COVID-19 ANALYSIS

Assignment Report: Data Analytics Using Python

July 2022

Pollyanna Hadley

Contents

1. Introduction

The purpose of this report is to present an analysis of COVID-19 data to the UK government.

The UK government intends to use identified trends and patterns in this report to inform its marketing campaigns promoting vaccination.

Objectives

The government has outlined the following as areas they would like to understand:

- Total vaccinations (first and second dose by region)
- Locations with the highest number of individuals partially vaccinated
- Locations which have the highest number of recoveries
- Trends of deaths over time
- Analysis of Twitter trends relating to COVID-19
- Trends in hospitalisation numbers

Report structure

The report discusses the approach to the analysis, patterns and trends, results and conclusions, and other recommendations.

A separate presentation and Jupyter Notebook has been prepared to support this report and details code specifics, observations and justifications.

2. Initial Approach

Covid cases and vaccinations data

The following steps were taken to load, clean and prepare the two datasets for initial analysis:

Data preparation and initial review

- The two .csv files loaded into Jupyter notebook along with relevant python libraries
- Data sets were explored separately to understand how the raw data is structured, including the shape of the data (number of rows, columns) and the data types stored.
- The first and last five lines of the data sets were previewed to get a flavour of how the data sets look.
- Checked for missing values
- Reviewed high level descriptive statistics, across all provinces and compared to the province with the highest vaccinations (Gibraltar)
- Grouped and aggregated data by province and dates

Initial insights and findings from data prepration:

- Data is available from January 2021 through to October 2021.
- Missing data:
    - Identified two rows in the cases set with missing data for Deaths, Cases, Recovered and Hospitalised.
    - The missing rows both relate to the same province, Bermuda, and are for two consecutive days – 21st to 22nd September 2020, and therefore it would appear these missing data sets are localised.
    - The lines will be left in for the purposes of the analysis since only two rows are unlikely to materially affect the dataset.
    - Recommend gaining a further understanding of how the missing data has occurred as this could indicate other points in the data set that may not be complete or accurate and could impact conclusions.
- Province categories:
    - 'Others' category found however the data does not specify which region this belongs to, making it difficult to make recommendations on this region, for example there could be sub-regions that should be analysed separately. The total sum of deaths and cases for this category appear to be outliers as they are significantly higher than the other provinces. Subsequent analysis has taken this into consideration.
- Highest/lowest vaccination rates:
    - Initial analysis shows Gibraltar had the highest number of fully and partially vaccinated individuals, whilst Saint Helena, Ascension and Tristan da Cunha had the lowest.
    - Given Gibraltar had the highest, a summary of descriptive statistics was generated for this entity, and compared to statistics across all provinces:

Gibraltar daily cases statistics summary:

	Deaths	*Cases*	*Recovered*	*Hospitalised*
Mean	40.21	2,237.11	1,512.82	1,027.63
Min	0	0	0	0
Max	97	5,727	4,670	4,907

All provinces daily cases statistics summary:

	Deaths	*Cases*	*Recovered*	*Hospitalised*
Mean	6,210.20	2,147,082	454.69	685.23
Min	0	0	0	0
Max	138,237	8,317,439	4,670	4,907

- Initial findings suggest that despite the highest vaccination rates, Gibraltar has a significantly higher daily average hospitalisations compared to all provinces. It would be interesting to explore this further to understand the relationship between the increase in hospitalisations and timing of the vaccinations. Additional information on whether the recorded hospitalisations are a direct result of COVID-19 only, or whether patients were being admitted for other reasons (e.g. a flu outbreak) but happened to be covid positive as well, would be useful.
- Average recovery rates in Gibraltar are higher than the average across all provinces, which could be an indication of the effectiveness of the vaccinations. However, this conclusion cannot be fully supported without additional information. For example, population demographics (such as age, gender) could also be a contributing factor to higher recovery rate.
- The summary above for all provinces highlight the anticipated skew of data resulting from the 'Others' province in the Deaths and Cases. Therefore for the purposes of the initial analysis these categories have been ignored.

Twitter data

The .csv file containing twitter data was imported into the supporting Jupyter notebook. Initial review and exploration of the data was undertaken to understand the structure and type of information contained in the set.

Initial analysis was performed to understand the count of favourited and retweeted posts. This is as far as this analysis was taken since more valuable insights could be obtained from reviewing the hashtags and text at this stage of the analysis, however analysis could be enhanced by revisiting these elements in the future.

Hospitalisation data

The partially completed functions provided by the consultant were reviewed and the moving average ('SMA') was plotted and reviewed.

3. Patterns and trends

Covid cases and vaccination data

The next stage of the analysis involved the following steps:

- Merging the two original datasets to form one set of data
- Adding additional columns to calculate:
  - Number of eligible individuals who have had one vaccination only (partially vaccinated)
  - Dose ratio - percentages of eligible individuals who are fully and partially vaccinated
- Grouping and aggregating dataset by Province and Date
- Reviewing the death and recovery rates over time, and by location

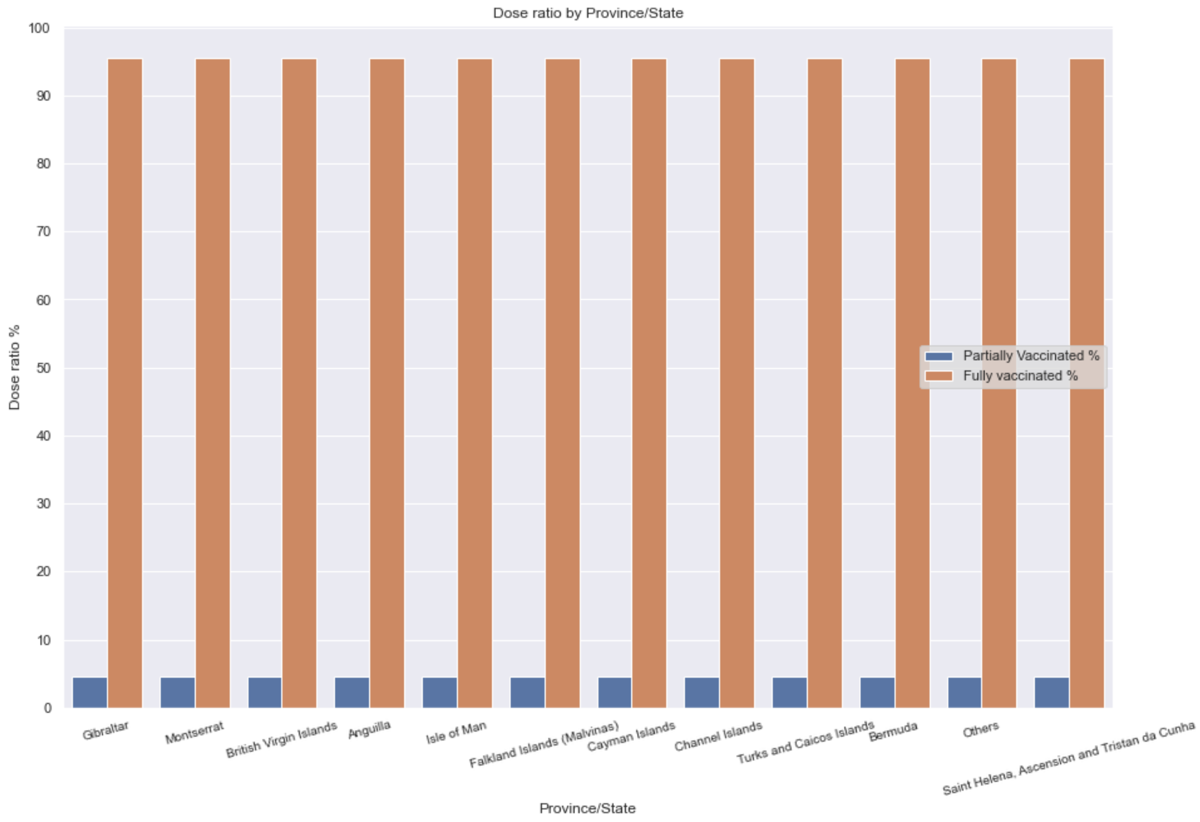
The Dose ratio by province

The table in Appendix 1 presents the findings when reviewing the dose ratios at a Province level.

The analysis shows that whilst Gibraltar has the highest number of individuals who are only partially vaccinated, Turks and Caicos Islands have the highest percentage of individuals who are only partially vaccinated.

Saint Helena, Acension and Tristan da Cunha have the lowest by number and percentage of partially vaccinated individuals.

A grouped barplot has been used to present dose ratios by region to easily compare:



This bar plot visualises that whilst there are differences in numbers of fully vaccinations between regions, there are no significant fluctuations in the dose percentage by province, with the percentage of fully and partially vaccinated individuals hovering around the 95.5% and 4.5% mark respectively.

Dose ratio over time

A further step was taken to review the change in ratio across all provinces over time – refer to Appendix 2.

The analysis shows a significant increase in the number of doses between January and April 2021 which then continues steadily and levels off by October 2021.

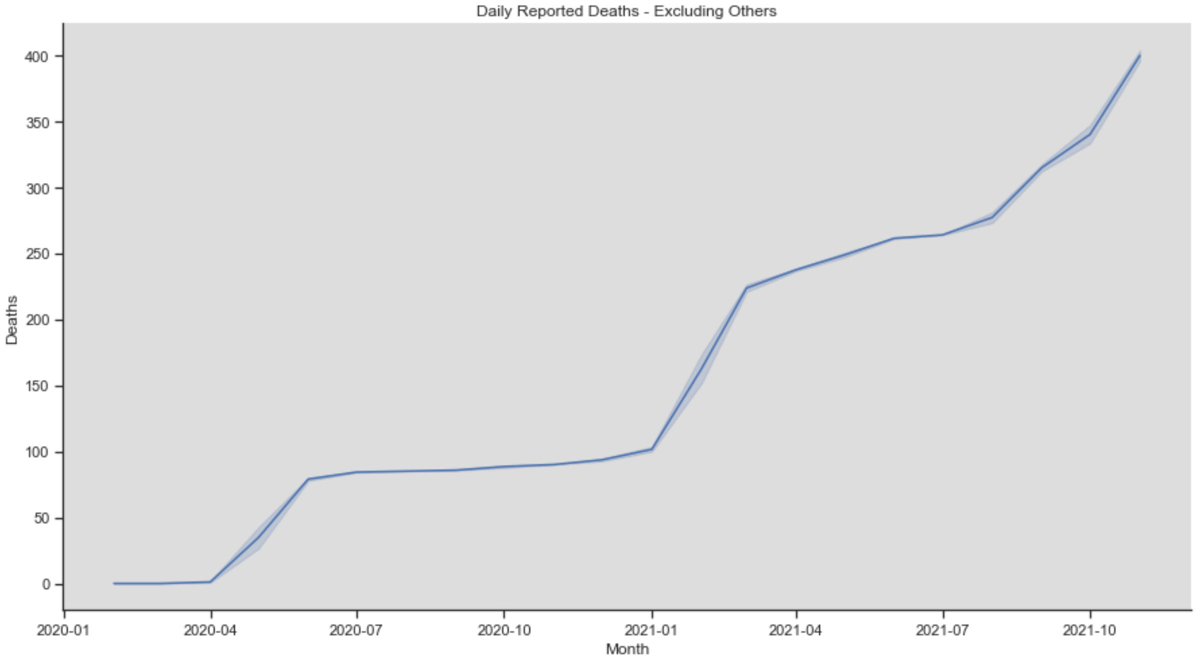
There is also a jump in the percentage of fully vaccinated individuals between March 2021 and May 2021. This likely represents the required time-frame needed between first and second doses.

In October 2021, both the number and percentage of partially vaccinated individuals begins to increase again. Data for later months through to 2022 would be valuable to understand the reasoning behind this to determine whether this is an indication of erroneous data, or other factors driving this increase, such as increased campaigning to encourage more unvaccinated to get their first does, or certain demographics (younger children) becoming eligible for the vaccine at this time.

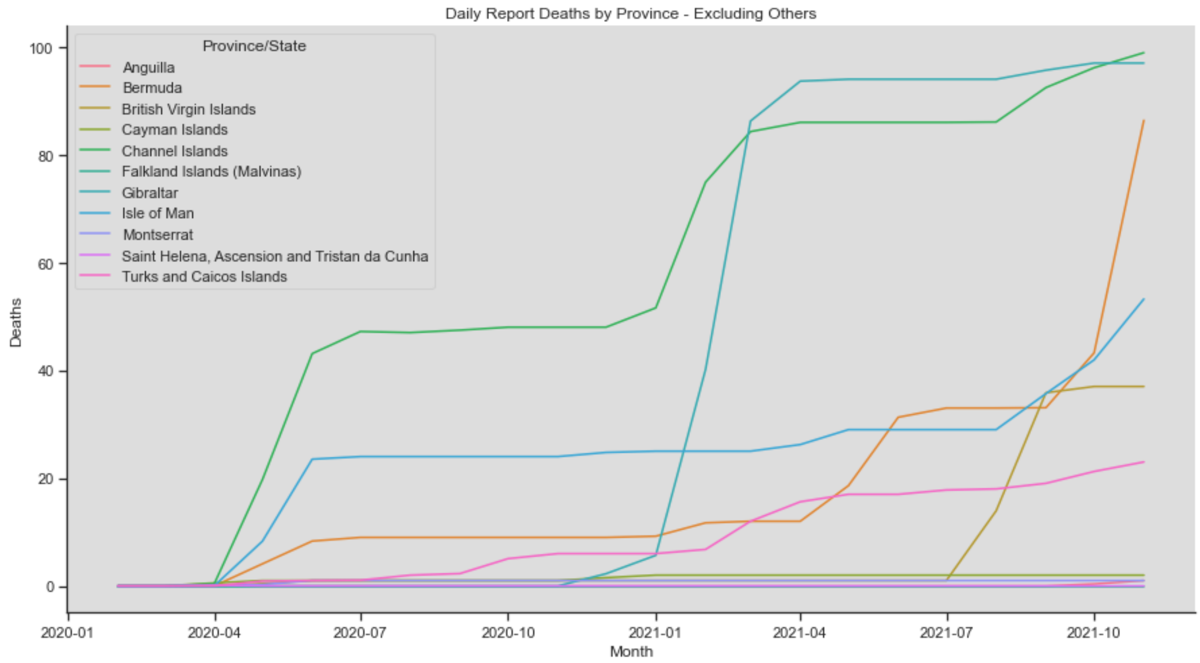
Daily death rates over time and by location

The daily death rates over time have been plotted. 'Others' category has been excluded from this analysis to prevent skew.

Use of a line plot easily visualises and helps identify trends when looking at data over a time series.



The line plots for all provinces suggests the overall daily death rates are continuing to rise past October 2021 and has not yet reached the peak.



Whilst the rate of increase of death rates has fluctuated, overall, there has been a continuing rise of daily rates throughout the period.

Daily death rates appear to be increasing sharply in Bermuda, and Isle of Man later in 2021. For other provinces, the rate of increase in daily death rates is being to flatten and there are early indications that the daily death rates are beginning to fall in Gibraltar.

Notably, the Gibraltar has seen a significant increase in daily death rates in the first quarter of 2021, with the daily rate flattening from around March 2021. I would wish to further explore the reasoning behind this and review in line with other available date. For example, possible reasons for this trend could include, for example, deaths not accurately being recorded until January 2021, timing of lifting of restrictions or a surge in cases from a variant.

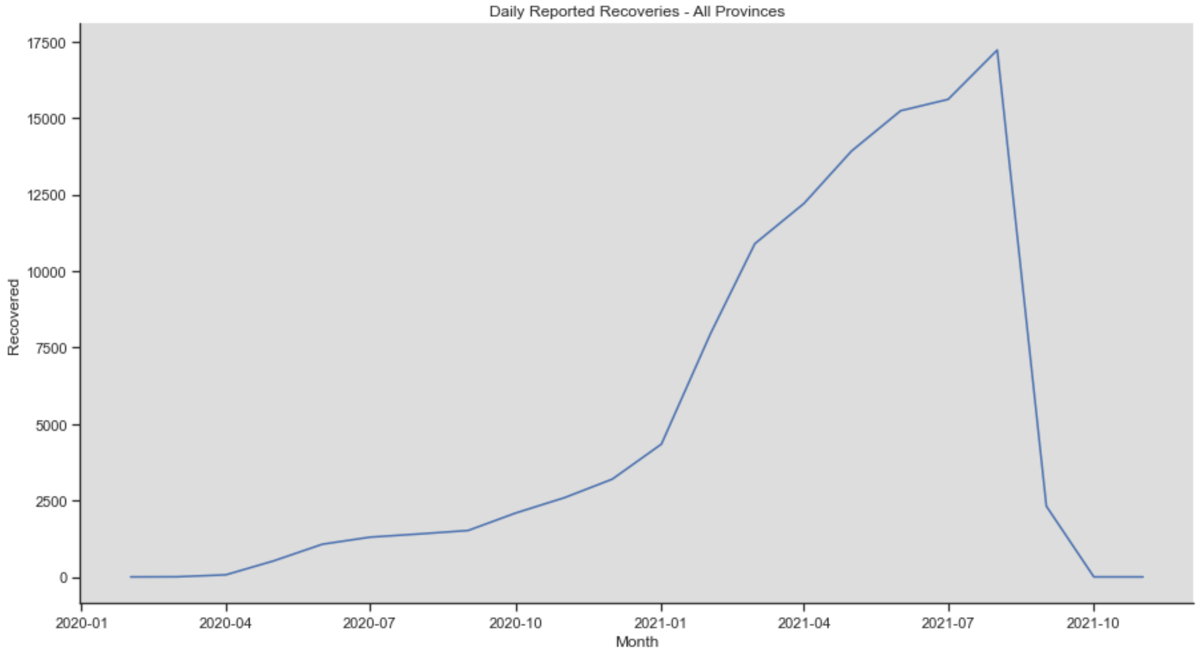
Number of daily deaths in Cayman, Falklands, Montserrat, Turks and Caicos, and Saint Helena, Ascension and Tristan da Cunha are significantly lower than the other provinces.

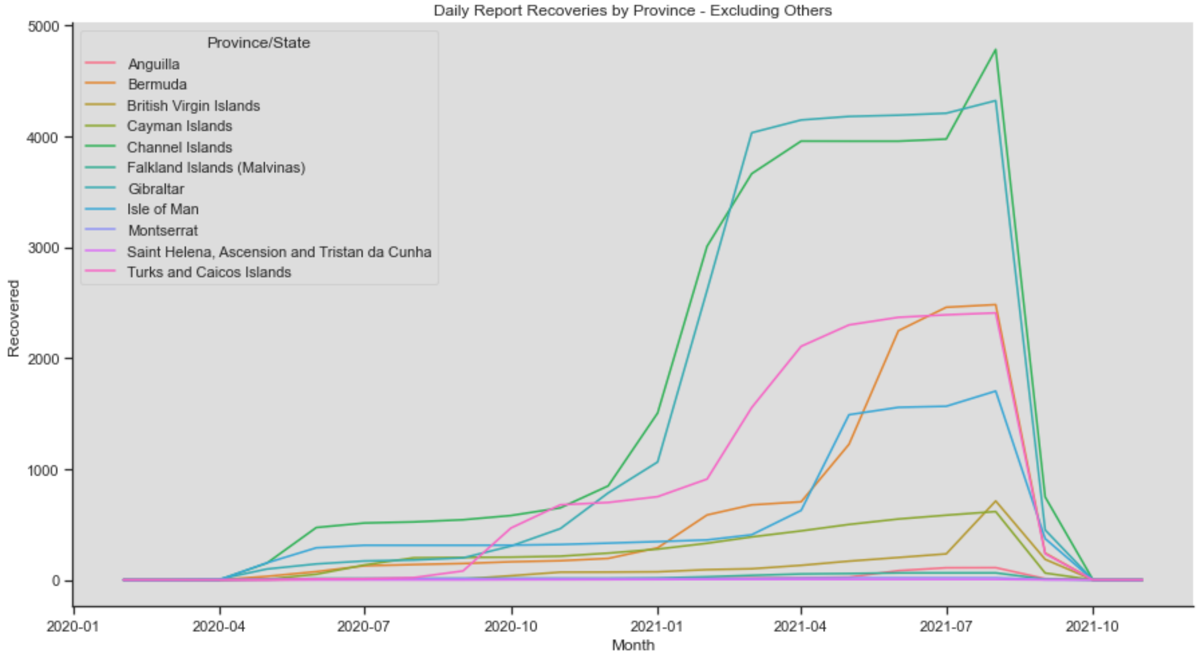
Further information on population size would enhance this analysis to determine the death rate per population and identify any significant variances at this level.

Daily recovery rates over time and by location

The daily recovery rates over time have been plotted. ‘Others’ category has been excluded from this analysis to prevent skew.

Use of a line plot easily visualises and helps identify trends when looking at data over a time series.





Overall, the line plots show a sharp rise in daily recoveries from January 2021 which corresponds to the date the vaccines were introduced.

On an individual province level, the significant rise in recoveries also aligns with slow of increase in daily deaths recorded for the provinces in the previous daily deaths graphs.

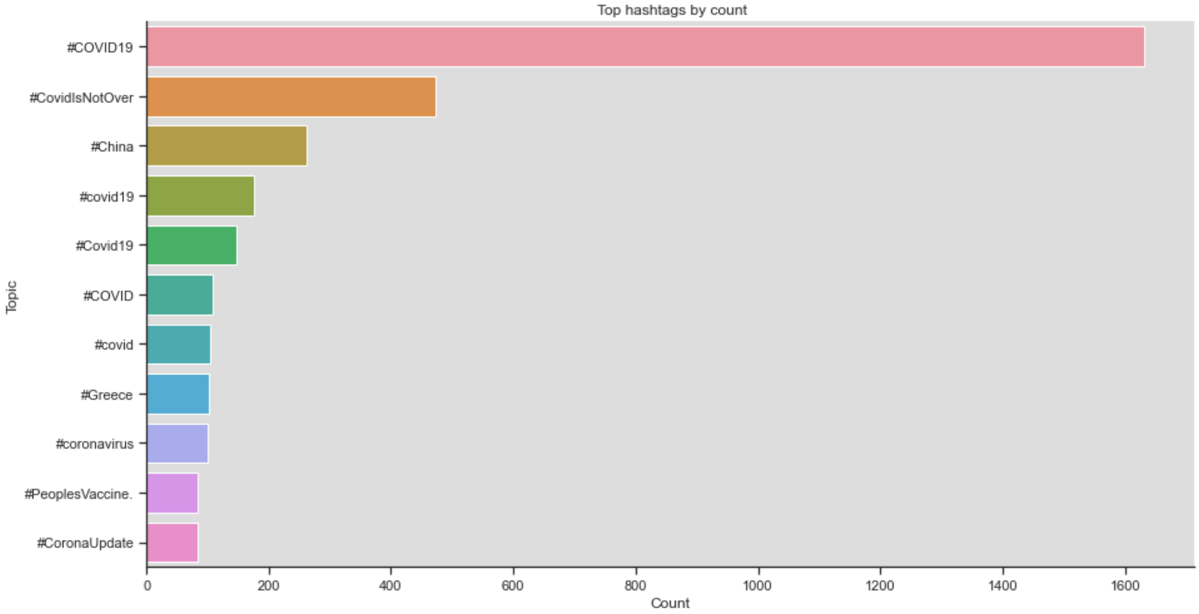
Recoveries drop significantly just before October 2021 across all provinces. This would suggest the data is not available in this dataset, rather than a significant decline in recoveries. Therefore the data after the peaks should be ignored for the purposes of this analysis.

Further analysis on the timing and lag between vaccinations and increase in vaccinations per province can be performed which will could support and corroborate a campaign to promote the uptake of the vaccine.

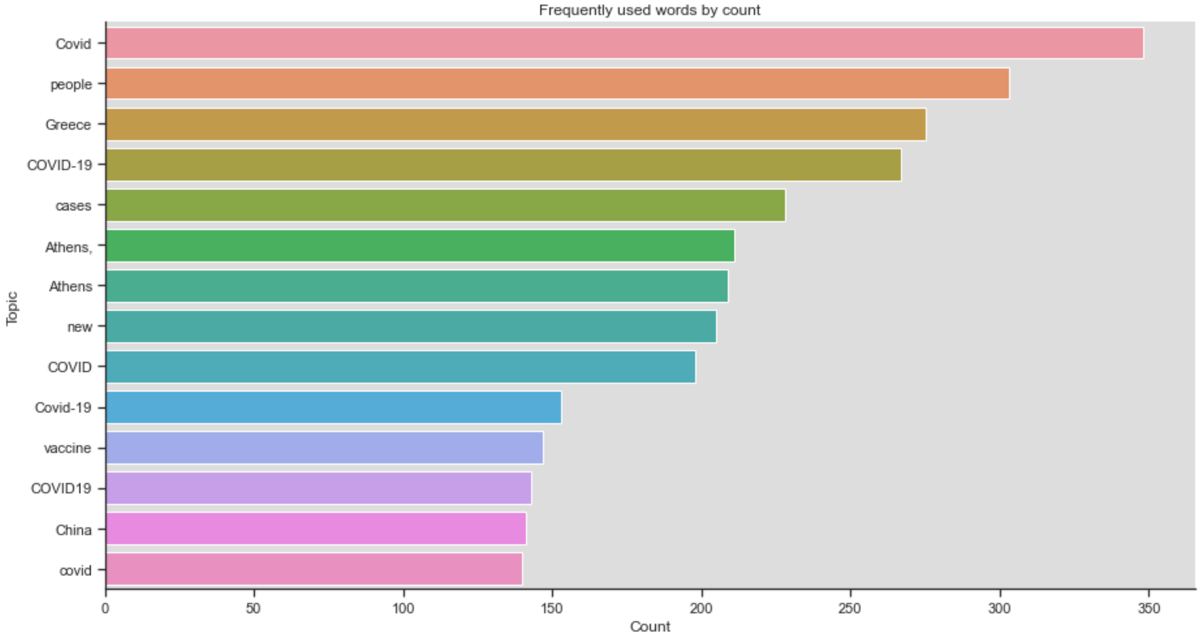
Twitter Data

Analysis of hashtags and text trends

The data was analysed to identify the top trending hashtags at the point of data collection. The following barplot presents these results.



Additionally, the text within the twitter posts was analysed to determine the top topics by reviewing the most frequently occurring words within the posts. Note, common English words (e.g. “and”, “the”) were filtered prior to the analysis which allows us to identify the theme and underlying topics.

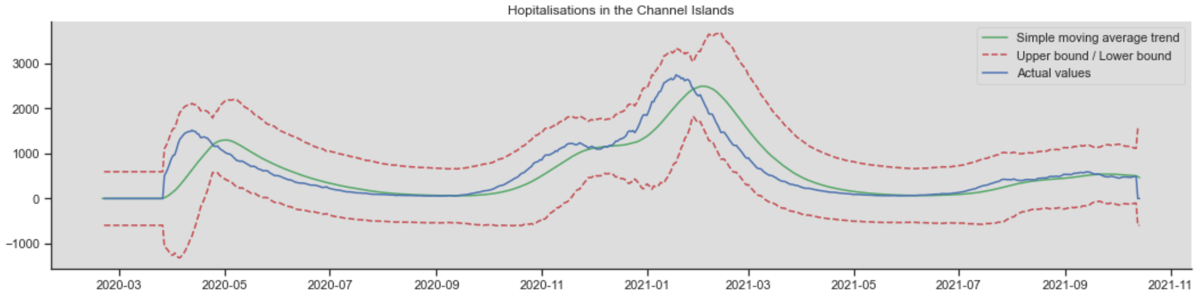


The charts show that in both cases, covid is a popular and trending topic on twitter, which suggests it could have a significant influence on the population. Therefore, the government will want to consider the use of Twitter and/or other social media channels as part of their marketing strategy.

However, I would recommend further analysis should be undertaken to determine whether the content of these trending posts discusses covid and vaccinations in a positive or negative manner, especially in light of recent media press surrounding fake news, as this will be able to further shape marketing to ensure it is implemented with maximum impact.

Hospitalisation data

The SMA was plotted using a 30 day window.



Using a window of 30 day smooths the data and helps highlight trends. We can see from the graph there are two main peaks – one in May 2020 and one in around February 2021.

These results align and correspond with the sharp increase in daily death rates in the Channel Islands.

Hospitalisations tail off following the last peak which could be linked to the timing of vaccination programme implementation. However there are early indications that they are beginning to increase again in October 2021.

4. Conclusions

Summary of findings

Findings have been summarised by revisiting the governments objectives:

Total vaccinations (first and second dose by region):

- Gibraltar has the highest number of partially vaccinated individuals
- Turks and Caicos have the highest percent of partially vaccinated individuals
- Saint Helena, Ascension and Tristan da Cunha have the lowest number and percentage of partially vaccinated individuals.
- Dose ratio is relatively consistent across all provinces.
- Increase in second vaccine uptake has tailed off to October 2021.

Analysis of covid deaths:

- The number of daily deaths have continued to rise throughout the period.
- Daily deaths appear to be increasing across all provinces other than Gibraltar which is showing signs the increase has slowed and may start to fall past October 2021.
- Suggests that the increase in daily deaths has not yet peaked.

Analysis of recovery rates:

- Highest number of recoveries seen in Gibraltar and Channel Islands.
- Increase in number of recoveries corresponds to the reduction in growth of daily deaths
- Sharp increases in recoveries correspond to introduction of vaccinations.

Analysis of Twitter trends:

- Covid is a top trending topic within on twitter in terms of hashtags and content of twitter posts.

Trends in Hospitalisation numbers:

- Two main peaks May 2020 and February 2021 which aligns to timing of the increase in daily deaths in this province.
- Responses to additional questions (refer to Appendix 3)

Recommendations and further analysis:

The report provides the requested information to assist government with the targeting marketing strategy in line with dose ratios, number of covid deaths and recoveries by location. To further enhance this analysis and optimize the strategy, I would recommend the following:

- Review the number of unvaccinated individuals by location, rather than just partially vaccinated
- Corroborating the trends between timing of implementation of vaccination programmes and reductions in hospitalisations and deaths, could strengthen and support the government's campaign by promoting the success and positive effects of the vaccine.
- Review and understand timing of other contributing factors that can help explain the patterns observed. For example, local laws, demographics, lockdown restrictions and quality of healthcare
- Look at percentages of partially vaccinated in relation to the entire population size of the province
- Understand the tone behind the covid trends on social media
- Explore more recent data for 2022

Appendix 1 – Vaccination data and dose ratio by region

Data has been sorted by the highest number of partially vaccinated individuals:

Province/State	Eligible	First dose	Partially vaccinated	Partially vaccinated %	Fully vaccinated %
Gibraltar	5,606,041	5,870,786	264,745	4.509532	95.490468
Montserrat	5,157,560	5,401,128	243,568	4.509577	95.490423
British Virgin Islands	4,933,315	5,166,303	232,988	4.509763	95.490237
Anguilla	4,709,072	4,931,470	222,398	4.509771	95.490229
Isle of Man	4,036,345	4,226,984	190,639	4.510048	95.489952
Falkland Islands (Malvinas)	3,587,869	3,757,307	169,438	4.509560	95.490440
Cayman Islands	3,363,624	3,522,476	158,852	4.509669	95.490331
Channel Islands	3,139,385	3,287,646	148,261	4.509640	95.490360
Turks and Caicos Islands	2,915,136	3,052,822	137,686	4.510122	95.489878
Bermuda	2,690,908	2,817,981	127,073	4.509363	95.490637
Others	2,466,669	2,583,151	116,482	4.509299	95.490701
Saint Helena, Ascension and Tristan da Cunha	2,242,421	2,348,310	105,889	4.509158	95.49084

Appendix 2 – Vaccinations and dose ratios over time

Year_Month	Eligible	First dose	Partially vaccinated	Partially vaccinated %	Fully vaccinated %
2020-01	0	0	0	NaN	NaN
2020-02	0	0	0	NaN	NaN
2020-03	0	0	0	NaN	NaN
2020-04	0	0	0	NaN	NaN
2020-05	0	0	0	NaN	NaN
2020-06	0	0	0	NaN	NaN
2020-07	0	0	0	NaN	NaN
2020-08	0	0	0	NaN	NaN
2020-09	0	0	0	NaN	NaN
2020-10	0	0	0	NaN	NaN
2020-11	0	0	0	NaN	NaN
2020-12	0	0	0	NaN	NaN
2021-01	102807	7009791	6906984	98.533380	1.466620
2021-02	424418	17988880	17564462	97.640665	2.359335
2021-03	4122064	28860884	24738820	85.717471	14.282529
2021-04	14565922	32075643	17509721	54.588839	45.411161
2021-05	25343318	37190595	11847277	31.855573	68.144427
2021-06	32656791	42574410	9917619	23.294789	76.705211
2021-07	37930766	44529811	6599045	14.819387	85.180613
2021-08	42518573	45801329	3282756	7.167382	92.832618
2021-09	44510420	46576914	2066494	4.436734	95.563266
2021-10	44848345	46966364	2118019	4.509651	95.490349

Appendix 3 – Response to additional questions

Question	Response
We have heard of both qualitative and quantitative data from the previous consultant. What are the differences between the two? Should we use only one or both of these types of data and why? How can these be used in business predictions? Could you provide examples of each?	Data typically falls between two types, qualitative and quantitative: • Qualitative data is based on interpretation, and incorporates different groupings and descriptions. This type of data allows us to understand why, how or what has happened when driven by behaviours or emotion. Examples include; colour, yes/no, emotions. • Quantitative data is numeric, countable or measurable information that can be used in calculations. It is this type of data that lends itself to the most statistical analysis and allows us to understand how many, how much or how often. Examples include; distance (km/m), time(m/h/s) Analysing both qualitative and quantitative data together will allow you to gain much deeper insights that using just one type of data alone. This is beneficial as you can begin to understand and align the 'What' and 'Why' together which will create better informed decision making.
We have also heard a bit about the need for continuous improvement. Why should this be implemented, it seems like a waste of time. Why can't we just implement the current project as it stands and move on to other pressing matters?	Continuous improvement is crucial in ensuring overall processes and overall quality is improved, for example by reducing inefficiencies and wasted time. Continual improvement can be made through small, but impactful changes, and is especially effective if an organisation can implement a culture throughout the business. Without continuously reflecting on how to improve processes, the organisation is unlikely to keep up within a dynamic environment (e.g. staying ahead of its competitors) and will repeat unfavourable elements of performance.
As a government, we adhere to all data protection requirements and have good governance in place. We only work with aggregated data and therefore will not expose any personal details. Have we covered everything from a data ethics standpoint? Is there anything else we need to implement from a data ethics perspective?	When dealing with data, it is important to ensure all individuals within the organisation are aware of and adhere to the legal requirements and governance practices. As well as this, you should consider not just the basic legal requirements , but also the ethical implications of handling data. I would recommend the implementation of a robust ethics framework which will ensure ethical data practices and uses meet the expectations of the public and other stakeholders. The framework should define data ownership and responsibility to cover who is responsible for areas such as data collection and processing. The framework should be regularly revisiting and tested to ensure it is being implemented effectively.