COVID-19 Report

Commission

This report aims to provide insights on the COVID-19 data and inform the UK government vaccination strategy.

Recommendations

It is suggested that the first marketing campaign should **focus on Gibraltar**, as this is the area with the largest number of people that have received only the first dose of the vaccine, and no second dose. It is also recommended that further campaigns should initially **avoid the Channel Islands**, as this is the area with the greatest number of recoveries.

A large peak in deaths can be observed across some regions during the winter months. Though there is not enough data to recognise a seasonal pattern, this may suggest a peak could occur next winter, and it is suggested that initial campaigns should target these affected areas (notably Gibraltar and Others).

The most common hashtag in the Twitter dataset is "**#covid19**", which also appears in some of the most-retweeted tweets in the dataset. As such, it is recommended to use this hashtag rather than "#coronavirus" or "#vaccinated" for any social media campaign.

It is difficult to determine whether hospitalisations have peaked yet using the current functions provided by the consultant, as these are focused on the current data rather than predicting future trends. It is recommended to carry out forecasting using an ARIMA method.

Further analysis is suggested, focusing in particular on insights from Twitter and time series forecasting, as well as investigation into some of the unexpected figures in the data provided.

Approach and Issues

Exploring the data

There is some missing data, most notably for dates 21/09/2020 - 22/09/2020 in Bermuda. Additionally, unexpected zero values for dates 13/10/2021 - 14/10/2021 suggest that some data may be captured on a delay, and data collection for recovered figures appears to stop at 04/08/2021 for most locations (to be thereafter recorded as zero).

Figures for "deaths", "cases", and "recovered" are assumed to be cumulative. Where data entry errors are presumed (if the reported figure is lower than at a previous date), this is corrected by setting these erroneous figures to the highest previous value.

The "hospitalised" figure is assumed to be the number of people currently in hospital with COVID-19 at that date (not new hospitalisations, and not cumulative). The "first dose" and "second dose"

figures are assumed to refer to the number of doses received on that date, and are also not cumulative. Note that to have received a second dose is to be considered fully vaccinated.

It is argued that any outliers should only be removed in extreme cases where there is good reason to suspect a data entry error. As such, box plots were examined, but no outliers were removed from this analysis.

Questions on the data

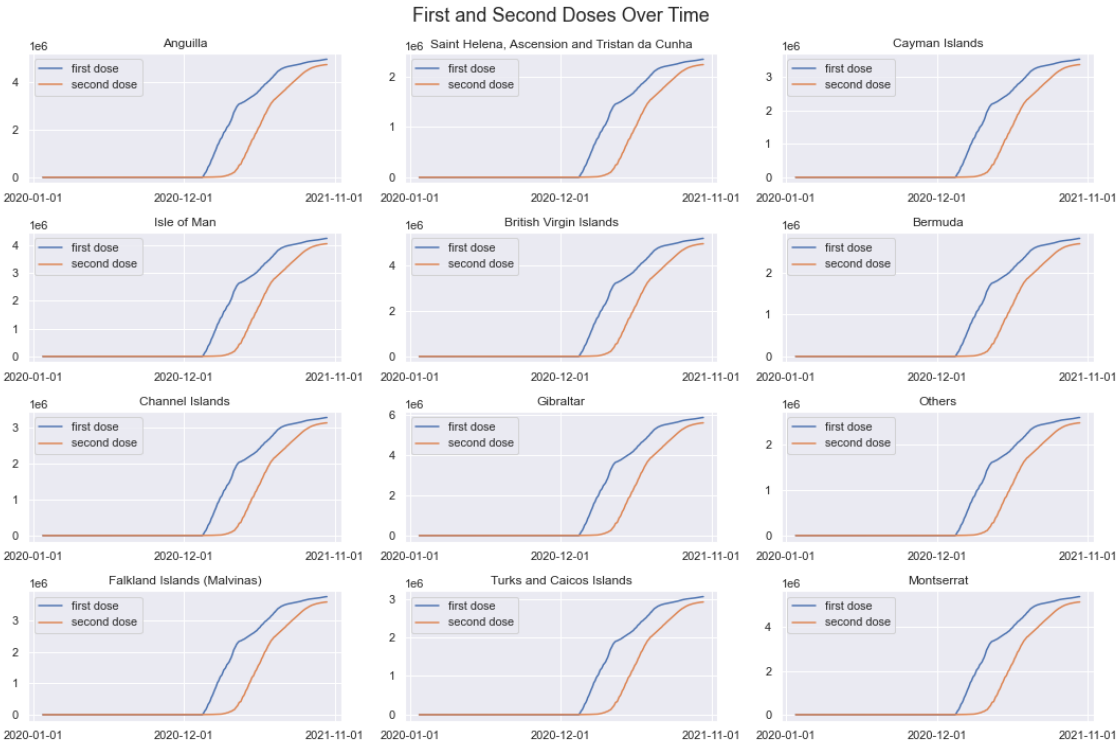
- This report includes analysis on an extract of Twitter data (3960 tweets) “relating to the #coronavirus hashtag”. Is there more information available on the context and criteria used to capture this data?
- There are 8,356,596 reported COVID cases in this data, with 99.5% of these occurring in the location “Others”. Are population figures available, so that COVID data can be analysed proportional to the population for each location?
- On some dates, hospitalised figures are higher than the total reported cases, and first and second doses appear to be rolled out on the same day. Are there any known gaps in the recording of data that could explain this?
- Figures for the first and second dose seem unreasonably high for each location. It appears that there is an issue with the quality of this data, or that it may require adjusting in some way. How has this data been captured?

In summary, trends in the data have been analysed in this report, **but please note that some figures may require adjustment**. Further investigation into these unexpected figures is strongly recommended.

Insights

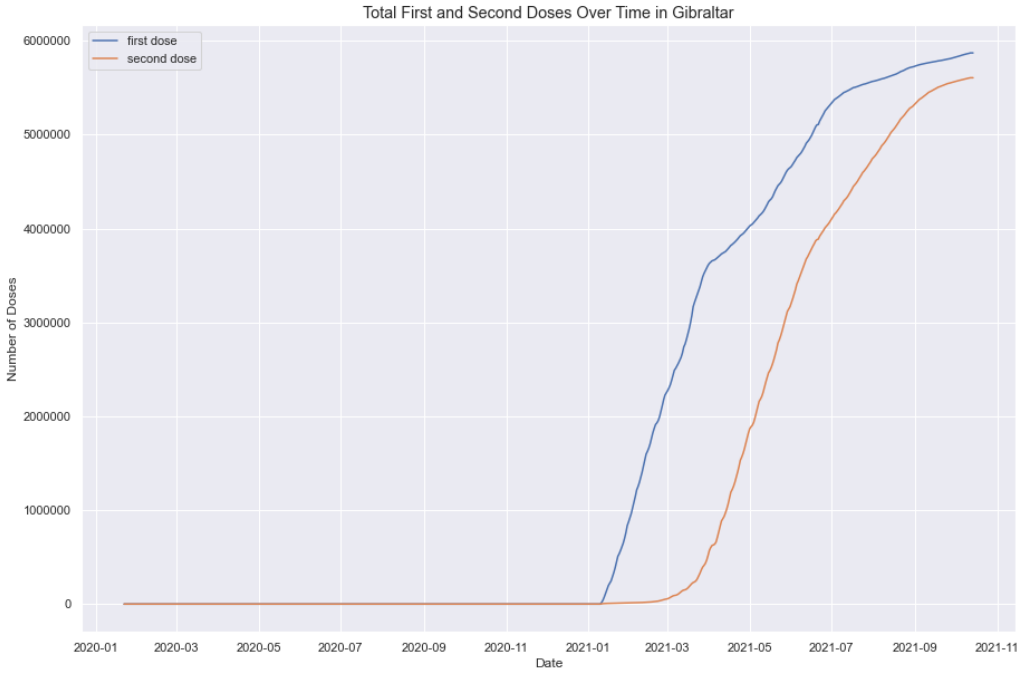
Vaccinations across regions

Examining the total first and second doses over time, we observe a similar trend across all regions.

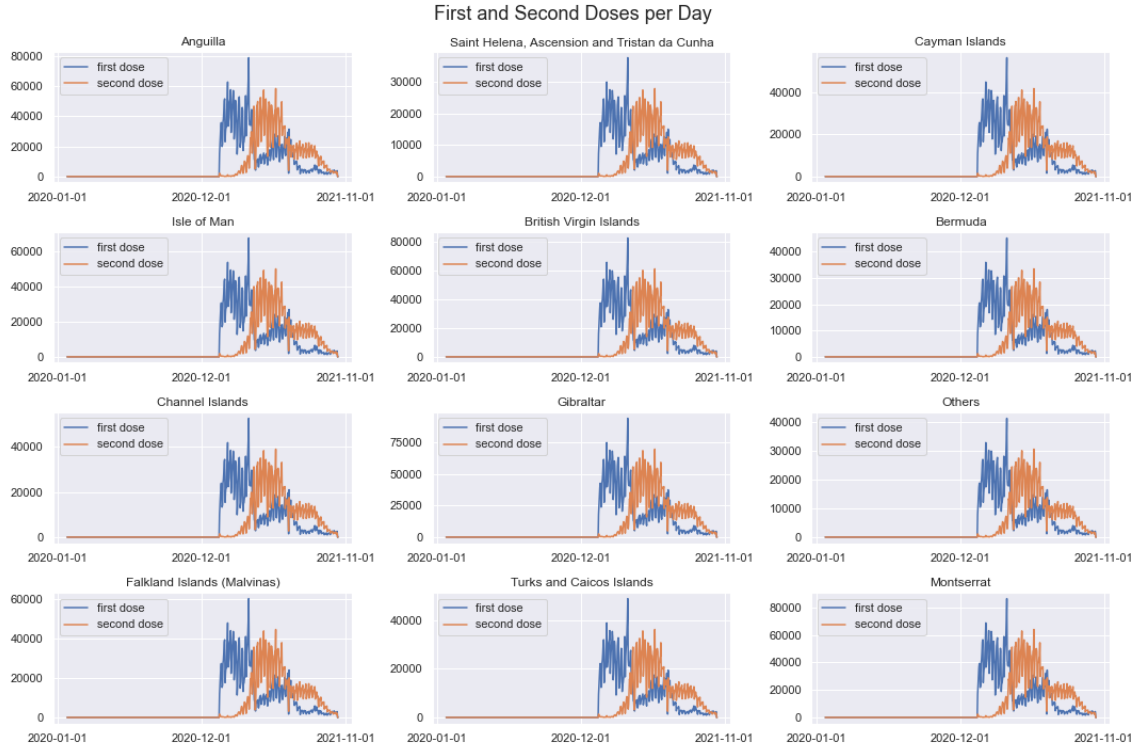


For both first and second doses, the curve increases sharply at first, but then becomes more level, tending towards a suspected upper limit (the finite population able to receive a dose). The second dose curve appears to imitate the first dose curve on a delay of around 2 months, which we may expect due to the recommended waiting period between COVID vaccines.

This can be clearly seen for Gibraltar in the plot below.



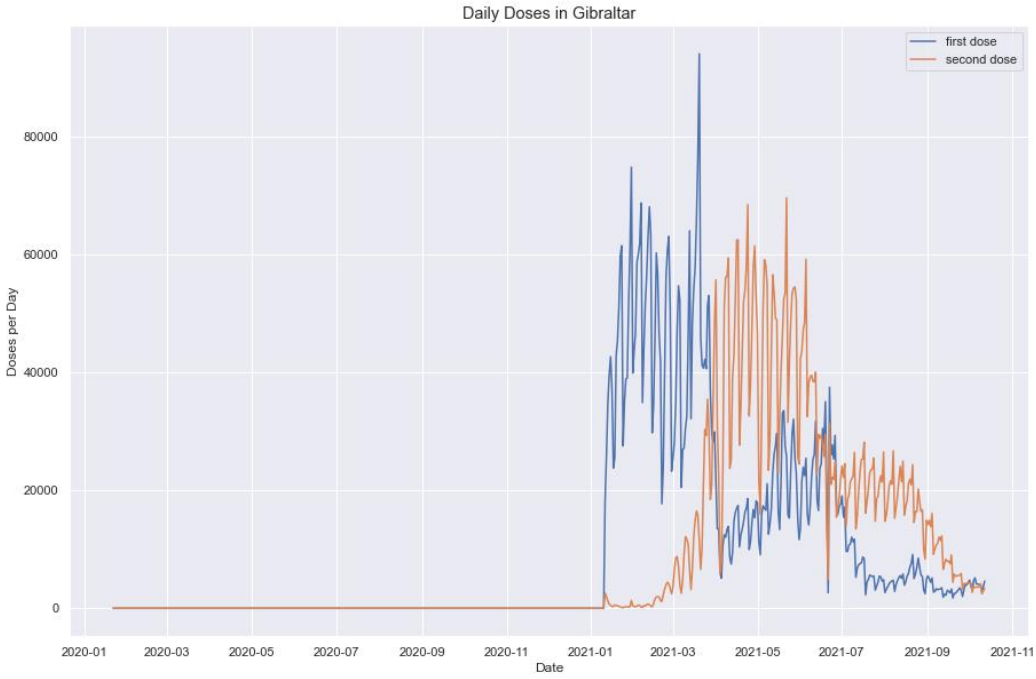
Examining the number of doses received on each day, we again observe a similar trend across all regions.



The trend for the second dose once again appears to roughly imitate that for the first dose, on a slight delay. After the initial peak at around 03/2021 (first dose) or 05/2021 (second dose), there is a general downward trend, with a notably sharp dip around one month later.

This may suggest a supply chain issue meaning that suddenly less vaccines were available around this date (followed by less people then being eligible for the second dose).

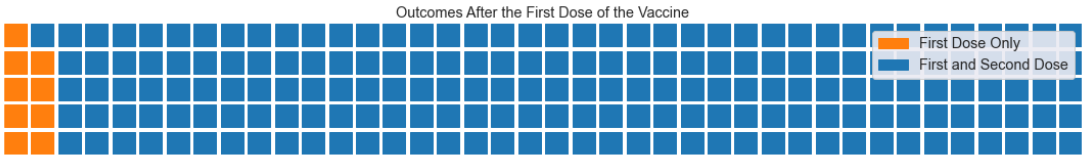
The dates 13/10/2021 and 14/10/2021 have been excluded from the plot for Gibraltar below, as we suspect that these zero values are due to a delay in the data capture process, and may be misleading.



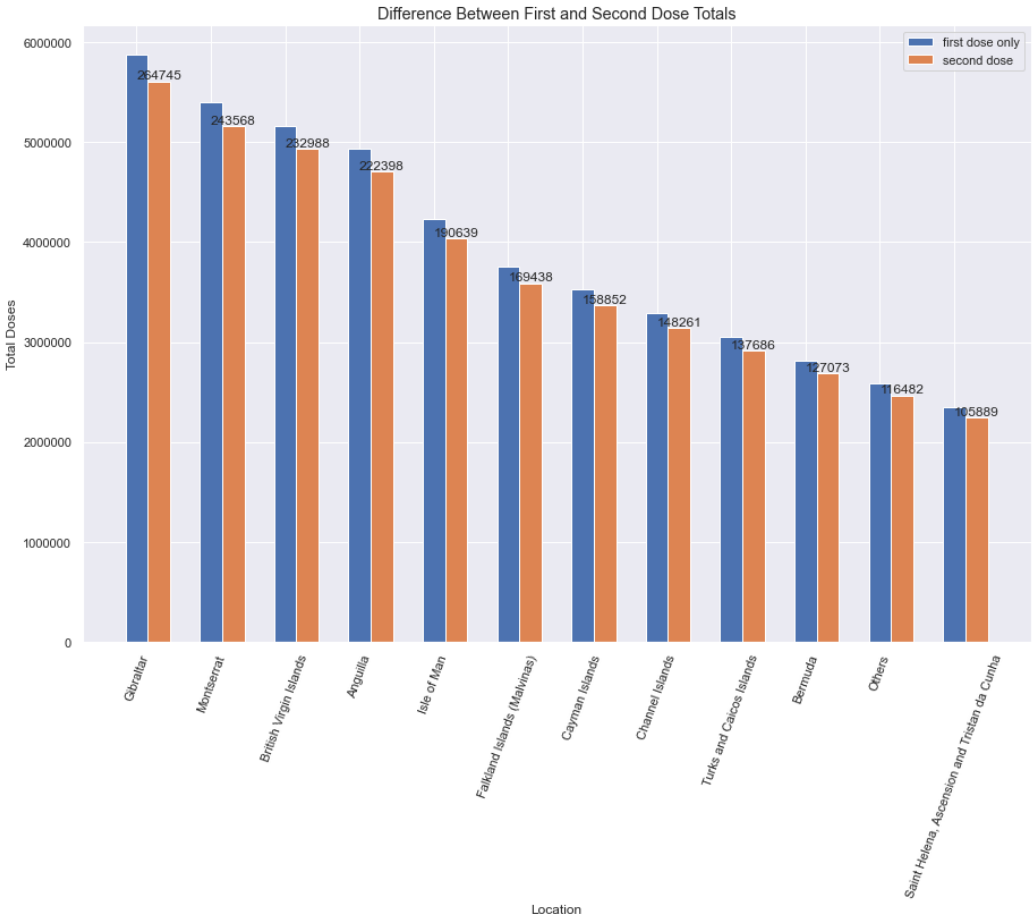
Marketing campaign

Low second dose uptake

Across all regions, 4.5% of people who received the first dose of the vaccine have not received the second dose of the vaccine. This is 9 in every 200 people, as demonstrated in the visualisation below.



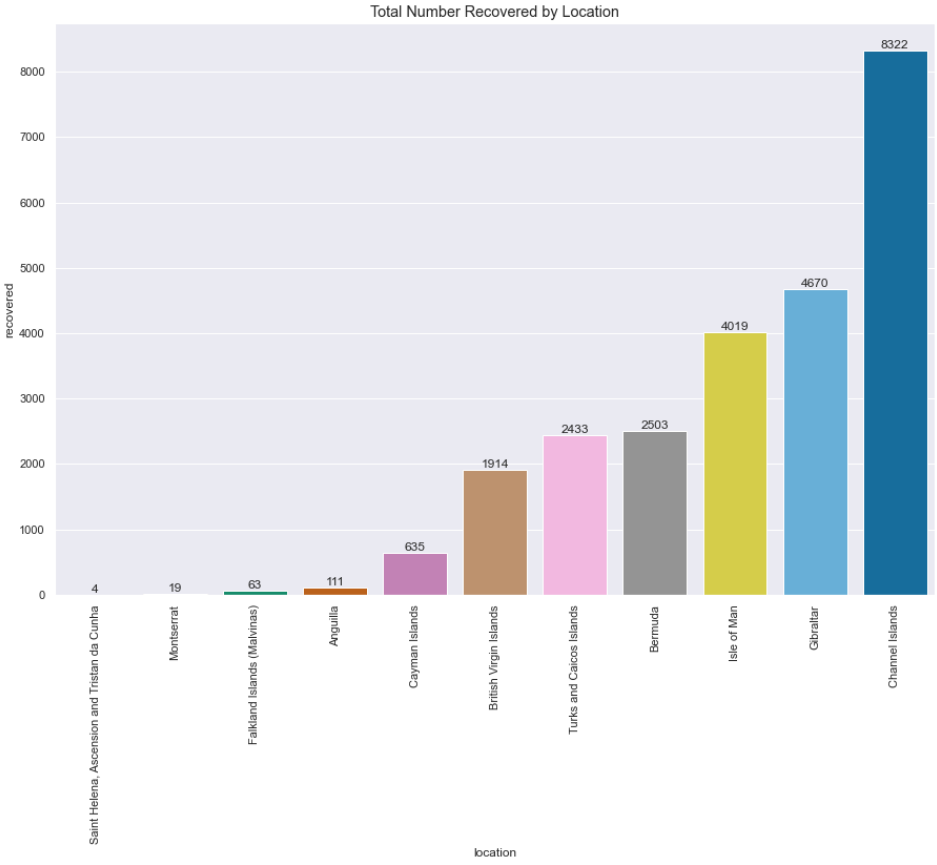
Although the percentage is the same across regions, the number of people affected is different, as displayed below. Gibraltar has the largest number of people who have received a first dose but no second dose (264,745 people).



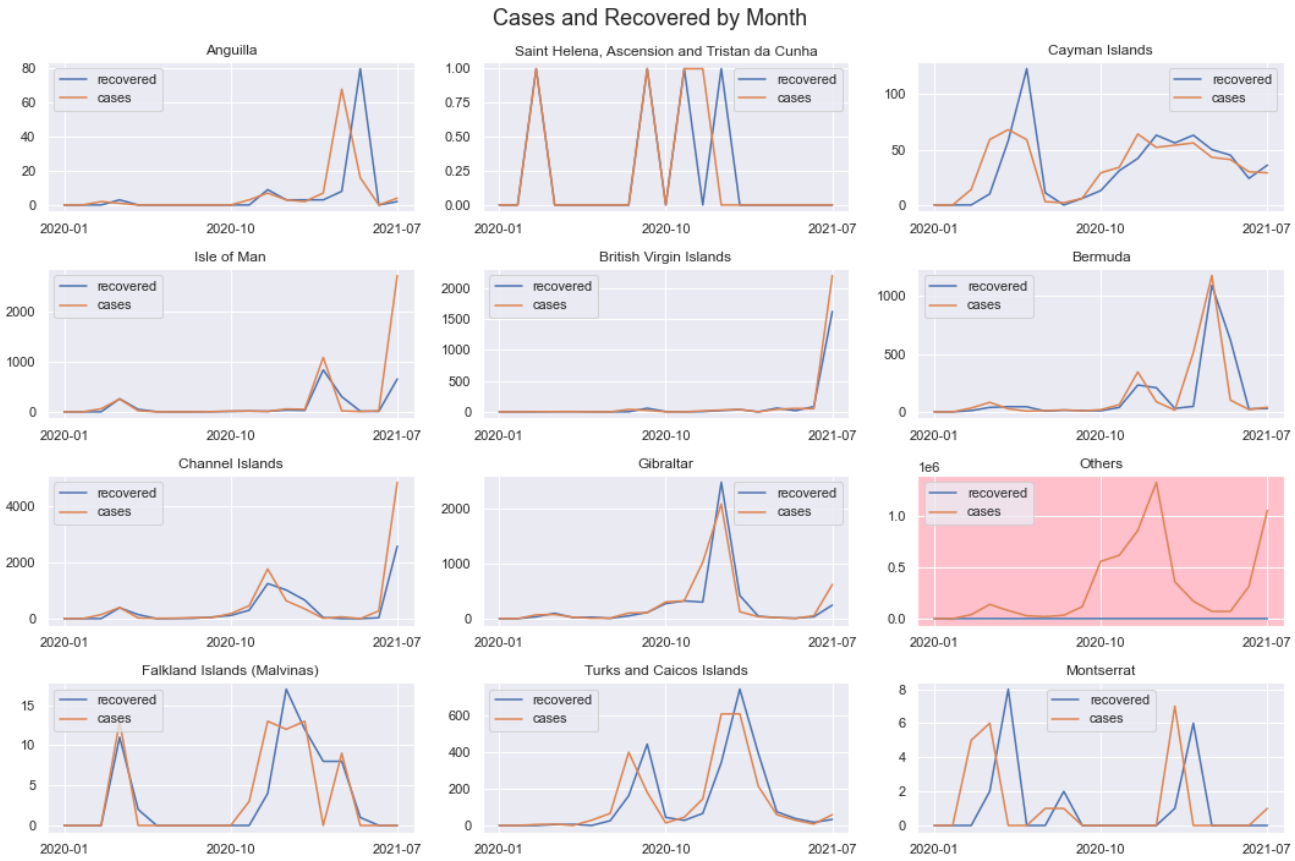
Greatest recoveries

Data collection for the “Recovered” figures appears to stop on 04/08/2021 for all locations except “Others”, which appears to stop on 12/04/2020 and has therefore been excluded.

As demonstrated below, the Channel Islands has the greatest number of total recoveries, with 8,322.



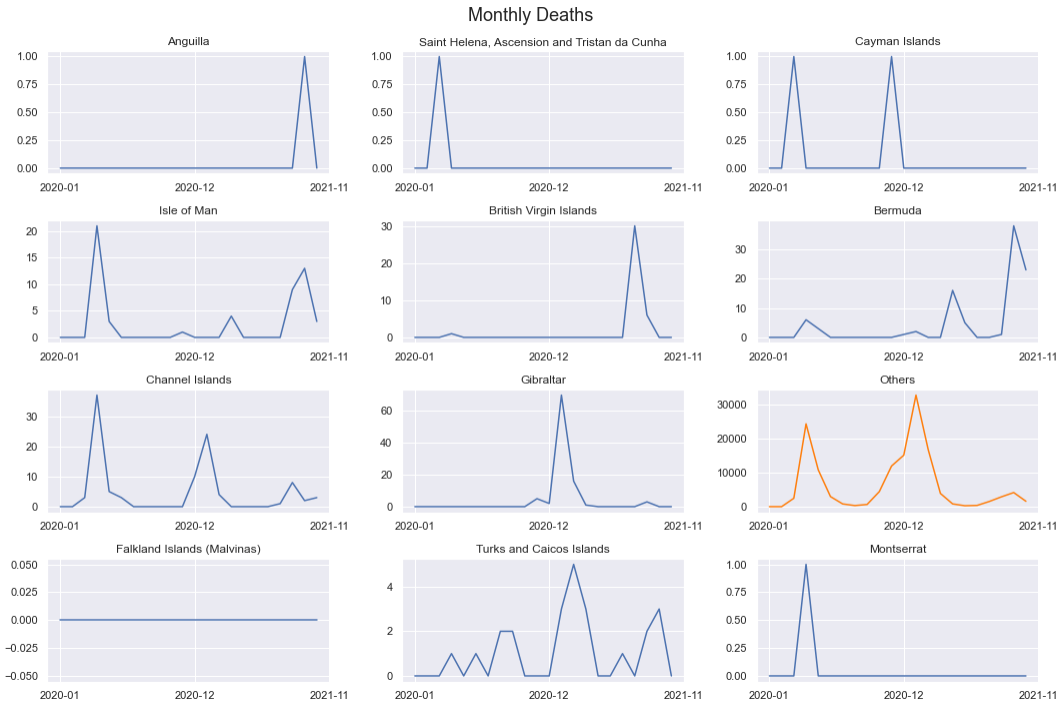
When examining the recovered figures by month, we excluded dates after 07/2021, as there was very little data on “Recovered” figures beyond this. It is clear that the number of recoveries tends to vary between months, and that recoveries generally seem to peak following a peak in COVID cases, which is as we might expect.



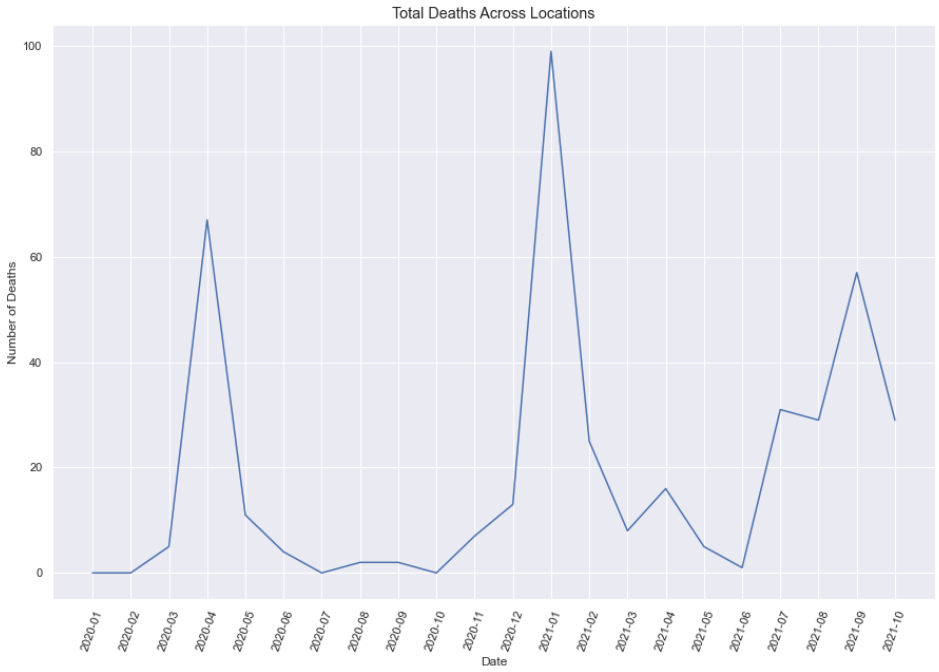
Deaths over time

Looking at the number of deaths per month for each location, displayed below, it appears that the deaths are not increasing steadily over time, but tend to appear in peaks.

A notable peak that can be seen across locations (such as Others, Channel Islands, and Gibraltar) is the peak at around 01/2021 – however, there is still potential for a greater peak in the future. In fact, although the data here is limited, we should consider that if there is any seasonal trend to COVID then this may suggest a similarly large peak may occur around 01/2022.



We look at the deaths per month for all locations combined (excluding “Others”, as the higher figures for this location would hide trends across other locations). Once again, we can notice the high peak at 01/2021.



Suggested strategy

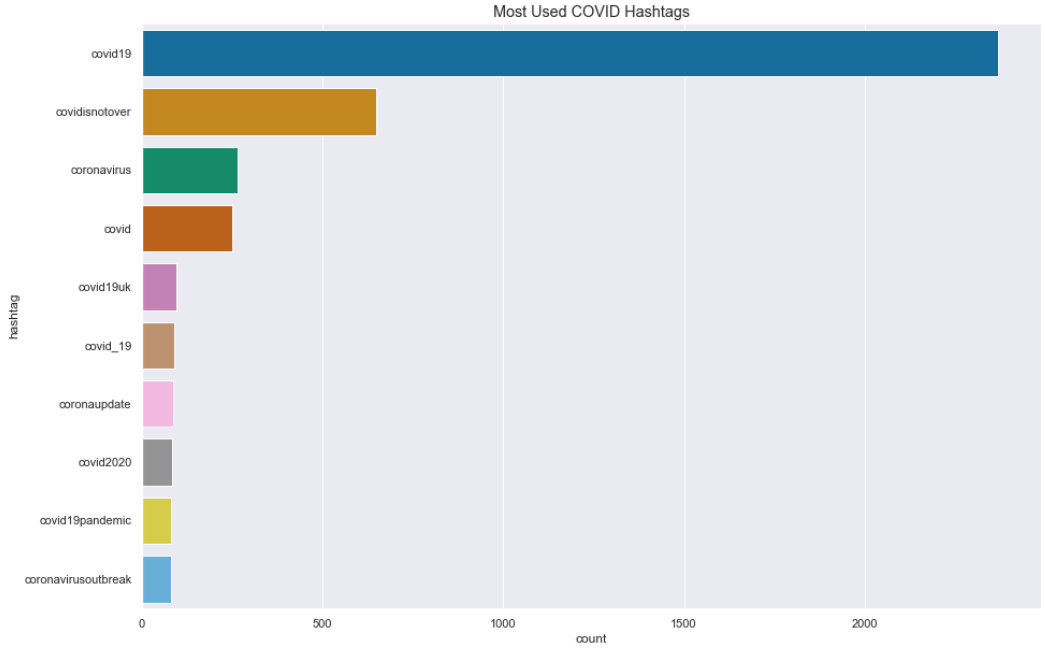
It is recommended that the first marketing campaign should be focused on Gibraltar, as this location has a large number of people who have not yet received the second dose.

There is also a suggestion that initial campaigns should focus on areas that experienced a large peak in deaths last winter, such as Gibraltar and “Others”, as they may experience this winter peak again. We could aim to roll out the marketing campaign and vaccine programme so that the vaccines are effective before this peak may arrive.

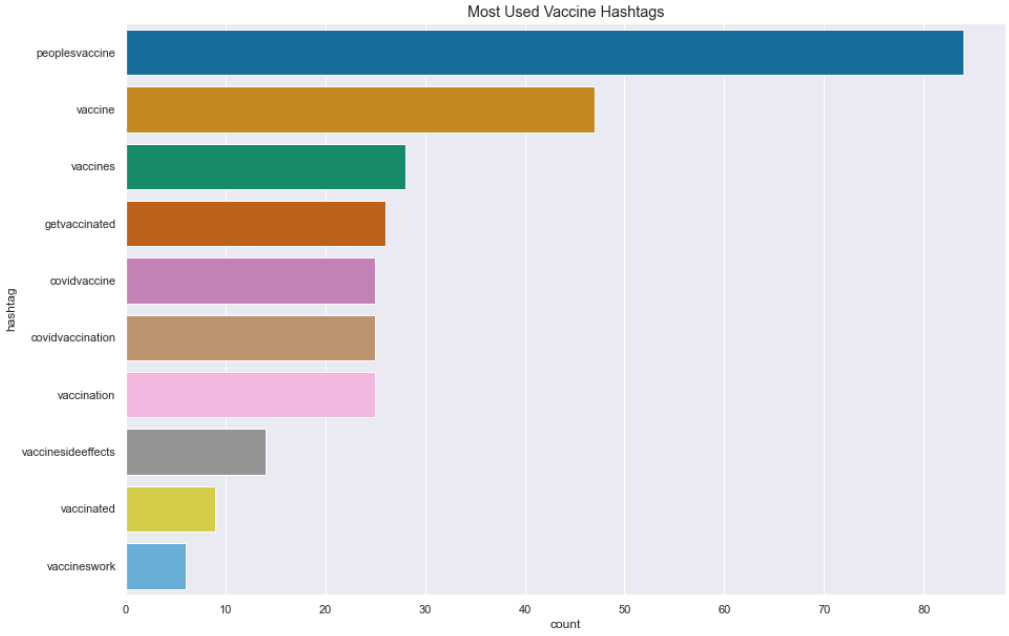
Twitter data: #coronavirus and #vaccinated hashtags

Capitalisation of hashtags was ignored for this analysis, so “#covid” and “#COVID” will be treated as the same hashtag.

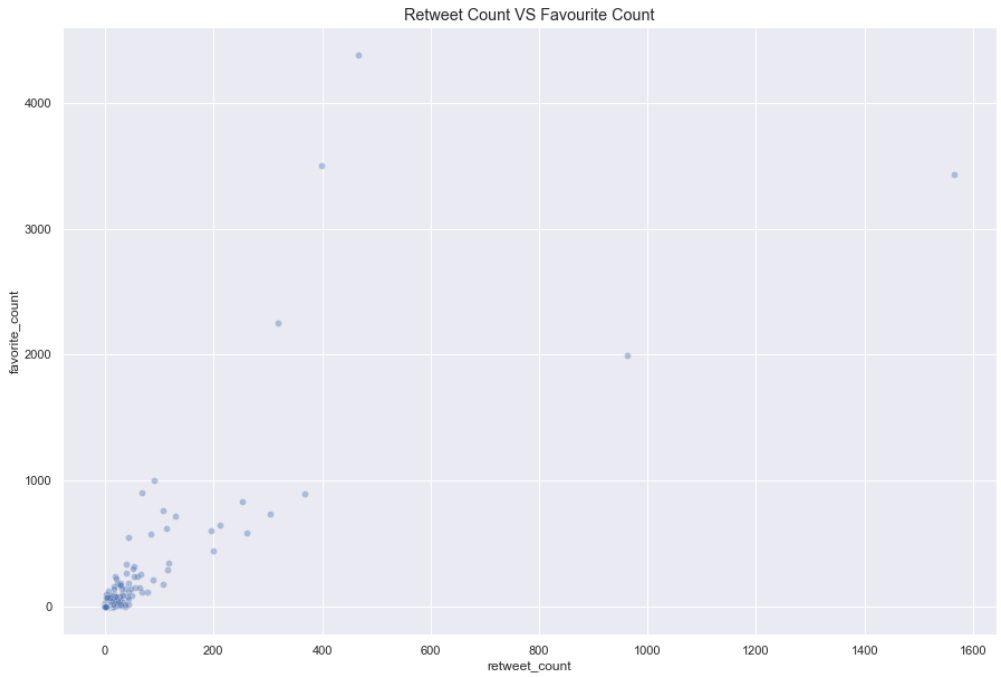
The most frequently occurring hashtags relating to COVID are displayed below. The search criteria were hashtags containing “corona” or “covid”. We can see below that “#covid19” is the most popular hashtag, appearing 2370 times in 3960 tweets.



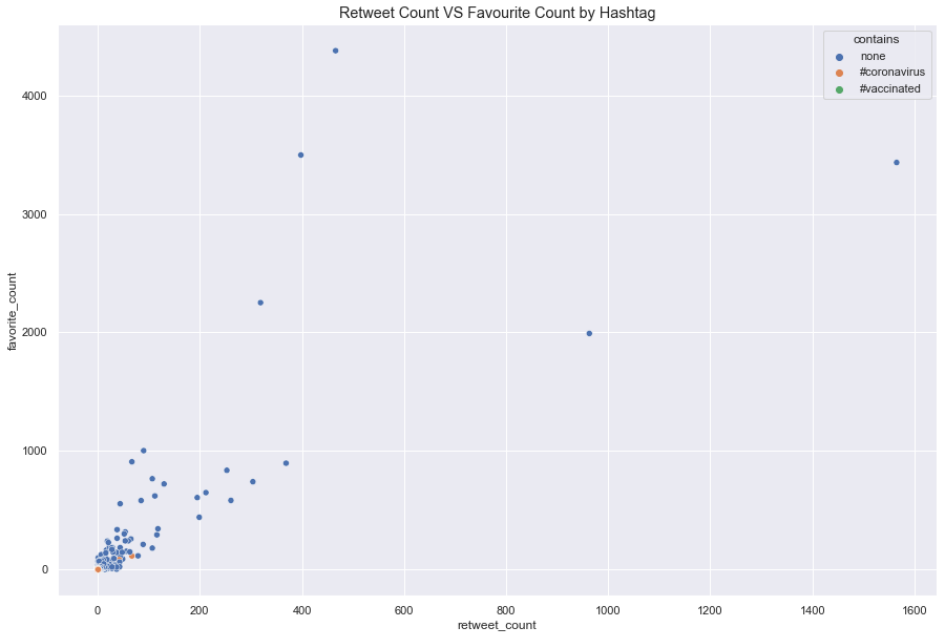
The most frequently occurring hashtags relating to vaccination are displayed below. The search criteria were hashtags containing “vaccin”. These vaccine tweets seem far less popular, as seen below.



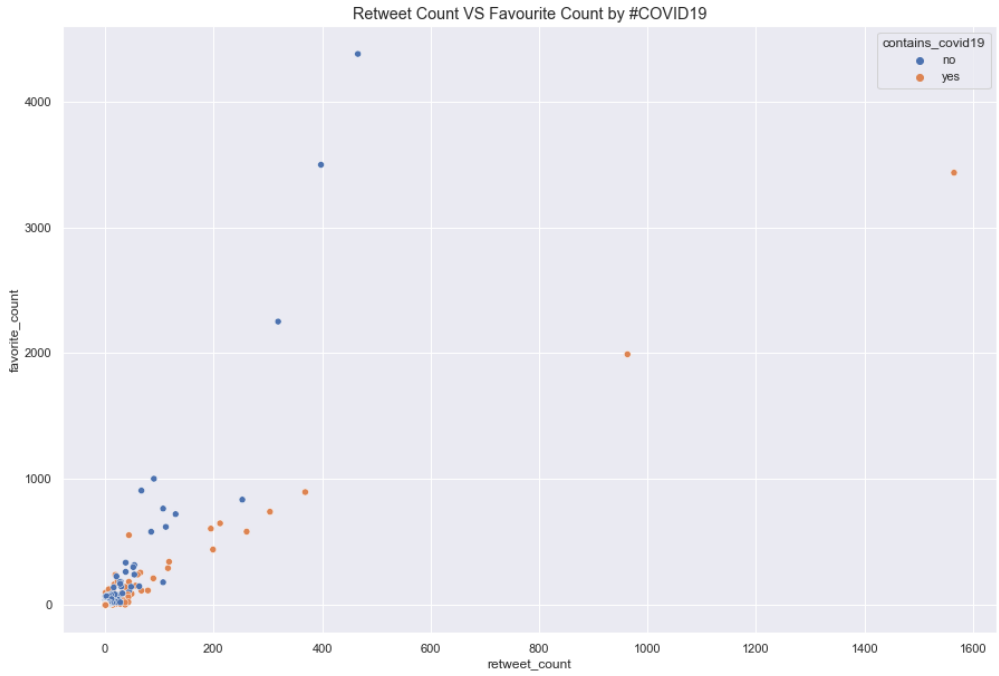
There appears to be a positive correlation between retweet count and favourite count (*r=0.813*), though interestingly the scatter appears wider at higher (extreme) values. We can also see from the plot below that there is a high density at the origin, implying that most tweets receive 0 retweets and 0 favourites.



There are no tweets in the dataset that contain both hashtags “#coronavirus” and “#vaccinated”. In fact, most tweets in this dataset do not contain either of these hashtags, especially at higher retweet or favourite counts, as demonstrated below.



In contrast, when investigating “#covid19” further, we observe that this hashtag seems to regularly appear in tweets with high retweet counts.



Twitter campaign

We recommend using the hashtag “#covid19” to launch any Twitter campaign, as this appears to be the most frequently used hashtag relating to coronavirus, and appears in some unexpectedly well-performing tweets.

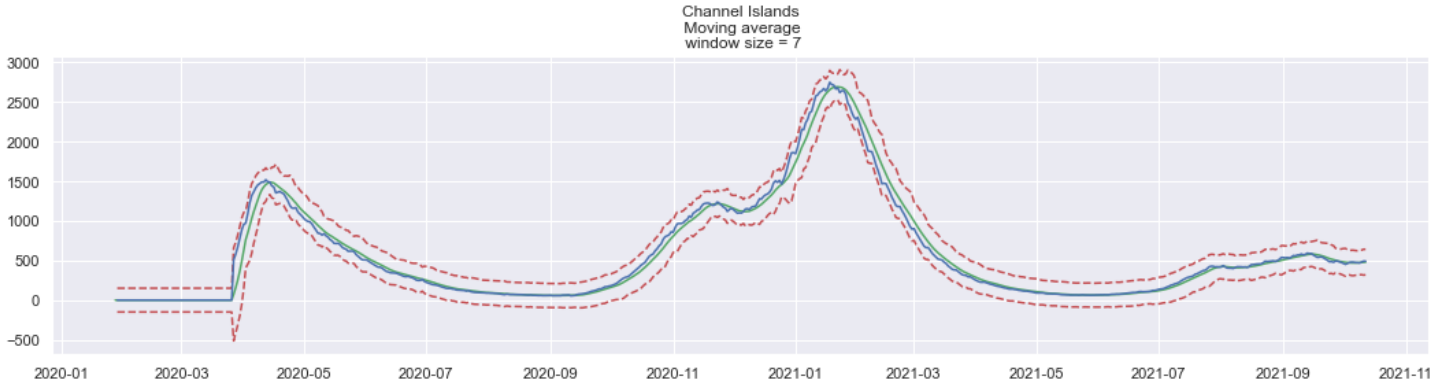
Further analysis is recommended in the form of performing a sentiment analysis on tweets from a large dataset to determine either “positive” or “negative” attitudes towards vaccines in different regions, so that marketing can be focused on those areas that may have a negative association.

Peaks in hospitalisations

Below is an outcome of the function created by the consultant. This displays the number of hospitalisations for the Channel Islands (blue) and the 7-day moving average (green), which attempts to predict the actual value by averaging the values of the last 7 days. Note that dates 13/10/2021-14/10/2021 have been excluded due to unexpected zero values which may appear misleading on this plot.

We notice a large peak in hospitalisations at around 02/2021 – a very similar trend exists across all regions. It does also appear that hospitalisations in the most recent months may be getting higher, possibly indicating another peak.

However, with the current function it is difficult to determine whether hospitalisations have peaked, as the function is focused on the current data rather than attempting to predict future figures. It is therefore recommended that an ARIMA method be applied for forecasting.



The table below, provided by the consultant, displays the dates where the actual hospitalisation numbers are most different to the 7-day rolling average figures (and thus have the highest “error”).

The previous dates all have a value of 0, followed by a very sudden leap to 509. This means that the rolling average for 27/03/2020 would be calculated as (0+0+0+0+0+0+509)/7 = 72.7. The sudden change compared to the previous six values means that the rolling average has a large error value in this case (436.3).

date	location	hospitalised	error
2020-03-27	Channel Islands	509.0	436.285714
2020-03-28	Channel Islands	579.0	423.571429
2020-03-29	Channel Islands	667.0	416.285714

Explaining concepts

Qualitative and quantitative data

Qualitative data is categorical, for example eye colour, which could be brown, or blue, or green. Quantitative data is numerical, for example the cost of a bottle of wine. Both data types can be used in business predictions, but they perform best in different scenarios and require different methods. Qualitative analysis may involve collecting feedback from customers or experts, and is useful if we lack historical data. Quantitative analysis uses historical data to learn patterns and predict future events, for example using previous sales figures to predict future sales.

Continuous improvement

Improving a data analysis project is likely to lead to more accurate conclusions, which in turn improves the return on investment for any campaign based on these insights. This is especially important when potential underlying issues in the data have been identified, or when areas suggested for further analysis could make marketing significantly more effective.

Data ethics

Aggregating data to avoid revealing personal details is good practice in data ethics, though not the only thing to consider. We also must ensure that any conclusions that we share from a data analysis are not poorly drawn, and that results have not been reported in a misleading way for a more interesting or more desirable result, or even unintentionally.

Summary

Recommendations for the vaccine marketing campaign, most notably the suggestion to campaign first in Gibraltar, and use popular hashtags such as “#covid19”, are presented in full at the start of this report, along with questions raised around potential issues with the current dataset.

Further analysis is recommended, particularly in the areas of forecasting future trends, and determining the sentiment behind tweets containing various hashtags.