UK COVID-19 Data Analysis

INTRODUCTION

Problem Statement

UK Government wants to increase the number of fully vaccinated people against COVID-19. To this end, the government wants to gather information and insights to help inform their marketing efforts to promote the vaccine.

Context

For the analysis, COVID-19 cases and vaccination information is provided for different provinces/states for the period, January 2020 to October 2021. The distinct names present in the data are the names of British Overseas Territories, along with a name 'Others'. The aim is to understand which regions to focus on for the government's marketing campaigns.

ANALYSIS

DATA

Two datasets are provided, one containing daily data of vaccination doses for each province/state, and the other with daily data of number of cases, deaths, recoveries and hospitalised individuals for each province/state. These are the key fields among other fields provided.

There is a name called 'Others' as a province/state in the dataset, so it is considered as mainland UK as an assumption. The key fields mentioned above have been used and the other fields have been excluded for the analysis.

Population data for each province/state is not provided in the dataset.

Findings

Several anomalies and inconsistencies are found both in daily data and aggregated values.

- Hospitalised numbers are more than number of cases both on a given date and when seen by province/state
- Hospitalised numbers are not only higher but they do not match with the difference between cases and recovered on a given date

- Number of cases, deaths and recoveries seem to be running totals but not the hospitalised numbers
- Recovered numbers are more than hospitalised numbers in some cases
- As population data is not provided in the dataset, to help determine the cause of anomalies, external data was gathered to cross check the numbers with the population numbers. A quick glance at the numbers (Figure 1 and Figure 2), tells us the actual population in each province/state is way lower than the aggregated values of doses or hospitalised numbers. This presents a challenge by preventing us to work with numbers per equal part of population (for ex. Cases per million)

Province/State	Hospitalised	Vaccinated	First Dose	Second Dose	Partially Vaccinated
Gibraltar	649459	5606041	5870786	5606041	264745
Montserrat	597486	5157560	5401128	5157560	243568
British Virgin Islands	571506	4933315	5166303	4933315	232988
Anguilla	545540	4709072	4931470	4709072	222398
Isle of Man	467605	4036345	4226984	4036345	190639
Falkland Islands (Malvinas)	415650	3587869	3757307	3587869	169438
Cayman Islands	389669	3363624	3522476	3363624	158852
Channel Islands	363690	3139385	3287646	3139385	148261
Turks and Caicos Islands	337710	2915136	3052822	2915136	137686
Bermuda	311547	2690908	2817981	2690908	127073
Others	285768	2466669	2583151	2466669	116482
Saint Helena, Ascension and Tristan da Cunha	259773	2242421	2348310	2242421	105889

Figure 1: Data from Analysis

	Name	Population
0	Anguilla	14,869
1	Bermuda	62,506
4	British Virgin Islands	31,758
5	Cayman Islands	69,656
6	Falkland Islands	3,377
7	Gibraltar	33,701
8	Montserrat	5,215
10	Saint Helena, Ascension and Tristan da Cunha,i...	5,633
16	Turks and Caicos Islands	38,191

Figure 2. Data source:
https://en.wikipedia.org/wiki/British_Overseas_Territories

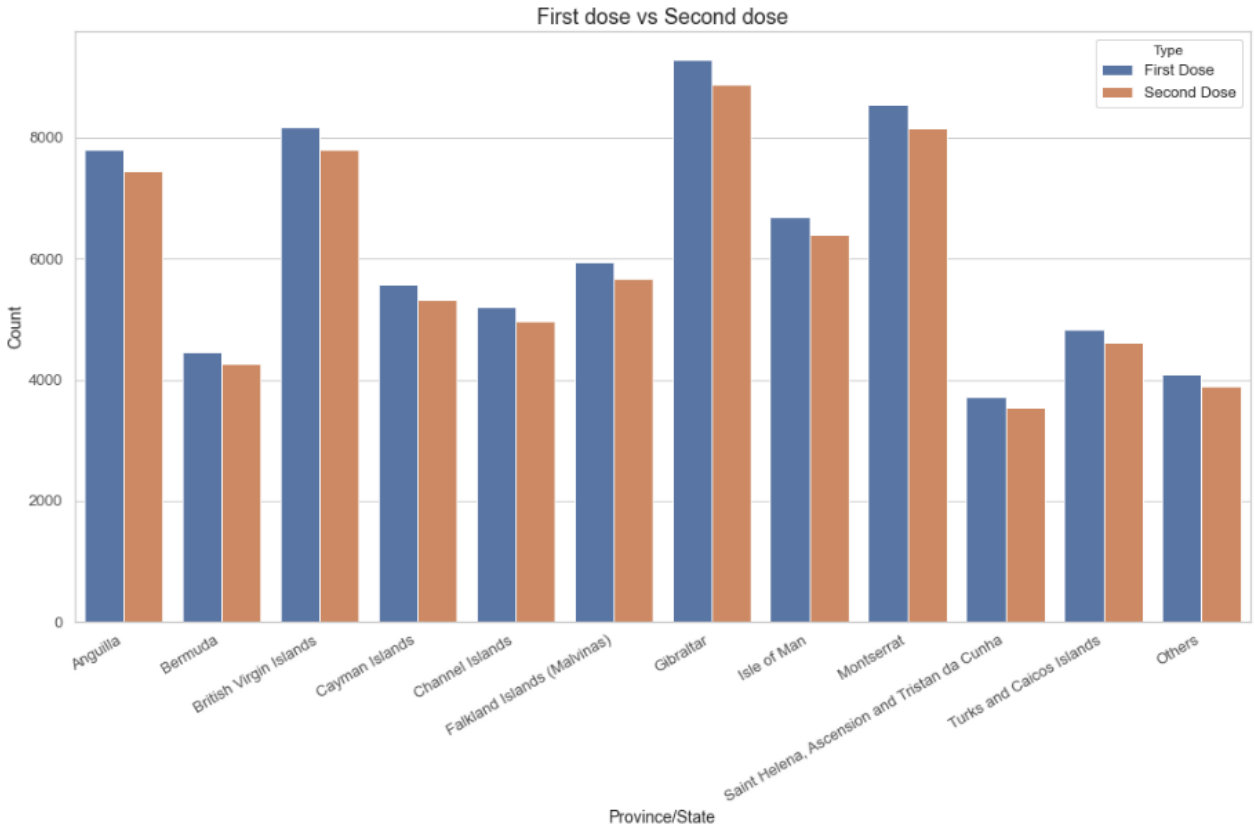

Figure 3 : First dose vs Second dose, from analysis

Exploring the data further, we can see in this graph (Figure 3), the number of first doses in comparison with second doses for each province/state.

It can be observed that the difference is not notably high for any particular region, while 'Gibraltar' shows the highest difference.

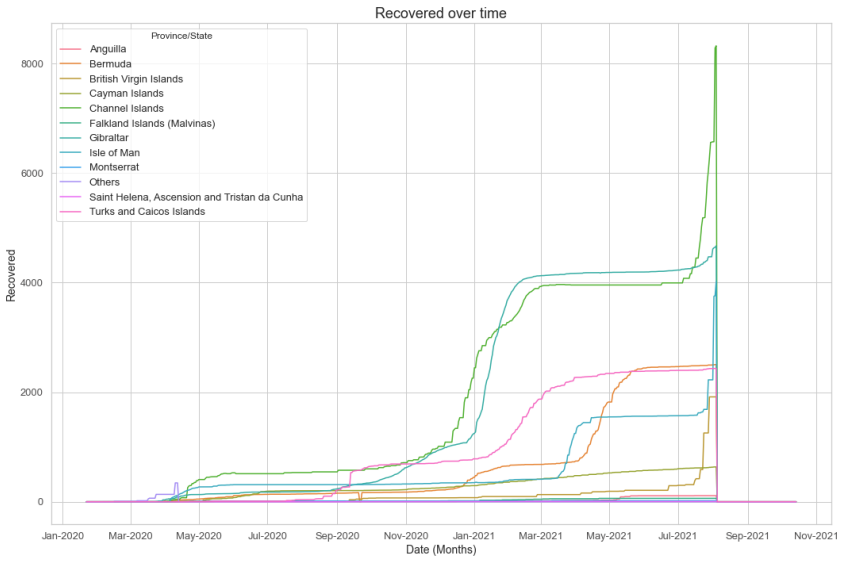

Figure 4: Recoveries over time, from analysis

The graph in Figure 4 shows the recoveries over time.

'Channel Islands' and 'Isle of Man' can be avoided from initial marketing campaign runs as their recovery numbers are relatively higher than other regions.

Deaths over time graph (Figure 5) shows that the deaths have not really reached a peak and seem to be increasing.

It is to note that these numbers are not numbers per equal part of the population and inconsistencies in the population and other numbers need to be taken into consideration.

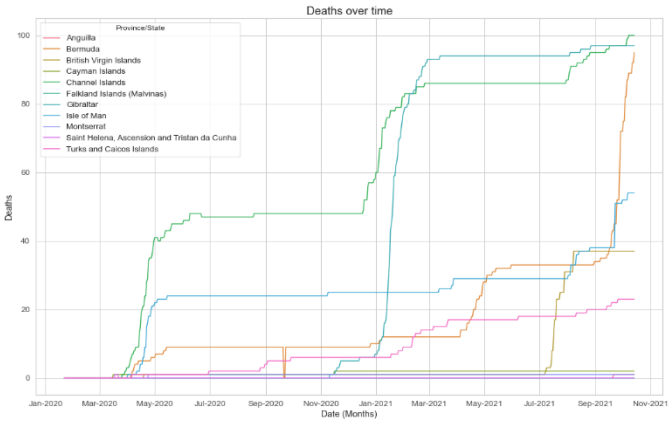

Figure 5: Deaths over time, from analysis

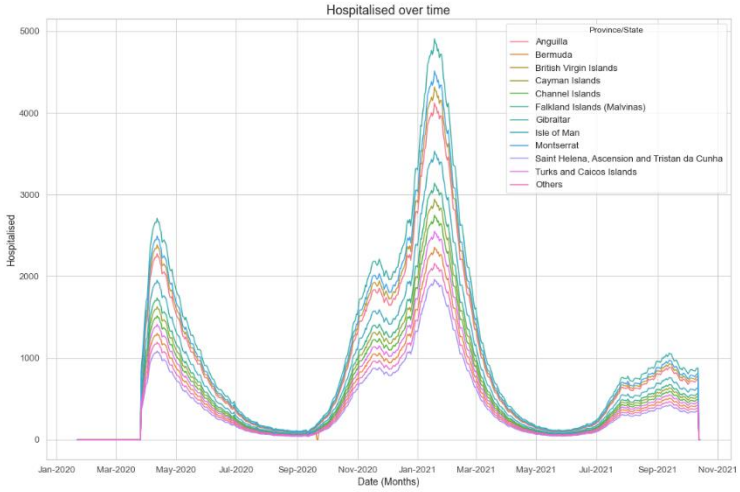

Figure 6: Hospitalised over time, from analysis

While hospitalised trends (Figure 6) show that a peak has been reached during Jan-Feb 2021, the patterns look extra ordinarily similar across all the provinces/states. This is another aspect contributing to the need for further inspection of the data.

Possible Causes

COVID-19 is a global pandemic and obtaining and keeping track of data is extremely challenging. The process is complex and has several factors contributing to right or wrong data. Although it is not possible to pinpoint the cause for inconsistencies in the data with initial analysis, further iterations can help determine the reasons. Looking at the results of this analysis, a few possible causes could be:

- Data collection without a single point of entry, leading to possible duplicated entries
- Inconsistent data collection processes or lack of integration across departments or regions
- Incomplete information or knowledge gaps
- Human error

RECOMMENDATIONS AND FURTHER STEPS

All the results presented can vary when the quality of the data is attended to and inconsistencies addressed. Investigating the data further in the following iterations is a necessary step to better understand the anomalies in the data, as problems in the data has a definite impact on the analysis. Some key questions towards that direction are:

- Looking for the availability of numbers per a part of population (For ex, cases per Thousand etc) gives better results to gather useful insights
- Understanding why certain fields are cumulative and some not
- Why the aggregated values are not matching with the population for each region
- Why the trends for hospitalised individuals across regions look extraordinarily similar
- Looking for the availability or establish a data dictionary to understand the meaning of certain fields like Recoveries.

Recommendations:

- Keeping in view of the limitations of the quality of the data, the results of analysis show that 'Gibraltar', 'Montserrat' and 'British Virgin Islands' are the regions with highest number of individuals who had first dose but no second dose.
- 'Channel Islands' and 'Isle of Man' are regions with most recoveries, so could be avoided for initial marketing campaigns.
- Data management and maintaining data quality as a continuous improvement process rather than one-time project.
- Investigating the source of data collection
- Looking at the benefits or need for a panel consensus about obtaining demographics of the individuals that are not personal data like age, gender etc and any other relevant data, which can be used for further analysis and also a possibility of conducting market research to complement it.
- Following ethical guidelines pertaining to all actions involving data.