**BATTLE OF THE POPULOUS**

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE - final assignment

*by Edward Jackson (25th August 2020)*

**Introduction**

The year 2020 has brought challenges to the entire global population. It is unprecedented for an event to affect the whole world at the same time but the current COVID-19 pandemic has made for some unique circumstances but very disparate responses.

As a school leader in the UK, I recognise the very real difficulties facing governments (both national and local), corporations, key services and – for the purposes of this project – education providers. I must be clear that I bring no political bias to this project but I’m certain that most of you will accept the impossible trade-off between ensuring that pupils access education (and secure their futures) and safeguarding the health of the population (without which there can BE no future). Note also that I make an arbitrary decision to highlight large Outstanding schools in this project; I stress that schools could be filtered in many different ways and, as someone who has never worked in a large or Outstanding school, I take this opportunity to celebrate the many excellent practitioners and leaders I have encountered in a variety of education settings.

In this project, there are many assumptions made about the variables affecting these decisions and the focus here is not on expert health advice. However, I will assume the role of local authority/government advisor, harnessing data to inform localised measures to balance both the safeguarding of health and maintaining education provision. This is not an exact science and no national government could claim to have the perfect solution as the situation inevitably changes from moment to moment.

Project goals

1. I will combine a map of Sheffield (a large city in the UK) with a GEOJSON overlay dividing the city into comparable population areas (i.e. relatively similar population in each). I will access and clean data to map the levels of infection in each of these areas. In planning for an inevitable increase in local lockdown measures this winter, I will access further data about school sizes and inspection ratings, dividing all schools into clusters; this will **assist the sharing of resource within clusters if infection interventions require school closures or central hubs to look after the children of key workers**. I will also highlight the large (above 400 pupils), Outstanding (inspection rating) schools to inform the capacity and location of expert support across the city and to help the local authority to target additional support as required.
2. Although the intentions listed above use APIs to access ‘current’ data, I recognise that this assignment requires the use of the Foursquare API ‘even if it is only Foursquare location data’ (cited from course instructions). Connection to this is patchy so I have limited my use of it to determine and **map the largest sporting venues to assist the local authority in enforcing infection interventions** (games behind closed doors, cancellations, other restrictions). Combining the Foursquare APIs and the other APIs outlined in the next section, I feel that this has allowed me to tackle my project aim and to present a coding framework with numerous add-ons possible in the future.

**Data**

Before explaining the rationale behind my choice of data, I summarise the data used in this project as follows:

* World map - openstreetmap.org (Open Licence) - Folium
* GEOJSON MSOA boundaries - ons.gov.uk (Open government licence v3.0) - via API
* Covid-19 infection data - coronavirus.data.gov.uk/cases (Open government licence v3.0) - via API
* School management and inspection data - assets.publishing.service.gov.uk (Open government licence v3.0) - via API
* Postcode-to-coordinates data - 'arcgis.geocoder' (Open Licence) - via API
* Sports venue location data - foursquare.com (Non-enterprise licence) - via API

Firstly, I must explain that Foursquare does not hold data on educational establishments for security reasons. Consequently, I have accessed data from other sources (using numerous APIs and I have used the Foursquare API to access location data for two big sporting venues only (as permitted in the course instructions) in order to alert the authorities of the need for cancellations/interventions where infection rates spike near and around their locations.

In order to present the infection rates clearly, I accessed a world map through Folium, centred on Sheffield (UK). I wanted to present infection rates clearly in a more localised format. I set up an API to access area boundaries through a GEOJSON file. This file provides geometry for each of the Middle layer Super Output Areas (MSOAs) attributed to the district of Sheffield (including the city). Last adjusted in 2011 in the UK, an MSOA makes for easier comparisons because it splits areas into similar population sizes (mean of 7200) i.e. denser, urban MSOAs will be smaller than sparser, more rural MSOAs. The COVID-19 infection data is updated by UK government on a daily basis, presenting data per week and a 7 day rolling figure. For my project, I use this data to present infection rates over 14 days (typically the length of quarantine in the UK) as I feel this gives a more certain indication of the public health interventions required (to arrest increases) and it reduces the 'noise' caused by short-term spikes.

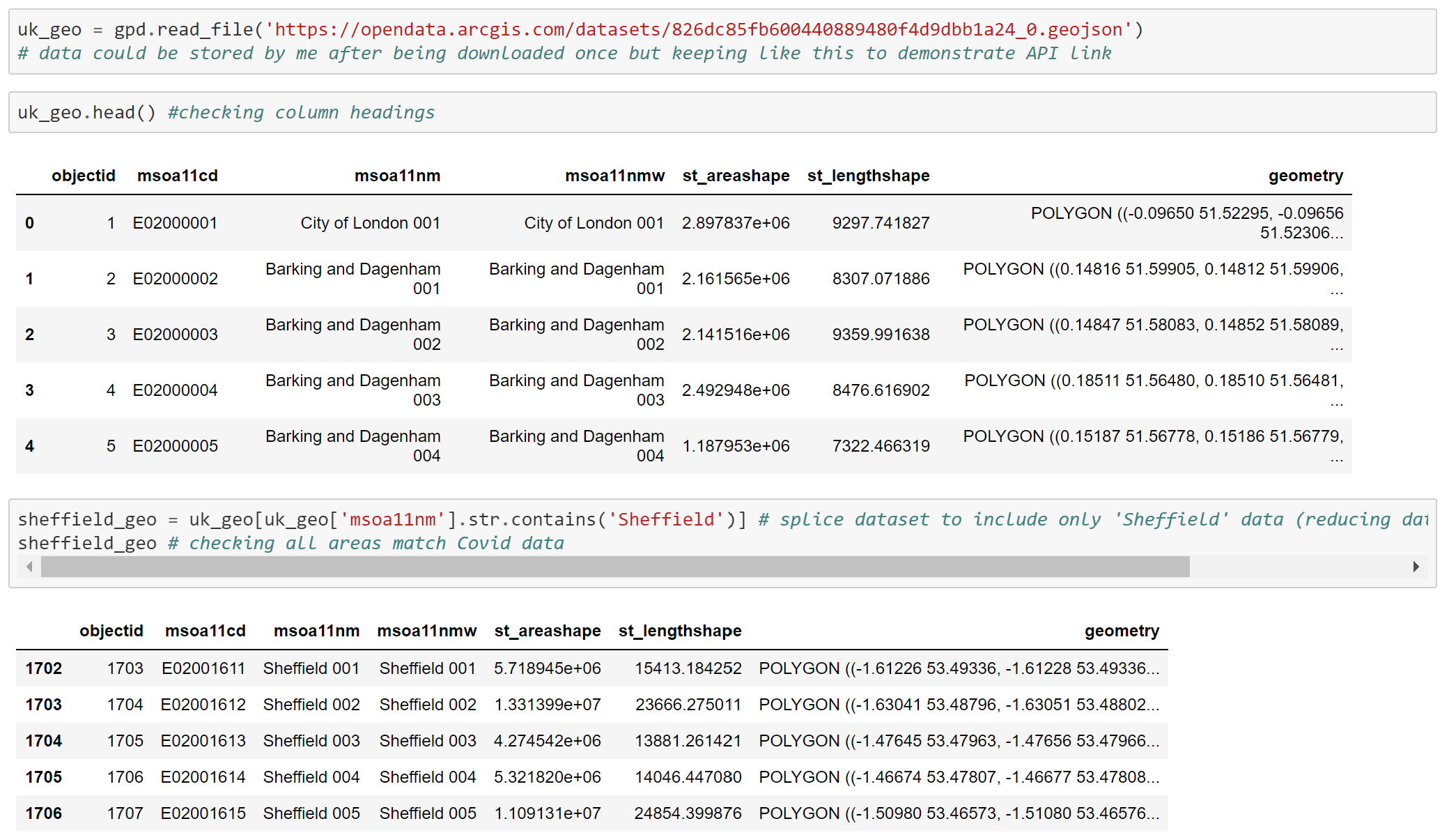
As an educator, I want to present a project which could assist local authorities in targeting educational support, mapping where larger, Outstanding schools are within the city. This will demonstrate where the greatest capacity for autonomous hub support already exists and where additional support should be planned in the case of infection increases disrupting local education services. This will also reinforce and promote the effective collaboration between schools and resources. This project could also be applied easily to any area of the UK within seconds by amending the District accessed from the various data.

**Methodology**

In this section, I will highlight the processes used and the decisions made throughout the practical part of this project. I will refer to sections in the Notebook directly.

Setup - From the outset, I imported and loaded the libraries required in the project. I used a prefix ';' at this stage to keep the notebook tidy, ensuring that the lengthy outputs generated by installations were not on display in the final presentation.

Importing, cleaning and shaping the COVID infection data - Accessing and presenting COVID-19 infection data is central to this project. Difficulties arose because the column headings in the data would change from day to day: one day the final column would be 'Last 7 days', then 'Latest 7 days', then missing altogether. My project presents infections over the most current (and complete) 2 week period so I tended to access 2 columns of complete weeks instead but this still needs adjustment as each new week is completed. Having selected the columns I required, and removing unallocated data, it was at this stage that I chose to reduce the dataload of the project by focusing on one city/district: Sheffield.



Importing, cleaning and shaping the geospatial data - splitting focus area into zones - The aim was always to present infection data on an interactive map. The GEOJSON file defining the geometry of each MSOA in the country is very large and takes some time to import via API. Given that I only required data for Sheffield at this point, it might have been prudent to save a local copy of a reduced GEOJSON file. One further difficulty with the GEOJSON file was that I had initially accessed an archived 2001 boundary file so its 'msoa01cd' key did not match the 'msoa11cd' key in the COVID data until I was able to import the 2011 (most current) boundary data. COVID and boundary data was merged (pictured above) to make it easier to check for missing data or mismatches and present data on the map. Indeed, I had to check the shape of 'Sheffield' COVID data and 'Sheffield' boundary data as initial attempts at mapping had holes: missing MSOAs. Again, this was due to having initially accessed an archived GEOJSON file.

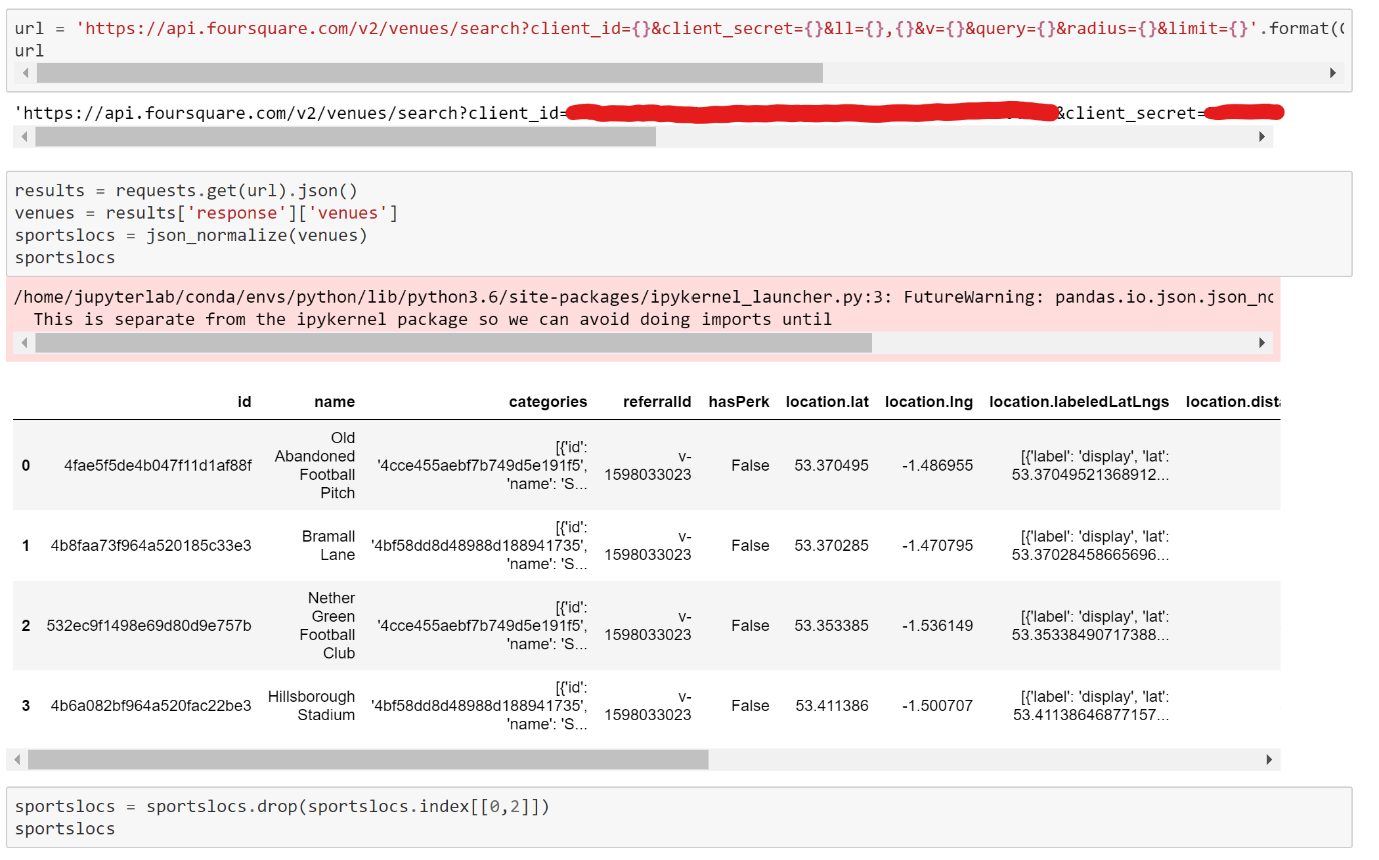
Creating a UK map, zoomed to the district and city of Sheffield, UK - Producing the first map through Folium was quite an easy step. However, the finer details proved much more time-consuming. In order to present infection data, I needed to comprehend the keys in the GEOJSON file for which I needed to do further research and, of course, further frustrated by using an 'out-of-date' boundary file initially. I was able to construct code to ensure that labels containing the MSOA name and infection data would appear when the mouse hovered over individual areas. Initial attempts also had a legend with a 'relative scale' i.e. the highest (but still very low) infection areas were displayed in red. I re-coded this to define the bins in 'absolute' terms to ensure that regular audiences would be able to trust what the colour coding would represent from day to day.

Importing, cleaning and preparing state schools' data, leaving school name, postcode and inspection Rating - As an educator, schools' data was always going to be critical to my project. Again, it might be easier to save this data locally rather than importing it via API each time the Notebook is loaded; the data is only updated every few months so the user would have to be mindful of this. For my purposes, the dataframe was reduced to desired variables only and Sheffield data was also isolated accordingly. A key issue at this stage was being able to plot each school on the map but the data file only contained postal codes rather than coordinates. I tried a number of free APIs to convert between the two (including api.postcodes.io) but they were even slower than the solution seen in the notebook. Further, there was a <space> within each postcode which was hindering this conversion. I added code to remove this <space>.

Filtering large Outstanding schools to provide 'area support' for children of key workers in case of lockdown - With school coordinates now added to the dataset successfully, I added markers to the original map. Again, I must stress that I am not saying that only large Outstanding schools have the capacity to be supportive; the chosen filter was purely arbitrary to demonstrate the feature and the flexibility for future use of the product.

Clustering schools in zones to aid sharing of resources and support in emergency situation - A further largely arbitrary decision was to divide schools into clusters. This was achieved using the algorithm 'K-means'. Originally, the elbow method for finding the 'best k' was used. However, I decided to override this decision based on my knowledge of the geography, catchments and dynamics between education providers. In my opinion, five or six clusters would be ideal in the case of Sheffield. Code was used to ensure that the school name and cluster number would also appear when the user clicked on each school marker on the map.

Importing location data for major sports venues using Foursquare API - Initial attempts to use the Foursquare API were frustrating and, given its apparent unreliability at times, I chose to use it in a limited way to access location data for large sporting venues in the city. The data provided for the venues was mostly irrelevant and even presented an abandoned football pitch as a venue. Once data was cleaned, two venues (the biggest professional football clubs in the city) were added to the final map.

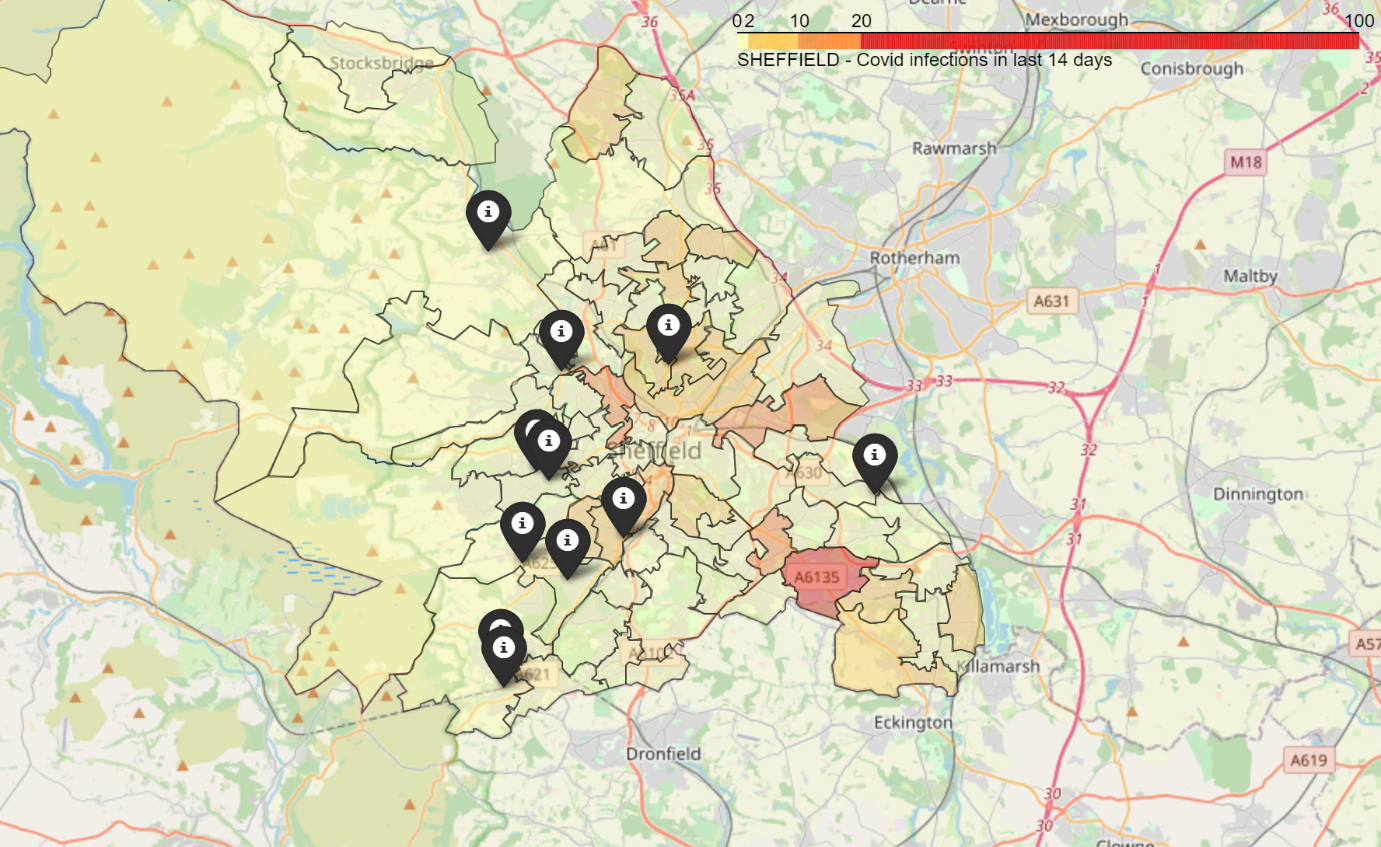


**Results/Discussion**

This project does not provide 'results' as such any more than we have global solution to the COVID-19 pandemic. Rather, the project is about potential impact and usefulness as a tool (amongst many) to inform on decisions aimed at reducing the spread of infection. Approaching this scientifically, I would prefer to conduct a user survey to evaluate the success of the project. However, my personal opinion is that the project achieves its goals but with some limitations:

Successes

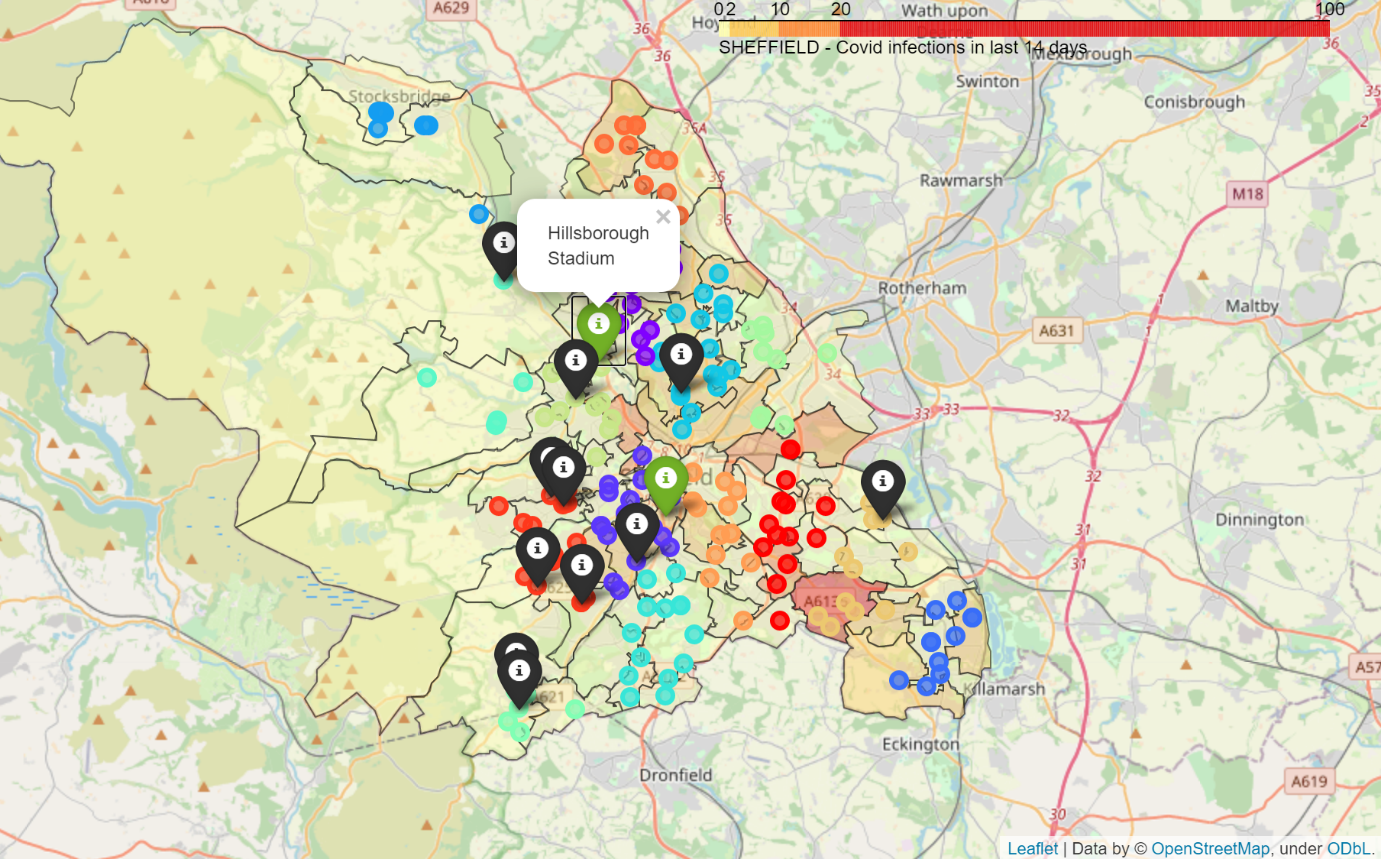
* the interactive map is easy to understand and information is presented clearly.
* the data presented is useful to the intended audience (local authority and education leaders) e.g. the map (below, extracted from Notebook) shows that there is a wealth of large Outstanding schools to the west of the city but emergency resources might be best targeted towards the east where capacity appears to be more limited.



* the amount of available data for schools (including columns not used in my dataframes) fires the imagination about how this project could be adapted in the future. It is highly flexible and could easily be tweaked to present data for other regions in England or even - given greater computing power - the whole country. Of course, there are also many other variables (e.g. staff absence, safeguarding history) affecting the capacity of individual schools to support their own pupils and/or other local schools.

Limitations

* the clustering of schools (below, extracted from Notebook) is somewhat arbitrary without an identified 'lead' school in each cluster to coordinate support efforts. Further work could be conducted to identify 'appropriate' schools near the centroid of each cluster, although there are many other reasonable solutions possible.



* the mapping of two sporting venues (green markers above) adds so much potential to extend this project. However, the information held in the Foursquare database for these venues was quite limited. My ideal would be to 'label' match information, including alerts on cancellations or restrictions.

**Conclusion**

This has been a long but enjoyable project and, in my opinion, it meets its goals and target audience. In the Results/Discussion section, I have highlighted some of its limitations but it is perhaps more useful to highlight these as positive areas of development without this initial, flexible framework. The COVID-19 pandemic is no fun for any of us and it has affected the liberties that post-war generations consider to be 'normal'. A key issue affecting populations across the globe is the clarity of messages from our leaders and getting the right information in the first place. Without a 'Dummies Guide to Covid-19', we must be innovative and patient. I hope that this project can inspire the postive, skilled and caring staff in our schools to lead the way on collaboration and 'reaching out' during this extraordinary passage in our lives.