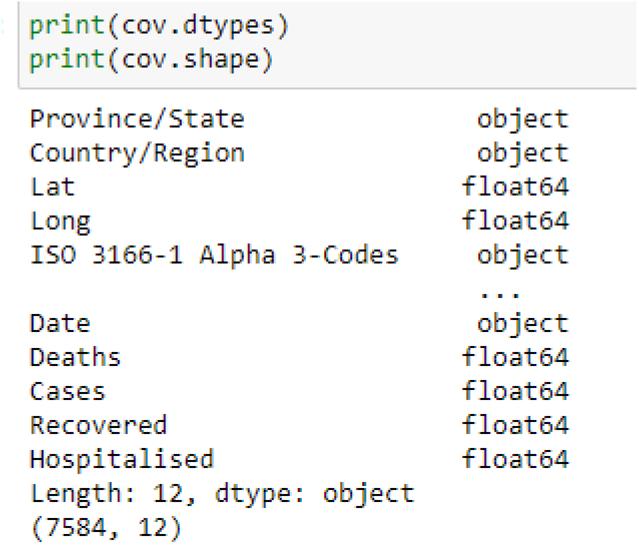
Data analytics using Python - UK Covid/vaccine data

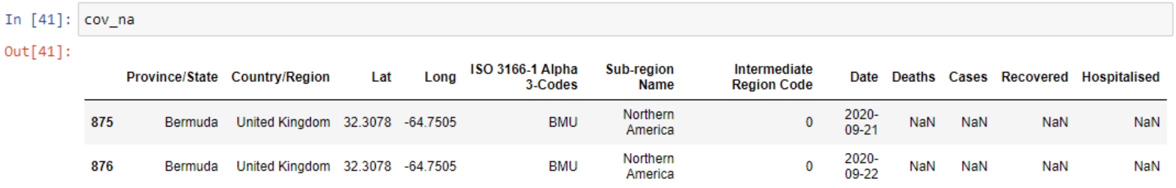
The 2 main datasets provided, contain information for 12 States/Provinces. There are only 2 rows with missing data overall.



```
print(cov.dtypes)
print(cov.shape)
```

```
Province/State              object
Country/Region              object
Lat                        float64
Long                       float64
ISO 3166-1 Alpha 3-Codes    object
                             ...
Date                        object
Deaths                     float64
Cases                      float64
Recovered                  float64
Hospitalised               float64
Length: 12, dtype: object
(7584, 12)
```

Figure 1: Determining the Data types in the DataFrame and number of rows and columns



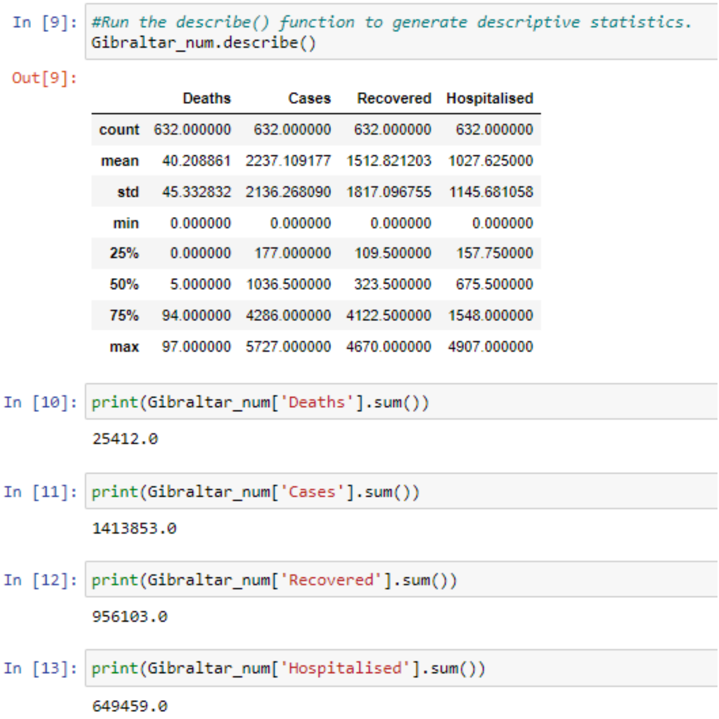
```
In [41]: cov_na
Out[41]:
```

	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

Figure 2: Determining the number of missing values.

The covid cases increases from the beginning (22/01/2020) all the way to the end (14/10/2021), however, it fluctuates and so it would be nice to be able to visualise this data with a line chart to see the changes over time. The recovered data stops from August 2021; is this an error? On their own, the descriptive statistics will not be very meaningful; what would help is to compare the results with other State's in the DataFrames and see how they compare.

The numerical data for Gibraltar seem to suggest the numbers are inflated. For example, the 'Vaccinated' sum is 5,606,041; Gibraltar's population in 2020 was only 33,691, so how can the total vaccinated exceed this number? The sum for 'Cases', 'Recovered' and 'Hospitalised' also exceed the population of Gibraltar. Where is the data from and is it reliable?



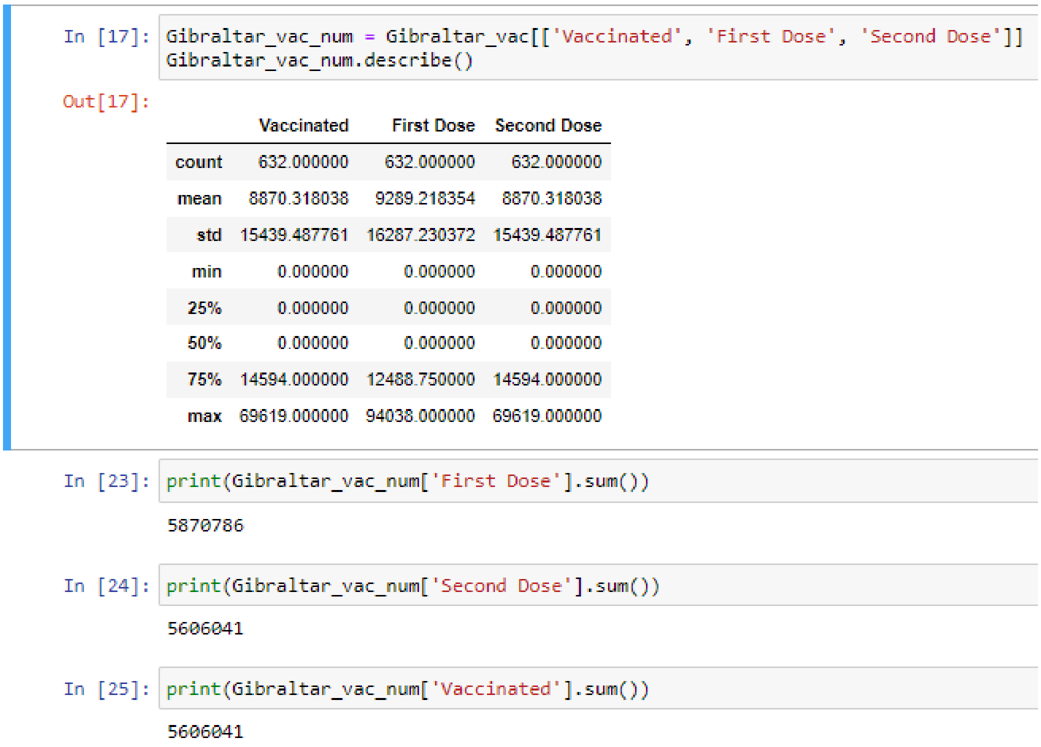
```
In [9]: #Run the describe() function to generate descriptive statistics.
        Gibraltar_num.describe()
```

Out[9]:

	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 [10]: print(Gibraltar_num['Deaths'].sum())

         25412.0

In [11]: print(Gibraltar_num['Cases'].sum())

         1413853.0

In [12]: print(Gibraltar_num['Recovered'].sum())

         956103.0

In [13]: print(Gibraltar_num['Hospitalised'].sum())

         649459.0
```

Figure 3: Running the descriptive statistics and sum functions on Gibraltar_num DataFrame



```
In [17]: Gibraltar_vac_num = Gibraltar_vac[['Vaccinated', 'First Dose', 'Second Dose']]
         Gibraltar_vac_num.describe()
```

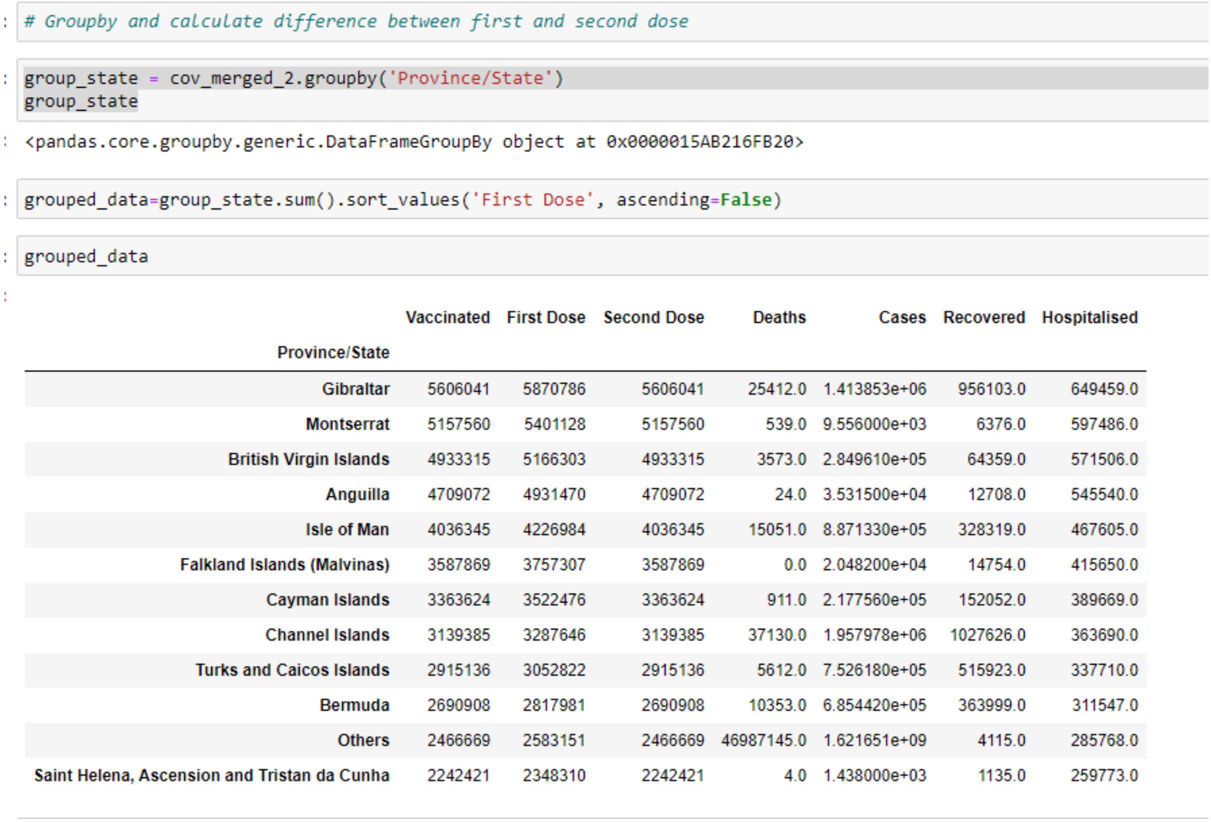
Out[17]:

	Vaccinated	First Dose	Second Dose
count	632.000000	632.000000	632.000000
mean	8870.318038	9289.218354	8870.318038
std	15439.487761	16287.230372	15439.487761
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000
75%	14594.000000	12488.750000	14594.000000
max	69619.000000	94038.000000	69619.000000

```
In [23]: print(Gibraltar_vac_num['First Dose'].sum())

         5870786

In [24]: print(Gibraltar_vac_num['Second Dose'].sum())

         5606041

In [25]: print(Gibraltar_vac_num['Vaccinated'].sum())

         5606041
```

Figure 4: Running the descriptive statistics on Gibraltar_vac_num DataFrame

After merging the 2 datasets together, we are able to view all the vaccination and covid data all in one data frame. The State that has received the highest number of First Doses is Gibraltar, although the State that has the highest percentage who have received the first dose but not the second dose is, ‘Saint Helena, Ascension and Tristan da Cunha’ and so this could be a State for the government to target to boost fully vaccinated numbers.



```
# Groupby and calculate difference between first and second dose
```

```
group_state = cov_merged_2.groupby('Province/State')
group_state
```

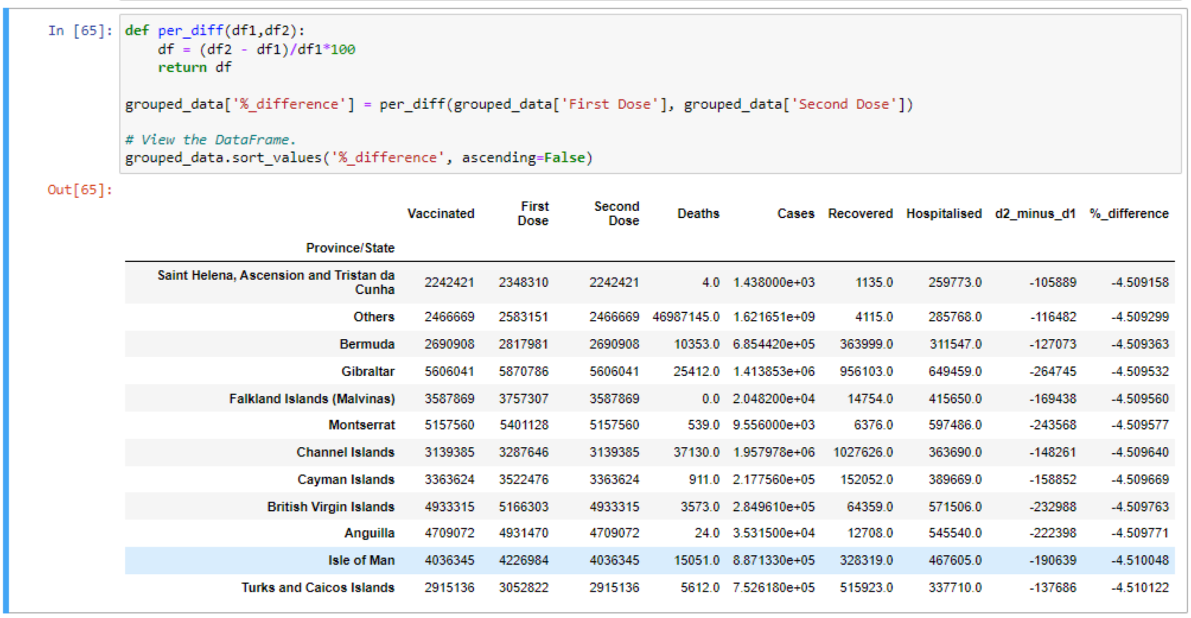
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x0000015AB216FB20>
```

```
grouped_data=group_state.sum().sort_values('First Dose', ascending=False)
```

```
grouped_data
```

	Vaccinated	First Dose	Second Dose	Deaths	Cases	Recovered	Hospitalised
Province/State							
Gibraltar	5606041	5870786	5606041	25412.0	1.413853e+06	956103.0	649459.0
Montserrat	5157560	5401128	5157560	539.0	9.556000e+03	6376.0	597486.0
British Virgin Islands	4933315	5166303	4933315	3573.0	2.849610e+05	64359.0	571506.0
Anguilla	4709072	4931470	4709072	24.0	3.531500e+04	12708.0	545540.0
Isle of Man	4036345	4226984	4036345	15051.0	8.871330e+05	328319.0	467605.0
Falkland Islands (Malvinas)	3587869	3757307	3587869	0.0	2.048200e+04	14754.0	415650.0
Cayman Islands	3363624	3522476	3363624	911.0	2.177560e+05	152052.0	389669.0
Channel Islands	3139385	3287646	3139385	37130.0	1.957978e+06	1027626.0	363690.0
Turks and Caicos Islands	2915136	3052822	2915136	5612.0	7.526180e+05	515923.0	337710.0
Bermuda	2690908	2817981	2690908	10353.0	6.854420e+05	363999.0	311547.0
Others	2466669	2583151	2466669	46987145.0	1.621651e+09	4115.0	285768.0
Saint Helena, Ascension and Tristan da Cunha	2242421	2348310	2242421	4.0	1.438000e+03	1135.0	259773.0

Figure 5: Grouped data by Province/State to show highest number of individuals with First Dose



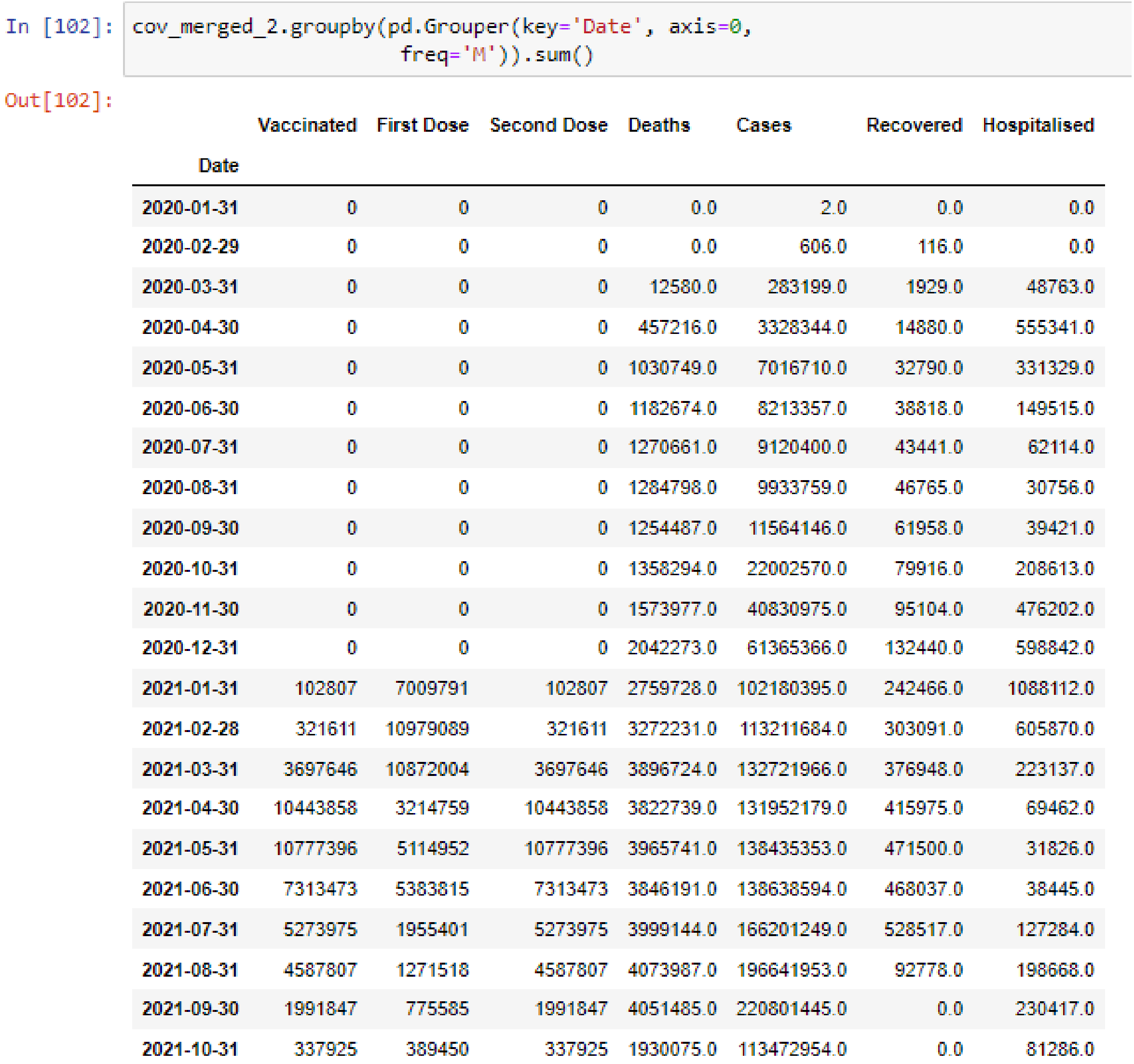
```
In [65]: def per_diff(df1,df2):
             df = (df2 - df1)/df1*100
             return df

         grouped_data['%_difference'] = per_diff(grouped_data['First Dose'], grouped_data['Second Dose'])

         # View the DataFrame.
         grouped_data.sort_values('%_difference', ascending=False)
```

Out[65]:

	Vaccinated	First Dose	Second Dose	Deaths	Cases	Recovered	Hospitalised	d2_minus_d1	%_difference
Province/State									
Saint Helena, Ascension and Tristan da Cunha	2242421	2348310	2242421	4.0	1.438000e+03	1135.0	259773.0	-105889	-4.509158
Others	2466669	2583151	2466669	46987145.0	1.621651e+09	4115.0	285768.0	-116482	-4.509299
Bermuda	2690908	2817981	2690908	10353.0	6.854420e+05	363999.0	311547.0	-127073	-4.509363
Gibraltar	5606041	5870786	5606041	25412.0	1.413853e+06	956103.0	649459.0	-264745	-4.509532
Falkland Islands (Malvinas)	3587869	3757307	3587869	0.0	2.048200e+04	14754.0	415650.0	-169438	-4.509560
Montserrat	5157560	5401128	5157560	539.0	9.556000e+03	6376.0	597486.0	-243568	-4.509577
Channel Islands	3139385	3287646	3139385	37130.0	1.957978e+06	1027626.0	363690.0	-148261	-4.509640
Cayman Islands	3363624	3522476	3363624	911.0	2.177560e+05	152052.0	389669.0	-158852	-4.509669
British Virgin Islands	4933315	5166303	4933315	3573.0	2.849610e+05	64359.0	571506.0	-232988	-4.509763
Anguilla	4709072	4931470	4709072	24.0	3.531500e+04	12708.0	545540.0	-222398	-4.509771
Isle of Man	4036345	4226984	4036345	15051.0	8.871330e+05	328319.0	467605.0	-190639	-4.510048
Turks and Caicos Islands	2915136	3052822	2915136	5612.0	7.526180e+05	515923.0	337710.0	-137686	-4.510122

Figure 6: Grouped data by Province/State to show highest percentage to not receive Second Dose after receiving First Dose

I grouped the date by month and aggregated the data to show the sum for each variable:



```
In [102]: cov_merged_2.groupby(pd.Grouper(key='Date', axis=0,
                                         freq='M')).sum()
```

```
Out[102]:
```

Date	Vaccinated	First Dose	Second Dose	Deaths	Cases	Recovered	Hospitalised
2020-01-31	0	0	0	0.0	2.0	0.0	0.0
2020-02-29	0	0	0	0.0	606.0	116.0	0.0
2020-03-31	0	0	0	12580.0	283199.0	1929.0	48763.0
2020-04-30	0	0	0	457216.0	3328344.0	14880.0	555341.0
2020-05-31	0	0	0	1030749.0	7016710.0	32790.0	331329.0
2020-06-30	0	0	0	1182674.0	8213357.0	38818.0	149515.0
2020-07-31	0	0	0	1270661.0	9120400.0	43441.0	62114.0
2020-08-31	0	0	0	1284798.0	9933759.0	46765.0	30756.0
2020-09-30	0	0	0	1254487.0	11564146.0	61958.0	39421.0
2020-10-31	0	0	0	1358294.0	22002570.0	79916.0	208613.0
2020-11-30	0	0	0	1573977.0	40830975.0	95104.0	476202.0
2020-12-31	0	0	0	2042273.0	61365366.0	132440.0	598842.0
2021-01-31	102807	7009791	102807	2759728.0	102180395.0	242466.0	1088112.0
2021-02-28	321611	10979089	321611	3272231.0	113211684.0	303091.0	605870.0
2021-03-31	3697646	10872004	3697646	3896724.0	132721966.0	376948.0	223137.0
2021-04-30	10443858	3214759	10443858	3822739.0	131952179.0	415975.0	69462.0
2021-05-31	10777396	5114952	10777396	3965741.0	138435353.0	471500.0	31826.0
2021-06-30	7313473	5383815	7313473	3846191.0	138638594.0	468037.0	38445.0
2021-07-31	5273975	1955401	5273975	3999144.0	166201249.0	528517.0	127284.0
2021-08-31	4587807	1271518	4587807	4073987.0	196641953.0	92778.0	198668.0
2021-09-30	1991847	775585	1991847	4051485.0	220801445.0	0.0	230417.0
2021-10-31	337925	389450	337925	1930075.0	113472954.0	0.0	81286.0

Figure 7: Table to show data grouped by month with sum for each variable.

This shows us that the number of vaccinated individuals goes up from Jan 2021 to May 2021 but then we see a fall thereafter. With the first dose numbers, there are fluctuations throughout the entire timeframe but in general we see that there is a decline in the numbers being vaccinated towards the end of time. We can also note that, around April/May 2021 there are drops in the number of hospitalised in the corresponding months (April to June 2021); this could be due to the positive effects of the vaccine and so fewer people required hospitalisations.

Using the Gibraltar subset, I used the melt() function to transform the data to be able to visualize the 'Cases', 'Recovered', 'Hospitalised' and 'Deaths' over time.

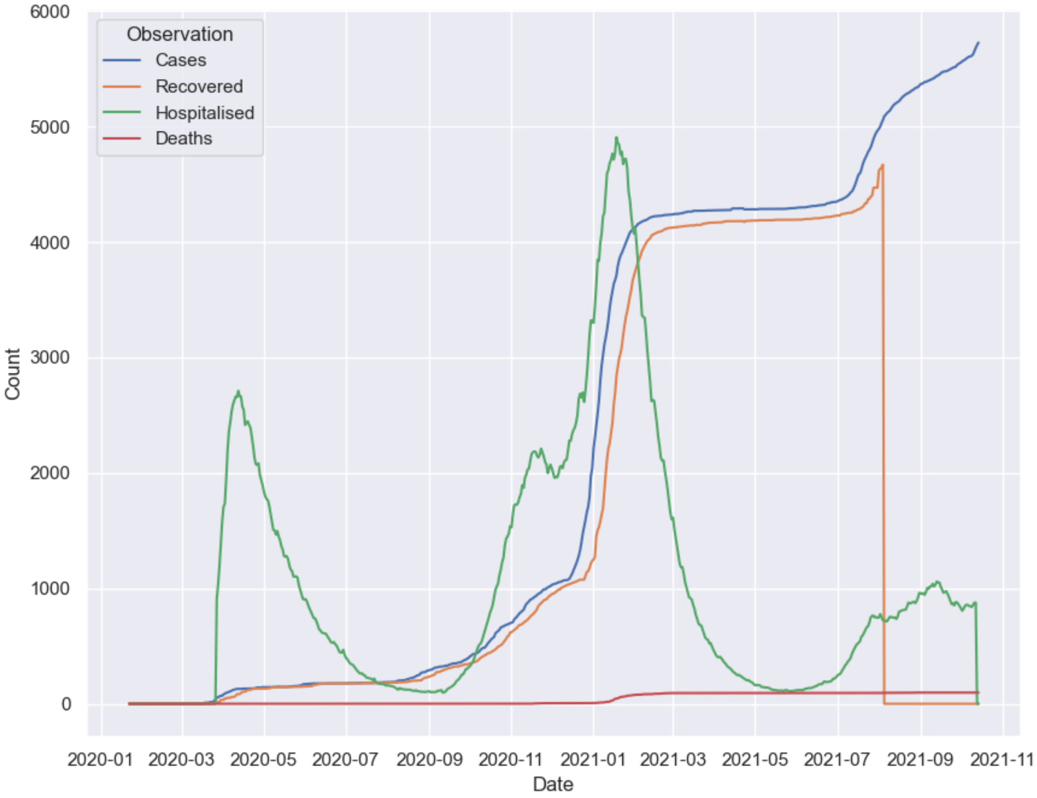


Figure 8: Plot to show Cases, Recovered, Hospitalised and Deaths for Gibraltar over time

This clearly shows what was mentioned earlier about irregularities with the dataset; how can hospitalised cases be higher than the number of Covid cases at any given timepoint? There is a steep drop in the number of recovered cases at one point; shouldn't the number of deaths increase at around the same time?

Calculating the percentage of people who are partially and fully vaccinated showed the same trend; fully vaccinated is 95.49% for all States. This again makes me question the integrity of the data; if the percentage of fully vaccinated is so high, the Government would not be working to boost its efforts to increase vaccination rates through marketing.

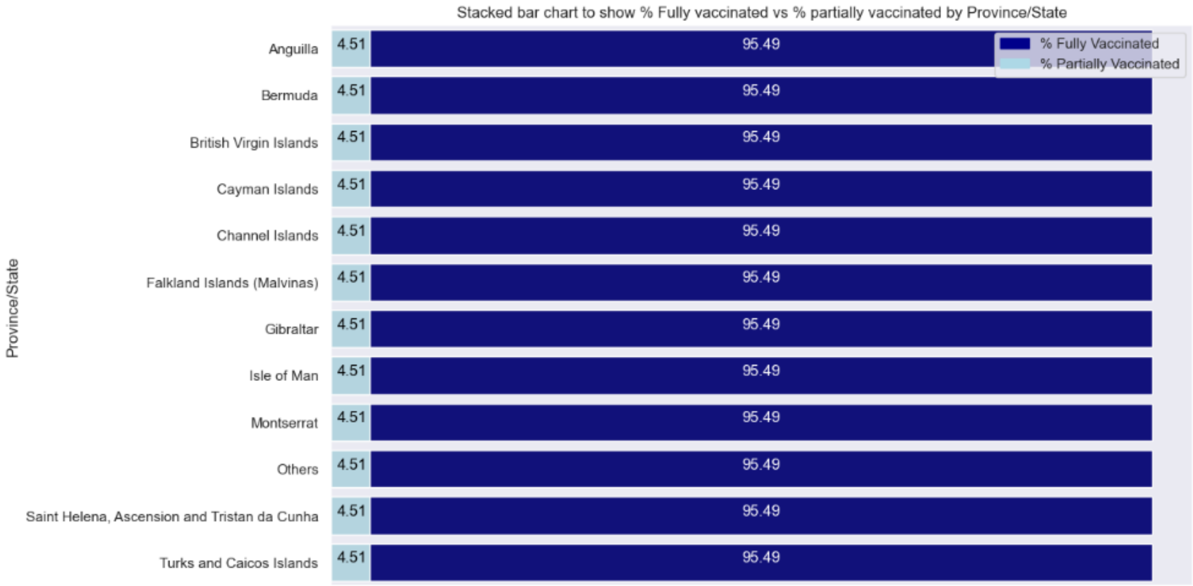


Figure 9: Stacked Bar Plot to show percentage of vaccinated and partially vaccinated for each State.

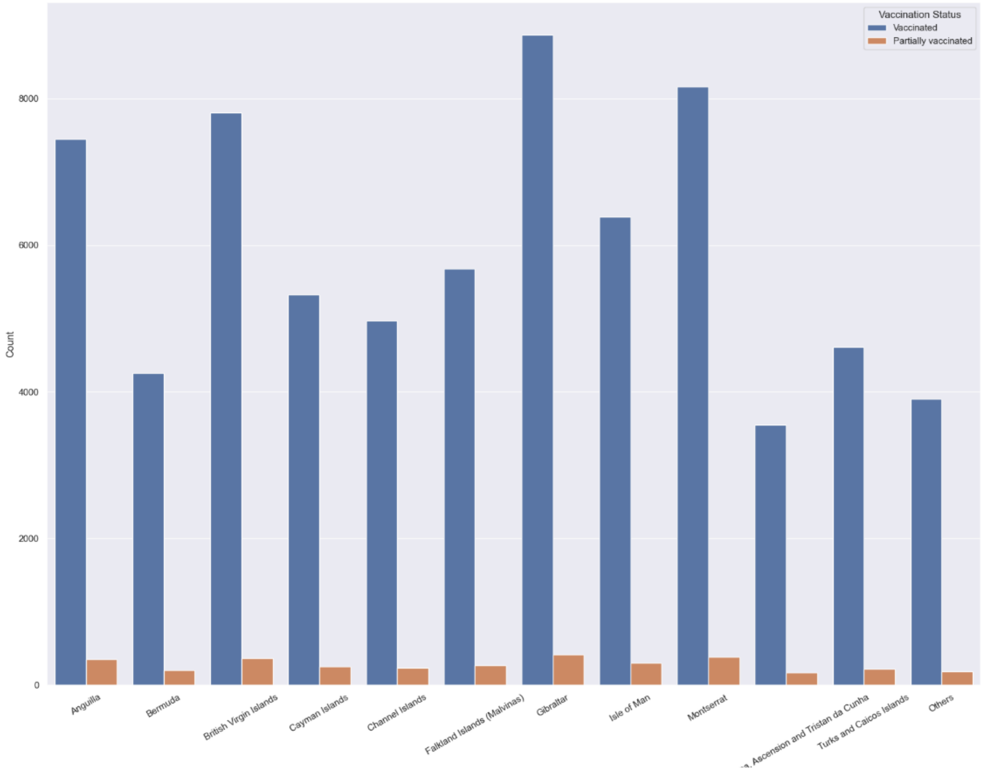
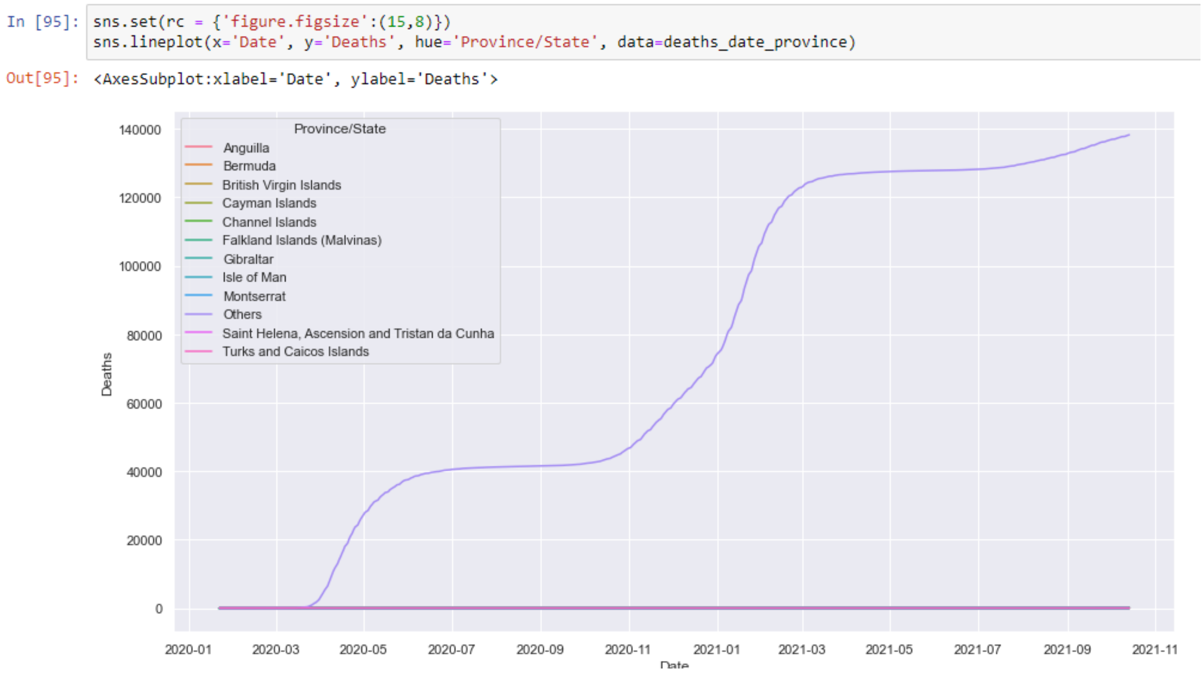


Figure 10: Bar Plot to show number of vaccinated and partially vaccinated cases for each State.

Whilst initially visualising the line plot for deaths grouped by Province/State over time, the plot was not showing correctly:



```
In [95]: sns.set(rc = {'figure.figsize':(15,8)})
         sns.lineplot(x='Date', y='Deaths', hue='Province/State', data=deaths_date_province)

Out[95]: <AxesSubplot:xlabel='Date', ylabel='Deaths'>
```

Figure 11: Skewed Line Plot for deaths grouped by Province/State.

Aggregating the deaths, I was able to identify the State that skewed the plot; 'Others'.

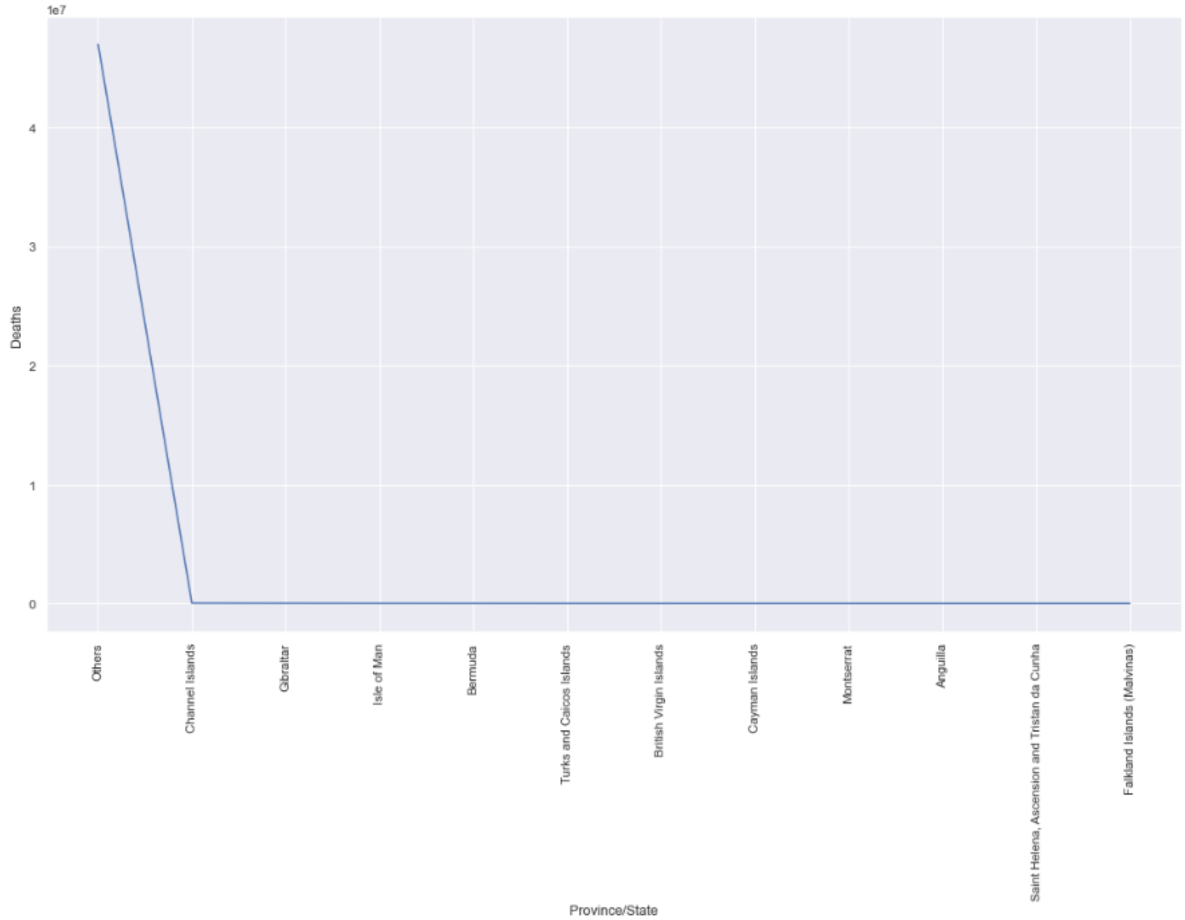
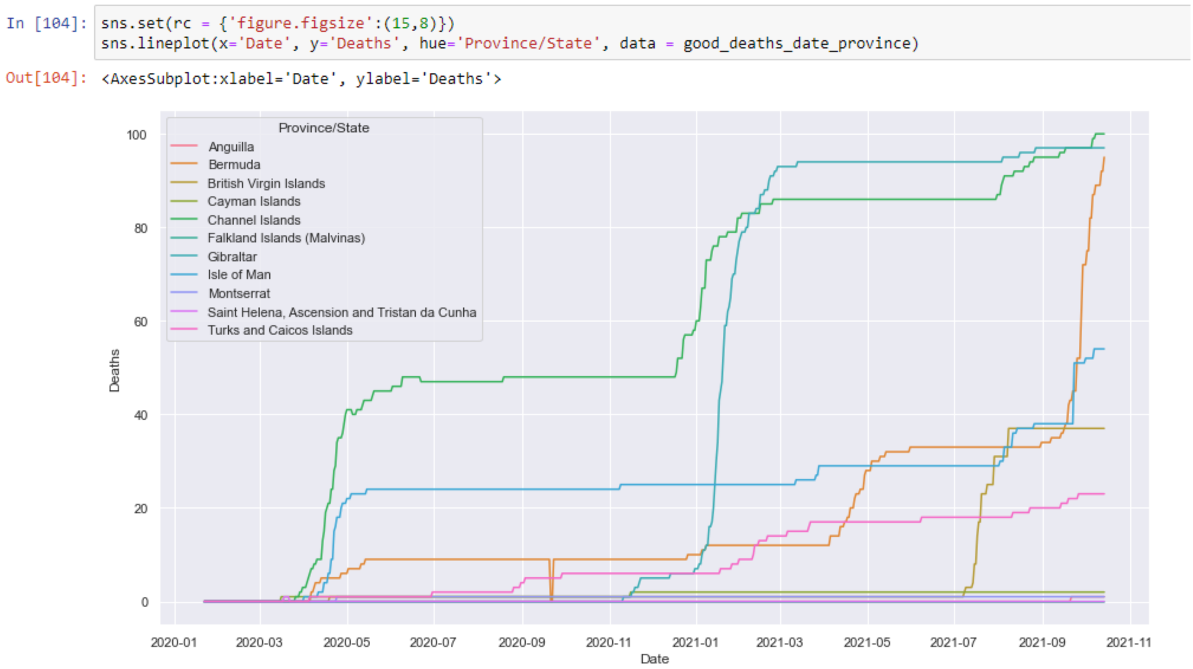


Figure 12: Skewed Line Plot identifying Province/State with skewed data for 'Deaths'.

Once I removed this State, the plot was showing up as I would expect it to:



```
In [104]: sns.set(rc = {'figure.figsize':(15,8)})
          sns.lineplot(x='Date', y='Deaths', hue='Province/State', data = good_deaths_date_province)

Out[104]: <AxesSubplot:xlabel='Date', ylabel='Deaths'>
```

Figure 13: Line Plot to show the Deaths over time grouped by Province/State.

By converting the dates to months/year, I was able to further improve on this visualisation:

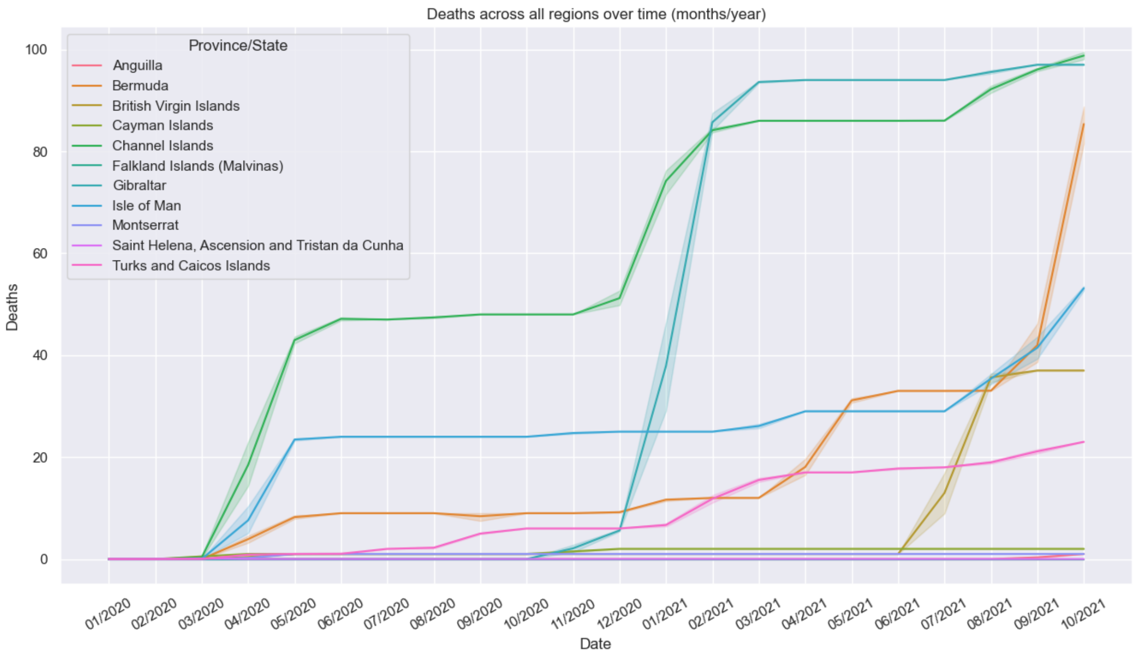


Figure 14: Line Plot to show the Deaths over time (month/year) grouped by Province/State.

This line plot shows that the death rates are increasing for all States and it doesn't seem to have reached the peak yet but starting to plateau. In addition, the deaths seem particularly low for some States; Falkland Islands has zero deaths, which doesn't seem correct and may need to be investigated. The States with the highest deaths are Channel Islands and Gibraltar.

Recovered cases are increasing over time for all States but then all of a sudden there is an abrupt drop for all regions; is there data missing from July 2021 onwards?

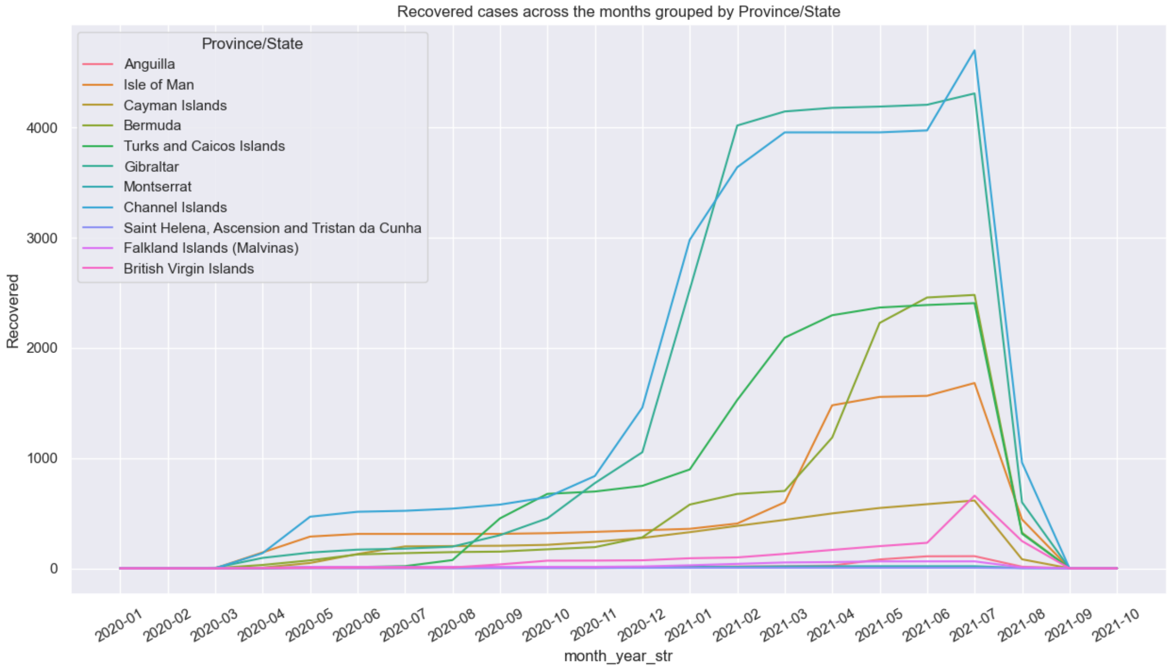


Figure 15: Line Plot to show the Recovered cases over time (month/year) grouped by Province/State.

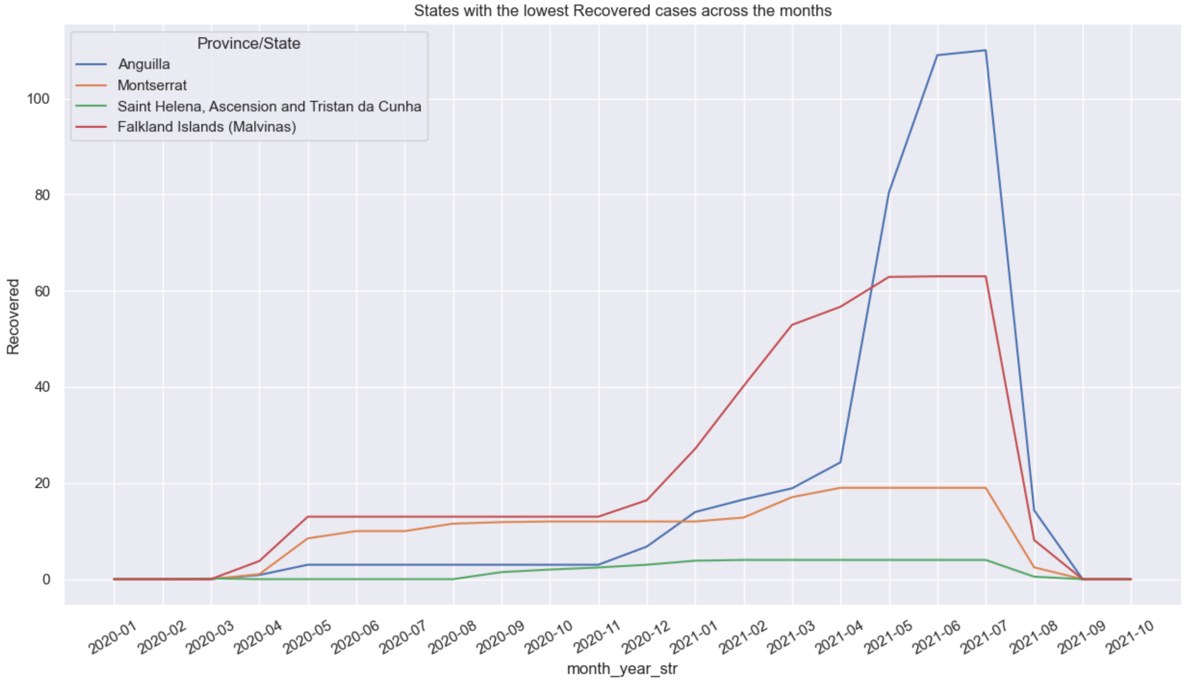
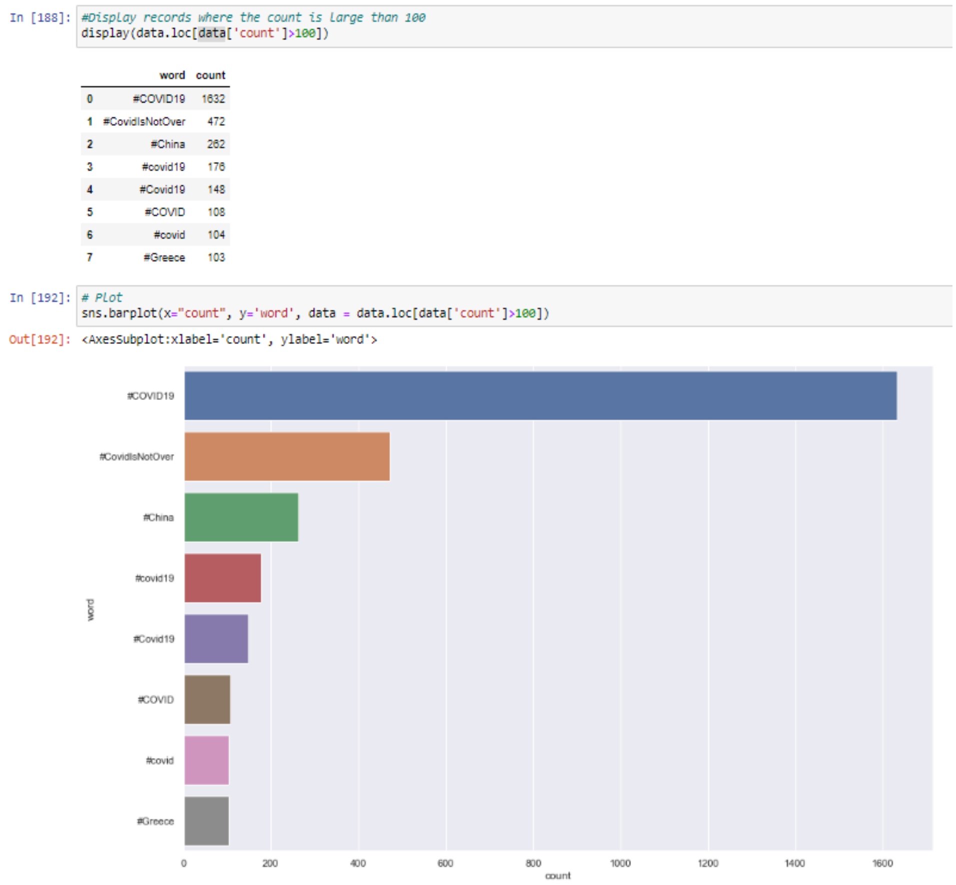


Figure 16: Line Plot to show the States with the lowest Recovered cases over time (month/year)

The Channel Islands had the highest number of recovered individuals over time, with Gibraltar briefly overtaking them from January 2021 to June 2021. Taking this information into consideration, the government could look to focus their vaccination drive at other regions where the recoveries is not so high (Figure 16).

After importing the ‘Tweets’ data into Python, I was able to identify the hashtags that were mentioned the most and visualise these as a bar plot:



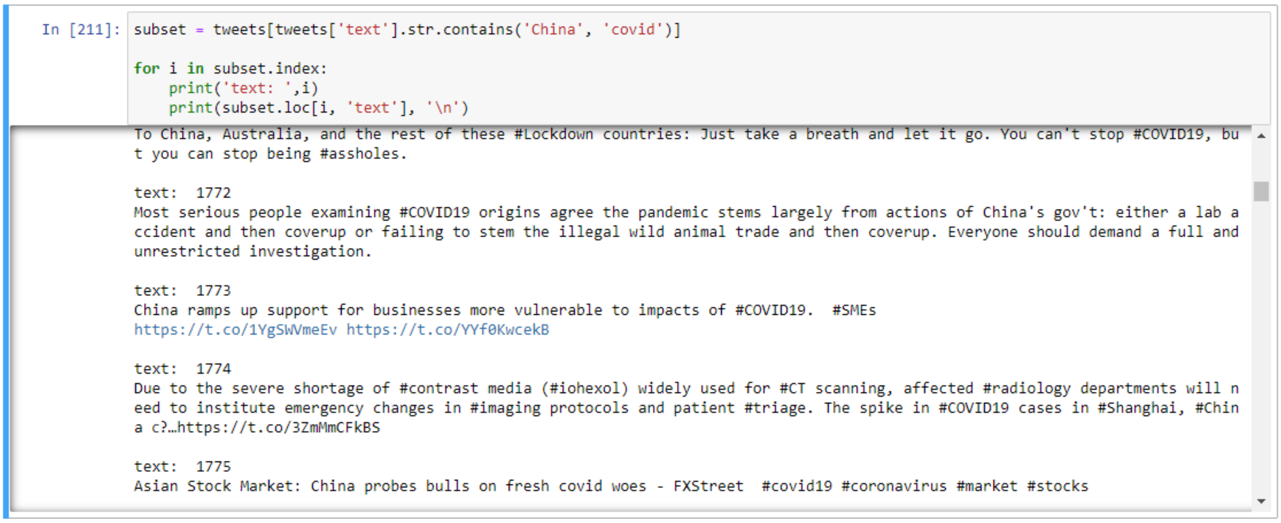
```
In [188]: #Display records where the count is large than 100
          display(data.loc[data['count']>100])
```

	word	count
0	#COVID19	1632
1	#CovidIsNotOver	472
2	#China	262
3	#covid19	176
4	#Covid19	148
5	#COVID	108
6	#covid	104
7	#Greece	103

```
In [192]: # Plot
          sns.barplot(x="count", y='word', data = data.loc[data['count']>100])
Out[192]: <AxesSubplot:xlabel='count', ylabel='word'>
```

Figure 17: Bar Plot to show highest trending hashtags.

#COVID19, #CovidIsNotOver and #China were the most common hashtags. This shows us our serious of a concern Covid was. You can see in the visualisation that 2 countries are in the plot; China and Greece. This is likely due to people tweeting about high covid cases in those countries at the time although we should investigate if these tweets are actually referring to covid.



```
In [211]: subset = tweets[tweets['text'].str.contains('China', 'covid')]

          for i in subset.index:
              print('text: ',i)
              print(subset.loc[i, 'text'], '\n')
```

```
To China, Australia, and the rest of these #Lockdown countries: Just take a breath and let it go. You can't stop #COVID19, bu
t you can stop being #assholes.

text:  1772
Most serious people examining #COVID19 origins agree the pandemic stems largely from actions of China's gov't: either a lab a
ccident and then coverup or failing to stem the illegal wild animal trade and then coverup. Everyone should demand a full and
unrestricted investigation.

text:  1773
China ramps up support for businesses more vulnerable to impacts of #COVID19.  #SMEs
https://t.co/1YgSWVmeEv https://t.co/YYf0KwcekB

text:  1774
Due to the severe shortage of #contrast media (#iohexol) widely used for #CT scanning, affected #radiology departments will n
eed to institute emergency changes in #imaging protocols and patient #triage. The spike in #COVID19 cases in #Shanghai, #Chin
a c?…https://t.co/3ZmMmCFkBS

text:  1775
Asian Stock Market: China probes bulls on fresh covid woes - FXStreet  #covid19 #coronavirus #market #stocks
```

Figure 18: Tweets with China mentioned in them.

Despite the concerns relating to the reliability of the data, which should be investigated further, we were able to draw some interesting trends and insights from the data that we have to hand. I found that the number of vaccinations is going down, after an initial increase. We have identified the States that have the highest percentage in terms receiving the first dose but not the second dose; these States are an easy target for the government in the first instance to boost vaccination numbers.

Lineplots helped to visualise the trends for all the States over time in terms of deaths and recoveries. I found that the deaths were increasing for all States and for some it looked as though the peak hadn't been reached yet. With regards to recoveries, I found that this was also increasing over time and I was able to identify the States with the highest recoveries and so these States could be targeted at a later stage by the government when looking to boost vaccinations, focusing initially at the States with lower recovery numbers. A visualisation that will help in boosting the vaccination rates is a scatterplot to show the relationship between vaccination and hospitalisation; this shows that as the number of vaccinations increases, the hospitalised numbers fall (figure 19). This is important as it helps to portray the positive effect of vaccines and in turn it will help reduce the burden on the NHS; this'll be a big plus point for the government.

To further the analysis, it would be useful to have population data available for all States. This would help us to interpret which State's actually had the best/worst vaccination uptake rates and also how serious the covid rates (deaths, hospitalisations, recovered) were.

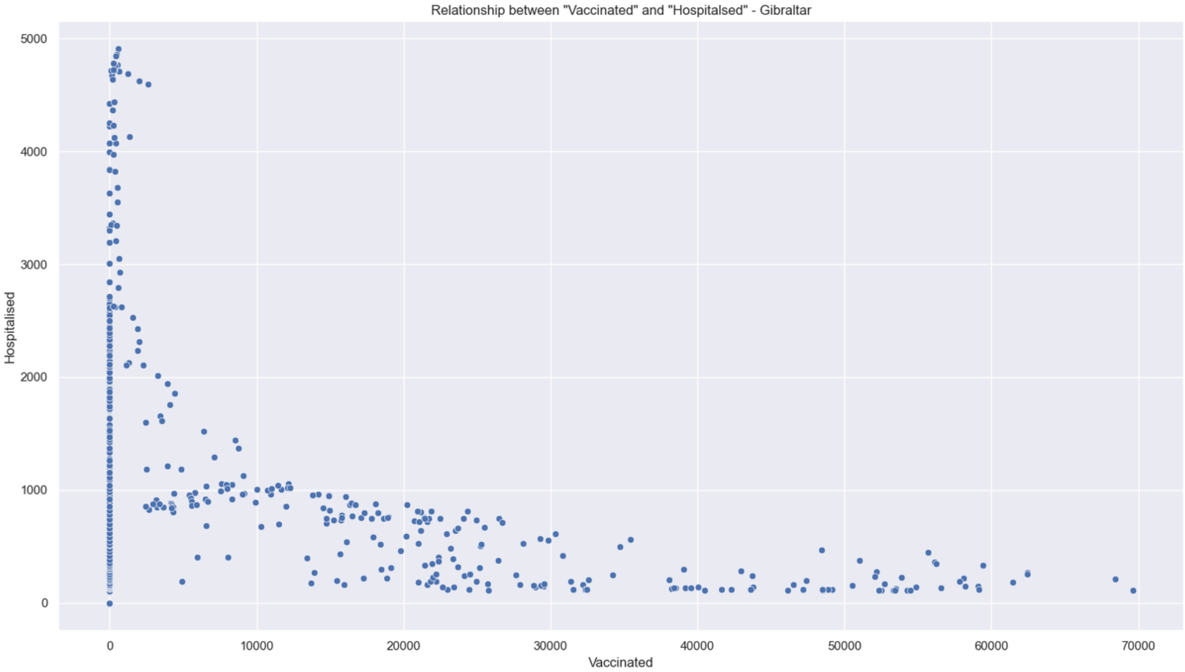


Figure 19: Scatterplot to show the relationship between Vaccinated and Hospitalised – Gibraltar data.

Word count (not including figures and charts) = 1088.