Data Analysis of COVID-19 in the UK

Objective:

The UK government wants to boast its efforts to increase vaccination rates through marketing campaigns to promote the COVID vaccinations. This analysis uses Python to analyse and identify trends and patterns from the data sets provided in order to help informing the UK government the best possible province for their next COVID-19 vaccinations campaign.

Stakeholders:

Public Health England (PHE)

Data Sources:

- covid_19_uk_cases.csv
- covid_19_uk_vaccinated.csv
- tweets.csv

These are the primary data provided by PHE. Secondary and tertiary data may be required to provide more exhaustive analysis.

Data Analysis Workflow

GitHub

A public GitHub repository titled “LSE_DA_COVID_Analysis” is setup to host the code, data files, Jupyter Notebooks and relevant documents for the data analysis to inform the UK about the COVID-19 vaccinations between January 2020 and October 2021. This GitHub repository forms part of the analysis workflow, allowing continuous analysis with additional data sets from different sources, contributions from other data analysts or researchers. It adds values to the project and organization, and allows further collaborations and the stakeholders to continuously investing in the project to help with their decision making process.

Python

Python version 3.10 is used for the analysis with the PEP 8 style and guidelines.

Assignment activity 2: Import and Explore Data

Assumptions:

- Numbers of First Dose, Second Dose and Vaccinated are daily recorded figures.
- Numbers of Deaths, Cases, Recovered and Hospitalised are cumulative.
- Numbers of unvaccinated is the difference between First Dose and Second Dose, used for identifying province/state eligible for the Second Dose.
- People who received the Second Dose is classified as fully vaccinated.
- Default index are used for used in some DataFrames

Data Summary:

- 7584 records in total data sets
- 12 UK provinces - Anguilla, Bermuda, British Virgin Islands, Cayman Islands, Falkland Islands, Gibraltar, Isle of Man, Monserrat, Others, Turks and Caicos Islands, Saint Helena, Ascension and Tristan da Cunha.

Data preparations:

Missing Values:

In Bermuda, there are 2 records found with missing values for Deaths, Cases, Hospitalised and Recovered fields dated 21 and 22 September 2020. Since they are of type numeric, they have been replaced with a 0.

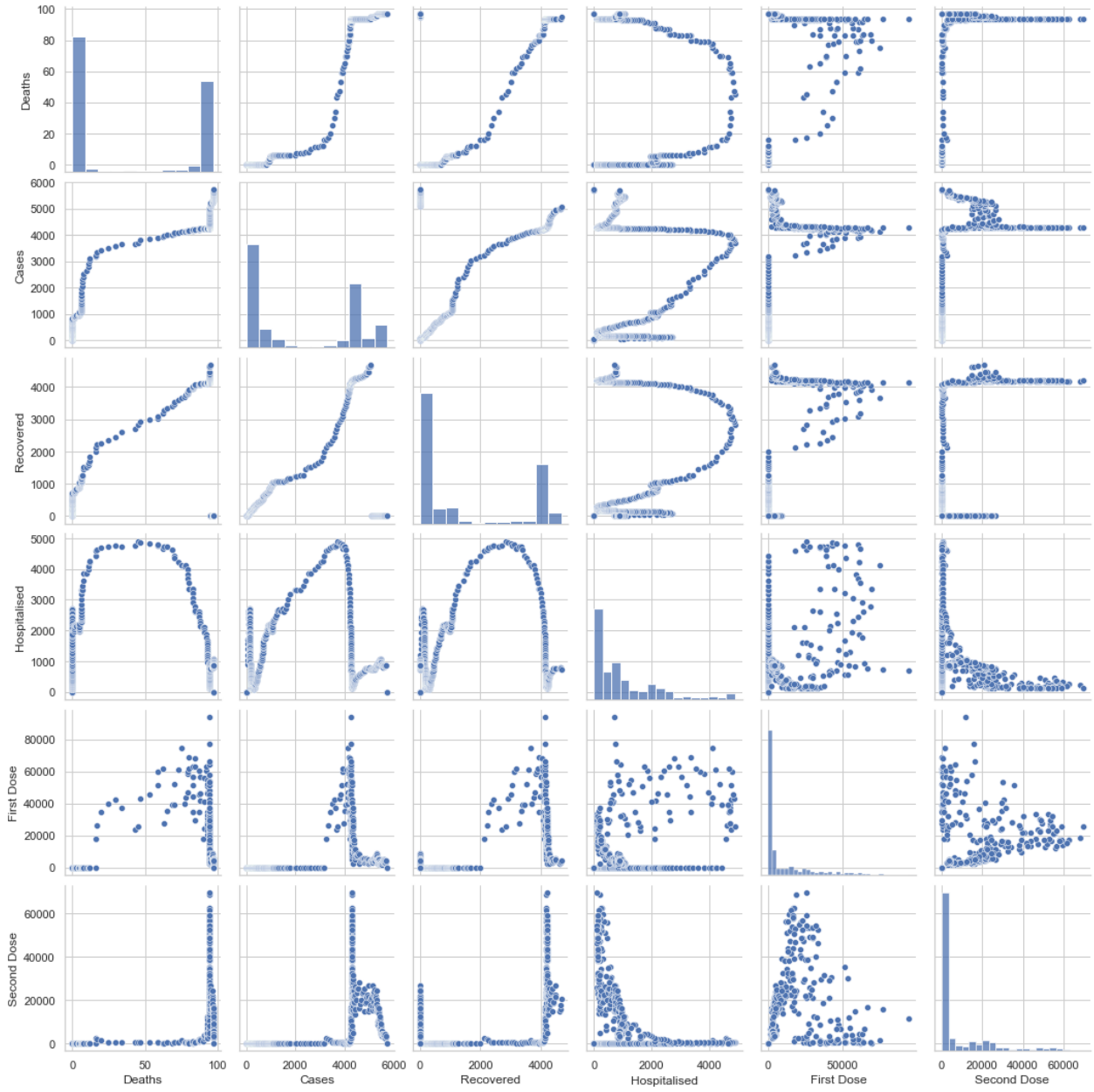
	Province/State	Country/Region	Lat	Long	ISO 3166-1 Alpha 3-Codes	Sub-region Name	Intermediate Region Code	Date	Deaths	Cases	Recovered	Hospitalised
875	Bermuda	United Kingdom	32.3078	-64.7505	BMU	Northern America	0	2020-09-21	NaN	NaN	NaN	NaN
876	Bermuda	United Kingdom	32.3078	-64.7505	BMU	Northern America	0	2020-09-22	NaN	NaN	NaN	NaN

Duplicates:

No duplicates have been identified.

Outliers Analysis:

Describe() function and pairplot are used to explore the outliers in the data sets and relationships between variables.



From the pairplot, all numbers have very strong clusters at the start because of the point of data collections. As First Dose increases, the numbers of Deaths, Cases and Recovered started to scatter and plateaux at the end, and Hospitalised numbers shift towards downward trends. Recovered remains low as the Second Dose hikes. There are some interesting relationships between variables such as First Dose and Hospitalised in this pairplot which we could explore more.

Assignment activity 3: Merge and analyse data

To explore the vaccination status, a subset of data based on Gibraltar is used to identify the number of people who are vaccinated, have received the first dose, and the second dose.

The first case of Covid in Gibraltar recorded on 04/03/2020. First recovered case recorded on 10/03/2020. Figures of hospitalisation only started on 27/03/2020. Records of vaccinations only started on 11/01/2021. This is a very interesting timeline, and it shows there is 10 months lapse between the outbreak of Covid in Gibraltar to the rollout of vaccination programme there.

Dates when the cases are first reported

Covid	Recovered	Hospitalised	Vaccination
04/03/2020	10/03/2020	27/03/2020	11/01/2021

As the vaccination programme is rolled out in phases by age groups and there is a recommended time lapse between first and second dose. It is important to take into account that the numbers used are aggregated for analysis, the rollout of vaccination programmes are by age groups, priorities and vulnerability of individuals. More information are needed to explore the trends and patterns of the following:

- Number of people in the difference age groups who have not received first dose
- Number of people in the age groups who have not received second dose 3 months after the first dose
- Number of people in the next age group for rolling out the next phase of the vaccination programme

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

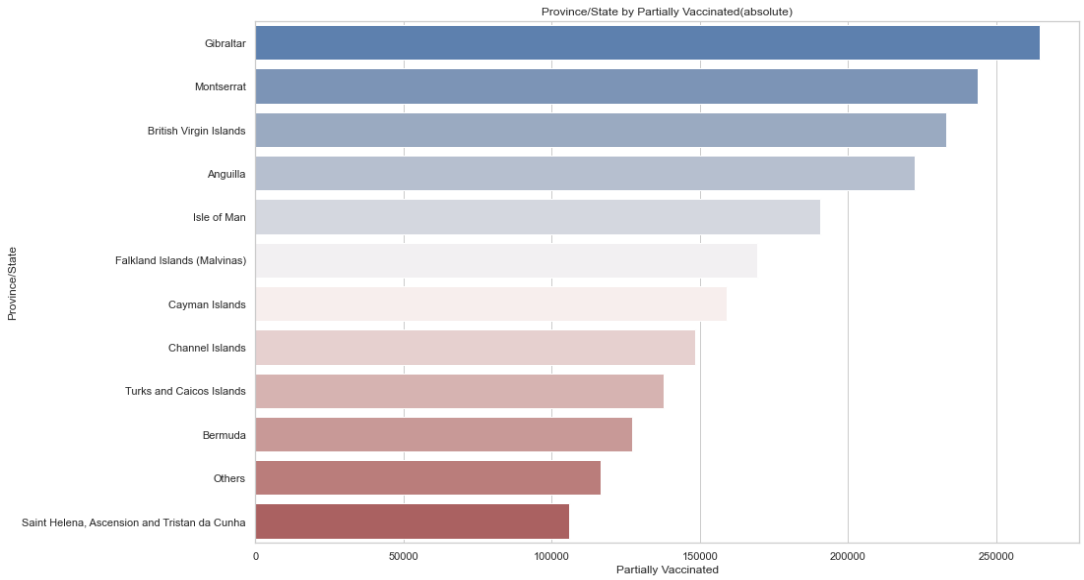
From the table, with absolute difference for individuals eligible for the Second Dose, Gibraltar has the highest. There is very insignificant change in relative percentage for the population eligible for second dose across all provinces. Turks and Caicos Islands has the highest percentage. A further analysis is recommended to explore the insights of this difference in trends and patterns.

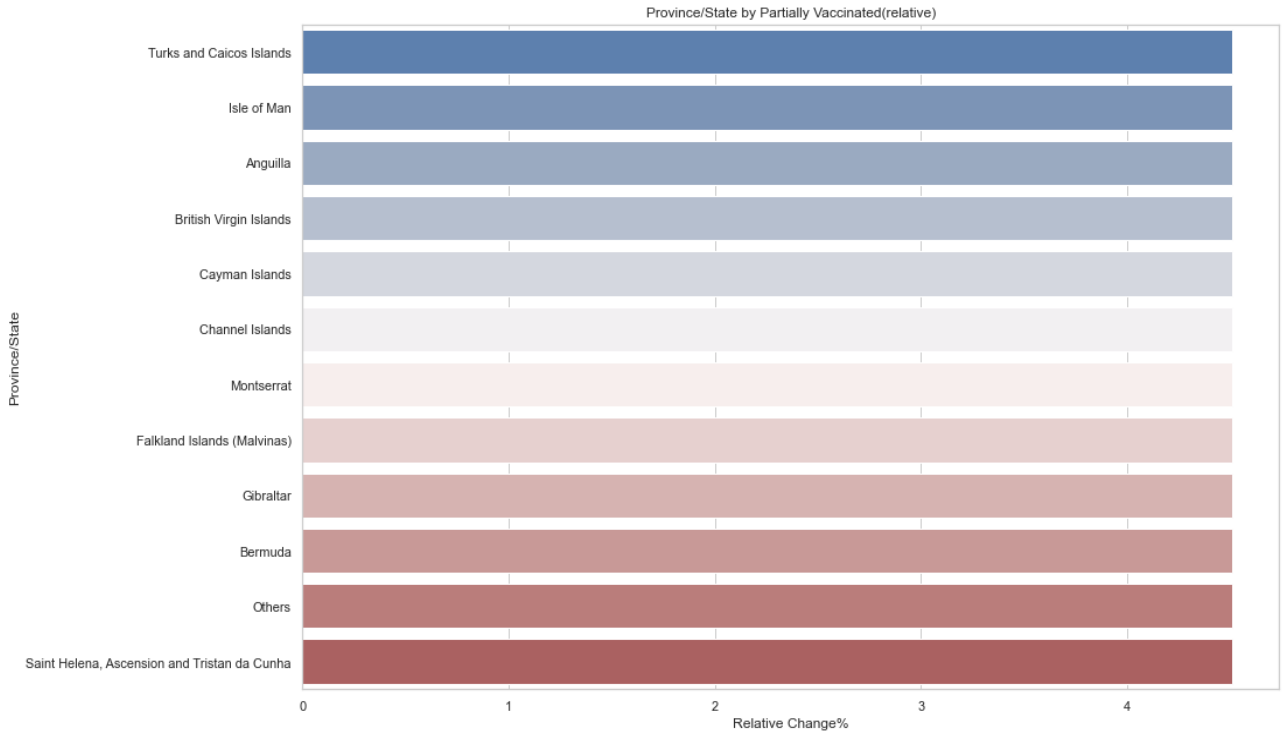
Province/State	Date	First Dose	Second Dose	Partially Vaccinated	Vaccinated
Anguilla	2021-01-11	15233	2181	13052	2181
Anguilla	2021-01-12	21804	1687	20117	1687
Anguilla	2021-01-13	29289	1023	28266	1023
Anguilla	2021-01-14	33253	552	32701	552
Anguilla	2021-01-15	35838	442	35396	442
Anguilla	2021-01-16	31094	275	30819	275
Anguilla	2021-01-17	19961	235	19726	235
Anguilla	2021-01-18	21428	479	20949	479
Anguilla	2021-01-19	36032	395	35637	395
Anguilla	2021-01-20	38168	358	37810	358
Anguilla	2021-01-21	43035	290	42745	290
Anguilla	2021-01-22	50216	191	50025	191
Anguilla	2021-01-23	51657	110	51547	110
Anguilla	2021-01-24	23126	86	23040	86
Anguilla	2021-01-25	29374	207	29167	207
Anguilla	2021-01-26	32661	180	32481	180
Anguilla	2021-01-27	32800	245	32555	245
Anguilla	2021-01-28	43514	186	43328	186
Anguilla	2021-01-29	51214	229	50985	229

The number of vaccinated individuals and individuals who have received second doses in Anguilla and in general are the same because they are individuals who have received both doses. The number of second doses varies quite significantly over time.

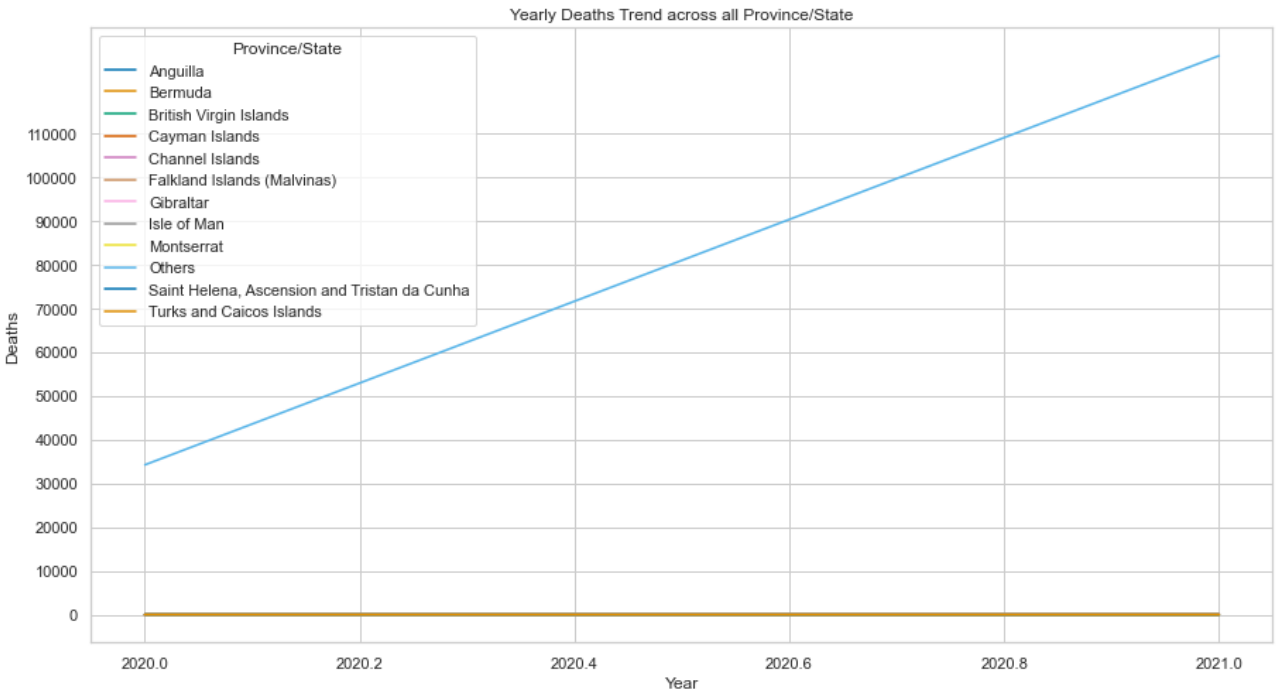
Assignment activity 4: Visualise and identify initial trends

The government is looking to promote second dose vaccinations. To identify the best possible area to test a new campaign, they look into partially vaccinated population, individuals who have only the first dose and are eligible the second dose.

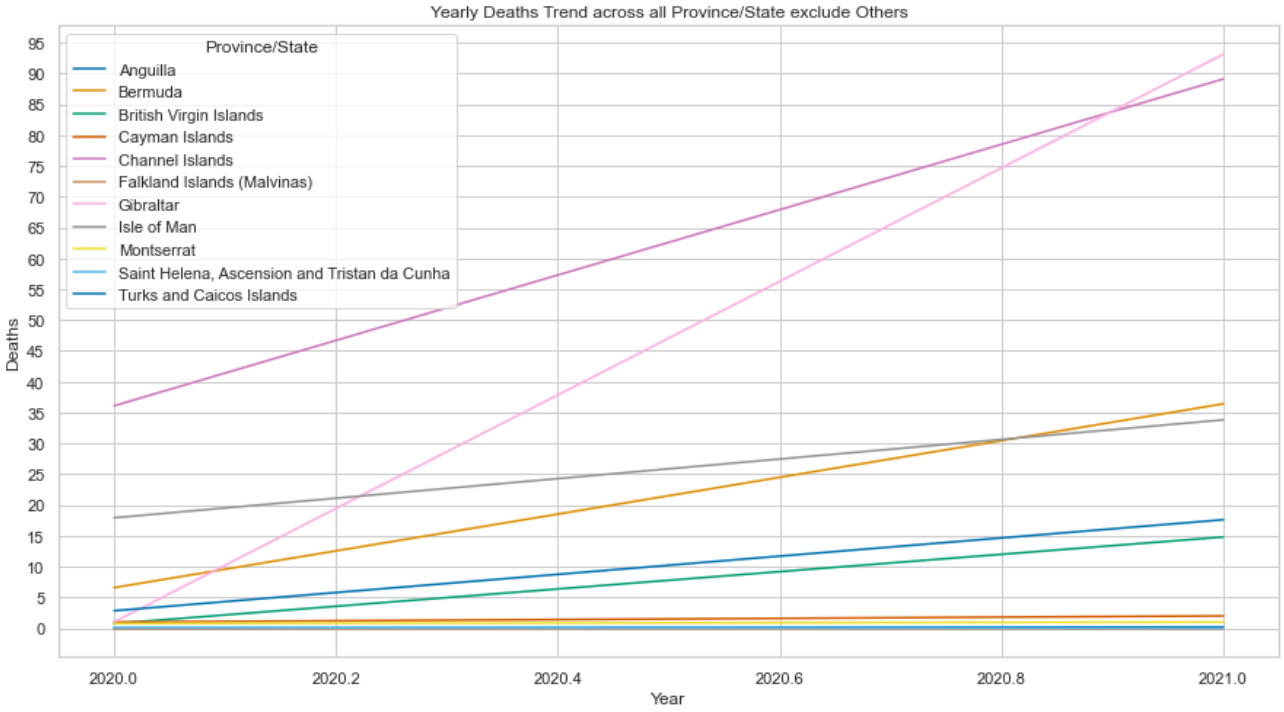




The barcharts of absolute and relative percentage give very different insights about the trends of the vaccinations across all provinces. Gibraltar has the highest vaccinated individuals, and Saint Helena has the lowest. Further analysis about the actual populations is recommended to identify if the number of vaccinated are proportional to the population of the provinces. There is very marginal different in the relative percentage change across all provinces.

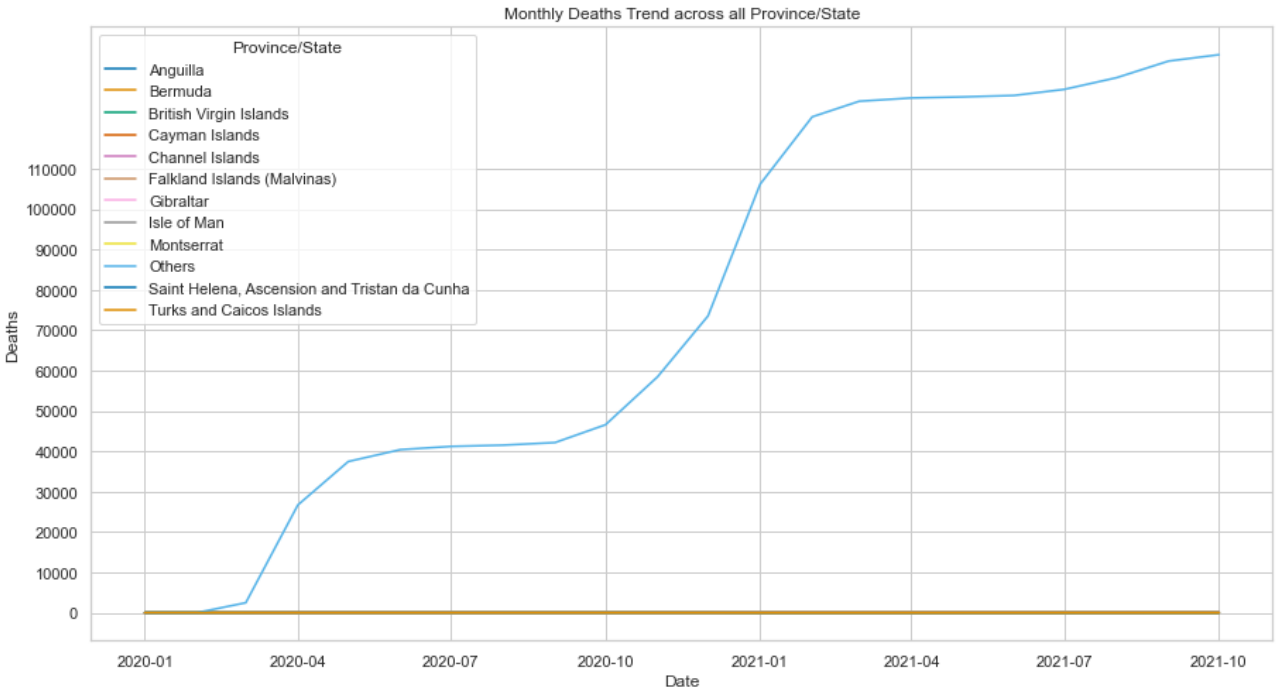


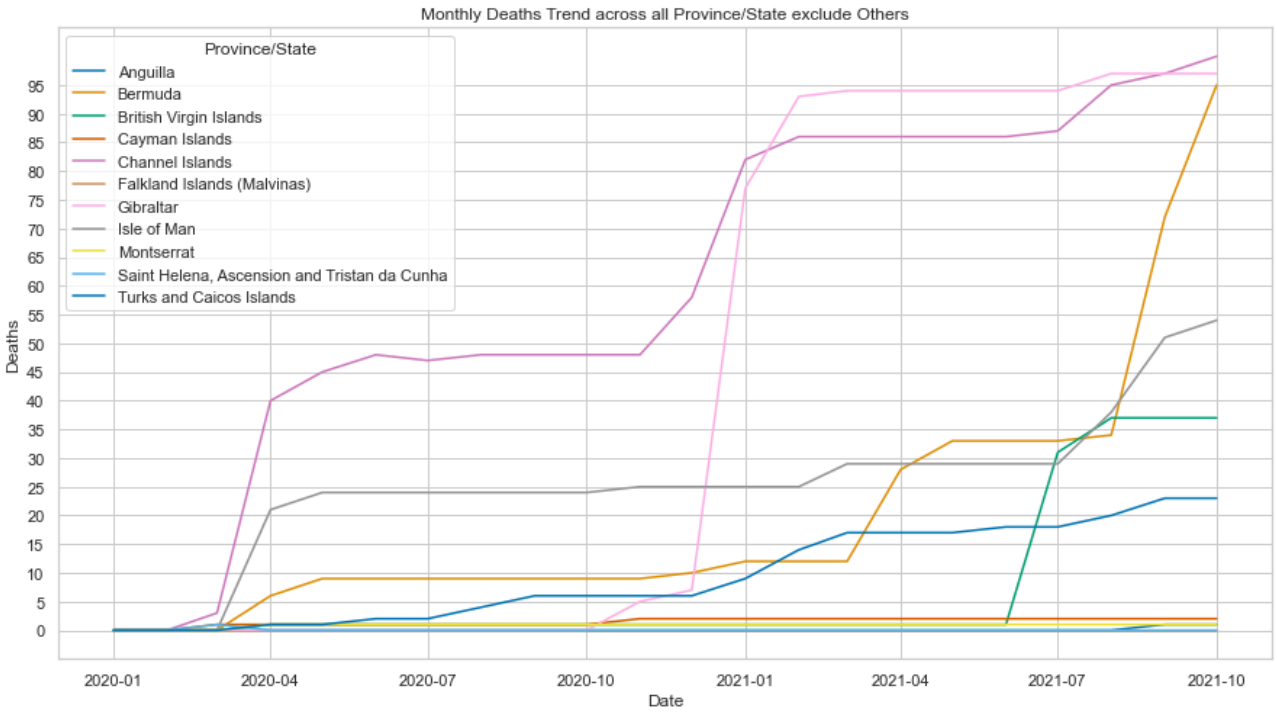
Others province dominates the yearly deaths trend across all provinces. It is impossible to visualise and identify any insights for the rest of the provinces. Others province has skewed the overall data with a high number of deaths, and should be removed and analysed separately.



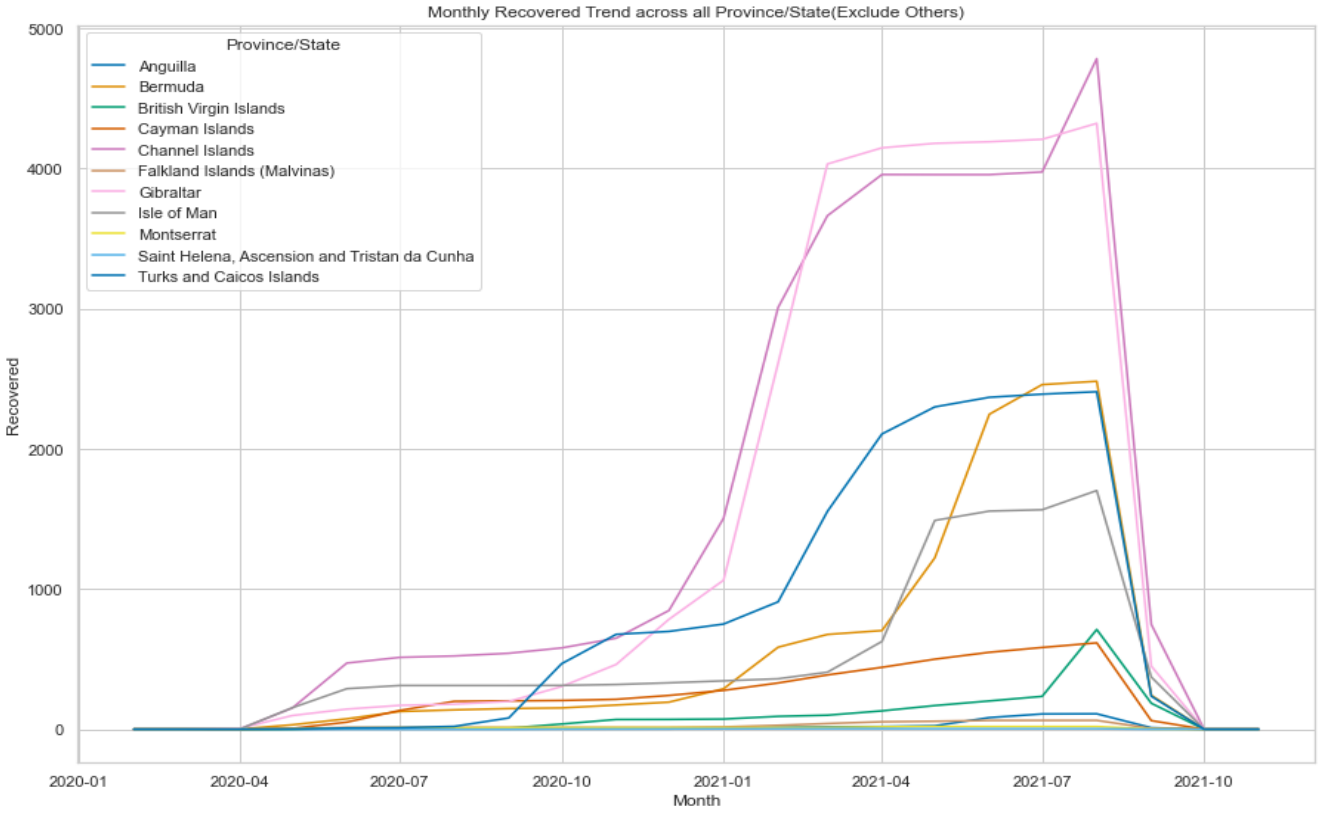
This lineplot without Others province shows upward deaths trends in Gibraltar and Channel Islands, followed by Falkland Islands. Provinces like Montserrat, Anguilla and Bermuda have rather low and flat deaths trends. There is lack of information about the populations of provinces during the period of analysis to provide more insights and to make an accurate analysis taking the context of pandemic into account.

To have a more granular views of the trends and patterns, visualise the deaths trend by months instead of years. The trends are smoother. The trends running across the data vary by provinces.

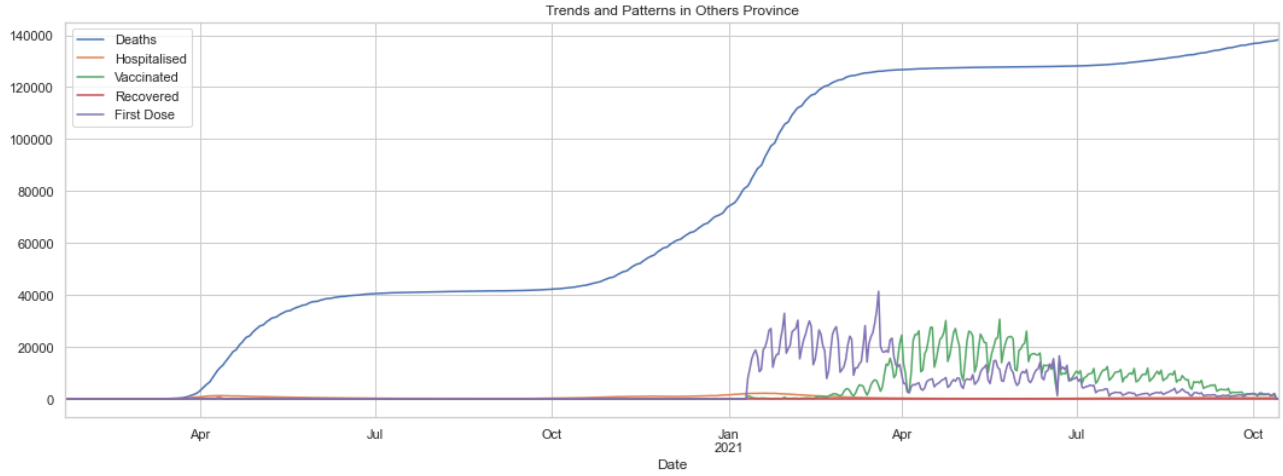




The death rates increase and vary at different period across all regions. Except Cayman Islands, Montserrat, Saint Helena which have relative low deaths, fewer than 5. Channel Islands peaks around March 2020, plateau for over months before another increase towards the end of the 2020. Gibraltar has an exponential increase from October 2020 and peaks around February 2021 before plateau. A peak is reached at October 2021 with data sets provided, more data will be needed to understand trends and patterns after October 2021.



Channel Islands and Gibraltar has had the most recoveries, and has been consistent over time before plateau between March and July 2021.



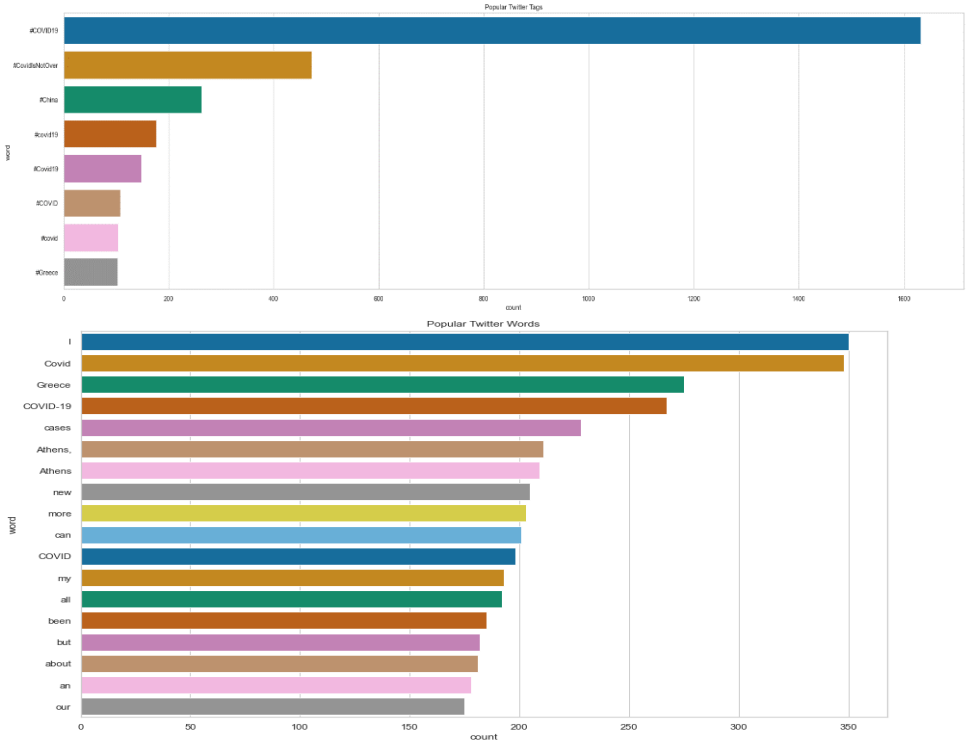
Others province shows very significant high death numbers. More information and analysis are needed to define what Others province include? There is a very big contrast against other trends including Recovered, First Dose, Vaccinated and Hospitalised.

To meet the guidelines for accessibility, I use palette for colourblind, a title, and simple design principles for visualisations. Dates on the X axis could be improved, to display vertically for instance.

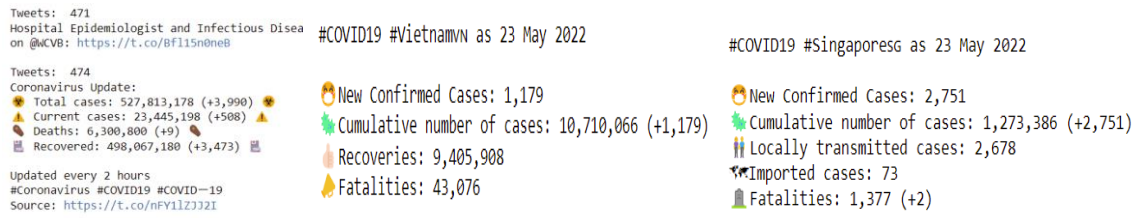
These visualisations give the government an overview and some insights of initial trends and patterns across all provinces. They are not sufficient to use in making informed decisions. More data from different sources, including secondary and tertiary, are recommended for more analysis and insights.

Assignment activity 5: Analyse the Twitter data

To have a more complete analysis, this project uses both quantitative and qualitative data. A sample of Twitter data with 3960 records are used to identify the most popular hashtags and keywords in relation to the covid outbreak.

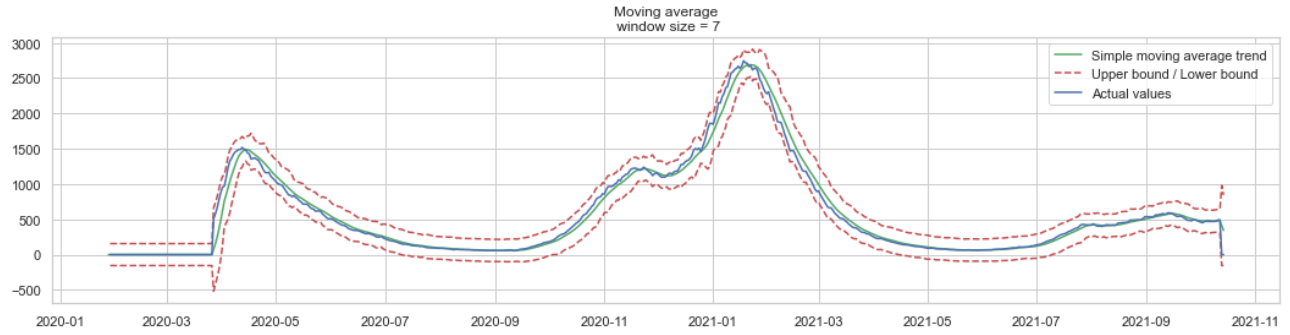


Hashtags contains words like “covid” return as the most popular tags in Twitter data. In order to understand the trends of vaccinations and covid, keywords like “cases”, “risk”, “dose”, “death”, “vaccination” are used to further capture relevant Tweet contents for analysis. Some interesting Tweets extracted like the following:

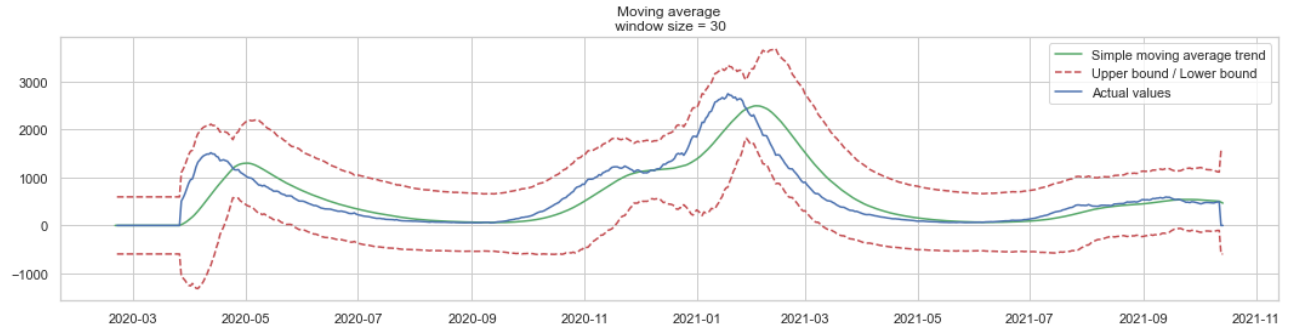


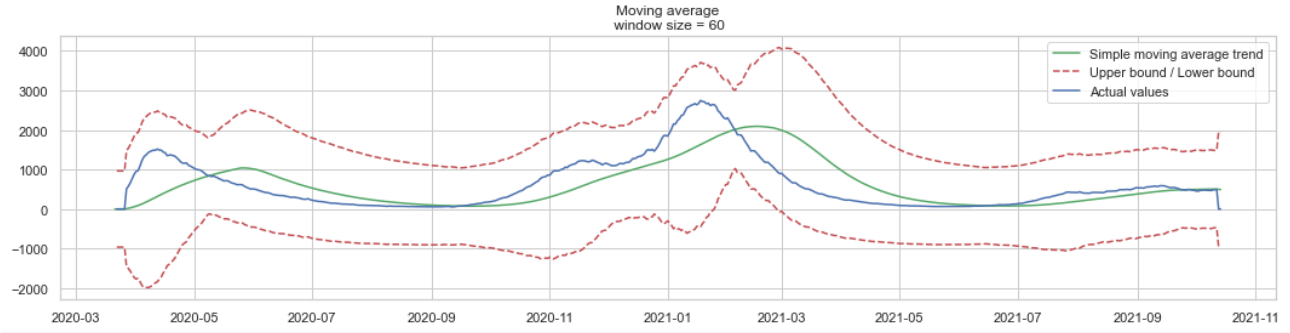
The sample of Twitter texts reveal more about international trends. The sample Twitter data is small. Twitter API scrapping is recommended to obtain a more up to date, relevant and larger data sets for better analysis and insights.

Assignment activity 6: Perform time-series analysis



Using the plot_moving_average() function, I passed in the Hospitalised data of Channels Islands, and set the rolling average as 7 days. The function plots a time-series. To prepare to meet the needs of any upward surges in hospitalisation, time-series forecasting is suitable for short and medium terms in this context. The time-series shows there are two waves in the past, the gradual surges took over 11 months to peak. The trend shows that it has been gradual increase in the hospitalisation, and more than 8 months have lapsed since the last peak on February 2021. I have decided to pass 30 days and 60 days rolling average to monitor the trends of hospitalisation for the next two months.





30 and 90 days forecast a smoother rise but the trends will not be a surge like the past wave in around January 2021.

3.1 What is the difference between qualitative and quantitative data?
Qualitative data, also known as categorical data. They are based on groups, interpretation, and description. Examples include provinces, tweet tags, tweet texts. Quantitative data are data which can be measured in numbers, and can be applied in statistical and mathematical manipulations. Examples include numbers of first dose, second dose, deaths and hospitalised.
Business can use quantitative data to find out how many, how much, or how often using calculations and apply statistical analysis to understand the past events and predict the probability of the future. Qualitative data helps business to understand why, how, or what happened behind certain behaviours, judgements and opinions, and make short term predictions.

3. 2 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 datasets from .csv files provided give some insights of the past trends and patterns of the covid vaccinations between January 2020 and October 2021 in the provinces. However, continuous improvements and adjustments are required because the Covid-19 is constantly changing, mutating and evolving with new variants. This project will, therefore, have to be maintained, updated and follow the changes in social, scientific, environment and other important contexts to provide good insights for business decision makings.

3.3 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)
Data ethics should be applied universally and have no boundaries even in the political context. They cannot be ignored. Even though the datasets are aggregated from the .csv files, data ethics have to apply to the use of data scrapped from TWEETS by filtering out and masking any personal details found in the Tweet texts.

Conclusion and Recommendations:

Government invests in this project with the aim to identify the best possible provinces for their next vaccination marketing campaign. Both quantitative and qualitative data are used in this analysis to identify trends and uncover insights. There are limitations with the datasets provided to help discovering the insights. Most data captured are cumulative, for instance Deaths, Hospitalised. Province like Others which if included skews the overall data, and has to be analysed separately and be excluded from the data sets. More information needed to understand what grouped the Others province. Data of populations in line with the data period Jan 2020 to October 2021 is recommended to source for understanding and refining the insights found so far. Twitter data sample is small, and trends found are more about international. Twitter API scrapping is recommended to capture larger and more relevant datasets for this project.

While no recommendation of the best possible province has been provided to target the next vaccinations campaign from the analysis so far. This project setup a solid ground in GitHub for further analysis and collaborations to help with the government decision making process in the future. While complying to data ethnics is important in the data analysis, recommendations are to source more data with a longer timeframe to help government to target and make good decisions. The datasets are captured during the pandemic between January 2020 and October 2021. Additional government policies about past vaccination programmes make available will help to break the numbers of individuals into groups for analysis as since during the pandemic individuals are prioritised by their vulnerability, disabilities, keyworkers and other factors.