Course 2: Final Assignment Jupyter Presentation

Javier Conde Pascual | 11th July 2022

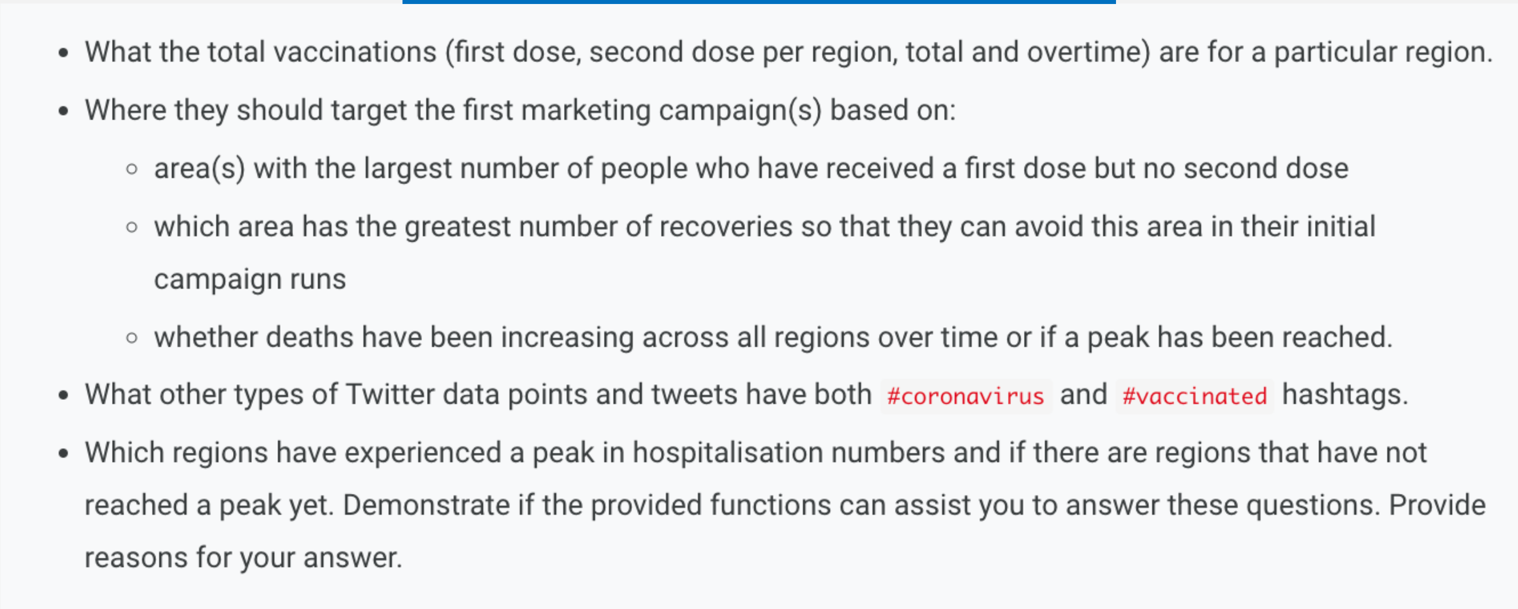
1. Scenario and system set up

Figure 1.1: Background, project goal and questions to answer

ROLE: assuming the role of data analyst working with the UK government to analyse COVID-19 data (from January 2020 to October 2021)

PROJECT GOAL: identify trends and patterns in the data that could inform a series of marketing campaigns to promote the vaccine. The ultimate target is increasing the number of fully vaccinated individuals (people who have received a first and second dose of the vaccine) through these campaigns

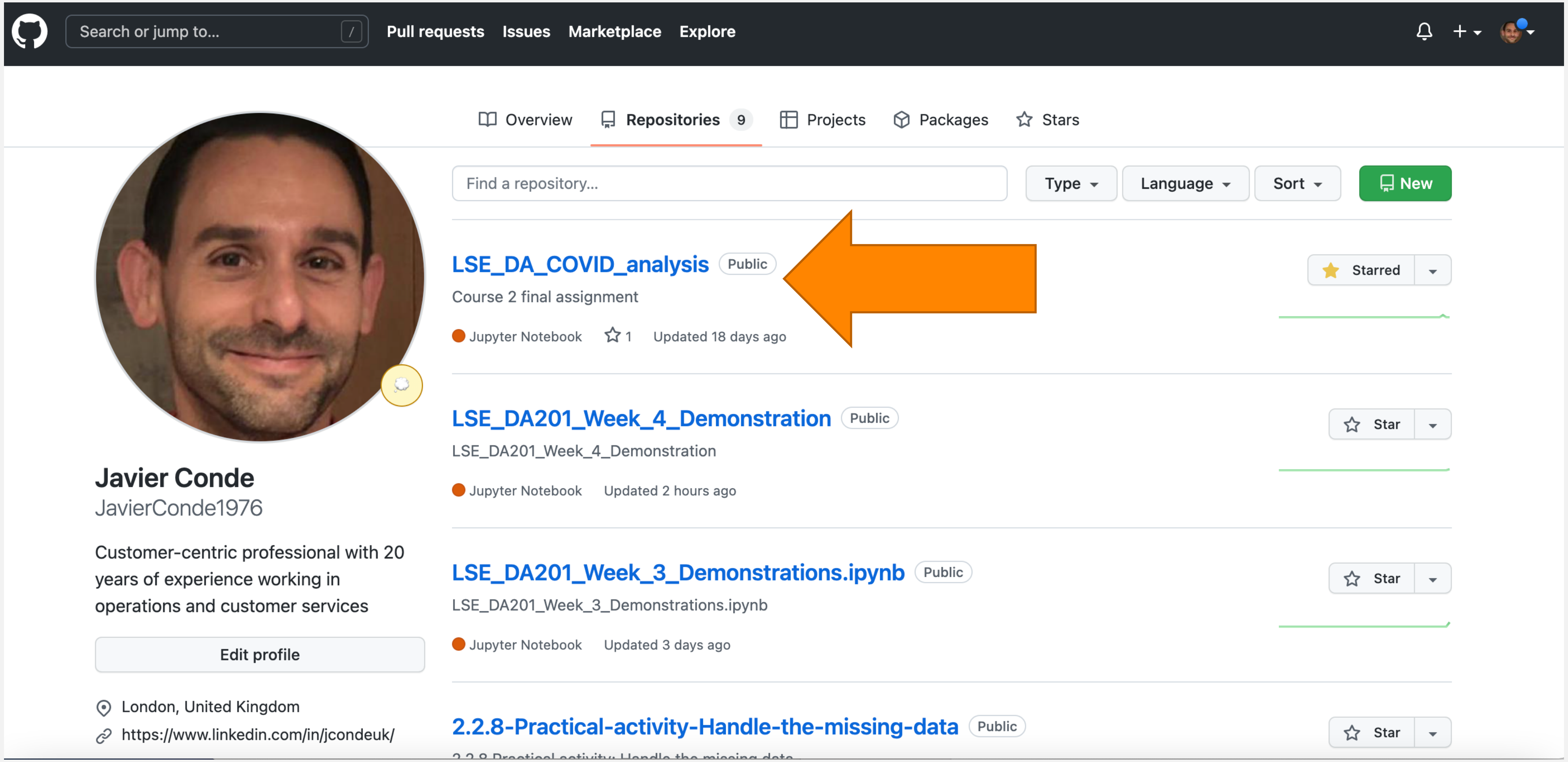
QUESTIONS TO BE ANSWERED:



- What the total vaccinations (first dose, second dose per region, total and overtime) are for a particular region.
- Where they should target the first marketing campaign(s) based on:
  - area(s) with the largest number of people who have received a first dose but no second dose
  - which area has the greatest number of recoveries so that they can avoid this area in their initial campaign runs
  - whether deaths have been increasing across all regions over time or if a peak has been reached.
- What other types of Twitter data points and tweets have both `#coronavirus` and `#vaccinated` hashtags.
- Which regions have experienced a peak in hospitalisation numbers and if there are regions that have not reached a peak yet. Demonstrate if the provided functions can assist you to answer these questions. Provide reasons for your answer.

Source: LSE (2022)

Figure 1.2: Project Github repository (I)



Search or jump to...

Pull requests Issues Marketplace Explore

Overview Repositories 9 Projects Packages Stars

Find a repository...

Type Language Sort New

Javier Conde
JavierConde1976

Customer-centric professional with 20 years of experience working in operations and customer services

Edit profile

London, United Kingdom
https://www.linkedin.com/in/jcondeuk/

LSE_DA_COVID_analysis Public
Course 2 final assignment
Jupyter Notebook ☆ 1 Updated 18 days ago
★ Starred

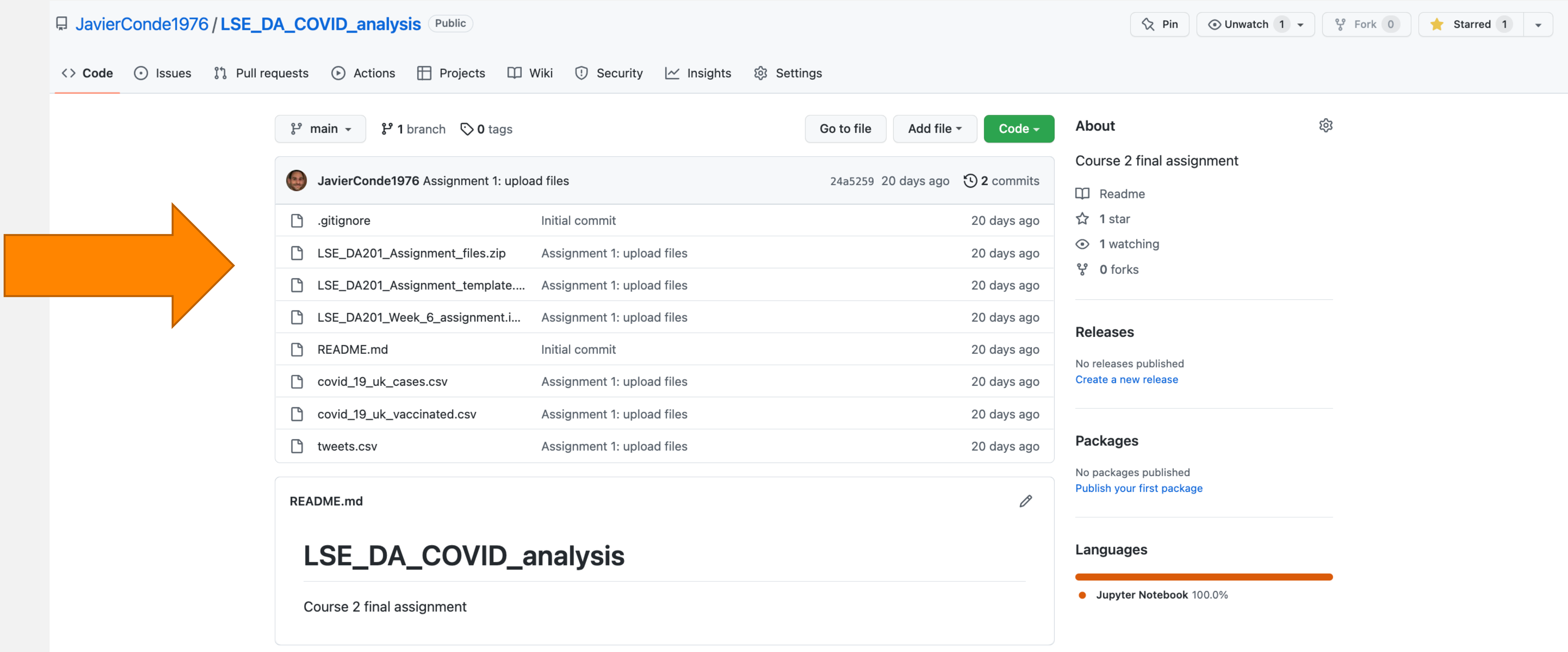
LSE_DA201_Week_4_Demonstration Public
LSE_DA201_Week_4_Demonstration
Jupyter Notebook Updated 2 hours ago
☆ Star

LSE_DA201_Week_3_Demonstrations.ipynb Public
LSE_DA201_Week_3_Demonstrations.ipynb
Jupyter Notebook Updated 3 days ago
☆ Star

2.2.8-Practical-activity-Handle-the-missing-data Public
2.2.8 Practical activity: Handle the missing data
☆ Star

Source: Javier Conde (2022)

Figure 1.3: Project Github repository (II)



JavierConde1976 / LSE_DA_COVID_analysis Public

1 branch 0 tags

JavierConde1976 Assignment 1: upload files 24a5259 20 days ago 2 commits

.gitignore	Initial commit	20 days ago
LSE_DA201_Assignment_files.zip	Assignment 1: upload files	20 days ago
LSE_DA201_Assignment_template....	Assignment 1: upload files	20 days ago
LSE_DA201_Week_6_assignment.i...	Assignment 1: upload files	20 days ago
README.md	Initial commit	20 days ago
covid_19_uk_cases.csv	Assignment 1: upload files	20 days ago
covid_19_uk_vaccinated.csv	Assignment 1: upload files	20 days ago
tweets.csv	Assignment 1: upload files	20 days ago

README.md

LSE_DA_COVID_analysis

Course 2 final assignment

About

Course 2 final assignment

Readme
1 star
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

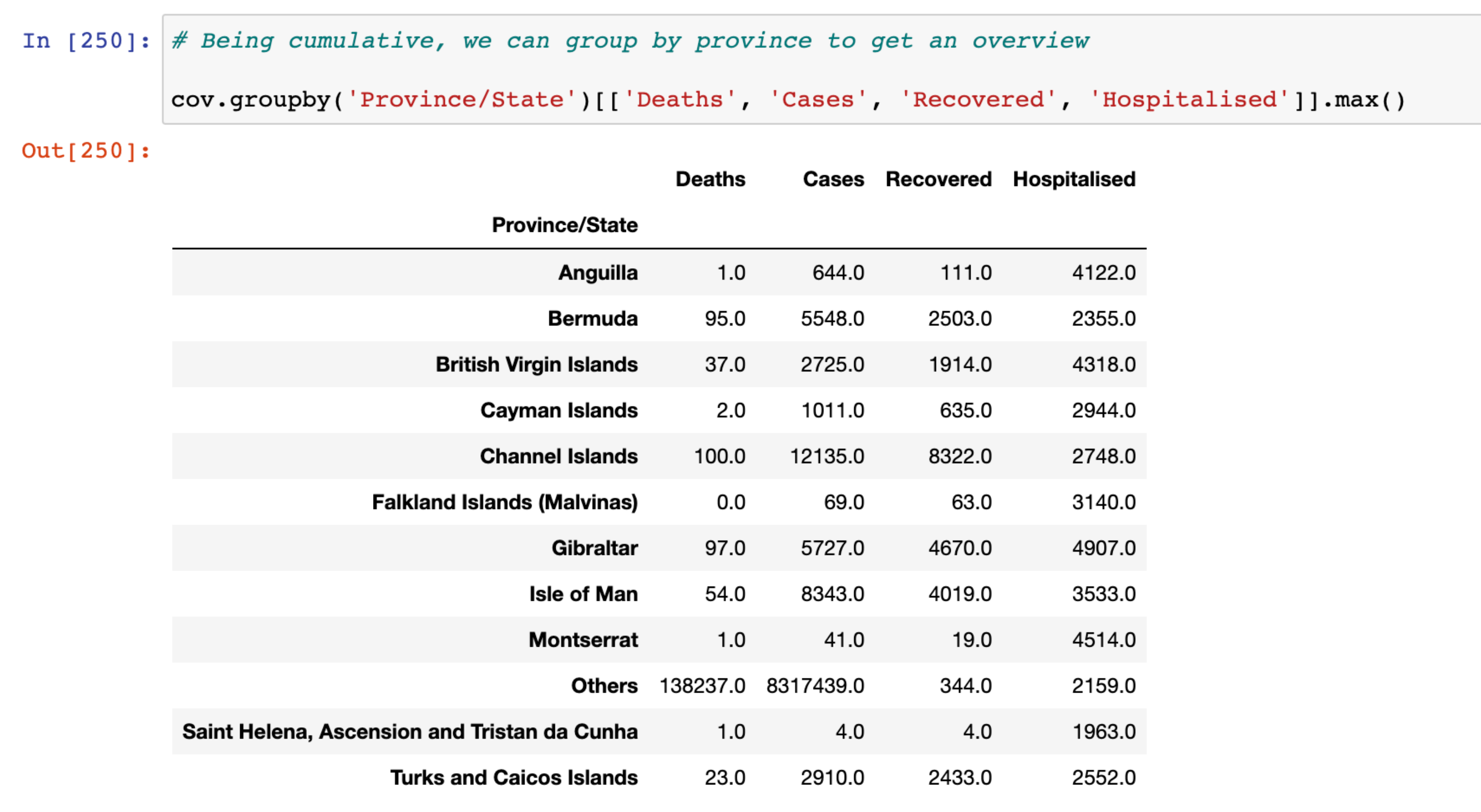
Languages

Jupyter Notebook 100.0%

Source: Javier Conde (2022)

2.1. First approach and insights to provided data

Figure 2.1: *Overview of COVID cases x UK province*

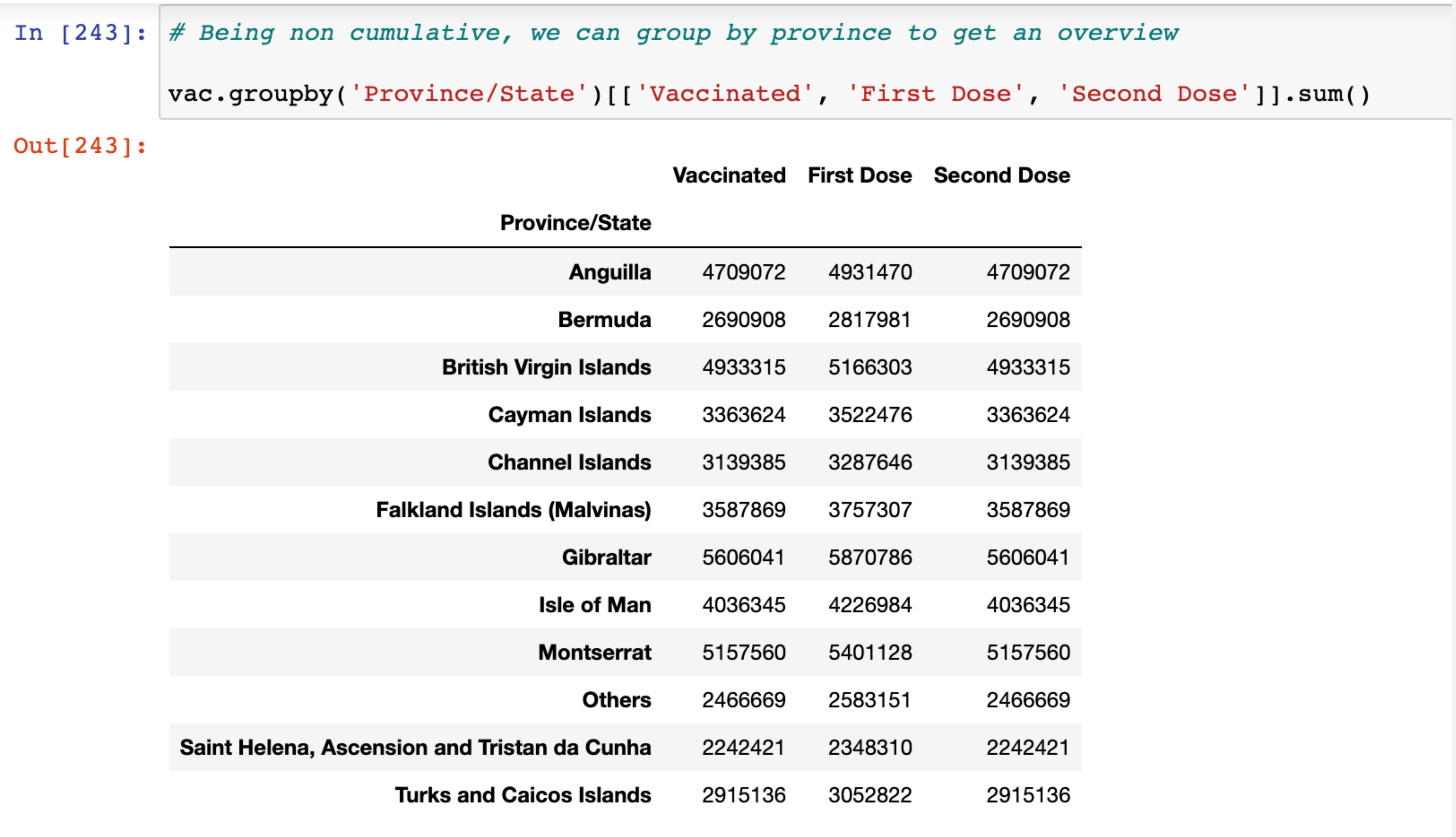


```
In [250]: # Being cumulative, we can group by province to get an overview

          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



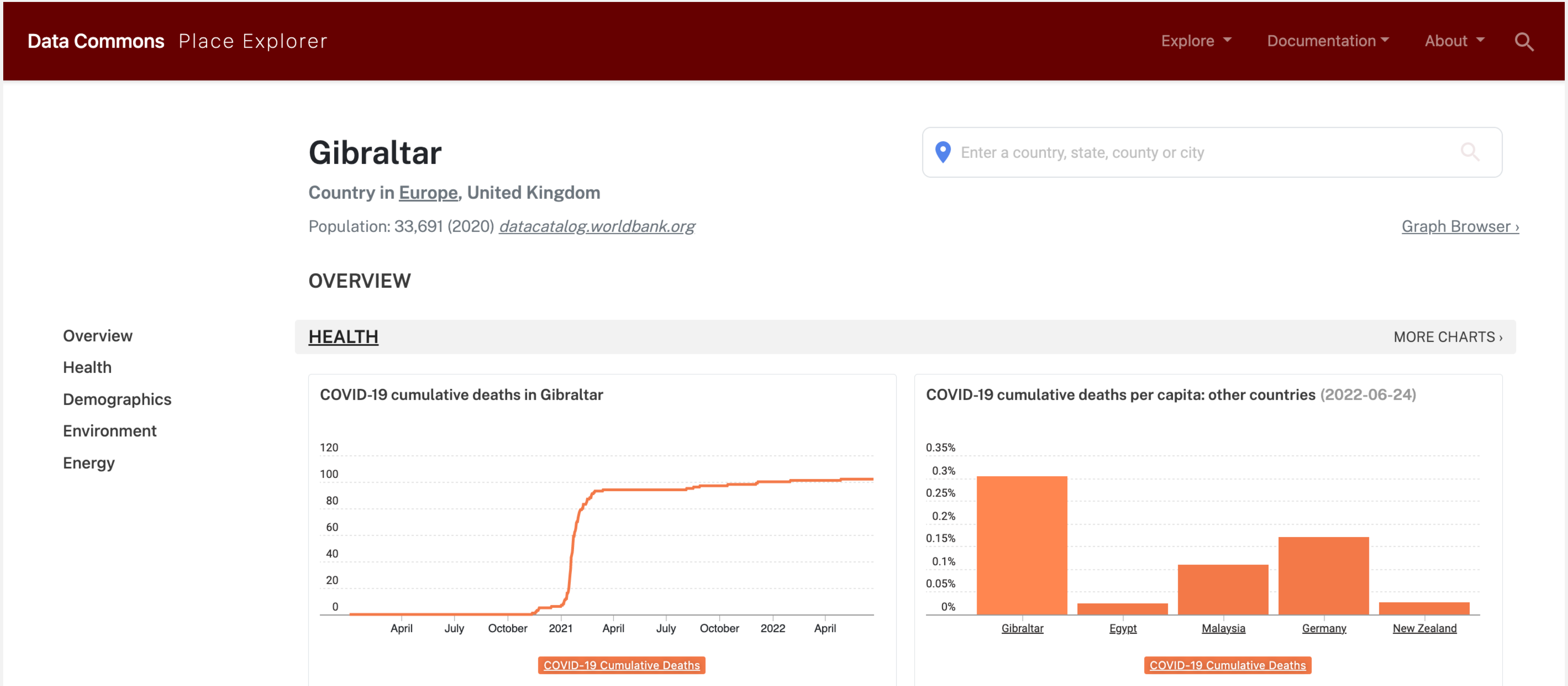
In [243]:
```
# Being non cumulative, we can group by province to get an overview

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



Data Commons Place Explorer
Explore
Documentation
About
Enter a country, state, county or city

Gibraltar

Country in Europe, United Kingdom

Population: 33,691 (2020) *datacatalog.worldbank.org*

Graph Browser ›

OVERVIEW

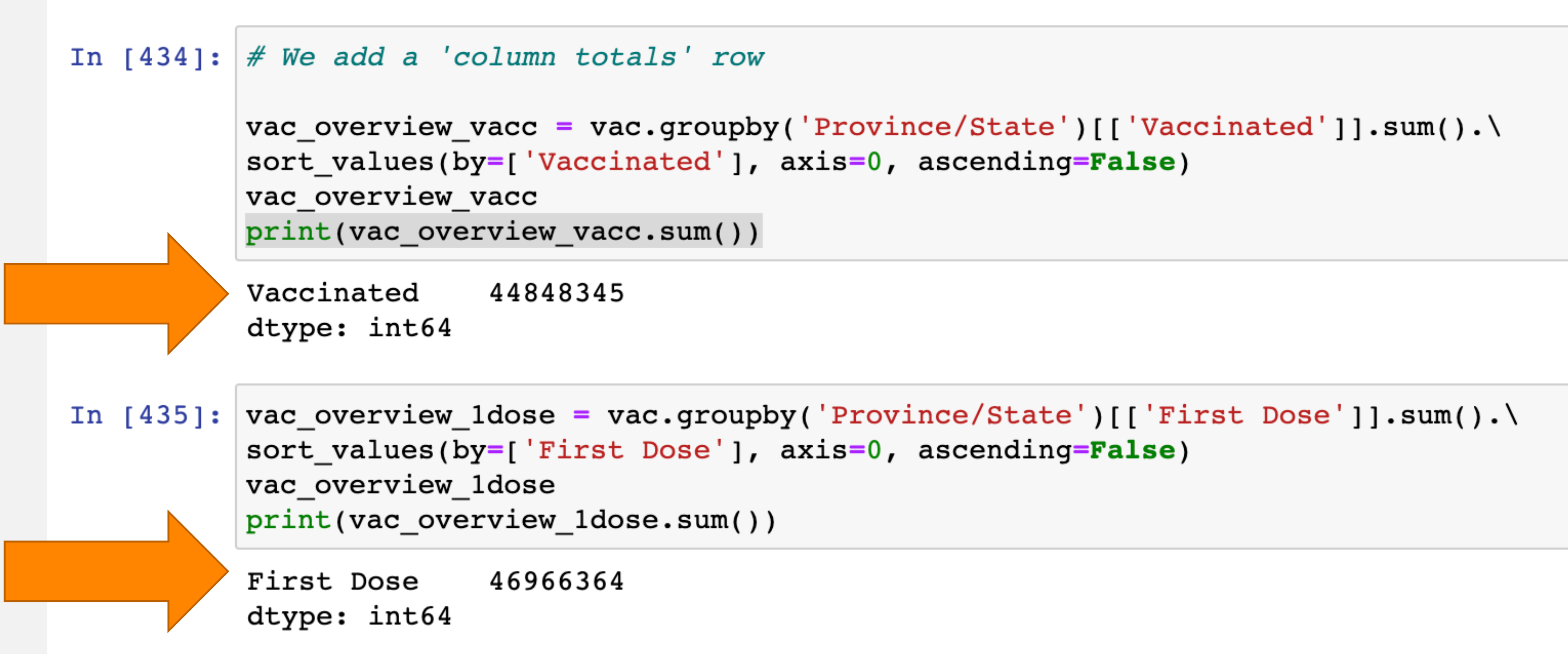
HEALTH MORE CHARTS ›

COVID-19 cumulative deaths in Gibraltar

COVID-19 cumulative deaths per capita: other countries (2022-06-24)

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)

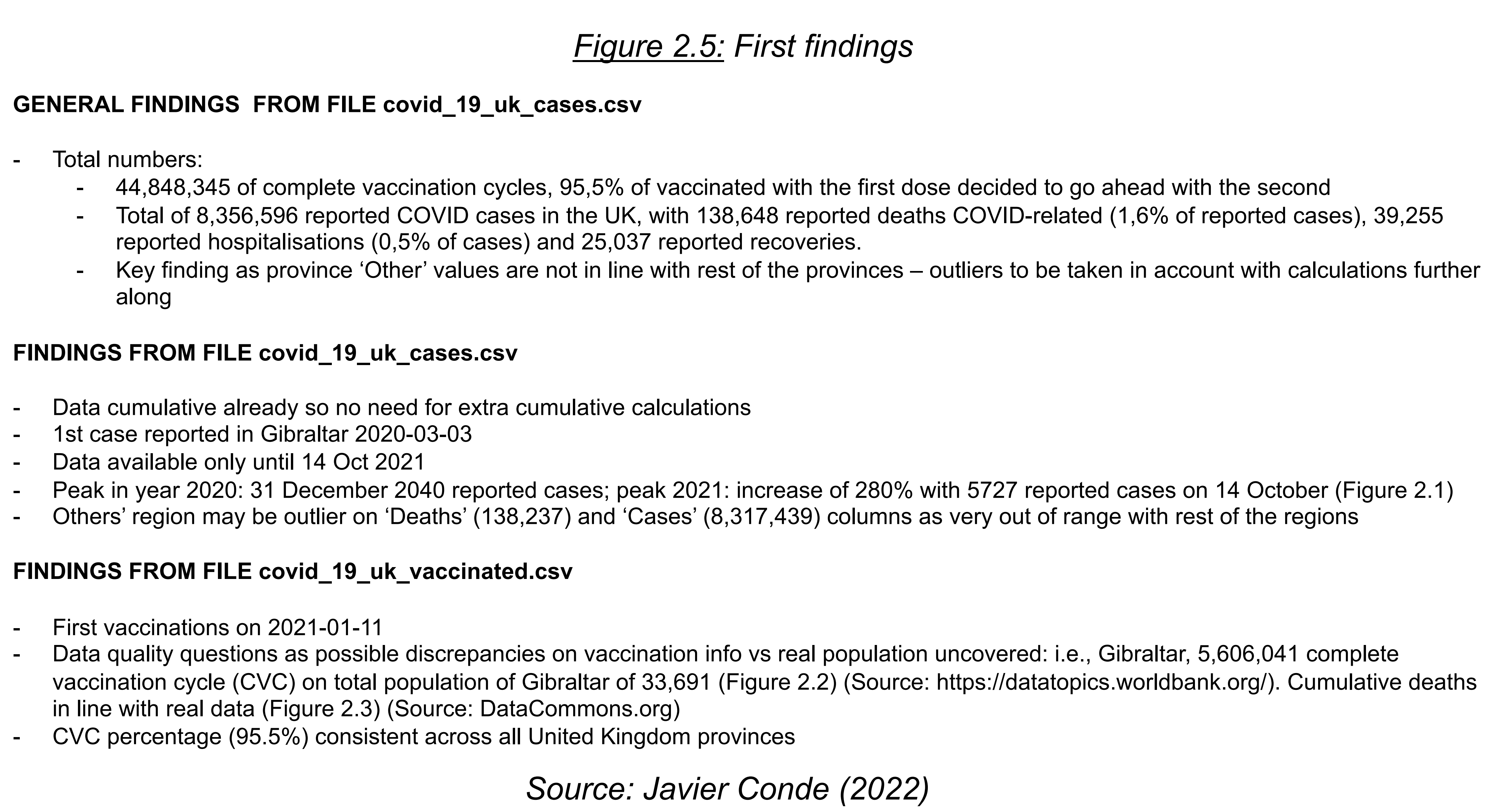


Figure 2.5: First findings

GENERAL FINDINGS FROM FILE covid_19_uk_cases.csv

- Total numbers:
    - 44,848,345 of complete vaccination cycles, 95,5% of vaccinated with the first dose decided to go ahead with the second
    - Total of 8,356,596 reported COVID cases in the UK, with 138,648 reported deaths COVID-related (1,6% of reported cases), 39,255 reported hospitalisations (0,5% of cases) and 25,037 reported recoveries.
    - Key finding as province ‘Other’ values are not in line with rest of the provinces – outliers to be taken in account with calculations further along

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

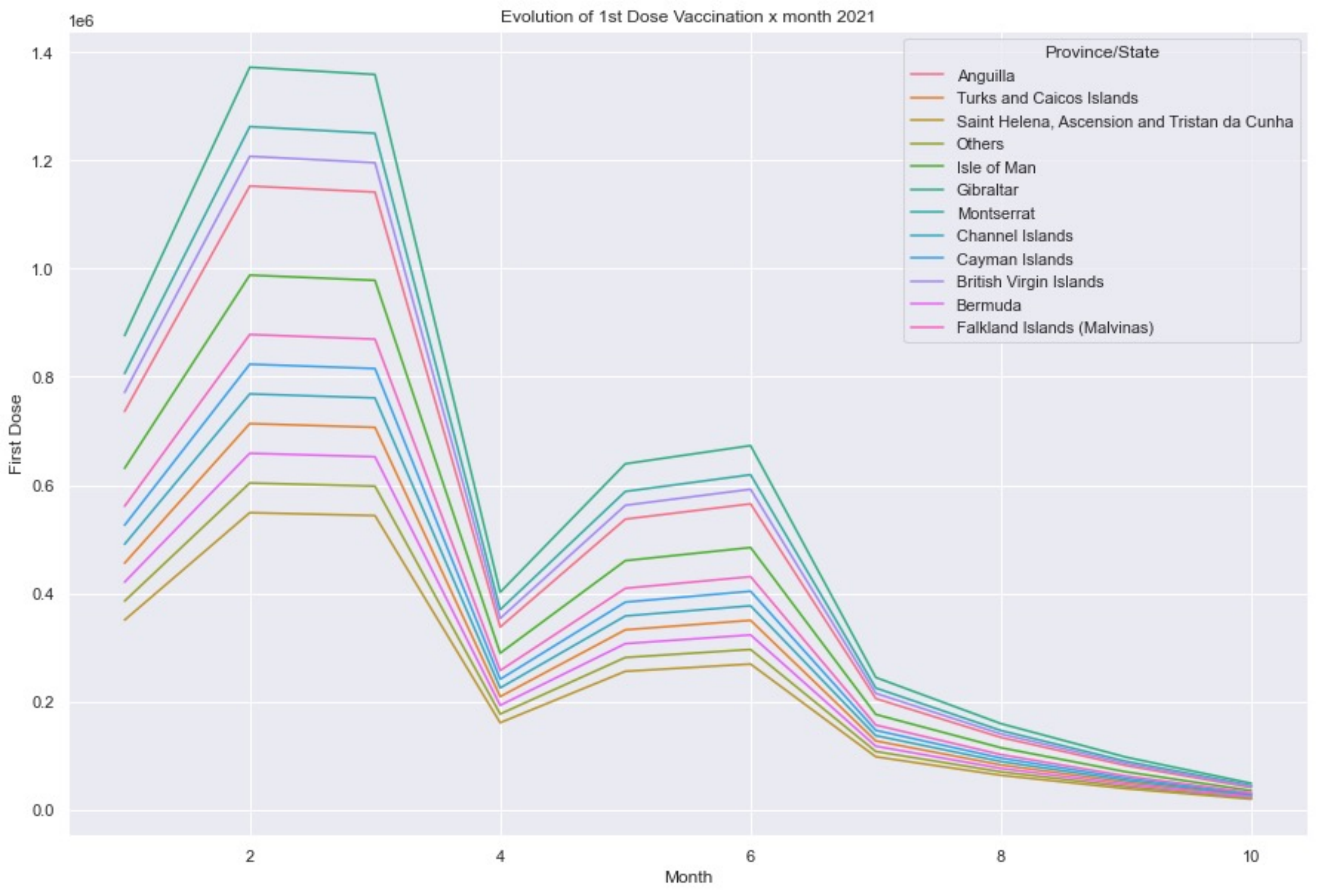
FINDINGS FROM FILE covid_19_uk_vaccinated.csv

- 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

Source: Javier Conde (2022)

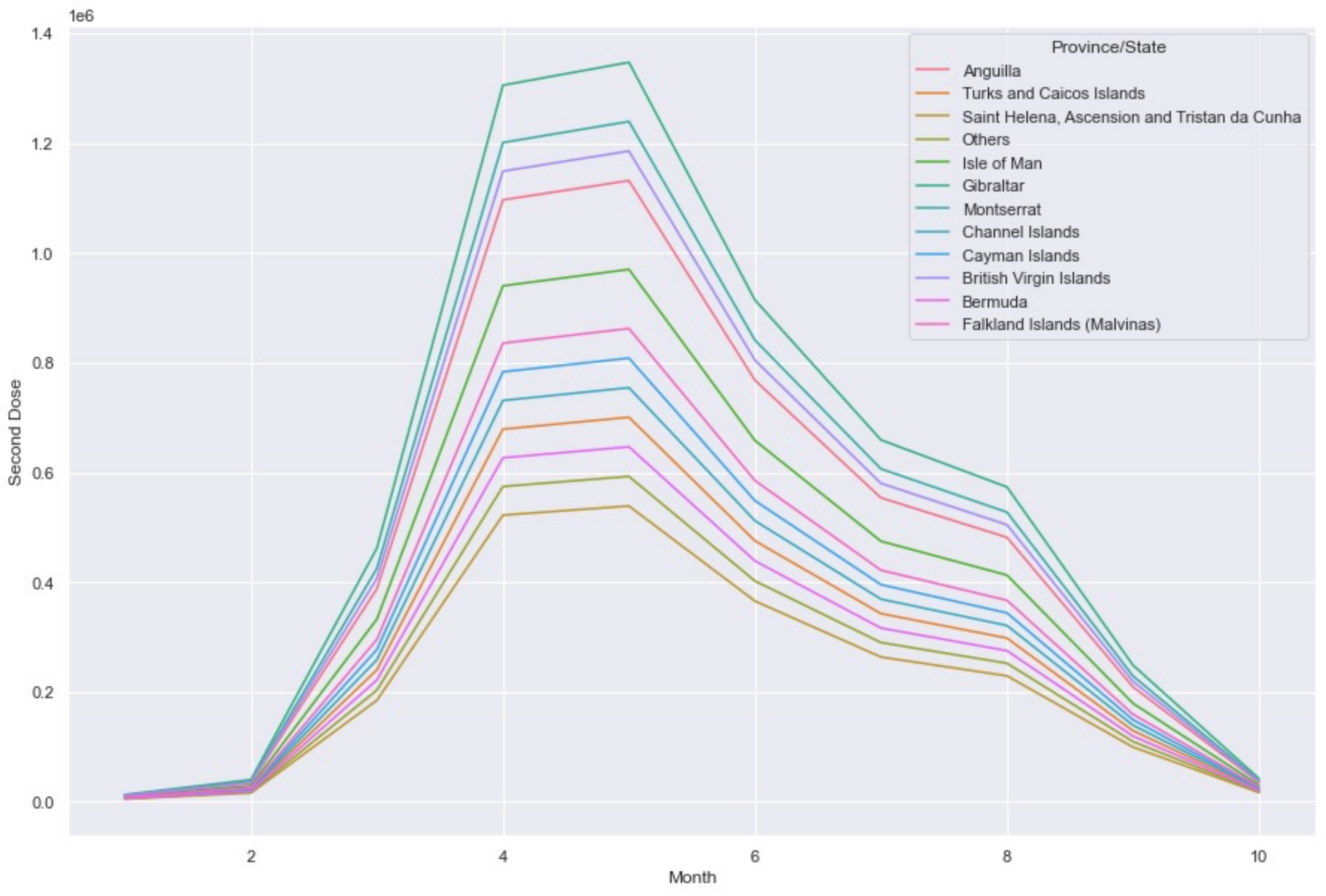
2.2. Vaccination - areas to target marketing campaigand first approach to visualisations

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



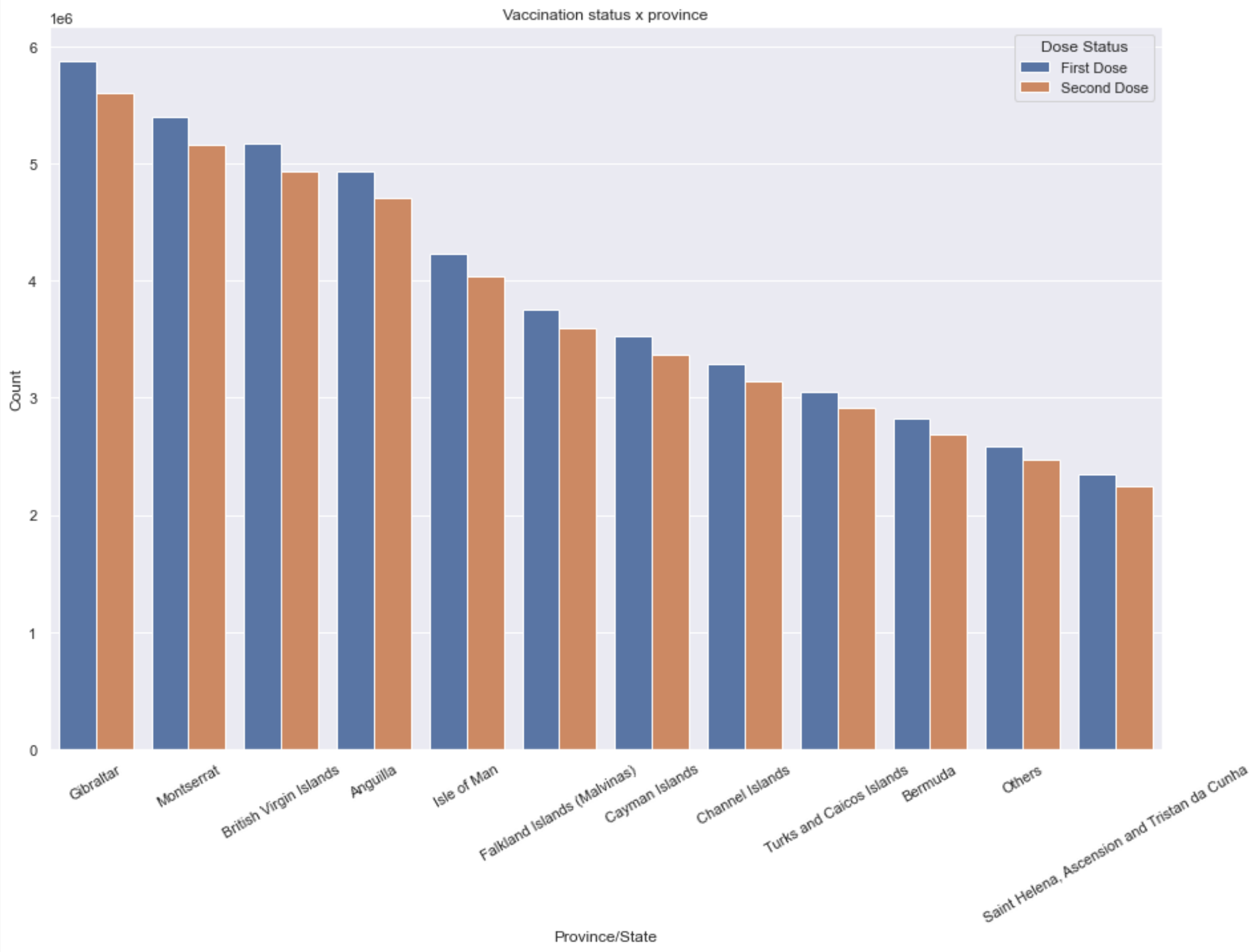
Source: Javier Conde (2022)

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



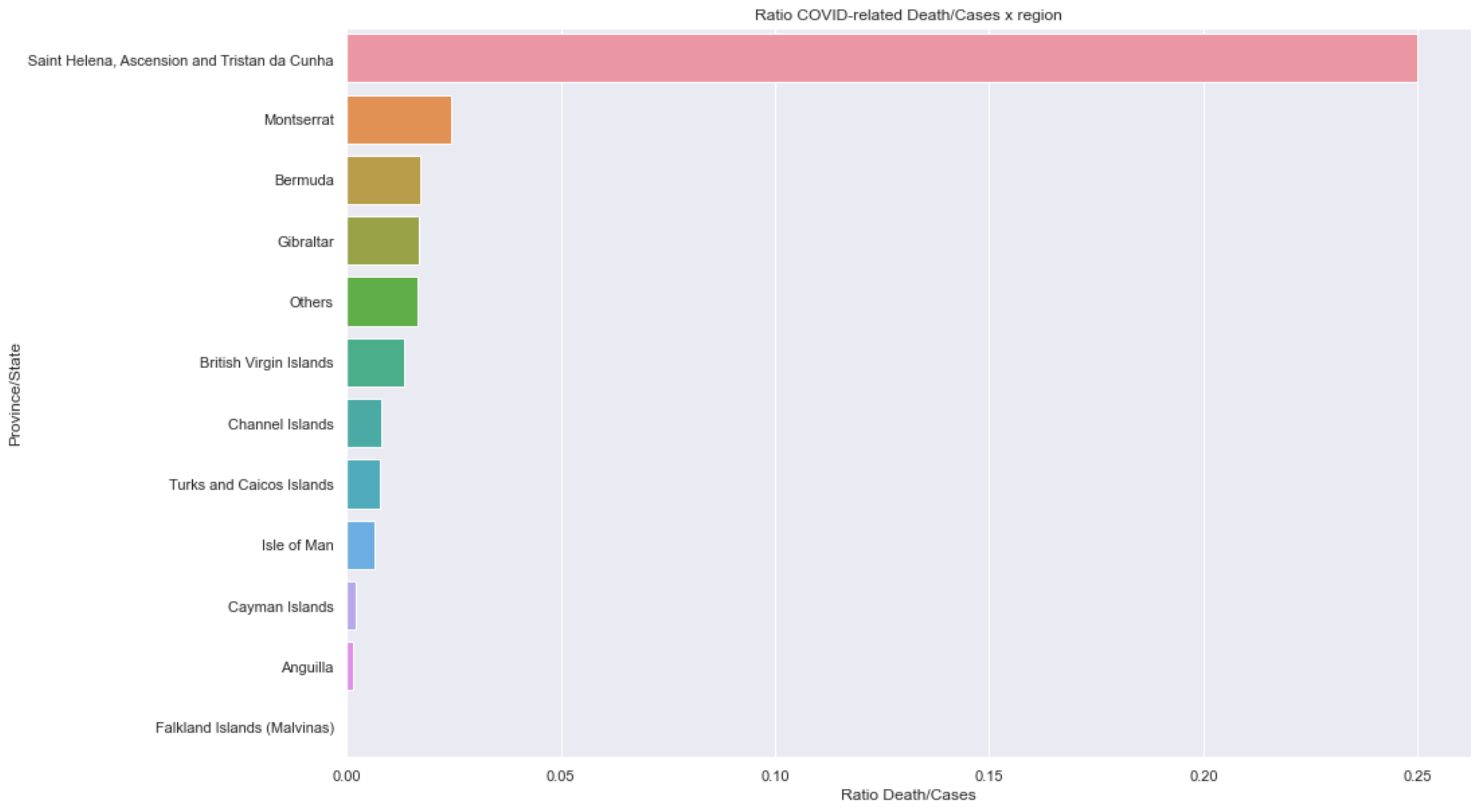
Source: Javier Conde (2022)

Figure 2.8: *Vaccination status x province*



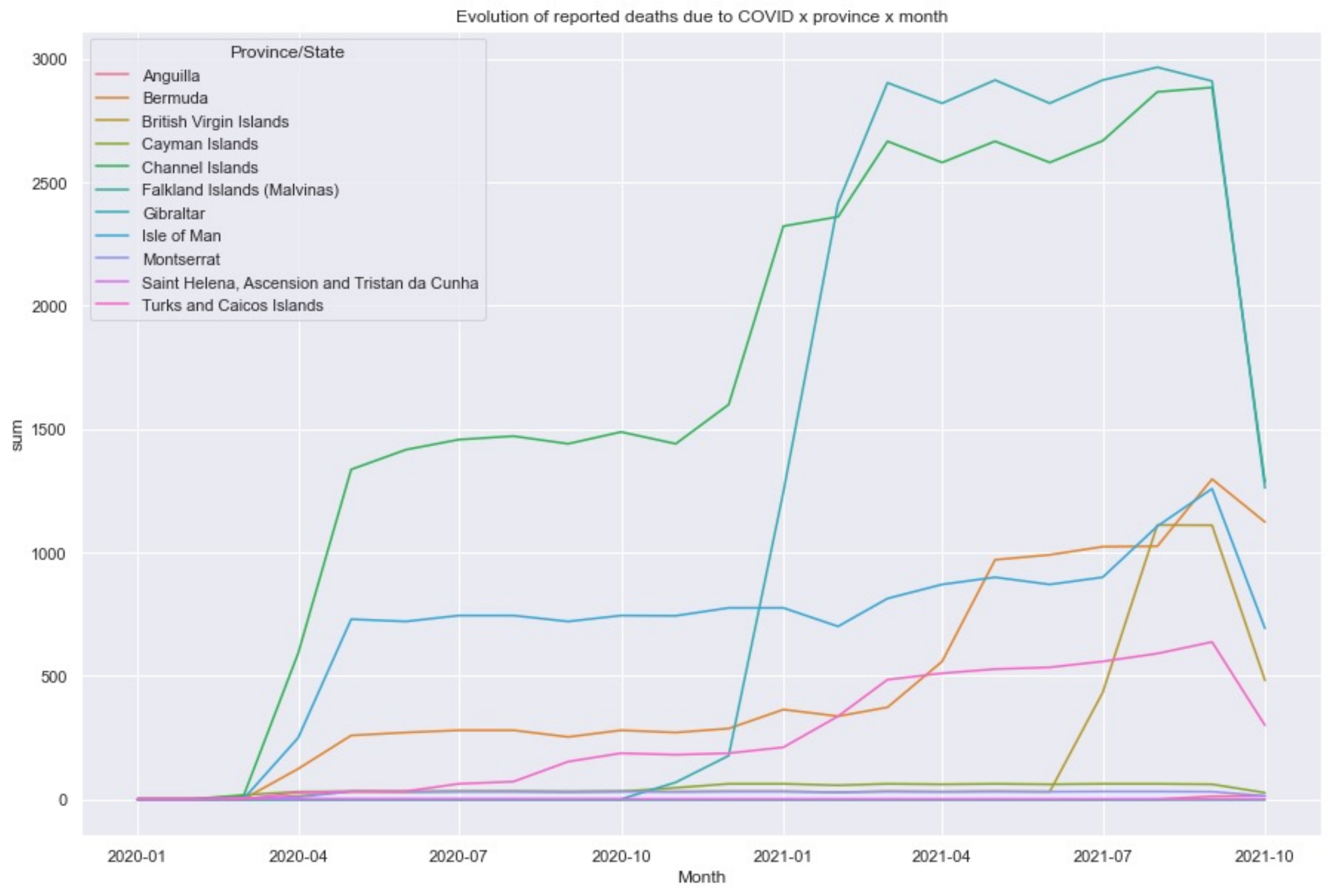
Source: Javier Conde (2022)

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



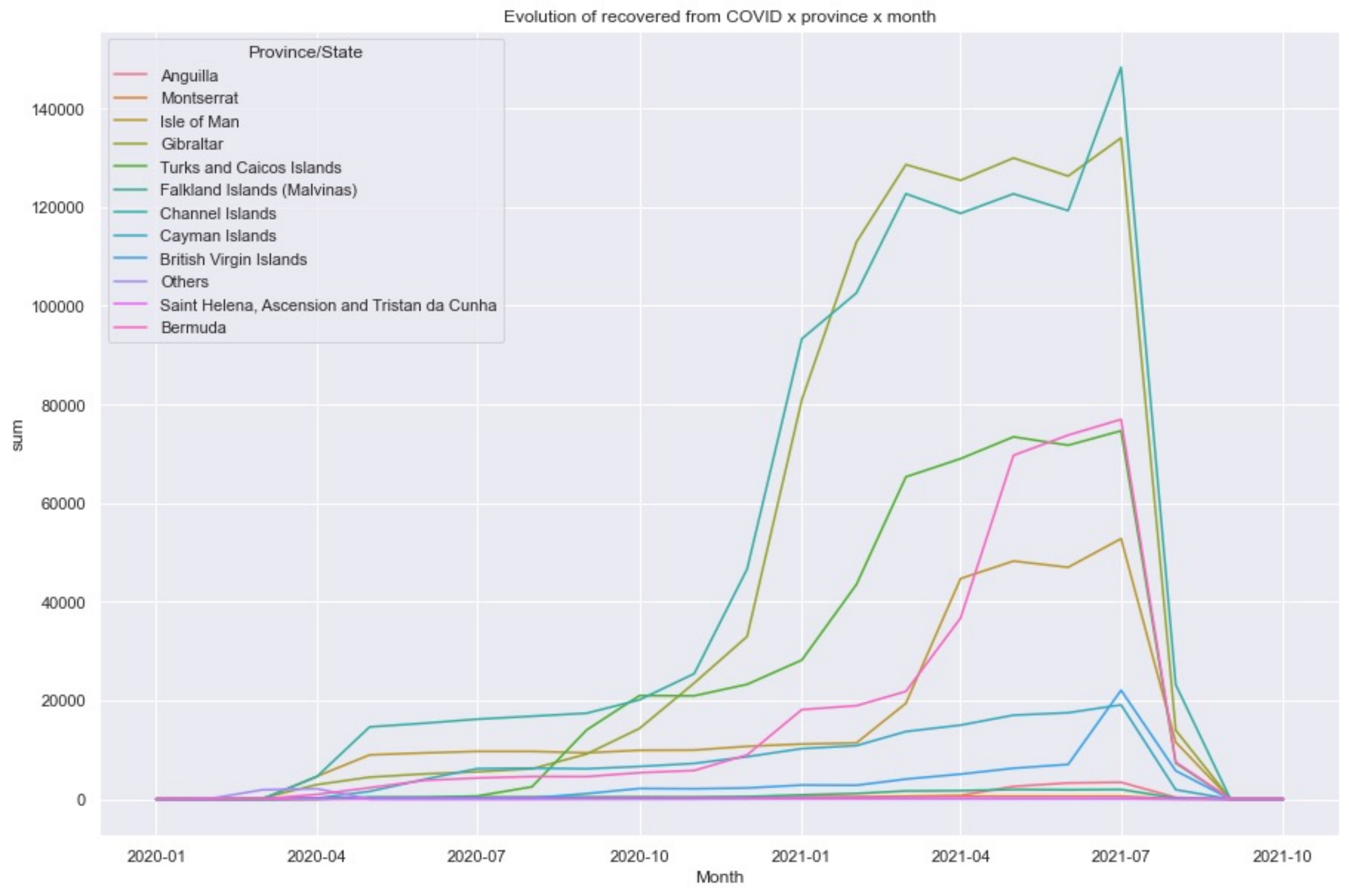
Source: Javier Conde (2022)

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



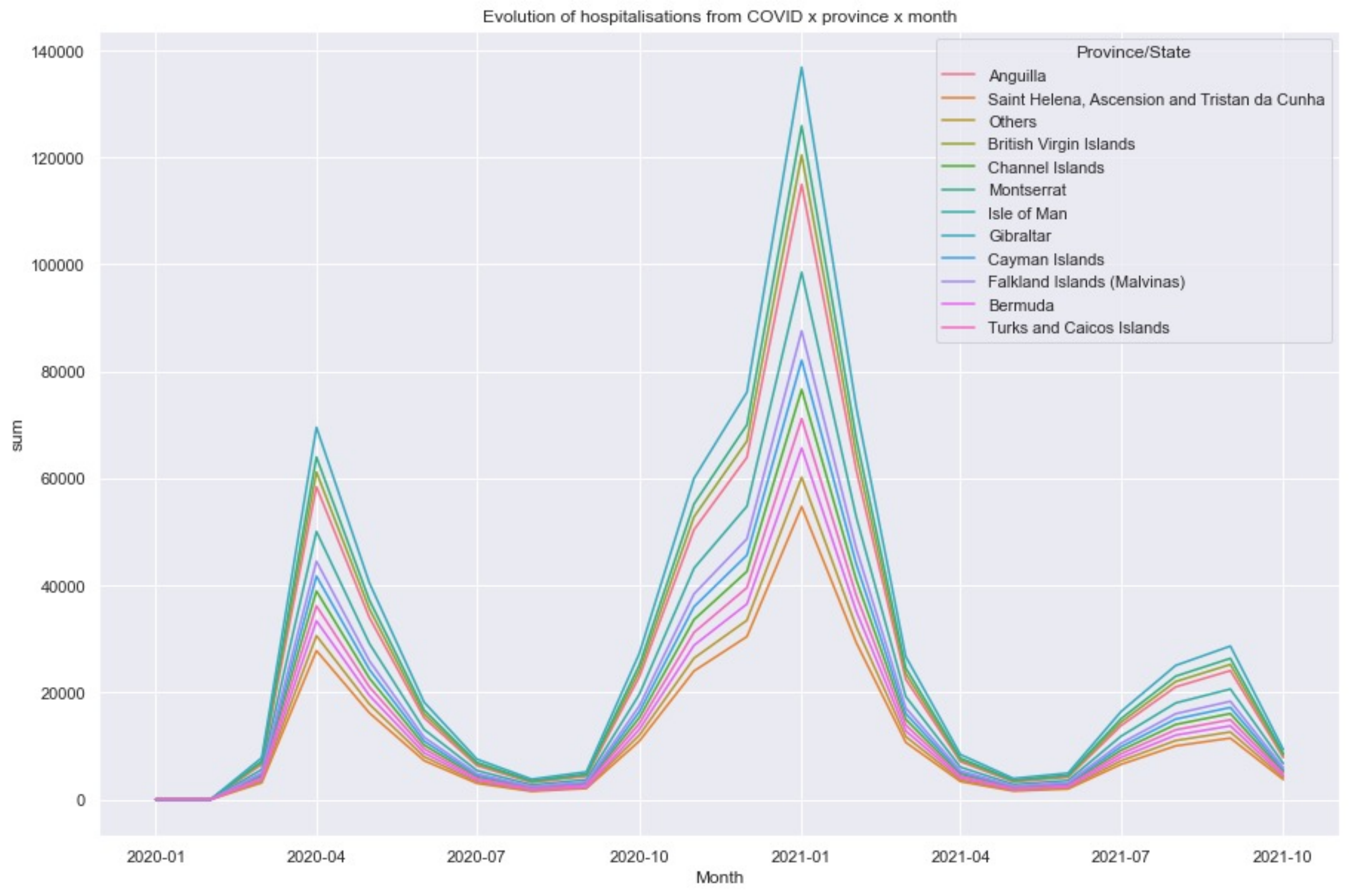
Source: Javier Conde (2022)

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



Source: Javier Conde (2022)

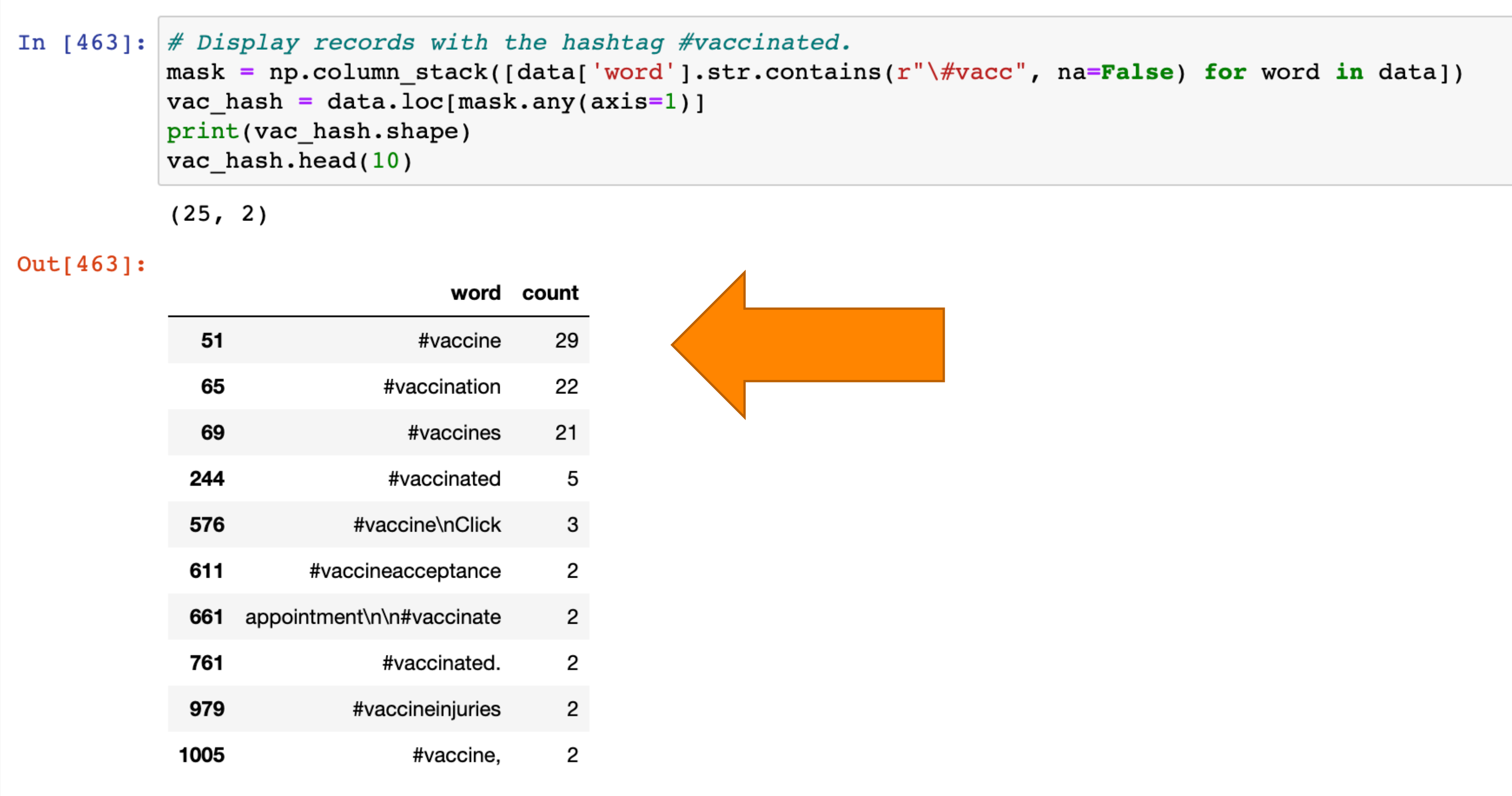
Figure 2.12: *Evolution of hospitalisations due to COVID x province x month*



Source: Javier Conde (2022)

2.4. Twitter data relating to the #coronavirus hashtag

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)
```

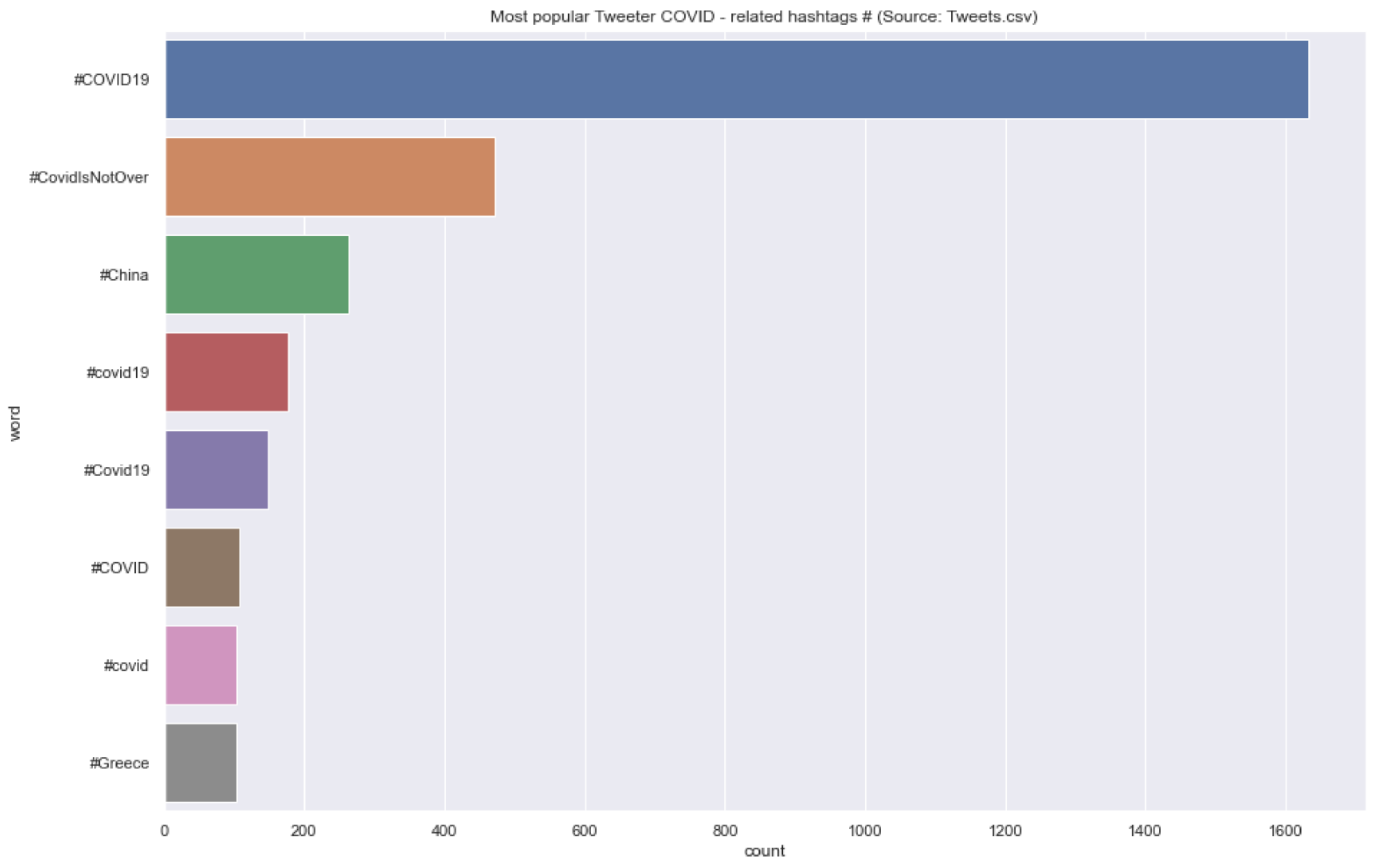
```
(25, 2)
```

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)

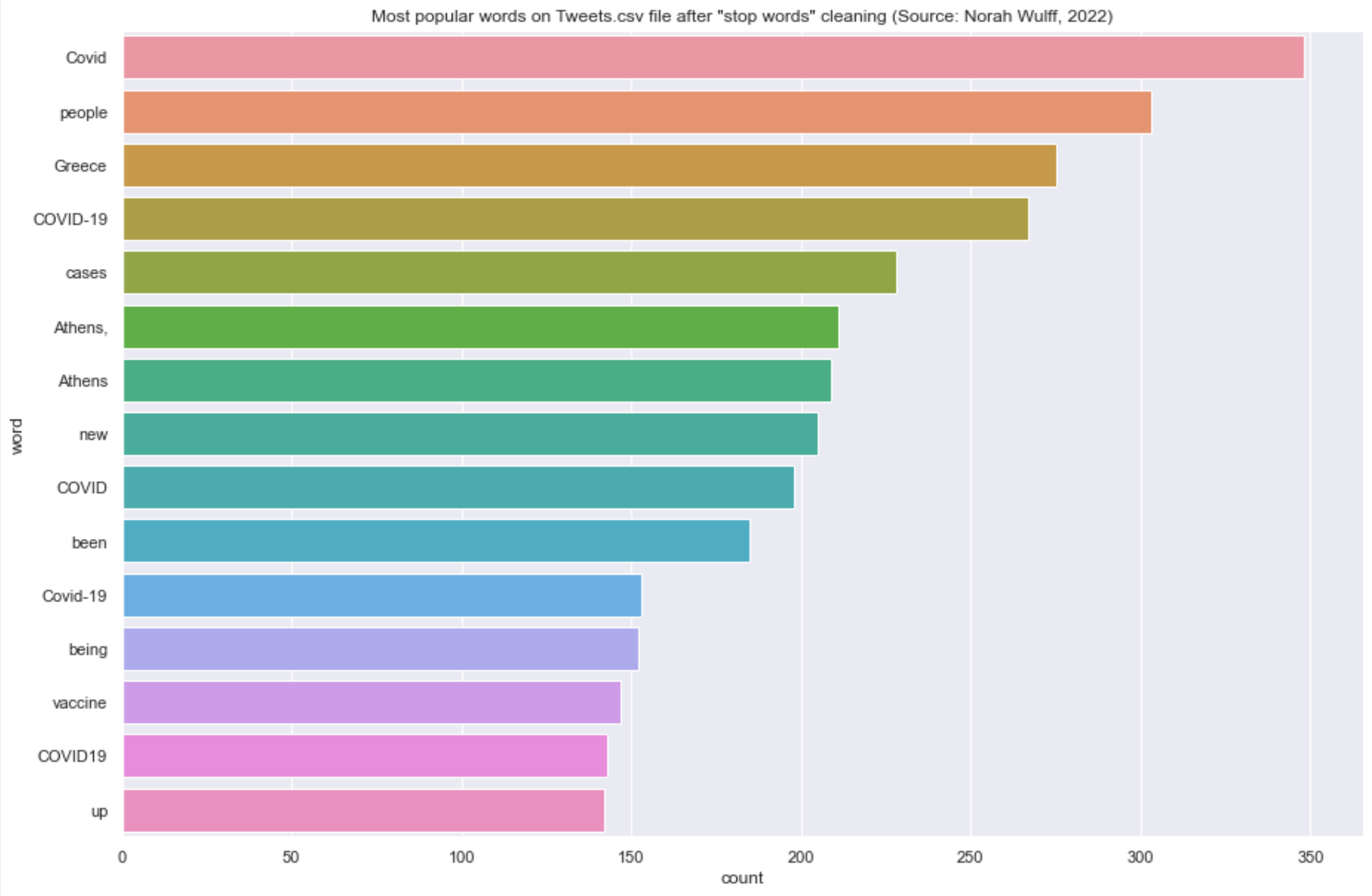
Figure 2.14: *Most popular Twitter COVID-related hashtags*



Source: Javier Conde (2022)

2.3. Visualisations exploration

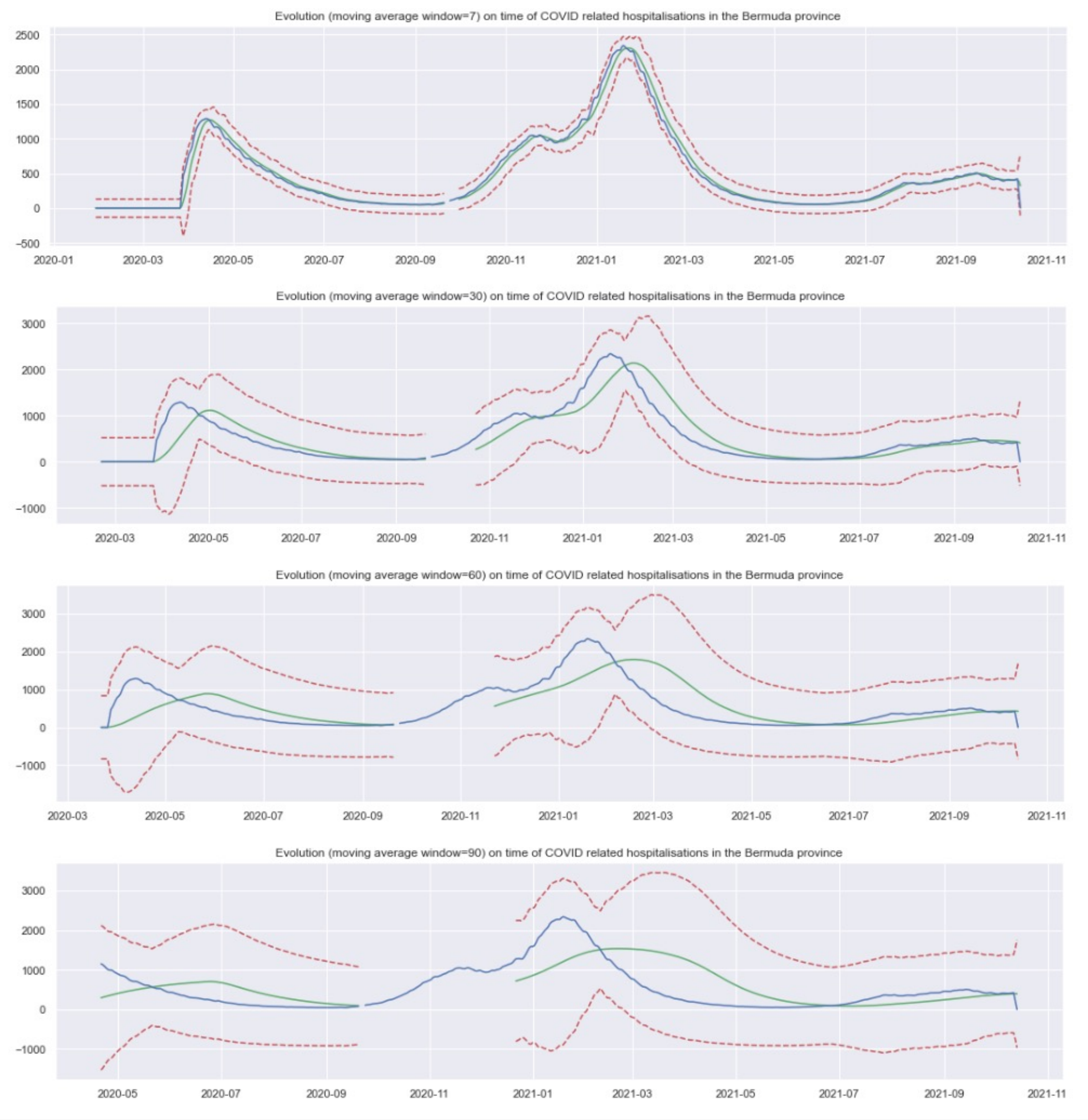
Figure 2.15: Most popular words on Twitter COVID-related after stop words cleaning

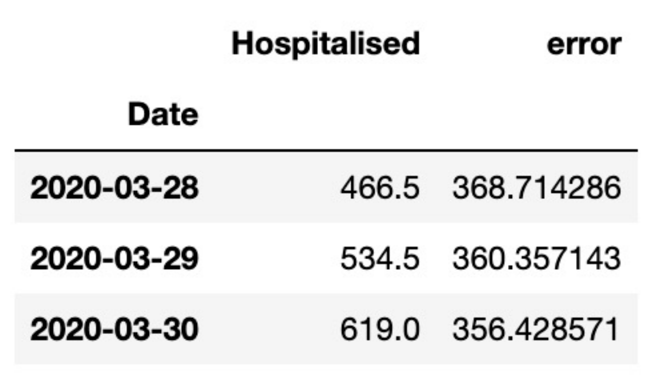


Source: Norah Wulff (2022)

2.5. Time-series analysis

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



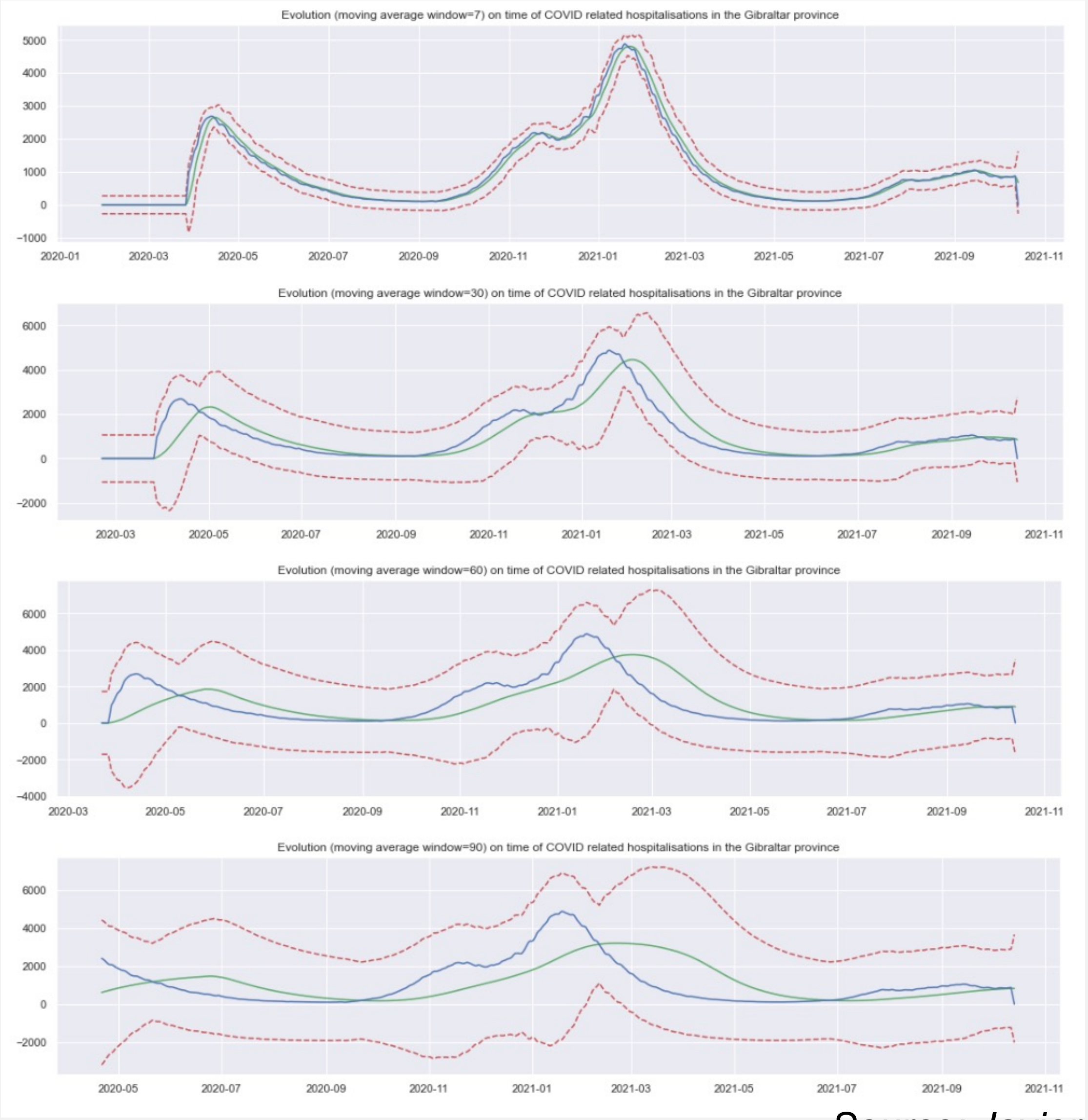


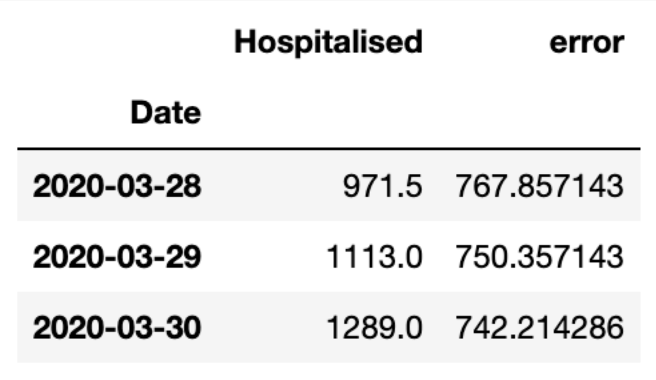
	Hospitalised	error
Date		
2020-03-28	466.5	368.714286
2020-03-29	534.5	360.357143
2020-03-30	619.0	356.428571

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*





	Hospitalised	error
Date		
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)

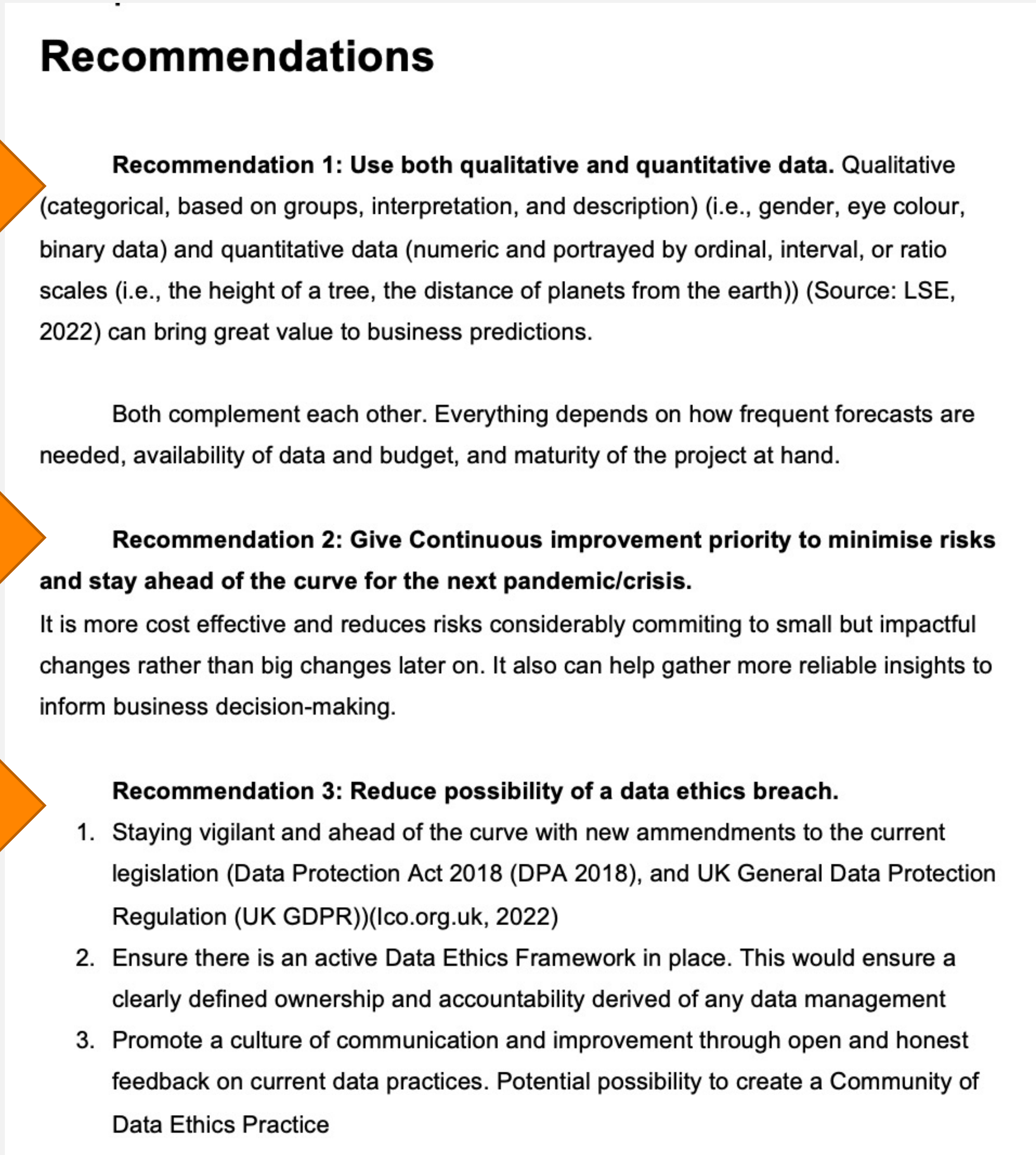
Figure 2.18: Further findings and insights

- Comparisons of cycles on vaccinations (evolution of 1st dose vs 2nd dose rollout) reach peak 4 months apart across the regions, which is in line with expected timelines
- Statistical insignificance of the rollout difference between regions (consistent 95.5%)
- Saint Helena, Montserrat, Bermuda and Gibraltar provinces with highest death/cases ratio. As only one death reported in both Saint Helena and Montserrat, Bermuda and Gibraltar are the provinces where we would advice to research aiming for a possible marketing campaign
- Deaths reaching early plateaus (early 2020) in Channel Islands, Turks, Isle of Man and British Virgin Islands with another increase in early 2021. Stabilisation of reported deaths coincides with rollout of vaccination program.
- Highest ratio reported recovered/ reported cases in Falklands (91.3%) and Turks (83.6%), followed by Gibraltar (81.5%) and Virgin Islands (70.23%). To consider when planning for the marketing campaign.
- Declining trend on hospitalisation from peak January 2021 while trends reverses with increase of reported recovered from the same date. To consider possible relation with vaccine rollout program.
- Only 0.8% of a total of 13,336 #coronavirus tweets analysed have hashtags #vacc-. Possible opportunity to research, moving forward with the marketing campaign, more popular COVID-related hashtags to be associated with (i.e. #COVID19 with 1,632 mentions)
- Moving average (MA) techniques can really help with forecasting and visualisation of data trends. Key feature is being able to estimate the error in a possible future prediction.

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)



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

Source: Javier Conde (2022)

Thank You