LSE DATA ANALYTICS CAREER ACCELERATOR

COURSE 2 ASSIGNMENT REPORT

Name : Kevin Wijaya Oey

Word Count: 1,100 words
(excluding references)

APPROACH

Organisation

- Covid cases and vaccinated data were merged using pd.merge( ) on python via a left join to ensure dataset completeness. A new primary key had to be made by concatenating the “Date” column and “Province/State” columns. Other columns did not represent the unique provinces well.
- The Dataframe was then subsetted by dropping unnecessary columns.
- Date values were formatted from string to datetime object for aggregation purposes.

Rationale behind Visualisations

- Line plots were utilised as they show how different metrics correlate by category over time.
- Tables were sorted in descending order to show the top province with a target metric.
- Colourblind palettes, legends and titles were employed to improve graph accessibility and expedite interpretation.

Outlier Management

- Outliers were not removed as they are meaningful in the context of a pandemic (eg may signal unresolved issues in data collection).
- In subsequent graphs the “Others” Province was removed as its values were too large and skewed the dataset.

ANALYSIS

WHERE TO TARGET FIRST MARKETING CAMPAIGNS

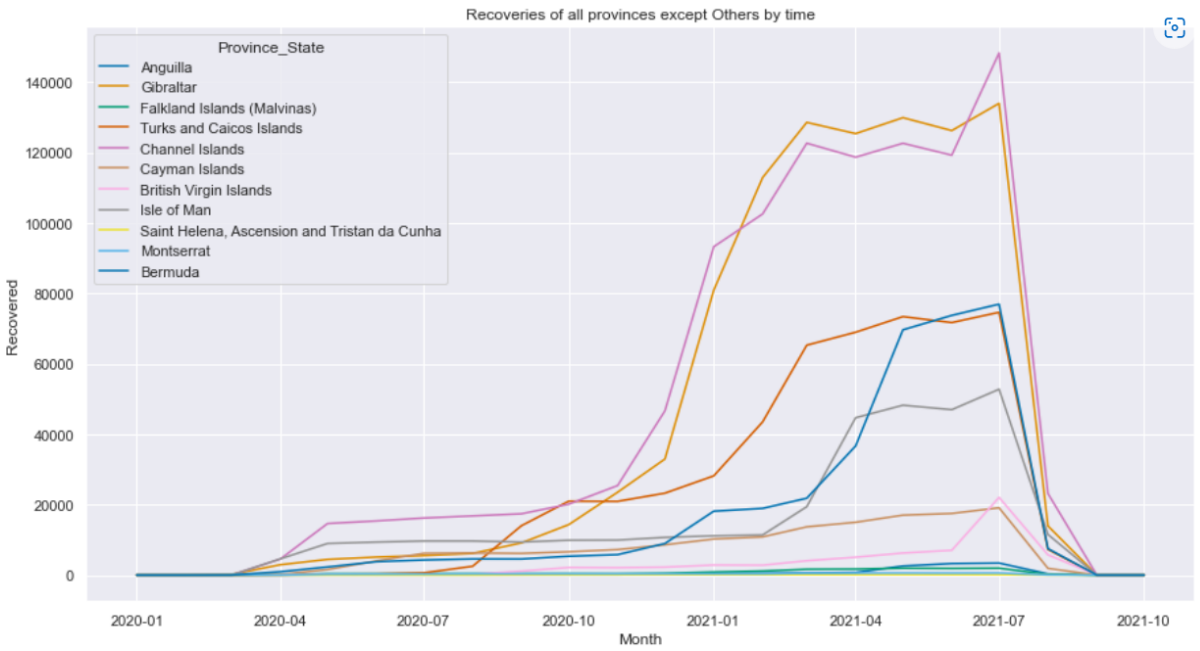
Province/State	First Dose	Second Dose	Difference in doses
Gibraltar	5870786	5606041	264745
Montserrat	5401128	5157560	243568
British Virgin Islands	5166303	4933315	232988
Anguilla	4931470	4709072	222398
Isle of Man	4226984	4036345	190639
Falkland Islands (Malvinas)	3757307	3587869	169438
Cayman Islands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

The data suggests that **Gibraltar** has the highest number of first doses (in absolute numbers).

Province/State	First Dose	Second Dose	Percentage of first doses only
Turks and Caicos Islands	3052822	2915136	4.510122
Isle of Man	4226984	4036345	4.510048
Anguilla	4931470	4709072	4.509771
British Virgin Islands	5166303	4933315	4.509763
Cayman Islands	3522476	3363624	4.509669
Channel Islands	3287646	3139385	4.509640
Montserrat	5401128	5157560	4.509577
Falkland Islands (Malvinas)	3757307	3587869	4.509560
Gibraltar	5870786	5606041	4.509532
Bermuda	2817981	2690908	4.509363
Others	2583151	2466669	4.509299
Saint Helena, Ascension and Tristan da Cunha	2348310	2242421	4.509158

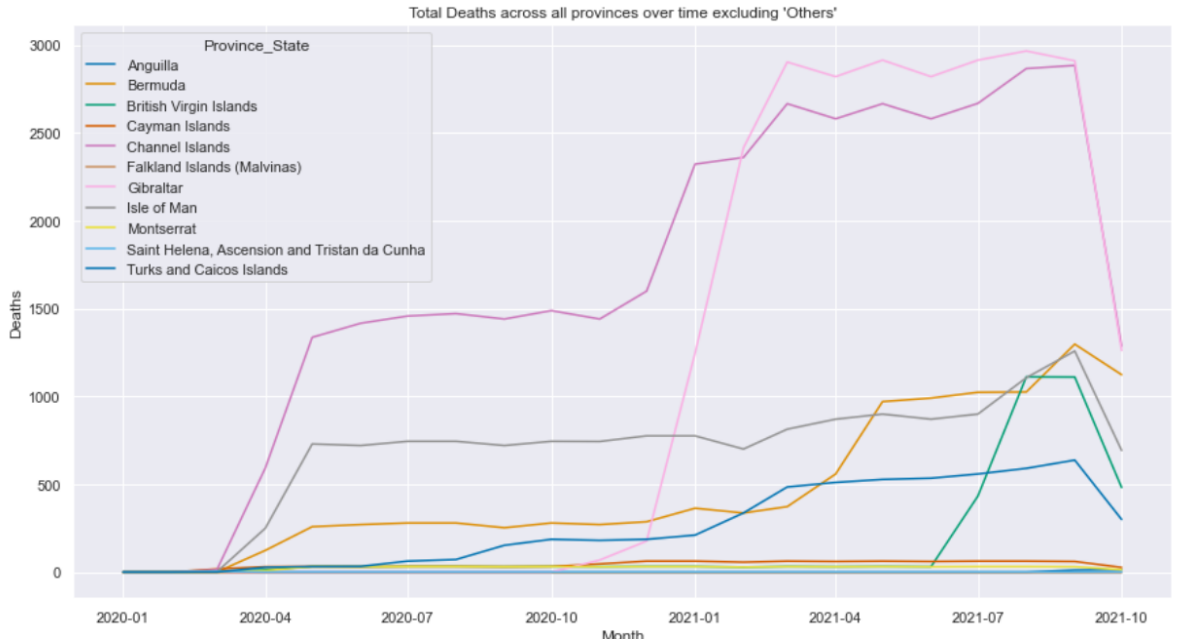
The **Turks and Caicos Islands** have the **highest percentage** of individuals who have received a first dose only.

AREAS WITH THE GREATEST NUMBER OF RECOVERIES



Recoveries **increase after October 2010** onwards (especially Gibraltar and the Channel islands), **peaking in July 2021** before plummeting to 0 in October 2021.

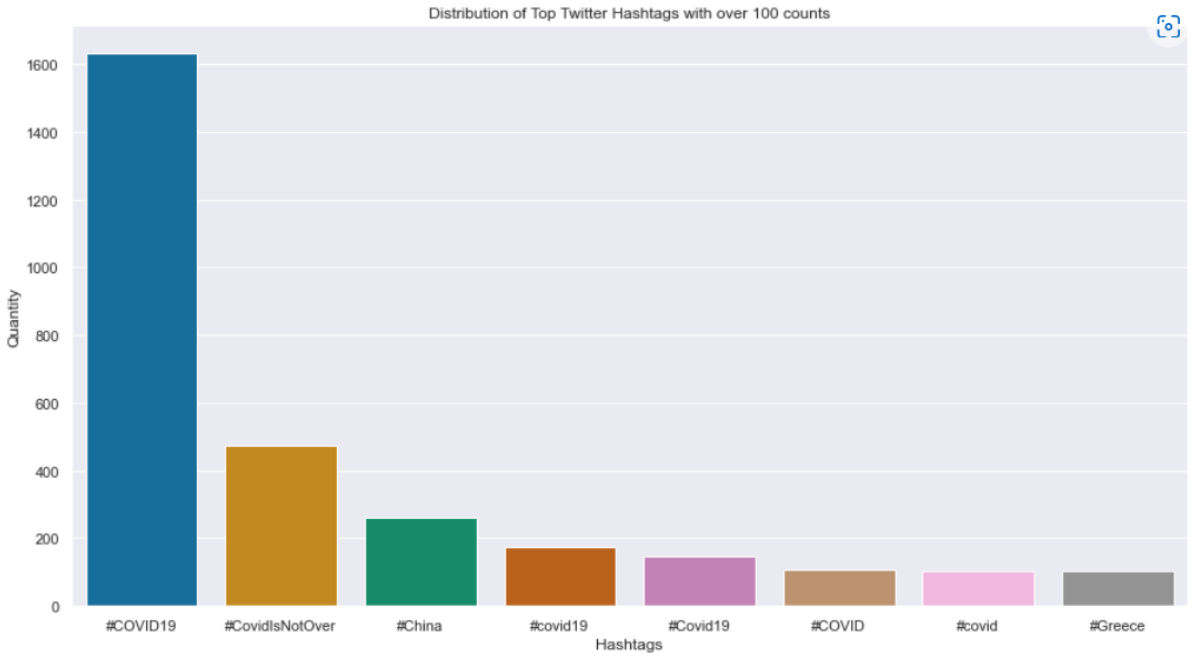
DEATH COUNTS PER REGION



Both Gibraltar and Channel Islands show a sudden increase in deaths from approximately December 2020 to mid February 2021, before plateauing and dropping at October 2021. The other provinces show a slight upward trajectory from 2020 to mid August 2021, before plummeting in October 2021.

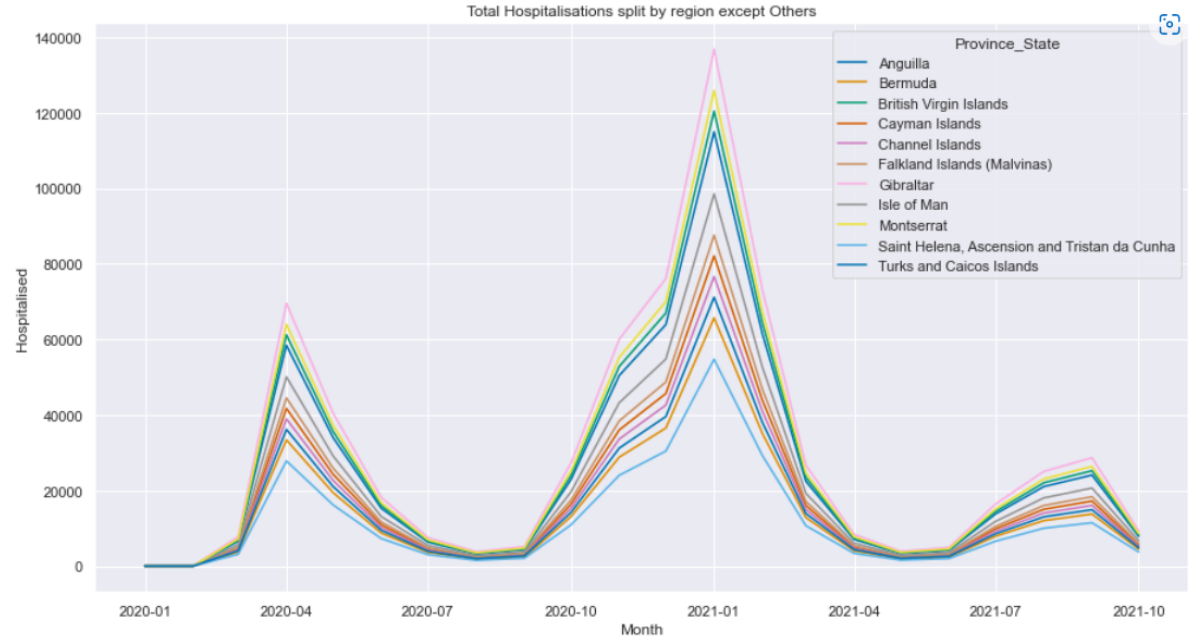
Whether deaths have reached a peak or not is inconclusive. It is possible for the Channel Islands and Gibraltar but less so likely for other provinces. Further monitoring across time is needed for clearer insights.

TWITTER DATA ANALYSIS



Covid- related hashtags are the most popular hashtags. However, while the public is aware of COVID as a danger, concerningly there aren't as much #vaccinations or #vaccine hashtags.

REGIONS THAT HAVE EXPERIENCED A PEAK IN HOSPITALISATION



The data suggests that 3 peaks have occurred in terms of Hospitalisations across all regions - in April 2020, January 2021 and approximately September 2021. The third peak shows a downward trend over time, implying that the peak is over. **None of the provinces** are experiencing a peak in Hospitalisations after October 2021.

LIMITATIONS

ANALYSIS AND VISUALISATIONS

- We could be missing out on certain customer segments by not scraping Facebook, instagram or other social media data.
- Many of the data visualisations are noisy when split to different provinces. Using a matrix showing region name by row and metrics as column could make it easier to compare / analyse trends.

DATA

- The cases dataset is missing values under the recovered or hospitalised columns from 5/8/2020 and 13/10/2020 onwards, implying that data collection may be incomplete.
- The data is unrealistic - it does not corroborate with actual covid data (eg different areas showing the same vaccination trends).
- The “Others” Province is a massive outlier and wholly skews the dataset.
- There are 2 missing values in the dataset for Bermuda on the 21st and 22nd of September 2020.

SUGGESTIONS FOR IMPROVEMENTS

- Split up "Others" into more provinces to increase granularity per region.
- Investigate further why outliers and zeroes exist. Is it an error in terms of data collection or are there contextual cues we are not aware of (eg no lockdowns implemented, or a superspreader event)?
- Provide more data columns (eg 7-day average, number of unvaccinated people)

QUESTIONS

QUANTITATIVE DATA, QUALITATIVE DATA AND FORECASTING

Qualitative data consists of data from observations that cannot be measured numerically (eg surveys). Quantitative data is data that can be measured numerically and has a certain order.

Forecasting methods differ according to data type. Working with Qualitative data would mean utilising the Delphi method or market research to gain opinions about a new service or policy. Similarly, working with quantitative data means utilising forecasts using time-series graphs and extrapolate from historic data or seasonal trends.

SIGNIFICANCE OF CONTINUOUS IMPROVEMENT

Continuous improvement is required because there will be bound to be failures or obstacles in any project and is necessary for an optimal result. Not implementing continuous improvement could lead to stagnancy and render current processes obsolete.

SIGNIFICANCE OF IMPLEMENTING A DATA ETHICS FRAMEWORK

Without a Data Ethics Framework, no one in our team will know how they are responsible for the data; we could risk team members leaking information unintentionally. A Data Ethics Framework aligned with company culture expedites ethical decisions.

ANALYSING DATA RELIABILITY

- **KURTOSIS VALUES**

```
# Find the excess kurtosis by subtracting 3
cov_vac_new_subset.kurtosis(axis=0)  -3
```

```
Vaccinated       2.427624
First Dose       4.016109
Second Dose      2.427624
Deaths          14.560852
Cases           25.268926
Recovered        5.032805
Hospitalised     1.326950
dtype: float64
```

According to https://pyshark.com/kurtosis-in-python/, a positive excess kurtosis indicates a leptokurtic distribution, which implies that all the dataset columns have extreme tails.

- **UNPAIRED T-TEST - Comparison with publicly available data from the WHO**

Assumptions

- Both datasets:
  - contain new daily cases
  - are independent

Approach & Results

- The average cases and deaths were aggregated across the same time period across both data sets (for the similar provinces only) and loaded as DataFrames.

We define two hypotheses:

- Null Hypothesis: μa = μb (the means of the WHO and provided dataset are equal)
- Alternate Hypothesis: μa ≠ μb (the means of both datasets are not equal)

If the means are equal, we can be confident that the data is reliable, otherwise we have to be sceptical about the data's reliability.

Average "Cases" p-value

```
P-Value:0.020504661348414038 T-Statistic:2.8061355238510046
```

Average Deaths p-value

```
P-Value:0.048374636300835334 T-Statistic:2.282355254100807
```

Conclusion

As the resulting P-values for both cases and deaths are lower than or equal to 0.05, there is enough evidence to reject the Null Hypothesis. The provided dataset's reliability is questionable due to its excessive kurtosis values and unequal means to WHO's data.

STRATEGIC RECOMMENDATIONS

- Avoid spending marketing budget in the Channel Islands
- Instead, allocate resources to increase the number of second doses in Turks and Caicos Islands, then Gibraltar (as it also has a very high recovery rate).
- Research on the best platform (eg print media) and incentive plan to convince first dose individuals to get their second jabs.
- Incentivise campaigns on Twitter or starting a #vaccinations or #second dose hashtags trend to encourage the public to share their vaccination experience.
- Deaths not peaking and dropping hospitalisations imply that the virus is circulating among vulnerable groups (eg the elderly), and that we need to divert hospitalisation resources to support the vulnerable.

ALTERNATIVES

- Using a dataset of questionable reliability to inform decisions could lead to unfavourable consequences, from misallocation of resources to further deaths.
- Consider procuring more accurate data from a new data provider. This will incur additional costs and time (two months for procurement and analysis).

REFERENCES PERUSED (Jupyter Notebook and report)

1. https://seaborn.pydata.org/generated/seaborn.lineplot.html
2. https://graphics.reuters.com/world-coronavirus-tracker-and-maps/
3. https://www.scribbr.com/statistics/t-test/
4. https://covid19.who.int/data
5. https://www.hackdeploy.com/python-t-test-a-friendly-guide/
6. https://pubmed.ncbi.nlm.nih.gov/31590904/
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7395797/
8. https://www.healthline.com/health/mental-health/covid-fatigue#signs-and-symptoms
9. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.kurtosis.html
10. https://pyshark.com/kurtosis-in-python/
11. https://seaborn.pydata.org/generated/seaborn.lineplot.html
12. https://stackoverflow.com/questions/6488928/where-are-the-ampersand-and-vertical-bar-characters-used-in-python
13. https://www.geeksforgeeks.org/filter-pandas-dataframe-with-multiple-conditions/
14. https://www.scribbr.com/statistics/statistical-tests/
15. https://seaborn.pydata.org/generated/seaborn.relplot.html
16. https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/
17. https://www.geeksforgeeks.org/python-pandas-dataframe-isin/
18. https://www.geeksforgeeks.org/python-pandas-series-value_counts/
19. https://forum.freecodecamp.org/t/attribute-error-axessubplot-object-has-no-attribute-savefig/460255
20. https://stackabuse.com/how-to-rename-pandas-dataframe-column-in-python/
21. https://notes.shanakadesoysa.com/Python/Pandas/MonthBegin_and_MonthEnd/
22. https://www.w3schools.com/python/python_for_loops.asp
23. https://www.askpython.com/python/examples/subset-a-dataframe