

LSE Data Analytics

Assignment 2

Data Analytics using Python

Overview

Scenario: Analysis of COVID-19 data (January 2020 – October 2021) for the UK Government's marketing campaigns to promote the vaccination.

Goal: Identify trends and patterns that can be used to inform UK Government's marketing approach to increase the number of fully vaccinated people.

Key questions:

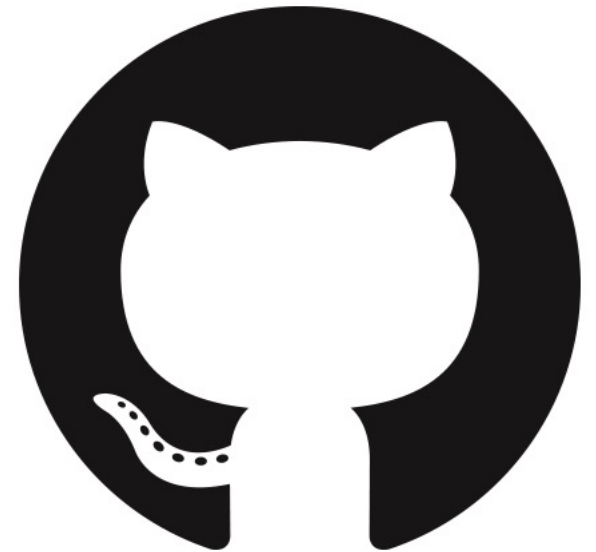
- What are the total vaccinations for a particular region?
- Where should the UK Government target the first marketing campaign(s)?
- What are the trending Twitter posts containing #coronavirus and #vaccinated hashtags?
- Which regions have experienced a peak in hospitalisation numbers? Are there regions that have not reached a peak yet?

1. Why GitHub?

GitHub is a *version control system (VCS)* which offers code hosting and repository services.

GitHub and other workflow tools can be used to:

- systematically manage changes in the code, programme or project;
- collaborate with other developers or as a team and handle complex projects across multiple time zones;
- establish a well-organised and reversible history to control changes over time;
- learn from other developers and build a professional portfolio as Data Analyst.



Source: <https://github.com/logos>

[Pull requests](#) [Issues](#) [Marketplace](#) [Explore](#)[ArSap7 / LSE_DA_COVID_analysis](#) Public[Pin](#)[Unwatch](#) 1[Fork](#) 0[Star](#) 0[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#) [Settings](#)[main](#) 1 branch 0 tags[Go to file](#)[Add file](#)[Code](#)

About



Analysis of the COVID-19 statistics to inform the UK government vaccination programme.

[Readme](#)

0 stars

1 watching








0 forks

Releases

No releases published

[Create a new release](#)**ArSap7** Add files via upload

49e8dac 1 hour ago 2 commits

| | | |
|---|----------------------|-------------|
|  .gitignore | Initial commit | 2 hours ago |
|  LSE_DA201_Assignment_template.... | Add files via upload | 1 hour ago |
|  LSE_DA201_Week_6_assignment.i... | Add files via upload | 1 hour ago |
|  README.md | Initial commit | 2 hours ago |
|  covid_19_uk_cases.csv | Add files via upload | 1 hour ago |
|  covid_19_uk_vaccinated.csv | Add files via upload | 1 hour ago |
|  tweets.csv | Add files via upload | 1 hour ago |

README.md

LSE_DA_COVID_analysis

Analysis of the COVID-19 statistics to inform the UK government vaccination programme.

The repository, or **GitHub repo**, is a folder where developers can store content related to the project in order to keep work organised.

2.1 Covid-19 Vaccination and Cases

General exploration process:

(1) Observe data types **info()**, records and numbers **describe()** and column features **value_count()**

(2) Identify missing data **isna()**

(3) Filter data for Gibraltar

```
cov[cov['Province/State']=='Gibraltar']
```

(4) Create Pandas DataFrame for Gibraltar

```
cov_Gibraltar = cov[cov['Province/State']=='Gibraltar']
```



Source: <https://commons.wikimedia.org/>

2.2 Covid-19 Vaccination and Cases (cont.)

General exploration process:

(5) Subset DataFrame for Gibraltar to explore particular columns

```
cov_Gibraltar_fcol = pd.read_csv('covid_19_uk_cases.csv',  
                                usecols=['Deaths', 'Cases', 'Recovered', 'Hospitalised'])
```

(6) Generate descriptive statistics

```
cov_Gibraltar_fcol.describe()
```



Source: <https://commons.wikimedia.org/>

2.3 Case Study: Gibraltar

To explore behaviour over time, a line graph could be added to visualise daily or/and monthly change in the number of “Deaths”, “Cases”, “Recovered” and “Hospitalised” individuals.

| | Province/State | Country/Region | Lat | Long | ISO 3166-1 Alpha 3-Codes | Sub-region Name | Intermediate Region Code | Date | Deaths | Cases | Recovered | Hospitalised |
|------|----------------|----------------|---------|---------|--------------------------|-----------------|--------------------------|------------|--------|--------|-----------|--------------|
| 3792 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2020-01-22 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3793 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2020-01-23 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3794 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2020-01-24 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3795 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2020-01-25 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3796 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2020-01-26 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4419 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2021-10-10 | 97.0 | 5626.0 | 0.0 | 858.0 |
| 4420 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2021-10-11 | 97.0 | 5655.0 | 0.0 | 876.0 |
| 4421 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2021-10-12 | 97.0 | 5682.0 | 0.0 | 876.0 |
| 4422 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2021-10-13 | 97.0 | 5707.0 | 0.0 | 0.0 |
| 4423 | Gibraltar | United Kingdom | 36.1408 | -5.3536 | GIB | Southern Europe | 0 | 2021-10-14 | 97.0 | 5727.0 | 0.0 | 0.0 |

2.4 Exploratory Analysis Good Practices

Key observations:

Total numbers vs trends

- Total numbers reveal only “frozen” patterns in data, while trends over time can offer additional insights that better explain observed behaviours.
- Decision-making should be based on the mixture of total numbers and analysis of past trends.

Data exploration

- **Data exploration** is an important stage in data analysis process which enables to summarise main characteristics of a dataset, identify initial correlations and irregularities, plot data to understand relationships and insights.
- Typical mistakes during the exploratory phase may include misunderstanding correlation and causation, focusing only in data and ignoring wider factors, disregarding missing and wrong values, etc.

3.1. Merging and Exploring Data

Futher exploration process:

(1) Merge “cov” and “vac” DataFrames

```
covid = pd.merge(cov, vac)
```

(2) Convert the data type of the Date column from object to DateTime

```
covid['Date'] = pd.to_datetime(covid['Date'])
```

(3) Clean up/drop unnecessary columns

```
covid_results = covid.drop(['Lat', 'Long', 'ISO 3166-1 Alpha 3-Codes', 'Sub-region Name', 'Intermediate Region Code'], axis = 1)
```

**General
exploration**



**Exploring data in the context of a
specific business question**

3.1. Merging and Exploring Data (cont.)

(4) Group data by “Province/State” and calculate the difference between “First Dose” and “Second Dose”.

```
covid_vaccination ['First Dose'] = covid_results.groupby('Province/State')['First Dose'].agg('sum')
covid_vaccination ['Second Dose'] = covid_results.groupby('Province/State')['Second Dose'].agg('sum')
covid_vaccination ['Difference per region'] = covid_vaccination ['First Dose'] - covid_vaccination ['Second Dose']
```

(5) Group data by “Province/State” and “Date” to calculate the difference between “First Dose” and “Second Dose” over time.

```
covid_vaccination ['First Dose'] = covid_results.groupby(['Province/State', 'Date'])['First Dose'].agg('sum')
covid_vaccination ['Second Dose'] = covid_results.groupby(['Province/State', 'Date'])['Second Dose'].agg('sum')
covid_vaccination ['Difference per region'] = covid_vaccination ['First Dose'] - covid_vaccination ['Second Dose']
```



3.2 Partially Vaccinated Individuals: Total Numbers vs Over Time

Difference per region (*Total Numbers*)

| 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 |

Difference per region (*Over Time*)

| | | First Dose | Second Dose | Difference per region |
|--------------------------|------------|------------|-------------|-----------------------|
| Province/State | Date | | | |
| Anguilla | 2020-01-22 | 0 | 0 | 0 |
| | 2020-01-23 | 0 | 0 | 0 |
| | 2020-01-24 | 0 | 0 | 0 |
| | 2020-01-25 | 0 | 0 | 0 |
| | 2020-01-26 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... |
| Turks and Caicos Islands | 2021-10-10 | 1437 | 1264 | 173 |
| | 2021-10-11 | 1896 | 1536 | 360 |
| | 2021-10-12 | 2395 | 1751 | 644 |
| | 2021-10-13 | 0 | 0 | 0 |
| | 2021-10-14 | 0 | 0 | 0 |

“Difference per region” is difference between “First Dose” and “Second Dose” columns which shows the total number of partially vaccinated individuals, i.e. people with the first shot only who are eligible for the second shot.

3.3 Initial Insights and Limitations

- **What insights can be gained from the data?**
 - Vaccination and Covid-19 statistics (i.e. deaths, cases, recoveries and hospitalisations) have increased over time for all regions.
 - There is no recorded data in the earlier stages of the pandemic.
 - For all regions, the total number of partially vaccinated individuals exceeded the total number of fully-vaccinated individuals.
- **Are there limitations or assumptions that needs to be considered?**
 - 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.
 - Total numbers and trends over time may yield conflicting insights which should be investigated further by visualising and comparing the results.
 - In analysing the relationship between vaccination and Covid-19 statistics, correlation may not necessarily imply causation.

4.1 Further Insights and Visualisations

- **What insights can be gained from the data?**

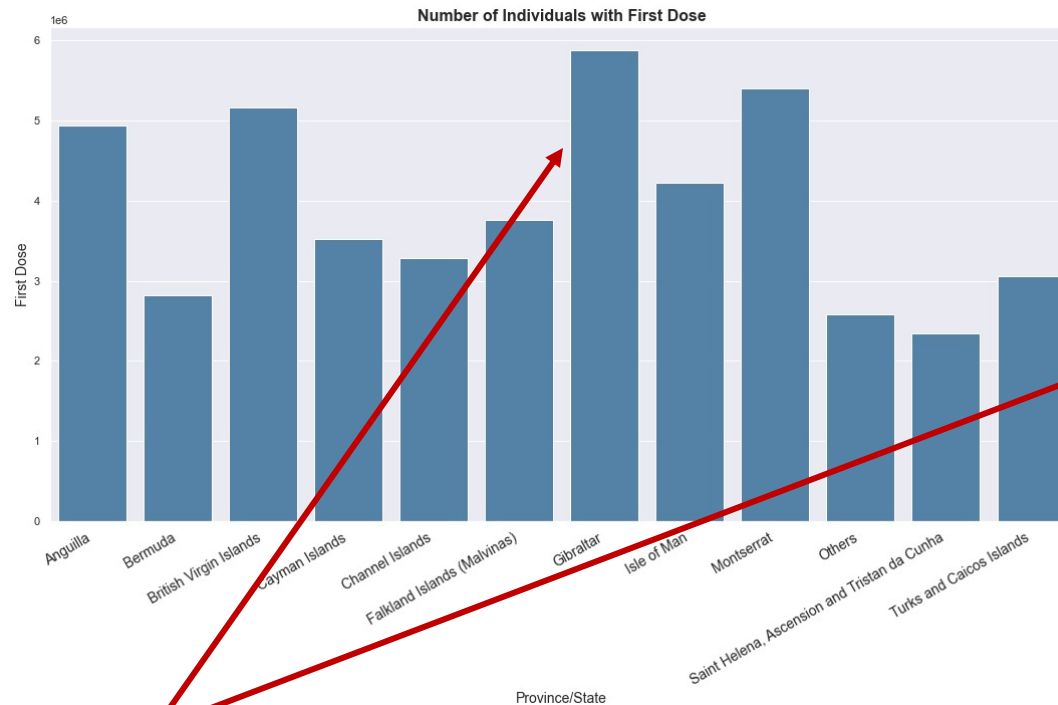
- Total number of vaccinations, deaths and recoveries across regions.
- Peaks and declines in vaccinations, deaths and recoveries over time.
- Comparison of daily and monthly observations.

- **Why do we need to consider other features?**

Apart from the vaccination statistics, other features such as the number of deceased and recovered individuals, could unfold wider trends and patterns in the data and thus help Data Analyst make more informed decisions.

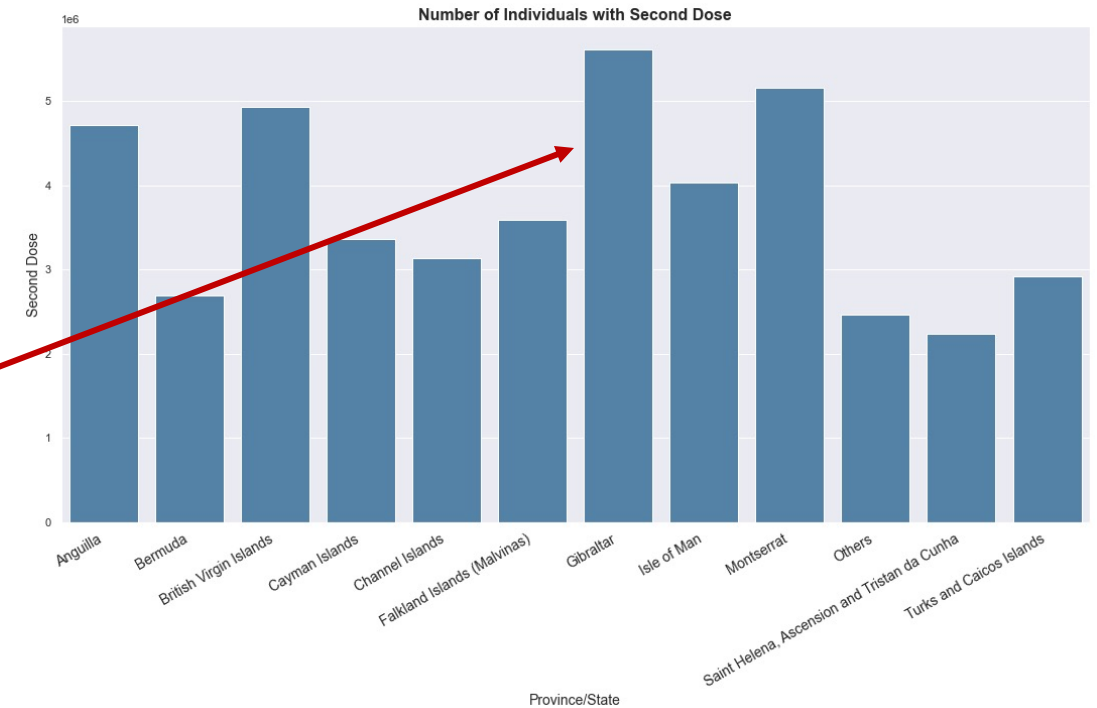
4.2 First Dose and Second Dose Comparison

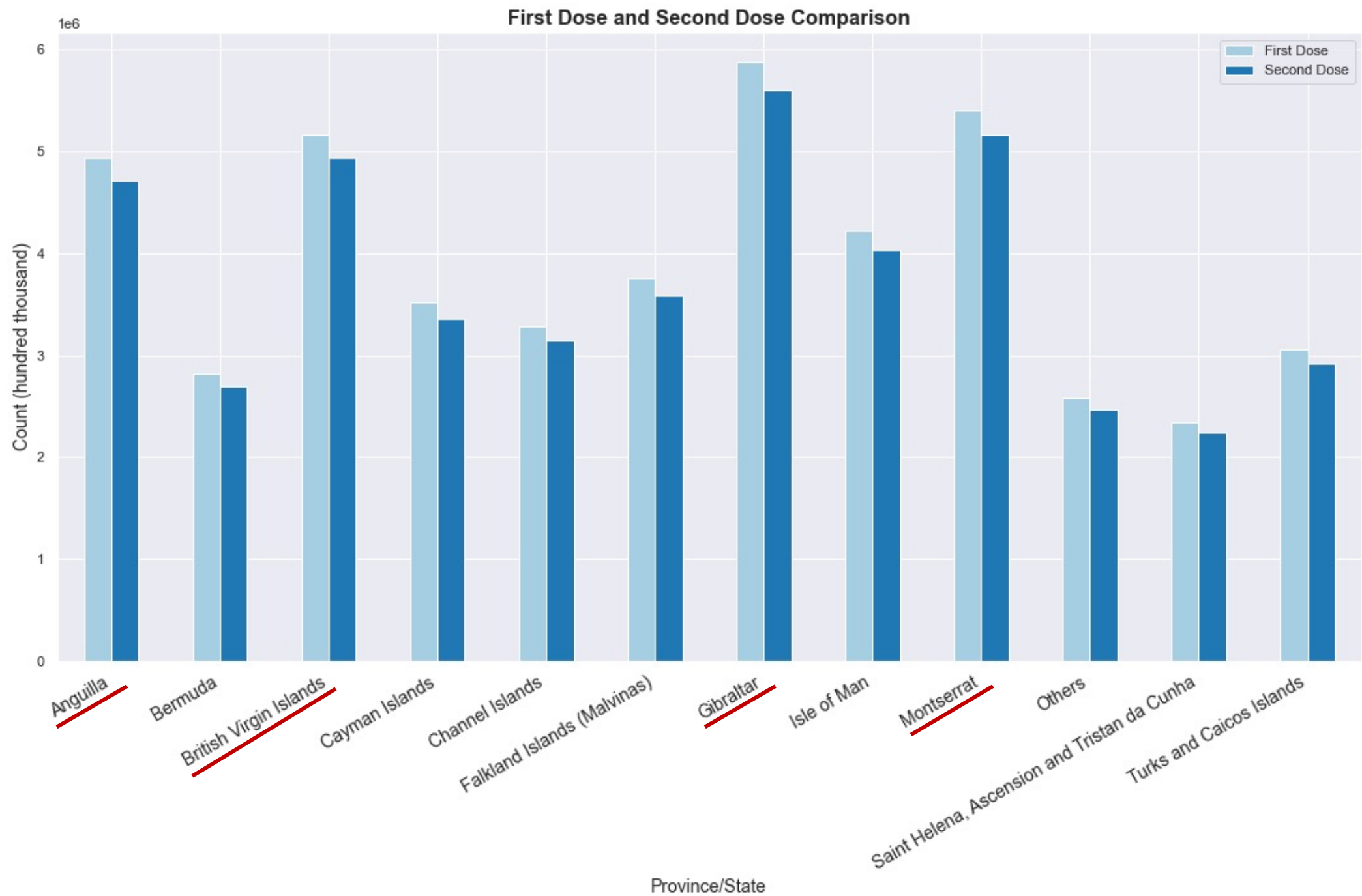
First Dose (*hundre thousand*)



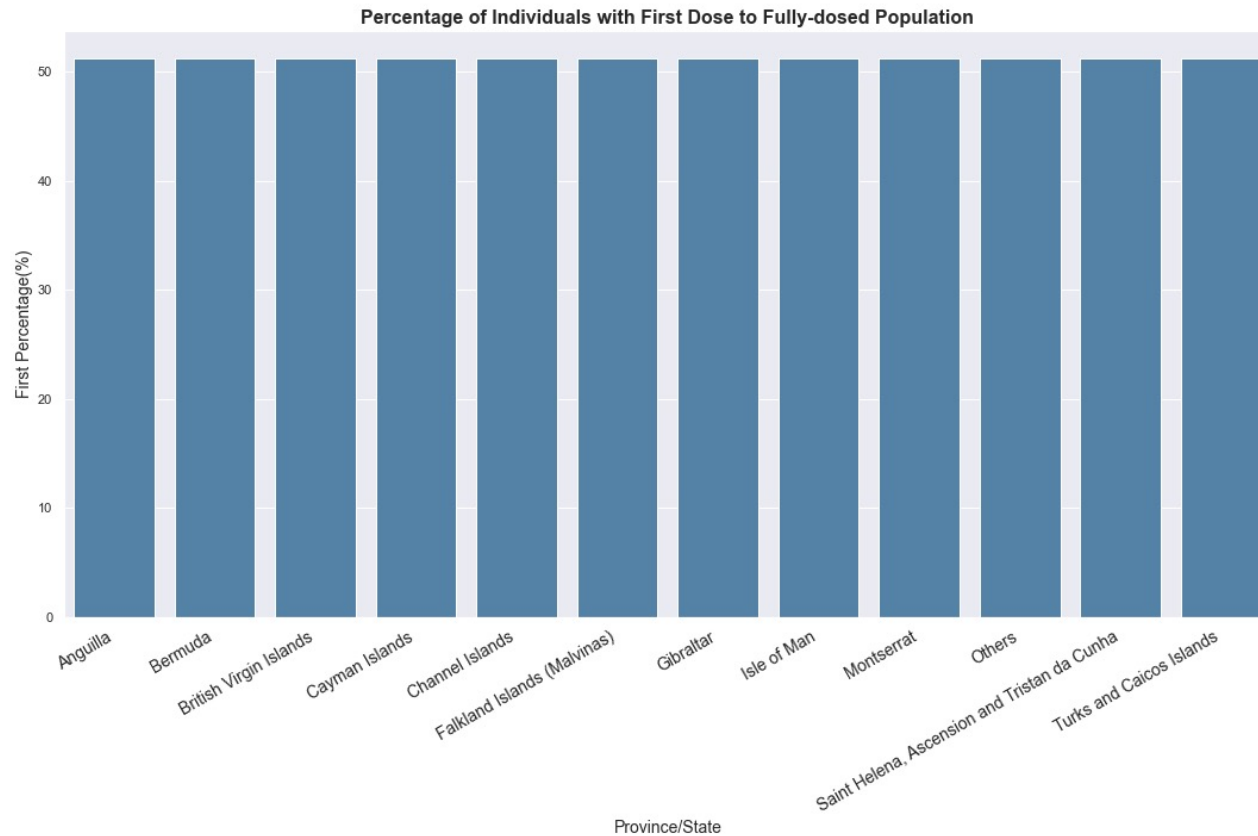
Gibraltar

Second Dose (*hundre thousand*)





4.3 First Dose to Second Dose Ratio

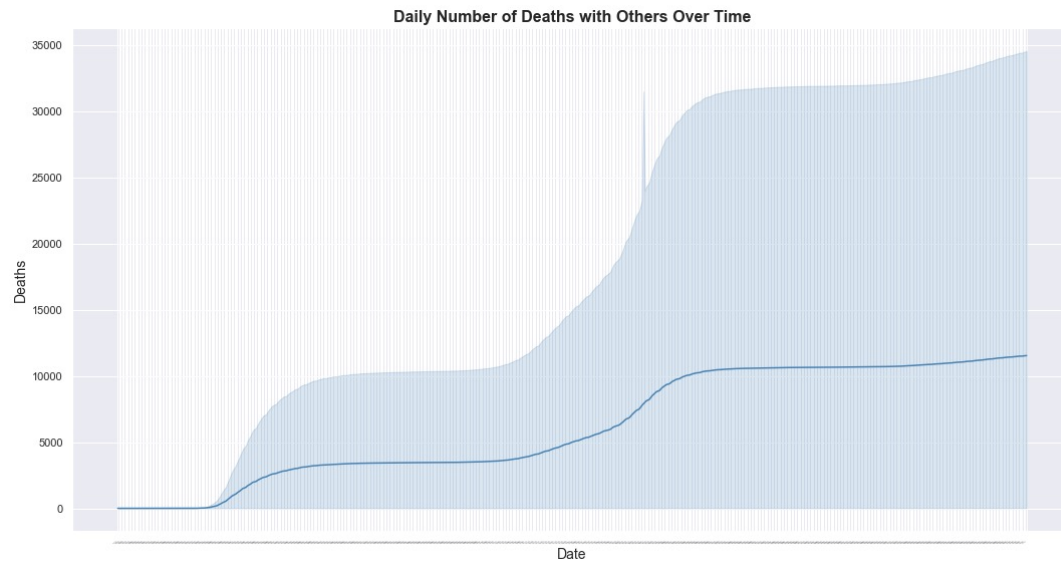


$$\text{First Percentage} = \frac{\text{First Dose}}{\text{Second Dose}}$$

$\approx 51\%$

4.4 Deaths and Recoveries (cont.)

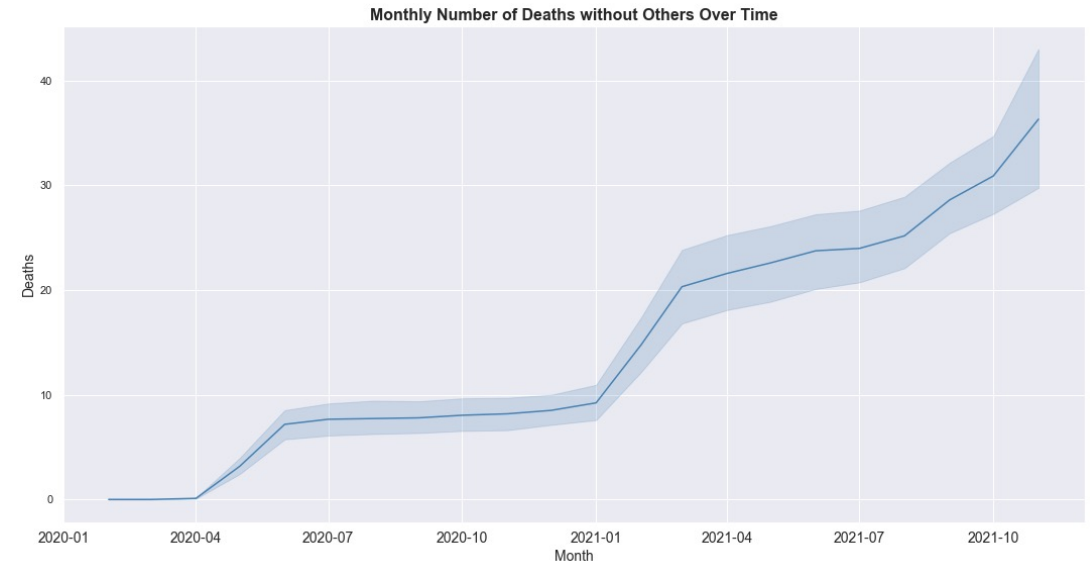
Daily Number of Deaths



DAY



Monthly Number of Deaths

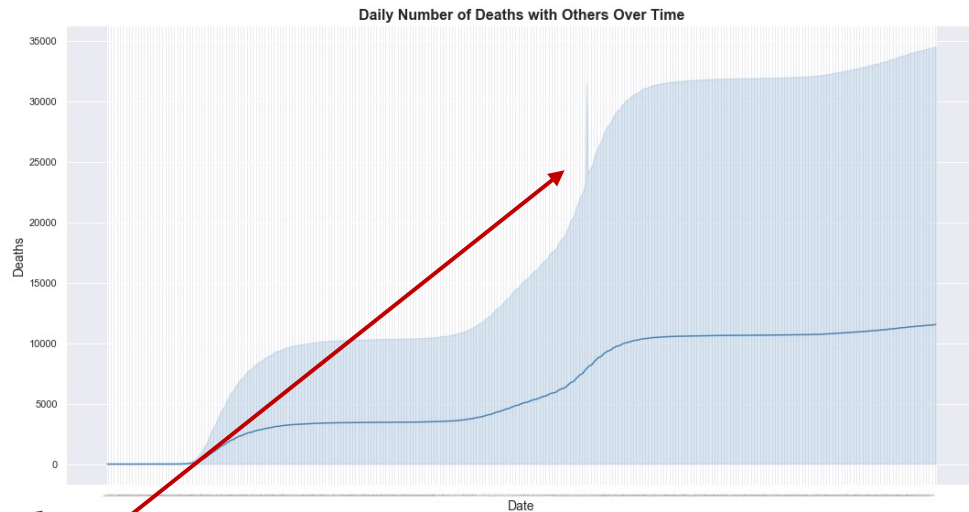


MONTH

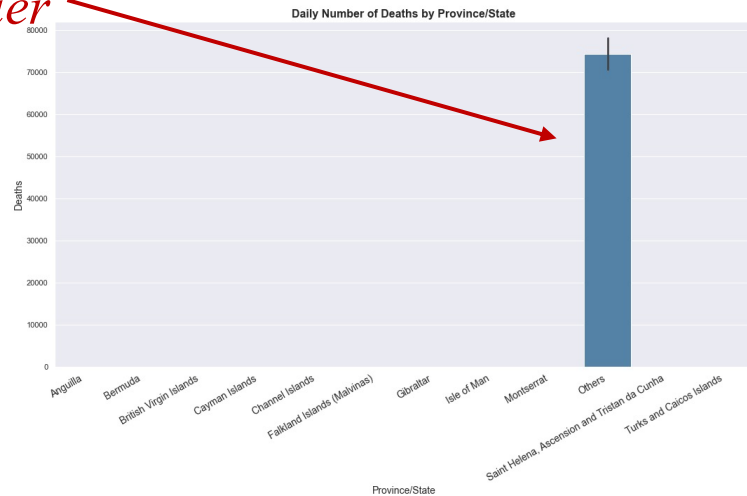


4.5. Removing Outliers

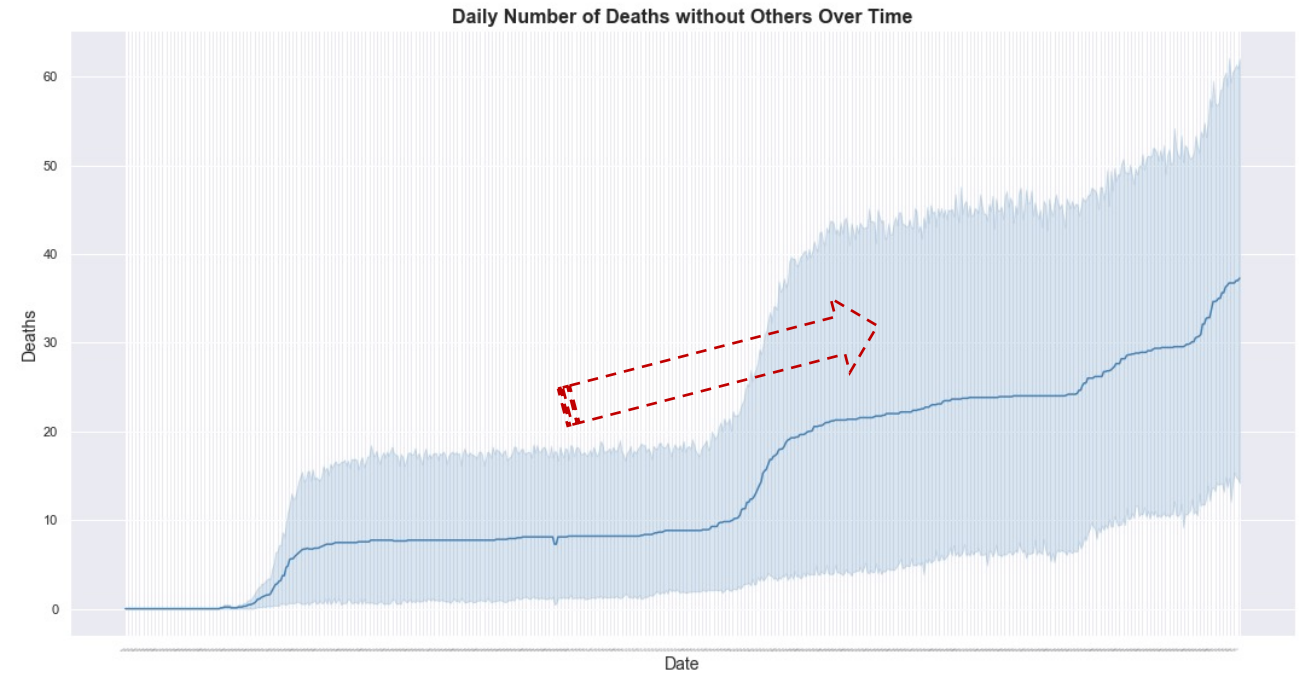
Before removing “Others”



outlier

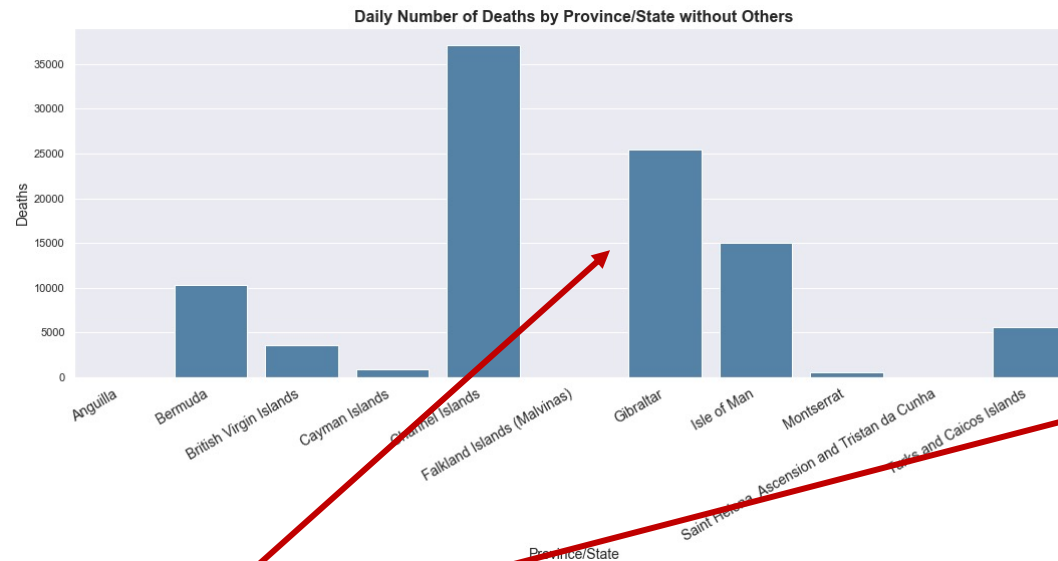


After removing “Others”

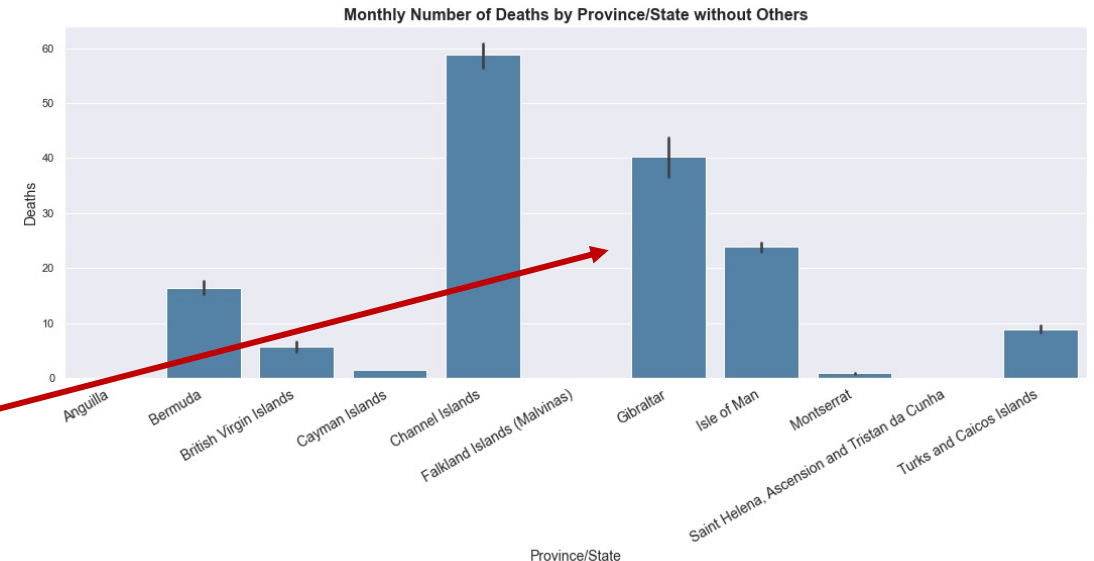


4.4 Deaths and Recoveries (cont.)

Daily Number of Deaths



Monthly Number of Deaths



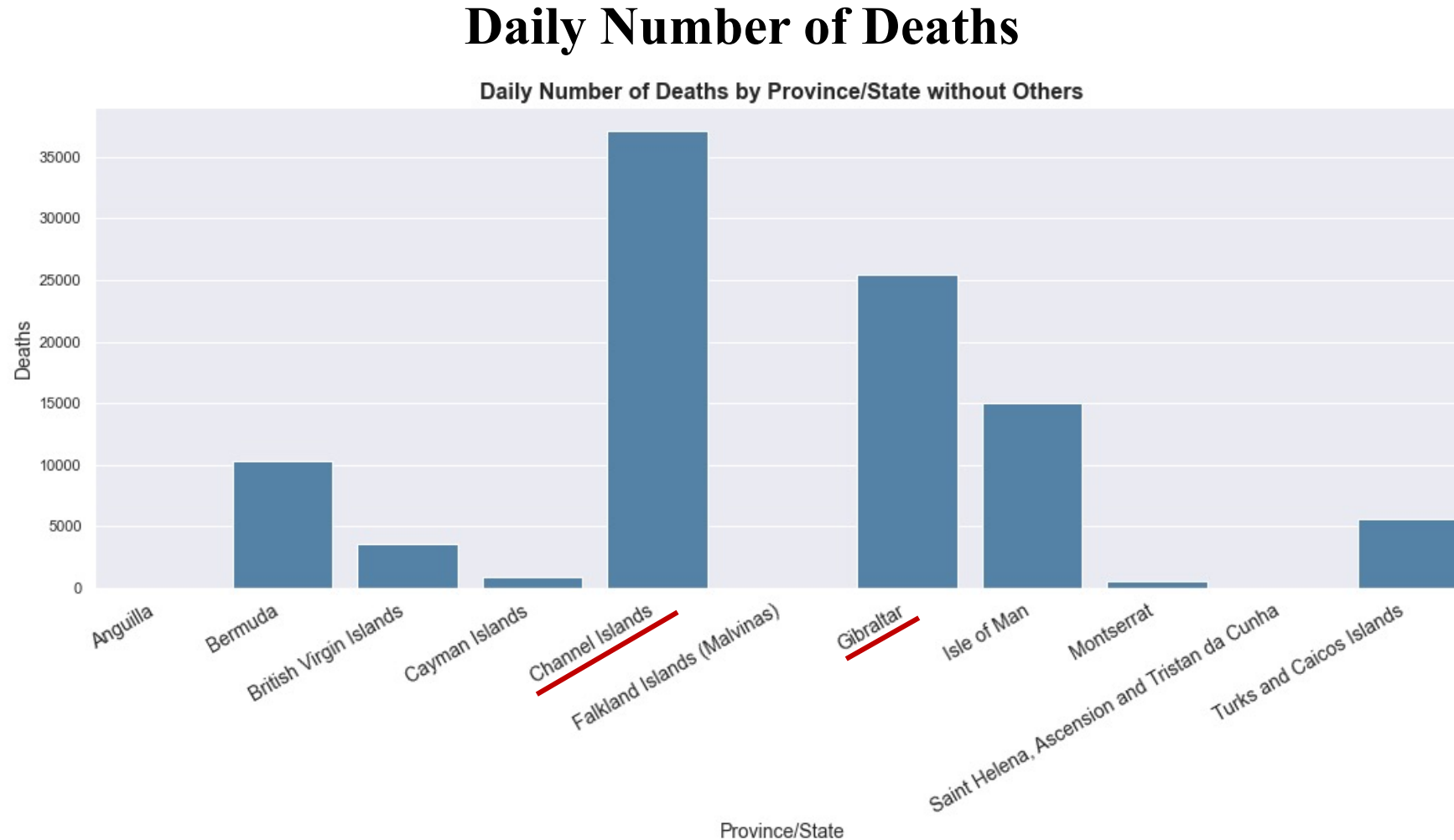
Gibraltar

DAY



MONTH

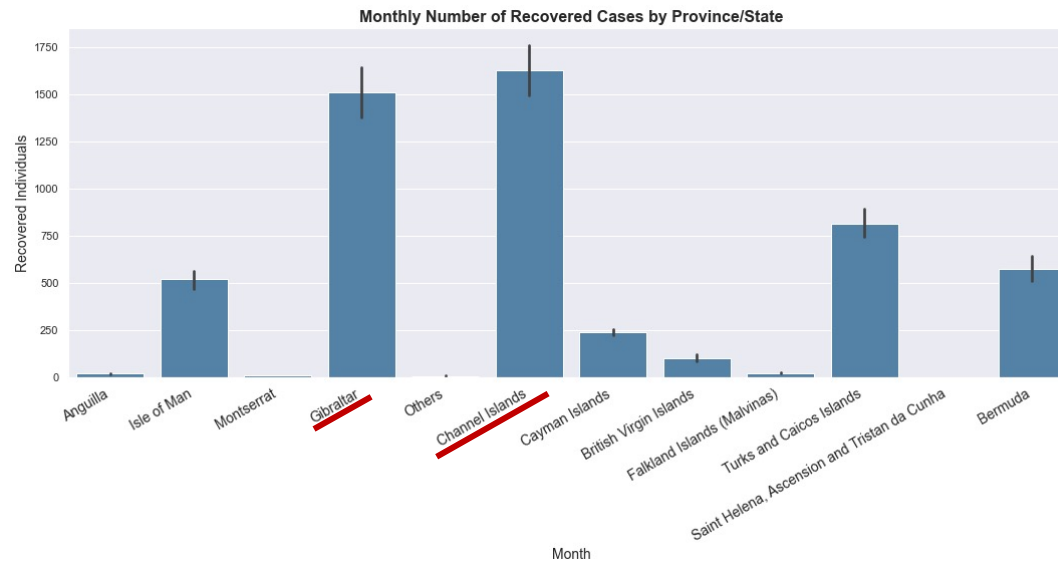
4.4 Highest Daily Number of Deaths by Province/State



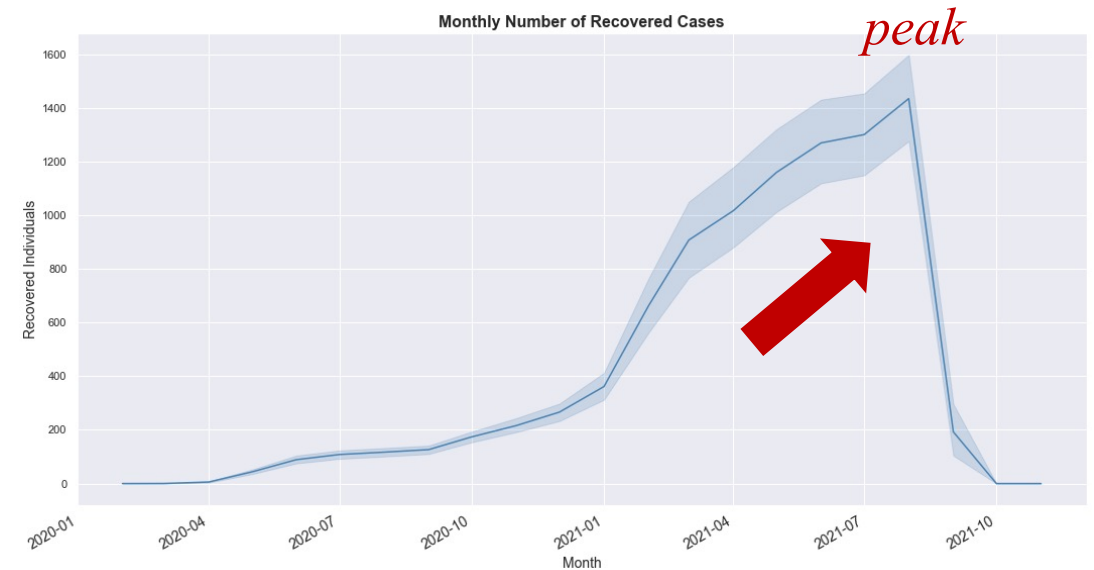


4.4 Deaths and Recoveries (cont.)

Monthly Number of Recoveries
(Total Numbers)



Monthly Number of Recoveries
(Over Time)



4.5 Data Analysis and Visualisation Good Practices

- **Why it is important to explore data and use different views?**

Data exploration can provide meaningful insights that explain the underlying causes behind the observed behaviours. By relying on a broad arrange of views, as well as evaluating a different variables and features, Data Analyst can improve the overall quality and effectiveness of decision-making.

Key suggestions:

1. Critically reflect on the insights and assumptions discovered during the earlier stages of data analysis.
2. Choose the right visualisation tool that will help you plot and identify initial insights.
3. Continuously review your analytical model and methods to better answer the question posed by a specific business problem.

5.1 External Data

- **What insights can be gained from the data?**

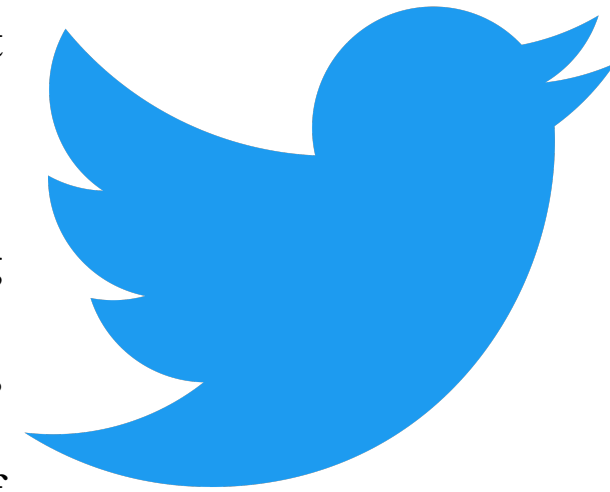
External data could be used to obtain information from various online sources and collect them to generate additional insight for an organisation.

- **What are the advantages and disadvantages of using external data?**

- External data may offer a nuanced solution to a business problem by outsourcing the specific information about certain topic in a certain period of time.
- It might be daunting in terms of technical considerations to analyse and harness extensive amounts of external data.
- Ethical concerns may arise in regards to privacy regulations and application of data scraping methods.

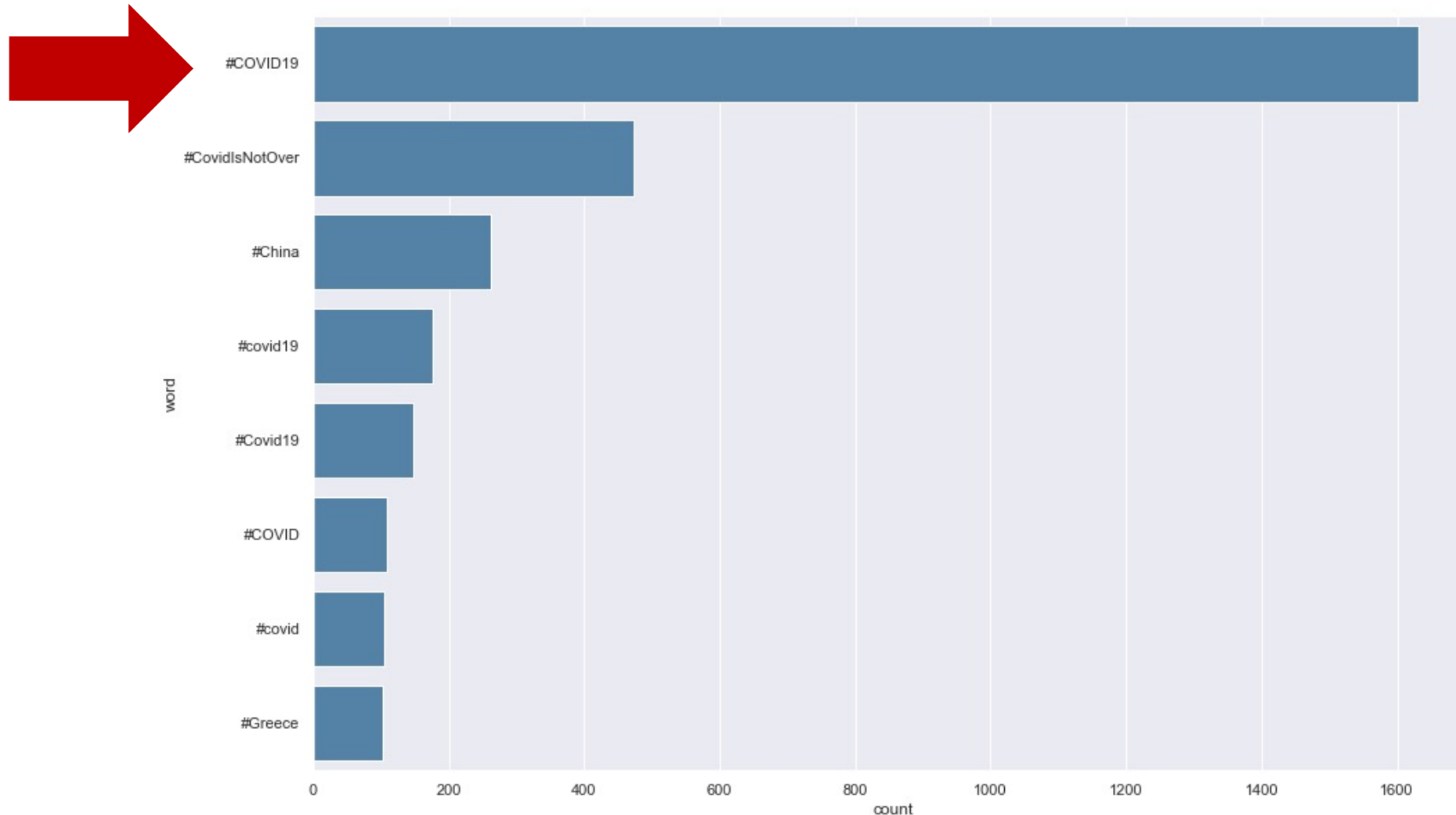
- **How would you suggest using external data in the project?**

The project may utilise Twitter-sourced data using APIs to evaluate the public awareness about the Covid-19 pandemic, determine the trending and widely discussed matters and gather popular sentiment on the UK government's vaccination campaign.



Source: <https://commons.wikimedia.org/>

5.2 Case study: #COVID19 Twitter Hashtags



6.1 Time-series Forecasting

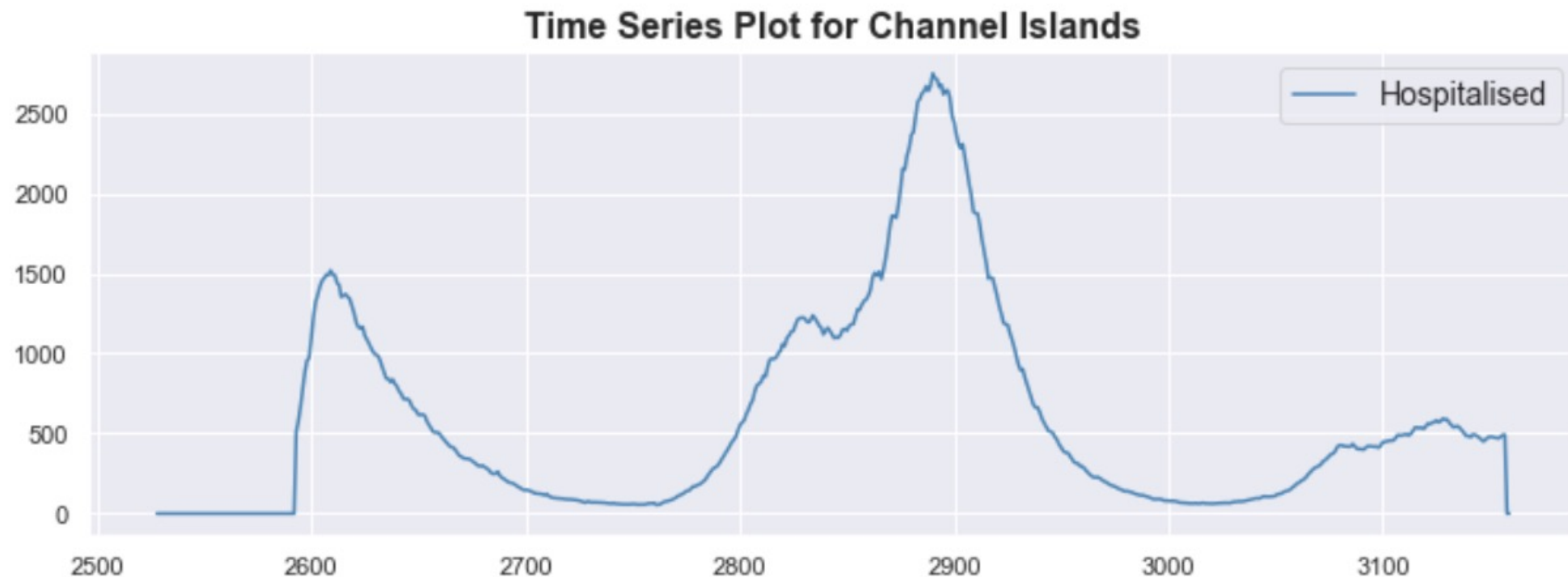
- **Time-series forecasting** is used to make short-to-medium-term forecasts to identify patterns, trends and any irregularity in data series.
 - Future predictions are based on the analysis of the existing time-series data which consist of comparable measurements recorded on a single variable over time.
- One of the common techniques used in forecasting is the **moving average** technique which eliminates seasonal effects and irregularities in data.
 - Moving average uses the weighted average of a number of consecutive points in the series.



6.2 Case Study: Hospitalisation Surge in the Channel Islands

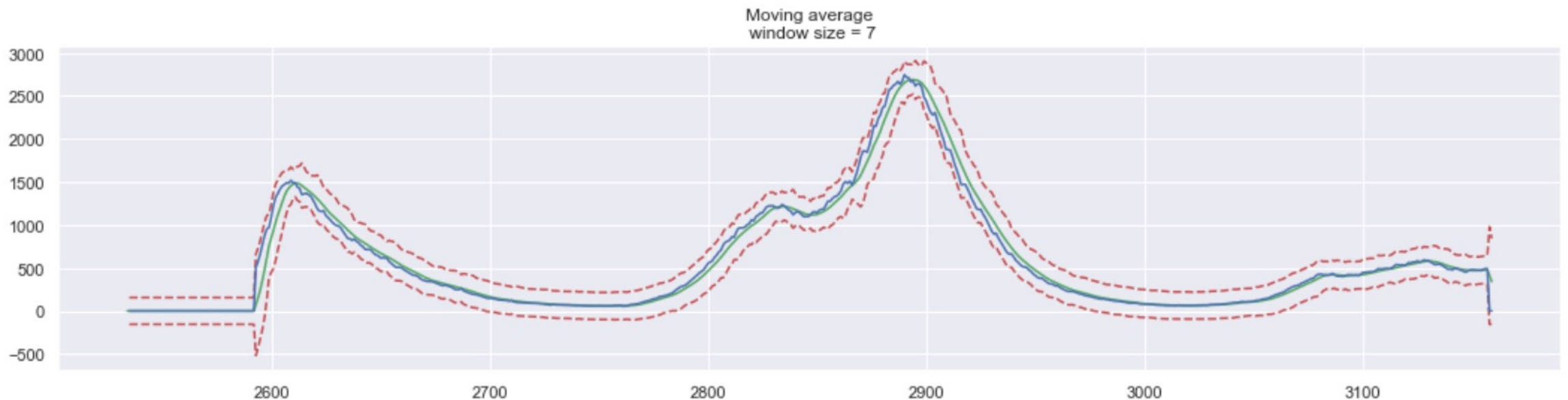


Time-series Analysis



6.2 Case Study: Hospitalisation Surge in the Channel Islands (cont.)

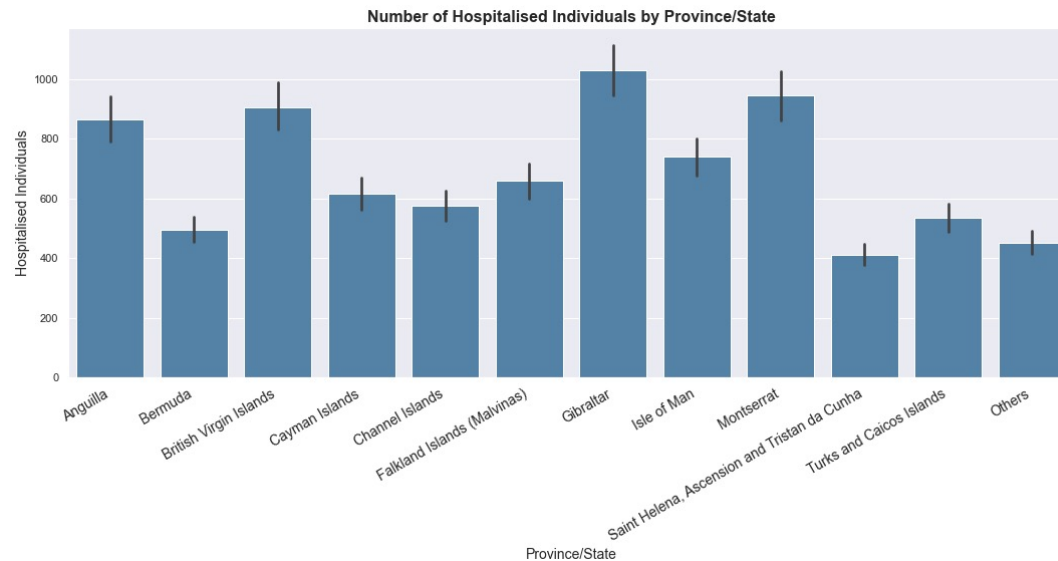
Moving Average



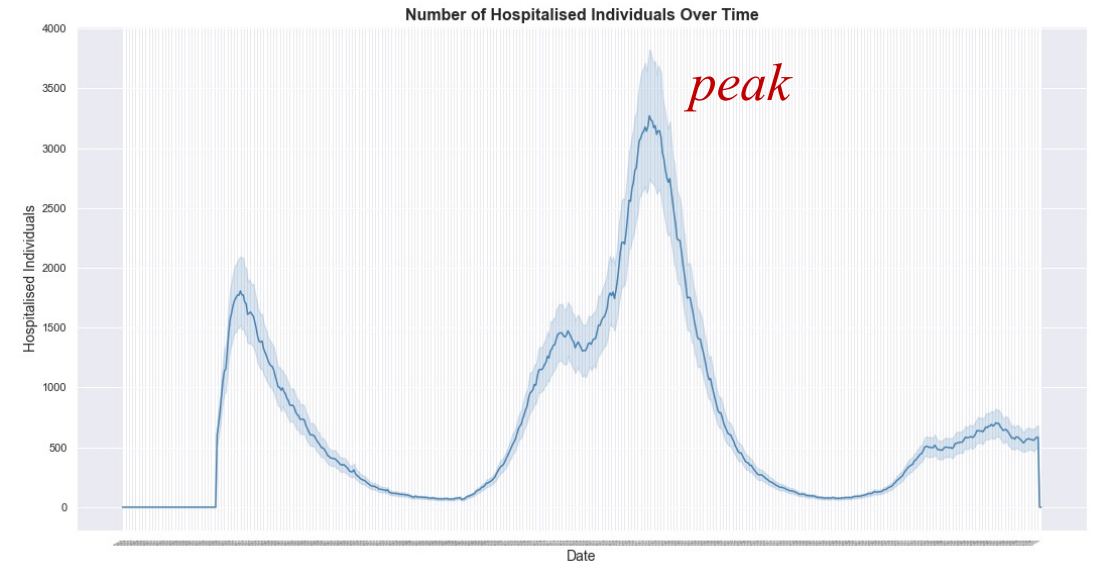


6.3 Surge in Hospitalisation Numbers

Hospitalised Individuals (*per Province/State*)



Hospitalised Individuals (*Over Time*)



6.4 UK Government Q&A

- **Question 1:** We have heard of both qualitative and quantitative data from the previous consultant. What are the differences between the two? Should we use only one or both of these types of data and why? How can these be used in business predictions? Could you provide examples of each?
- **Question 2:** We have also heard a bit about the need for continuous improvement. Why should this be implemented, it seems like a waste of time. Why can't we just implement the current project as it stands and move on to other pressing matters?
- **Question 3:** As a government, we adhere to all data protection requirements and have good governance in place. We only work with aggregated data and therefore will not expose any personal details. Have we covered everything from a data ethics standpoint? Is there anything else we need to implement from a data ethics perspective?

Question 1: We have heard of both qualitative and quantitative data from the previous consultant. What are the differences between the two? Should we use only one or both of these types of data and why? How can these be used in business predictions? Could you provide examples of each?

- ***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 2: We have also heard a bit about the need for continuous improvement. Why should this be implemented, it seems like a waste of time. Why can't we just implement the current project as it stands 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.
- ***Continuous improvement*** yields overall **more efficient processes, higher quality and a reduction in wasted time and effort**. 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: As a government, we adhere to all data protection requirements and have good governance in place. We only work with aggregated data and therefore will not expose any personal details. Have we covered everything from a data ethics standpoint? Is there anything else we need to implement from a data ethics perspective?

- 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.



Debrief

- What are the total vaccinations for a particular region?

Total numbers of vaccinated individuals are reflected in the “First Dose” and “Second Dose” columns.

- Where should the UK Government target the first marketing campaign(s)?
 - *Area(s) with the largest number of people who have received a first dose but no second dose.*

Anguilla, British Virgin Islands, Gibraltar and Montserrat have the highest “Difference per region”.

Recommendation 1: To prioritise vaccination campaign(s) in these areas to increase the total number of fully-vaccinated individuals.

| 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 |



Debrief (cont.)

- *Area(s) with the greatest number of recoveries.*

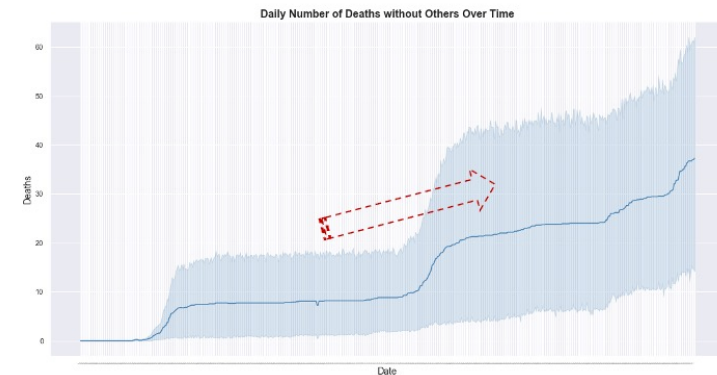
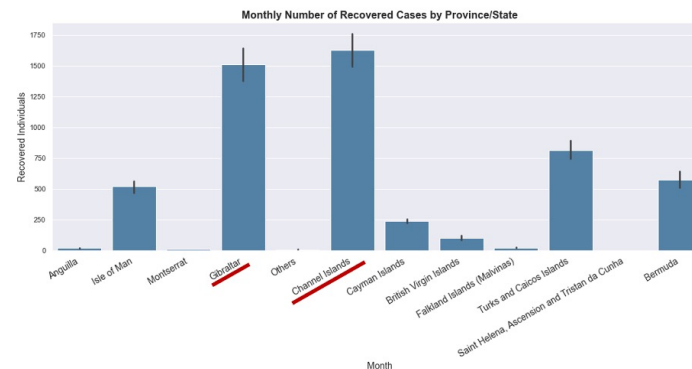
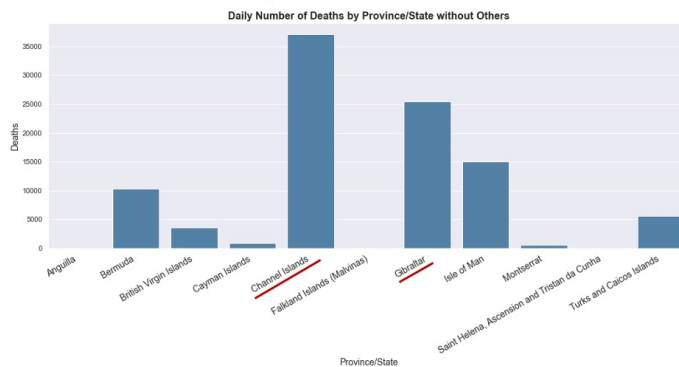
Gibraltar and the Channel Islands have the greatest number of recovered individuals, as opposed to relatively insignificant recoveries in other regions.

Recommendation 2: To avoid these area in the initial vaccination campaign runs that promote Second Dose.

- *Have deaths been increasing across all regions over time or has a peak been reached?*

Aggregated number of deaths has been increasing over time. Deaths have not reached a peak yet, unlike the aggregated number of recoveries across all regions.

Recommendation 3: To target Gibraltar and the Channel Islands as these two regions have the greatest number of daily and monthly deaths, which indicates a high rate of infected individuals.





Debrief (cont.)

- **What are the trending Twitter posts containing #coronavirus and #vaccinated hashtags?**

The top trending hashtags related to Covid-19 pandemic were: #COVID19, #CovidIsNotOver, #covid19, #Covid19, #COVID, #covid.

- **Which regions have experienced a peak in hospitalisation numbers? Are there regions that have not reached a peak yet?**

Channel Islands have reached a peak in the number of hospitalised individuals. Anguilla, British Virgin Islands, Gibraltar and Montserrat have had the highest hospitalisation numbers compared to other regions. The plotting of hospitalisations over time reveals that all regions have reached a peak.

