Report: Covid 19 Vaccinations for UK Government

1. Introduction

The analysis on the Covid case study was conducted using the analysis tools and approaches taught during the modules 1-6 in Course 2 Analysing Data with Python. Python libraries were used that were within the scope of the project and the Jupyter Notebook was uploaded to the GitHub Repo.

2. Background / Context of the Business Problem

The UK government wants to promite the vaccination program with a marketing campaign in order to increase the proportion of people who are vaccinated.

What I was aspiring to answer with this analysis is: Which regions are most likely to be successfully targeted by a marketing campaign to promote the vaccinations?

To answer this question, I wanted to find out whether a region has the highest proportion of people who are unvaccinated and who have the highest mortality rate, as this region would have a higher likelihood of success.

3. Setting up a GitHub Repo

The first task was to set up my personal Github Repository with which I was going to manage the project.

The most helpful aspects of GitHub is the versioning control and the collaboration tools. This makes it very useful to not accidentally overwrite important files and keep track of who changed what based on which reasons.

4. Data Exploration

The Data exploration phase is important for the analyst to get familiar with the datasets and spot potential data quality issues. I started at high level and explored the datatypes and the min-max values to get a big picture.

I performed these with the print(), describe(), min(), max(), head() and tail() functions. After the basic exploration I go into the details and subset the dataframe. This is considered best practise as it saves computational resources and makes it easier to work with. The Gibraltar province was selected to perform the data exploration in detail. The subset can then be used to perform additional calculations,

and grouping. In this case I first determined the total number of people enrolling in the vaccination program and how the proportion of the first and second dose vaccination has changed.

The exploration should raise more questions to be explored at the later stages of the project. I would recommend coming back to the notes and questions as the project progresses.

5. Exploratory Analysis on the whole dataset

For this stage I was trying to understand the development over time and aggregate the numbers by month. I was subsetting the dataframes and only kept the relevant columns on the subsets so that it is easier to work with them. Then I performed the grouping to month level on the whole dataframe and split it also by the different regions.

The proportion of people receiving a first dose vs second dose spike from April 2021 and again in July 2021. The numbers are questionable since the proportion of the people receiving a second dose is larger than people receiving a first dose.

Therefore, the assumption that people receiving a second dose are within the population of people receiving a first dose may be questioned. However as there is no clarity on this, I would not try to get lost in the details but view the data as it changes over time.

Considerations at this stage should be whether the metrics and calculations are the correct ones that would lead to any actionable insights. Also, to which extend the data should be grouped or aggregated is very important as it would lead to different decisions depending on how the data is displayed.

6. Data Visualization

In the data visualization step I was hoping to find some more insights regarding the trends and developments over time.

I created various subsets to produce the visuals that I believed would solve the questions. For the first visual I tried to see the differences in the first vs second dose population but there were no actionable insights from that. I used the formula given in the assignment tasks and if that was correct then all regions had the same proportion of people who are fully vaccinated vs the ones who are yet to receive the second dose.

Understanding the death rates gave a bit more clarity on the trends but looking at the recoveries I had again more doubts regarding the dataset at hand. I tried to distinguish the regions with labels but realized it would be easier to do it after exporting the visual.

7. External Data

The advantages to external data is do add more depth to the analysis. Diversity in the data that is included can provide hints or blind spots that the analyst may not have looked at. It can also add

credibility to the analysis if the trends for external data correlates to the findings from the analysis. Of course, correlation does not mean causation but in certain cases it may be helpful.

The disadvantages are the resources it takes for web-scrapping. Authorization and security as well as ethical considerations need to be included when performing web-scrapping and this can take up a considerable amount of work. Depending on the project scope or budget this may outweigh the additional benefits of including external data.

For this project external data would make sense to include to understand the population's attitude towards the vaccination program. The marketing campaign and message will differ significantly if the population has a rather positive or negative view of the vaccination program. Therefore, an analysis of the twitter data across the different Provinces/States could help. However, people who use tweets are not necessarily a representative sample, so I would also be careful before committing strongly for an action based on the twitter data only. A randomized survey could be a better assessment of this.

8. Conclusion and Recommendations

As the death rates are still increasing for a few regions, one consideration could be to focus the marketing for the vaccination program on Gibraltar and Channel Islands.

These are also the regions with the highest number (absolute) where people only received the first dose of the vaccine.

Another approach could be to look at the sub-regions and focus on those sub-regions where the population who have received a first dose is the largest which is Latin America and the Caribbean.

However, after working with the dataset throughout the project, there are so many questions around the validity of the data, that I do not feel in a position to make a recommendation and action step based on these insights gained from the data. Therefore, my recommendation is to restart the project, and collect the data again. Once the data is collected, we can run the dataset through the same set of codes and compare the results with the first initial findings.