COVID_Assignment_Course_2

Week 2 Observations The data in the Gibraltar dataframe doesn't look correct. There are more hospitalisations than cases from 2020-03-27. The data in the deaths, cases and recovered columns look cumulative whilst the hospitalised values are not. It is assumed that this is the data for that specific date. this is also the case for the full cov dataframe

There are default indexes on both the vac and cov dataframes, the data in both looks like it is recorded in the same order (ie index 0 of the cov dataframe corresponds to index 0 of the vac dataframe). This could be useful if possible to merge the dataframes together. There are columns which are unnecessary for this analysis such as Lat, Long and Sub-region name. I am going to use the Province/State to compare different areas.

Are there any visualisations that could be added here to make it easier? Line charts of changing variables over time would be useful as this would make it easier to see how this data is behaving.

What would I like to explore further? I would like to be able to merge the cov and vac dataframes to allow me to analyse the impact vaccinations has on deaths, cases and hospitalisations. I would also like to analyse the speed of the vaccine roll out in different states by viewing the data in a line graph over time.

Week 3 Notes

Gibraltar has administered the most vaccinations in total, but also has the highest number of people who have only had one dose. It is worth noting at this point that the summed values seem very high (5.8 million 1st doses administered in Gibraltar).

The data for province/state "Others" doesn’t seem correct as the ratio of deaths to recovered is not consistent with other provinces. This data has been dropped from the analysis as I feel I cannot trust it.

The merged dataframe contains cumulative data for some columns and non-cumulative data for other columns.

The percentage of people who after the first dose went on to have the second dose is consistent across all provinces at around 95%. I am unable to determine what percentage of the population has been vaccinated as I do not have the population data for each province.

A column for deaths per number of cases has been created to compare different Provinces as deaths data alone cannot be used as it is unknown what percentage of the Province population this is.

The dataframes now seem in a state that will allow the data to be plotted on visualisations that can be analysed easier. This will allow the business questions to be answered.

Week 4 Notes

Any other observations regarding the data?

- Some of the summed figures seem very large and not correct. For the purpose of this assignment they have been used.

What would your future data requirements be?

- Population data would be useful to analyse what proportion of the population has been vaccinated/infected/passed away.

Business Questions to answer Area(s) with the largest number of people who have received a first dose but no second dose

- In terms of numbers of people, Gibraltar has the largest number of people who have only had one dose, but they have vaccinated a lot more people. In all Provinces approximately 95% of the people who have had one dose have gone on to have the second.

Which area has the greatest number of recoveries so that they can avoid this area in their initial campaign runs

- Channel Islands have had the greatest number of recoveries

Whether deaths have been increasing across all regions over time or if a peak has been reached?

- The death rates are still on the rise across the majority of Provinces. Bermuda having the sharpest incline which is demonstrating a lot of deaths each month. The deaths per case in Bermuda is also relatively high, so this is a Province to focus vaccinations on.
- There are more deaths per positive test in Saint Helena but this is skewing the data as there are only 4 positive tests.
- Gibraltar has a high number of vaccinated residents, and the deaths are levelling off

What affect did converting the date to months have?

- Converting the date to months smoothed out the lines and made it more appealing to the viewer.

Are the visualisations of good quality?

- I think that the visualisations could be improved by increasing the font size of the labels and ticks, the auto colouring system used by Seaborn could be difficult for someone with colour blindness to decipher.

Week 5 Observations and Notes COVID related hashtags were trending a lot on Twitter from the dataset provided. One insight I have found when scraping data from websites, particularly hashtags on Twitter is to ensure that the data you trend allows for a range of different text formats as well as words and phrases that may be insightful. For example, COVID-19 could be written as Covid. Other hashtags that are insightful are the number of vaccination related tweets that were made.

Week 6 Observations and Questions

Question 2 The code is returning the biggest difference between the 7 day mean and the actual number hospitalised on that particular day. This data is useful as it shows spikes in hospitalisations (ie. when the virus was at it's most dangerous. It could also be used to analyse if there are any social, seasonal or environmental reasons for the spike the hospitalisations on these days)

Question 3 Question 3.1 What is the difference between qualitative and quantitative data? How can these be used in business predictions? Qualitative data is usually gathered by asking questions and then using the answers to perform analysis. There are no statistics and results are based on opinions or judgements of the people analysing the data. An example of qualitative data would be the results of an employee engagement survey. It can be used in business predictions to analyse which services/products customers like by analysing their feedback comments for example.

Quantitative data can be used for statistical analysis as the data is in numeric form. This means that historic data can be analysed to predict future trends. An example of quantitative data is good pieces produced per hour in a manufacturing process. This can be used in business predictions to predict future trends based on previous data, for example sales of items in different seasons.

Question 3.2 Can you provide you observations around why continuous improvement is required, can we not just implement the project and move on to other pressing matters? Continuous improvement is required to increase the quality and processes involved in a project. It's aim is to reduce waste to make a process more efficient. For data analysis, CI can be used to make small but impactful changes to help the business make better decisions.

Question 3.3 As a government, we adhere to all data protection requirements and have good governance in place. Does that mean we can ignore data ethics? We only

work with aggregated data and therefore will not expose any personal details? (Provide an example of how data ethics could apply to this case; two or three sentences max) We cannot ignore data ethics. Although aggregated data is used, this data has been formed by using individuals data. An example of how ethics could apply to this case is location data. Tracing to track the virus means that peoples locations are constantly being monitored. This data has to be handled ethically and legally.