Dan Thorneloe 14/07/2022

LSE Data Analytics Accelerator
Course 2 Assignment

Context and Problems Statement

The government wants to identify trends and patterns that can be used to inform its marketing approach to increase the number of fully vaccinated people to COVID-19.

Problem statement: maximise vaccine marketing effectiveness (suitable KPIs could be increase in daily vaccinations and reduction in % of first dose recipients not taking second for example)

Specific questions are:

- What the total vaccinations are for a particular region.
- Where they should target the first marketing campaign(s)
- What tweets have both #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

There are other factors that could influence vaccine take-up e.g., demographics, logistics and vaccine availability which are not considered here. Also, only Twitter was looked at in terms of media not other channels or mediums.

Exploring data

Three data sources:

1. Covid cases, hospitalisations, recovered, and deaths by State in Great Britain.
2. Vaccination volumes (Vaccinated, First Dose, Second Dose) by State
3. Twitter data sample

Approach: data was profiled using functioning such as info(), describe(); tabulated, visualised in simple line plots and aggregated to provide a feel for the data, spot missing data and outliers e.g., data from the 'Others' province/state.

Initial findings:

- 2 rows with null values present in data set for Bermuda (Deaths, Recovered, Hospitalisation, Cases data missing).
- 'Others' Province is assumed to be GBR given the volumes. This has values higher than other Provinces so may cause issues with visualisation.
- 'Recovered' data for GBR is blank. No suitable proxy available so left blank.
- Vaccinated counts = second does count so duplicate data
- All provinces show the same ratio of second to first dose (checked also in excel) so seems to be modelled not actual data
- It is not clear from the data dictionary what the values are actually showing:
    - Initially I surmised that, given shape of lineplot, Cases, Recovered and Deaths were cumulative data (positive gradient) and Hospitalisations, First Dose, Second Dose were daily totals (noisy data, declining over time). Therefore, the max of cumulative data would equal the total not the sum.

  - However, comparing totals in this way created erroneous results e.g. more Hospitalised than Cases so surmised that all data was daily and should be summed to get totals
  - However, summing generates totals that are far higher than the population of the provinces e.g., if the data is daily vaccinations, then in a single day, 20/3/2021, 94038 receive a first does. 3x the population of Gibraltar. However, dose data cant be cumulative as values fall over time

Outcomes of data exploration

- HEALTH WARNING. I cannot reconcile some of the data. Further investigation of data source is required. However, to proceed I have assumed all data is daily and the trends will still be of value
- Volume are higher than populations can sustain so assume this is example data for this Assignment. In real-world data source would be interrogated further so resolve anomalies.
- 'Others' Province (assume GBR) excluded form final analysis. Recommend separate analysis.
- Missing values in Bermuda replaced with mean values for each relevant column for Bermuda

There seemed to be some debate about the accuracy of the data on the community so please note assumptions above when considering findings.

Initial Trend Findings (Gibraltar data only)

- Hospitalised shows spike in April 2020 with only a relatively minor jump in Cases. This could be explained as COVID being new and protections not in place. However, there was no accompanying rise in Deaths so worth exploring further e.g. non-covid hospitalisations or Deaths data issue
- Cases rise sharply from end of 2020
- Hospitalisations rise sharply in Q1 2021
- Deaths rise sharply in Q1 2021
- Recovered follows Cases with lag. N.B. data stops being recorded August 2021.
- Immunisation programme starts Q1 2021
- High rates of immunisation leads to hospitalisation falling and Deaths holding steady but interestingly not falling. Requires further analysis
- Hospitalisations and Deaths begin to rise again in Q4 2021 as either vaccine effectiveness falls (particularly the 4.5% not taking the second dose) or new variant arrives

Dan Thorneloe 14/07/2022

Response to asked questions

What the total vaccinations (first dose, second dose per region) are for a particular region.

Fig 1: First dose per region

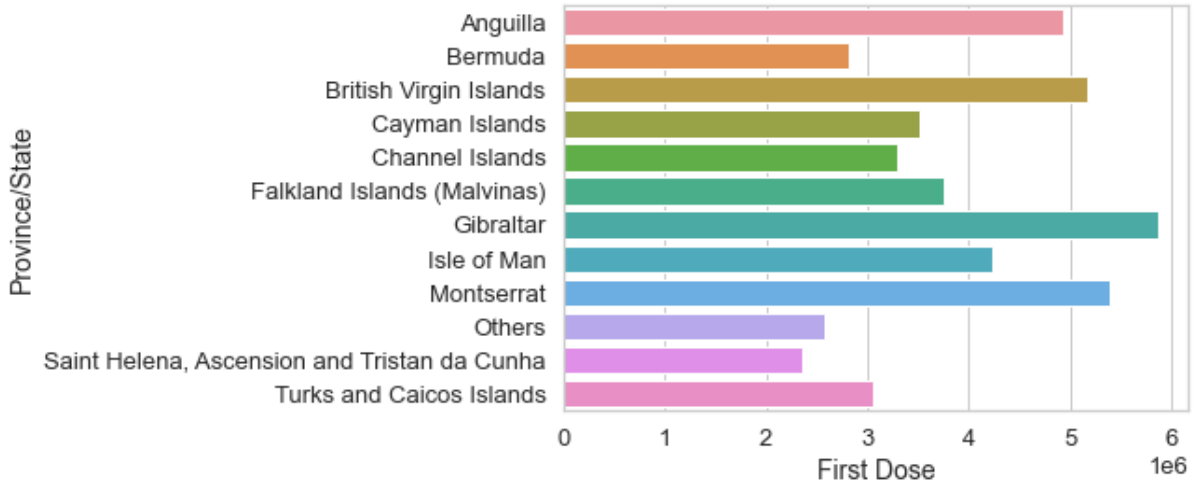
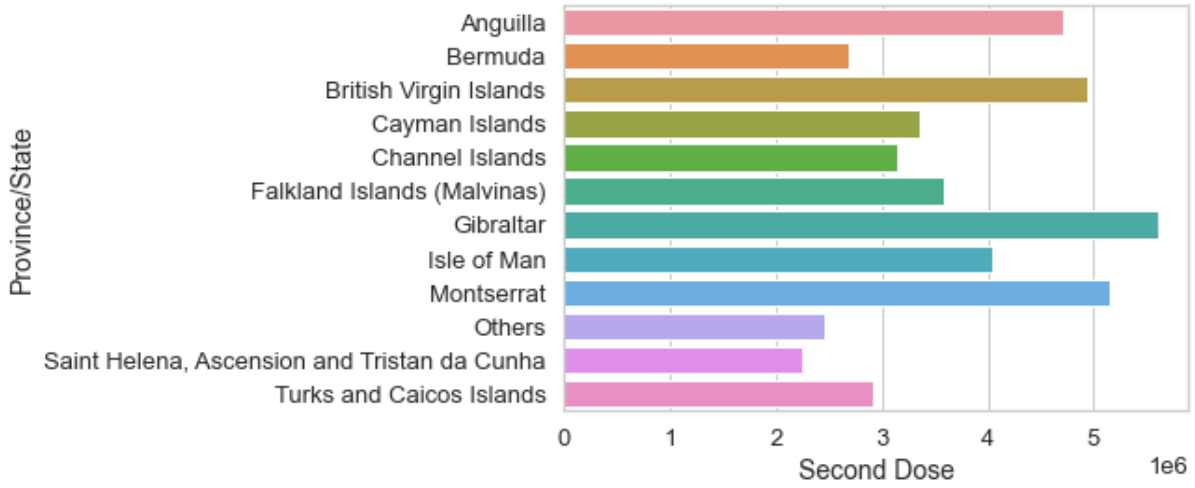


Fig2: Second Dose per region

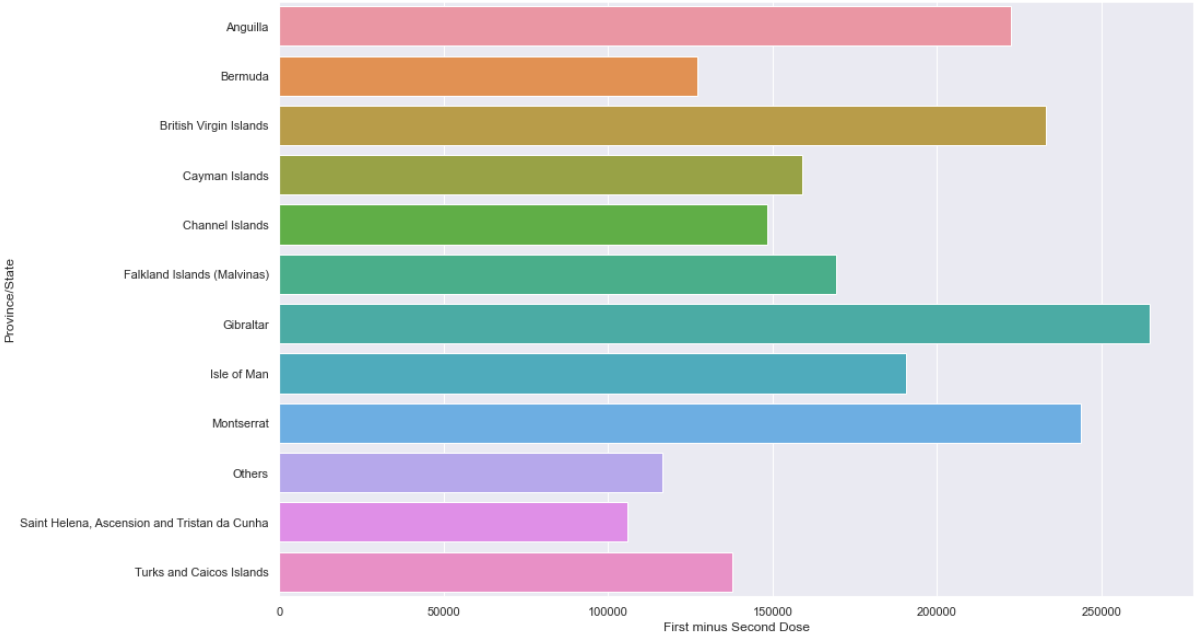


The rate of Second dose to First Dose was 95% for all regions

Where they should target the first marketing campaign(s) based on:

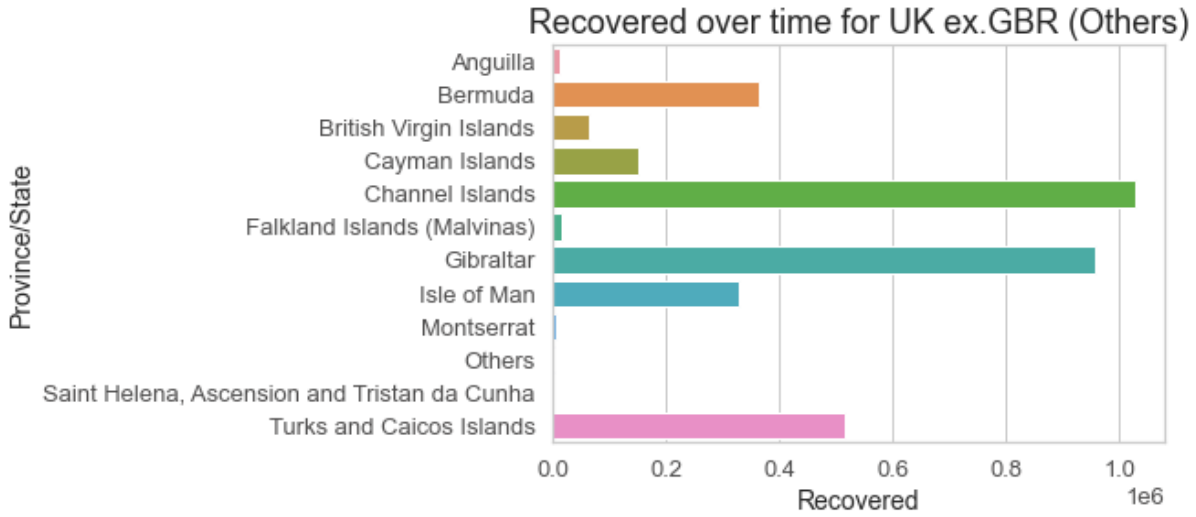
- Area(s) with the largest number of people who have received a first dose but no second dose – Gibraltar. Overall 4.5% of first doses are not getting there second.

Fig. 3: Difference between the first dose and second dose by region



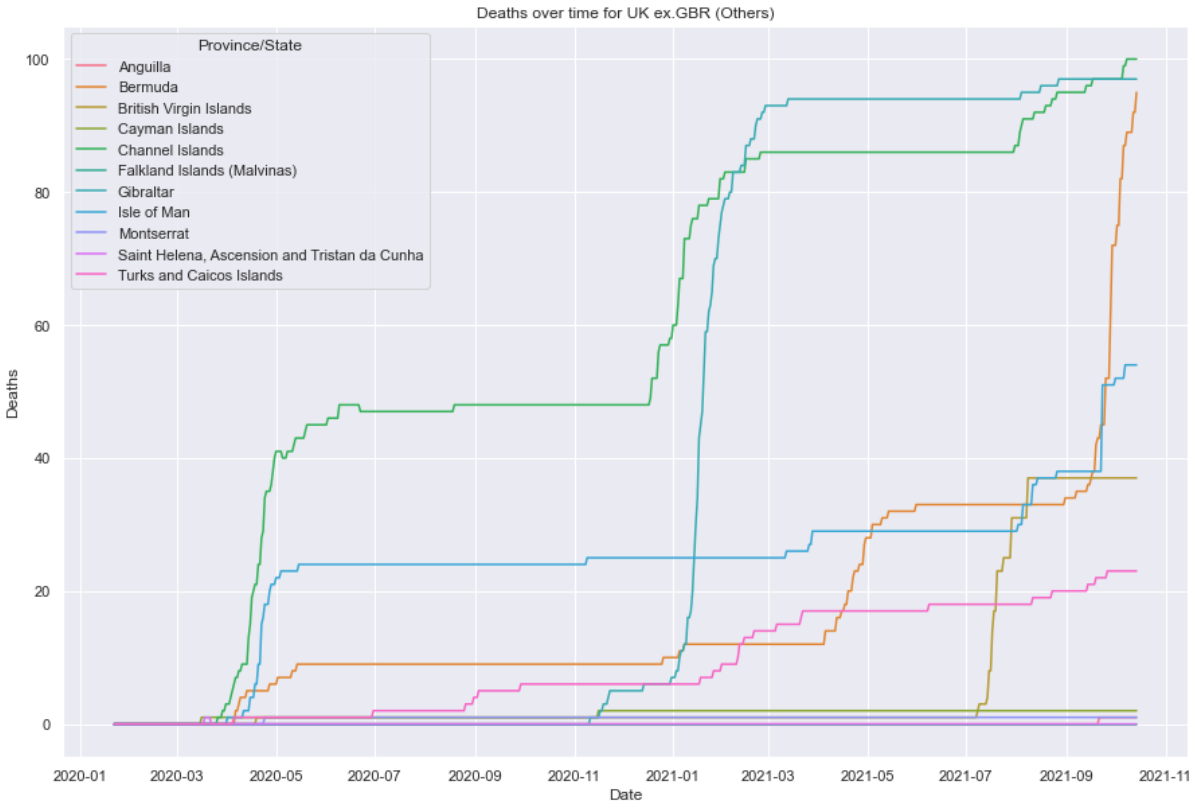
- Which area has the greatest number of recoveries so that they can avoid this area in their initial campaign runs – Channel Islands and Gibraltar

Fig 4: Total Recovered by Province



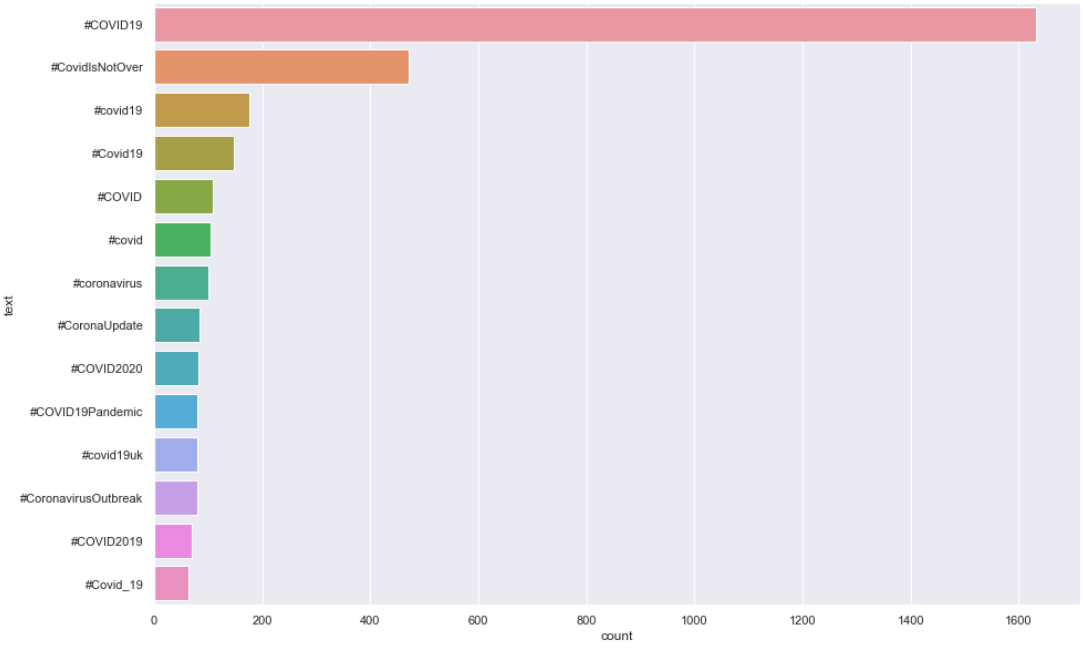
- Whether deaths have been increasing across all regions over time or if a peak has been reached - Deaths are increasing sharply for Bermuda so most urgent. Channel Islands are rising after plateau so worth observation. Isle of Man also on rise.

Fig. 5 Daily death volumes over time

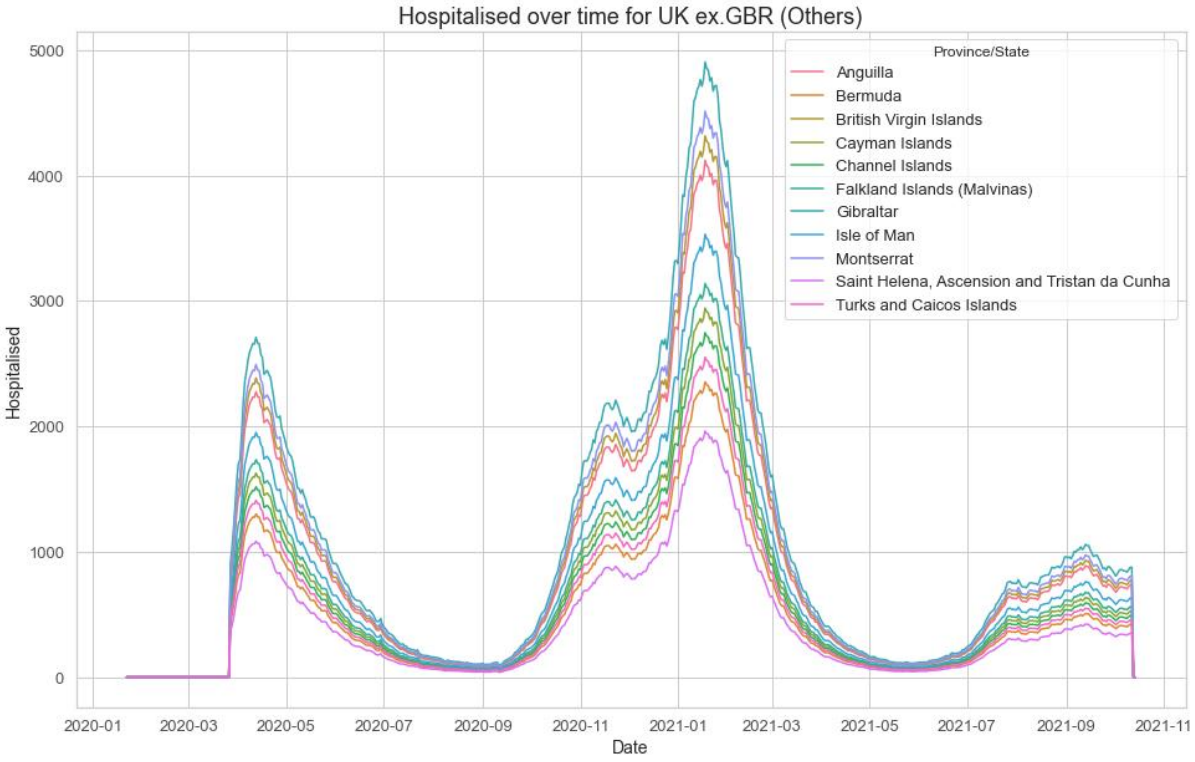


- What tweets have both #coronavirus and #vaccinated hashtags – #COVID19 is by far the most popular hashtag related to COVID-19 #vaccinated does not rank in this view.

Fig 6: highest trending ##s relating to COVID-19 or vaccinations



- Which regions have experienced a peak in hospitalisation numbers and if there are regions that have not reached a peak yet – on the data available peak hospitalisation have been reached. However, cases, deaths and hospitalisations are rising again for a number of regions for latest data so cannot know if a further peak is pending.



Conclusion

- The government should focus on Gibraltar where we see the highest absolute gap between First and Second Doses. Their social media campaign should include the #COVID19 to reach the widest audience.
- However, death rates are rising fastest in Bermuda, the Channel Islands and Isle of Man which warrant attention

Further investigation improvements

- Investigate data sources to clear up anomalies identified earlier
- Use total population data (from csv or api) to investigate % of population
- I would add in population data to compare these values to the overall population i.e., what % of population is vaccinated.
- GBR, as largest geography, would require regional (e.g. county) breakdown to be useful
- Improve visuals by aggregating monthly values