ASSIGNMENT 2:

DATA ANALYTICS USING PYTHON

TUNA TUNCAY

LSE Data Analytics Career Accelerator
Spring'22

TABLE OF CONTENTS

INTRODUCTION

PROBLEM DEFINITION:

Covid has been the major problem of all economies for the past 2 years. As the data analyst working with UK government, I am expected to identify trends and patterns that will guide through vaccine promotion.

METHADOLOGY

DATA CLEANING & THOUGHT PROCESS:

1. CVS files are imported as pandas dataframe to Jupyter notebook.
2. Column types, column names and shapes are analysed for further actions.
    a. Cases data has 2 missing lines for Bermuda region on dates 21-22.09.2020. Deaths, cases, recovered and hospitalised numbers are not provided.
    b. Hospitalised values at the last 2 dates (13-14/10/2022 are not captured in data. Value is 0 for all state/provinces.
    c. Recovered values are not captured in data after 5/8/2021. Value is 0 for all state/provinces.
    d. Vaccination values at the last 2 dates (13-14/10/2022 are not captured in data. Value is 0 for all state/provinces.
    e. Tweets dataframe's in_reply_to_xxx and quoted_status_xxx columns are all empty.
3. Missing values/incorrect 0 values are excluded in analysis by putting specific filters on analysis. Master data is not amended.
4. I maintained 2 Jupyter notebooks during 6-week analysis. One is used to complete weekly tasks and the second one used for further analysis/trial and error. I decided to share my further analysis copy as final work.

KEY ASSUMPTIONS:

- Deaths: Values are cumulative sum. Daily values are calculated separately as needed and labelled as daily death.
- Cases: Values are cumulative sum. Daily values are calculated separately as needed and labelled as daily cases.
- Recovered: Values are cumulative sum. Daily values are calculated separately as needed and labelled as daily recovered.
- Hospitalised: Values show number of people at the hospital on specific day. New hospitalised numbers are calculated separately showing net add/loss of people in hospital when needed and labelled as net hospitalised.
- Vaccinated: Values show number of people vaccinated on specific day. Cumulative vaccinated numbers are calculated separately and labelled as fully vaccinated.
- Fully vaccinated rate: Calculated as first dose owners over second dose owners. Treated as they are the same people.

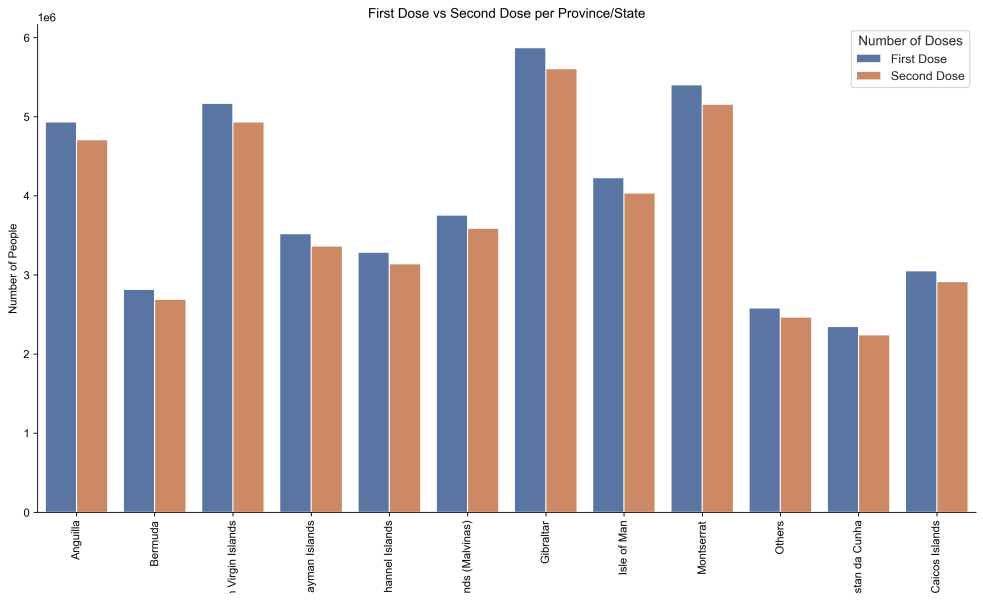
KEY FINDINGS

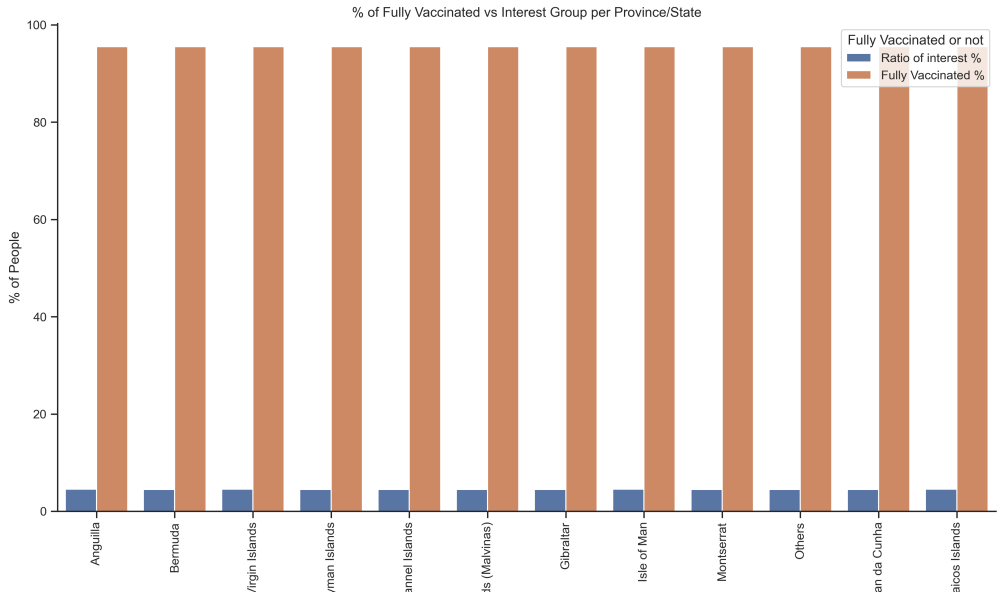
Vaccination rate is calculated as first dose takers over second dose takers. The percentage of fully vaccinated people is almost the same for all regions. Gibraltar has the biggest number of nominal vaccinated value.

This makes the work of the government hard to make decision. Further analysis with regards to population would be valuable. Since the vaccination rates by itself is not enough by itself to make any decision, I further evaluated with death, case, and hospitalisation numbers.

Number of People Vaccinated in Province/State breakdown

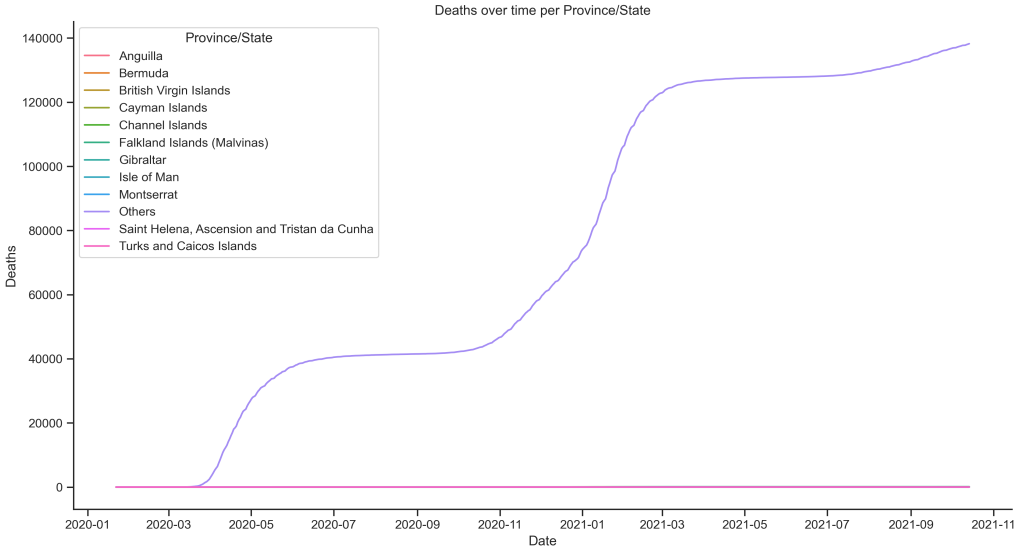
Province/State	Vaccinated	First Dose	Second Dose	Delta	not_fully_vaccinated %
Turks and Caicos Islands	2915136	3052822	2915136	137686	4.510122
Isle of Man	4036345	4226984	4036345	190639	4.510048
Anguilla	4709072	4931470	4709072	222398	4.509771
British Virgin Islands	4933315	5166303	4933315	232988	4.509763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

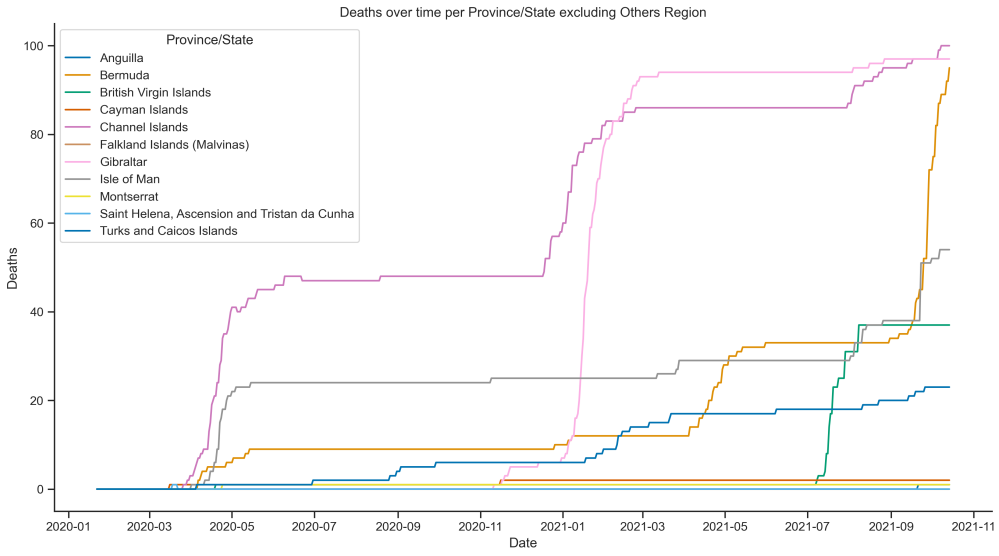




Death numbers in the data are cumulative sums. Therefore, it is expected that numbers are increasing over time. However, the pace and trend are important here. Deaths in other region had most of the values therefore analysed data with and without others region.

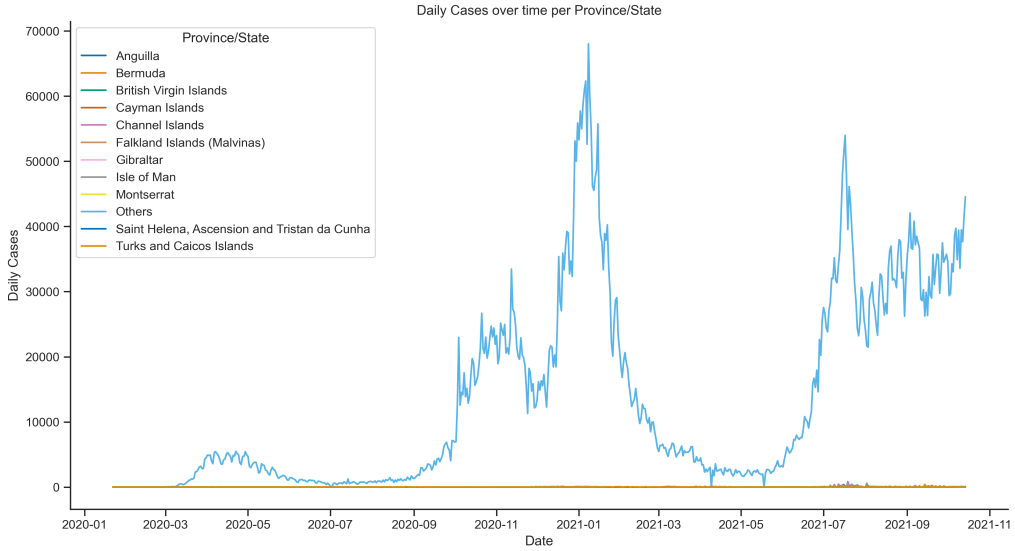
- There are 2 main death cycles for almost all provinces.
- Anguilla, Isla of Man, Bermuda started having a sharp increase in death numbers at the beginning of the pandemic (03/2020) whereas Turk & Caicos, Gibraltar kept themselves almost safe till Q42020. British Virgin Island remained safe from death almost till mid 2021.
- Gibraltar province witnessed deaths later. However, it is the worst hit province after others region.

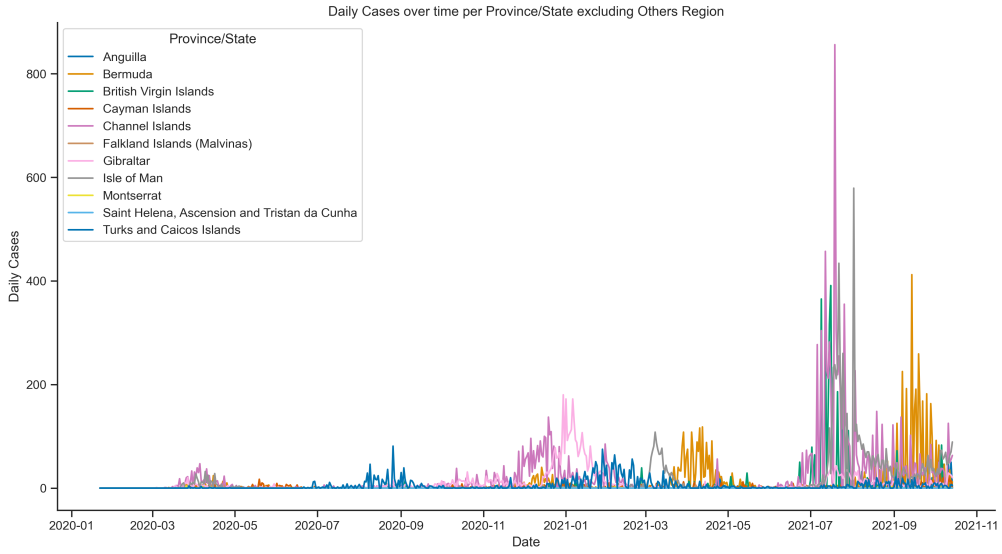




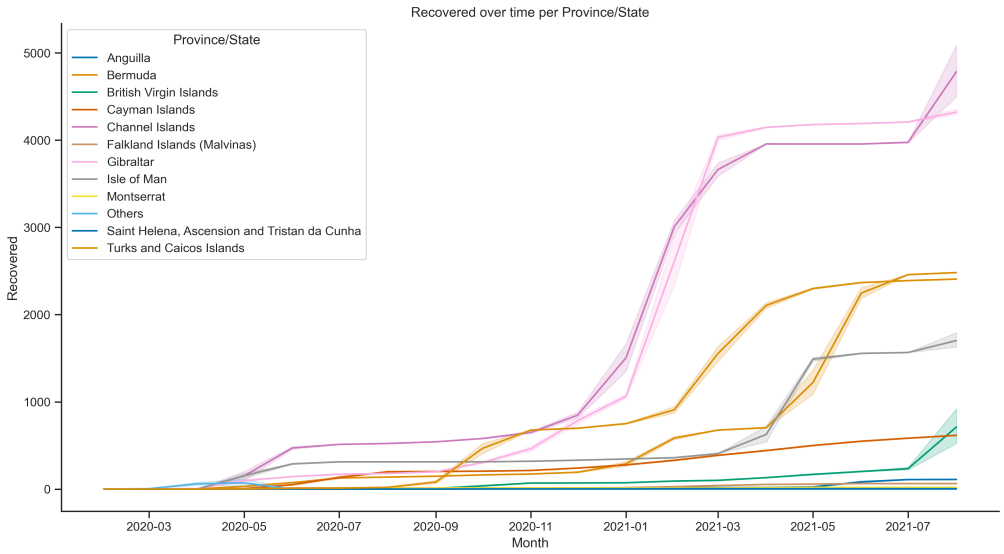
It is easier to check case numbers over time on daily values. Different from death numbers, I can spot clear 3 peaks here. It is important to note that even the daily case numbers are bigger in each peak compared to previous one, I don't see sharp increase in death numbers in the last peak. This should be the positive affect of vaccination. Plot showing all regions except others have too many data points. I would recommend providing further analysis per region when needed. The shape is similar in all regions

(with 3 peaks however timing is slightly delayed in some regions like Bermuda, which is also the case in death numbers.

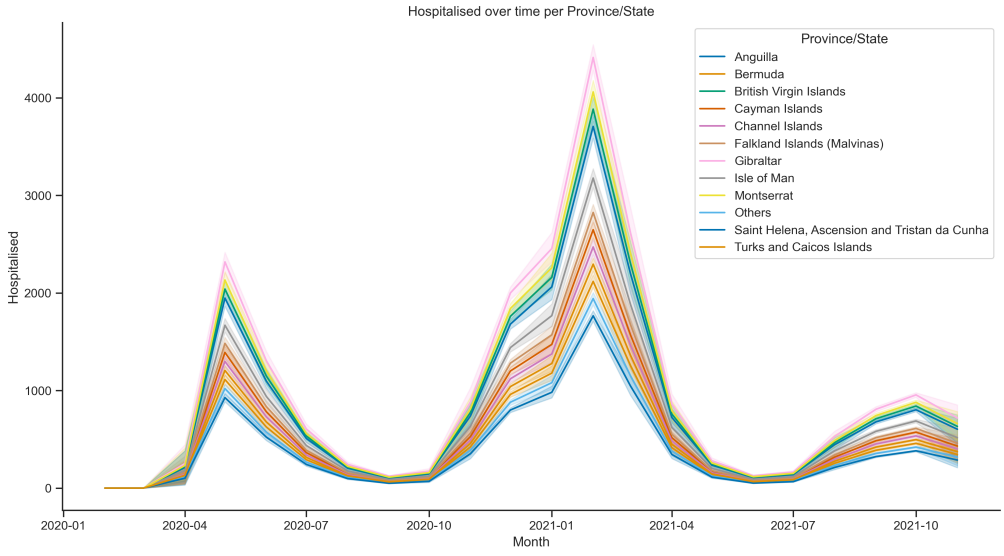


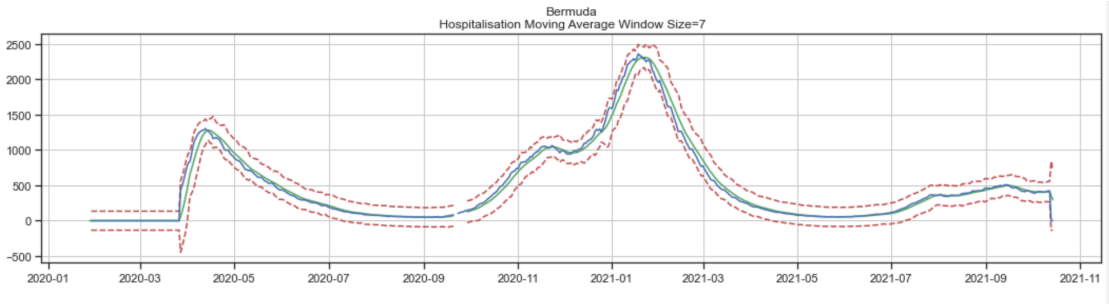


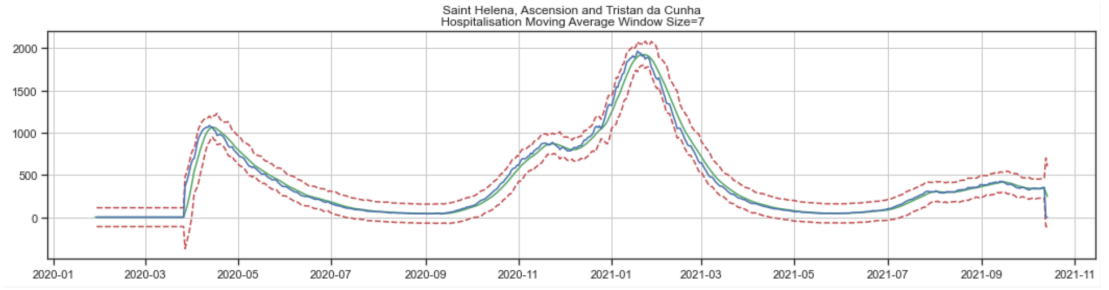
Hospitalised and recovered numbers are pretty much following the same pattern for all state/provinces. Although the case numbers are increasing and having additional peaks, by the effect of vaccinated people are recovering.

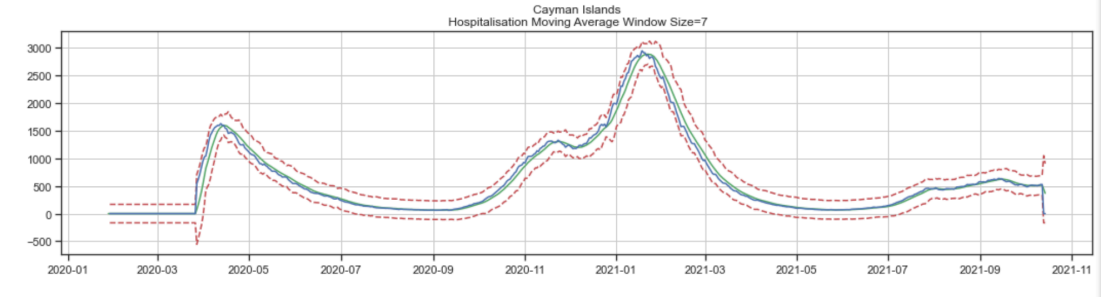


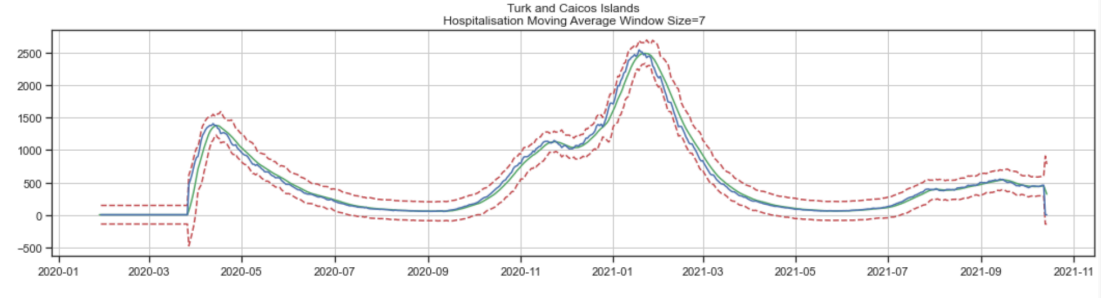
Hospitalised numbers are perfectly aligned with each other for each. All regions are reaching their peak the same time which makes me sceptical about the hospitalised numbers. I continue my analysis on hospitalised data using moving averaging forecast model with a window of 7 days. All regions seem to have reached their peak already.

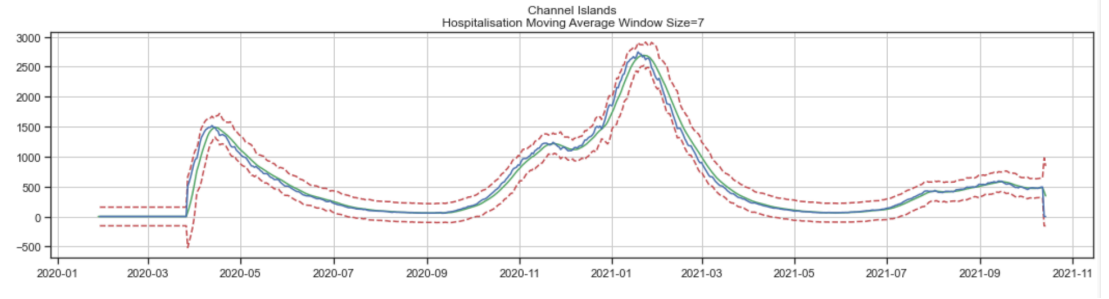


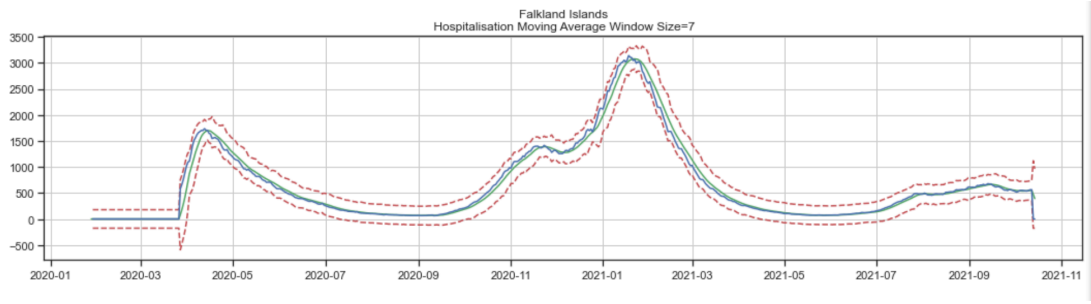

Bermuda
Hospitalisation Moving Average Window Size=7
2500
2000
1500
1000
500
0
−500
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11

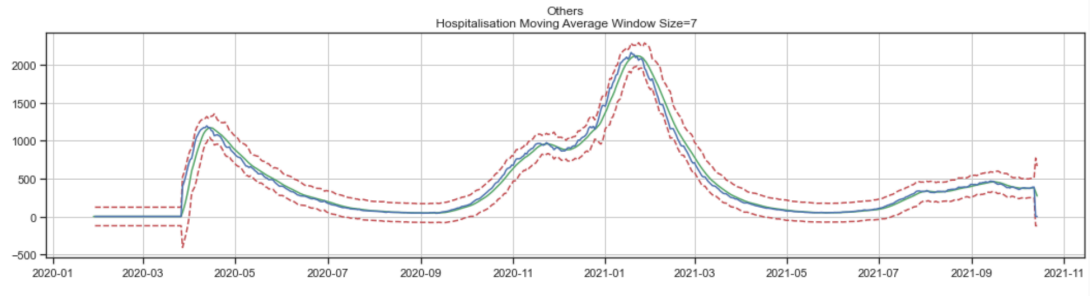

Saint Helena, Ascension and Tristan da Cunha
Hospitalisation Moving Average Window Size=7
2000
1500
1000
500
0
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11

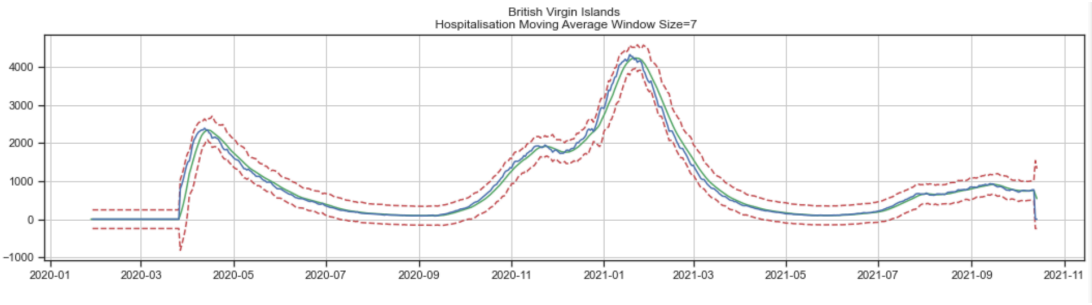

Cayman Islands
Hospitalisation Moving Average Window Size=7
3000
2500
2000
1500
1000
500
0
−500
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11

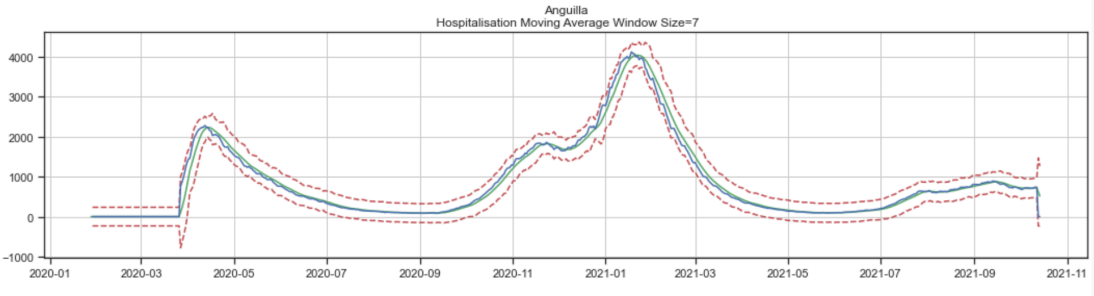

Turk and Caicos Islands
Hospitalisation Moving Average Window Size=7
2500
2000
1500
1000
500
0
−500
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11

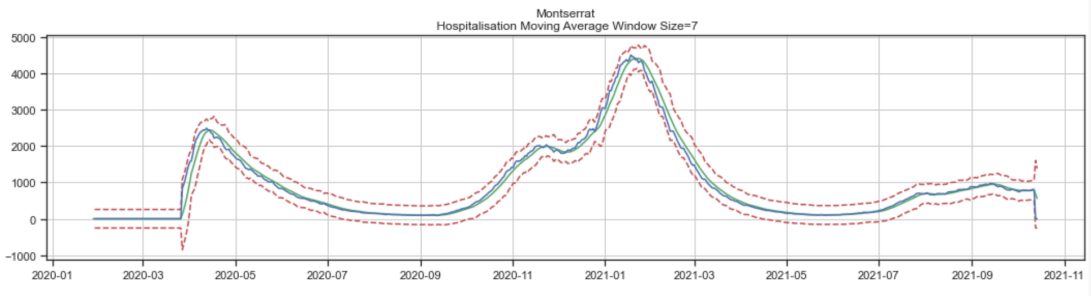

Channel Islands
Hospitalisation Moving Average Window Size=7
3000
2500
2000
1500
1000
500
0
−500
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11

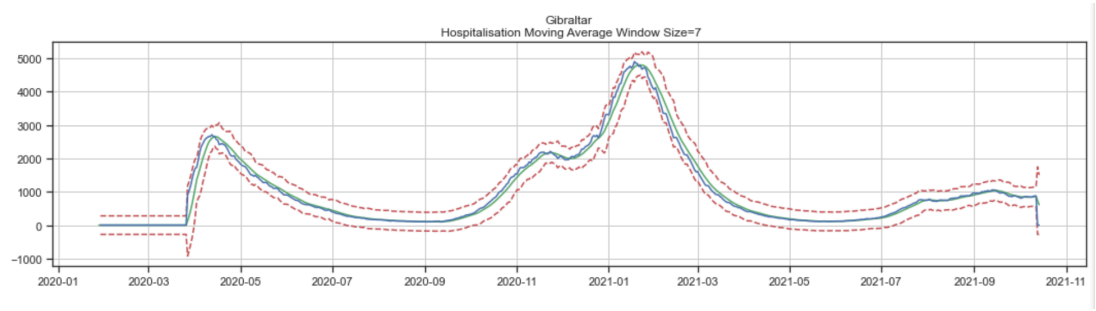

Falkland Islands
Hospitalisation Moving Average Window Size=7
3500
3000
2500
2000
1500
1000
500
0
−500
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11


Others
Hospitalisation Moving Average Window Size=7
2000
1500
1000
500
0
−500
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11


British Virgin Islands
Hospitalisation Moving Average Window Size=7
4000
3000
2000
1000
0
−1000
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11

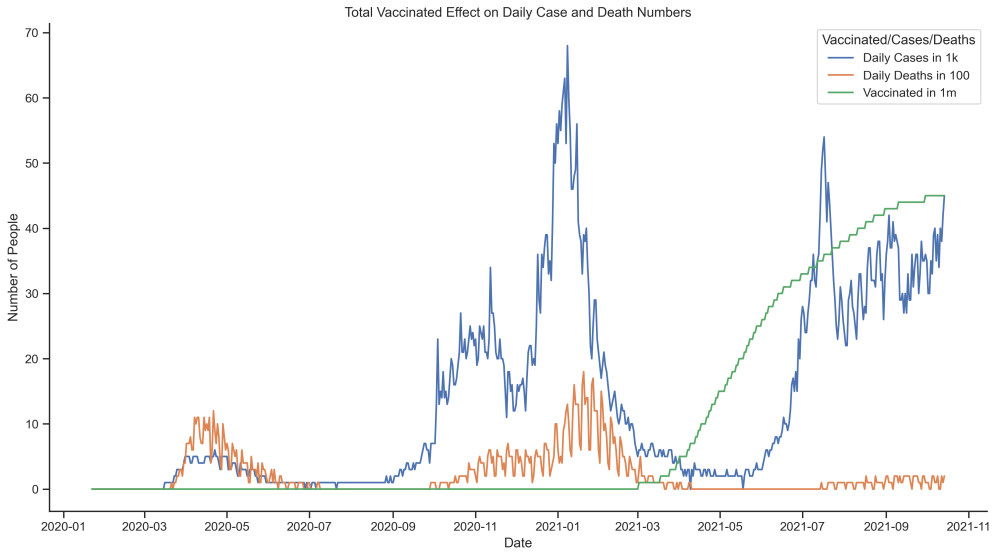

Anguilla
Hospitalisation Moving Average Window Size=7
4000
3000
2000
1000
0
−1000
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11


Montserrat
Hospitalisation Moving Average Window Size=7
5000
4000
3000
2000
1000
0
−1000
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11



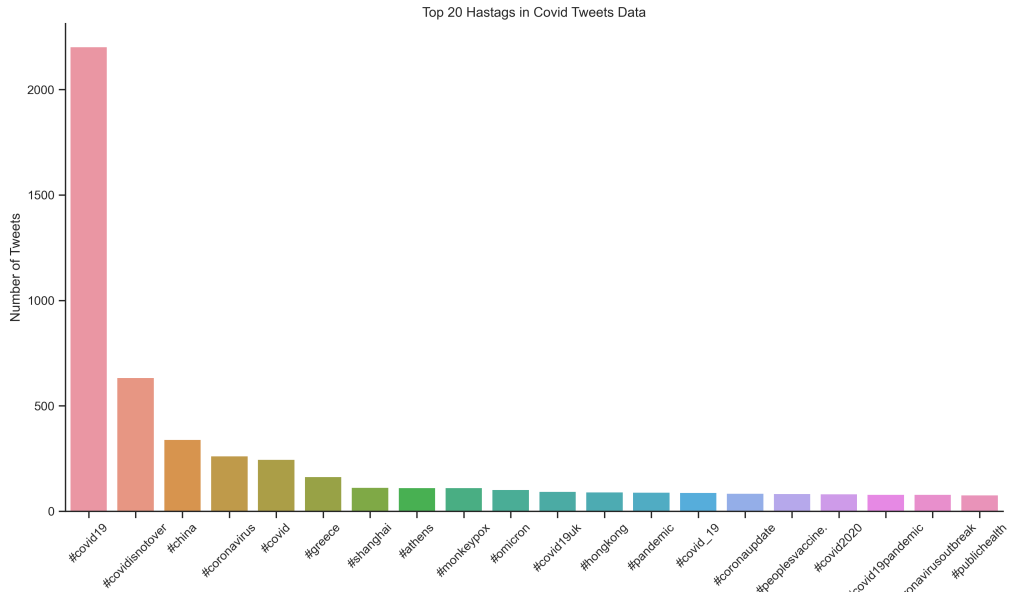
Looking at the case, death, and vaccinated numbers separately in charts helped to identify trends and patterns on Province/State level. I further analysed their relationship with each other to have a better understanding. Below graph shows daily deaths in 100 people, daily cases in 1000 people and vaccinated people in 1 million people.

It is no brainer that cases and deaths are positively correlated with each other before vaccines. However, vaccine clearly changed the picture. After vaccine rollout, even the case numbers surge, death numbers are not increasing dramatically. In the below graph you can see the effect of vaccines clearly.

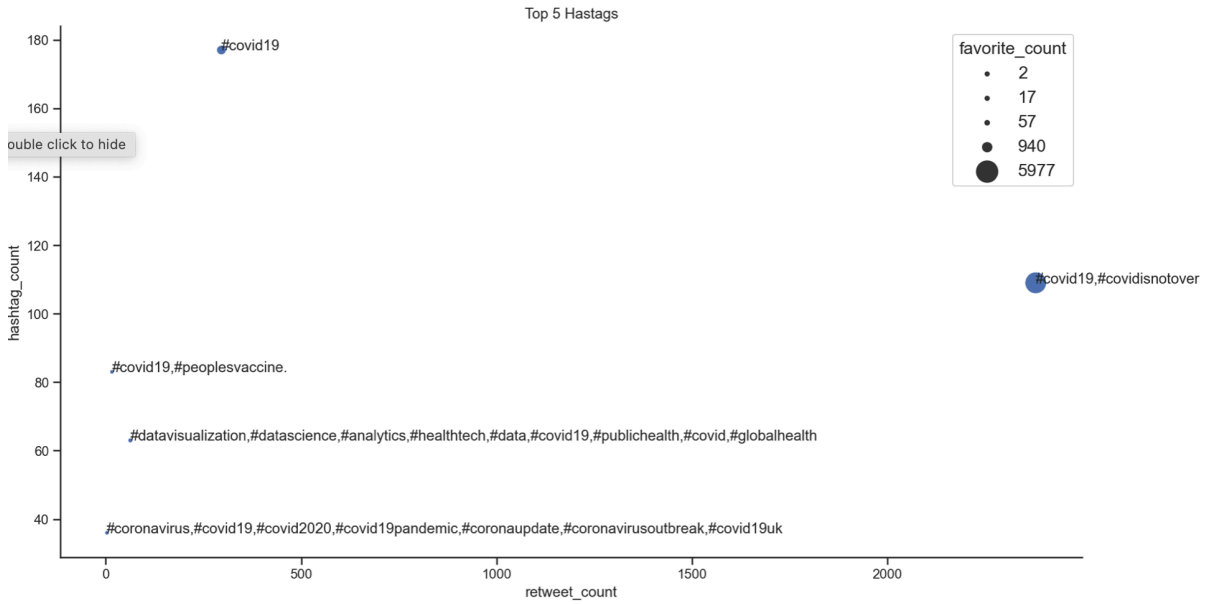


It is important to note that while case numbers continue to rise, vaccination numbers are slowing down. Since we don't have the population data, it is not clear that percentage of vaccinated population in the regions. The effect of vaccine over deaths is clear, therefore there should be something else causing the slow down. I will try to explain this with twitter data.

From the tweets data, I mainly focus on hashtags. Below you may find the top 20 hashtags discussed in the given dataset. I lowercase all hashtags in my list so that I can group same hashtags that has the same letters but expressed in lower/upper case.



Top 20 hashtags are almost all related to Covid name. The vaccine hashtag is only captured once in top 20. I realised some tweets include more than 1 hashtags and usually people express themselves with several hashtags. Below is the scatter plot that shows the relationship between hashtag count, retweet count and favourites. You can see that now #peoplesvaccine is within the top 5. People use #peoplesvaccine hashtag with #covid19 hashtag. Hashtag count for #peoplesvaccine is the third one but it is not retweeted and favourited by others.



CONCLUSION

The timeline for Covid 19 pandemic has been unique for different regions in UK. However, pattern and trends are pretty much similar for all. Some had the biggest hit at the beginning of pandemic, some had a chance to delay a bit but no matter what all had been badly affected before vaccine rollout.

Vaccines changed the whole picture. Although case numbers continue to rise with new variants and testing capabilities (rapid tests etc) death curve is flattened.

All provinces have similar vaccinations rates when we make the calculation over all first dose takers. I would like to analyse further with the population data to see how many percentages of the population is vaccinated and is there any differences on that. That would be helpful for me to suggest UK government where to focus on their marketing campaign further.

Data from twitter has been very helpful to understand why vaccination numbers are lately flattened. Although the timing of tweets data (15-23/05/2022) and vaccine data set (Jan'2020-Oct'2021) is not same it gives me an initial idea to analyse further. I would love to check twitter hashtags and trending topics historically for similar dates if I have an excess to that data.