United Kingdom Covid Data Jan 2020 – Oct 2021

Background & Context

The UK government requires analysis of Covid data as it is planning to launch a series of marketing campaigns to promote uptake of the vaccine by identifying trends in the data to inform the approach, such as;

- What the total vaccinations (partial, full, by area and over time).
- Where they should target the first marketing campaign(s) based on:
    - area(s) with the largest number of people who are partially vaccinated.
    - which area has the greatest number of recoveries so it can be avoided
    - whether deaths have been increasing across all regions over time or if a peak has been reached.
- Analyse Tweets to show most popular hashtags and review data points associated with #coronavirus and #vaccinated hashtags.
- Which regions have experienced a peak in hospitalisation numbers and if there are regions that have not reached a peak yet.

Analytical Approach and Data Limitations

Much of the analytical approach is documented in the accompanying Jupyter notebook and can be followed there, however some key elements will be highlighted in this report

Assumptions/Understanding of Data

Where not clear in the data dictionary, though EDA the following assumptions are made;

- Cases, death and recovery data are cumulative, hospitalisations is patients in hospital each day, vaccination data is daily.
- Province/State looks to be British Overseas territories, with the Province/State "Other" being mainland UK – assumed this based on the large cases numbers and lat/long coordinates pointing to Scotland.
- Anyone who is counted as receiving a second dose of vaccine has been assumed to have received a first and therefore be fully vaccinated.
- As part of EDA, box plot visualisations were created for each Province/State to see the shape of the key data points (cases, deaths, recovered, hospitalised, first dose vaccine, second dose vaccine) (see relevant png files on GitHub or within the Jupyter notebook). Key points of this work are.
    - The data for most of the provinces was too small to give any meaningful insight to outliers
    - Whilst there are outliers in some larger provinces (Others, Channel Islands, Gibraltar etc.) for cases, deaths, from what we know of these data points in a pandemic (ie exponential growth) it will not follow a normal pattern and there is possible for a large range in the data. As a result, I was happy the data looked reasonable and performed no cleaning with regards to this.
    - An interesting discovery was made on hospitalised and vaccine data – all Provinces and States had the same shape and similar numbers which looked odd.
    - Recovery data had many outliers and numbers looked low in relation to cases in every Province/State. On further investigation the data also ends in August 2021. It does not look of good quality
    - The twitter data provided was used to look for the most popular hashtags and the most popular words used within tweets. The data set was 3,959 tweets so a reasonably small sample size. The tweets were all taken from May 15-23, 2022, so after all of the covid data previously used. There was no indication in the data as to tweet location

Data Limitations and Errors

Following the box plots highlighting some strangeness in the vaccine and hospitalisation data further analysis into the data showed that the values looked incorrect. I found an external data source for these data points[1] (excluding hospitalisations), downloaded the csv and worked in Excel to review. Bringing in population data and normalizing key data points I could show that this data provided did indeed look wrong by Province/State– eg. Gibraltar normalized population is 0.0004, cases & deaths both 0.0007 but hospitalization & vaccination is 1.0. The externally sourced data looked much more sensible ash show below.

[1] https://ourworldindata.org/explorers/coronavirus-data-explorer?zoomToSelection=true&time=earliest..2021-10-14&facet=none&uniformYAxis=0&pickerSort=desc&pickerMetric=total_cases&Metric=People+vaccinated+%28by+dose%29&Interval=Cumulative&Relative+to+Population=false&Color+by+test+positivity=false&country=GIB~AIA~GBR~VGB~CYM~FLK~MSR~SHN~TCA~IMN

	Provided Data						From Our World in Data				
	Normalised Pop	Normalised Cases	Normalised Max in Hosp	Normalised Deaths	Normalised First Dose	Normalised Second Dose	Normalised Pop	Normalised Cases	Normalised Deaths	Normalised First Dose	Normalised Second Dose
Anguilla	0.0002	0.0001	0.7334	0.0000	0.7333	0.7333	0.0002	0.0001	0.0000	0.0002	0.0002
Bermuda	0.0009	0.0007	0.1332	0.0007	0.1333	0.1333	0.0009	0.0007	0.0007	0.0009	0.0009
British Virgin Islands	0.0004	0.0003	0.7999	0.0003	0.8000	0.8000	0.0004	0.0003	0.0003	0.0003	0.0003
Cayman Islands	0.0009	0.0001	0.3332	0.0000	0.3333	0.3333	0.0009	0.0001	0.0000	0.0011	0.0011
Channel Islands	0.0024	0.0015	0.2666	0.0007	0.2667	0.2667	0.0024	0.0000	0.0000	0.0022	0.0019
Falkland Islands (Malvinas)	0.0000	0.0000	0.3998	0.0000	0.4000	0.4000	0.0000	0.0000	0.0000	0.0000	0.0000
Gibraltar	0.0004	0.0007	1.0000	0.0007	1.0000	1.0000	0.0004	0.0007	0.0007	0.0008	0.0008
Isle of Man	0.0012	0.0010	0.5333	0.0004	0.5333	0.5333	0.0012	0.0010	0.0004	0.0013	0.0014
Montserrat	0.0000	0.0000	0.8665	0.0000	0.8667	0.8667	0.0000	0.0000	0.0000	0.0000	0.0000
Saint Helena, Ascension and Tristan da Cunha	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
Turks and Caicos Islands	0.0005	0.0003	0.2001	0.0002	0.2000	0.2000	0.0005	0.0003	0.0002	0.0005	0.0005
Others	1.0000	1.0000	0.0666	1.0000	0.0667	0.0667	1.0000	1.0000	1.0000	1.0000	1.0000

Therefore, additional questions I would ask would focus on this data and its veracity if recommendations are to be made based on it. I would go back to the government and ask them to double check the data and reissue if necessary.

For the purposes of this report, I am using the data as provided.

Visualisation Design

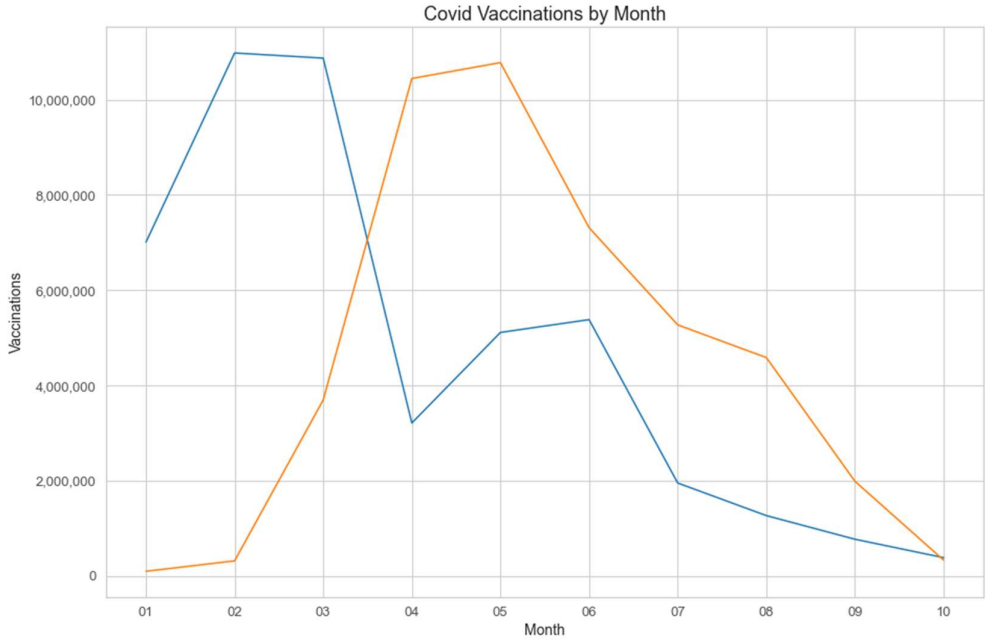
For the visualisations I have tried to keep a standard size and layout where possible, keeping them clean and easy to read using a white background, titles and legends where applicable. There are limitations to the visualisations that can be done in Python (mainly around formatting and interactivity) so I would suggest for further explanatory visualisdations do be done in bespoke software such as Tableau.

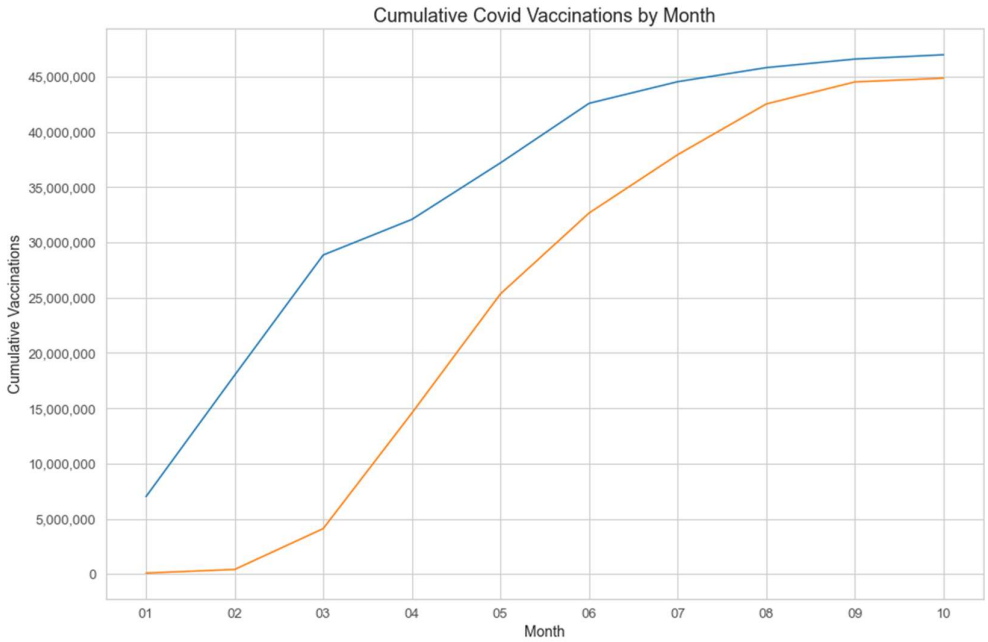
Patterns, Trends, Insights and Recommendations

Vaccinations - Total

Total vaccinations currently stand at 47m people having received the first dose and 44.8m people single vaccinated. The take up over time is shown in fig.1. The first month of vaccination data is Jan 2021, with sharp growth in first doses meaning around 30m people were vaccinated within 3 months, from mid-way through March second doses started to be given out more widely (delay v first dose mostly down to not having many eligible as would generally be 12 weeks after first dose) and showed a similar quick roll out so that by October 95% of those who have had a first dose are now double vaccinated.

fig.1





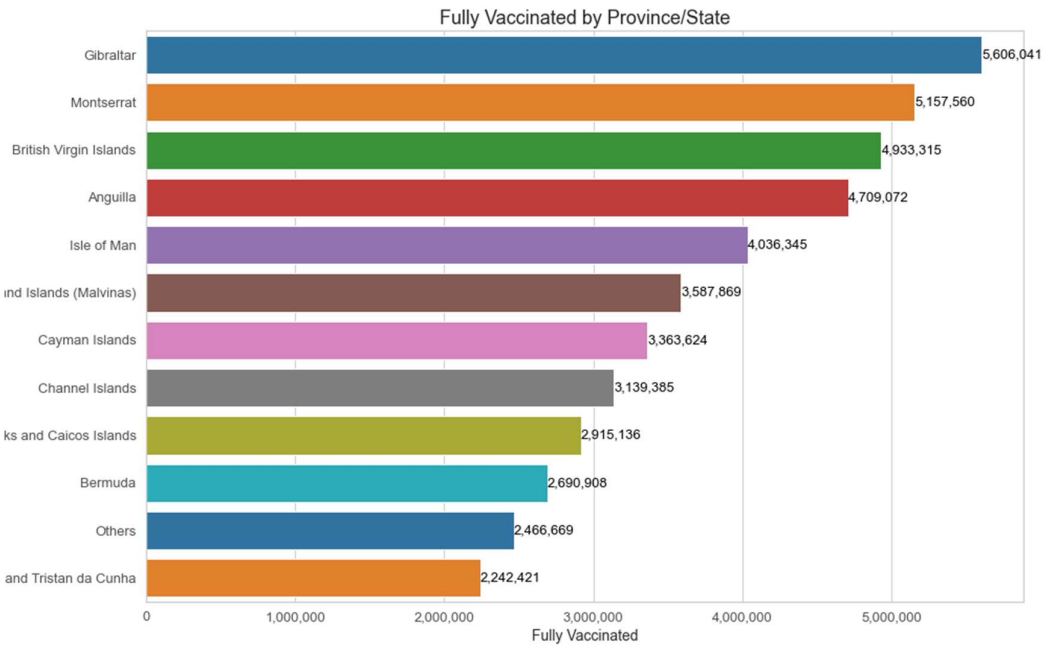
Overall take up of the vaccine looks positive with a slight plateau now as the number of the population who haven't been vaccinated falls. There are still 2m people to target to get their second vaccination

Case, Vaccine, Death and Recovery Data by Province/State

Vaccinations

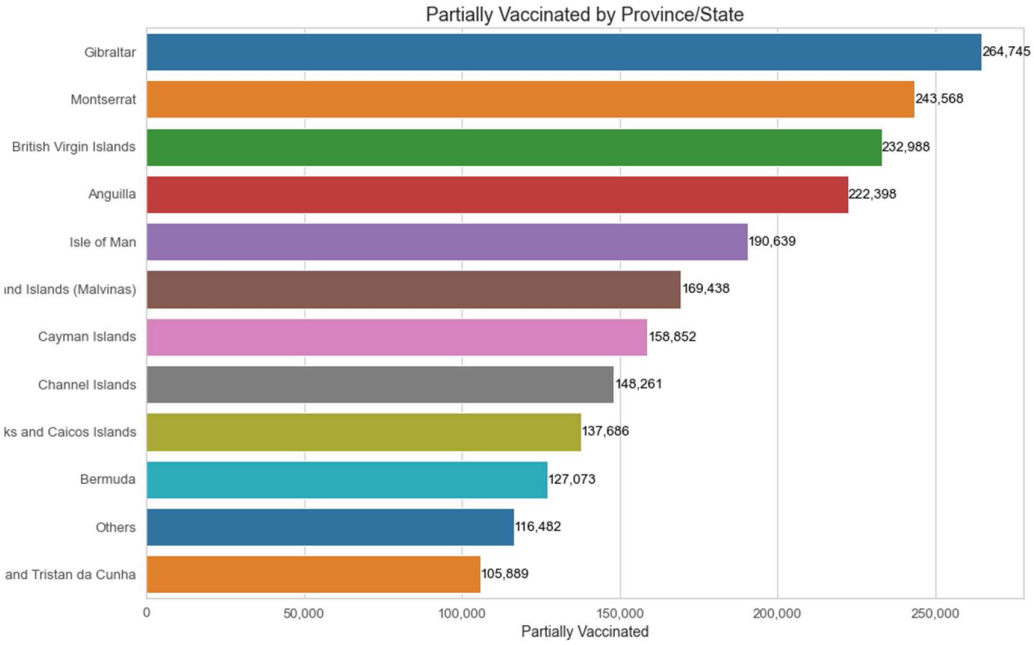
The Province/State breakdown is interesting – Gibraltar seems to have given the most full vaccines at 5.6m (fig.2). Assuming this to be true, with a population of about 34,000, it would mean each person has been vaccinated 164 times. Another indicator that there has been something go wrong with the Province/State split of the data.

fig.2



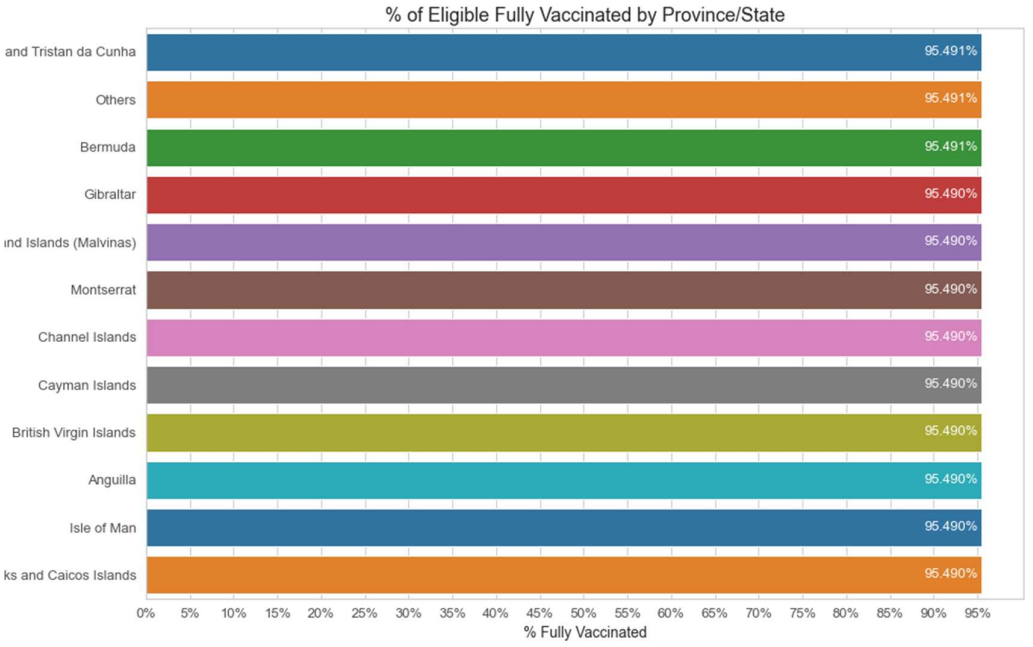
Partially vaccinated numbers follow a similar pattern from a Province/State point of view. Indications are that Gibraltar has the largest number of people who have had a first dose but no second dose with 265,000 (fig.3).

fig.3



When looking at the Province/State data in terms of % of people fully vaccinated (ie eligible for a second dose and have had a second dose) then across the board it is nearly identical (fig.4)

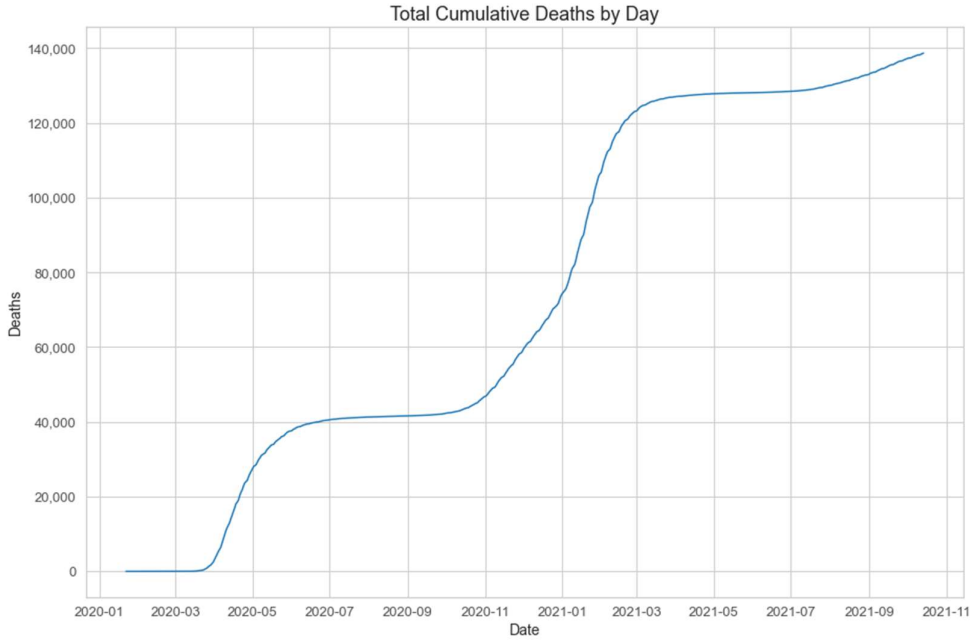
fig.4



Deaths

Total deaths increase over. There are distinct patterns in the death data (fig.5) so far with two separate peaks around the first and second waves of the pandemic (Apr-Jul 2020 and Nov 2020-Mar 2021)

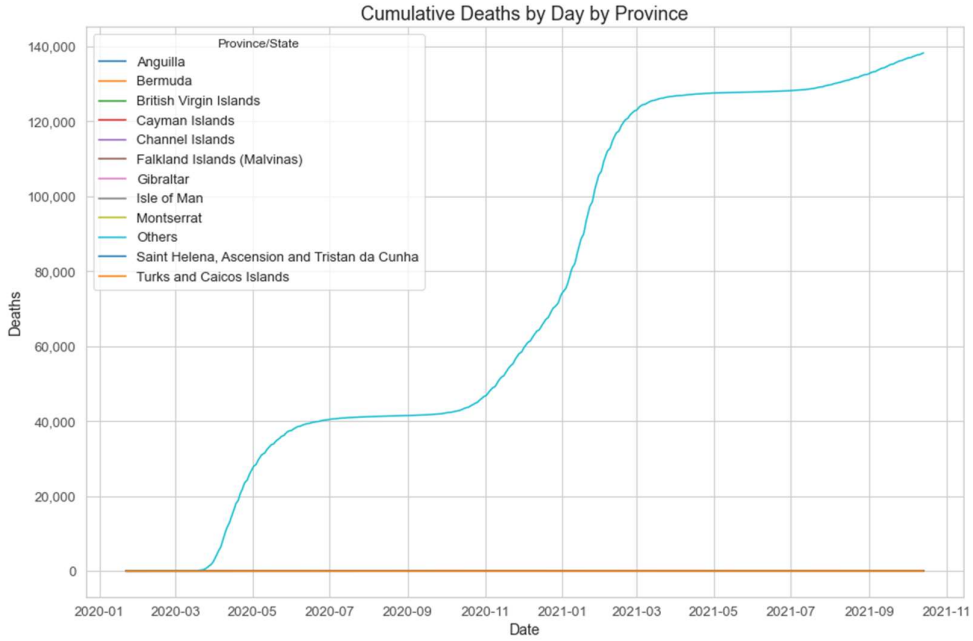
fig.5



Looking at historical data, and the slight uptick we are seeing at the end of the current dataset, it would indicate that we could expect to see another peak of deaths as we approach winter 2021 and a peak has not been reached.

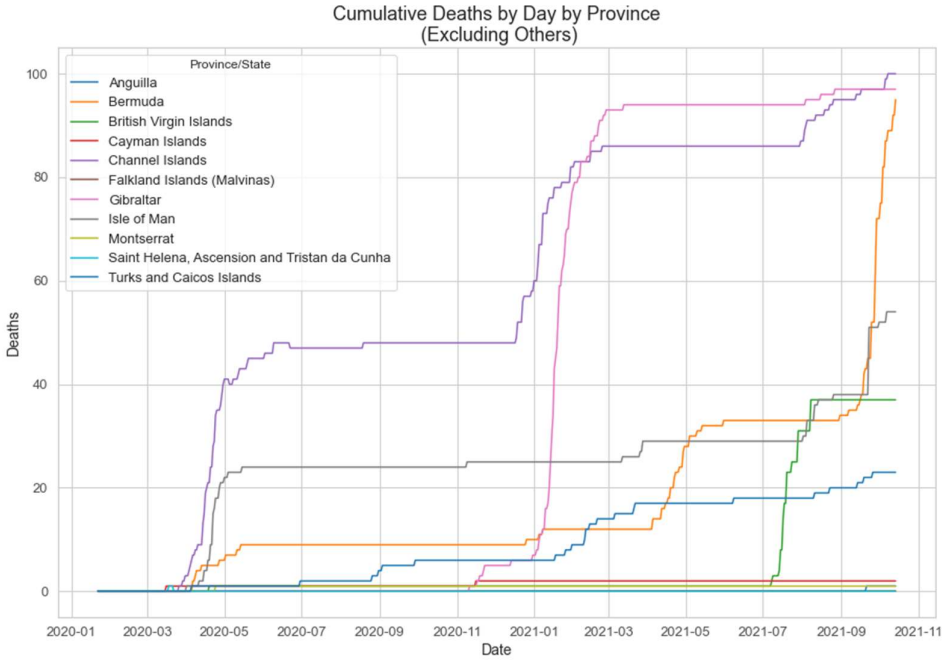
Splitting the death data by Province/State we see the same pattern, heaving influenced by the “Others” (mainland UK) Province/State which has an overwhelming majority of the deaths (fig.6).

fig.6



Removing others there are some interesting trends that become visible in the remaining Province/State fields (fig.7)

fig.7



Deaths in the Channel Islands mirror that of mainland UK the closest, Gibraltar had a large peak in the middle but looks to have now flattened, Anguilla has had the steadiest, most consistent rise over time. The areas I would most want to focus on are Bermuda, Isle of Man and British Virgin Islands – all three have had a recent sharp rise in Deaths.

Cases are the main leading indicator of deaths, and I would say out of the three areas Bermuda and Isle of Man look of most concern (fig.8) with cases yet to peak so more deaths to follow. The government may want to focus the vaccine push in those areas to look to control that rise.

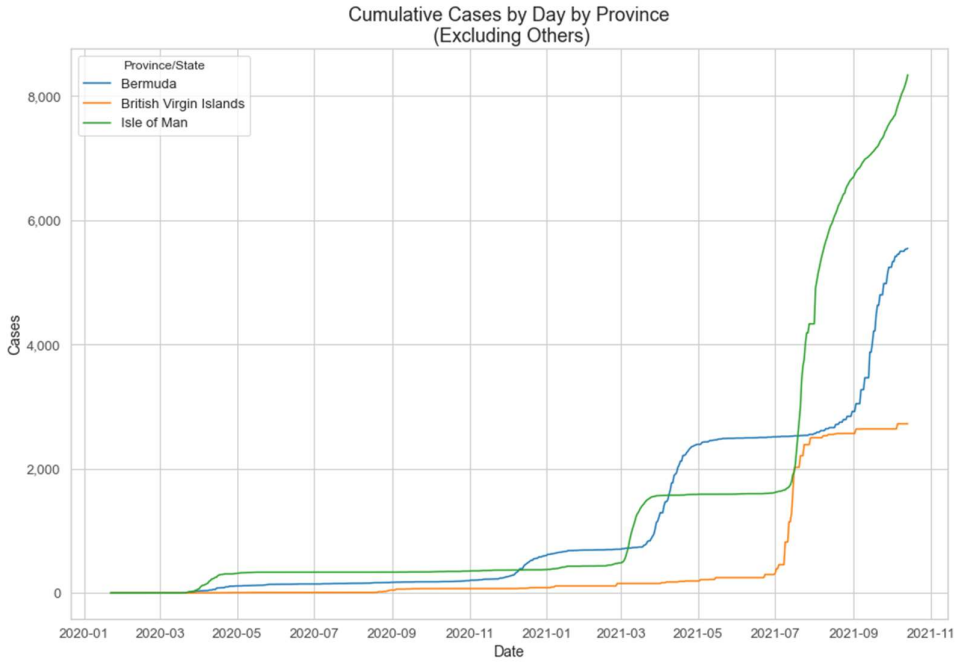


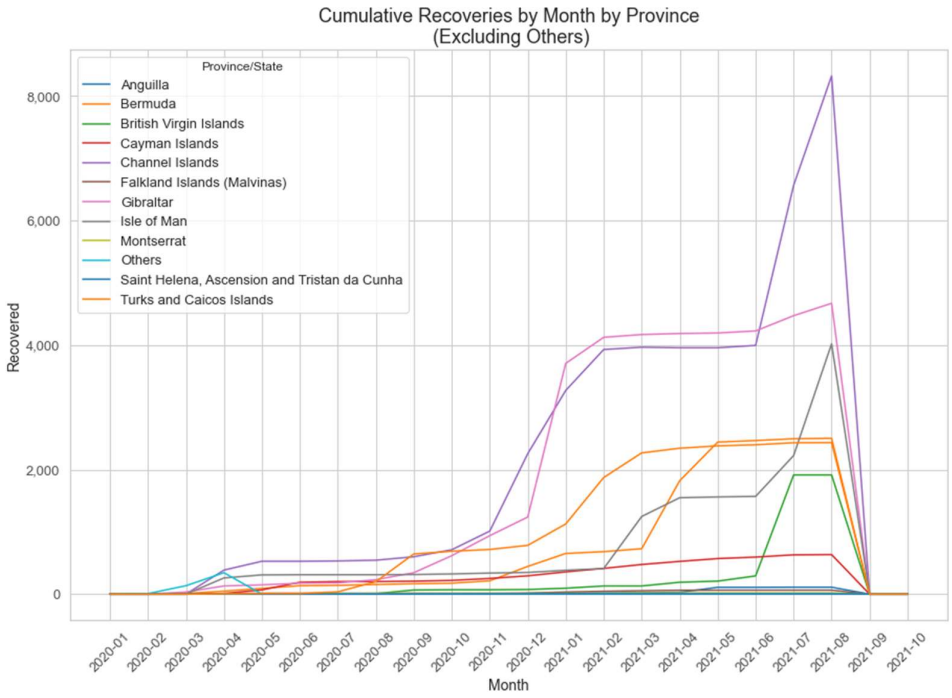
fig.8

Recoveries

The recovery data as mentioned previously looks to have both errors in it (numbers look very low - 25,000 total recovered 8.3m cases with 139,000 deaths suggests some 8.1m live cases) and is incomplete (the data stops on 4 Aug 2021). I would therefore not rely on it to make any decisions around where to target the campaign for vaccination.

However, from what we can see (fig.9) most recoveries are in the Channel Islands, Gibraltar and Isle of Man and that would indicate these areas would be avoided for the campaign. Interestingly Bermuda is quite low in recovered which, aligned with the previous data might indicate it as a good area to target.

fig.9

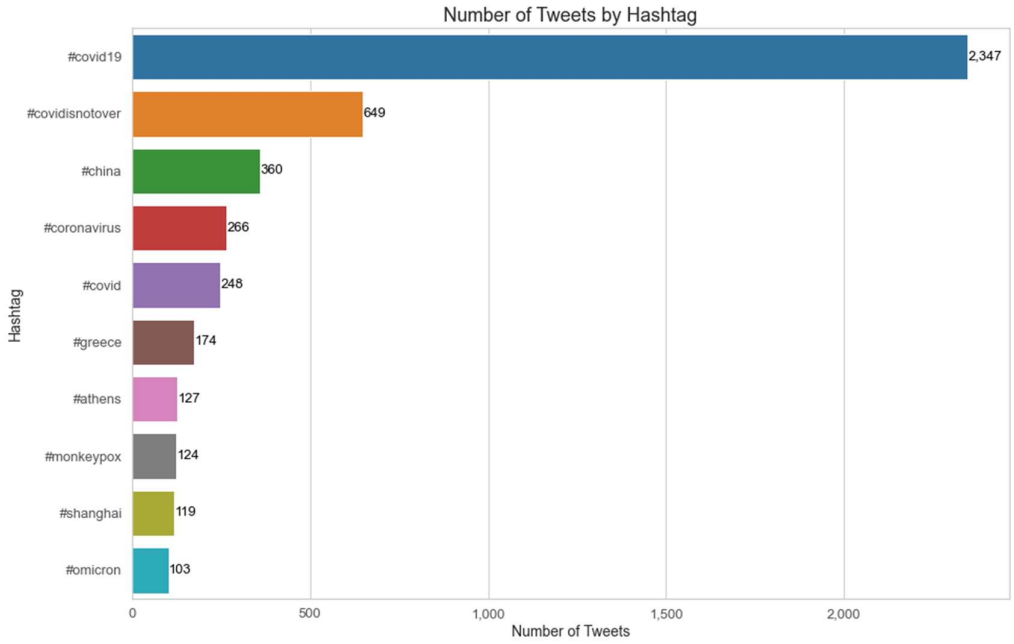


Twitter Data

Most Tweeted Hashtags

The hashtags were separated out of the text of the tweet, and then counts performed on them to find with the most tweets (fig.10) and the list was heavily dominated by Covid – with #covid19 the most tweeted hashtag and #covid and #coronavirus featuring high in the list, showing that the pandemic is still at the forefront of people's minds.

fig.10



A further measure of trending topic on Twitter is looking at retweets and favourites, these can indicate the reach or spread of a tweet (eg. 1 tweet could be shared 10,000 times). Whilst the top two tweets from the previous list are also the most shared and favourited, there are some interesting additions to show what else is trending on Twitter over that time - #misinformation was 4th most retweeted from just 2 tweets and #tedros 9th from 1 tweet. More investigation could be done on these as to why and what message the tweets are conveying.

fig.11

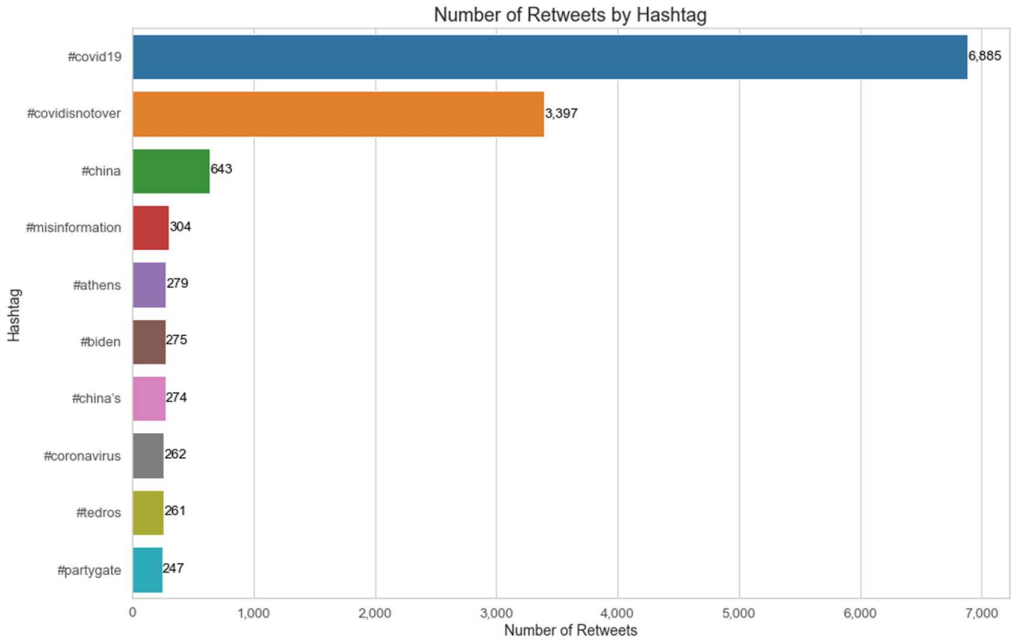
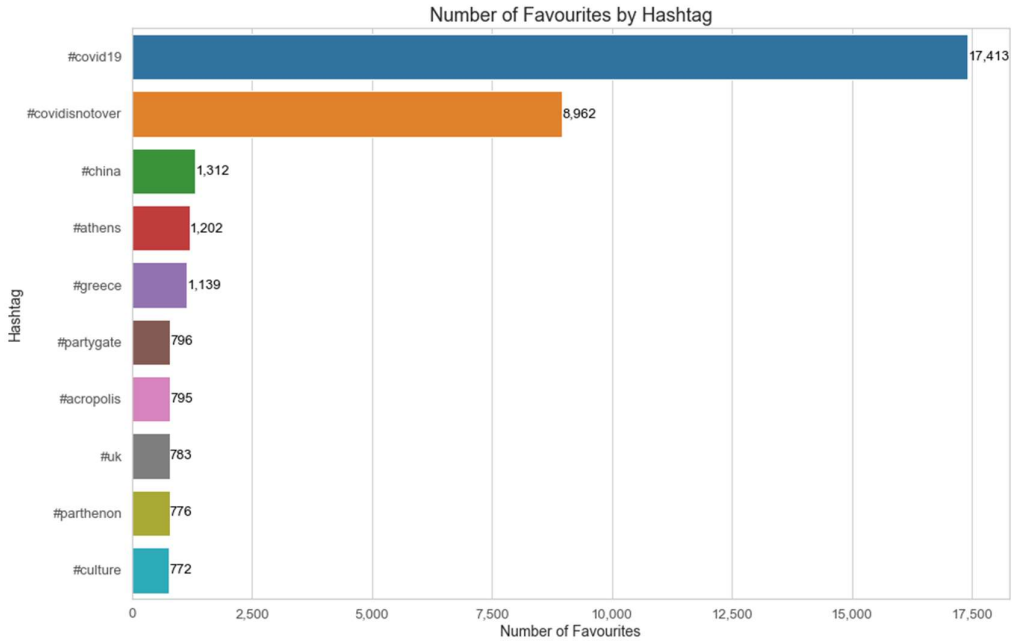


fig.12



#coronavirus and # vaccinations

#coronavirus and #vaccinated do not feature highly on the most tweeted or shared lists

	hashtag	tweet_count	retweet_count1	favourite_count1
142	#vaccinated	9	1	7

	hashtag	tweet_count	retweet_count1	favourite_count1
3	#coronavirus	266	262	424

There are very few tweets under #vaccination in particular – this could indicate that it is not a subject in people's minds and the government might need to work hard to push the message.

In addition to searching hashtags, a search was done on words or phrases excluding hashtags to see if vaccination or associated words might be more part of the conversation if not an explicit hashtag.

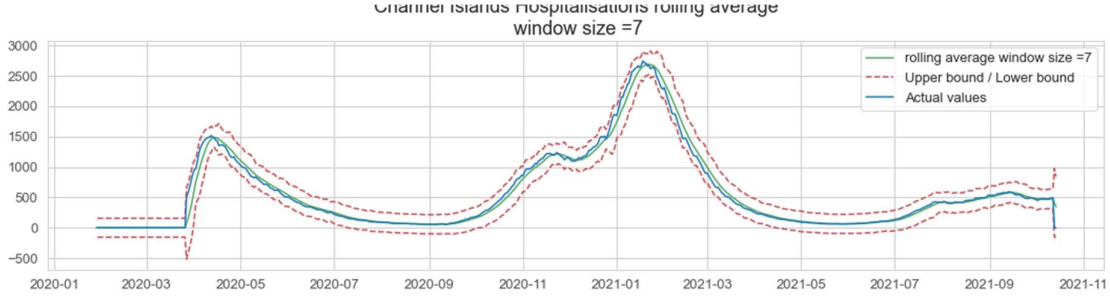
	text	tweet_count
14	vaccine	196
59	vaccines	94
177	vaccination	50
183	vaccinated	50
1153	unvaccinated	9
1566	vaccinations	6
2988	covid-vaccine	3
4789	vaccinate	2

There were a few more tweets with this in the text, 423 tweets in total that would make it in the top 5 most frequent words so evidence that vaccination is being discussed but not used as hashtags.

Hospitalisations – Channel Islands

Specifically looking at Channel Island data we are attempting to ascertain whether hospitalisations have peaked yet and whether they are on the increase or decline (fig.13)

fig.13



The trend over time is similar to the death data with a couple of peaks but looks to have stabilized now for a few months, which could be the impact of vaccinations reducing the severity of symptoms. The rolling weekly average is useful to smooth out day of the week effects (eg. Less reporting at weekends and is generally close to the daily data.

The error rate (fig.14) is quite useful to chart whether cases are on the rise or fall, as we can see below the sharp rises and falls from each peak and confirmation of the current stability of hospitalisations.

fig.14

Conclusions, Recommendations and Further Analysis Required

The government has instigated this project to identify the best areas for a targeted vaccination marketing campaign.

Firstly, based on current data, a solid recommendation cannot be made due to reservations as to the quality of data at a province level – I recommend a thorough review of current data and potentially web scraping of the Our World in Data site could be done to correct this.

There are some useful data points - vaccination data at total level looks fine and shows good take up over time, even if it has begun to tail off. Further research needs to be done as to the reasons for this and action can be taken based off this. I would also want to see more segmentation to the data – age, risk levels of patients etc to see if there is any particular demographic not taking the vaccine.

Case and Death data looks reliable. From this, areas of concern look to be Bermuda and British Virgin Islands that could be targets for a campaign, as they are seeing a recent rise in both and a peak has not yet been reached, even if we can't ascertain their vaccination status. Additional data here that would support decision making would be to get population weighted numbers to standardize the data set for comparison.

The qualitative twitter data is useful to see what is on the minds of the public in terms of hashtags and conversations, there is definitely more publicity that could be done on vaccinations and this would be a good place for it to be promoted. However there is no location metadata there so we are unable to see this level and would recommend that for further investigation.