LSE_DA201_ASSIGNMENT_REPORT_MADURAA_KANNAN

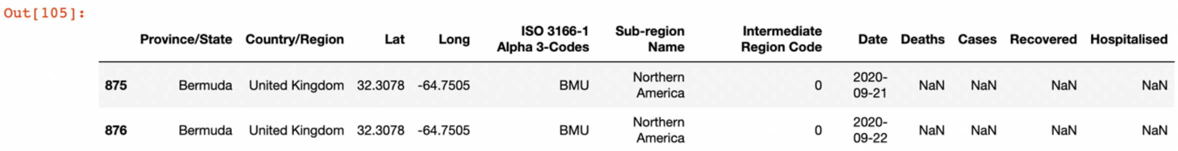
Term 2 Assignment

Maduraa Kannan

The UK government is aiming to increase the number of fully vaccinated individuals by launching marketing campaigns to promote the vaccine. They want to do this through analysing the provided dataset and identifying trends and patterns. The analysis will focus on:

- Vaccinations per regions
- The area with the greatest number of recoveries
- Hospitalisation data
- If deaths have been increasing across all regions overtime, or if a peak has been reached.
- Hashtags used on twitter.

I initially explored the dataset by importing the covid_19_uk_cases.csv and the covid_uk_vaccinated.csv files into python and loaded them into DataFrames on python. I then analysed the datatypes of the columns, the shape and size of the data sets, as well as any possible missing data. There were 8 missing data in covid_19_uk_cases.csv. The rows containing the missing data are shown below.

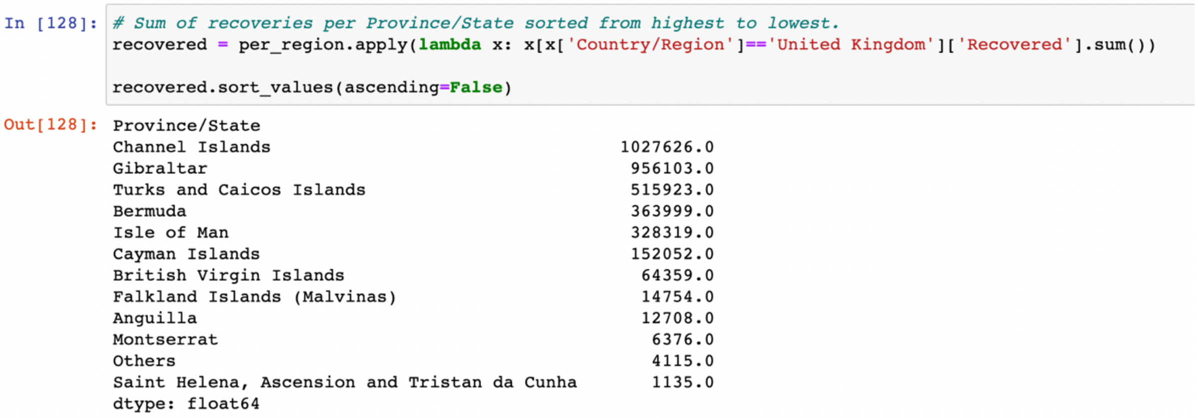


Out[105]:

	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

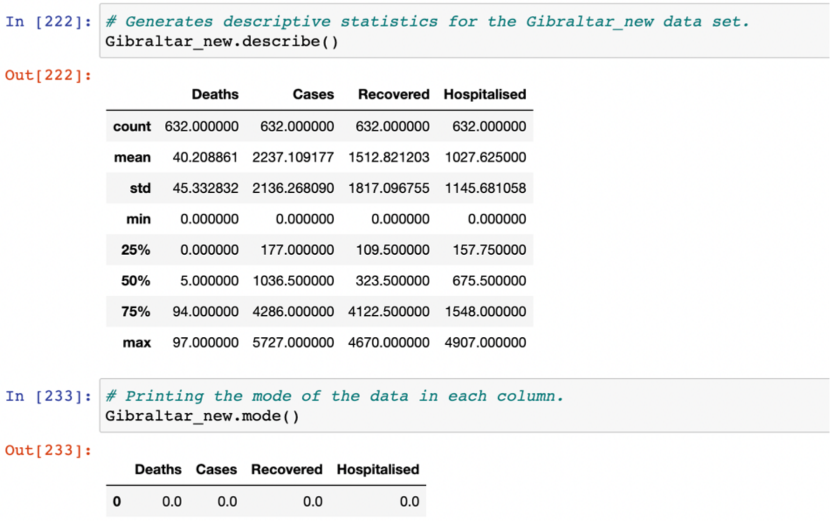
To simplify the analysis process, I merged the covid cases and vaccinated dataset, removed any unnecessary columns (‘Lat’, ‘Long’, ‘ISO 3166-1 Alpha 3-Codes’, ‘Sub-region Name’, ‘Intermediate Region Code’), and changed any incorrect datatype (‘Date’).

I wanted an overview of the data provided to give me insight into the regions I should investigate further. I used the lambda function to split the data up by ‘Province/State’ and for each of these regions found the sum of cases, deaths, recoveries, first doses, second doses, and total number of people with first doses but not their second doses.



```
In [128]: # Sum of recoveries per Province/State sorted from highest to lowest.
          recovered = per_region.apply(lambda x: x[x['Country/Region']=='United Kingdom']['Recovered'].sum())

          recovered.sort_values(ascending=False)
```

```
Out[128]: Province/State
          Channel Islands                                 1027626.0
          Gibraltar                                        956103.0
          Turks and Caicos Islands                         515923.0
          Bermuda                                          363999.0
          Isle of Man                                      328319.0
          Cayman Islands                                   152052.0
          British Virgin Islands                            64359.0
          Falkland Islands (Malvinas)                       14754.0
          Anguilla                                          12708.0
          Montserrat                                         6376.0
          Others                                             4115.0
          Saint Helena, Ascension and Tristan da Cunha       1135.0
          dtype: float64
```



```
In [222]: # Generates descriptive statistics for the Gibraltar_new data set.
          Gibraltar_new.describe()
```

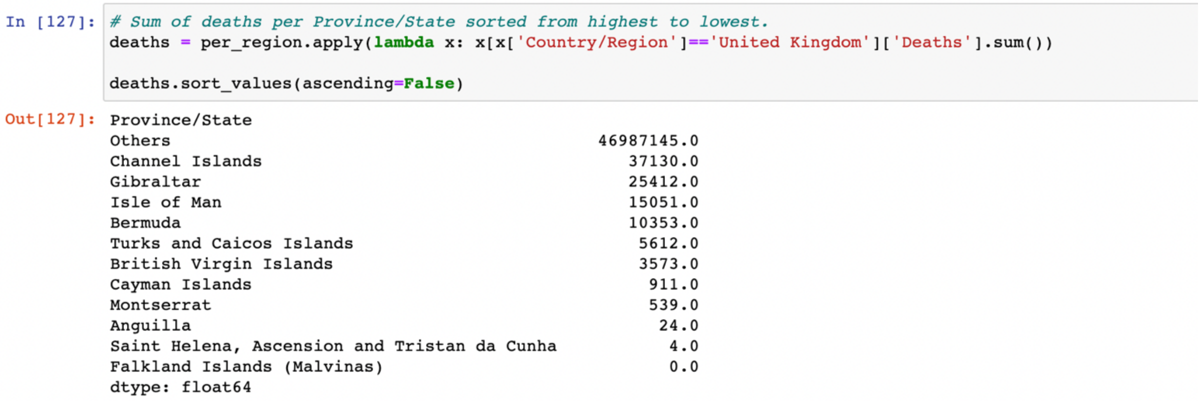
Out[222]:

	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 [233]: # Printing the mode of the data in each column.
          Gibraltar_new.mode()
```

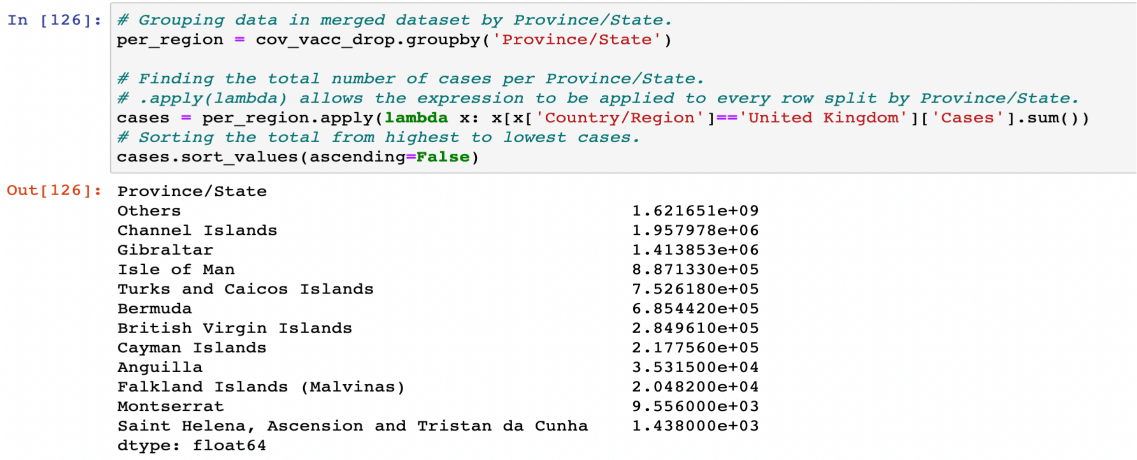
Out[233]:

	Deaths	Cases	Recovered	Hospitalised
0	0.0	0.0	0.0	0.0

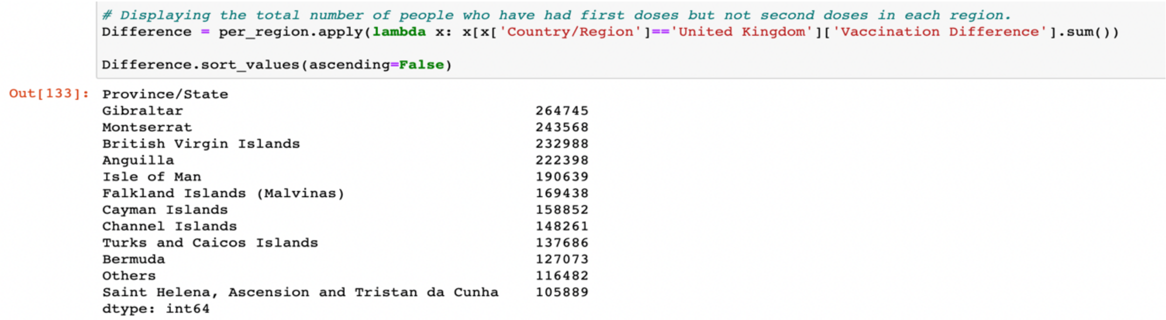


```
In [127]: # Sum of deaths per Province/State sorted from highest to lowest.
          deaths = per_region.apply(lambda x: x[x['Country/Region']=='United Kingdom']['Deaths'].sum())

          deaths.sort_values(ascending=False)
```

```
Out[127]: Province/State
          Others                                          46987145.0
          Channel Islands                                    37130.0
          Gibraltar                                          25412.0
          Isle of Man                                        15051.0
          Bermuda                                            10353.0
          Turks and Caicos Islands                            5612.0
          British Virgin Islands                              3573.0
          Cayman Islands                                       911.0
          Montserrat                                           539.0
          Anguilla                                              24.0
          Saint Helena, Ascension and Tristan da Cunha           4.0
          Falkland Islands (Malvinas)                            0.0
          dtype: float64
```

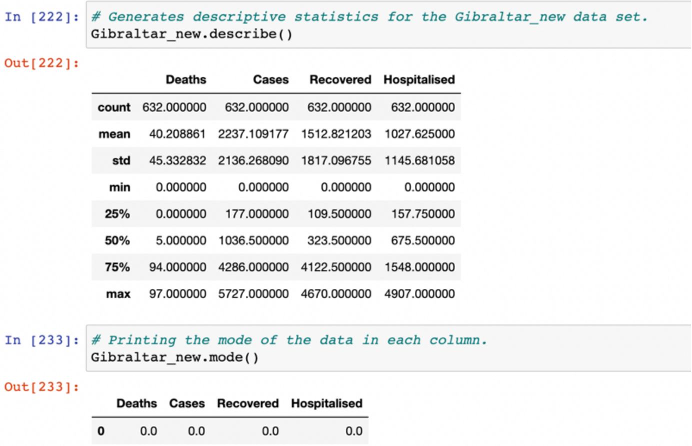


```
In [126]: # Grouping data in merged dataset by Province/State.
          per_region = cov_vacc_drop.groupby('Province/State')

          # Finding the total number of cases per Province/State.
          # .apply(lambda) allows the expression to be applied to every row split by Province/State.
          cases = per_region.apply(lambda x: x[x['Country/Region']=='United Kingdom']['Cases'].sum())
          # Sorting the total from highest to lowest cases.
          cases.sort_values(ascending=False)

Out[126]: Province/State
          Others                                          1.621651e+09
          Channel Islands                                 1.957978e+06
          Gibraltar                                       1.413853e+06
          Isle of Man                                     8.871330e+05
          Turks and Caicos Islands                        7.526180e+05
          Bermuda                                         6.854420e+05
          British Virgin Islands                          2.849610e+05
          Cayman Islands                                  2.177560e+05
          Anguilla                                        3.531500e+04
          Falkland Islands (Malvinas)                     2.048200e+04
          Montserrat                                      9.556000e+03
          Saint Helena, Ascension and Tristan da Cunha    1.438000e+03
          dtype: float64
```



```
           # Displaying the total number of people who have had first doses but not second doses in each region.
           Difference = per_region.apply(lambda x: x[x['Country/Region']=='United Kingdom']['Vaccination Difference'].sum())

           Difference.sort_values(ascending=False)

Out[133]:  Province/State
           Gibraltar                                       264745
           Montserrat                                      243568
           British Virgin Islands                          232988
           Anguilla                                        222398
           Isle of Man                                     190639
           Falkland Islands (Malvinas)                     169438
           Cayman Islands                                  158852
           Channel Islands                                 148261
           Turks and Caicos Islands                        137686
           Bermuda                                         127073
           Others                                          116482
           Saint Helena, Ascension and Tristan da Cunha    105889
           dtype: int64
```

This initial summary caused me to explore the data for Gibraltar. This is because the initial analysis showed it had the second highest number of deaths and had the highest number of people who received their first dose, but not their second dose. This large difference between the individuals in Gibraltar who are fully vaccinated versus those who have only had one dose of the vaccine means that the people in Gibraltar are more likely to engage with the campaign and spread the campaign further within the region.



```
In [222]: # Generates descriptive statistics for the Gibraltar_new data set.
          Gibraltar_new.describe()
```

Out[222]:

	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

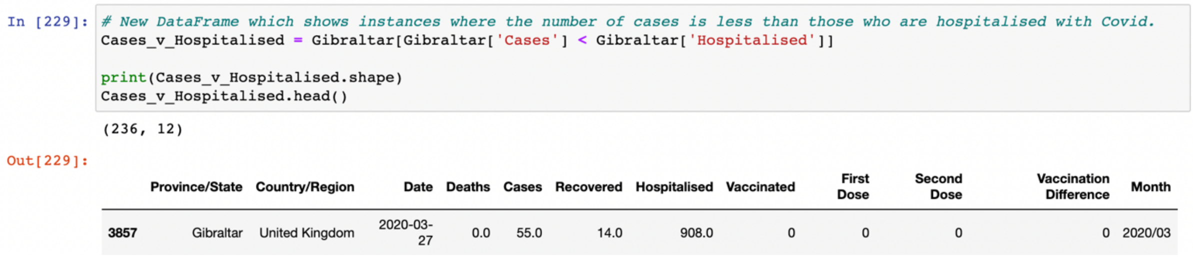
```
In [233]: # Printing the mode of the data in each column.
          Gibraltar_new.mode()
```

Out[233]:

	Deaths	Cases	Recovered	Hospitalised
0	0.0	0.0	0.0	0.0

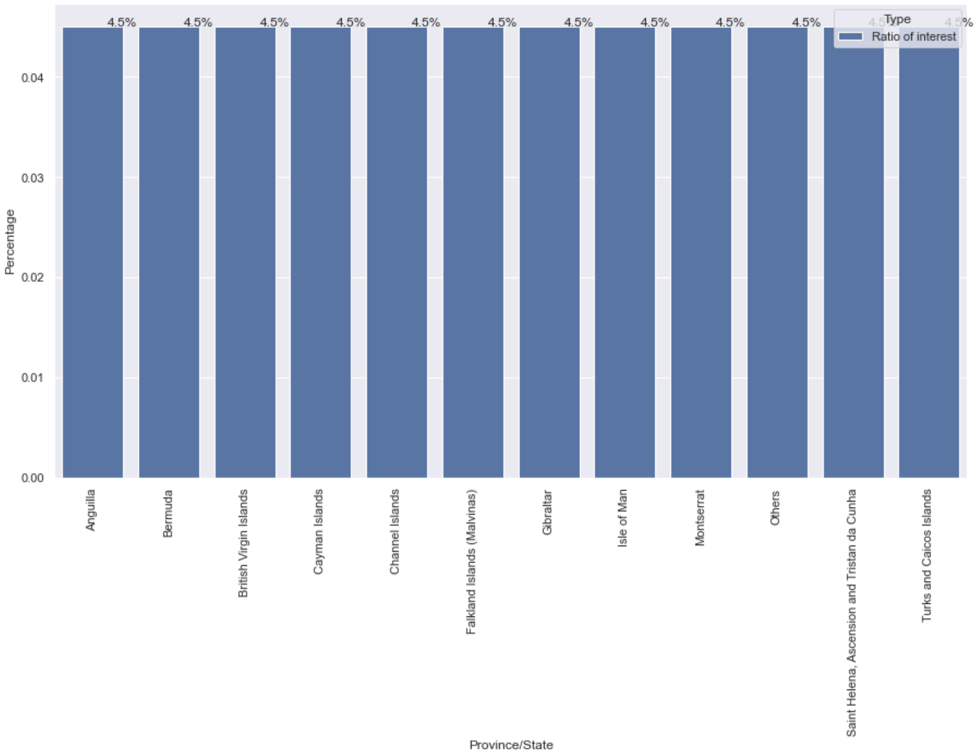
The data from Gibraltar (as seen above) showed that deaths, cases, recoveries, and hospitalisation were positively skewed (Mode < Median < Mean). This means that majority of the data values lie below the mean, with there being more of a spread in the data for the higher values in the data. This could mean that more of the outliers exist above the upper limit compared with the lower limit. I performed a mini analysis on the outliers in deaths for all provinces/regions. If I had more time, I would have performed this analysis to validate the statement.

However, the data for Gibraltar also seemed to be incorrect in some areas. The data consisted of 236/632 rows which indicated that on certain days the number of Covid cases were less than the number of people who were hospitalised with Covid. This could be a result of reporting errors which were not uncommon, especially during the beginning of the pandemic where there weren't enough Covid tests available, and where in some instances thousands of covid cases were missed when tallying the daily figures (https://www.bbc.co.uk/news/uk-54412581).



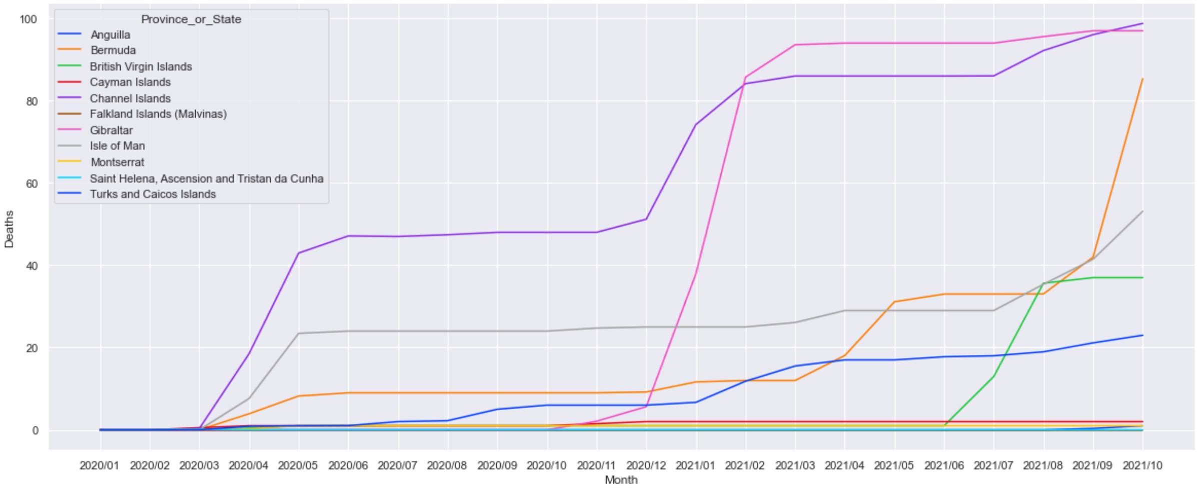
```
In [229]: # New DataFrame which shows instances where the number of cases is less than those who are hospitalised with Covid.
          Cases_v_Hospitalised = Gibraltar[Gibraltar['Cases'] < Gibraltar['Hospitalised']]

          print(Cases_v_Hospitalised.shape)
          Cases_v_Hospitalised.head()

          (236, 12)

Out[229]:
```

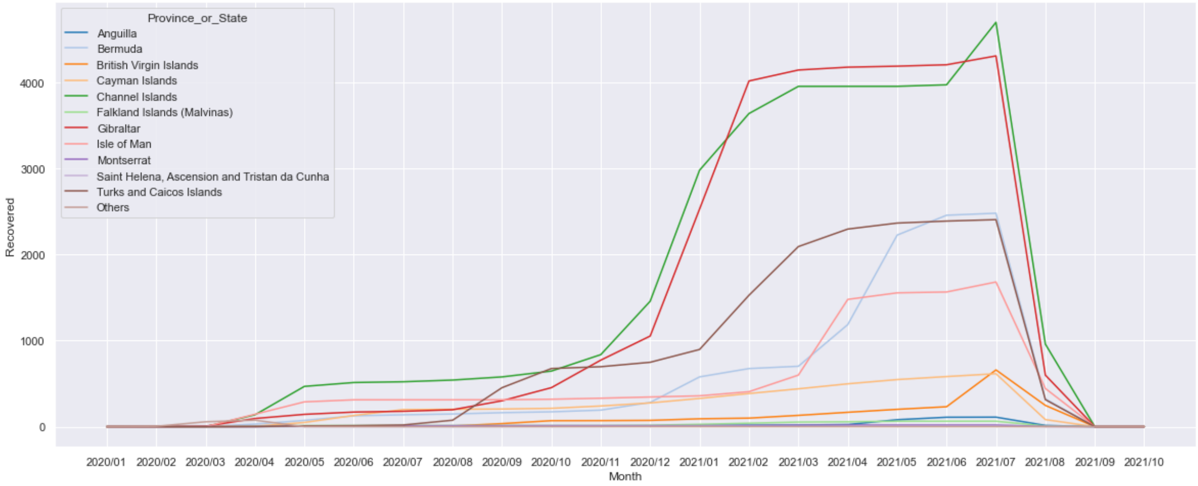
	Province/State	Country/Region	Date	Deaths	Cases	Recovered	Hospitalised	Vaccinated	First Dose	Second Dose	Vaccination Difference	Month
3857	Gibraltar	United Kingdom	2020-03-27	0.0	55.0	14.0	908.0	0	0	0	0	2020/03



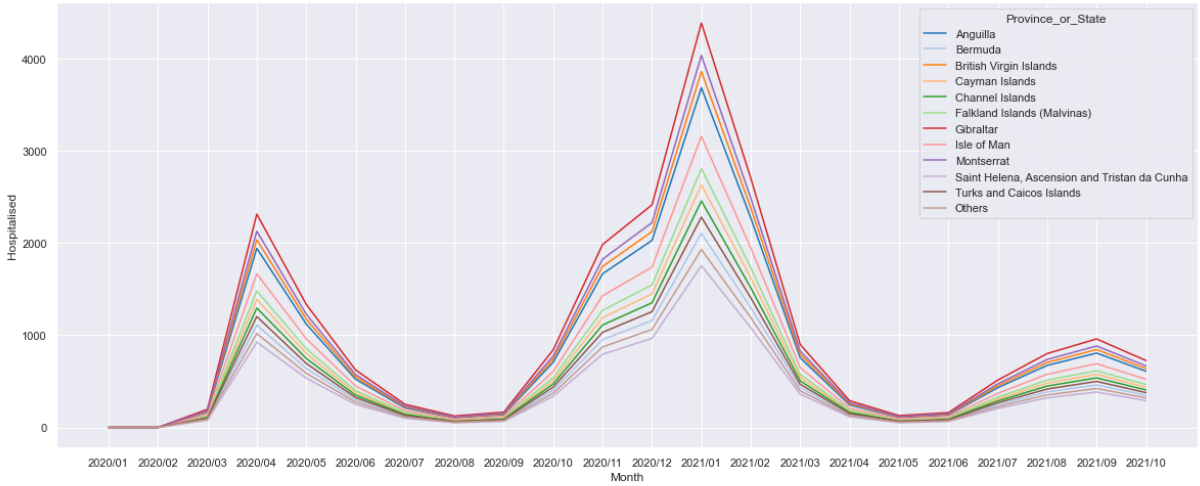
This bar chart shows the ratio of interest per region. It illustrates that every region has the same proportion of individuals eligible for the second dose. Could this simply be due to these individuals not yet being eligible for their second dose, or reluctance due to anti-vaccine propaganda and issues caused from controversies surrounding the AstraZeneca vaccine?

The graph below shows a line plot used to display the trend of deaths across all regions over time. Gibraltar has the highest rate of increase from 2020/12 to 2021/02, and although its death rate is increasing in 2021/10, the gradient is much smaller, and thus less steep. This contrasts with Channel Islands and Montserrat, where the number of deaths is still rising. Montserrat has a much steeper gradient than Channel Islands, and has risen a considerable amount in one month, whereas Channel Islands death rate has increased steadily over four months. Montserrat has a higher number of people who have only had their first dose and has a higher number of individuals who have been vaccinated compared to Channel Islands. It is possible that in Montserrat people's confidence in the vaccine, regardless of not being fully vaccinated, resulted in more socialising without precaution, thus leading to a rapid increase in the number of deaths.

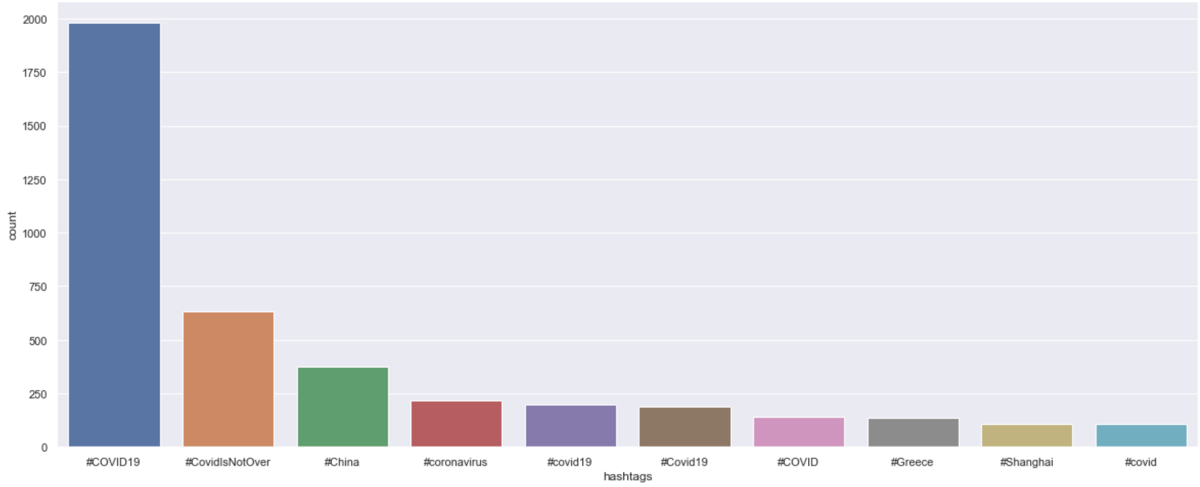




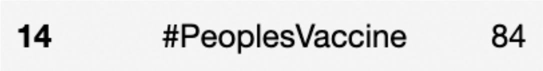
The line plot above shows the recovered cases across the months. This line plot looks like the line plot for the Deaths dataset with Channel Islands having the highest peak in recoveries; however, the peaks differ by 3 months. For Gibraltar and Channel Islands, the recovered cases hit their peak in 2021/07, whereas the death cases hit their peak in 2021/10. One potential cause for this lag in deaths could be a result of the lag in deaths caused by the delay in time it took for a patient to become critically ill (https://covidtracking.com/analysis-updates/how-lagging-death-counts-muddied-our-view-of-the-pandemic]).



The line plot above shows the number of hospitalised cases across the month. There are three peaks across all the regions. Highest peak occurs in 2021/01, followed by a peak in 2020/04, and then 2021/10. Gibraltar has consistently had the highest number of hospitalisations every month. Additionally, this data seems to contradict the deaths over time. How can hospitals have their lowest peak in 2021/10, but regions have their highest deaths in 2021/10?



This bar chart shows the top trending hashtags on twitter. #COVID19 is the most popular hashtag. Thus, marketing campaigns would spread further if this hashtag was incorporated.



14	#PeoplesVaccine	84

#PeoplesVaccine is the largest hashtag relating to vaccines, and thus can also be employed in the campaign.

To conclude, the data suggests that Gibraltar might be a good place for the government to start its marketing approach to increase the number of fully vaccinated people. This is due to Gibraltar having the highest number of fully vaccinated individuals, as well as the highest number of individuals who have only had their first dose. Additionally, the high death rate also suggests that the people of Gibraltar are more likely to respond well to the campaigns.

Another interesting region to consider is Isle of Man as it comes third for the highest number of deaths and comes fifth for the sum of people who recovered from Covid, and for the total number of people who had their first dose, but not their second dose. Due to the high death rate, and its relatively low number of recoveries, marketing campaigns may also be successful here.

Moreover, the twitter analysis suggests Covid marketing campaigns should contain the hashtag, #COVID19, as this is the most popular COVID related hashtag.