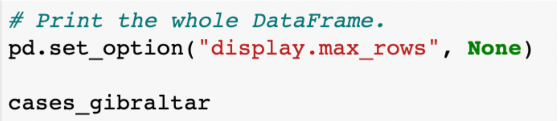
LSE DA201 Assignment Report

I started with importing and exploring the data to get an idea of the data and identified two rows in the cases dataframe with missing values. The Deaths, Cases, Recovered, Hospitalised are missing for Bermuda on September 21st and 22nd 2020.

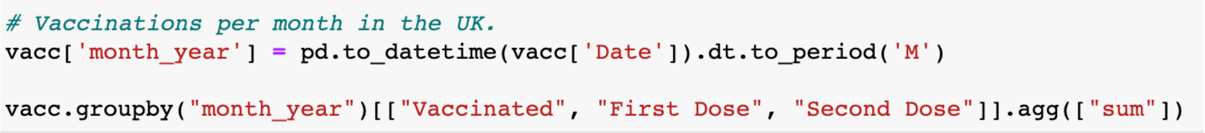
Using the code below I was able to see the full dataframes.



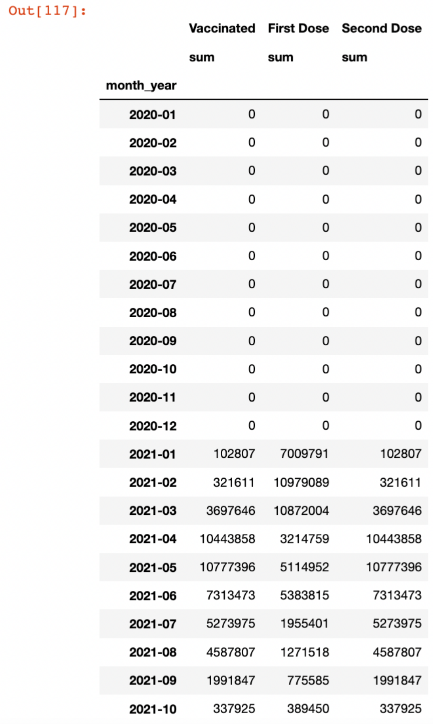
```
# Print the whole DataFrame.
pd.set_option("display.max_rows", None)

cases_gibraltar
```

I noticed that there were no hospitalisations until 2020-03-26 and suddenly 908 people hospitalised in one day. It seems it was not being recorded before March 26th. Sense checked this with Anguilla. data and saw the same pattern. The same with vaccinations; no first doses before January 11th but on that day, there are both first and second doses. There must be missing data as it is not possible for anyone to get their first and second dose on the same day.

I used this code to group the vaccinations by month.



```
# Vaccinations per month in the UK.
vacc['month_year'] = pd.to_datetime(vacc['Date']).dt.to_period('M')

vacc.groupby("month_year")[["Vaccinated", "First Dose", "Second Dose"]].agg(["sum"])
```



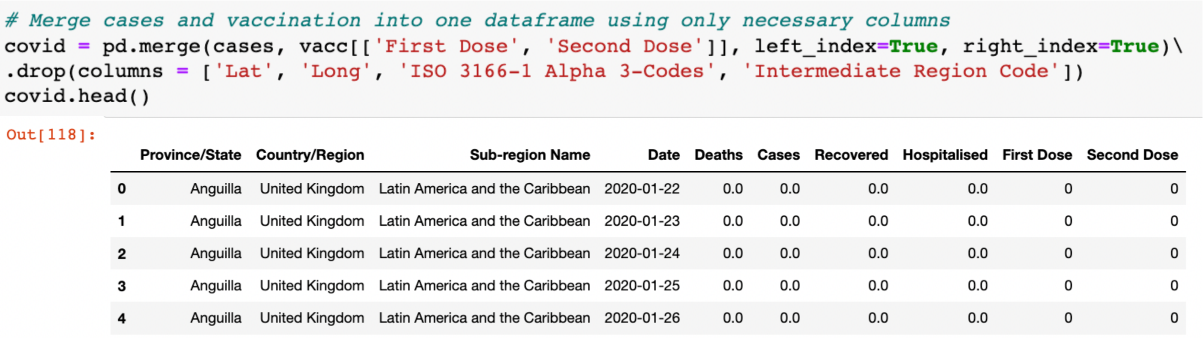
Out[117]:

	Vaccinated	First Dose	Second Dose
	sum	sum	sum
month_year			
2020-01	0	0	0
2020-02	0	0	0
2020-03	0	0	0
2020-04	0	0	0
2020-05	0	0	0
2020-06	0	0	0
2020-07	0	0	0
2020-08	0	0	0
2020-09	0	0	0
2020-10	0	0	0
2020-11	0	0	0
2020-12	0	0	0
2021-01	102807	7009791	102807
2021-02	321611	10979089	321611
2021-03	3697646	10872004	3697646
2021-04	10443858	3214759	10443858
2021-05	10777396	5114952	10777396
2021-06	7313473	5383815	7313473
2021-07	5273975	1955401	5273975
2021-08	4587807	1271518	4587807
2021-09	1991847	775585	1991847
2021-10	337925	389450	337925

I noticed in the dataframe that the Vaccinated and Second Dose columns are the same, so there is no need for the vaccinated column.

I also noticed that first doses administered per month peaked in February and March with over 10 million doses per month, and second doses administered per month peaked after that in April and May. The drop in first doses administered in April compared to March would indicate that at that point most eligible people had received their first dose and so there was a greater need to administer second doses.

I merged the cases and vaccinations tables to get one comprehensive covid table.

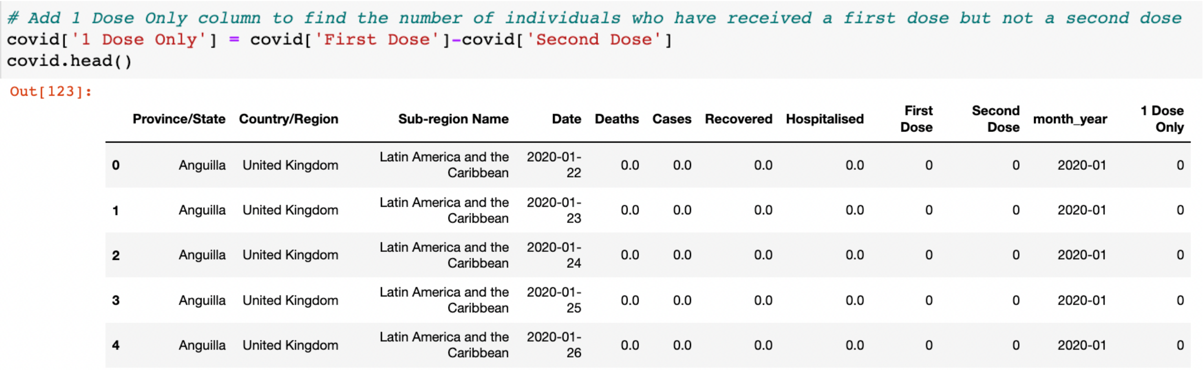


```
# Merge cases and vaccination into one dataframe using only necessary columns
covid = pd.merge(cases, vacc[['First Dose', 'Second Dose']], left_index=True, right_index=True)\
.drop(columns = ['Lat', 'Long', 'ISO 3166-1 Alpha 3-Codes', 'Intermediate Region Code'])
covid.head()
```

Out[118]:

	Province/State	Country/Region	Sub-region Name	Date	Deaths	Cases	Recovered	Hospitalised	First Dose	Second Dose
0	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-22	0.0	0.0	0.0	0.0	0	0
1	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-23	0.0	0.0	0.0	0.0	0	0
2	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-24	0.0	0.0	0.0	0.0	0	0
3	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-25	0.0	0.0	0.0	0.0	0	0
4	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-26	0.0	0.0	0.0	0.0	0	0

I then added a new column calculating the number of partially vaccinated people.

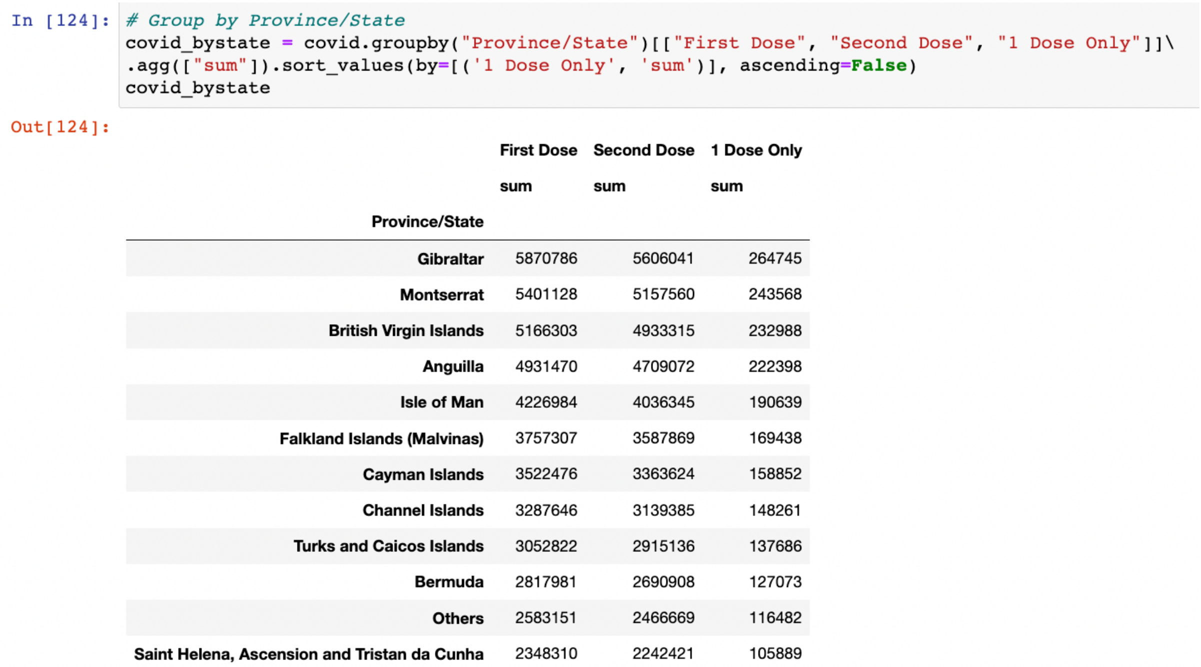


```
# Add 1 Dose Only column to find the number of individuals who have received a first dose but not a second dose
covid['1 Dose Only'] = covid['First Dose']-covid['Second Dose']
covid.head()
```

Out[123]:

	Province/State	Country/Region	Sub-region Name	Date	Deaths	Cases	Recovered	Hospitalised	First Dose	Second Dose	month_year	1 Dose Only
0	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-22	0.0	0.0	0.0	0.0	0	0	2020-01	0
1	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-23	0.0	0.0	0.0	0.0	0	0	2020-01	0
2	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-24	0.0	0.0	0.0	0.0	0	0	2020-01	0
3	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-25	0.0	0.0	0.0	0.0	0	0	2020-01	0
4	Anguilla	United Kingdom	Latin America and the Caribbean	2020-01-26	0.0	0.0	0.0	0.0	0	0	2020-01	0

I then grouped the vaccine data by state.



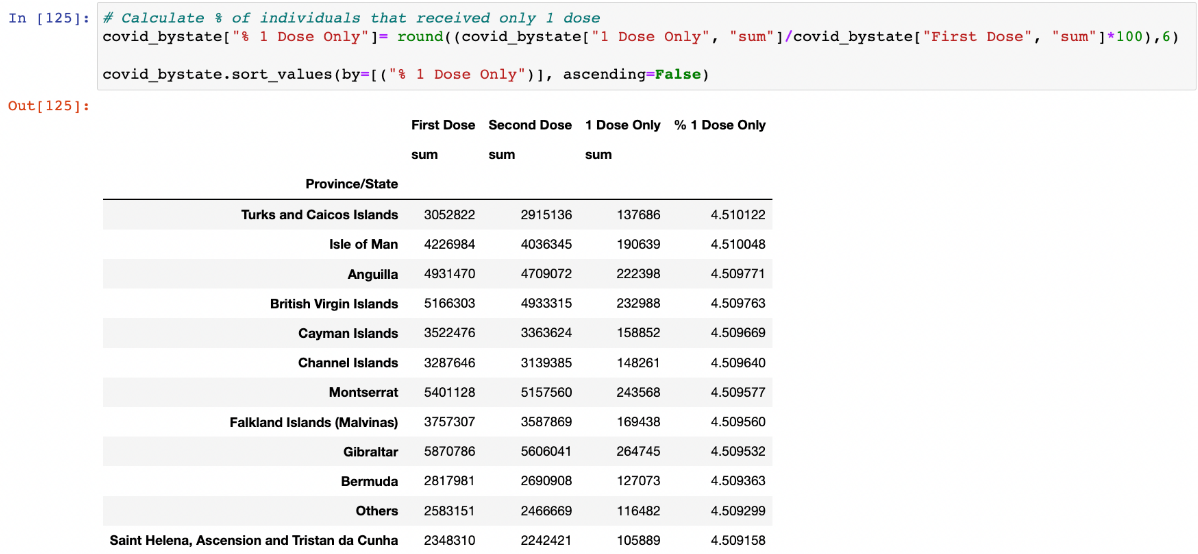
```
In [124]: # Group by Province/State
          covid_bystate = covid.groupby("Province/State")[["First Dose", "Second Dose", "1 Dose Only"]]\
          .agg(["sum"]).sort_values(by=[('1 Dose Only', 'sum')], ascending=False)
          covid_bystate
```

Out[124]:

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

The table allowed me to identify that Gibraltar has the highest number of individuals who have received the first dose, second dose, and first dose only.

I then calculated the percentage of people that are partially vaccinated.

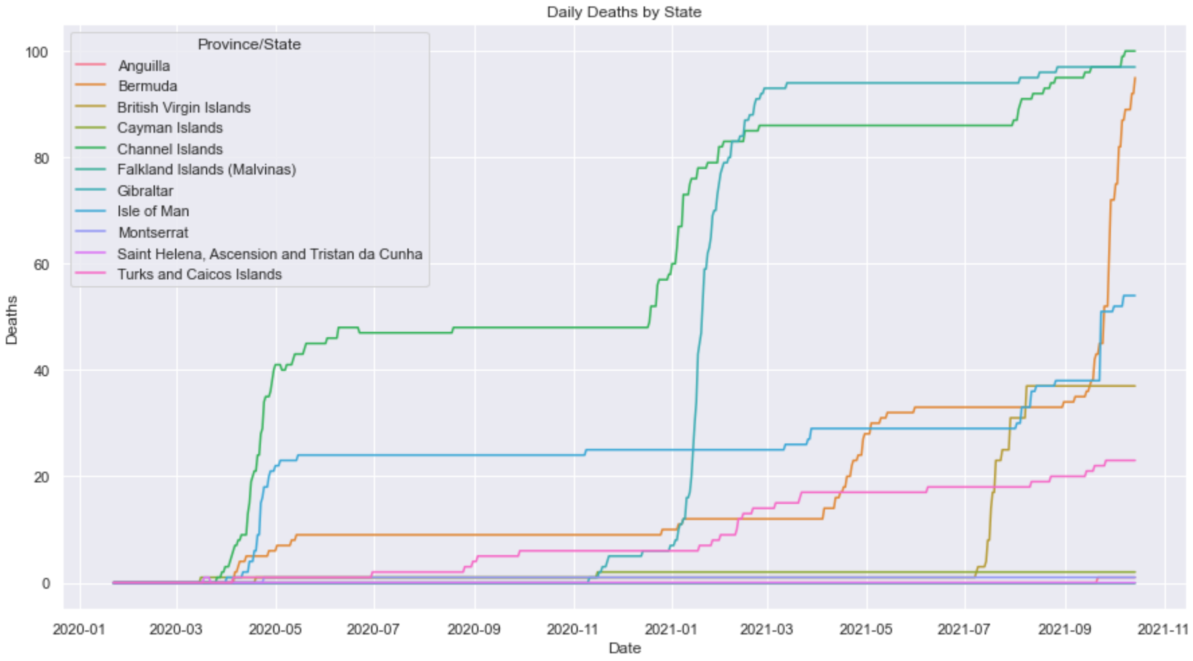


```
In [125]: # Calculate % of individuals that received only 1 dose
          covid_bystate["% 1 Dose Only"]= round((covid_bystate["1 Dose Only", "sum"]/covid_bystate["First Dose", "sum"]*100),6)

          covid_bystate.sort_values(by=[("% 1 Dose Only")], ascending=False)
Out[125]:
```

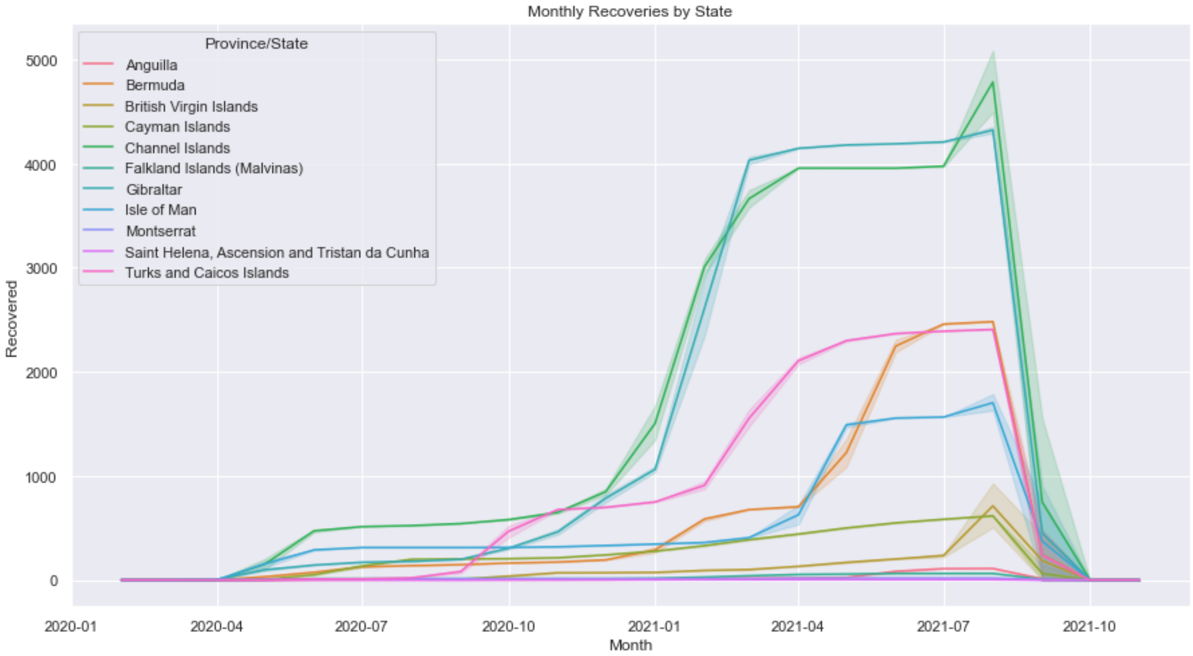
	First Dose	Second Dose	1 Dose Only	% 1 Dose Only
	sum	sum	sum	
Province/State				
Turks and Caicos Islands	3052822	2915136	137686	4.510122
Isle of Man	4226984	4036345	190639	4.510048
Anguilla	4931470	4709072	222398	4.509771
British Virgin Islands	5166303	4933315	232988	4.509763
Cayman Islands	3522476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

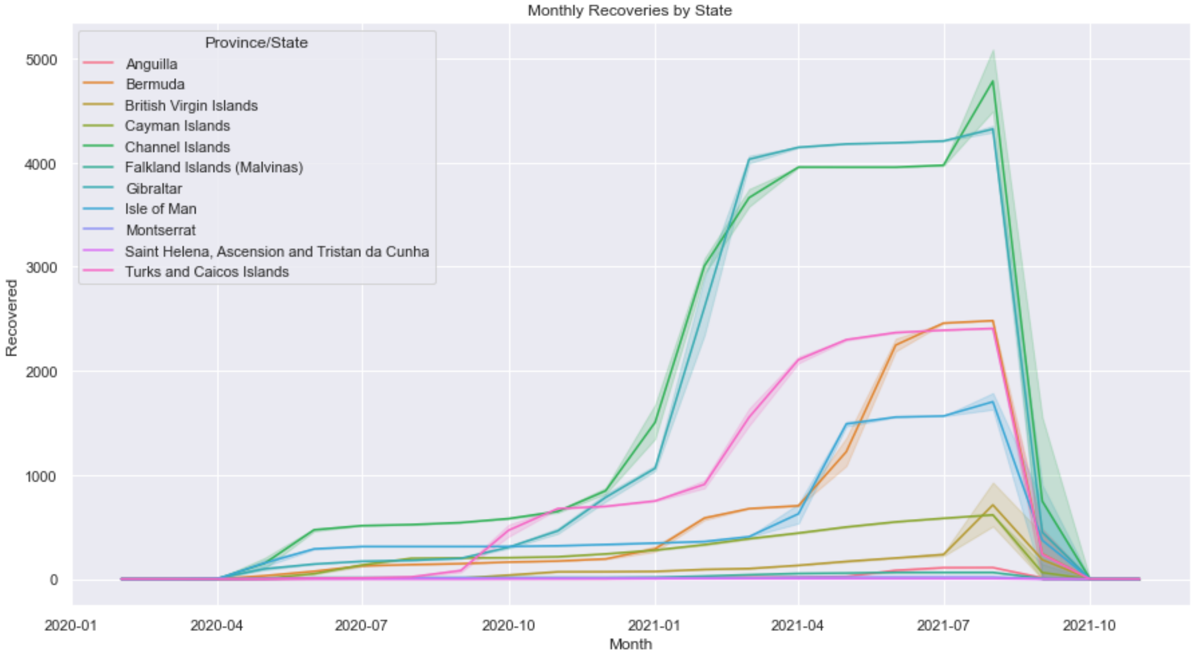
About 4.5% of vaccinated people have received only 1 dose, the remaining 95.5% are fully vaccinated. This is consistent across all regions, but Turks and Caicos Islands had the highest percentage of partially vaccinated people.

I then plotted the deaths, and recoveries by state over time.



At first I plotted the daily data but converting the Date into Months provides a clearer picture by smoothing out the lines on the line chart. This does not take away any important information as the daily changes are not important in this case while looking at data spanning two years.





Findings:

- Gibraltar and Channel Islands have the highest numbers of deaths, recovered cases. Possibly just due to bigger populations.

- The region 'other' skews the data and is not particularly useful when looking at the differences between regions and so should be removed when looking at the data grouped by region.

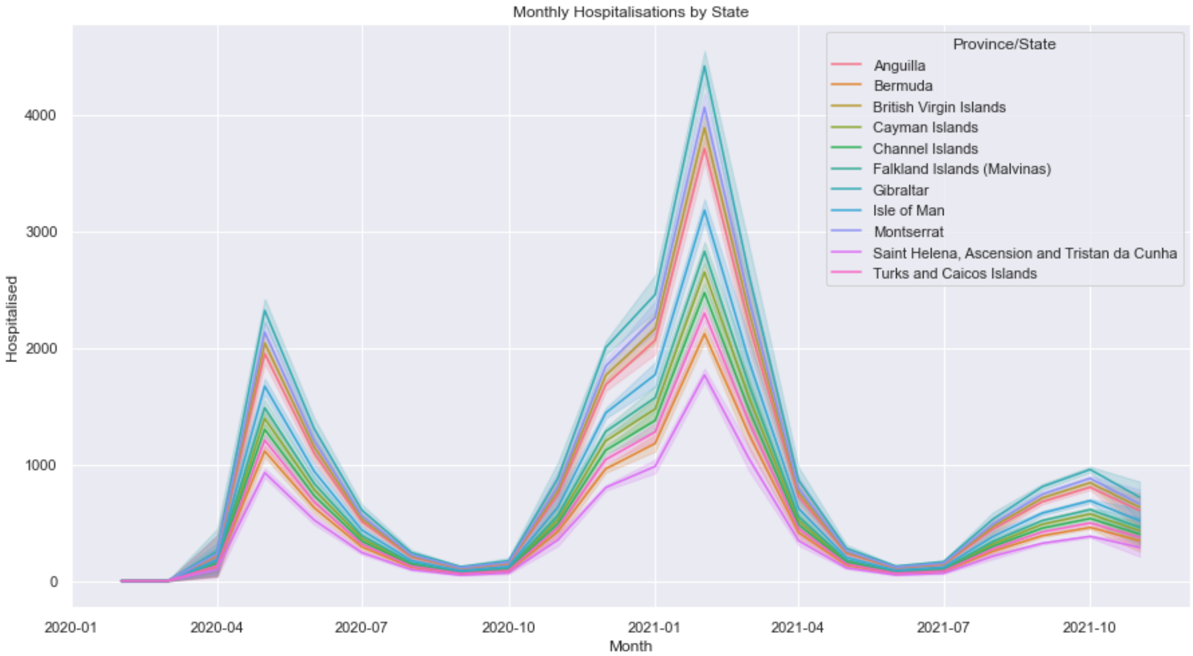
- Deaths are still on the rise in many regions such as the Channel Islands, Bermuda, Isle of Man and Saint Helena. In Gibraltar and British Virgin islands the deaths seems to have plateaued, however they have not yet decreased.

- The Channel Islands seem to be slightly ahead of other regions with spikes in deaths showing first in the Channel Islands and the other regions quickly follow.

- The Channel Islands has had the most recoveries and has been consistently with the exception of November 2020 when Turks and Caicos had the most recoveries, and March 2021 to July 2021 when Gibraltar had the most recoveries.

- These visualisations show which regions possibly need the most urgent attention due to the number and trajectory of deaths, such as Bermuda and Ise of Man where deaths are increasing rapidly and recoveries have dropped.

To give the Deaths and Recoveries data more meaning, it would be useful to have the total population per state. Then we could calculate the percentage of the population that is not fully vaccinated, the percentage of the population that has already recovered from covid and more.

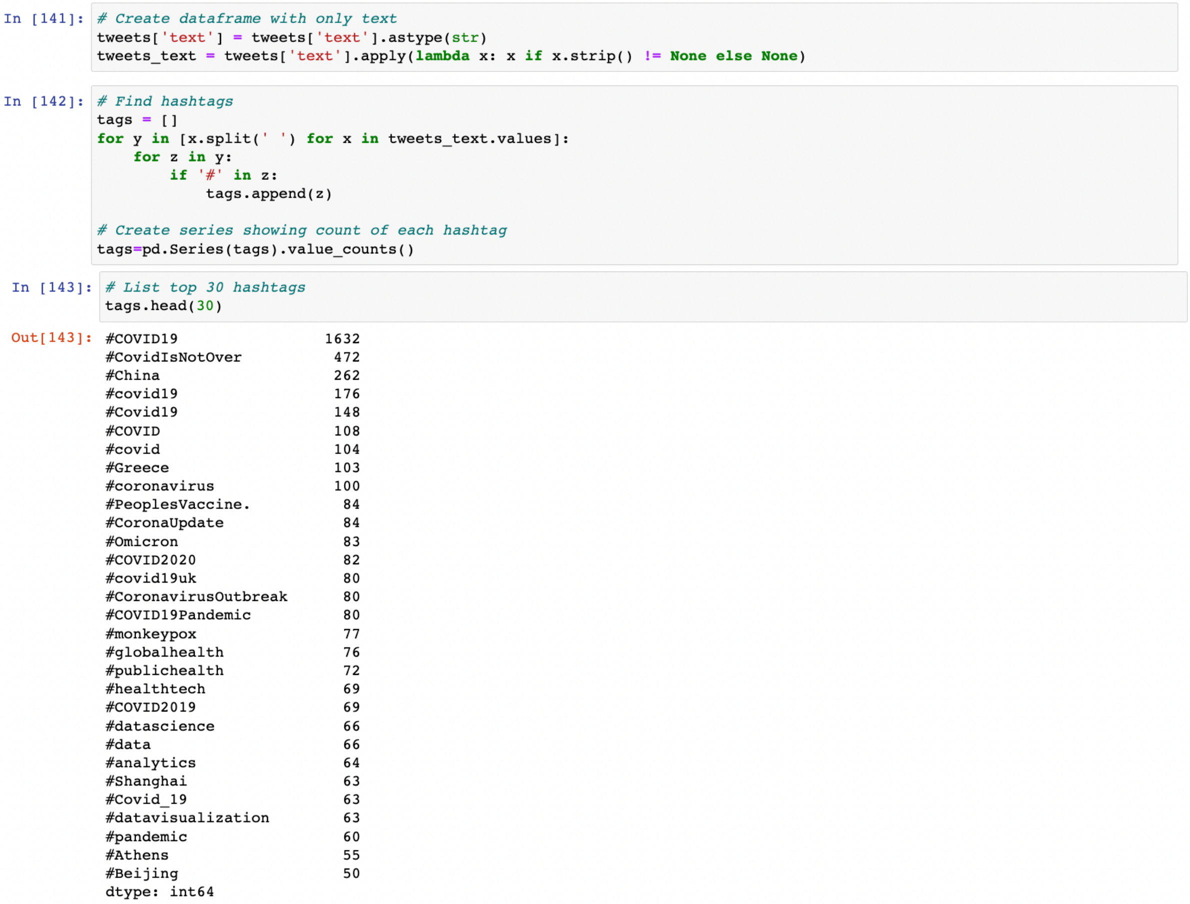


Hospitalisation Findings:

- All regions follow a similar pattern which shows the latest peak was in October 2021, with November figures showing a slight decline in hospitalisations. Gibraltar consistently has the highest number of hospitalisations, but that could be due to having a bigger population than the other states. Again, having population data would give more context to be able to interpret the data. When looking at hospitalisations data it would also be important to have data on hospital capacity; although the graph above may show a state does not have the highest number of hospitalisations, if that state is running out of hospital beds, that might be a more urgent area to focus the vaccination campaign.

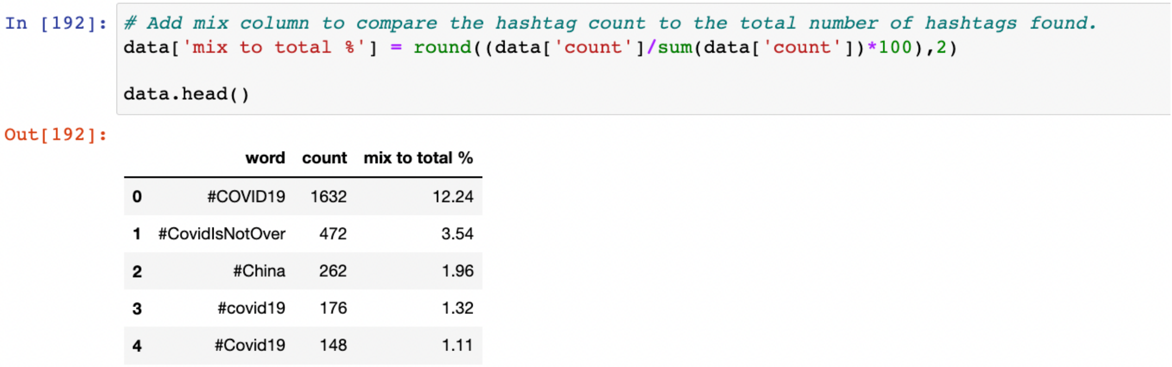
Twitter analysis:

I imported the twitter data and isolated the hashtags.



```
In [141]: # Create dataframe with only text
          tweets['text'] = tweets['text'].astype(str)
          tweets_text = tweets['text'].apply(lambda x: x if x.strip() != None else None)

In [142]: # Find hashtags
          tags = []
          for y in [x.split(' ') for x in tweets_text.values]:
              for z in y:
                  if '#' in z:
                      tags.append(z)

          # Create series showing count of each hashtag
          tags=pd.Series(tags).value_counts()

In [143]: # List top 30 hashtags
          tags.head(30)

Out[143]: #COVID19                  1632
          #CovidIsNotOver            472
          #China                     262
          #covid19                   176
          #Covid19                   148
          #COVID                     108
          #covid                     104
          #Greece                    103
          #coronavirus               100
          #PeoplesVaccine.            84
          #CoronaUpdate               84
          #Omicron                    83
          #COVID2020                  82
          #covid19uk                  80
          #CoronavirusOutbreak        80
          #COVID19Pandemic            80
          #monkeypox                  77
          #globalhealth               76
          #publichealth               72
          #healthtech                 69
          #COVID2019                  69
          #datascience                66
          #data                       66
          #analytics                  64
          #Shanghai                   63
          #Covid_19                   63
          #datavisualization          63
          #pandemic                   60
          #Athens                     55
          #Beijing                    50
          dtype: int64
```

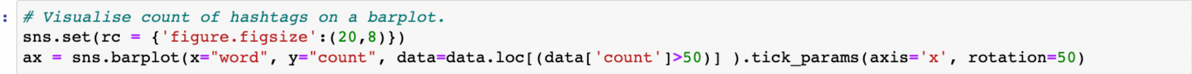
I calculated the hashtag mix to total.



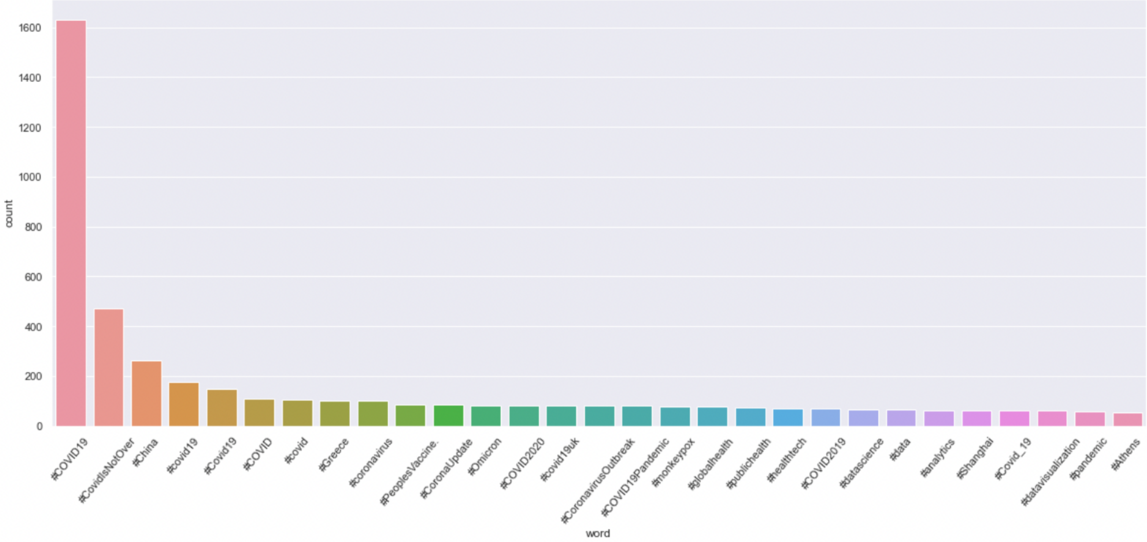
```
In [192]: # Add mix column to compare the hashtag count to the total number of hashtags found.
          data['mix to total %'] = round((data['count']/sum(data['count'])*100),2)

          data.head()

Out[192]:
```

	word	count	mix to total %
0	#COVID19	1632	12.24
1	#CovidIsNotOver	472	3.54
2	#China	262	1.96
3	#covid19	176	1.32
4	#Covid19	148	1.11

And then plotted all the hashtags used over 50 times.



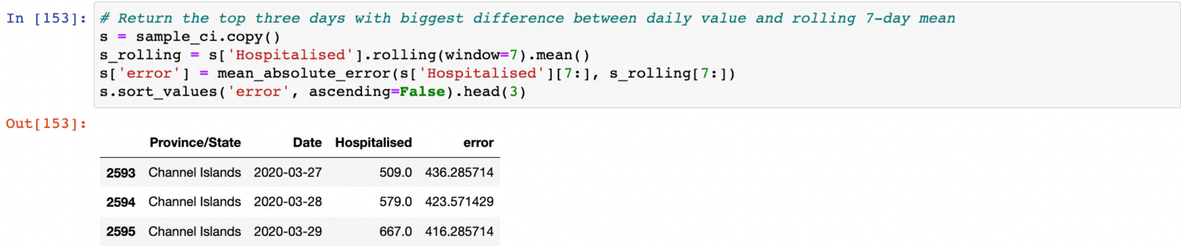
```
: # Visualise count of hashtags on a barplot.
  sns.set(rc = {'figure.figsize':(20,8)})
  ax = sns.barplot(x="word", y="count", data=data.loc[(data['count']>50)] ).tick_params(axis='x', rotation=50)
```



Twitter findings:
- #Covid19 is by far the most frequently used hashtag, accounting for 12.24% of tweets.

- The majority of hashtags (used more than 50 times) included the words covid, corona, pandemic, omicron, vaccine, and health. This is important to note for any further analysis on Twitter trends as these key words should be included in order to not exclude important and relevant data.

- It would be interesting to analyse the tweets further and understand the context around these hashtags to get an understanding of peoples' sentiment; are they concerned about contracting covid? are they fed up with covid rules? are they pro or anti vaccine? This can inform where to focus efforts to convince people to get vaccinated.



```
In [153]: # Return the top three days with biggest difference between daily value and rolling 7-day mean
          s = sample_ci.copy()
          s_rolling = s['Hospitalised'].rolling(window=7).mean()
          s['error'] = mean_absolute_error(s['Hospitalised'][7:], s_rolling[7:])
          s.sort_values('error', ascending=False).head(3)
```

Out[153]:

	Province/State	Date	Hospitalised	error
2593	Channel Islands	2020-03-27	509.0	436.285714
2594	Channel Islands	2020-03-28	579.0	423.571429
2595	Channel Islands	2020-03-29	667.0	416.285714

The code above shows the 3 days with the biggest difference between the 7 day moving average and the real value. This is useful to know in case you were to use the moving average forecasting technique, the forecast could be inaccurate by up 436 hospitalisations a day. Depending on what exactly the forecast will be used for, that might be an acceptable margin of error, or it might be too large and then you would know to try to use a different forecast technique to produce a more accurate forecast.

What is the difference between qualitative and quantitative data? How can these be used in business predictions?

- Qualitative data refers to categorical data such as a name, country or color, which are written as words or symbols. Instead Quantitative data refers to numerical data such as time, size or distance which are written as numbers.
- Historical quantitative data can be used for statistical analysis to identify patterns and trends to make predictions of future events, sales, traffic etc. so that businesses can prepare accordingly.
- Qualitative data can be used in forecasting when there is not a lot of historical data available to perform statistical analysis on. This includes gathering opinions from either experts or potential customers to gauge future performance of a product or service.

Can you provide you observations around why continuous improvement is required, can we not just implement the project and move on to other pressing matters?

- The process of continuous improvement allows you to review the project, taking in any feedback from stakeholders and then work on solving any issues identified, delivering a new and improved version of the project. Without a process of continuous improvement issues or requests for additional information or features can go unanswered as there are no resources dedicated to monitoring and improving the project.

As a government, we adhere to all data protection requirements and have good governance in place. Does that mean we can ignore data ethics? We only work with aggregated data and therefore will not expose any personal details? (Provide an example of how data ethics could apply to this case; two or three sentences max)

- Other than just complying with legal requirements, it is recommended that any organisation that works with data implements a data ethics framework which helps everyone within the organisation that works with data comply with data ethics principles set out by the organisation. Although the data you work with does not contain personal details, data ethics also applies to the ways that data is collected, analysed and shared.