LSE DA Course 2: Data Analytics using Python

Carl Szabo

The UK government was to focus on increasing vaccinations rates. They have provided us with files on covid illnesses and on vaccinations across different areas of the country. The government is planning on launching a series of marketing campaigns to promote the vaccine. To inform the approach, we are asked to identify trends and patterns in the data that could assist in these efforts.

Assignment Activity 2: Exploring the data

After initially loading the covid and vaccine data, I wanted to better understand the amount of data (e.g., number of rows and columns) we would be working with. I also wanted to check for any missing values in both datasets.

From here it's important to understand how the data is structured, what type of data is in each column, and then having a view of the data itself. I noticed the data type for the 'Date' column is an object and therefore may need to be changed, and some of the columns may not be necessary for the next stage of analysis (e.g., location coordinates).

Looking at Gibraltar as an example, we can run some descriptive statistics to get a sense of the data ranges in each column. Given the vaccine wasn't available, I decided to filter based on date to get a better sense of averages for administrating the vaccine. It's also here we can notice some of the columns appear to be cumulative whilst others are daily.

Exploring the data, not only in aggregate but also over time can provide two different pictures. Therefore, it's appropriate to do both. It would be nice to visualise this data, perhaps with a line plot, to help identify additional trends we might miss if just looking at data tables.

Assignment Activity 3: Combine and continue to explore the data

Next, I merged the covid and vaccination datasets to try and identify relationships using the index to avoid duplication. I also fixed the data type of the 'Date' column so I can use it more easily in the next stage of analysis. From here, I removed some of the unnecessary columns before the next stage of analysis. This will set us up to explore the regions in a bit more detail.

I next wanted to subset the data so I could better understand vaccination rates by region to understand where the government might want to target. Looking at differences between the total number of people who have had their second dose vs the total number with the first dose we can identify how many are still yet to receive that second dose.

In absolute terms, Gibraltar, Montserrat, British Virgin Islands (BVI), and Anguilla came up as the regions with the highest number of people with only the first dose. Percentages across all regions looked similar, but as we progress, we can investigate these four regions in a bit more detail.

Assignment Activity 4: Where should the government be targeting for their campaign?

Now we are getting to the point where we can really dive in. The first step is to isolate and visualise the highest number of individuals with only one dose, not only in absolute figures but also in percentage terms.

Next step is to create a new dataframe to understand the relationship between deaths, hospitalisations, cases and recoveries over time. The 'Other' region disturbs this analysis for Deaths initially, but after excluding it we see Channel Islands, Gibraltar and Bermuda as the most affected here. In terms of hospitalisations, the pattern here is consistent across regions.

Cases are also distorted by the 'Other' region, but after removing we see the Channel Islands, Isle of Man, Gibraltar, and Bermuda as having the most cumulative cases. With recoveries, Channel Island, Isle of Man and BVI have seen recent rises here.

With the recent recoveries data for BVI, and the total number of deaths being low for Montserrat and Aguilla, it would seem Gibraltar makes the most sense for the initial campaign.

Finally, we want to visualise the overall relationship between Cases, Deaths and vaccinations. I decided to use daily data for all 3 and then scale the numbers to improve visibility. What we found was that the vaccine drive has done a great job in reducing deaths overall. The recent spike in cases has not been met with much of a rise in deaths, so this is really great to see.

Assignment Activity 5: Analysing Twitter data

Turning to the sample of tweets received via a csv file, we were able to explore the 3900 rows for trending hashtags to quantify sentiment. From that data we can clearly see that COVID is still on people's minds. 6 of the top 8 trending hashtags featured ~ covid in one version or another.

This is informative as we know this is still a topic of conversation via social media channels like twitter. It would appear advertising through this route could be suitable as people are paying attention to the matter.

As a next step, it would be great to access data to confirm if people in Gibraltar are also talking about COVID too before running the campaign. If not, why not? What are they talking about? This would be helpful to know in order to shape the messaging.

External data can be a valuable source of confirmation. It can also open our eyes to something we didn't see in the internally sourced data.

Assignment Activity 6:

Picking up on where the consultant left off in their work, I think the 7-day moving average is a useful way to smooth the data and highlight trends.

After graphing the data by region, we notice similar patterns across the different localities. While I think it was valuable to look at the data this way, in the end I'm not sure we gained much from the analysis.