Jake Evans

Introduction

The UK government wants a data-driven analysis with Python to understand COVID-19 trends to inform a marketing strategy to increase fully vaccinated individuals. This investigation aims to understand the vaccinated individuals both fully and only first dose by region. Death, hospitalisation, and recovery data will be analysed, plus public sentiment towards vaccinations through Twitter data. The government can then prioritise regions and increase the effectiveness of marketing campaigns to increase the vaccination rate.

Analytical Approach

The head function demonstrated that the Covid Cases and Vaccination data provided contain insightful quantitative data. There are 3 data types: object, float, and integer. They both have 7,584 rows and 12 and 11 columns respectively. They contain values for Province/States by date. Cases, Deaths and Recovered numbers are cumulative over time, unlike the other columns. These are useful columns to help prioritise the marketing campaign, however, region population could add greater context. The data lacks vast historical data for any forecasts, especially with large external events increasing the volatility of data such as outbreaks or vaccines.

Values for Bermuda 21/09/2020 and 22/09/2020 were missing in the Cases data. The back fill method was chosen most appropriate to replace them. Hospitalisation numbers stop being recorded from August 2021 onward, meaning recent data is unreliable. There are no missing values in vaccinated data.

Covid Cases and Vaccinated data frames were merged using the first 8 columns. Every column was dropped except Province/State, Date, and the integer columns. Additional columns were generated to answer the business questions, total of first dosed only individuals and percentage of first or fully dosed individuals. Appropriate groupings and aggregations were applied to date or region. The 'Date' column was converted from an object to datetime.

Within the Tweets data, FOR loops were applied to strip the hashtags from tweets then added to a list to be filtered for keywords to gauge public sentiment. Qualitative data is not as measurable, plus there is no region data, affecting this data's relevance in the decision-making.

A rolling-mean function is provided to assist understanding hospitalisation peaks. The merits of this function in helping the government will be tested.

Visualisations

Aggregating by region helped determine locations for further analysis and generate bar charts. 'Others' is a province/state that could skew the data with 138,237 deaths (next highest being 100), thus it's removed to improve visualising the combined regions. Aggregation and bar charts also lack context or could be outdated. Grouped bar charts illustrate totals and percentages of first only and fully dosed individuals of each province/state allowing comparison for the different regions. The graphs use complementary colours to distinguish the different values. A bar chart with a colourblind palette was selected to illustrate the most used COVID hashtags for ease of comparison, but the plot cannot display all values.

Jake Evans

Observations of trends over time are needed to prioritise regions. A function was created to plot each column against time in a line plot, helping understand if region deaths have peaked, are recovering or if vaccinations have increased. These were created overall and by each region to help vaccination prioritisation. Line plots effectively tell a story over time, however lack of data in some provinces cause misleading figures. 'Date' was transformed to be monthly, allowing for better interpretation.

Patterns/Trends

From descriptive statistics, Gibraltar had the most hospitalised individuals of any region on a day (4,907), prompting further analysis. Viewing deaths over time, 96% of Gibraltar's deaths occurred in a 3-month window before plateauing.

The regions with most first dose only individuals are Gibraltar, Montserrat, and the British Virgin Islands, with large differences between regions. The ratio of individuals who received one dose for each region is 4.5%, making this not useful for government decisions. Unvaccinated data would help give further insight to these findings.

Deaths plotted over time sees cycles of increases then plateaus but has not yet peaked. Each region peaked except Isle of Mann, Channel Islands and Turks & Caicos Islands, indicating they should be highly considered as marketing targets.

As 95.5% are fully vaccinated, expectedly they increase rapidly on the line plots. After introducing vaccines, recoveries increase and deaths level out, highlighting the effectiveness of the vaccine for marketing messages.

The Channel Islands had the most recoveries (8,322), largely from having more cases. The Isle of Mann and Bermuda have lower recovered to new cases ratios (48% and 45% respectively) than the Channel Islands (69%), making them potentially at higher risk, however, absent recovery data from August 2021 onward, decreases the reliability of these metrics.

The top trending hashtag was '#covid19' (2,201 times) then 'covidisnotover' (635). The most popular hashtags including 'corona' or 'vaccine' was '#coronavirus' (262) and '#peoplesvaccine' (84) respectively. #COVID19 was used in 56% of tweets, showing it was heavily discussed. To further research, favourites and retweets for each hashtag could be measured and grouped by region.

The function created by the consultant generates a timeseries graph to forecast hospitalisation numbers. It was tested with hospitalisation numbers for the Channel Islands on a 7-day window. The function is helpful in viewing actual values vs. the 7-day mean. To be effective, there needs to be plenty of historical data. Ultimately, there is little data on COVID, and with the introduction of the vaccine, previous data may not be relevant either. It is good in visualising the status of hospitalised individuals and short-term decisions but has limitations for long term forecasting.

Recommendations

- Channel Islands and Isle of Mann have high cases and deaths haven't peaked, therefore should be prioritized.
- Gibraltar, Montserrat, and British Virgin Islands should be next as they have the largest one dose population.

Jake Evans

- Use data on increasing recoveries and deaths plateauing as a marketing message.
- Increase vaccination uptake potential by finding regions of tweets.
- Use the rolling-mean function to make short term decisions only.

To improve:

- Track vaccination rate/recovery metrics after marketing campaign launch and procure unvaccinated population data.
- Improve data collection methods to reduce inconsistencies in the data, i.e., missing recovery data.