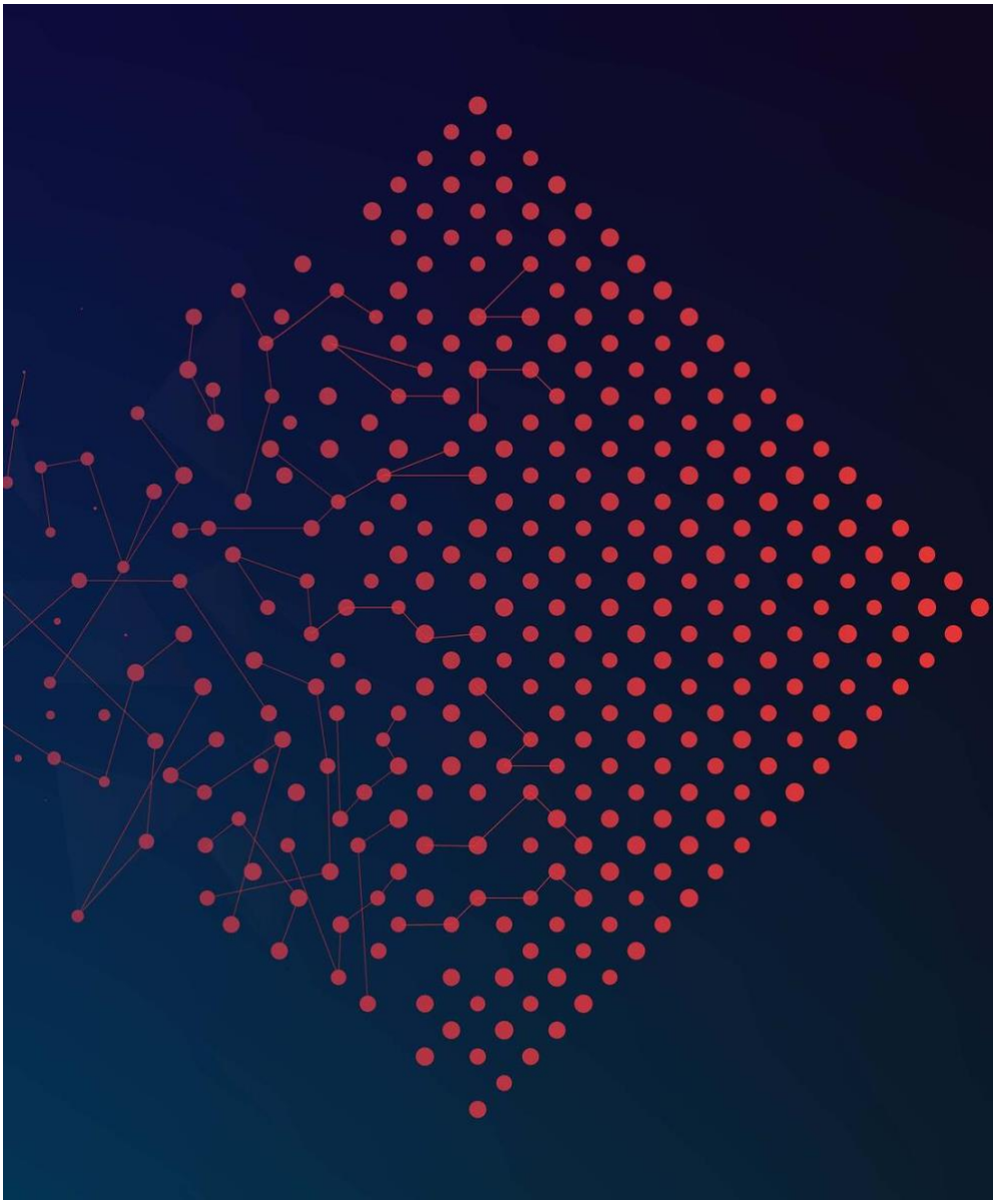
ASSIGNMENT 2:
COVID-19 UK GOVERNMENT PROJECT
JULY 2022



Javier Conde Pascual

LSE Data Analytics Career Accelerator
London School of Economics and Political Science

I, Javier Conde Pascual, certify that this is an original piece of work. I have acknowledged all sources and citations. No section of this assignment has been plagiarized.

TABLE OF CONTENTS

Chapter 1. Scenario and system set up 1

Chapter 2. Analysis approach 3

2.1 First approach and insights to provided data 4

2.2 Vaccination - areas to target marketing campaign and first approach to visualisations 7

2.3 Visualisations exploration 10

2.4 Twitter data relating to the #coronavirus hashtag exploration 12

2.5 Time-series analysis and data smoothing 14

Chapter 3. Recommendations 17

Reference list 19

List of figures 20

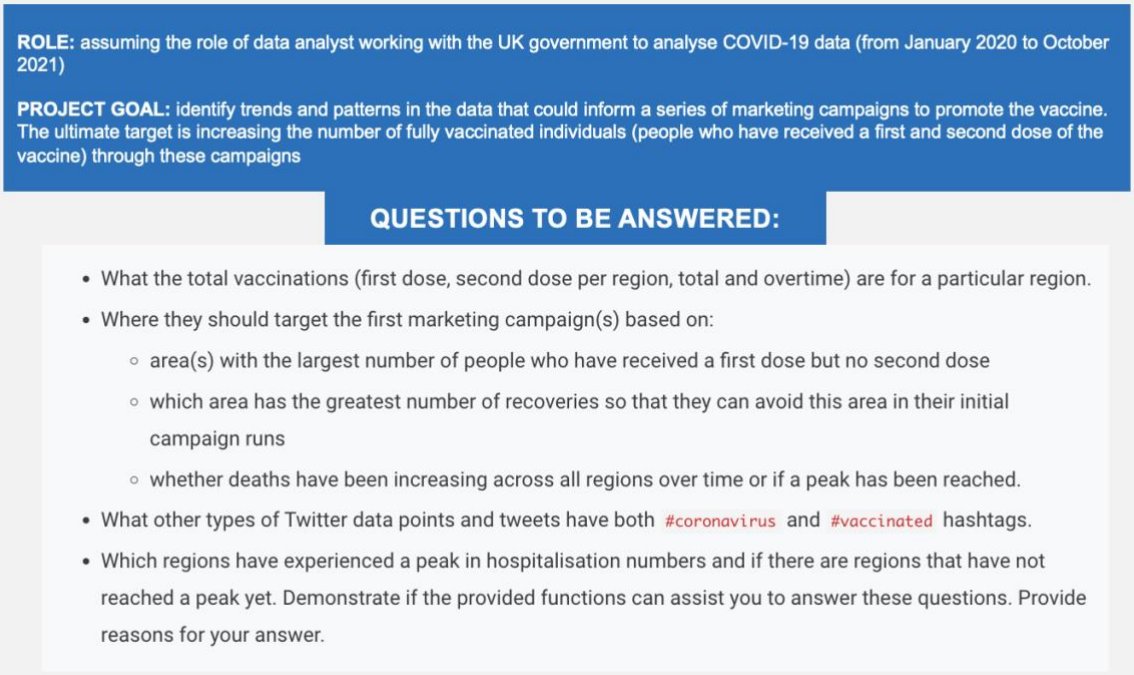
Word count: 1061
Own words: 1061

Chapter 1

Scenario and system set up

Background, scenario, and context provided beforehand for this report are available here. The business problem and key questions for the analyst team and the data are also provided, Figure 1.1. below. This helps also formulating the data analytics problem, building a process to follow accordingly.

Figure 1.1: Background, project goal and questions to answer



Source: LSE (2022)

All files generated for this project are stores in GitHub, a distributed centralised version system control (CVS). These tools can add solid and measurable value to organisations. Having reliable repositories to control the different versions and updates to the code is essential in the day-to-day work of the team.

Github will help to control overall project progression, different steps we are accomplising along the way and any changes we make to the original Python code block.

Now once we have prepared the central Python repository (a Jupiter Notebook) in Github where the project will be coded, let’s study the analytical approach used.

Chapter 2

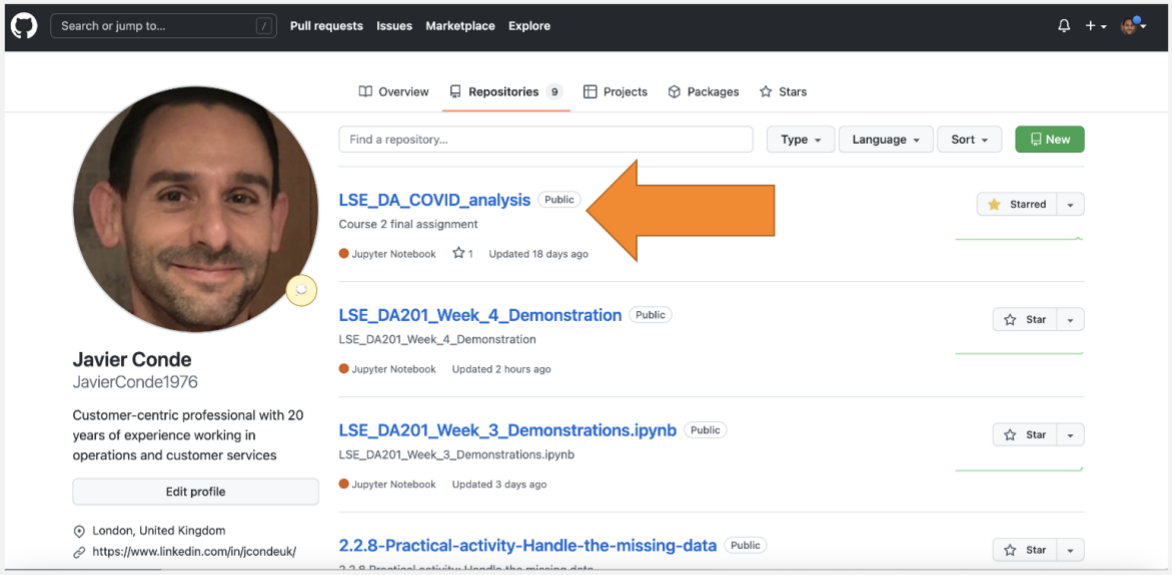
Analytical approach

This approach has four key points:

1. Bringing stakeholders and team together through understanding the importance and risk-reducing advantages of methodical and thorough data analysis and data quality processes.
2. Provide stakeholders with rigorous, accurate and realistic analysis on the data provided through Python and answer all their questions.
3. Assessing the quality of the data and results obtained
4. All visualizations to be created as accessible as possible through very legible fonts (Arial), big font sizes (10-12 onwards when possible), clear head titles and colors that are color-blind and color-contrast friendly among other features.

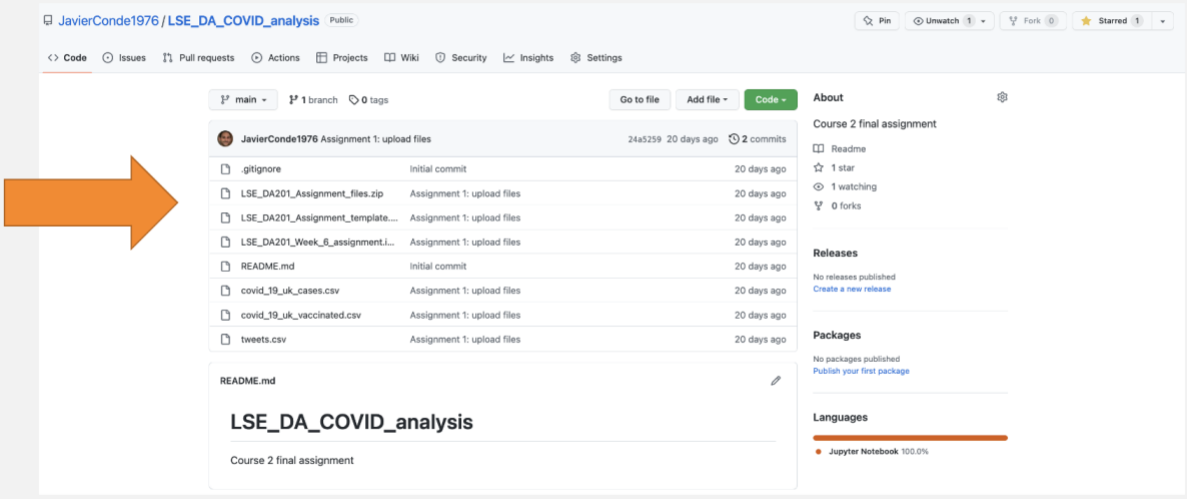
File preparation for Github and Phyton (Anaconda/Jupyter environment) is completed (Figures 1.2 and 1.3 below). Let's now analyze and visualize the data provided with Python.

Figure 1.2: Project Github repository (I)



Source: Javier Conde (2022)

Figure 1.3: Project Github repository (II)



Source: Javier Conde (2022)

2.1 First approach and insights to provided data

Two files provided analyzed for insights:

1. covid_19_uk_cases.csv (COVID cases in the UK in per province)
2. covid_19_uk_vaccinated.csv (vaccination data)

Let's import them to the Jupyter project Notebook, along with the appropriate libraries needed for statistical analysis and visualization.

Figure 2.1: Overview of COVID cases x UK province

```
In [250]: # Being cumulative, we can group by province to get an overview
          cov.groupby('Province/State')[['Deaths', 'Cases', 'Recovered', 'Hospitalised']].max()
```

Out[250]:

Province/State	Deaths	Cases	Recovered	Hospitalised
Anguilla	1.0	644.0	111.0	4122.0
Bermuda	95.0	5548.0	2503.0	2355.0
British Virgin Islands	37.0	2725.0	1914.0	4318.0
Cayman Islands	2.0	1011.0	635.0	2944.0
Channel Islands	100.0	12135.0	8322.0	2748.0
Falkland Islands (Malvinas)	0.0	69.0	63.0	3140.0
Gibraltar	97.0	5727.0	4670.0	4907.0
Isle of Man	54.0	8343.0	4019.0	3533.0
Montserrat	1.0	41.0	19.0	4514.0
Others	138237.0	8317439.0	344.0	2159.0
Saint Helena, Ascension and Tristan da Cunha	1.0	4.0	4.0	1963.0
Turks and Caicos Islands	23.0	2910.0	2433.0	2552.0

Source: Javier Conde (2022)

Figure 2.2: Overview of vaccinations x UK province

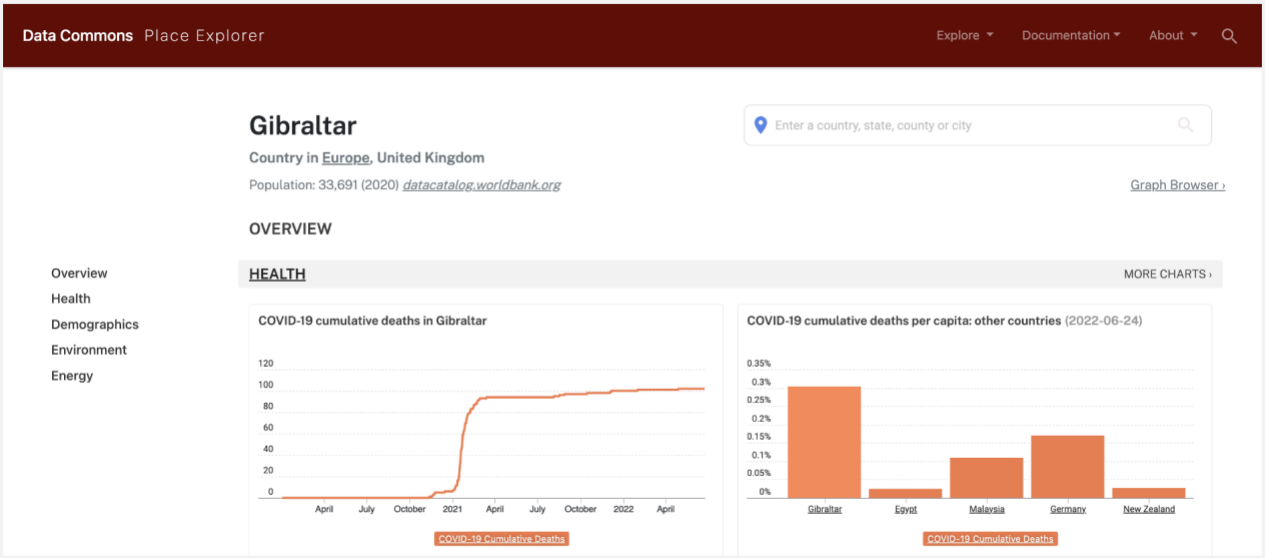
```
In [243]: # Being non cumulative, we can group by province to get an overview
          vac.groupby('Province/State')[['Vaccinated', 'First Dose', 'Second Dose']].sum()
```

Out[243]:

Province/State	Vaccinated	First Dose	Second Dose
Anguilla	4709072	4931470	4709072
Bermuda	2690908	2817981	2690908
British Virgin Islands	4933315	5166303	4933315
Cayman Islands	3363624	3522476	3363624
Channel Islands	3139385	3287646	3139385
Falkland Islands (Malvinas)	3587869	3757307	3587869
Gibraltar	5606041	5870786	5606041
Isle of Man	4036345	4226984	4036345
Montserrat	5157560	5401128	5157560
Others	2466669	2583151	2466669
Saint Helena, Ascension and Tristan da Cunha	2242421	2348310	2242421
Turks and Caicos Islands	2915136	3052822	2915136

Source: Javier Conde (2022)

Figure 2.3: Cumulative deaths reported in UK province Gibraltar



Source: DataCommons.org (2022)

Figure 2.4: Overview of total country vaccinations

```
In [434]: # We add a 'column totals' row

          vac_overview_vacc = vac.groupby('Province/State')[['Vaccinated']].sum().\
          sort_values(by=['Vaccinated'], axis=0, ascending=False)
          vac_overview_vacc
          print(vac_overview_vacc.sum())

Vaccinated    44848345
dtype: int64

In [435]: vac_overview_1dose = vac.groupby('Province/State')[['First Dose']].sum().\
          sort_values(by=['First Dose'], axis=0, ascending=False)
          vac_overview_1dose
          print(vac_overview_1dose.sum())

First Dose    46966364
dtype: int64
```

Source: Javier Conde (2022)

Figure 2.5: First findings

FINDINGS FROM FILE covid_19_uk_cases.csv

- Data cumulative already so no need for extra cumulative calculations
- 1st case reported in Gibraltar 2020-03-03
- Data available only until 14 Oct 2021
- Peak in year 2020: 31 December 2040 reported cases; peak 2021: increase of 280% with 5727 reported cases on 14 October (Figure 2.1)
- Others' region may be outlier on 'Deaths' (138,237) and 'Cases' (8,317,439) columns as very out of range with rest of the regions

FINDINGS FROM FILE covid_19_uk_vaccinated.csv

- First vaccinations on 2021-01-11
- Data quality questions as possible discrepancies on vaccination info vs real population uncovered: i.e., Gibraltar, 5,606,041 complete vaccination cycle (CVC) on total population of Gibraltar of 33,691 (Figure 2.2) (Source: https://datatopics.worldbank.org/). Cumulative deaths in line with real data (Figure 2.3) (Source: DataCommons.org)
- CVC percentage (95.5%) consistent across all United Kingdom provinces

ADVICE FOR TEAM & STAKEHOLDERS: Always important to keep an eye on data quality, understanding data shape & data types at all times and looking for:

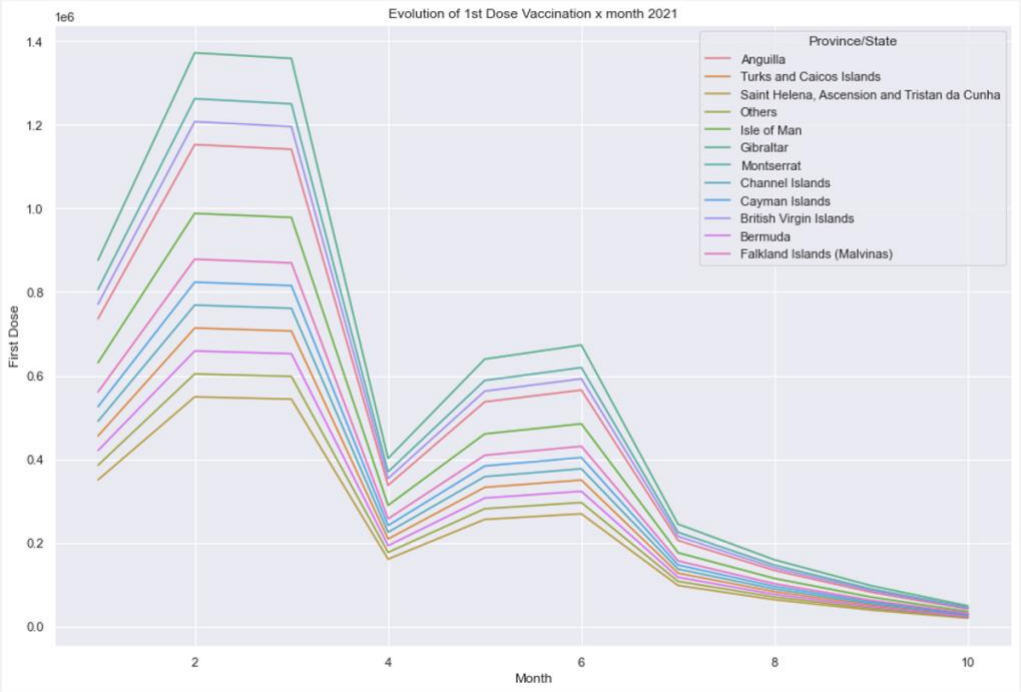
1. Data types
2. Inconsistencies in column names
3. Possible duplicate entries
4. Inconsistent data entry & spelling
5. Start looking for possible outliers with the .describe() method)
6. Ways to improve code quality (PEP8 Style Guide)

Source: Javier Conde (2022)

2.2 Vaccination - areas to target marketing campaign and first approach to visualisations

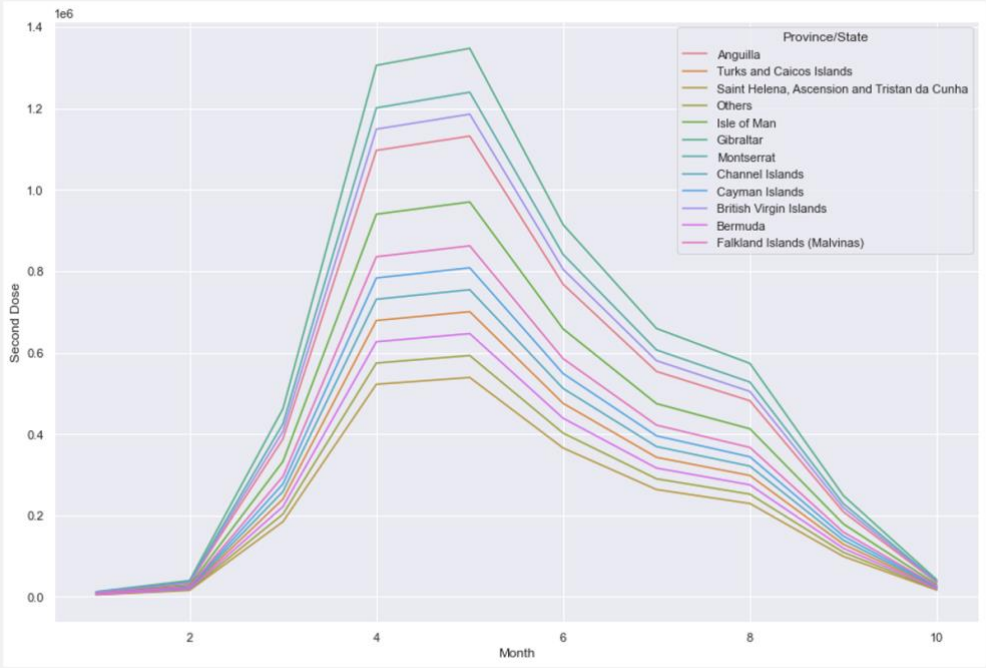
Let's now visualize the evolution of vaccinations with time per region to understand where we should target a possible marketing campaign. We drop 'Others' as per our previous findings (Figure 2.5).

Figure 2.6: Evolution of 1st Dose Vaccination x month 2021



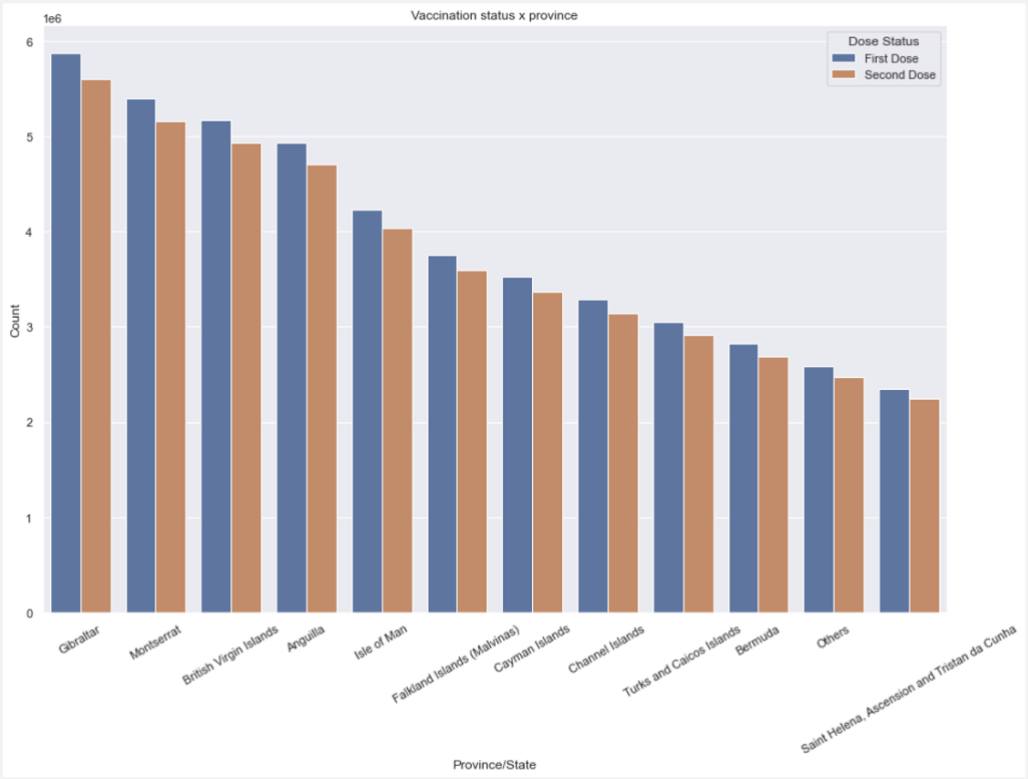
Source: Javier Conde (2022)

Figure 2.7: Evolution of 2nd Dose Vaccination x month 2021



Source: Javier Conde (2022)

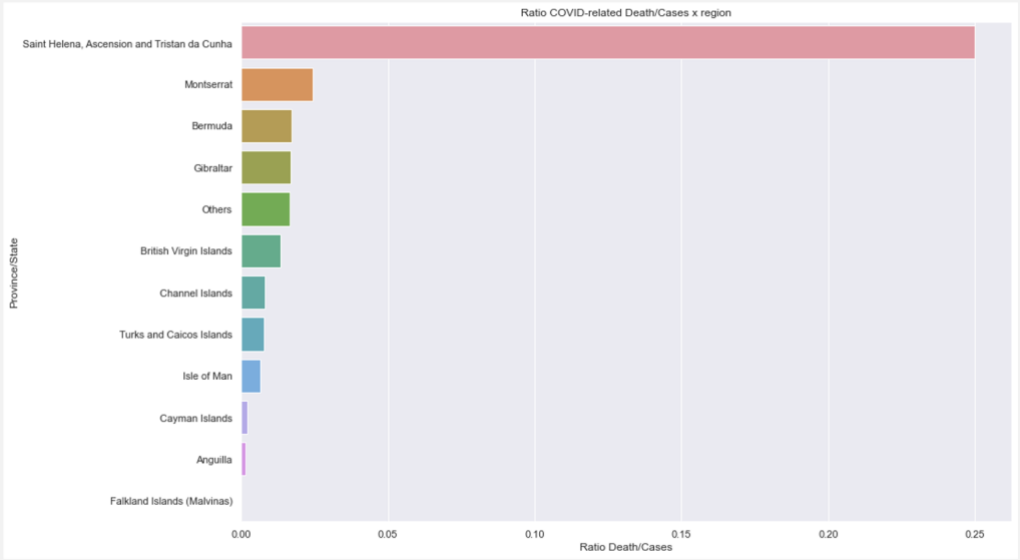
Figure 2.8: Vaccination status x province



Source: Javier Conde (2022)

Visualizations show consistency across regions of 4.51% of population vaccinated with a first dose missing the second dose. This difference across provinces is statistically insignificant. Previous analysis on Deaths to Cases ratio per region shows that Saint Helena, Montserrat, Bermuda, and Gibraltar are the regions with the highest ratio (Figure 2.9). Saint Helena and Montserrat have only reported one COVID-related death each, which render this data statistically insignificant. Advice would be to investigate Bermuda and Gibraltar to launch the campaign. 'Others' province is again included as the ratio calculation is consistent here with the other provinces.

Figure 2.9: Ratio COVID-related Death/Cases x province

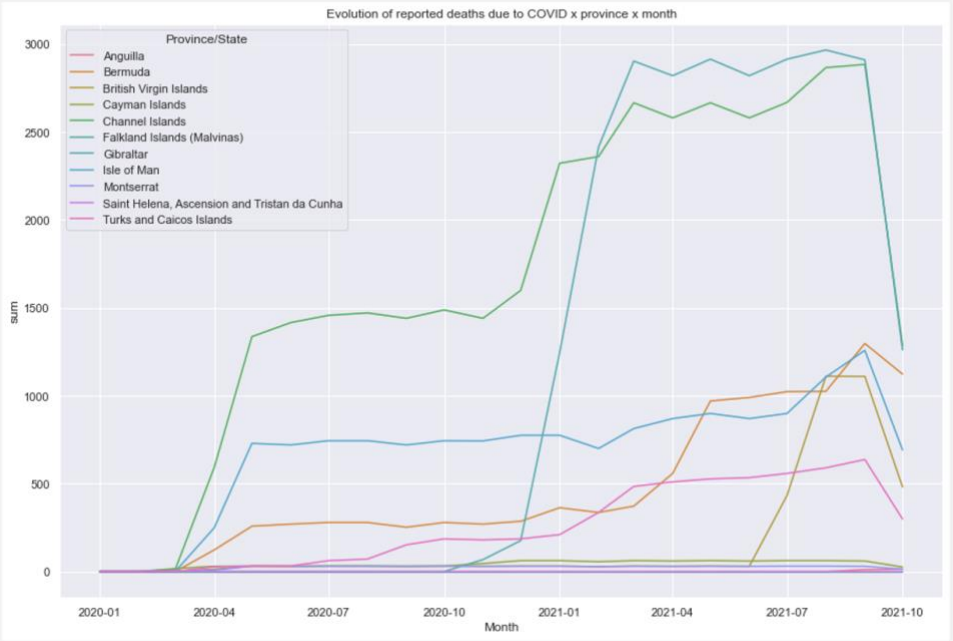


Source: Javier Conde (2022)

These are all absolute figures, so let's work now the evolution of vaccinations and cases per region with time.

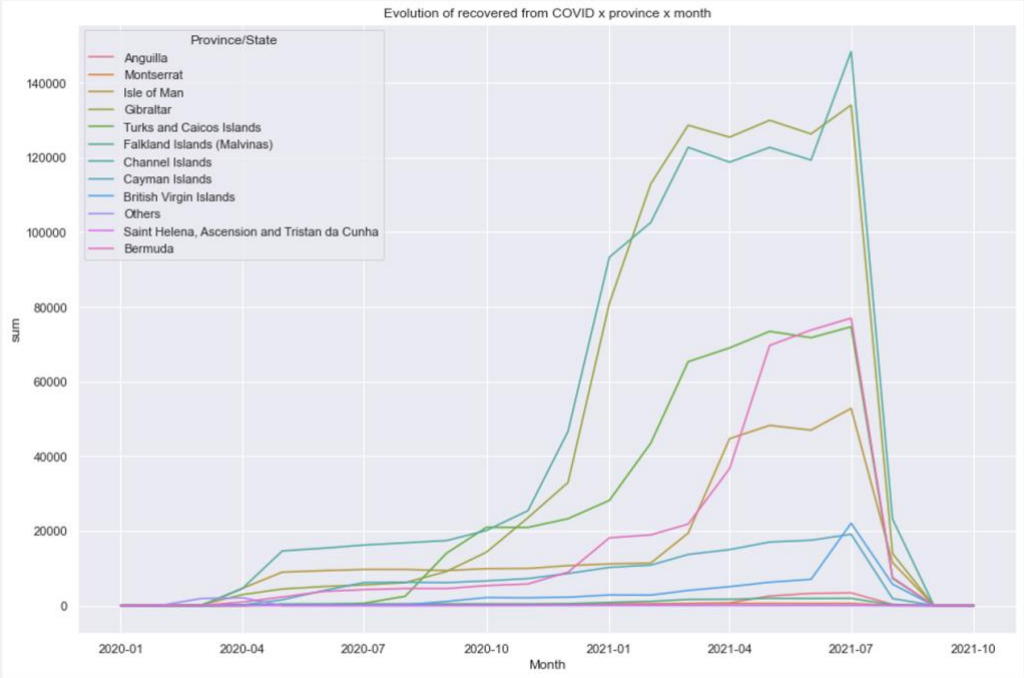
2.3 Visualisations exploration

Figure 2.10: Evolution of reported deaths due to COVID x province x month



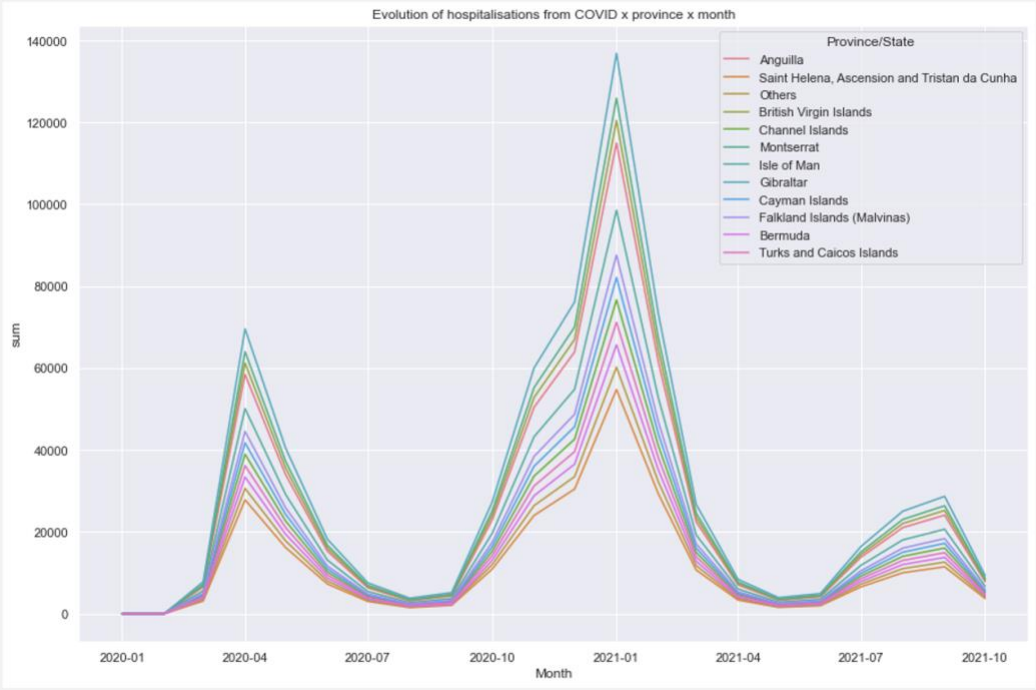
Source: Javier Conde (2022)

Figure 2.11: Evolution of recovered from COVID x province x month



Source: Javier Conde (2022)

Figure 2.12: Evolution of hospitalisations due to COVID x province x month



Source: Javier Conde (2022)

There is here strong evidence that from February 2021 the number of COVID related deaths show signs of plateauing across all regions. Simultaneously, hospitalizations fall sharply (with a small increase in September 2021, possibly as the summer holidays end) and recovered patients show a trend to stabilize or increase. This suggests that the peak of the pandemic may be over with an increase when autumn starts.

Moving forward, it is key to get a basic understanding of people's sentiment about the pandemic. As it is a popular and widely used tool, let's approach and analyze Twitter posts-hashtags on #coronavirus.

2.4 Twitter data relating to the #coronavirus hashtag

One file ('Tweets.csv') is provided for analysis.

The dataset size (3960 records) is too small to make meaningful use of 'retweets'/'favorite' counts, although they may be useful to evaluate messages in the future. There may be opportunities to identify relevant creators and successful hashtags (Source: Norah Wulff, 2022).

Also, only 105 hashtag mentions (0.8%) of a total of 13,336 were #vacc-related within the main #coronavirus (Figure 2.13). More information may be needed on when the "Twitter.csv' was created to uncover opportunities to associate #vacc-related hashtags with others, possibly more popular (i.e. #COVID19 (1,632 mentions), #CovidIsNotOver (472 mentions), #China (262 mentions)) (Figures 2.14, 2.15). This should be taken in great account for future data wrangling iterations of Twitter search for this project, as it could return great value in the shape of information on public sentiment about the relationship between vaccination and COVID.

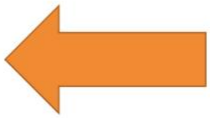
Figure 2.13: Mentions of hashtags containing #vacc- within #coronavirus

```
In [463]: # Display records with the hashtag #vaccinated.
mask = np.column_stack([data['word'].str.contains(r"\#vacc", na=False) for word in data])
vac_hash = data.loc[mask.any(axis=1)]
print(vac_hash.shape)
vac_hash.head(10)
```

```
(25, 2)
```

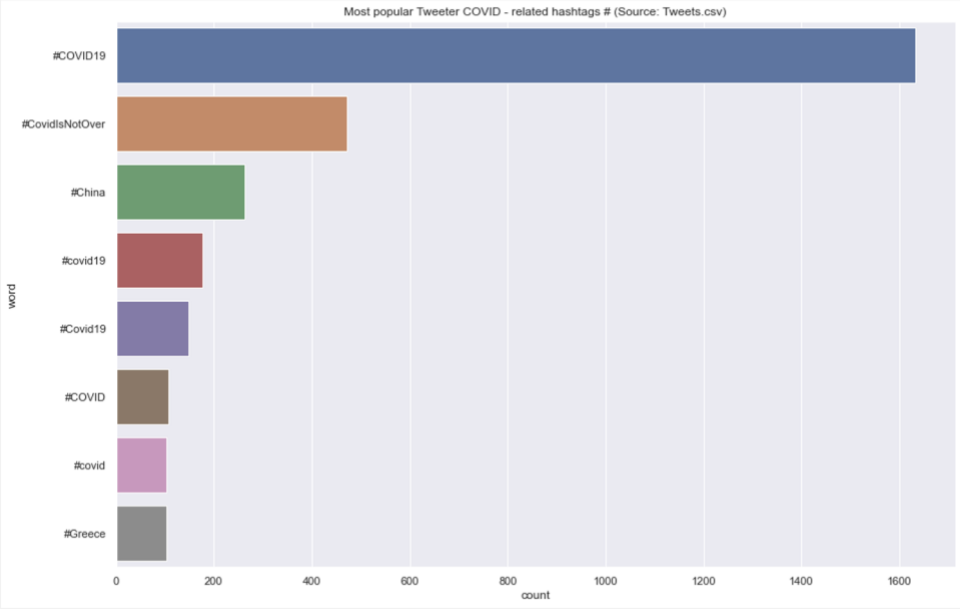
Out[463]:

	word	count
51	#vaccine	29
65	#vaccination	22
69	#vaccines	21
244	#vaccinated	5
576	#vaccine\nClick	3
611	#vaccineacceptance	2
661	appointment\n\n#vaccinate	2
761	#vaccinated.	2
979	#vaccineinjuries	2
1005	#vaccine,	2



Source: Javier Conde (2022)

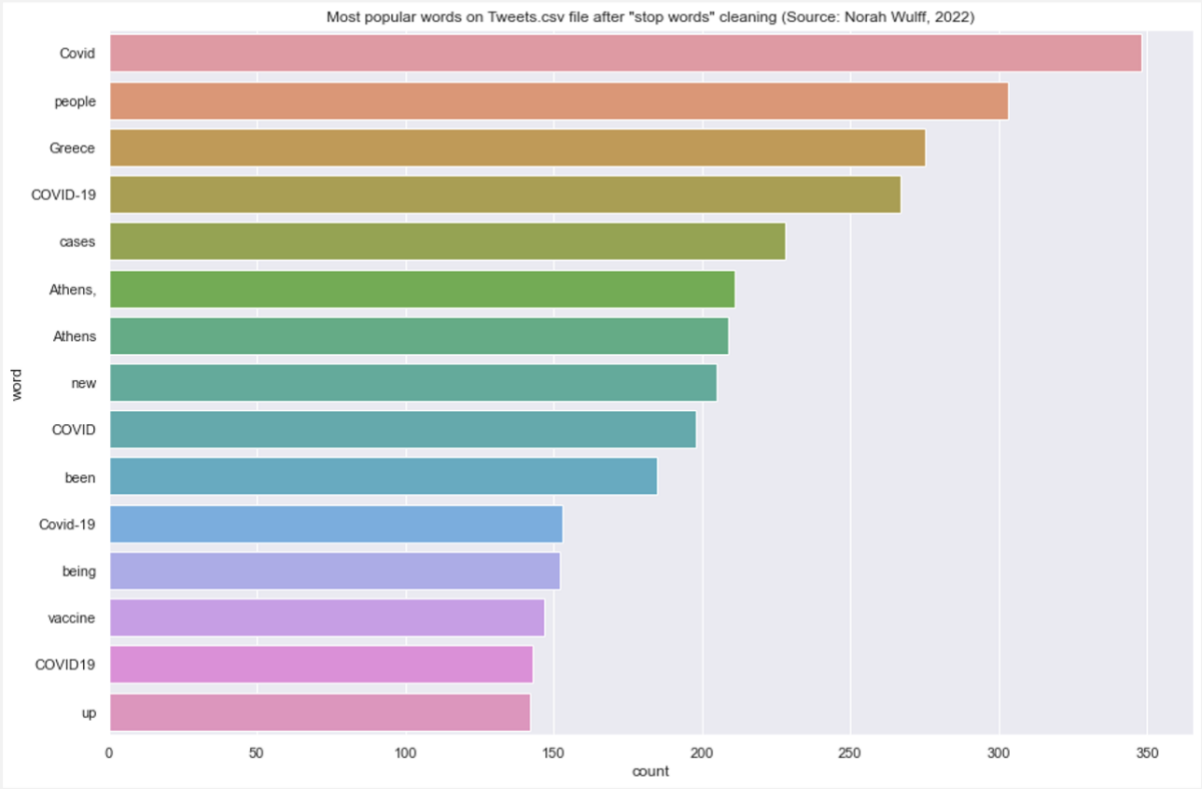
Figure 2.14: Most popular Twitter COVID-related hashtags



Source: Javier Conde (2022)

To finish this analysis, let's investigate the different ways in which Python could help us improving the visualization on time series plots through smoothing and Moving Average (MA) processes.

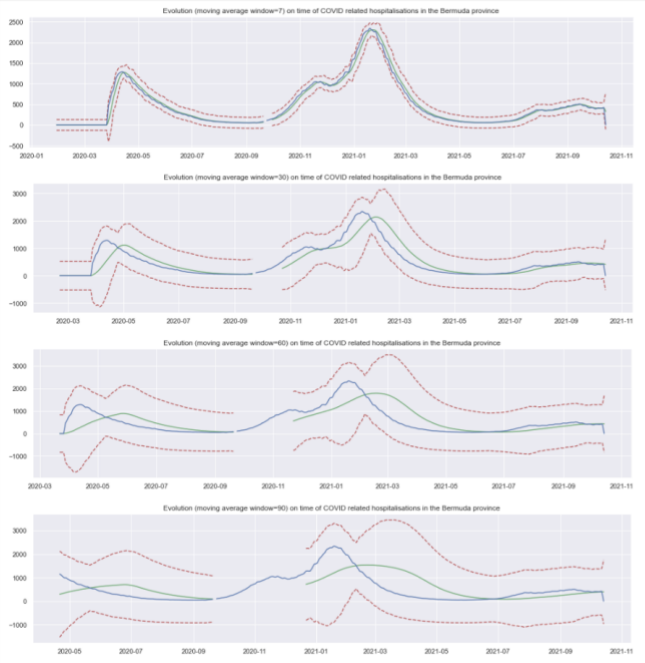
Figure 2.15: Most popular words on Twitter COVID-related after stop words cleaning



Source: Norah Wulff (2022)

2.5 Time-series analysis and data smoothing

Figure 2.16: Evolution (MA window 7/30/60/90) on COVID hospitalisations in the Bermuda province

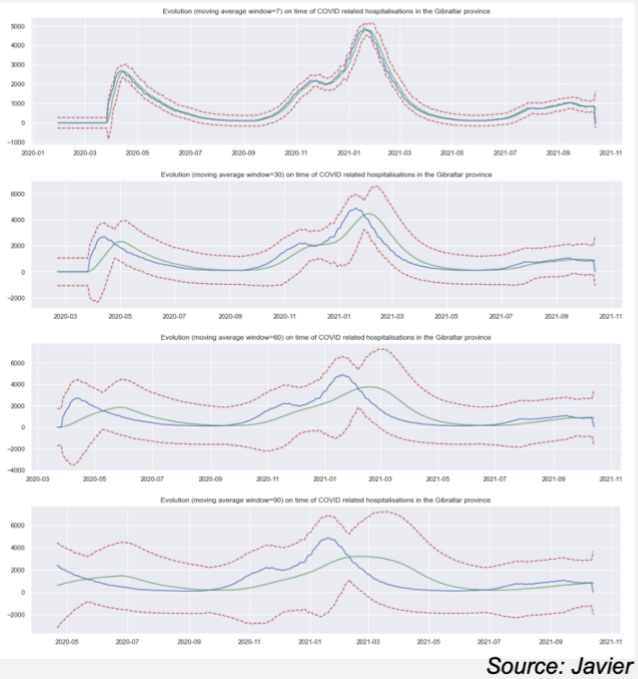


Date	Hospitalised	error
2020-03-28	466.5	368.714286
2020-03-29	534.5	360.357143
2020-03-30	619.0	356.428571

Top three days with biggest difference between daily value and rolling 7-day mean

Source: Javier Conde (2022)

Figure 2.17: Evolution (MA window 7/30/60/90) on COVID hospitalisations in the Gibraltar province



Date	Hospitalised	error
2020-03-28	971.5	767.857143
2020-03-29	1113.0	750.357143
2020-03-30	1289.0	742.214286

Top three days with biggest difference between daily value and rolling 7-day mean

Source: Javier Conde (2022)

These findings evidence that great measurable value could be added in the shape of hospitalization forecasting per province. Advice would be ensuring the quality and quantity of data available, allocating more resources when available.

Let’s explore to finish this report the different recommendations to improve identification of trends, patterns and forecast possibilities while ensuring data ethics integrity in future iterations of this project.

Chapter 3

Recommendations

Recommendation 1: Use both qualitative and quantitative data. Qualitative (categorical, based on groups, interpretation, and description) (i.e., gender, eye colour, binary data) and quantitative data (numeric and portrayed by ordinal, interval, or ratio scales (i.e., the height of a tree, the distance of planets from the earth)) (Source: LSE, 2022) can bring great value to business predictions.

Both complement each other. Everything depends on how frequent forecasts are needed, availability of data and budget, and maturity of the project at hand.

Recommendation 2: Give Continuous improvement priority to minimise risks and stay ahead of the curve for the next pandemic/crisis.
It is more cost effective and reduces risks considerably commiting to small but impactful changes rather than big changes later on. It also can help gather more reliable insights to inform business decision-making.

Recommendation 3: Reduce possibility of a data ethics breach.

1. Staying vigilant and ahead of the curve with new ammendments to the current legislation (Data Protection Act 2018 (DPA 2018), and UK General Data Protection Regulation (UK GDPR))(Ico.org.uk, 2022)
2. Ensure there is an active Data Ethics Framework in place. This would ensure a clearly defined ownership and accountability derived of any data management
3. Promote a culture of communication and improvement through open and honest feedback on current data practices. Potential possibility to create a Community of Data Ethics Practice

Reference List

- Ico.org.uk (2022) *Guide to Data Protection.* Available at: https://ico.org.uk/for-organisations/guide-to-data-protection (Accessed 7 July 2022)
- DataCommons.org (2022) *Gibraltar: COVID-19 cumulative deaths.* Available at: https://datacommons.org/place/country/GIB?category=Health (Accessed 25 June 2022)

List of Figures

- Figure 1.1: Background, project goal and questions to answer (LSE, 2022)
- Figure 1.2: Project Github repository (I) (Javier Conde, 2022)
- Figure 1.3: Project Github repository (II) (Javier Conde, 2022)
- Figure 2.1: Overview of COVID cases x UK province (Javier Conde, 2022)
- Figure 2.2: Overview of vaccinations x UK province (Javier Conde, 2022)
- Figure 2.3: Cumulative deaths reported in UK province Gibraltar (DataCommons.org, 2022)
- Figure 2.4: Overview of total country vaccinations (Javier Conde, 2022)
- Figure 2.5: First findings (Javier Conde, 2022)
- Figure 2.6: Evolution of 1st Dose Vaccination x month 2021 (Javier Conde, 2022)
- Figure 2.7: Evolution of 2nd Dose Vaccination x month 2021 (Javier Conde, 2022)
- Figure 2.8: Vaccination status x province (Javier Conde, 2022)
- Figure 2.9: Ratio COVID-related Death/Cases x province (Javier Conde, 2022)
- Figure 2.10: Evolution of reported deaths due to COVID x province x month (Javier Conde, 2022)
- Figure 2.11: Evolution of recovered from COVID x province x month (Javier Conde, 2022)
- Figure 2.12: Evolution of hospitalizations due to COVID x province x month (Javier Conde, 2022)
- Figure 2.13: Mentions of hashtags containing #vacc- within #coronavirus (Javier Conde, 2022)
- Figure 2.14: Most popular Twitter COVID-related hashtags (Javier Conde, 2022)

- Figure 2.15: Most popular words on Twitter COVID-related after stop words cleaning (Javier Conde, 2022)
- Figure 2.16: Evolution (MA window 7/30/60/90) on COVID hospitalizations in the Bermuda province (Javier Conde, 2022)
- Figure 2.17: Evolution (MA window 7/30/60/90) on COVID hospitalizations in the Gibraltar province (Javier Conde, 2022)