# DA201 Assignment: Report.

## Introduction

The UK Government is seeking to boost uptake of COVID-19 vaccinations through a data driven marketing campaign with the objective to increase the number of individuals who have taken two doses of the vaccine. The government is seeking trends from its data on COVID-19 to inform an appropriate marketing strategy.

## Objective

Use the data provided to generate novel and compelling insights. Particularly, create stunning data visuals and incisive analysis that will meet the government's stated objective.

### Key Assumptions

1. Audience familiar with basic statistical concepts.
2. Data provided whilst thematically real world is masked and not intended to be accurate.
3. The government is focusing on increasing the number of individuals taking up the vaccine with the broader objective of improving public health and reducing ancillary impacts.

### Analysis Approach

All data wrangling, analysis was done using on Jupyter Notebooks. The output along with the underlying data are stored on my GitHub Repository.

Initially, I spent some time exploring the data characteristic. In the process, it became apparent that there were some issues:

1. The scale of the data is enormous; the magnitude of the difference in scales across the dataset would make effective visualisations near impossible. A degree of data transformation techniques would need to be deployed; I used log scales.
2. The data for the region "Others" was clearly the UK Mainland given its relative scale. The magnitude of the difference in scale of the data here versus the other regions is even greater and risks masking meaningful insights if not appropriately re-factored. Again log scales were useful.
3. Alternative would have been to exclude "Others" altogether. I decided against this given it is undoubtedly centric to the total number of observations. Aside from yet another scale problem, the data from this region did not stand out versus the rest of the sample. I removed this relative analysis from the submitted notebook as it was not called for within the rubric of the assignment; I did not want to distract the audience with excessive information.
4. I filtered each of the two quantitative datasets for a smaller subset to run basic analytics so that I could use the process to better inform my analytic approach to the broader data.
5. I have made detailed notes within the accompanying Jupyter notebook to record my approach to analysing the data, the rationale for the selection of specific analytical methods and the codes deployed against each analytical objective. These are omitted here in view of the limited word count available. Similarly, visualisations are not reproduced.
6. Data was analysed using time-series, aggregation and correlation techniques.
7. I iteratively discovered neater/briefer codes for the same analytic objective. I intentionally did not replace the earlier codes so as to generate an audit trail.

8. I cleaned the data lightly after deploying standard statistical methods for outlier identification. I felt that the native distribution should be retained as far as possible:
    a. It is so imperfect that it is very easy to get heavy handed and thus entirely alter it beyond its use case.
    b. It might lead to losing the ability to generate any insights at all and thus risk not delivering versus the government's stated objectives.
9. I wanted to generate a stunning and centric visualisation that would underpin my storyboard. Progressively it became evident that there were some linkages between the variables that don't immediately stand out. Asides the previously stated scale problems, some linkages were time lagged whilst others were through second or third order relationships which to be truly meaningful require much longer and ideally cleaner time-series data. I eventually landed on creating a correlation matrix which I feel does fulfil the brief.

## Visualisations and insights

1. We cannot analyse what drives first dose uptake with the data available.
2. A targeted marketing campaign is an appropriate technique. COVID-19 is by far the most trending topic on social media after two years since the onset of the pandemic. Such platforms should generate a disproportionately high number of eyeballs on the campaign content.
3. The data shows that all regions are already heavily vaccinated; the incremental gain to public health will be in ensuring that regular boosters are taken up when offered and individuals become eligible. There is little room to discernibly increase vaccination uptake beyond the c. 95% who are already double jabbed in all the regions.
4. Vaccination itself as a hashtag doesn't trend in our Twitter data; it is not in the top 10 hashtags. Campaign should focus on the bigger story.
5. The focus needs to be on the success of the vaccines in beating the worst effects of COVID-19, dramatically reducing hospitalisations and potentially deaths rather than beating case numbers

## Patterns and trends

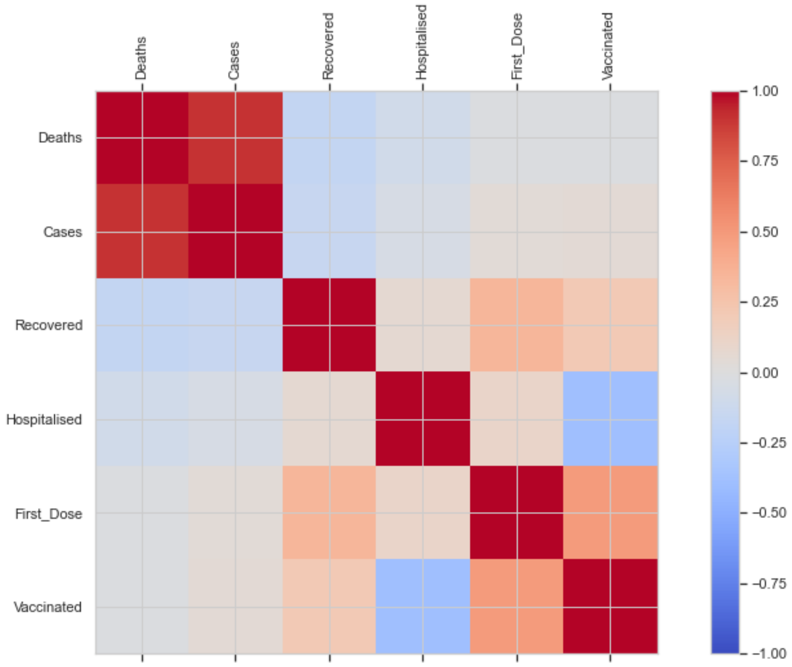
1. Although deaths and hospitalisations peaked a long time ago, they are again uptrending.
2. A government campaign to boost uptake of vaccines is timely.
3. There is a slow decline in the uptake of the second dose; the population should be reminded of the vaccine's effectiveness.
4. There is lagged negative correlation between the worst effects of COVID-19 and the relative uptake of both first and second dose of the vaccine.
5. Hospitalisations during the recent peak were much lower than previous peaks in 2020 when the vaccine was not available.
6. Any campaign should target all regions: the broad direction of the data in all regions are in sync. If resources are constrained, the campaign should set out by targeting the region with the highest absolute numbers which is "Others".

## Recommendations

1. Have a targeted marketing campaign on social media across all regions.
2. It is timely do so now: there has been a recent decline in uptake of the second dose even whilst COVID-19 is still highly trending in public opinion.
3. In it, tell the story about the potential risks being generated by the failure to take up the second dose.
    a. Rise in cases results in more deaths; vaccines may reduce severity of cases.

b. Hospitalisations and accompanying strain on public infrastructure significantly reduced once population is fully vaccinated.

4. Use a novel and impact visualisation such as the one below to support the storyboard.



***Get the jab & beat the blues.***

Further analysis is required as follows:

1. First dose uptake over total population.
2. Introduce more recent, real data including for the period when clinical trials were run.
3. Much deeper analysis of data required from Twitter and other social media for the entire duration of the pandemic. Need to understand if there is any link between social media sentiment and uptake of vaccines.
4. Need to introduce spatial autocorrelation techniques.
5. Continue to extensively collect and mine data to explore predictive capabilities.

***1124 words***