LSE Data Analytics Career Accelerator

Course-2

Assignment

Assignment Cover Sheet

Submitted by: Atif Habib Syed

Date Submitted: 11-July-2022

Module Title: Course-2

Date/Year of Module: 2022

Word Count:

Number of Pages: **12**

Question: *You are a data analyst working with the UK government to analyse COVID-19 data (from January 2020 to October 2021). As part of its goal to increase the number of fully vaccinated individuals (people who have received a first and second dose of the vaccine), the government is planning to launch a series of marketing campaigns to promote the vaccine. The government wants to identify trends and patterns that can be used to inform its marketing approach to increase the number of fully vaccinated people.*

Declaration: *"I declare that this work is entirely my own in accordance with the academia's guidelines on plagiarism and collusion. All external references and sources are clearly acknowledged and identified within the contents.*

No substantial part(s) of the work submitted here has also been submitted by me in other assessments for accredited courses of study, and I acknowledge that if this has been done it may result in me being reported for self-plagiarism and an appropriate reduction in marks may be made when marking this piece of work."

Table of Contents

Background

A divine retribution, man-made virus or a mere twist of global fate, whatever Sars-Cov-2 a.k.a. Covid-19 is, it wreaked havoc on a scale that one is unable to find in living memory. World was halted at its feet where busy intersections like Shibuya (Japan), Times Square (New York, USA) or SOHO (London) presented a picture of a Zombie town. That said, thanks to researchers and pharmaceutical companies, Vaccines were produced on a scale that would aid global life, as we knew it, to be restored to somewhat pre-pandemic levels.

This report/ assignment is focused on Covid-19 related data including its emergence and related effects such as deaths and hospitalizations and vaccination drives.

The Data

Data is focused on 12 territories/ Provinces of the UK. Time period covered in the dataset is from Jan 2020 to October 2021 (22 Months). Dataset provides details of identified case, deaths, hospitalizations, recoveries along with vaccination status for the population of each Province for those 22 months. We will explore this dataset in upcoming sections and present our analysis.

The Analysis Tool

We will be using PYTHON as out data analysis tool. The Pythonic libraries that will be deployed for this data analysis are as following:

1. NumPy for mathematical functions
2. Pandas for Data Framing
3. Seaborn for Visualization
4. Matplotlib for Visualization

The Assumptions

Following are the initial assumptions that have been made about the dataset:

1. Daily numbers are not cumulative
2. ‘Others’ Province(s) are UK mainland provinces
3. Numbers provided in the dataset are correct and have not been counter checked by any external source

1.0 Data Exploration

The Dataset contains relevant and some irrelevant (for this analysis) information. Irrelevant information includes geographical location data of the provinces along with international regional codes and names. As this information is not required for our analysis, we will discard/drop this information during our analysis in Python.

The data doesn't have a default index, however, during initial data loading from the CSV files, Python assigned index to data frame.

An initial look at the data reveals that in January 2020, there were no identified case in the provinces and hence there was no need for a vaccine, which obviously wasn't discovered at the time either. However, if we look at the records in October 2021, it is visible that daily number have increased to gigantic proportions. Specifically, vaccination numbers have increased significantly. This indicates that UK government's vaccination drive is on the rise in all provinces. In addition, that may also suggest that the vaccine supplies are increasing rapidly.

There are eight values missing in 'COVID' (Cov) Data Frame. These values belong to Index number 875 and 876, Province: Bermuda and columns Deaths, Cases, Recovered and Hospitalized. Dates of the missing Data are 21st of September, 2021and 22nd of September, 2022 (which is obviously a Typo)

1.1 Filtering and Sub-setting Gibraltar

To sample verify the data, Gibraltar was filtered from both Data Frames (Cov & Vac). We will start by filtering Gibraltar from both data frames and dropping unnecessary columns such as geographic location information etc.

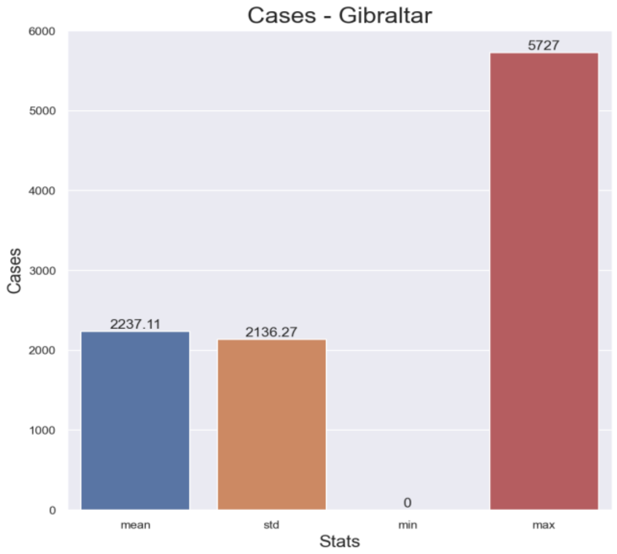
'Cov' Dataframe

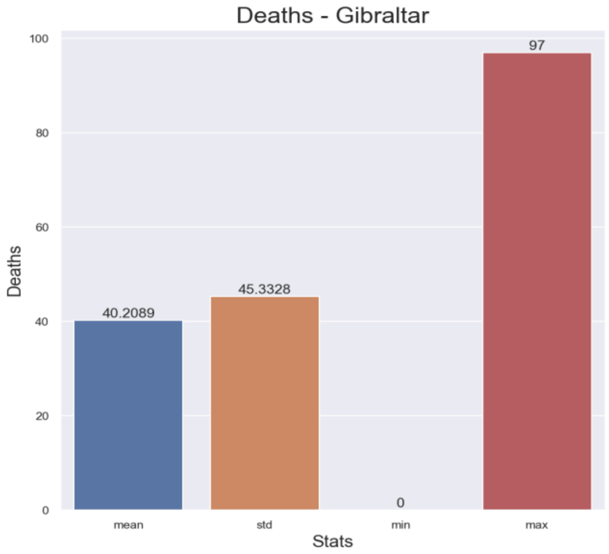
Looking at the Gibraltar's 'Cov' data frame, it appears that first cases of Covid-19 were identified in early March 2022 and the number gradually increases over the period of 20 Months (October 2021).

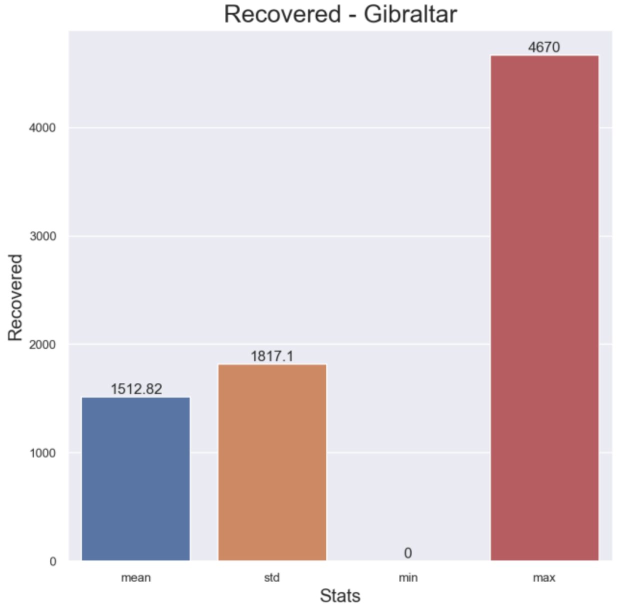
The number of 'Cases' don't seem to reach a peak as they are still on the rise till 14th of October 2021 (last date when data was collected). Same is the case with 'Deaths' and 'Recovered', however, data for 'Recovered' is not seemed to be collected since early August 2021 till October 2021.

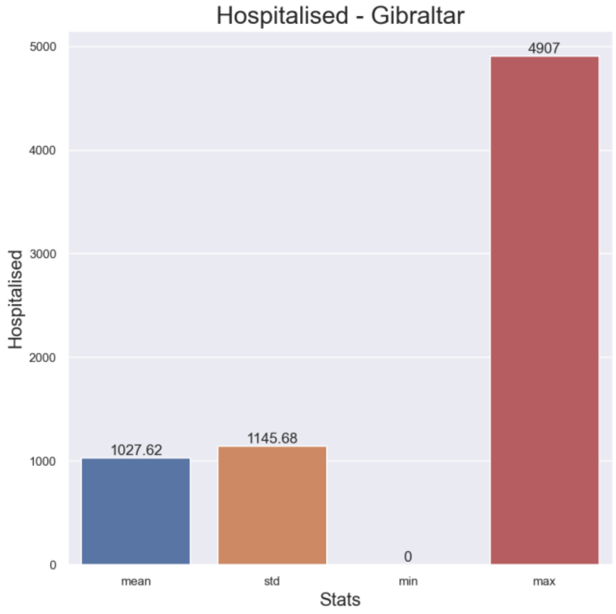
One strange phenomenon seems to occur if we look at identified 'Cases' and 'Hospitalized' numbers. Number of people who are 'Hospitalized' are way greater than number of identified 'Cases'. There can be a plausible explanation for this, that is due to confusing/unknown/unconfirmed symptoms of the virus, all the people who were ‘Hospitalized’ at the time did not have confirmed Covid-19 virus, however, were “considered” to have Covid-19 to rule out the possibility.

A brief look at the statistics (mean, standard deviation, minimum and maximum values) for all four categories (Cases, Deaths, Recovered and Hospitalized) provide a brief insight into Gibraltar’s dataset.





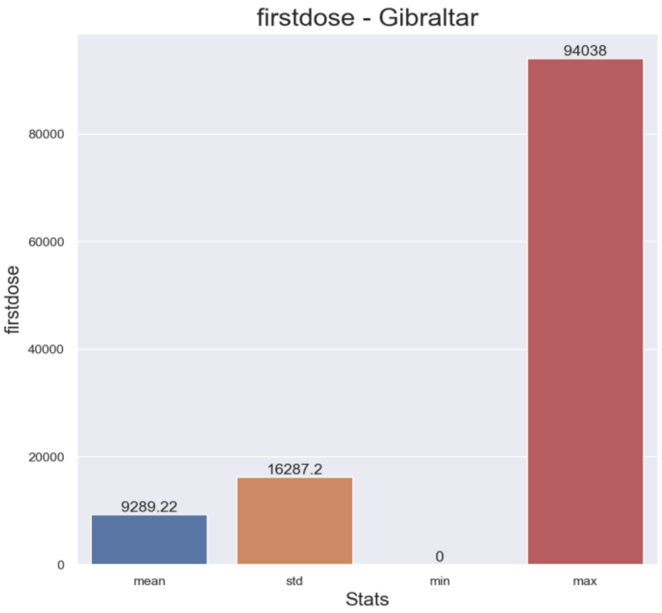


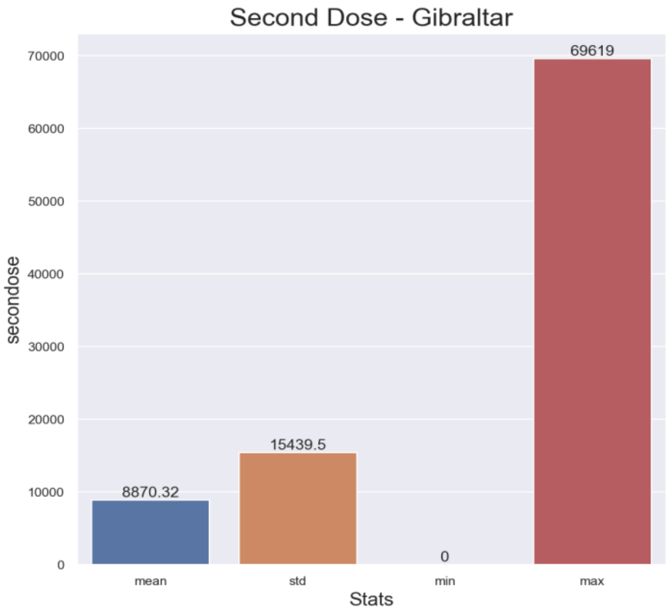


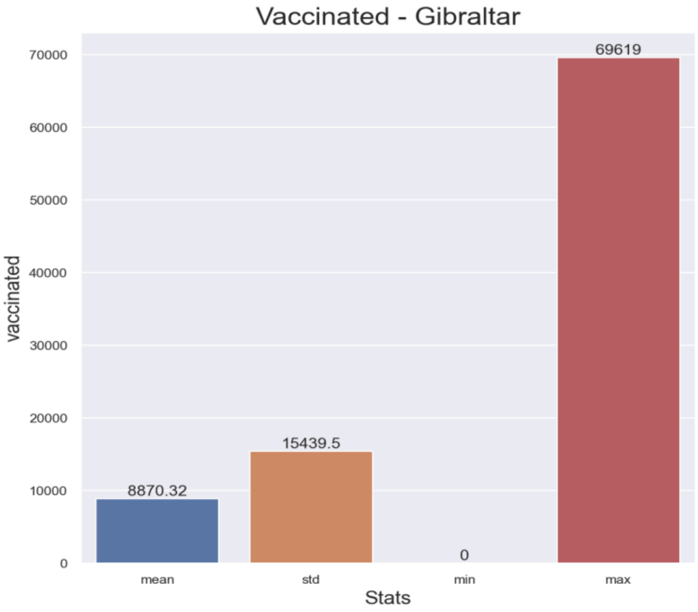
‘Vac’ Dataframe

Analyzing Gibraltar’s ‘Vac’ dataframe, it can be observed that data abruptly appears in early January 2021. A brief look at the data shows that daily vaccination numbers start to decline in September 2021, an assumption is made here that this is because majority of Gibraltar’s population has been vaccinated by September 2021.

A brief look at Gibraltar’s visualizations will assist in understanding the data:



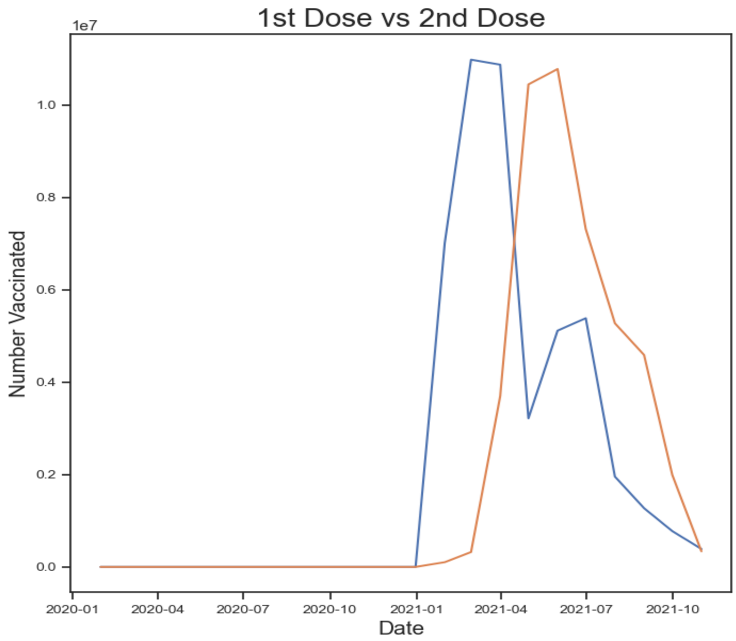


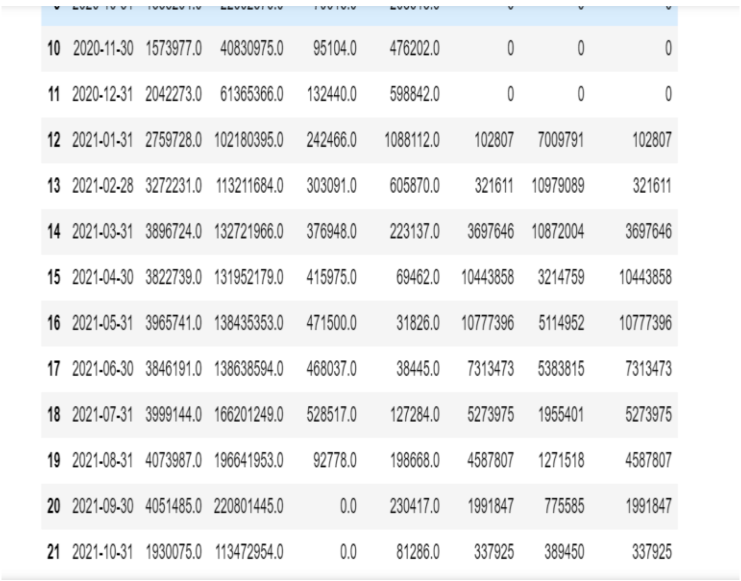


2.0 Merging and Analyzing Data

When both data frames are combined, it opens multitude of opportunities for analysis. In this activity, we will explore vaccination status (first and Second doses) over time and Province/State wise.

2.1 Vaccination Status over time





[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0	0
10	2020-11-30	1573977.0	40830975.0	95104.0	476202.0	0	0	0
11	2020-12-31	2042273.0	61365366.0	132440.0	598842.0	0	0	0
12	2021-01-31	2759728.0	102180395.0	242466.0	1088112.0	102807	7009791	102807
13	2021-02-28	3272231.0	113211684.0	303091.0	605870.0	321611	10979089	321611
14	2021-03-31	3896724.0	132721966.0	376948.0	223137.0	3697646	10872004	3697646
15	2021-04-30	3822739.0	131952179.0	415975.0	69462.0	10443858	3214759	10443858
16	2021-05-31	3965741.0	138435353.0	471500.0	31826.0	10777396	5114952	10777396
17	2021-06-30	3846191.0	138638594.0	468037.0	38445.0	7313473	5383815	7313473
18	2021-07-31	3999144.0	166201249.0	528517.0	127284.0	5273975	1955401	5273975
19	2021-08-31	4073987.0	196641953.0	92778.0	198668.0	4587807	1271518	4587807
20	2021-09-30	4051485.0	220801445.0	0.0	230417.0	1991847	775585	1991847
21	2021-10-31	1930075.0	113472954.0	0.0	81286.0	337925	389450	337925

We can see that vaccination drive in the UK started in January 2021, where number of first doses peaked in March/ April 2021 reaching to 10 million. Whereas number of second doses gained momentum around April 2021 reaching its peak of around 10 million in May 2021. There has been a slight rise in First doses around July 2021 but it subsided accordingly.

2.2 Vaccination status in Provinces

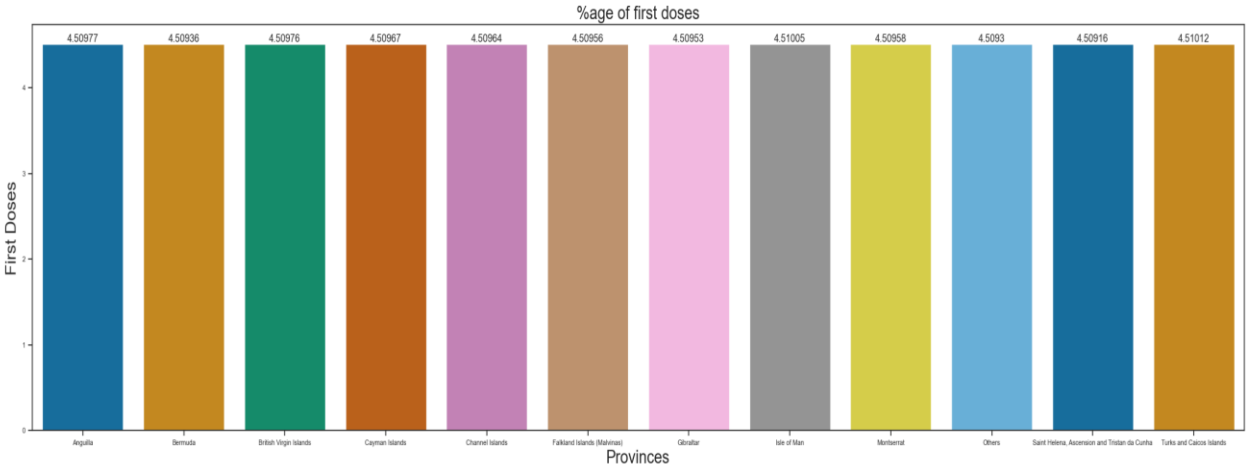
Highest number individuals that received first doses but not the second dose is in Gibraltar with 222,398 of individuals still waiting for their second dose. Looking at the percentages of individuals who received their first dose but not the second one, surprisingly, all provinces have a similar percentage of 4.5%. Considering the fact that data is correct, this can be accorded to UK government’s aggressive and synchronized vaccination drive which aimed at vaccinating maximum number of people in all its provinces.

3.0 Visualise and Identify Initial Trends

As UK’s government intend to identify and, consequently, target the areas for their vaccination drive, we need to identify trends such as deaths, recoveries, and vaccinations across regions and over time through data grouping.

3.1 Second Dose Eligibility

To assist the government to test their campaign to increase second dose in more than one area, we first need to find number of first and second doses that have already been administered to the population.

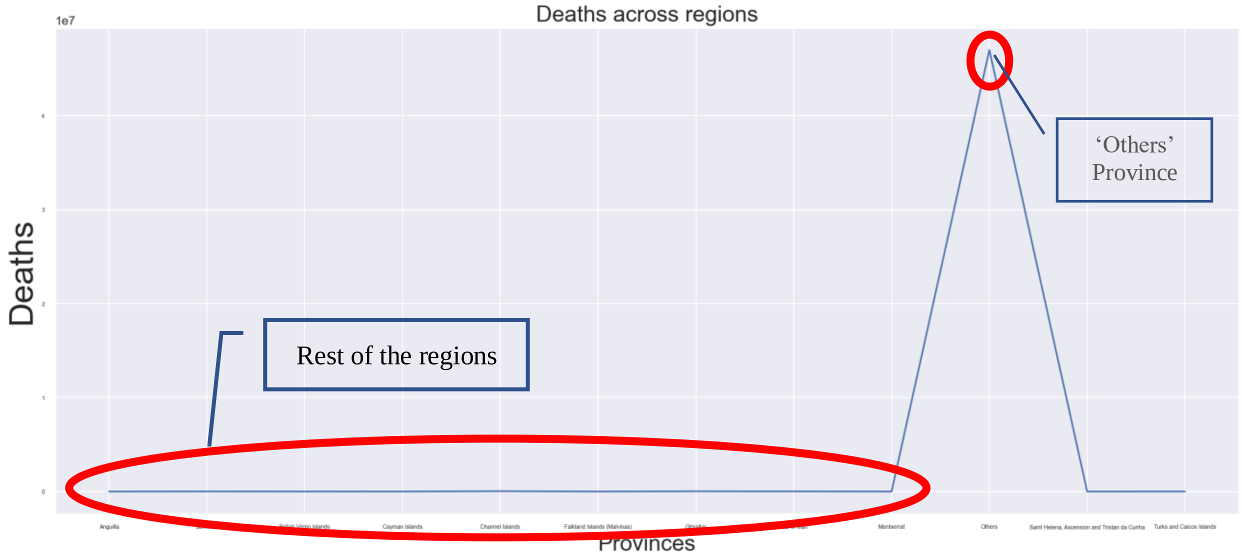


Looking at the data and visualization, it is interesting to note that percentage of population, across all regions, that is eligible for second dose, is same around 4.5%. Considering the size of the population of each region, this discovery can be accounted to UK’s government’s synchronized vaccination drive that was able to vaccinate 95% of the population within a period of 10 months.

3.2 Groups responsible for skewness

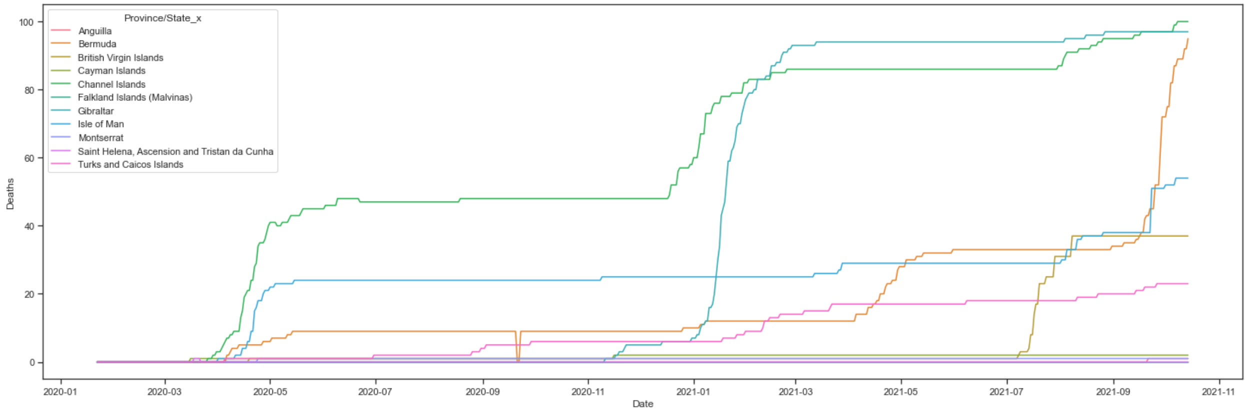
Looking at number of deaths across regions, ‘Others’ province shows a drastic deviation from rest of the regions. Even if it is assumed that ‘Others’ province include England, Scotland, Wales and Northern Ireland, number of deaths is approximately touching 46 million, which is 70% of

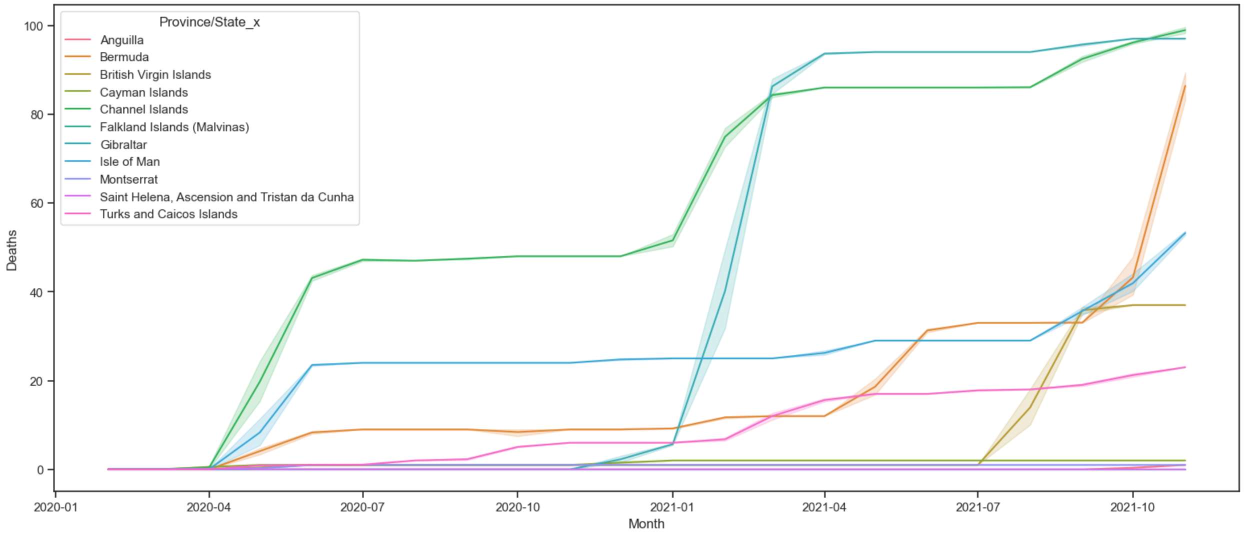
UK mainland's total population. Therefore, it can safely be assumed that deaths data from 'Others' province is erroneous and should not be considered while performing our analysis.



3.3 Affect of converting Dates (Days) into Months

Lineplots became much smoother after converting dates (days) into months, following figures will show the difference before and after the conversion:

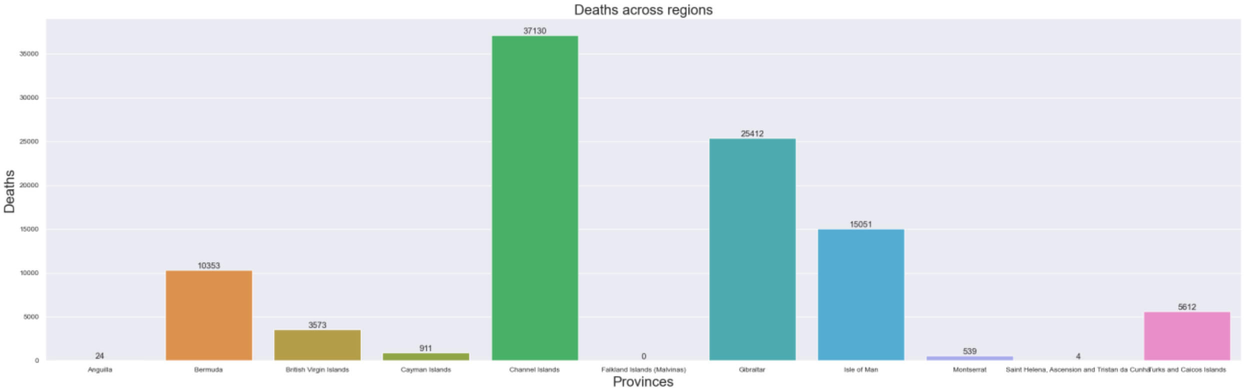




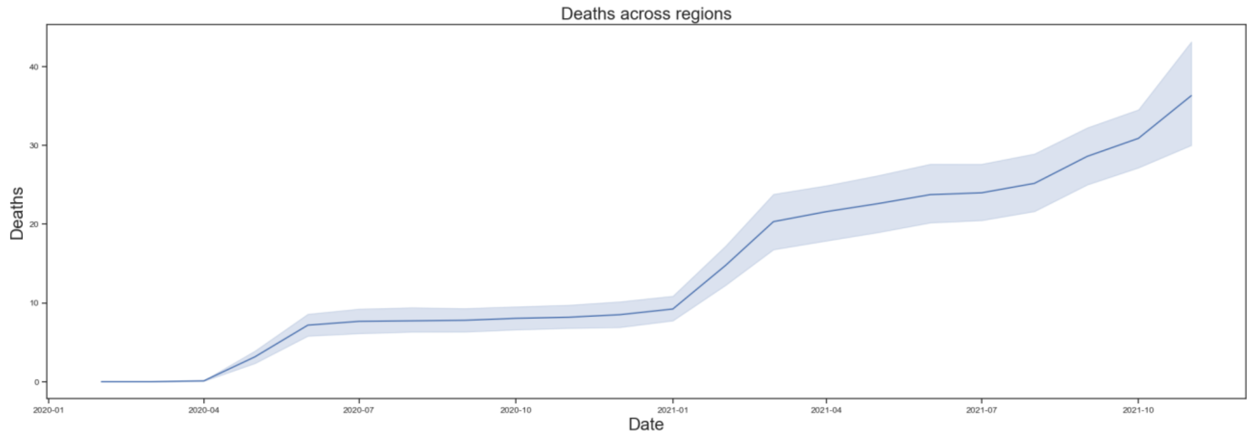
3.3 Number of Deaths across regions and trend

Number of deaths in some regions have either reached their peak or plateaued, whereas in others such as Bermuda, Channel Islands, Turks and Caicos Islands, Isle of man and Anguilla, death rate is still on the rise. Looking at Bermuda, death rate is steeply rising till last data was collected.

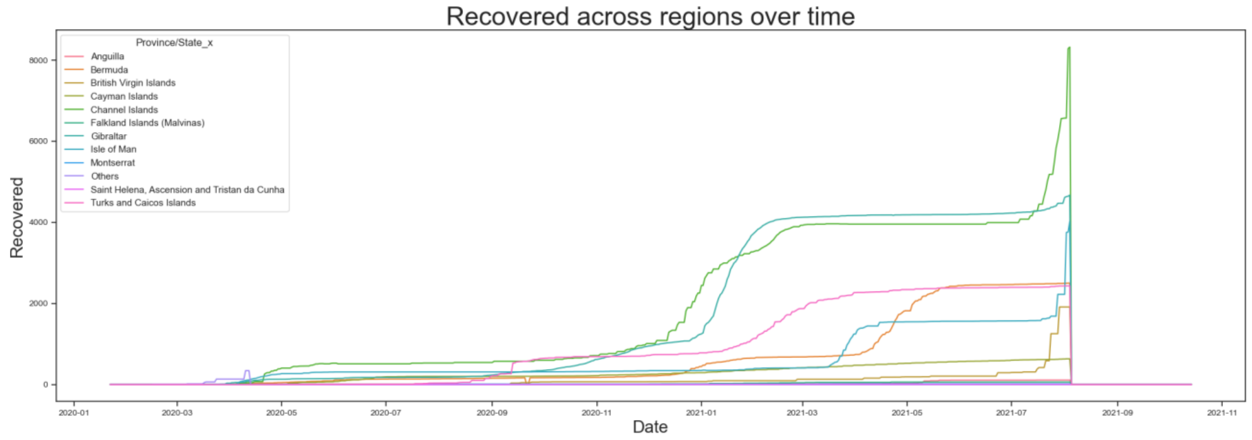
Looking at the total number of deaths since the start of the pandemic, Channel Islands suffered the greatest number of deaths (37,130) followed by Gibraltar (25,412) and Isle of Man (15,051)



If we look at total number of deaths over time (Deaths vs Time), it is visible that number of deaths is still increasing till the time of data collection.



3.4 Recoveries in regions



The region with most recoveries, over time and in total, is Channel Islands. However, this hasn't been consistent over time as there is a somewhat gradual rise around January 2021 which can be credited to vaccinations. However, there is a steep rise in recoveries around July/ August.

As there is no data collected after August 2021, so there is a steep fall in all regions to zero in September 2021.

3.5 Helping Government making decisions

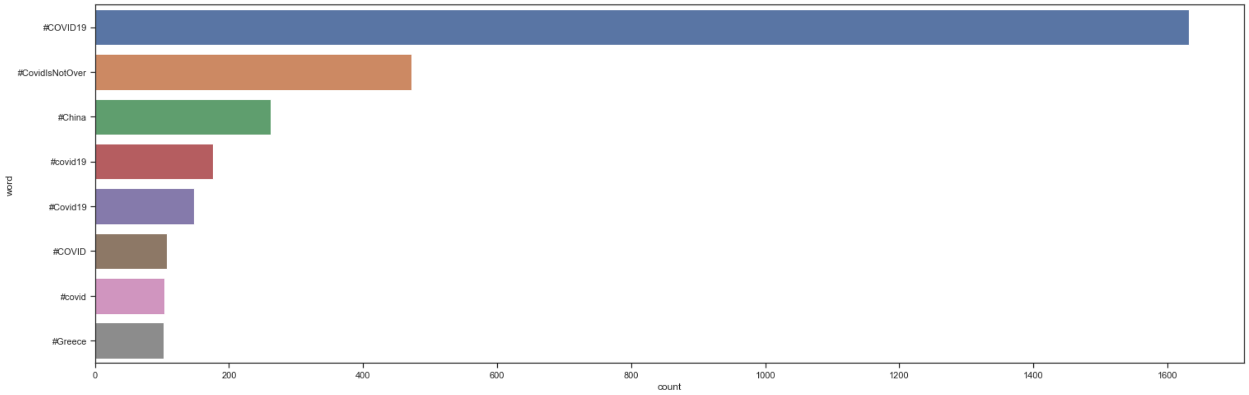
UK Government's aim is to target regions where number of deaths are still on the rise and number of recoveries is not improving. Taking specific examples, we can see that number of deaths in Channel Islands, Isle of Man and Gibraltar are the highest in total, however, Bermuda death toll rises steeply from July to October.

In case of recoveries, Channel Islands and Isle of Man, have the greatest recovery rate over time and in total, which is an indicator UK Government to initiate their campaign from areas where recovery rates are slow.

4.0 Analyzing Twitter Data

Social Media, nowadays, is a gold mine for organizations or institutions to understand general public’s sentiment towards them or a phenomenon that is being focused.

In our case, we will focus on real time data on twitter related to Covid-19 and vaccination status. After pulling the data from twitter, it is apparent that top trending topic is Covid-19, followed closely by ‘Covid is not over’ and other variants of ‘Covid’. This can assist UK government to correlate their internal (official) data and discover the differences or discrepancies between both data.



To make external/ Social Media data analysis more objective in our case, it would be more precise to find trending topics region by region. The reason for this is that trending topics for each region are different. In case, region is selected as the UK, we might receive top trends for mainland UK but not for other territories. Therefore, in order to collect accurate social media data, it would be appropriate to explore region wise data.