases and vaccination uptake too, which would again give us an indication of whether cases or vaccine uptake are plateauing.

Merging and analysing the data

Before this can be done, we will need to combine the two DataFrames. The merge() function was used to join the DateFrames using a left outer join. I could have used a full outer join, based on the initial data exploration, which identified the same start and end dates and the same 7,584 rows of data. However, I decided on a left join to the 'cases' DataFrame to ensure all vaccination data were matched to the case collection data.

Results

Question 1: What are the total vaccinations received by a region and how trends have developed over time?

Data analysis was conducted on data for the Gibraltar region. The following observations were recorded:

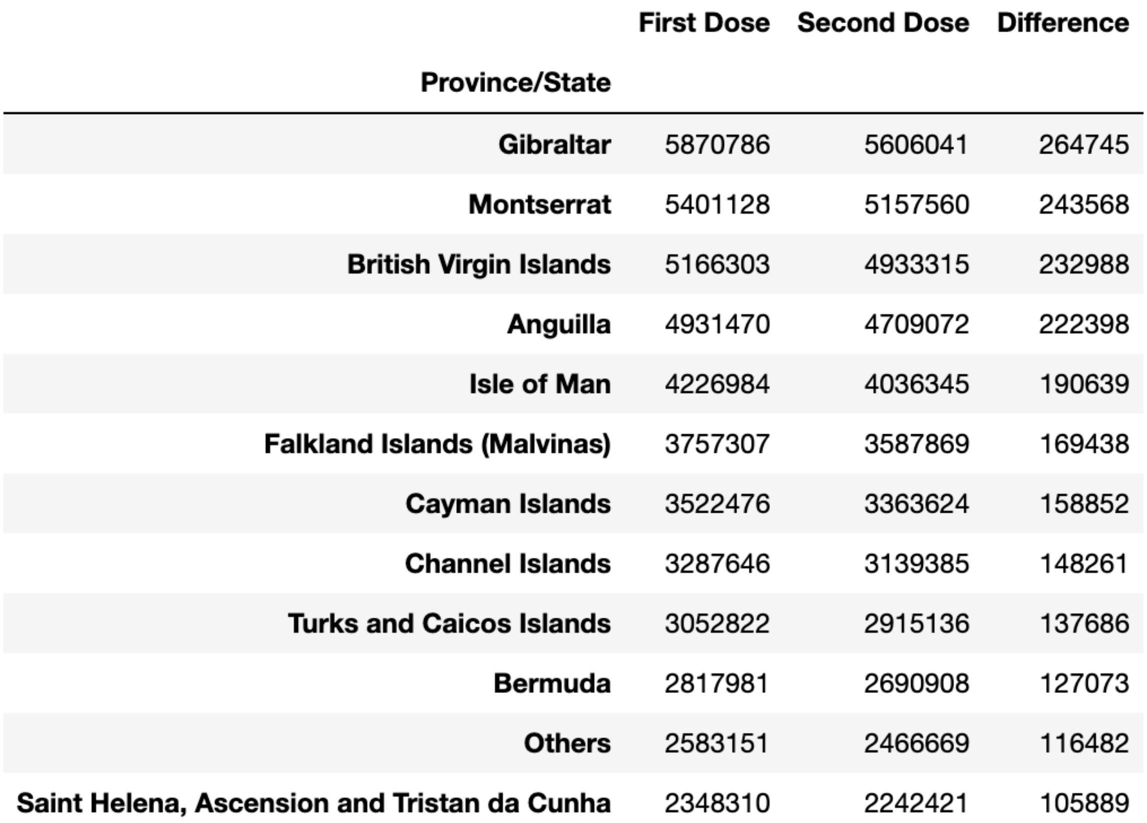
- Data for Gibraltar was collected over 632 days from 22 January 2020 to 14 October 2022 - this looks like it might be consistent with other regions given the minimum and maximum date values recorded for the full data sets.
- Across this time period (632 days), the average (mean) of daily cases is 2,337 in Gibraltar. However, the first case was not reported until 4 March 2020 so the average will be slightly higher if we were to convert the 0 to NAN. I have decided against doing this for the moment so we have a comparative measure across regions for the 632 days of data collection.
- There have been a total of 97 deaths in Gibraltar. The first death reported was on 11 November 2020.
- The maximum number of daily cases was 5,727. This was recorded on the last date of data collection, which would indicate the peak hasn't yet been reached.
- There is missing data in the 'Recovered' column: data collection stops on 5 August 2021. This results in a significant three-month gap in the data. I will need to review whether this is a consistent pattern in other regions (i.e. missing not at random - MNAR). If this is the case then, rather than trying to extrapolate the data based on other data (e.g. cases), I will treat it as unreliable and not use it as part of the analysis.
- In terms of vaccinations, the 'Vaccinated' column shows the number of people who are fully vaccinated (having received two doses). The shape of this data would indicate this is daily data rather than cumulative because it peaks and then reduces over time.
- The total number fully vaccinated and the total number who have received a second dose match (5,607,041), which acts as a good check for data quality.
- The total number of people recorded as fully vaccinated is over 5.6m which, in a real-world situation, I would want to sense check against population data. In 2020, the population of Gibraltar was 33,691[1], significantly less than the number recorded to be fully vaccinated.
- This would raise serious data quality issues but, for the purpose of this assignment, it will be ignored.
- There are 264,745 people who have yet to have a second dose after their first. It isn't clear yet whether this is an acceptable lag, given how long people would need to wait between a first and second dose.

[1] World Bank, 2020 - Data Commons, Gibraltar

Question 2: Where should marketing efforts be focused?

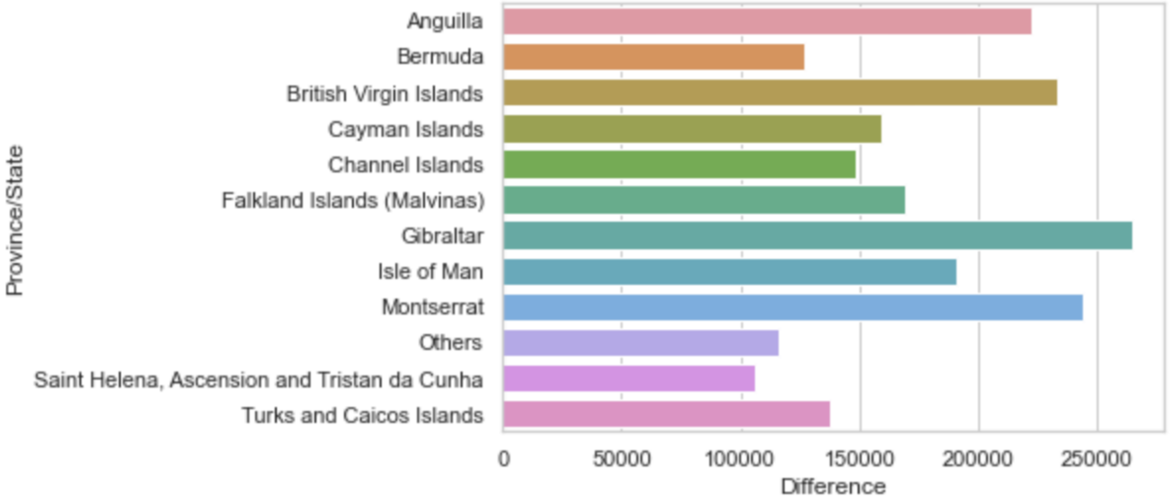
Data were grouped by region and analysed to identify the areas where there is the biggest variation between people receiving their first dose but not their second. As you can see from Table 1(and the associated bar plot 1), below, Gibraltar has the highest difference by volume.

Table 1 - variation by region of people who have had their first vaccination but not their second.



	First Dose	Second Dose	Difference
Province/State			
Gibraltar	5870786	5606041	264745
Montserrat	5401128	5157560	243568
British Virgin Islands	5166303	4933315	232988
Anguilla	4931470	4709072	222398
Isle of Man	4226984	4036345	190639
Falkland Islands (Malvinas)	3757307	3587869	169438
Cayman Islands	3522476	3363624	158852
Channel Islands	3287646	3139385	148261
Turks and Caicos Islands	3052822	2915136	137686
Bermuda	2817981	2690908	127073
Others	2583151	2466669	116482
Saint Helena, Ascension and Tristan da Cunha	2348310	2242421	105889

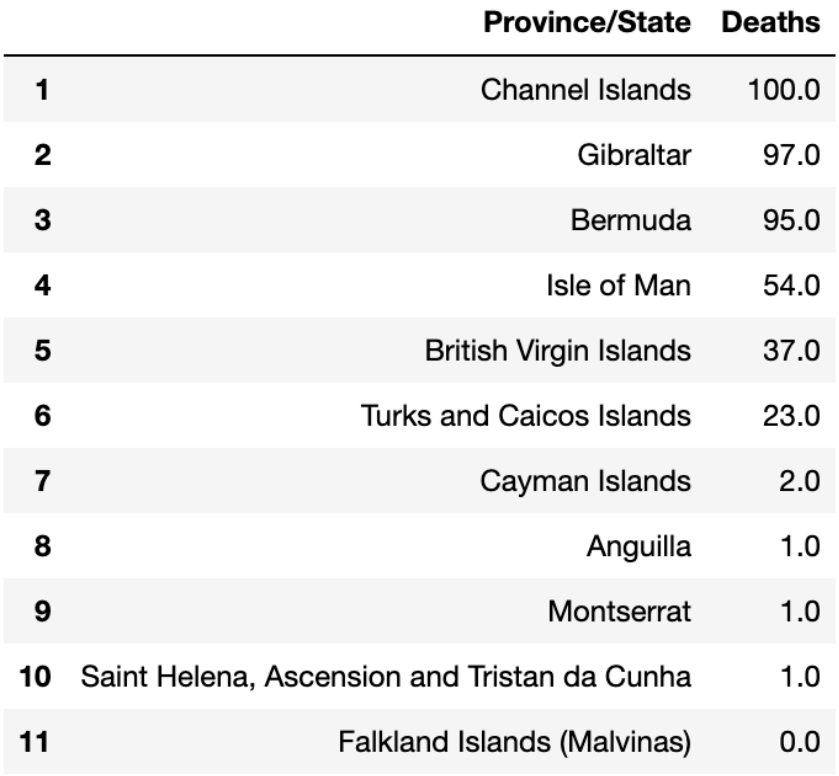
Plot 1 - variation by region of people who have had their first vaccination but not their second.



However, as a percentage difference by region, there is little very little variation between those having had their first dose and the second dose - see Plot 1.

The regions with the highest volumes of deaths are the Channel Islands, Gibraltar and Bermuda - all of which have had over 90 deaths based on the cumulative data captured. See Table 2.

Table 2 - cumulative deaths by region.

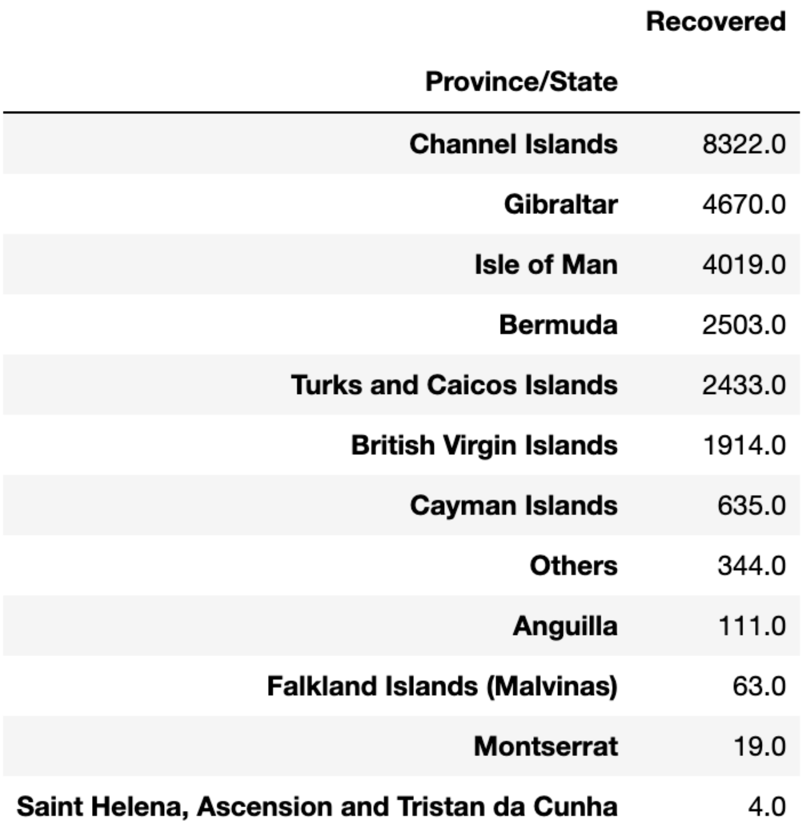


	Province/State	Deaths
1	Channel Islands	100.0
2	Gibraltar	97.0
3	Bermuda	95.0
4	Isle of Man	54.0
5	British Virgin Islands	37.0
6	Turks and Caicos Islands	23.0
7	Cayman Islands	2.0
8	Anguilla	1.0
9	Montserrat	1.0
10	Saint Helena, Ascension and Tristan da Cunha	1.0
11	Falkland Islands (Malvinas)	0.0

Visualising this data via a simple lineplot has been difficult because the 'Others' category skewed the data too much and I haven't been able to filter this out in the plot. This would need further exploration to determine whether peaks are plateuring over time.

Although the Channel Islands has had the highest number of deaths, it has also had the highest number of recoveries - see Table 3. Though, as noted above, three months of data is missing from 5 August 2021 through to the end data collection point so we should treat this insight with caution.

Table 3 - cumulative recoveries by region.



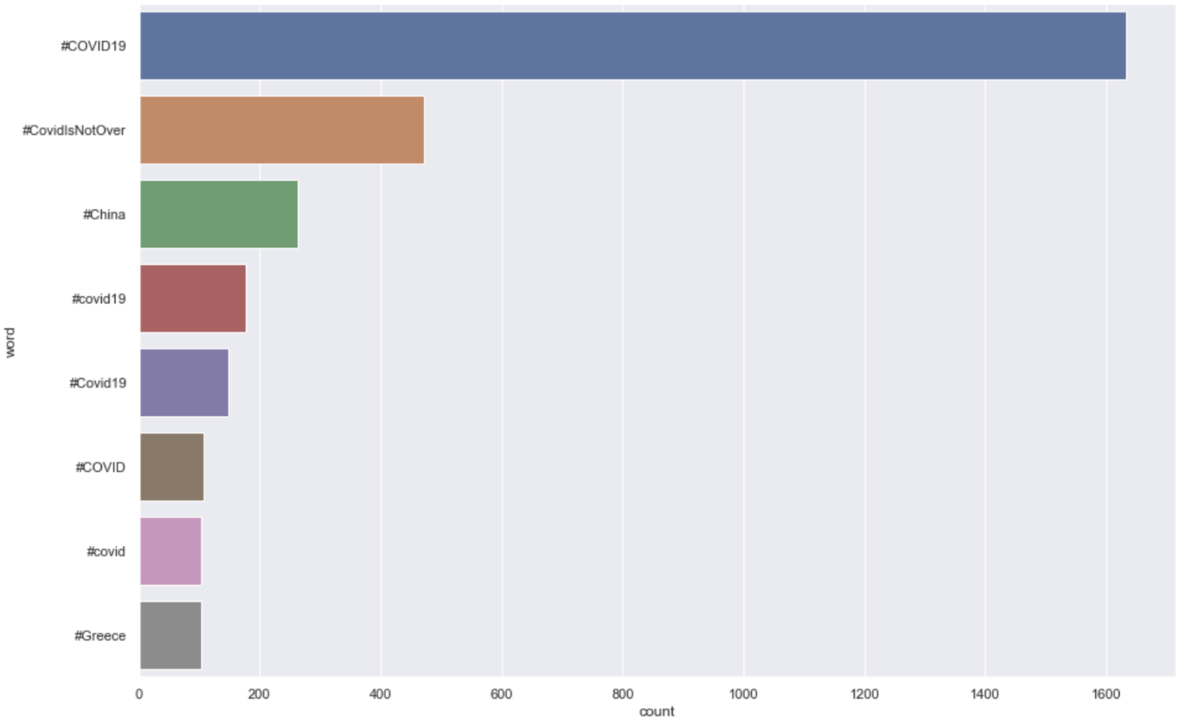
	Recovered
Province/State	
Channel Islands	8322.0
Gibraltar	4670.0
Isle of Man	4019.0
Bermuda	2503.0
Turks and Caicos Islands	2433.0
British Virgin Islands	1914.0
Cayman Islands	635.0
Others	344.0
Anguilla	111.0
Falkland Islands (Malvinas)	63.0
Montserrat	19.0
Saint Helena, Ascension and Tristan da Cunha	4.0

Question 3: What trends can be identified by analysing unstructured Twitter data?

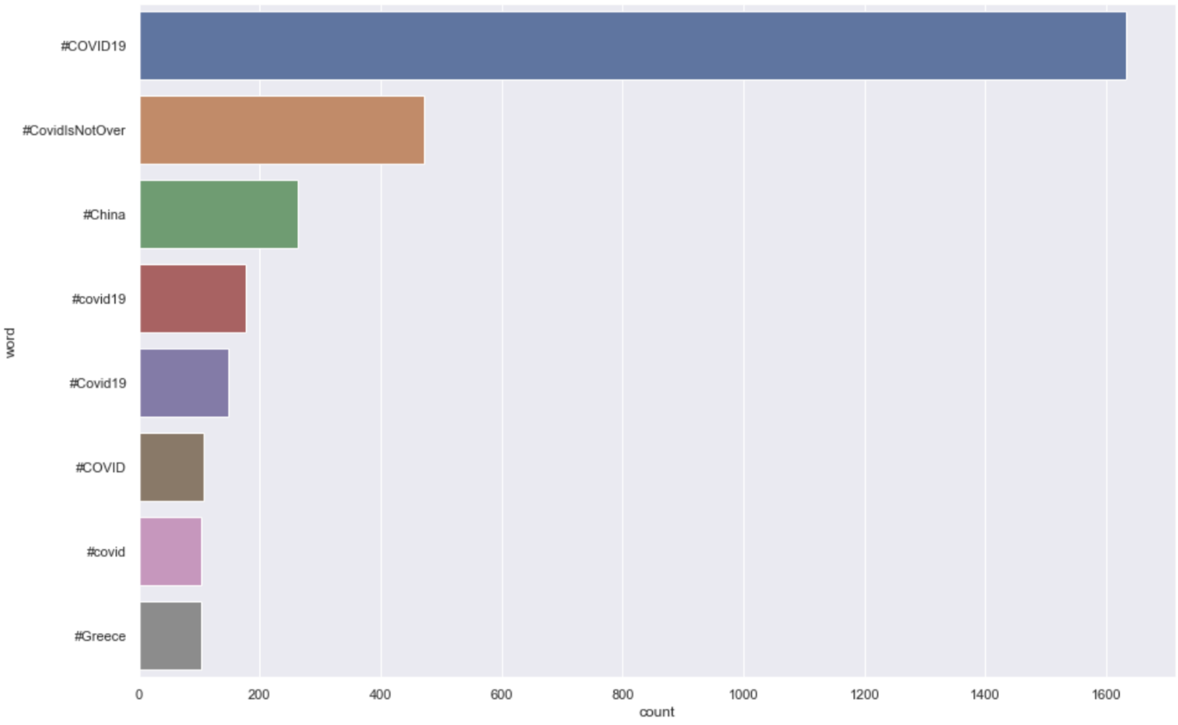
With the support of an expert Twitter analyst on the team we have also looked at Twitter trends in the tweet .csv file. This analysis concluded the dataset was too small to make any meaningful use of retweet and favourite counts.

The most popular hashtag used in tweets was #COVID19, as can be seen in Plot 2. Plot 3 illustrated the most popular keywords used in the text of tweets - and would indicate that there was likely to be something interesting going on in Greece at the time the data was extracted.

Plot 2 - popular Twitter hashtags



Plot 3 - popular Twitter keywords

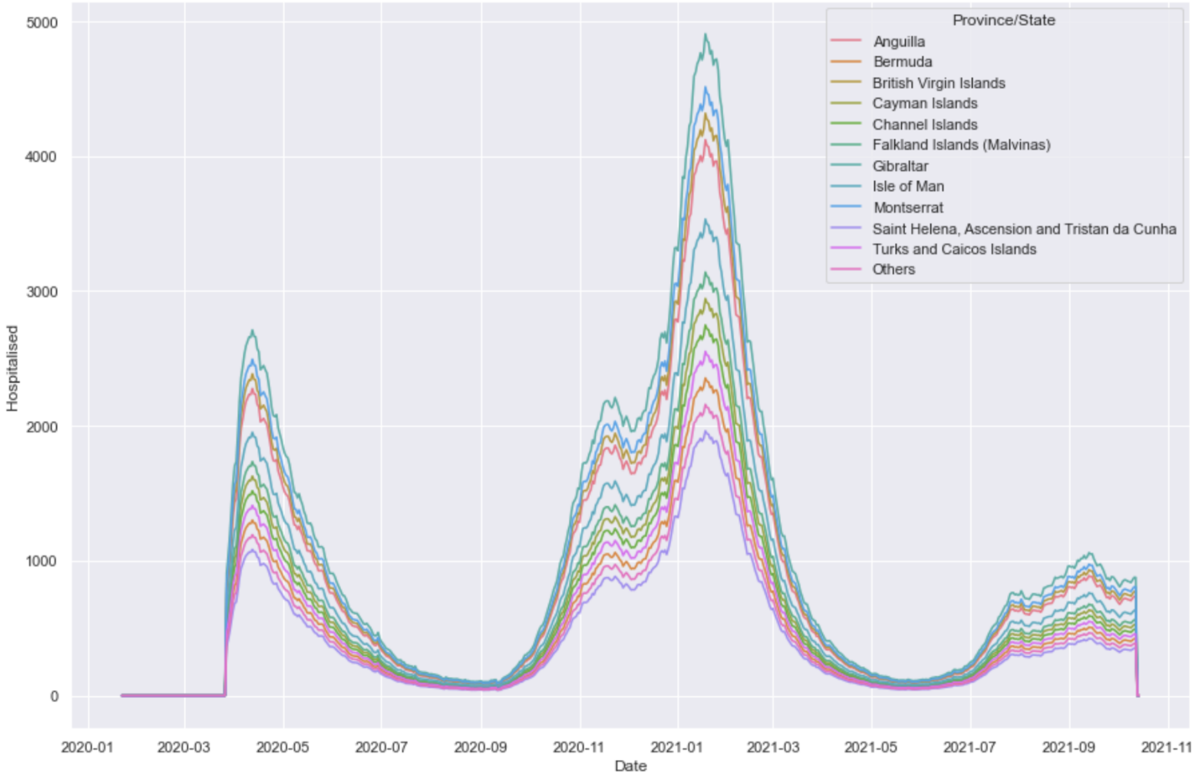


This type of data could be very useful for extracting real-time trends and in bringing to life experiences on the ground of people talking about Covid-19.

Question 4: Which regions have received a peak in hospitalisation numbers and which haven't yet reached this peak?

All regions have had a peak in hospitalisations as can be seen from Plot 4. The largest of these peaks for all regions was between January and March 2021.

Plot 4 - hospitalisation trends by region

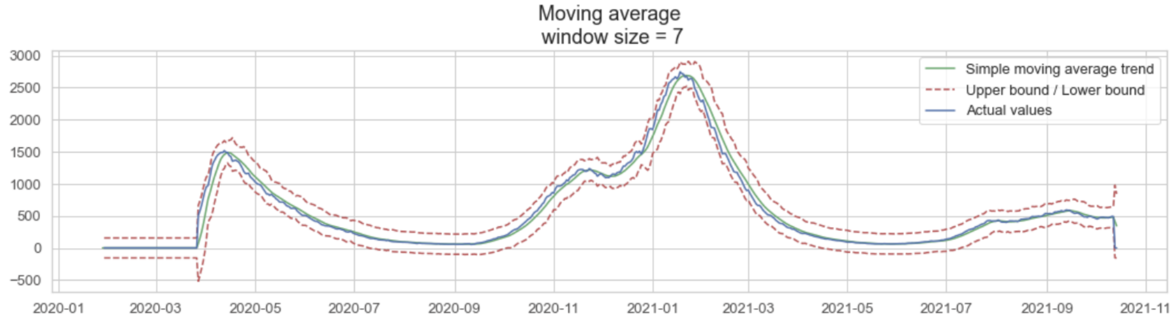


To help ensure hospitals are prepared for any upcoming surges in hospitalisations, analysis was conducted on data for the Channel Islands to forecast hospitalisation rates. This was done using a simple moving average method to capture the average change in the data series over time and to eliminate the effects of seasonal irregularities.

Plot 5 visualises the change over time using a 7 day window. This data would indicate that peak hospitalisation in the Channel Islands has been passed and there is a broad plateuing of hospitalisations.

Additional questions asked by the management team can be found in the file called LSE_DA201_Week_6_assignment_notebook_Nikki_Stopford.

Plot 5 - time-series analysis of hospitalisation rates for the Channel Islands



Conclusions and recommendations

The quality of the data is questionable given the inconsistency identified between the volumes of vaccinations received and external population data. As a next step, I would recommend this data is compared to other public Covid-19 data sets.

Based on the analysis of this data, I would recommend that marketing efforts should be focused on Gibraltar. The percentage variation between people who have had the first vaccination but not their second is similar by region. However, Gibraltar has the highest volume of people who have had their first vaccination but not their second - 264, 745 in total.