Assignment Course2
Written Report
Alan Ireland

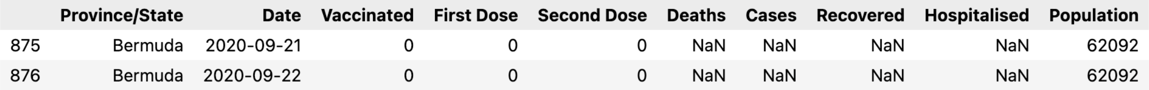
Approach:

-Import, clean, wrangle the data.
-Analyse and visualise the data.
-Highlight key findings from data.
-Answer assignment question and make recommendation: Where should UK govt target first for it's marketing campaign?

Import, clean, wrangle the data.

After importing the data there wasn't much cleaning to be done. Only two rows in the cases data were partially empty and I decided to keep this in the data set because they did contain data for vaccinations (below). If this caused problems later I could strip them out. During cleaning I checked the shape and type of the data set, merged the data sets, and confirmed any dates had the correct dtype. I removed the province 'Others' since we have no information on this and are not told where it is from.
As I analysed the data I also used the max absolute scaling method of normalisation, grouped data into months, used descriptive statistics, applied my own user defined functions. See notebook for specifics.

null_values (FourthRev data):

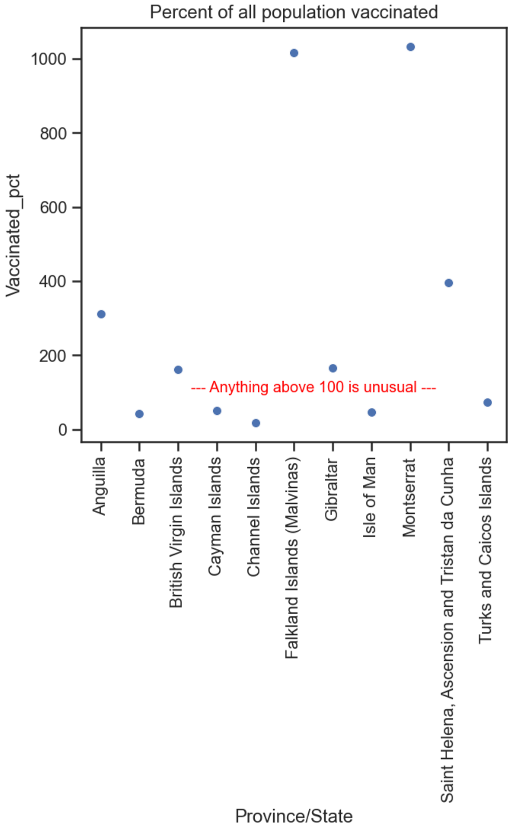


	Province/State	Date	Vaccinated	First Dose	Second Dose	Deaths	Cases	Recovered	Hospitalised	Population
875	Bermuda	2020-09-21	0	0	0	NaN	NaN	NaN	NaN	62092
876	Bermuda	2020-09-22	0	0	0	NaN	NaN	NaN	NaN	62092

-Analyse and visualise the data

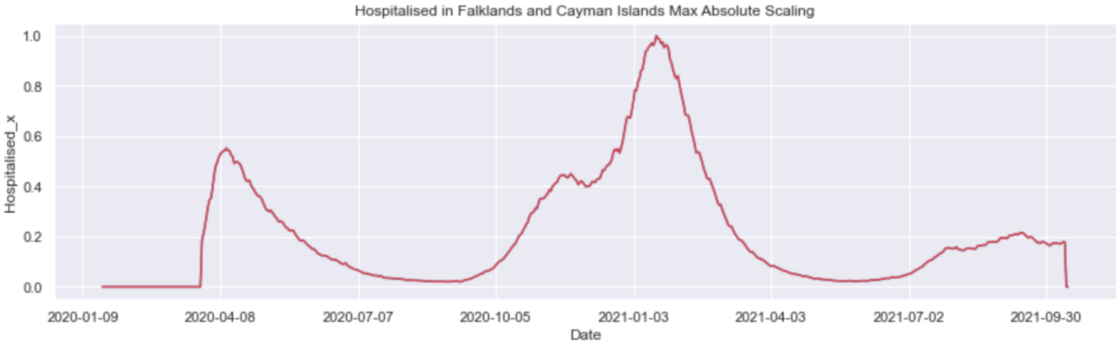
Simple initial inspection of the data quickly revealed serious flaws in the data. The vaccinated.csv data has been made up. The numbers are way too large.

Percent of all population vaccinated (FourthRev data):



The cases.csv data has also been made up. The normalised cases data is identical despite provinces being vastly different in location and population. This is completely implausible. The chart below shows two different data sets but the lines are on top of one another. Despite the provinces being ~8,200km apart and Cayman Islands has a population nearly 20x that of Falklands.

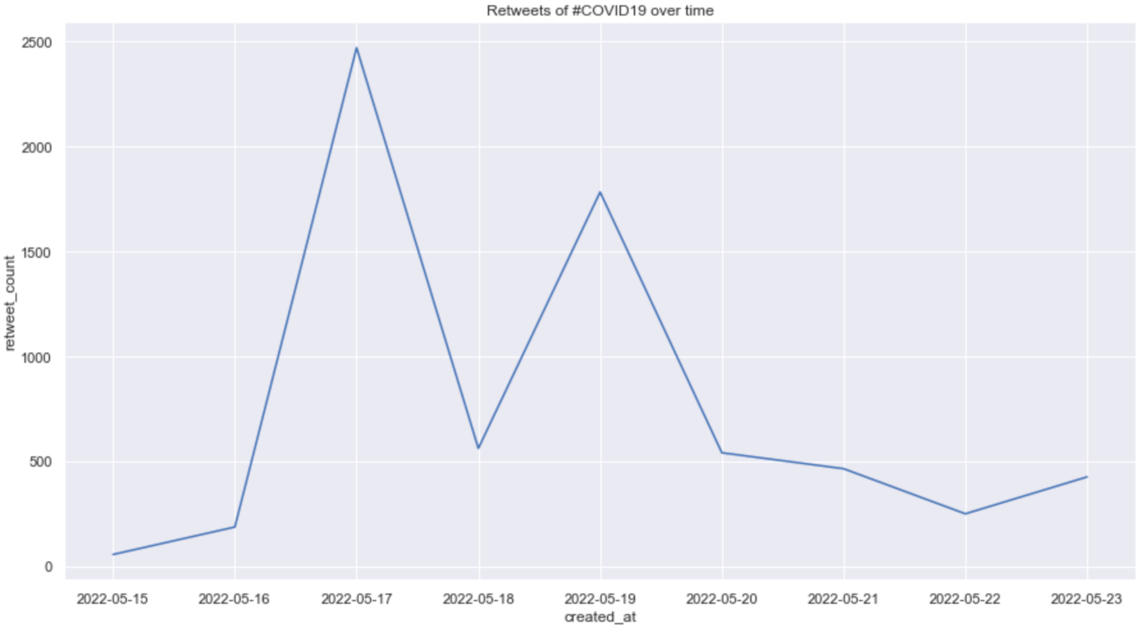
Hospitalisations normalised in Falklands and Cayman Islands time series (FourthRev data):



For additional reasoning and comments and a statistical check, see notebook.

I undertook some rudimentary analysis of the twitter data set to search for patterns and themes.

There is much more I would to have investigated with more experience and skills, see end of notebook.



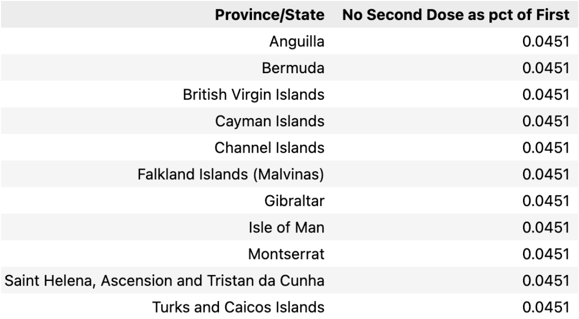
-Highlight key findings from data & Answer assignment question.
Where should UK govt target first for it's marketing campaign?
Recommendation: Do not focus the marketing campaign on any one area since they all have the same need for second doses.

The data cannot be compared between provinces in absolute terms because absolute data must be put in context of population. Even if the data wasn't laughable, we couldn't conduct comparisons across provinces in absolute terms because this biases the data set to the large provinces. i.e. how could the Falkland Islands produce a meaningful signal from absolute data when it's less than 2% the size of The Channel Islands?

When we compare data between provinces on a relative basis, the only sensible relative comparison we can produce is the percentage of first dose recipients that need a second dose. The data shows all provinces have exactly the same 4.51% of first dosers needing a second dose.

Thus, using the data and guidelines provided with this assignment, and accepting the data as is, the inescapable conclusion is there is no particular province the UK govt should focus on for they are all the same need.

Percent of first dose recipients in need of second dose by Province (FourthRev data):



Province/State	No Second Dose as pct of First
Anguilla	0.0451
Bermuda	0.0451
British Virgin Islands	0.0451
Cayman Islands	0.0451
Channel Islands	0.0451
Falkland Islands (Malvinas)	0.0451
Gibraltar	0.0451
Isle of Man	0.0451
Montserrat	0.0451
Saint Helena, Ascension and Tristan da Cunha	0.0451
Turks and Caicos Islands	0.0451

It is a shame the data has been created badly like this because the real COVID data will be much more interesting. I think there are also wider ethical considerations about using and producing knowingly fake data on a real topic as sensitive as COVID but I accept that is a question of personal taste / core values.
People died for refusing a vaccine because they believed in conspiracy theories not based on data. It seems strange effort has been expended generating fake data when the real data published by governments, including at daily press conferences throughout the pandemic, is widely available. It also reminds me of an earlier warning from the course notes “garbage in – garbage out”.

Therefore so ends the work I will do using the FourthRev fake data set. I have compiled monthly data and visualised as a chart in line with the assignment requests. I do not feel like there is any insight in this work because I have shown the data is fake and the absolute numbers are absurd.

Report Word Count=705

What follows in the rest of this report is me following my own curiosity to scrape the real COVID data from the web; to analyse that; to try identify some demographic trends; to try reproduce some of the assignment tasks if time permits.

Approach:
-Import, clean, wrangle the real data.
-Analyse and visualise the real data.
-Highlight key findings from the real data including, where possible, completing comparable assignment requests using real data.
-Compare the two data sets and highlight differences/findings.

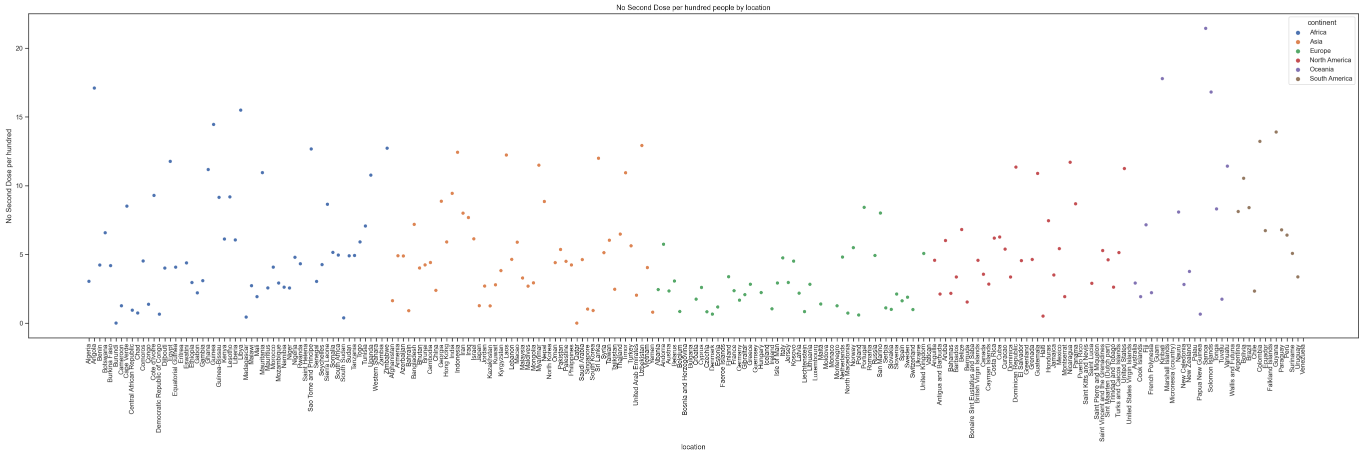
-Import, clean, wrangle real world data
I imported global COVID data from Our World In Data website which I have used before for my job and know is a reliable source. I reduced the dataset to only those fields I am interested in. I kept all countries in the data set because I was curious what I would find.
I also imported GDP per capita by country data from the WorldBank website to merge to the real data set.

A couple of issues were identified during cleaning. First, that some countries update their data less frequently than others. Second, that the cumulative data is only written to the dataset when it changes, else blank, so I may need some kind of ffill or time based mean.
In addition, some assignment provinces are unavailable. The ‘Channel Islands’ is not present in the real data but it’s constituent members, Guernsey and Jersey, are so those are included. Montserrat and Saint Helena produced no data at all so it is not present in my data set. The Falklands data is included but limited. Interestingly the real data is missing in the smallest of provinces which intuitively makes sense.

-Analyse and visualise the data

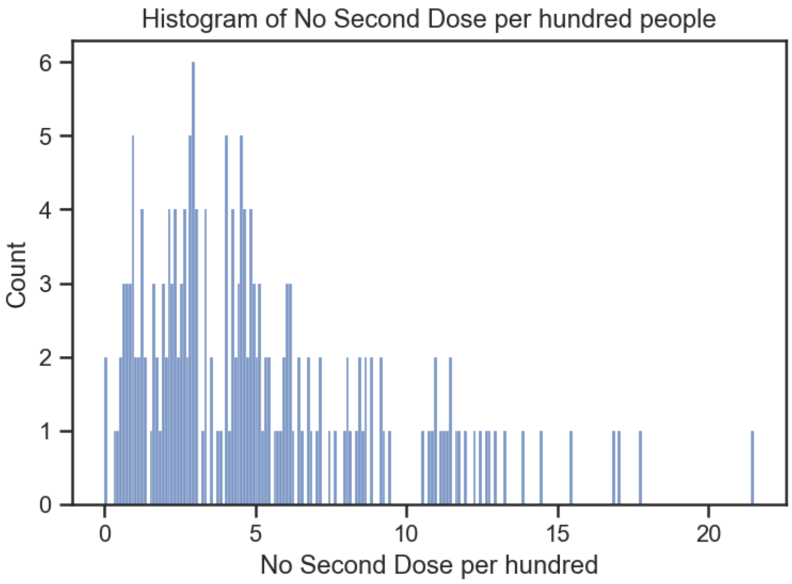
I applied the assignment question to the worldwide real COVID data. To get a quick sense of the data I charted the percentage of the population that had a first dose but no second dose. This is what the original assignment asked for. I charted this for all countries to get a quick skim of the data and see what the distribution is like, coloured by continent to see if there are any major divergences (below).

Note, this is not intended to be readable, it's a quick viz of everything in the dataset.

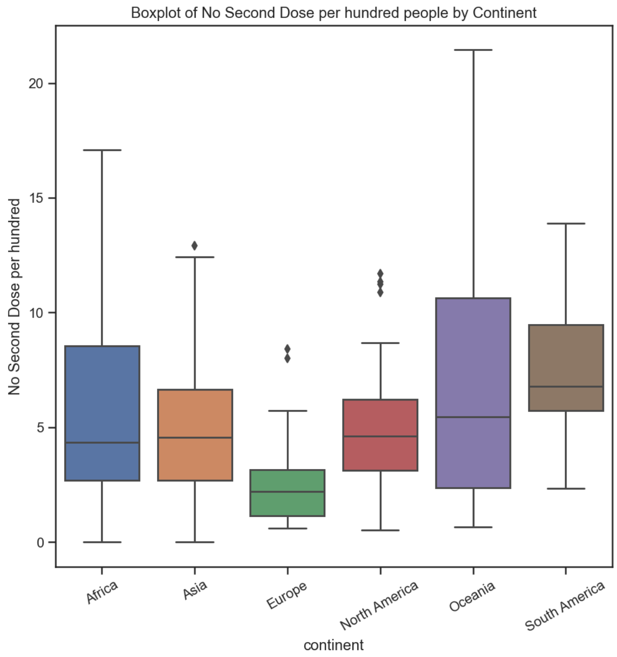


Immediately the distribution of real world COVID data looks more interesting than the FourthRev data.

I examined the overall distribution of those needing a second dose as percent of population using a histogram:

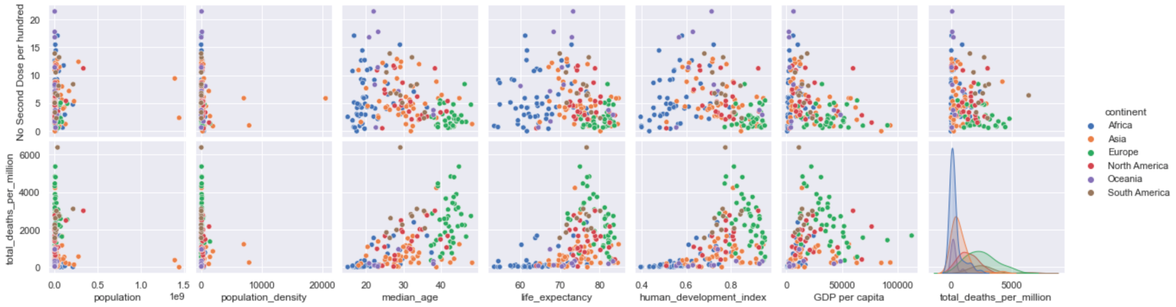


And also drilled this down by continent as a box plot:



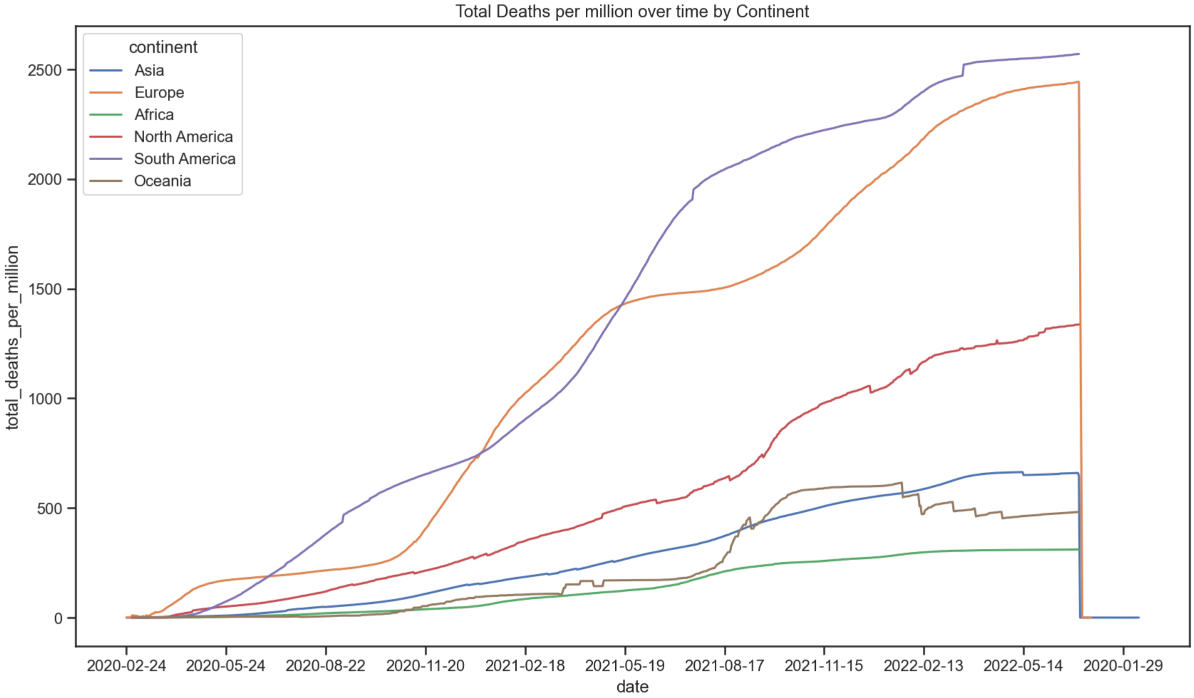
I generated a pair plot of those needing a second dose, number of deaths, against key demographic variables including GDP per capita and life expectancy (below).

My observations are included in the notebook.

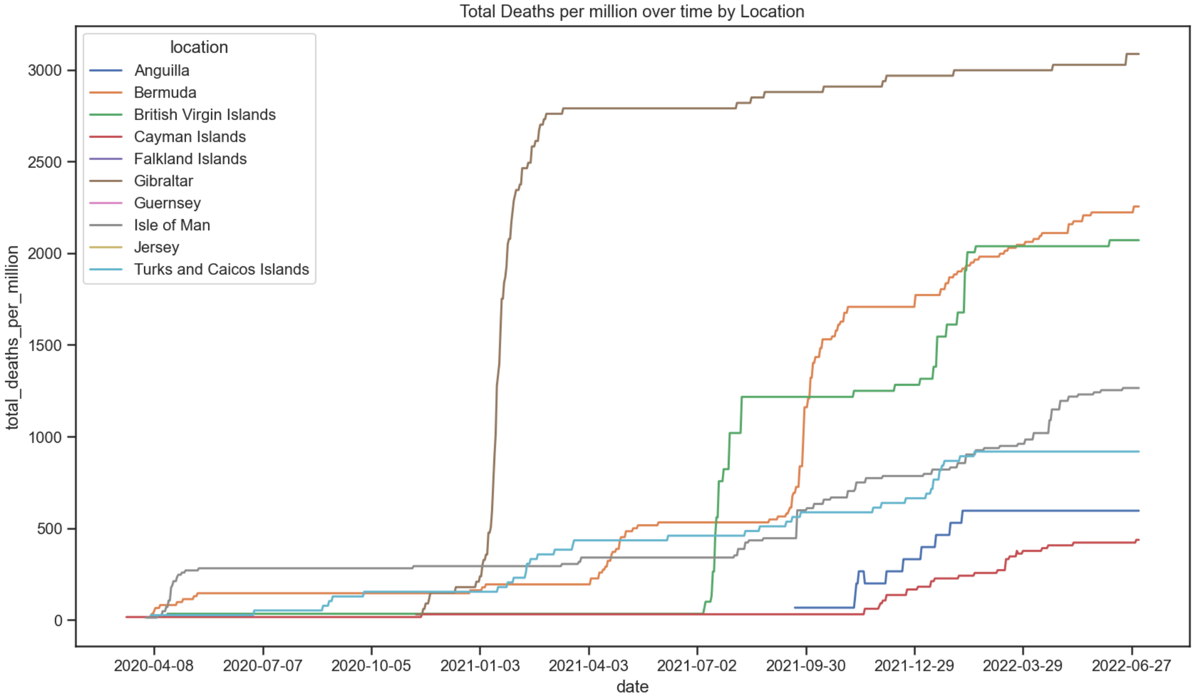


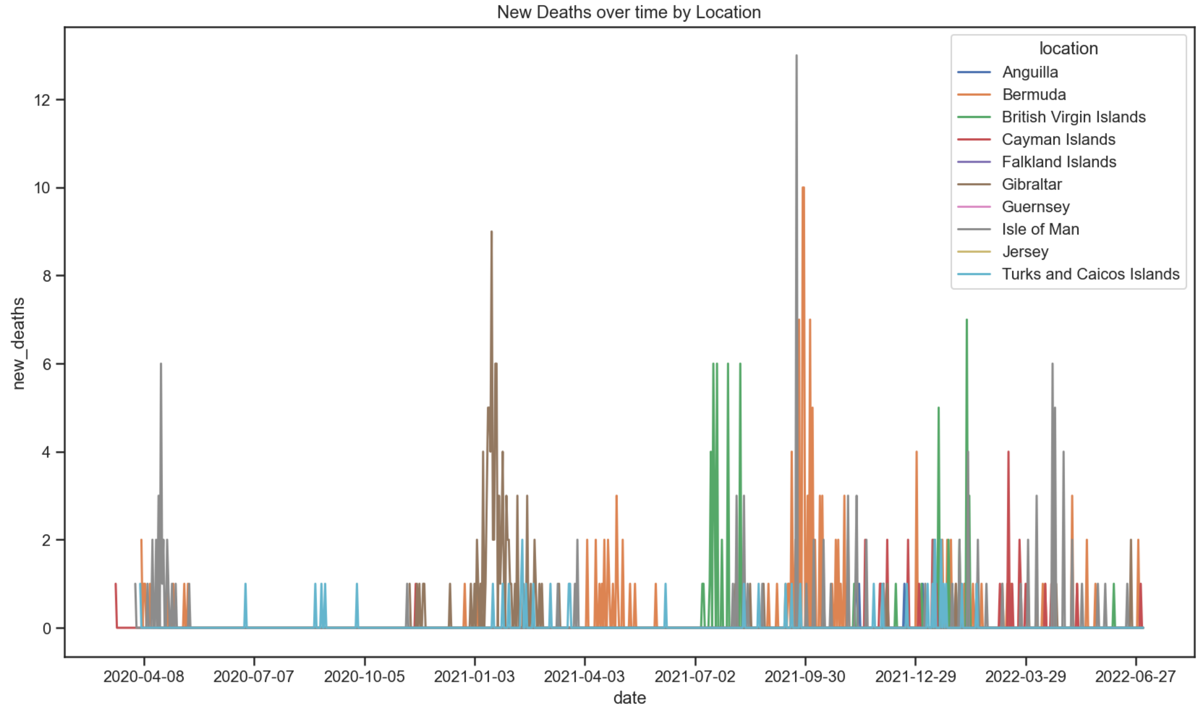
To match assignment tasks I created the following charts.

Total deaths per million as a time series by continent:



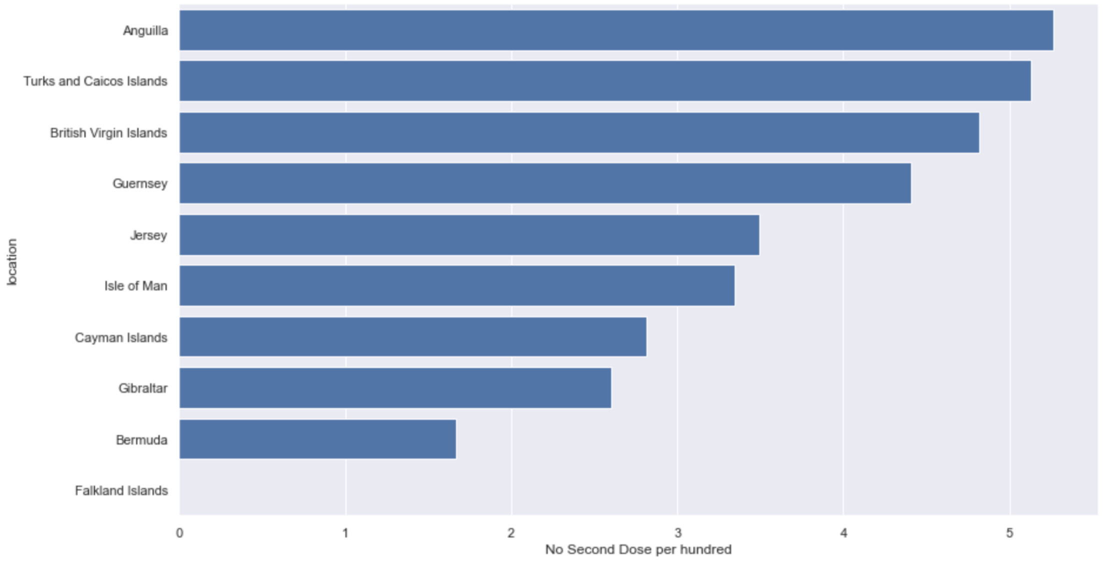
Total deaths per million as a time series by assignment province (per notes, those provinces where real world data available):



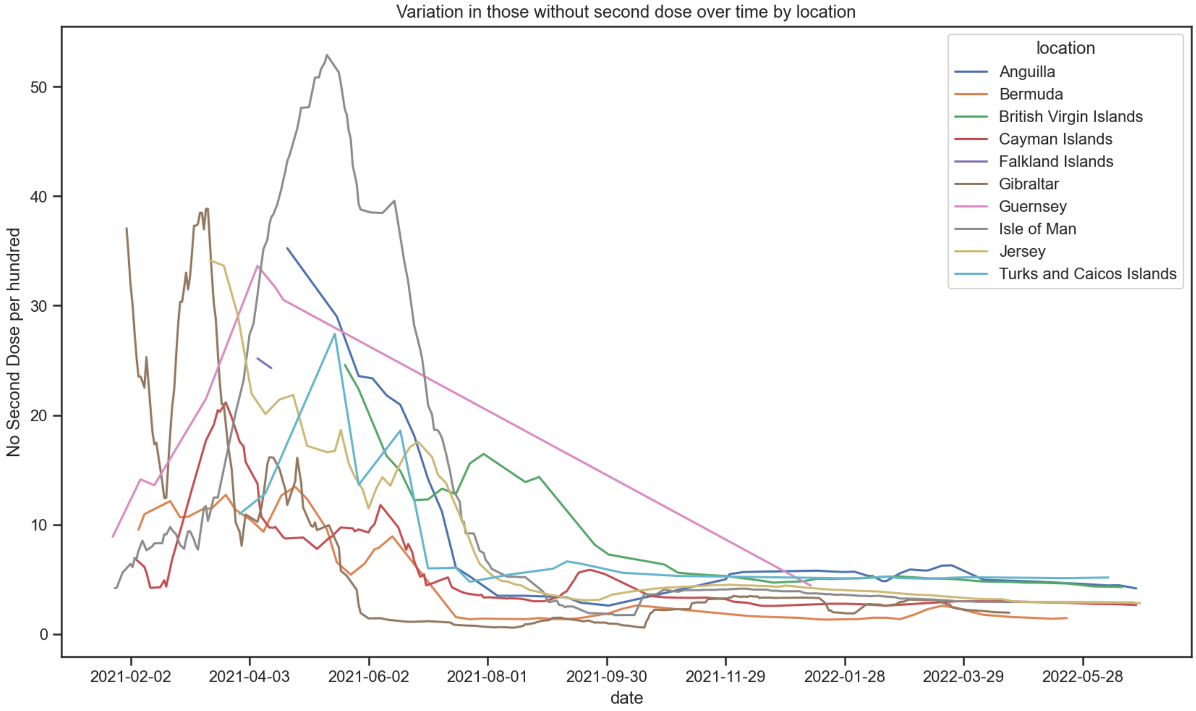


All have clearly identifiable differences/patterns, see notebook.

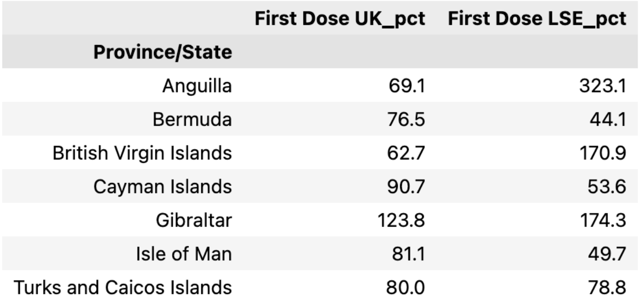
This bar chart using real data shows where the UK govt should first focus it's second dose marketing campaign.
Real world recommendation: UK govt should focus it's marketing efforts on Anguilla, Turks&Caicos, BVI, and Geurnsey.



And this time series shows how the percentage of those needing a second dose has varied over time.



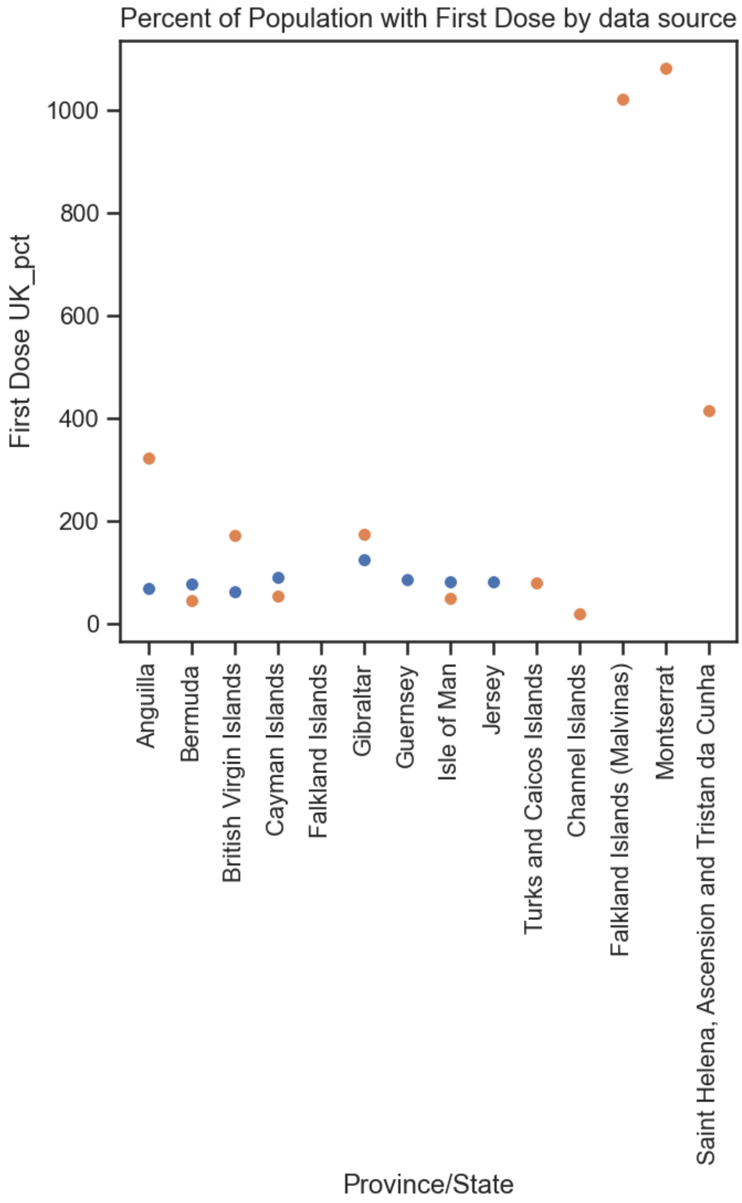
I compared the FourthRev fake data to the real world data using percent of population with first dose. This was the most sensible metric to use because ‘total doses’ in the real world = first dose + second dose + boosters and gets big.



	First Dose UK_pct	First Dose LSE_pct
Province/State		
Anguilla	69.1	323.1
Bermuda	76.5	44.1
British Virgin Islands	62.7	170.9
Cayman Islands	90.7	53.6
Gibraltar	123.8	174.3
Isle of Man	81.1	49.7
Turks and Caicos Islands	80.0	78.8

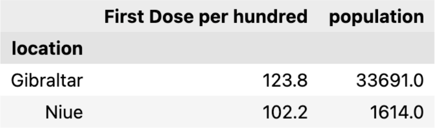
As a scatterplot visualising the differences:

Yellow = FourthRev fake data
Blue = real world data



Lastly I wanted to know if there were locations where more than 100% of the population had received a first dose in the real data.

The answer is yes and those countries are:



	First Dose per hundred	population
location		
Gibraltar	123.8	33691.0
Niue	102.2	1614.0

I contemplated why this might be the case and offered an explanation by highlighting three relevant population subgroups in Gibraltar.

Thus, in the end, the assignment came full circle. It turned out that of only two places in the world that have unusual COVID vaccine data, one of them was used for our assignment!