COVID-19 Analysis

Yiying Zhong

Scenario

This project aims to help the government to better promote COVID-19 vaccinations across regions. The UK government is planning to launch a series of marketing campaigns to increase the number of fully vaccinated individuals (people who have received two doses of the vaccine). By looking at COVID-19 data (from January 2020 to October 2021), the government wants to identify trends and patterns that helps to inform its marketing approach. In particular they want to understand:

- What the total vaccinations (first dose, second dose per region, total and overtime) are for a particular region.
- 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
    - which area has the greatest number of recoveries so that they can avoid this area in their initial campaign runs
    - whether deaths have been increasing across all regions over time or if a peak has been reached.
- What other types of Twitter data points and 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.

Project access

This project is available for access on GitHub (https://github.com/yinggz12/LSE_DA_COVID_analysis). GitHub repo is a great way to track changes and handle snapshot of code and project. This is especially useful for projects

involving coding, as version control systems allows user to revert back to previous copy when error introduced during editing or accidental deletion.

About the data

The two main datasets used in this analysis are COVID-19 cases (covid_19_uk_cases.csv) and vaccinations (covid_19_uk_vaccinations.csv) in the UK. Below shows details of variables included in the dataset.

covid_19_uk_cases.csv

Column	Description
Province/State	The state or province, for example, Bermuda and Gibraltar.
Country/Region	The country, for example, United Kingdom.
Lat	The latitude (location) coordinates of the Province/State.
Long	The longitude (location) coordinates of the Province/State.
ISO 3166-1 Alpha 3-Codes	The three-letter country abbreviation as defined in the ISO 3166 standard published to represent countries, for example, BMU for Bermuda and GIB for Gibraltar.
Sub-region Name	The subregion of the province/state, which is a part of a larger continent usually based on location, for example, Bermuda and Gibraltar's subregions are Northern America and Southern Europe, respectively.
Intermediate Region Code	The codes for the intermediate regions (regions that are neither rural nor urban), for example, the Intermediate Region Code for Anguilla is 29.
Date	The date the figures are attributed to, written using the format: YYYY/MM/DD.
Deaths	The number of deaths attributed to COVID-19.
Cases	The number of cases where individuals are COVID-19 positive.
Recovered	The number of cases where COVID-19 positive individuals have recovered.
Hospitalised	The number of cases where COVID-19 positive individuals are hospitalised.

covid_19_uk_vaccinations.csv

Column	Description
Province/State	The state or province, for example, Bermuda and Gibraltar.
Country/Region	The country, for example, United Kingdom.
Lat	The latitude (location) coordinates of the Province/State.

Long	The longitude (location) coordinates of the Province/State.
ISO 3166-1 Alpha 3-Codes	The three-letter country abbreviation as defined in the ISO 3166 standard published to represent countries, for example, BMU for Bermuda and GIB for Gibraltar.
Sub-region Name	The subregion of the province/state, which is a part of a larger continent usually based on location, for example, Bermuda and Gibraltar's subregions are Northern America and Southern Europe, respectively.
Intermediate Region Code	The codes for the intermediate regions (regions that are neither rural nor urban), for example, the Intermediate Region Code for Anguilla is 29.
Date	The date the figures are attributed to written using the format: YYYY/MM/DD.
Vaccinated	The number of individuals who are fully vaccinated. 'Fully vaccinated' means that the individual has received two doses of the vaccine.
First Dose	The number of individuals who received the first dose of the vaccine.
Second Dose	The number of individuals who received the second dose of the vaccine.

Validating data

The two datasets recorded COVID cases and vaccinations data from 2020-01-22 to 2021-10-14. There are 7584 records found in both datasets, in which includes data from 12 different regions (Anguilla, Bermuda, British Virgin Islands, Cayman Islands, Channel Islands, Falkland Islands (Malvinas), Gibraltar, Isle of Man, Montserrat, Saint Helena, Ascension and Tristan da Cunha, Turks and Caicos Islands, Others).

COVID cases dataset quick look:

	Province/State	Country/Region	Lat	Long	ISO 3166-1 Alpha 3-Codes	Sub-region Name	Intermediate Region Code	Date	Deaths	Cases	Recovered	Hospitalised
0	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-22	0.0	0.0	0.0	0.0
1	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-23	0.0	0.0	0.0	0.0
2	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-24	0.0	0.0	0.0	0.0
3	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-25	0.0	0.0	0.0	0.0
4	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-26	0.0	0.0	0.0	0.0

Vaccination dataset quick look:

	Province/State	Country/Region	Lat	Long	ISO 3166-1 Alpha 3-Codes	Sub-region Name	Intermediate Region Code	Date	Vaccinated	First Dose	Second Dose
0	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-22	0	0	0
1	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-23	0	0	0
2	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-24	0	0	0
3	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-25	0	0	0
4	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-26	0	0	0

In the process of data validation, two null records found in COVID cases data. They are from Bermuda on date 2020-09-21 and 2020-09-22, these two records are then replaced with previous record using forward fill method.

	Province/State	Country/Region	Lat	Long	ISO 3166-1 Alpha 3-Codes	Sub-region Name	Intermediate Region Code	Date	Deaths	Cases	Recovered	Hospitalised
875	Bermuda	United Kingdom	32.3078	-64.7505	BMU	Northern America	0	2020-09-21	NaN	NaN	NaN	NaN
876	Bermuda	United Kingdom	32.3078	-64.7505	BMU	Northern America	0	2020-09-22	NaN	NaN	NaN	NaN

In COVID cases data, variables Deaths, Cases and Recovered are displayed as cumulative data while hospitalised and vaccination data (Vaccinated, First Dose and Second Dose) are daily data.

Initial insights

To get a better understanding of the data, we subset the data to only Gibraltar region. From Gibraltar, a total of 632 records found. Throughout the whole timeline of this subset, there 97 deaths and 5727 cases recorded, but hospitalised number have a maximum of 4907 on a daily record. After looking into the data, it suggests that cases recorded are not very accurate, as on some days the number of hospitalised COVID positive individuals are greater than the total number of cases.

	Deaths	Cases	Recovered	Hospitalised
count	632.000000	632.000000	632.000000	632.000000
mean	40.208861	2237.109177	1512.821203	1027.625000
std	45.332832	2136.268090	1817.096755	1145.681058
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	177.000000	109.500000	157.750000
50%	5.000000	1036.500000	323.500000	675.500000
75%	94.000000	4286.000000	4122.500000	1548.000000
max	97.000000	5727.000000	4670.000000	4907.000000

In the vaccinations data (Gibraltar only), actual record of data only starts appearing in January 2021, the vaccination program only begins then. Variable Vaccinated and Second Dose are

equivalent as Vaccinated is defined as people who have received a second dose of vaccine. After aggregating the data, we found that in Gibraltar region, there are a total of 5606041 people fully vaccinated. This raises question, as Gibraltar only has a population around 30k (2020), and total number of people fully vaccinated in the region is over 5 million.

If we ignore this factor for now, we can see that vaccinations over time matches as expected. Number of people who receive the first dose quickly increases in the first few months, reaching its peak on month 3, while number of people receiving the second dose lags behind 2 months. This is due to the time gap between the first dose and the second dose of vaccinations. The number gradually decreases as most people are vaccinated.

		Vaccinated	First Dose	Second Dose
Year	Month			
2020	1	0	0	0
	2	0	0	0
	3	0	0	0
	4	0	0	0
	5	0	0	0
	6	0	0	0
	7	0	0	0
	8	0	0	0
	9	0	0	0
	10	0	0	0
	11	0	0	0
	12	0	0	0
2021	1	12851	876224	12851
	2	40201	1372385	40201
	3	462203	1358999	462203
	4	1305483	401847	1305483
	5	1347172	639366	1347172
	6	914184	672977	914184
	7	659247	244424	659247
	8	573475	158939	573475
	9	248983	96945	248983
	10	42242	48680	42242

Further investigations

After looking at a subset of the data, the analysis is extended to the whole data. This begins with merging two dataset, using Province/State, Country/Region and Date as key. Then only keep useful columns for this analysis for easier access, the columns used are: Province/State, Country/Region, Date, Vaccinated, First Dose, Second Dose, Deaths, Cases, Recovered, Hospitalised.

Vaccination data are aggregated (sum) per month, as the same approach that applied for Gibraltar subset. The table below shows a similar pattern as the subset, where first dose reaches peak quickly within the first 3 months, and second dose lags behind 2 months.

Year	Month	Vaccinated	First Dose	Second Dose
2021	1	102807	7009791	102807
	2	321611	10979089	321611
	3	3697646	10872004	3697646
	4	10443858	3214759	10443858
	5	10777396	5114952	10777396
	6	7313473	5383815	7313473
	7	5273975	1955401	5273975
	8	4587807	1271518	4587807
	9	1991847	775585	1991847
	10	337925	389450	337925

After looking at aggregation per month, we turn to look at aggregation per region where this can help to find which region is best to target for promoting vaccinations. The table below shows total vaccinations per region and it is sorted by region with the highest number of people received first dose of vaccine only (First Dose- Vaccinated). Here we can see that Gibraltar region has over 250k of people only received first dose, based on this result, Gibraltar would be the top one to target when comes to vaccination campaigns.

	Province/State	Vaccinated	First Dose	Second Dose	First Dose only
6	Gibraltar	5606041	5870786	5606041	264745
8	Montserrat	5157560	5401128	5157560	243568
2	British Virgin Islands	4933315	5166303	4933315	232988
0	Anguilla	4709072	4931470	4709072	222398
7	Isle of Man	4036345	4226984	4036345	190639
5	Falkland Islands (Malvinas)	3587869	3757307	3587869	169438
3	Cayman Islands	3363624	3522476	3363624	158852
4	Channel Islands	3139385	3287646	3139385	148261
11	Turks and Caicos Islands	2915136	3052822	2915136	137686
1	Bermuda	2690908	2817981	2690908	127073
9	Others	2466669	2583151	2466669	116482
10	Saint Helena, Ascension and Tristan da Cunha	2242421	2348310	2242421	105889

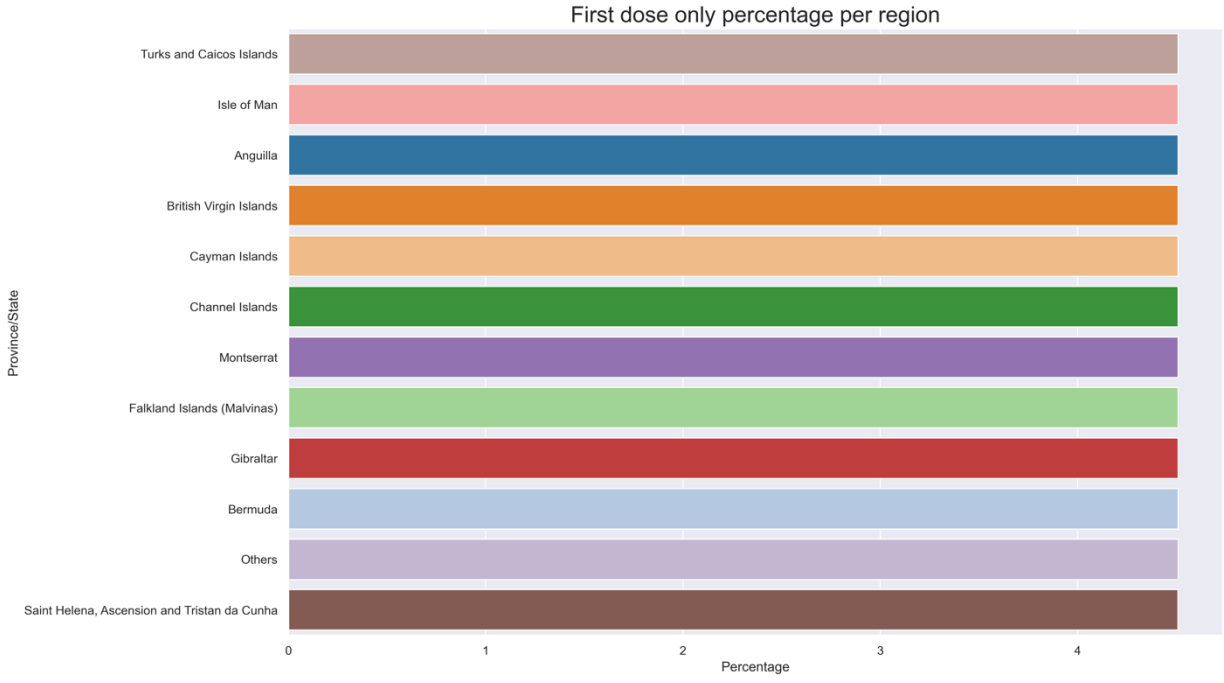
To see a comparison between fully vaccinated and first dose only, a percentage is calculated. The formula used is First dose only/First dose. Here we can see that, despite the actual number of people received first dose only varies a lot, the percentage is relatively similar, all around 4.5% of people received first dose only. This percentage is surprisingly even across regions, and not too high compared to people who has received a vaccine. This suggests that the government's initial market campaign distributed evenly across all regions and equally effective.

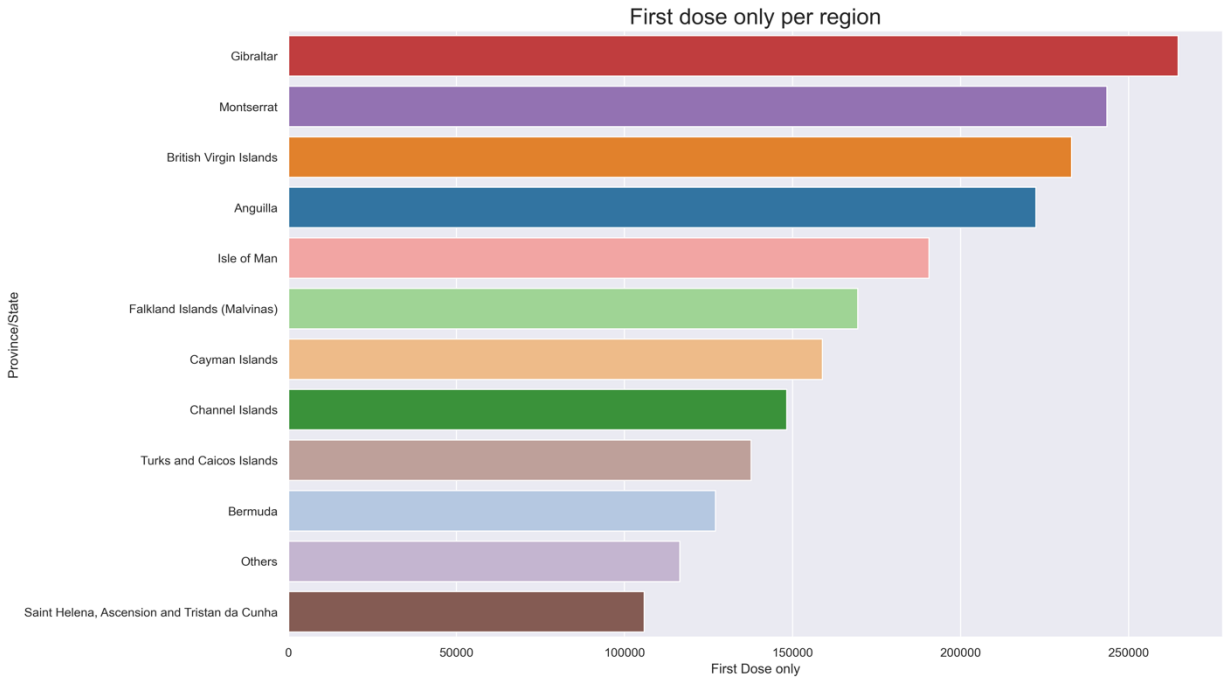
	Province/State	Vaccinated	First Dose	Second Dose	First Dose only	Percentage
11	Turks and Caicos Islands	2915136	3052822	2915136	137686	4.510122
7	Isle of Man	4036345	4226984	4036345	190639	4.510048
0	Anguilla	4709072	4931470	4709072	222398	4.509771
2	British Virgin Islands	4933315	5166303	4933315	232988	4.509763
3	Cayman Islands	3363624	3522476	3363624	158852	4.509669
4	Channel Islands	3139385	3287646	3139385	148261	4.509640
8	Montserrat	5157560	5401128	5157560	243568	4.509577
5	Falkland Islands (Malvinas)	3587869	3757307	3587869	169438	4.509560
6	Gibraltar	5606041	5870786	5606041	264745	4.509532
1	Bermuda	2690908	2817981	2690908	127073	4.509363
9	Others	2466669	2583151	2466669	116482	4.509299
10	Saint Helena, Ascension and Tristan da Cunha	2242421	2348310	2242421	105889	4.509158

Visualise

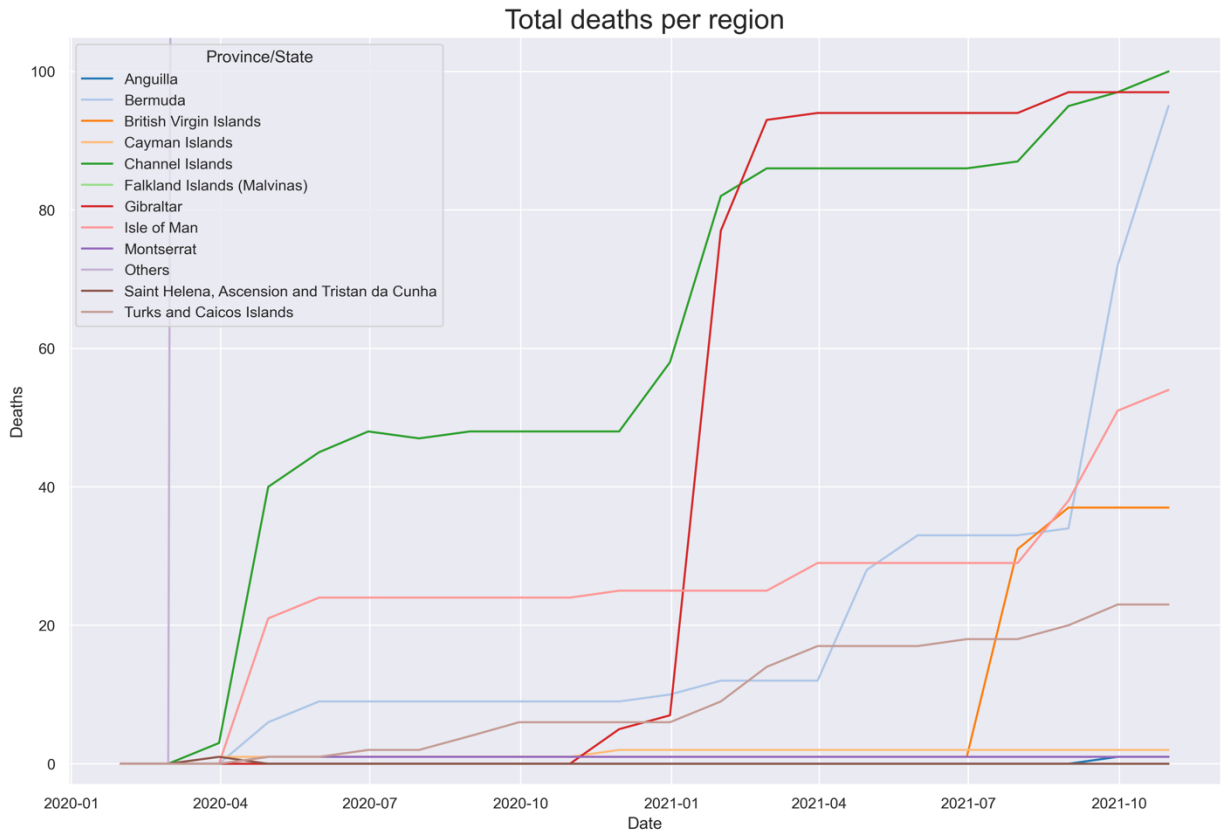
As numbers in tables are more difficult to see patterns and trends. Bar plots are created for the two tables displayed above.

We can see that from the two plots below, when showing first dose only as a percentage, no difference between regions can be seen by eye. However, when displayed as actual numbers a clear distinct ranking shown. In Saint Helena, Ascension and Tristan da Cunha has around 100k of people who received first dose only, while Gibraltar has more than double amount of people not fully vaccinated.

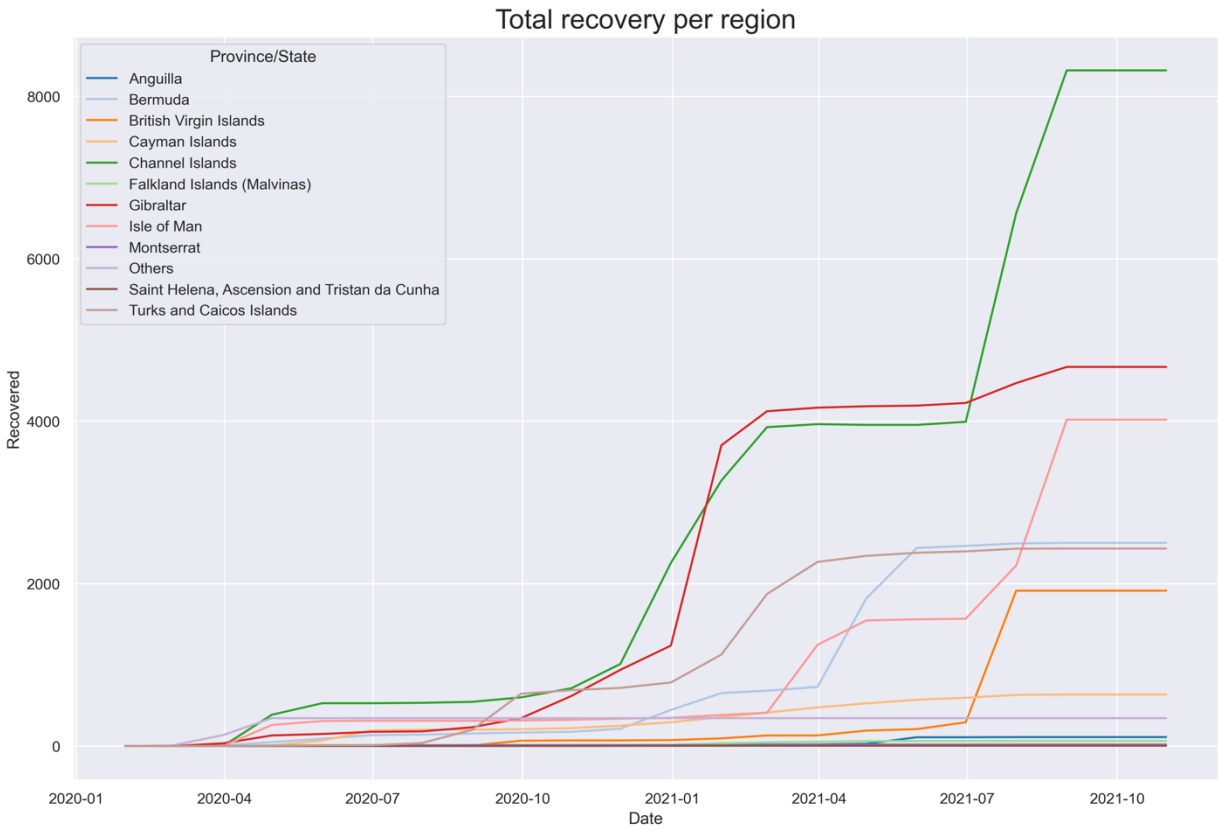




To help better decide which region to target deaths and recovery data are also considered. Deaths and recovered variables are smoothed out by taking monthly data instead of daily. Here, the approach is to group the data by region and date (monthly) then take the maximum value. On the initial plot of the timeseries, deaths count in Others are much higher than other regions, causing the plot to be skewed. The plot below limits the y-axis so that trends for most regions can be shown.

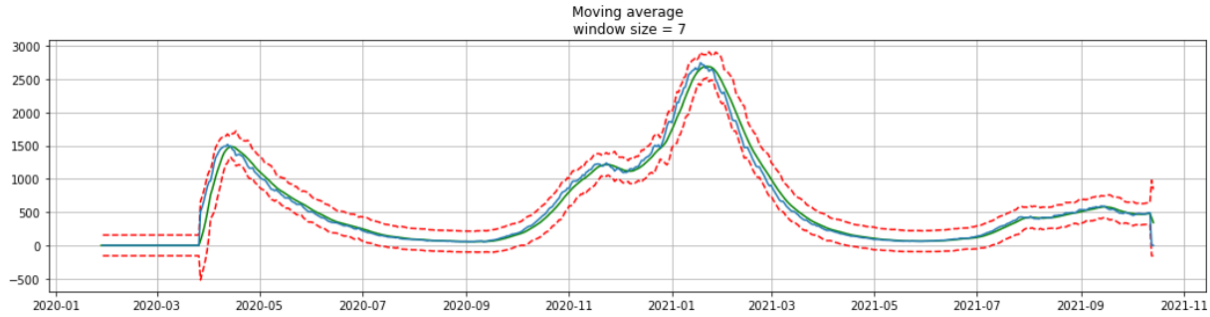


Data from Others region are not removed at this stage, as not enough evidence shown that it is an outlier, and this can also keep same colour coding for all regions related plot. We can see that for most regions deaths count reaches its second peak around March and April 2021.



Vaccinations started in January 2021, from the recovery plot suggests that there is a large number of recoveries after vaccination program starts. The total number of recoveries then flattens out, this can also suggest a slowdown of COVID cases. However, looking at the total deaths count, despite a similar pattern of flattening out, the number did not seem to be lower than the previous peak in May 2020.

The moving average plot below shows daily hospitalised data for Channel Islands. This plot shows that while hospitalised number peaks around January 2021 and February 2021, it drops much quickly than the first peak in April 2020. This further supports that since the vaccination program started in 2021 it has helped to reduce the number of COVID-19 patients.



Additional Data

Additional data such as tweets and trending tags are also looked at to see if it can bring further insights to help the government's marketing strategy. The dataset tweets.csv, is a limited sample of twitter data (data ranges from 2022-05-15 to 2022-05-23), but the approach to analysis this dataset can be repeated on a richer dataset with similar structure.

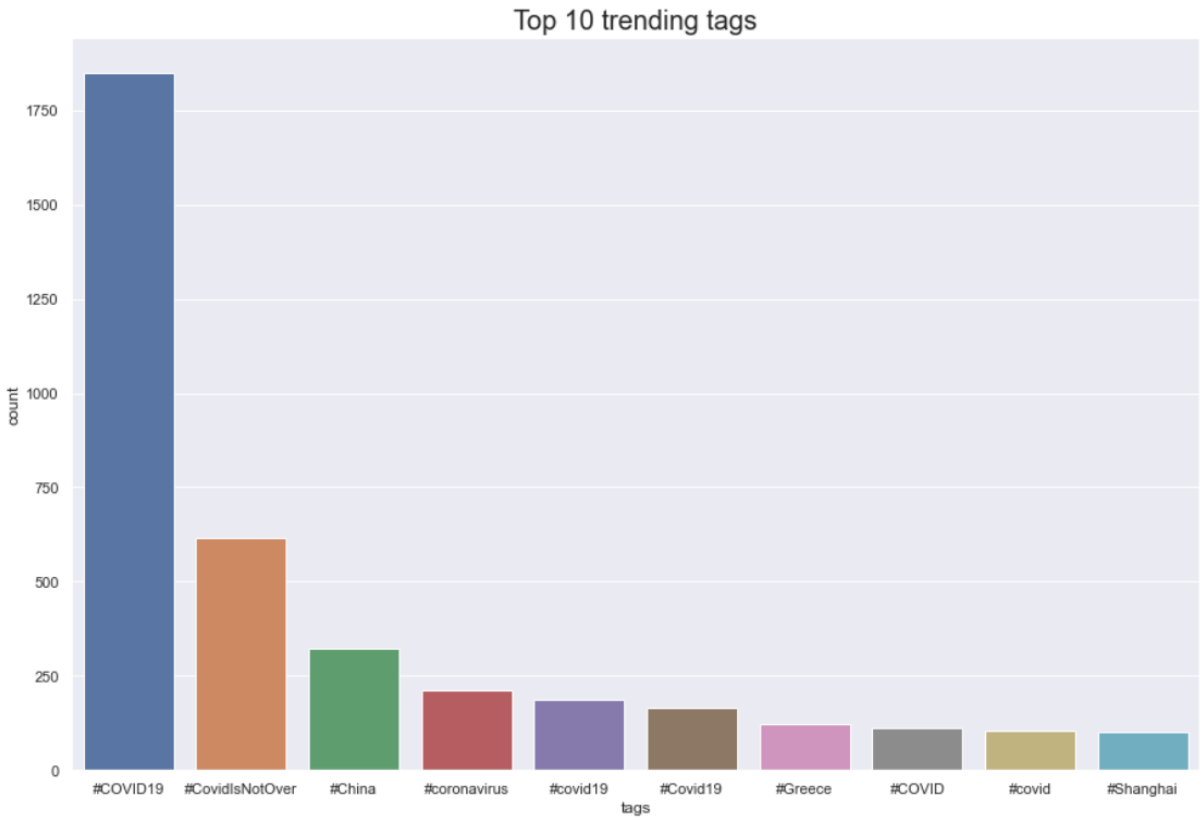
Column	Description
created_at	The date and time when the tweets and the related data were extracted off Twitter with the help of Twitter APIs. For example, DD/MM/YYYY.
id	The integer representation of the unique identifier for the tweet. This number is greater than 53 bits and some programming languages may have difficulty interpreting it. Using a signed 64-bit integer for storing this identifier is safe.
id_str	The string representation of the unique identifier for the tweet.
text	The tweet from the feed of an individual. For example, an untruncated or lengthened tweet might look like: RT @IranNewsUpdate1: #IranProtests #Tehran, #Iran's capital—contract medical staff rally outside the Parliament [Majlis], protesting offici…
truncated	Whether the tweets have been truncated (shortened) or not. For example, TRUE or FALSE.
source	The origin where the tweet was sent or posted from. For example, Twitter for iPhone means the tweet originated from an iPhone.
in_reply_to_status_id	This field is null when the represented tweet is not a reply. If it is a reply then this field will contain the integer representation of the original tweet's ID.
in_reply_to_status_id_str	This field is null when the represented tweet is not a reply. If it is a reply then this field will contain the string representation of the original tweet's ID.
in_reply_to_user_id	Nullable. If the represented tweet is a reply, this field will contain the integer representation of the original tweet's author ID. This will not necessarily always be the user directly mentioned in the tweet. For example: "in_reply_to_user_id":6253282

in_reply_to_user_id_str	Nullable. If the represented tweet is a reply, this field will contain the string representation of the original tweet's author ID. This will not necessarily always be the user directly mentioned in the tweet. For example: "in_reply_to_user_id_str":"6253282"
in_reply_to_screen_name	Nullable. If the represented tweet is a reply, this field will contain the screen name of the original tweet's author. For example: "in_reply_to_screen_name":"twitterapi"
coordinates	This field will be null when present, representing the geographic location of this tweet as reported by the user or client application.
is_quote_status	Indicates whether this is a quoted tweet (a tweet that is a retweet with a comment) or not.
retweet_count	The number of times the tweet was retweeted.
favorite_count	The number of times the tweet was marked as a favourite.
favorited	If the tweet was favourited by the user that made the api call or not.
retweeted	Indicates whether this Tweet has been retweeted by the authenticating user. Example: "retweeted":false
lang	Language, for example, en for English.
possibly_sensitive	Whether the tweet contains content that may be considered sensitive such as violence, pornography, or illegal activities that may be harmful to users or other parties.
quoted_status_id	This field contains the integer value tweet ID of the quoted tweet (when the tweet is a quote tweet.)
quoted_status_id_str	This field contains the string representation tweet ID of the quoted tweet (when the tweet is a quote tweet.)

The focus in this tweets data is in column created_at and text, these are the date of extraction of related tweets and the contain of the tweet for individuals. There are over 4700 tags founds with keywords such as covid or corona in this short sample of twitter data, while keyword vaccine only found 236. Below shows the top ten trending tags over this period.

Twitter data analysis can be extended by looking at sentiment analysis of the content of tweets. This can help the government to understand social sentiments of COVID-19, vaccination programs etc.

Using external source of data can help the creditability of methods used in analysis, however this can be time and monetary costly, as extra time needs to be committed and different tools maybe used.



Conclusion

Based on the two COVID dataset, the percentage of vaccinated with first dose only is surprisingly even across 12 regions, only 4.5% not have the second dose. However, when converted to actual number, in Gibraltar has over 260,000 people not been vaccinated with second dose, followed by Montserrat region (only 20,000 less than Gibraltar). From the recovery timeseries, neither of Gibraltar or Montserrat have the greatest number of recoveries.

On this result, the government can target Gibraltar as the priority to promote second dose vaccinations. However, there are questions raised about these two datasets. Gibraltar for example, total number of people fully vaccinated is over 100 times of its population. There are also instances where cumulative data deaths and recovered number drops below previous records. Another “error” found in the data is that the number of individual tested positive (cumulative data) is less than the number of people hospitalised (daily data). These are all evidences that doubt the accuracy and reliability of the dataset. Hence, the interpreted results are unlikely to be accurate.

Despite not having reliable data, the data is well structured. This means that once another more accurate data has been collected (and follow the same data structure as the datasets used in this project), this analysis can easily be regenerated by following the same procedures and approach used in this project. This is the benefit of using python for analysis and keep coding generic.

Further improvements

While the approach of this project can be reused for similar datasets, more can be added to the analysis.

For example, when visualising recovery data, the number of cases can also be added to the graph to show the “actual” effect of recoveries. A large number of recoveries may not necessarily mean good effect of vaccinations, as a large number of covid cases will generally lead to a higher number of recoveries compared to other regions. Population data can also be used to get an idea of what population of people are not vaccinated.

Twitter data is not looked in extensively in this project, this can also be extended. Sentiment analysis can be used to analyse social sentiments of government’s vaccination program. This may give feedbacks to how vaccinations can be improved.