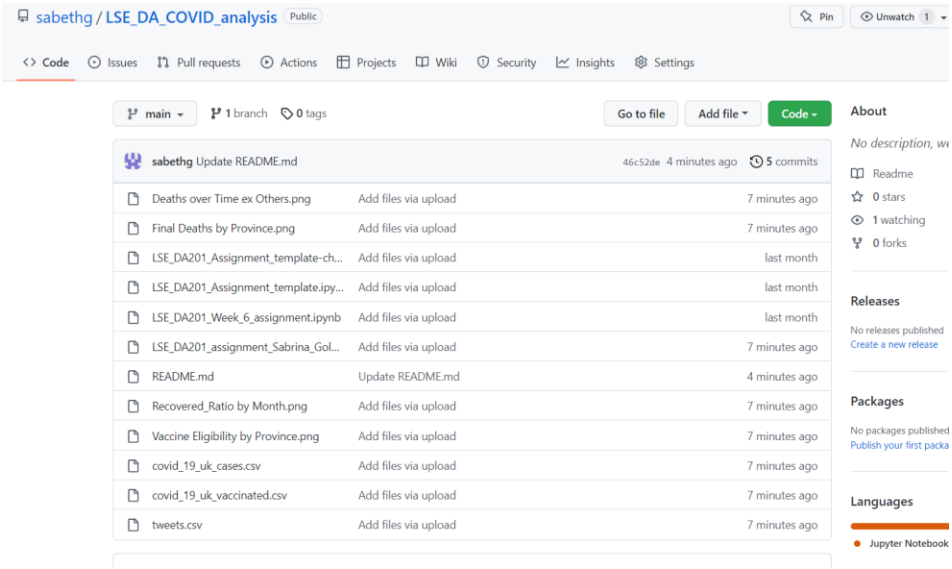
Report for LSE 201 Final Assessment

Activity 1

Github screenshot



Activity 2

I analysed two datasets comprising 632 days of data collected during the first 22 months of the pandemic. The goal of the analysis was to identify trends in key metrics (deaths, cases, recovered, first dose vaccine uptake, second dose vaccine uptake) across 12 UK provinces/states.

The cases_subset data contains one categorical variable (Province/State) and five numeric variables (Date, Deaths, Cases. Recovered, and Hospitalised).

The vaccination_subset data contains one categorical variable (Province/State) and three numeric variables (Date, First Dose, and Second Dose). The “Vaccinated” column was omitted as the interpretation of “fully vaccinated” has evolved over the course of the pandemic as further booster shots have been recommended in some instances. As the data in this column was redundant with the data in the Second Dose column, Ichose to keep this one only, for its interpretability.

As can be seen by examining the missing data per column, quite a bit of data were either missing to begin with or were removed during data cleaning.

For instance, it appears as though daily hospitalisations were not recorded until 27/03/2020, presumably because knowledge concerning Covid’s threat to public health were still emerging. This is inferred by the sudden jump from 0 to quite large values from 26/03/2020 to 27/03/2020. Rather than treat these data as reflecting 0 hospitalisations, they were coded as missing.

All provinces are missing data for daily Recovered from 05/08/2021 to the end of the data collection period.

St Helena has incorrect data for Deaths, Cases, and Recovered for the dates 18/03/2020 - 21/03/2020. Therefore, these were coded as missing.

The geographic area coded as ‘Others’ is missing Recovered data from 14/04/2020.

Based on these inconsistencies, it may be better to exclude Recovered from further analysis

St Helena has flat case numbers from 24/12/2020 through 14/10/2021. This is extremely unlikely to be correct, so these were coded as missing.

St Helena and Montserra appear to have incorrect data for Deaths. For instance, despite topping 4000 hospitalisations, Montserra reports only 1 death. St Helena reports 0 deaths despite topping 2000 hospitalisations. Given the known mortality estimate for Covid and the number of deaths/hospitalisation in other provinces,I do not think these data are reliable. Although, early in the pandemic, recording 0 deaths is likely correct, I cannot infer where the incorrect data begin. Therefore, deaths should be removed for these two provinces.

Gibraltar doesn’t report the first death until 11/11/2020, despite having had a couple of waves of increased hospitalisations, topping approximately 2700. Given Covid’s mortality, this seems unlikely. It is difficult to know the best way to handle this observation, given that the province’s later Death’s data seem reasonable. One method is to focus all analysis of deaths for Gibraltar on dates beginning December 2020. I have removed death data prior to this date (for Gibraltar). If further information can clarify death’s data prior to December 2020, then perhaps it can be included in future analyses.

Case numbers across provinces seem underestimates based on hospitalisations. As different provinces may have used different testing and reporting policies to measure cases, Imay be comparing apples and oranges trying to gain insight from these data. It may be better to exclude Cases from further analysis, unless Ican supplement the data with policy information which explains variability within this column

All provinces are missing Recovered and Hospitalised data for 13-14/10/2021.

Vaccination information begins 11/01/2021, with some individuals already reported as “fully vaccinated.” As a true beginning to these data should include first doses only, with fully vaccinated and second doses coming later, this cannot be correct. Therefore, fully vaccinated, first dose, and second dose data prior to 11/01/2021 are coded as missing.

This data set is also missing information from 13th and 14th October 2021. Therefore, in practical terms, the cases and vaccinated subsets of data finish on 12th October, 2021.

Cases_subset data for Gibraltar was examined to identify how deaths, cases, and hospitalisations evolved over time.

Activity 3

The cases_subset and vaccination_subset data were merged, as they contained data that overlapped in time and Province/State. A column was added to show the difference in uptake of the first and second dose of the vaccine. This metric is important as it allows us to identify whether eligible people are getting second doses of the vaccine and whether this pattern varies by Province/State. I can also identify how the stages of the vaccine roll out evolve over time. A positive vaccine difference indicates more people receiving the first dose than the second dose. A negative vaccine difference indicates more people receiving the second dose.

Activity 4

Using the vaccine difference metric, Gibraltar has the largest absolute difference in first and second dose vaccine uptake. Uptake of the first dose was greater than second dose in the early months of vaccine roll out. As the programme continued, it was more common for second dose uptake to outnumber first dose (that is, more people were getting their booster than were getting the vaccine for the first time).

It should be noted that the vaccine difference metric does not account for population differences. Larger vaccine differences should be expected when first dose uptake is large. A way to eliminate this bias is to look, instead, at the percentage of second dose vaccine uptake relative to the first.

The percentage of second dose uptake among the eligible population was nearly identical across Provinces/States (if this weren't an assessment, I would be suspicious that the data were intentionally generated, but I will assume this is just a wild coincidence). This suggests each Province/State has been equally successful in getting people to uptake the second dose.

In making a recommendation about where to target a new campaign, I would want more information about how "tried and tested" the campaign is. If the campaign is known to be effective (e.g., it has been tried and tested elsewhere and there is good reason to think it will work in this context), then the goal should be to test it where it can encourage the greatest number of people to get the second dose (i.e., Gibraltar). If the campaign is being piloted, then the goal should be to test it in a relatively small population, so that money is not wasted on a campaign that might not work.

Activity 5

The text of tweets was parsed to identify what hashtags people were using and how frequently they were being used. The programming for this is still under development, but indicative code for visualising findings has been included. In addition to the frequency of each count, I would like to examine clusters of similar tweets (e.g, #covid, #covid19) to identify the most effective hashtags for a topic (e.g., which has a higher count?). I would also look at dependencies between hashtags (e.g., given #covid, what is the most likely # to co-occur?) .

Activity 6

I used the plot_moving_average function on Hospitalisation data for the Channel Islands to visualise the basic trend in Hospitalisation numbers with less noise than using raw data. I created a bespoke function to compute mean absolute error and tested this with sample arrays. In practice, this could be useful for assessing model accuracy by comparing predicted and actual values.