**London districts and schools**

**Introduction**

Despite Brexit, London is a great city to live. It is one of the most prominent financial, cultural, historical and social centres of the world, which offers many exciting job opportunities for young people.

However, as most of the new parents in their late twenties and early thirties will tell you is how high the housing and childcare costs are here. The schools are divided into very expensive private schools and free state schools. Good quality free schools are scarce and hard to get into. A child must typically leave within one mile radius from a school to get accepted and it is not uncommon for families to move closer to the school of their choice to increase their admission chances.

This Data Science capstone project covers the following:

* Analysing London postal districts venue types and clustering the districts by their respective venues’ popularity.
* Analysing London school venues and locating state primary schools on a map, only mixed or girls only schools are selected and only the ones which have an ‘outstanding’ OFSTED (Office for Standards in Education, Children's Services and Skills) rating. The resulting map should display a 1 mile radius around the schools. It will be of help to parents to research their local schools and to plan their move if needed.

**Data**

For the purposes of the project the following data has been used:

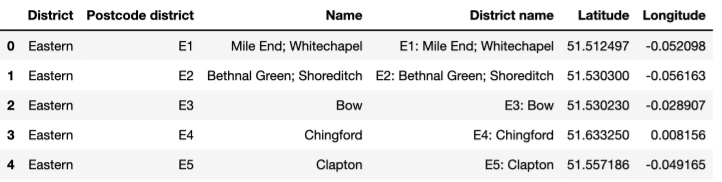
1. A list of the London’s postal districts, their names and postal IDs. The list was obtained from [1]. Python Nominatim functionality was used to obtain each district’s latitude and longitude coordinates.
2. The most popular venues and their types (e.g. a coffee shop or a theatre) for each of the London’s postal districts. These were sourced using the Foursquare API functionality. The popularity of each venue type was determined by the number of occurrences of that particular venue type in a district.
3. List of schools for each of the London’s district was initially sourced with the Foursquare API. However, this approach appeared to be not ideal: a) it displayed all the venues which included the word ‘school’ in the name, even locations such as bus stops or driving schools, which are irrelevant for the purpose of this project; b) it didn’t not display information about the schools like their ratings, if a school is private or state, primary or secondary, etc.
4. Data on London schools, their URNs, names, types, OFSTED ratings and postcodes. The data was taken from [2]. Nominatim functionality was used to obtain each schools’ latitude and longitude coordinates.

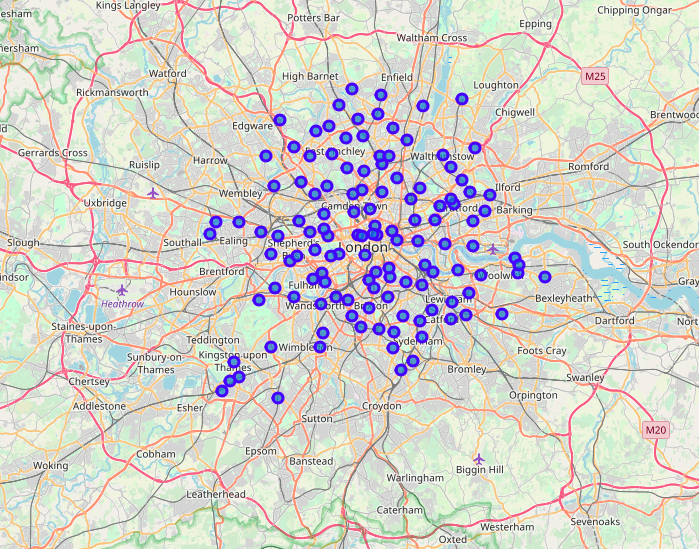
**Methodology**

*London Postal Districts*

Firstly, a csv file with 125 London postal districts and their coordinates was downloaded as a data frame using the *Pandas* library. The coordinates had been previously calculated by the *Nominatim* functionality of the *Geopy* library and stored in a csv file to avoid recalculating the latitudes and longitudes every time the code is run.

The following figure is the head of the downloaded data frame:



Secondly, the *Folium* package was used to create a map of London with the 125 postal districts marked on it:

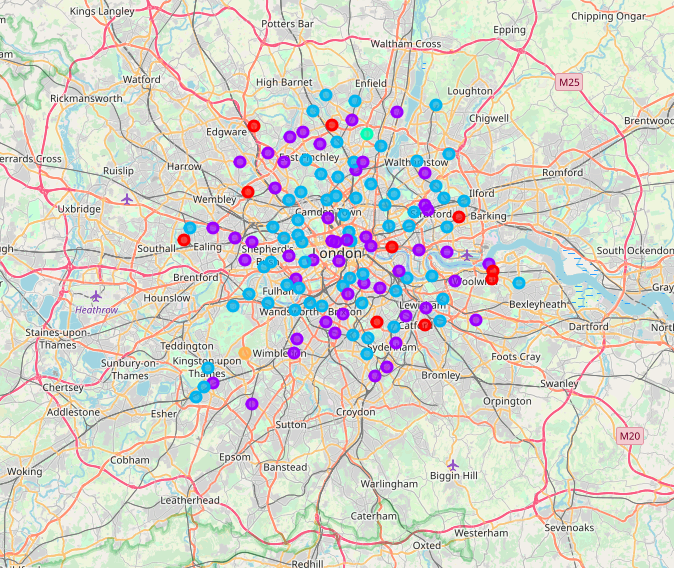
Thirdly, the *Foursquare API Explore* functionality was used to find the most popular venues and their types for each of the 125 districts; the limit on the maximum number of venues to be extracted was set to 50.

Altogether, there were 326 unique types of venues found. Below is the word cloud generated using the *Wordcloud* library to display the most popular venue types found in London:



Fourthly, the data was manipulated to select the 10 most popular venue types in each of the districts and to prepare it for the K-Means Clustering exercise.

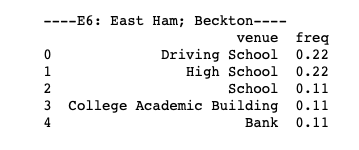
Lastly, the K-Means Clustering algorithm from the *Sklearn* library was run with K=5 and the results displayed in the map with each cluster assigned a different colour:



*London Schools*

Initially, the *Foursquare API Search* functionality was used to search for the ‘state schools’ within London postal districts. However, this approach did not yield great results for the following reasons:

a) it displayed all the venues which included the word ‘school’ in the name, even locations such as bus stops or driving schools, which are irrelevant for the purpose of this project:



b) it didn’t not display information about the schools like their ratings, if a school is private or state, primary or secondary, etc.

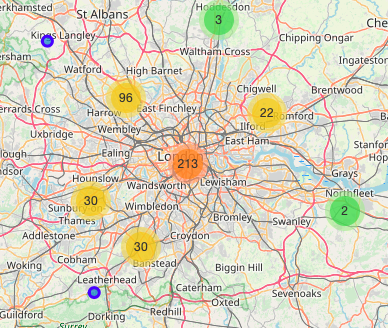
The schools data from [2] was then downloaded and cleaned:

* Irrelevant columns deleted;
* Only kept schools which satisfy the following criteria:
  + Located in Inner and Outer London;
  + State schools (Maintained schools);
  + Mixed or Girls only;
  + Have an outstanding OFSTED rating (Overall effectiveness of 1).

After cleaning the data only 398 schools left.

The data contained the postcodes of each of the schools and the *Nominatim* functionality of the *Geopy* library was run to obtain the schools’ latitudes and longitudes.

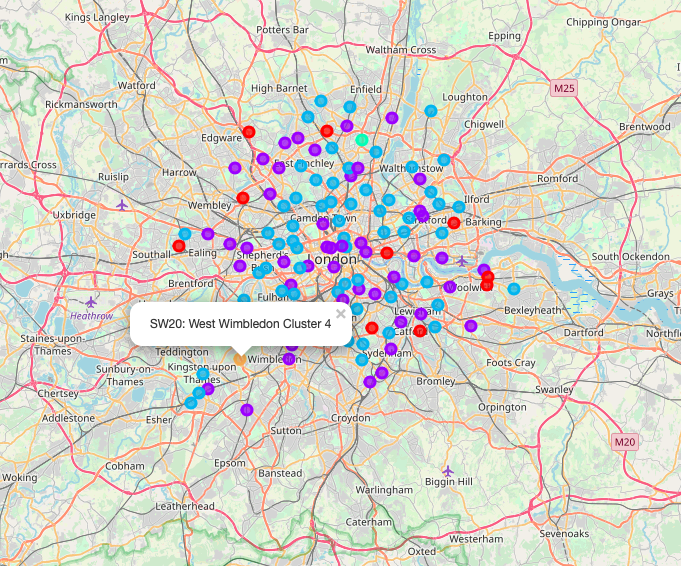
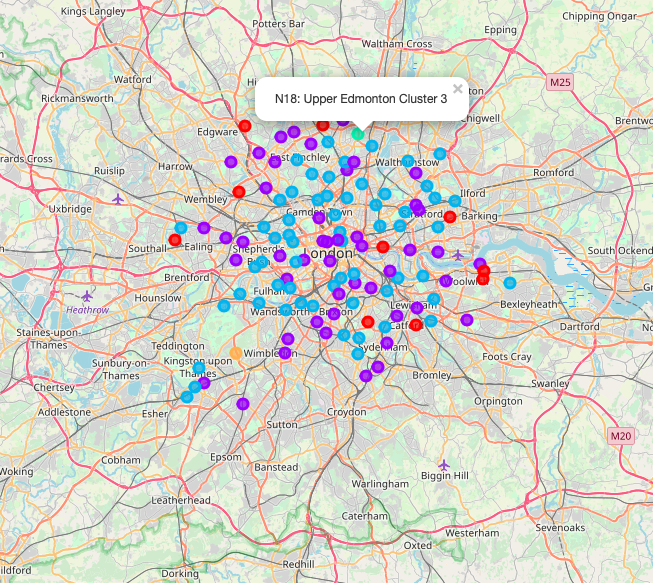
Cluster markers were added to the map of London with 1km radius circles around each of the schools:

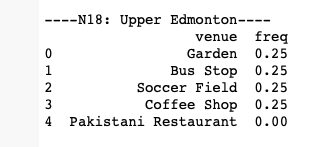


**Results / Conclusions**

*London Postal Districts*

There were two districts which were assigned a separate cluster each: West Wimbledon and Upper Edmonton:

**

On closer examination, it appeared that these two districts have rather untypical popular venues compared to the other districts: stables, golf course and gardens. Moreover, as one can see from the frequencies, there were not so many other venues in these districts.

As for the other districts, the results look sensible. Unsurprisingly, as we could see from the word cloud, the most popular venue types in London are dominated by pubs, coffee shops, cafes and grocery shops. The K-Means Cluster algorithm has split the districts accordingly, with most central locations split in 2 clusters and some more rural areas assigned a different cluster.

*London Schools*

Judging by the map we can see that most of the outstanding state schools are located in Central, South-West and North parts of London. This seems sensible: these are considered to be the more expensive areas, with higher property prices, as people tend to pay a premium to live next to good schools.

Sadly, there are fewer outstanding schools in the East and South-East.

The resulting map is a useful tool, as it enables the user to zoom in and out to do the following:

* check the locations of the outstanding state schools;
* check how close they are to the user’s current home to estimate how difficult it is to get admission into one of the schools. If you live in the overlapping circles it means there is more than one school next to you, so there are more chances to get admitted;
* check where to move, if a family has a preferred school of their choice to maximise their chances of getting admitted.

**Discussion / Next Steps**

To gain a better picture one could also do some further analysis based on his circumstances, for example;

* Narrow down the schools further (e.g. by religion denomination or by territory).
* Instead of using the ‘Overall effectiveness’ rating one could use a different metrics like, GCSEs, A-level results or admissions to universities.
* Check the demographics of each area and calculate the ratio of good schools to the number of children – the lower the ratio, the harder it is to get admitted.
* Create a Choropleth maps showing the number of schools and the property prices next to them.
* Analyse the property prices or rental yields next to the preferred schools in more detail.
* Collect as much data as possible about the preferred list of schools.

**References**

1. London postcodes: <https://www.doogal.co.uk/london_postcodes.php>
2. Data on London schools: <https://www.compare-school-performance.service.gov.uk/schools-by-type?step=default&table=schools&region=all-england&for=ofsted>