Data Analytics Report: UK Government COVID-19 Vaccinations Programme

1. Overview

This report is focused on analysing COVID-19 data and provides a policy recommendation to inform the vaccination marketing strategy conducted by the UK Government. The patterns and trends covered in this Report provide an insight into the pandemic landscape from January 2020 to October 2021 and aim to assist the UK Government in its campaign to enhance the COVID-19 vaccine acceptance rate and ultimately increase the number of fully vaccinated individuals across the British Overseas Territories.

2. Importing and Exploring Data

Section 2 of the Report has conducted a descriptive analysis of the data provided for the COVID-19 cases and vaccination rate across the British Overseas Territories by completing the following steps:

- *First*, the Report used the info (), describe (), shape and value_counts () methods to observe data types, records and numbers, and column features.
- *Second*, the isna () function was applied to identify missing data in the given datasets.
- *Third*, the Report applied filtered data for a particular region (i.e., Gibraltar) and proceeded to create Pandas DataFrame based on the data for Gibraltar.
- *Finally*, the Gibraltar DataFrame was subsetted to create aggregated columns displaying the change in the total number of deceased, infected, recovered and hospitalised individuals over time.

The following trends and patterns were observed while conducting a preliminary analysis:

a. The initial observations after exploring "cov" and "vac" Pandas DataFrame indicate the non-recorded data for COVID-19 cases and vaccination in the "covid_19_uk_cases" and "covid_19_uk_vaccinated", respectively. The given datasets have overlapping columns, including the geographic location (e.g., "Province/State, "Country/Region", etc.) and time (i.e., "Date").
b. From the implementation of the head () function, it is evident that the default index for DataFrame is the range of numbers starting at 0 [see Appendix A].
c. The records for the COVID-19 cases and vaccination over the first months of the pandemic are mostly absent from both datasets, which could be explained by either the absence of recorded cases or the failure to report new cases.
d. The number of vaccinated individuals changed exponentially over time, which could be indicative of the (1) successful public health campaign conducted by the UK government to curb the pandemic *or/and* (2) increased awareness among the population of the potential benefits of the vaccine.
e. There are 2 rows (Index 875 and 876) with missing or non-assigned values in the "covid_19_uk_cases" DataFrame in 12 columns. The missing values are in the "Deaths", "Cases", "Recovered", and "Hospitalised" columns and belong to Bermuda. Respectively, there are 0 rows with missing values present across 11 columns in the "covid_19_uk_vaccinated" DataFrame [see Appendix B].
f. The filtering of "cov" DataSet indicates there were no COVID-19 instances in Gibraltar during 22 January 2020–26 March 2020, as evidenced by "0.0" across the first 65 rows in the "Deaths", "Cases", "Recovered", "Hospitalised" columns [see Appendix C].

3. Merging and Analysing Data

Section 3 of the Report has completed the following steps:

- Merging the given DataFrames into and exploring a new "covid" DataFrame using the info () function.
- Converting the data type of the "Date" column from object to DateTime.
- Cleaning up unnecessary columns and creating a new "covid_results" DataFrame that meets the set parameters.
- Adding calculated features to the merged Data Frames applying groupby () and agg () function to display the difference between First and Second Dose vaccinations as a total ("Province/State") and overtime ("Date").

Based on the observed trends in the vaccination dataset measured both as totals and percentages, the Report has discovered the following insights:

a. Gibraltar has the highest number of individuals who have received First Dose but not Second Dose, i.e., "Difference per region" = 264,745 [see Appendix D].
b. The Turks and Caicos Islands have the highest percentage of individuals who have received First Dose but not a Second Dose, i.e., "Difference%" = 4.723142% [see Appendix E].
c. In the dataset in general, the number of vaccinated individuals and individuals who have received the first and second doses has increased over time. For example, Anguilla saw a doubling in the number of partially vaccinated individuals and a dramatic increase in the number of individuals with "First Dose" and "Second Dose" between October 10–13.

Further exploration will focus on other statistics pertaining to the COVID-19 pandemic, including mortality and recovery rate observed as aggregated numbers and over time.

4. Visualising and Identifying Initial Trends

Further analysis conducted in **Section 4** provides a side-by-side comparison of the fully and partially vaccinated population to inform the second dose vaccination campaign planned by the UK government.

Vaccination

Gibraltar has had the highest number of people who have received a first dose and not a second dose. That is, Gibraltar has been least successful in encouraging the population to get the Second Dose, as reflected in the "Difference per region" indicator [Figure 1] measuring the total number of partially vaccinated individuals eligible for the Second Dose.[1]

[1] *See COVID-19 vaccination assumptions below*

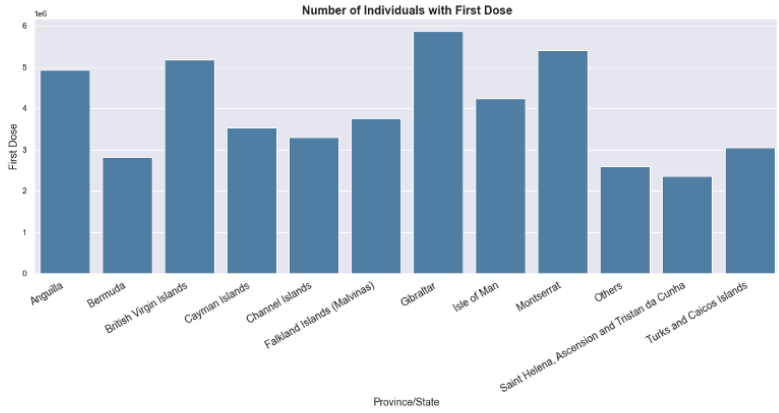
The calculation of the difference between first and second follows the two major assumptions: (a) only individuals with the First Dose are eligible for a Second Dose (b) individuals with the First Dose only are considered partially vaccinated as opposed to fully-dosed population with both shots.

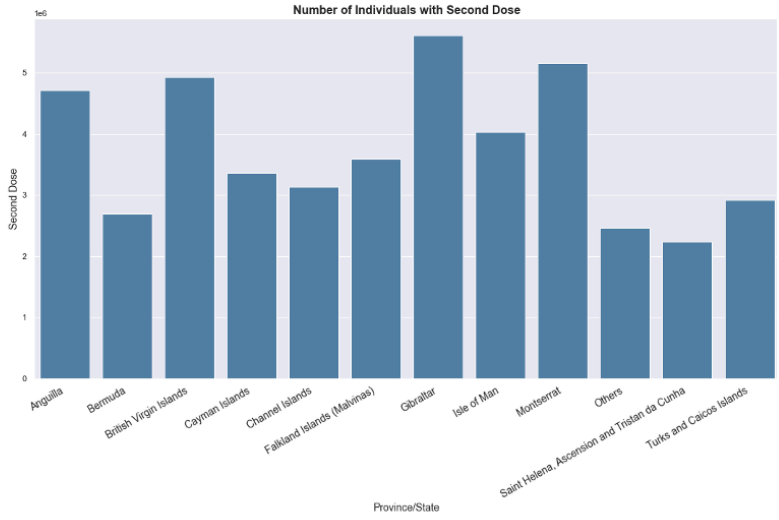
Figure 1

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

In terms of total numbers, the four provinces including Gibraltar, Montserrat, British Virgin Islands and Anguilla have the highest actual number of individuals with the First Dose and Second Dose [Figure 2].

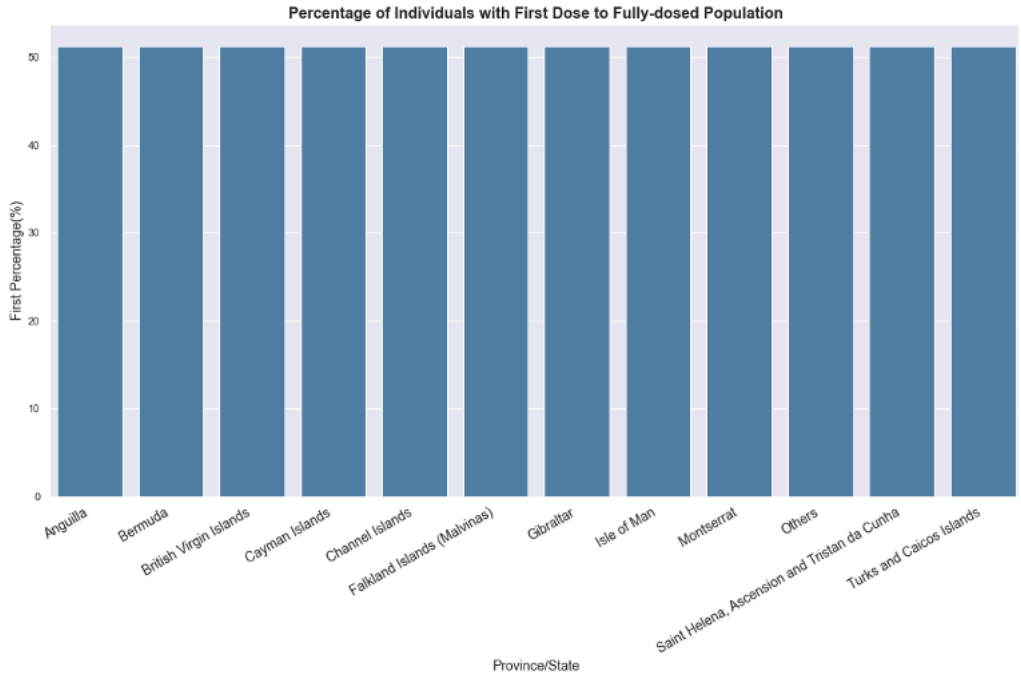
Figure 2





Regarding relative numbers, there has been no significant difference in partially vaccinated people across Provinces/States, with the ratio of individuals with First Dose to the fully-dosed population being approximately equal to 51% [Figure 3]. Similarly, "Rate of Interest" showing the percentage difference between the individuals eligible for the Second Dose (i.e., "Difference per region") and individuals with the First Dose only has remained roughly at 4.5% [Figure 4].

Figure 3



Province/State	First Dose	Second Dose	Difference per region	Ratio of Intererest	Fully-dosed	First Percentage
Anguilla	4931470	4709072	222398	4.509771	9640542	51.153452
Bermuda	2817981	2690908	127073	4.509363	5508889	51.153345
British Virgin Islands	5166303	4933315	232988	4.509763	10099618	51.153450
Cayman Islands	3522476	3363624	158852	4.509669	6886100	51.153425
Channel Islands	3287646	3139385	148261	4.509640	6427031	51.153417
Falkland Islands (Malvinas)	3757307	3587869	169438	4.509560	7345176	51.153396
Gibraltar	5870786	5606041	264745	4.509532	11476827	51.153389
Isle of Man	4226984	4036345	190639	4.510048	8263329	51.153524
Montserrat	5401128	5157560	243568	4.509577	10558688	51.153401
Others	2583151	2466669	116482	4.509299	5049820	51.153328
Saint Helena, Ascension and Tristan da Cunha	2348310	2242421	105889	4.509158	4590731	51.153291
Turks and Caicos Islands	3052822	2915136	137686	4.510122	5967958	51.153544

Figure 4

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

On balance, the UK Government should target Gibraltar among the first regions to promote the benefits of full immunisation, although the region has been rather successful in total numbers of vaccinations but is falling behind in terms of relative numbers.

Deaths and Recoveries

Additional observation has considered a change in the number of deaths and recoveries over time to evaluate the existing disparities between the different provinces/states. Data visualisation has revealed the following patterns:

Figure 5

The daily number of deaths follows the upward trend, with several peaking points over time and a sudden surge mid-year indicative of the outliers in data.

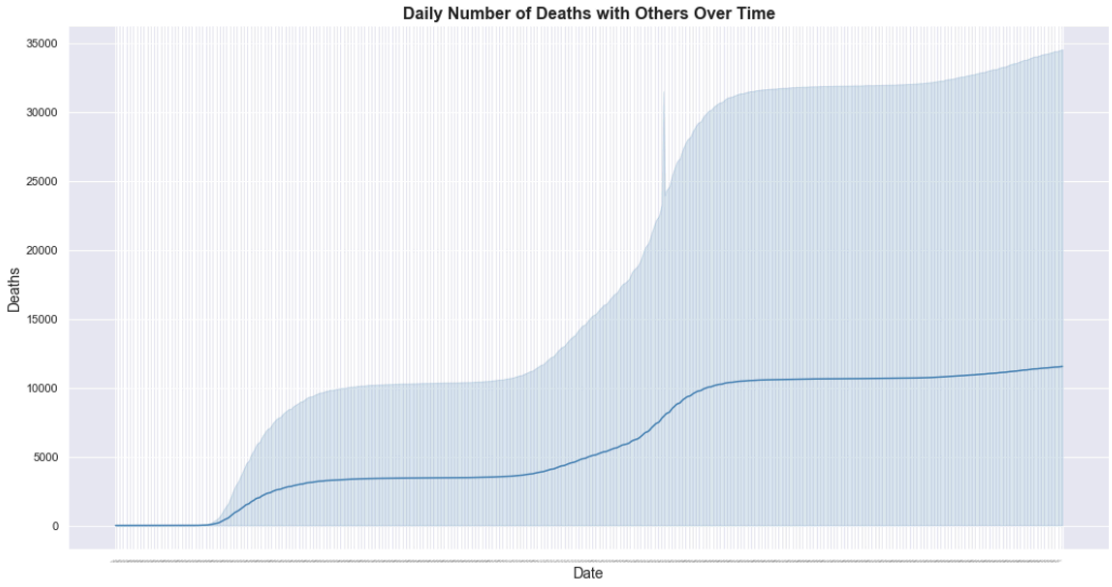


Figure 6

Aggregated deaths statistics follow similar patterns observed in vaccination data, with the provinces such as Anguilla and Montserrat that lead in total vaccination numbers having the lowest or close to zero daily and monthly number of deaths. At the same time, Gibraltar and the Channel Islands prominently exceed other provinces in terms of both daily and monthly deaths, despite showing comparably good results in total vaccination numbers.

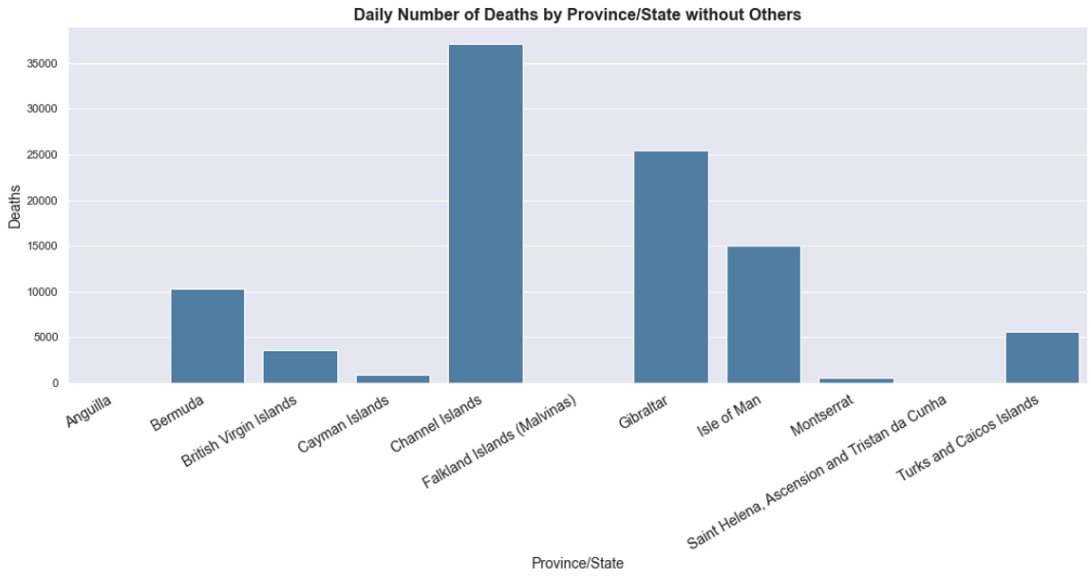


Figure 7

When measured on the monthly basis, the pattern remains predominantly unchanged, although the conversion into the average monthly numbers resulted in a smoother curve with more accurate measurements of peak data.

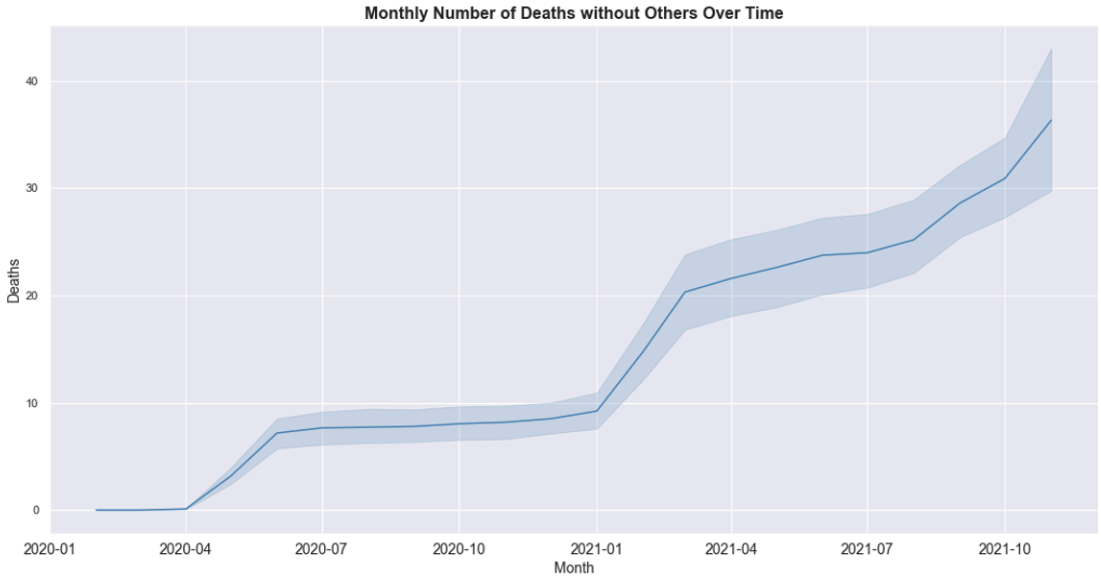
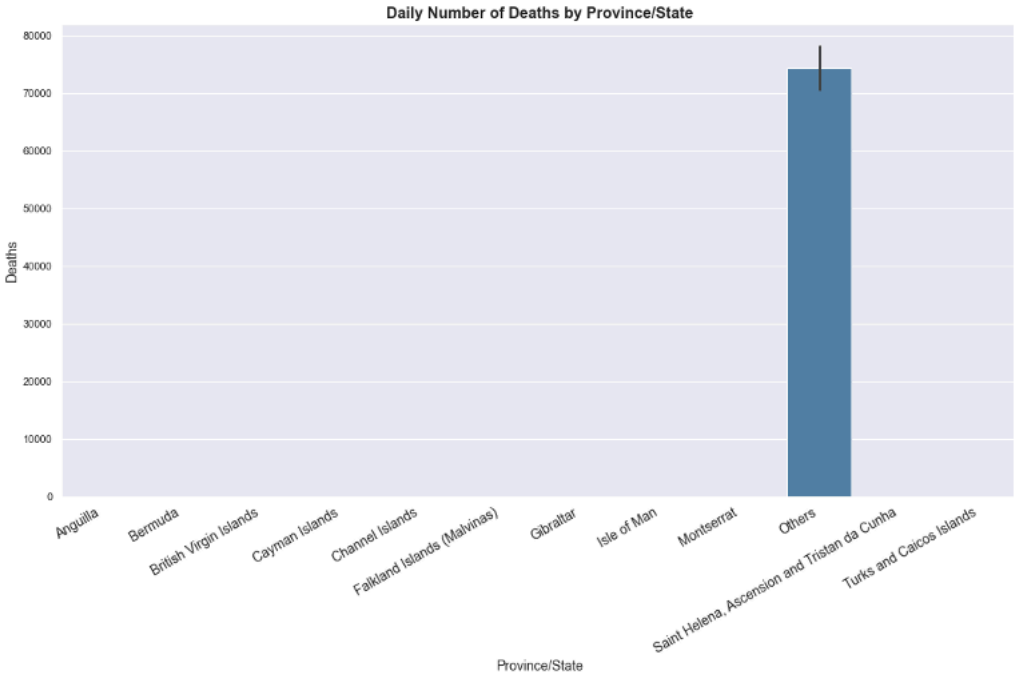


Figure 8

To be able to observe the pattern, certain groups of data considered outliers, such as 'Others' that notably skewed the overall picture, need to be separated and analysed in isolation.



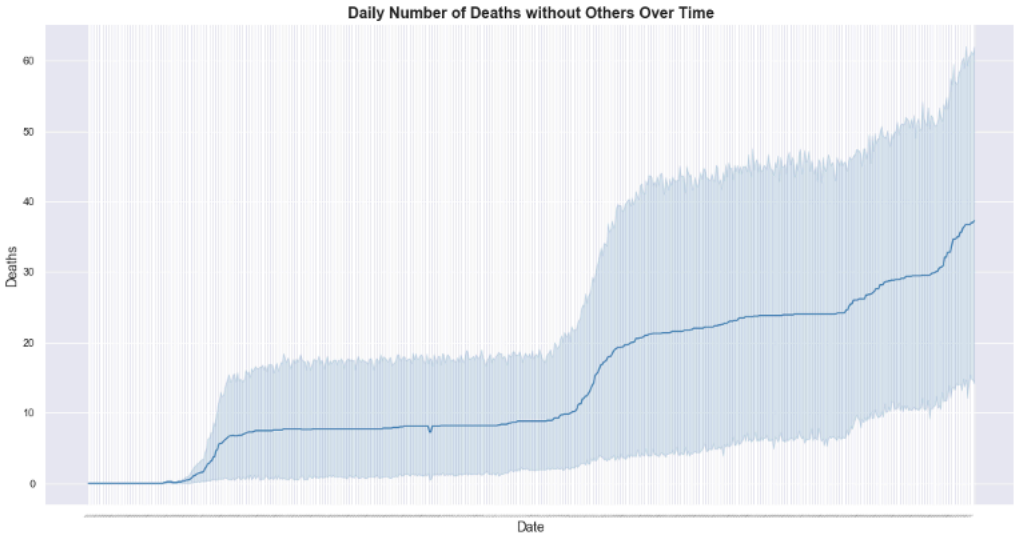


Figure 9

Monthly number of recovered cases has an upward trend, with a dramatic increase in January 2021. The number of recovered individuals continued to grow and reached the top in July-August to plummet in October 2021.

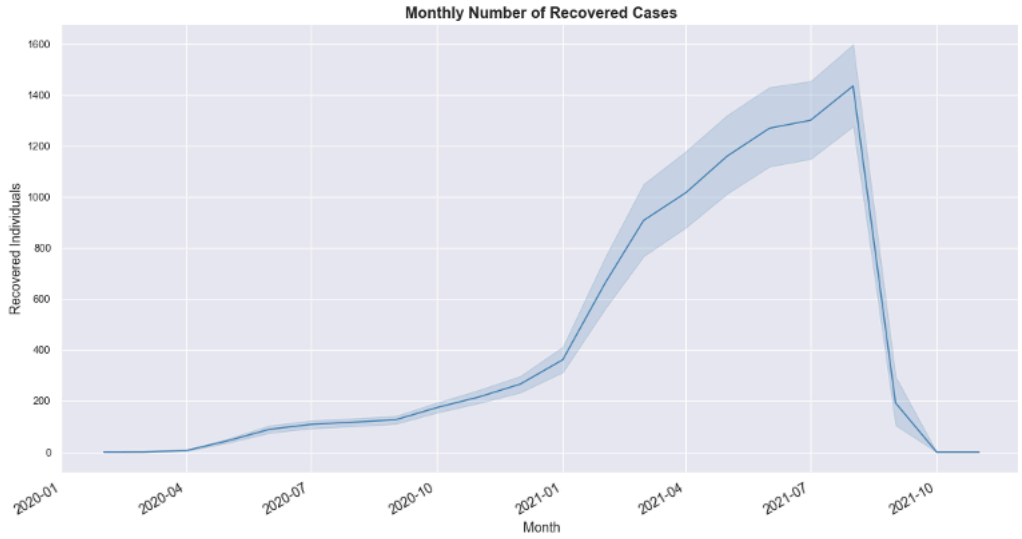
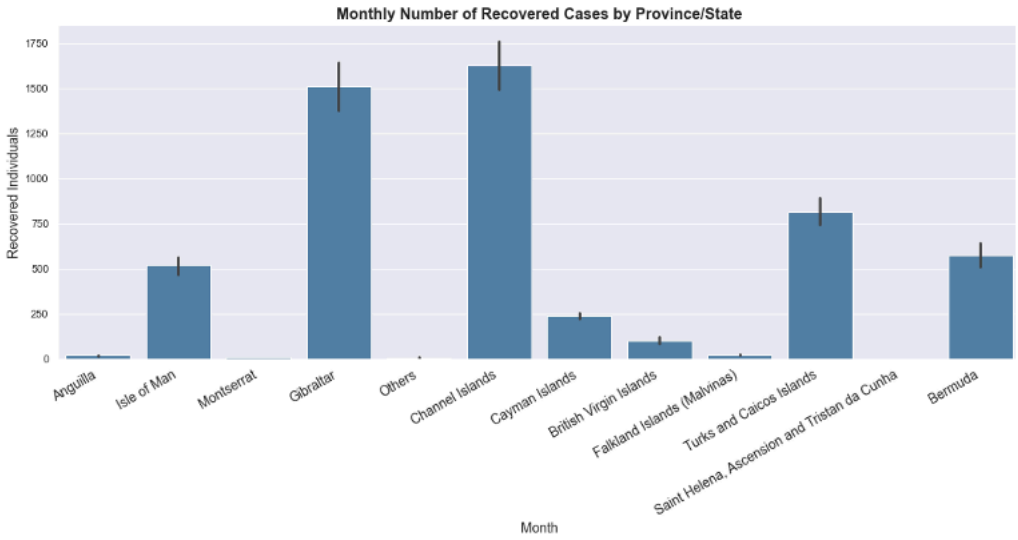


Figure 10

The number of recovered cases consistently remained higher for Gibraltar and Channel Islands.



Having considered additional features such as mortality and recovery statistics, the Report has discovered the following insights:

a. There is a consistent upward trend in vaccination, deaths and recovery cases as the pandemic progresses. The data for 'Others' is skewed and should be removed.

b. The line graph displaying the number of deaths and recoveries has a smoother curve after converting Date into Months.
c. The death rates have increased across regions over time, reaching a peak in October 2021.
d. The percentage of the first dose to fully vaccinated individuals, or "First Percentage", is equal to 51%.
e. Gibraltar has had both the highest total vaccination (First Dose and Second Dose) and the highest number of deaths (daily and cumulative).
f. The Channel Islands have had the most recoveries, which remained consistent over time.
g. The Report has achieved a relatively good quality of visualisations. Apart from the Date-to-Month conversion, the smoothens of the lines could be further improved by cleaning the data and removing outliers.
h. The visualisations provided in the Report could be used to determine the target regions for the UK Government's vaccination campaigns, as well as assess the overall impact of the COVID-19 pandemic and successful management of post-pandemic recovery.

5. Analysing Twitter Data

Section 5 of the Report looks at the external data to complement the existing COVID-19 datasets while reviewing the provided file and step-by-step demonstrating how the external function could be used to (a) count retweets and favourites, (b) identify Twitter hashtags containing the word "COVID" and (b) visualize the top trending hashtags.

Retweets and favourites

Figure 11

The report has explored the structure of the dataset and counted the retweeted and favourite Twitter posts.

```
In [72]: tweets.retweet_count.value_counts()
Out[72]: 0      2818
         1       570
         2       190
         3        96
         4        54
                ...
         49        1
         33        1
         261       1
         212       1
         52        1
         Name: retweet_count, Length: 67, dtype: int64

In [73]: tweets.favorite_count.value_counts()
Out[73]: 0       2240
         1        682
         2        252
         3        174
         4        100
                 ...
         41         1
         116        1
         741        1
         3433       1
         301        1
         Name: favorite_count, Length: 117, dtype: int64
```

Top trending hashtags

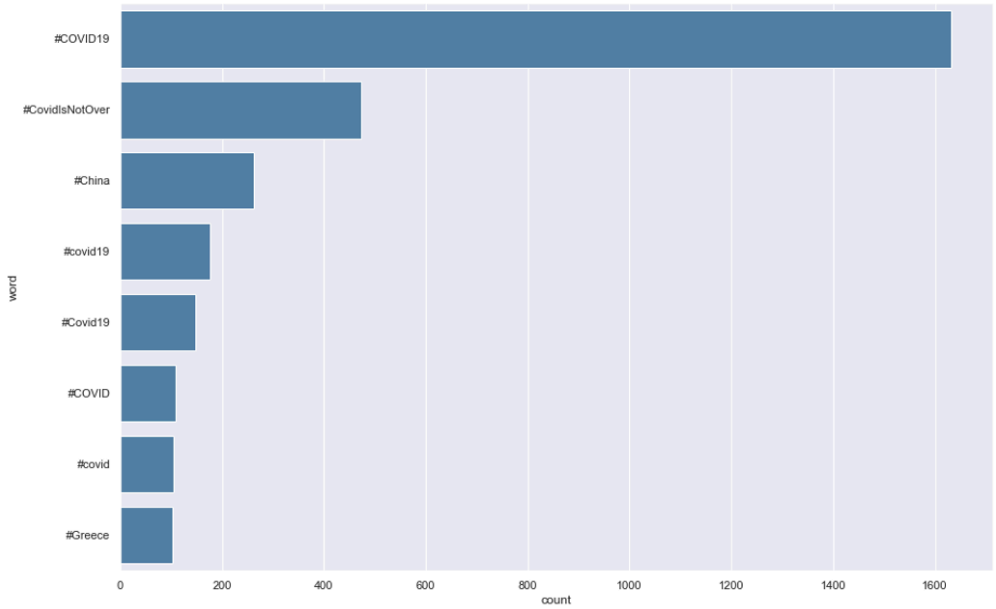
Figure 12

Having created a list of values containing the # symbol, the Report creates a Pandas Series to count the values in the list. Here, the Report looked at the first 30 #tags used in the messages.

```
#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
```

Figure 13

The bar chart indicates that #COVID19 was one of the trending hashtags when the tweets and the related data were extracted off Twitter using Twitter APIs.



While a limited dataset is insufficient to generate a meaningful inference from the identified tweets and hashtags trends, the Report makes several recommendations regarding the future use of Twitter API:

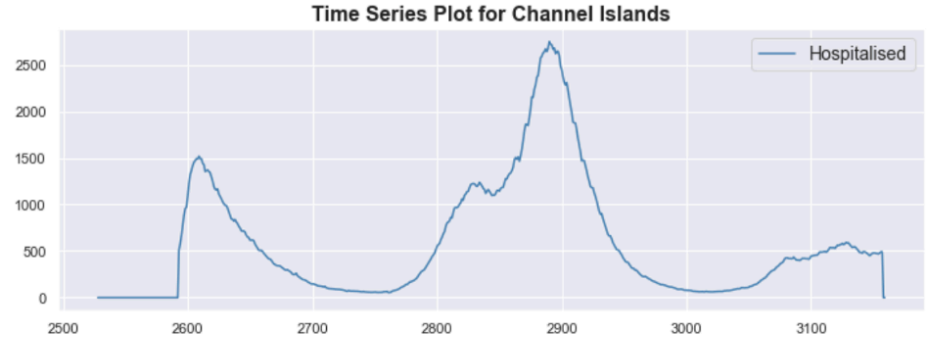
- Counted retweets and favourites could be applied to evaluate Twitter messages by identifying the original posters or predicting the hashtags that have a higher probability of being noticed.
- The identified trending hashtags contained both the search word and related words, which could be used to refine future searches to provide richer insights.
- Although Twitter-sourced data remains a popular strategy in use in the business world, Twitter API is based on live and current events which mean that the outputs could vary.
- Accessing the Twitter API could inform future government marketing campaigns by gauging public interest and sentiment on the proposed public health policies.

6. Time-series forecasting

The **final section** of the Report evaluated and conducts a time-series analysis using Python code to look at the data irregularities.

Figure 14

In particular, a **time-series analysis** was performed to look at surges in the number of hospitalised individuals in the Channel Islands.



Part 1: Instruction Given to Consultant

The two functions provided by the consultant, plot_moving_average () and mean_absolute_error (), must have been used to conduct time-series forecasting.

Figure 15

By plotting the ***moving average***, the consultant would eliminate any irregularity in the data series.

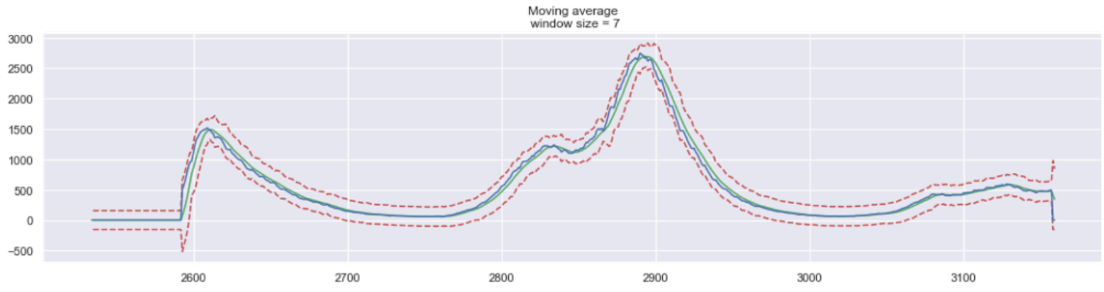


Figure 16

The consultant would calculate the ***mean absolute error*** to measure the average magnitude of the errors in the forecasted dataset.

	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

With a dramatic rise in the hospitalisation cases during the beginning and middle of the pandemic, the Report recommends the UK Government prioritise marketing campaign(s) in the Channel Islands to alleviate Covid-19 symptoms and decrease the hospital admittance rate.

Part 2: Additional Questions

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

Qualitative data measures types, or categories, and can be represented by names, symbols or numeric codes. It relies on interpretation and description, which helps understand why, how or what happened behind certain observed patterns and behaviours. Quantitative data is represented by measures numeric values suitable for statistical analysis. It shows measurable quantities used for calculations and helps answer how many, how much or how often questions behind the data. Both qualitative and quantitative data could be used in business predictions. For example, forecasting modelling could be categorised into qualitative and quantitative techniques based on the type of data. Common examples of qualitative forecasting models include the Delphi method, market research and Panel consensus. The two widely-spread quantitative techniques are time-series forecasting (e.g., exponential smoothing, etc.) and causal analysis (e.g., regression, econometrics, etc.).

Question 3.2 (double click cell to edit)

- 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?

In data analytics and forecasting, continuous improvement reflects the strive to perform better with each new version by introducing consistent and incremental enhancements. When the proactive changes are implemented effectively across the organisation, a continuous improvement mindset can benefit teams, departments, and individuals in the organisation. Although implementing a project at the current stage might arguably save time, continuous improvement yields overall more efficient processes, higher quality and a reduction in wasted time and effort. Furthermore, continuous improvement practices could generate impactful changes that produce more reliable insights to inform decision-making and thereby enabling an organisation to stay competitive and achieve strategic goals.

Question 3.3 (double click cell to edit)

- 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?

A data ethics framework guides the responsible data use in business, government and wider public applications. The concept implies the application of ethical considerations and responsible data governance throughout the processes and procedures related to managing, using and protecting data. The use of aggregated data may indeed protect the identity and personal details, thus conforming to best data protection practices and adhering to the public's perception and expectation of privacy. However, ensuring an ethical collection of data constitutes only a part of the ethical data governance, which entails other important practices from a data ethics standpoint. These not only guarantee that the way data is collected and stored complies with data privacy legislation but also guard against any unauthorised dissemination of data or technology and prevent third-party from misusing private information. In addition, data protection could also include the training programmes to raise data ethics awareness among the employees as well as the installation of effective endpoint, network and email filters to prohibit malware or dangerous files from entering the government database.

Conclusion

In considering trends and patterns identified in the COVID-19 cases and vaccination statistics, the Report has made several suggestions to inform the UK Government's marketing campaigns. As COVID-19 remains an important concern for public health, the Report recognises the benefits of a data-driven approach in enhancing national response in the post-pandemic time.

Appendix

Appendix A

```
# View the first five rows for 'covid_19_uk_cases.csv'.
cov.head()
```

	Province/State	Country/Region	Lat	Long	ISO 3166-1 Alpha 3-Codes	Sub-region Name	Intermediate Region Code	Date	Deaths	Cases	Recovered	Hospitalised
0	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-22	0.0	0.0	0.0	0.0
1	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-23	0.0	0.0	0.0	0.0
2	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-24	0.0	0.0	0.0	0.0
3	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-25	0.0	0.0	0.0	0.0
4	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-26	0.0	0.0	0.0	0.0

```
# View the first five rows for 'covid_19_uk_vaccinated.csv'.
vac.head()
```

	Province/State	Country/Region	Lat	Long	ISO 3166-1 Alpha 3-Codes	Sub-region Name	Intermediate Region Code	Date	Vaccinated	First Dose	Second Dose
0	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-22	0	0	0
1	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-23	0	0	0
2	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-24	0	0	0
3	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-25	0	0	0
4	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-26	0	0	0

Appendix B

In [12]:
```
# Determine the number of missing values for 'covid_19_uk_cases.csv'.
cov_na = cov[cov.isna().any(axis=1)]
cov_na.shape
# There are 2 rows with NaN values in 12 columns of the cov DataFrame.
```

Out[12]: (2, 12)

In [103]:
```
cov_na = cov[cov.isna().any(axis=1)]
display(cov_na)
```

	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

In [13]:
```
# Determine the number of missing values for 'covid_19_uk_vaccinated.csv'.
vac_na = vac[vac.isna().any(axis=1)]
vac_na.shape
# There are 0 rows with NaN values in 11 columns of the vac DataFrame.
```

Out[13]: (0, 11)

Appendix C

```
In [20]: #Subset the Gibraltar DataFrame that you have created consisting of the following columns: Deaths, Cases, Recovered and
         # Selecting few columns.
         cov_Gibraltar_fcol = pd.read_csv('covid_19_uk_cases.csv',
                                          usecols=['Deaths', 'Cases', 'Recovered', 'Hospitalised'])

         # Printing the DataFrame.
         cov_Gibraltar_fcol
```

57	0.0	0.0	0.0	0.0
58	0.0	0.0	0.0	0.0
59	0.0	0.0	0.0	0.0
60	0.0	0.0	0.0	0.0
61	0.0	0.0	0.0	0.0
62	0.0	0.0	0.0	0.0
63	0.0	0.0	0.0	0.0
64	0.0	0.0	0.0	0.0
65	0.0	0.0	0.0	763.0
66	0.0	2.0	0.0	869.0
67	0.0	2.0	0.0	1000.0
68	0.0	2.0	0.0	1165.0

Appendix D

```
In [15]: # Difference between first and second dose shows the total number partially vaccinated people by province/state, i
         covid_vaccination ['Difference per region'] = covid_vaccination ['First Dose'] - covid_vaccination ['Second Dose']
         covid_vaccination
```

Out[15]:

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

Appendix E

```
In [10]: # Difference% between first and second dose shows the percenrage of partially vaccinated people by province/region
         covid_vaccination ['Difference%'] = covid_vaccination ['Difference per region'] / covid_vaccination ['Second Dose'
         covid_vaccination
```

Out[10]:

Province/State	First Dose	Second Dose	Difference per region	Difference%
Anguilla	4931470	4709072	222398	4.722756
Bermuda	2817981	2690908	127073	4.722309
British Virgin Islands	5166303	4933315	232988	4.722747
Cayman Islands	3522476	3363624	158852	4.722644
Channel Islands	3287646	3139385	148261	4.722613
Falkland Islands (Malvinas)	3757307	3587869	169438	4.722525
Gibraltar	5870786	5606041	264745	4.722495
Isle of Man	4226984	4036345	190639	4.723060
Montserrat	5401128	5157560	243568	4.722543
Others	2583151	2466669	116482	4.722239
Saint Helena, Ascension and Tristan da Cunha	2348310	2242421	105889	4.722084
Turks and Caicos Islands	3052822	2915136	137686	4.723142