Capstone Project Report

Stephanie Lin

IBM Data Science

19 July 2020

Location recommendation

for a residential apartment development project in London

1. Introduction

1.1 Background

London, as a global real estate market, attracts many overseas investors. In the meantime, many research reports point to the fact that there is a shortage of housing in the capital. Triggered by the aftermath of COVID-19, the British government announced further policy aids to help stimulate further demand for housing nationwide. As such the ongoing housing shortage problem in the capital could intensify further, creating an investment opportunity for overseas developers.

Against this backdrop, an international real estate developer has approached us to investigate and make recommendation on plausible locations in the Greater London area for a residential development project. The extra help is needed due to their limited knowledge of neighbourhoods outside the Central London area. The scope of this project is therefore expanded beyond Central London due to concerns over affordability in the city's prime central area.

The developer's plan is to build two to four apartments for young professionals or young families with working parents who are still keen to live in the city (thus not yet moved out to suburban areas).

1.2 Strategy

Given that the target buyers of these apartments are young families, possibly with kids, we think locations with close proximity to schools is a key consideration. Areas with more choices for different schools and different types of schools (e.g. private vs. public) are likely to be favoured by buyers. And therefore should be preferred locations for this residential development project.

2. Data

There are three data sources in this project:

- All postal areas in Greater London, including the name of the postal area and its corresponding postcodes. Data source: www.worldpostalcode.com
- Geo-coordinates of postal areas in London (latitude and longitudes). Data source: Geopy library
- List of school venues in London and their categories, latitudes and longitudes. Data source: Foursquare API

2.1 Data cleaning

According to data scrapped from worldpostalcode.com, there are 720 postal areas in Greater London. Matching these postal areas against the Geopy library in order to find out the geo-coordinates of these locations, we observed some outliers (fig 1). This could either due to errors from worldpostalcode.com or Geopy library.

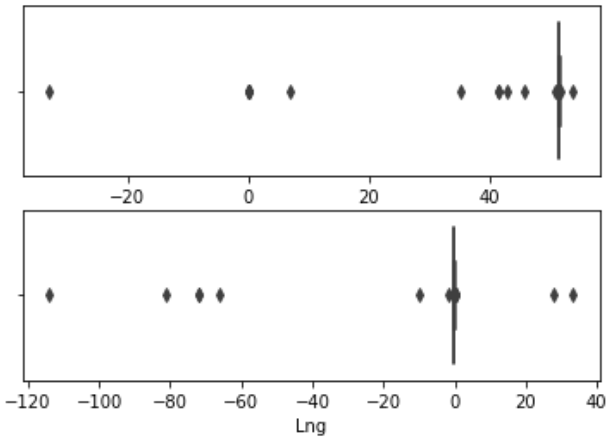

Fig1: Latitude and Longitude outliers in the Greater London geo-location dataset

Given we have a relatively large dataset (720 rows) and there are only 40 outliers, the outliers were simply removed without further investigation.

3. Methodology

3.1 Postal areas in Greater London

After the removal of outliers, all 680 locations under consideration are being plotted on the map (Fig 2).

Fig 2: Postal areas in Greater London

As part of the strategy (refer back to section 1.2 strategy), proximity to schools is a key factor to consider. I have queried all school-related venues via the Foursquare API.

Among the 680 postal areas requested, there are 99 postal areas without any schools. However given the areas locate relatively close to each other (see Fig 1), this could be fine if there are schools in close proximity. Therefore the analysis further divides all Greater London postal areas into a number of clusters to form ‘neighbourhoods’.

3.2 Clustering Greater London postal areas

I have opted for the K-means model to cluster the 680 geo-locations of Greater London postal areas. However before we run the model, it is important to understand the optimal number of clusters that would give us least residual errors without unnecessary extra computation. Therefore I have fitted the model with the number of clusters ranging from 1 to 80 (Fig 3).

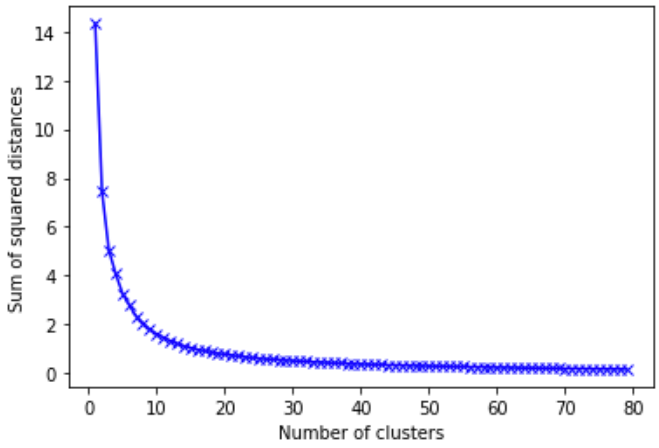

Fig 3: Number of clusters vs. K-means model sum of squared distances

30 clusters appears to be a good optimal point because any further increases in clusters require additional computational power but yields little extra reduction in error (sum of squared distances).

The K-means model with k=30 (cluster=30) is therefore being applied on the Greater London geo-location dataset to return 30 clusters. These clusters are being coloured on the map (Fig 4)

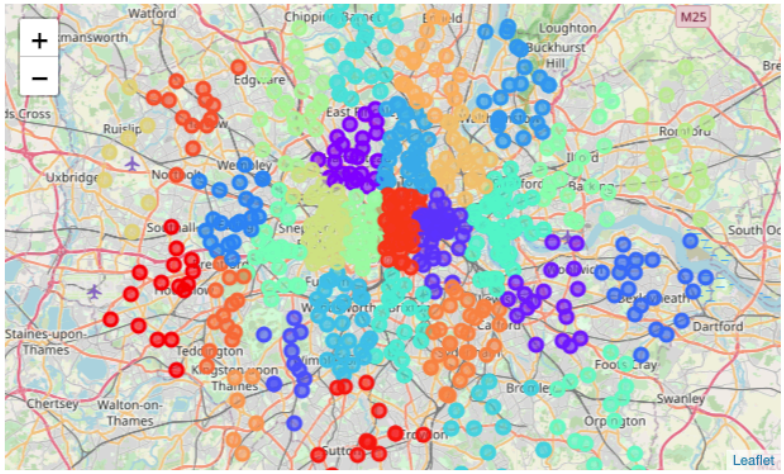

Fig 4: Clustered Greater London map

3.3 Cluster analysis

Once all postal areas have been assigned with a cluster, we can unpack the dataset with each row representing a school, and its category according to

Foursquare, as well as some additional corresponding data against this school. For example its corresponding postal area, postcodes, cluster, etc.

Upon observation, there are two types of schools that should not be included for the purpose of this exercise:

- Type 1: Irrelevant schools that are not conventional education institutions. For example, pet training centres or driving schools.
- Type 2: College or university related premises. These are removed as university enrolment are not bounded by residential addresses and many people tend to move away from home when they attend universities or colleges.

As a results, venues that satisfy the above two types are removed and I have concluded a dataset contains 1179 relevant schools (Fig 5).

After grouping this dataset by clusters we can observe that clusters 28, 20, 3, 8 and 22 have the highest number of schools (Fig 6), and therefore are preferred neighbourhoods for this project.

	PostCode	Areas	Lat	Lng	School_check	School_name	School_category	clusters
1	NW8	Abbey Road	51.458009	0.134958	1	Uplands Primary School	School	5
2	NW8	Abbey Road	51.458009	0.134958	1	Crook Log Primary School	Elementary School	5
3	SE2	Abbey Wood	51.487621	0.114050	1	Abbeywood Nursury School & Children's Centre	Nursery School	5
4	SE2	Abbey Wood	51.487621	0.114050	1	Alexander Mcleod Primary School	Middle School	5
6	SE2	Abbey Wood	51.487621	0.114050	1	Boxgrove Primary School	Nursery School	5

Fig 5: A snapshot of dataset with all relevant schools returned.

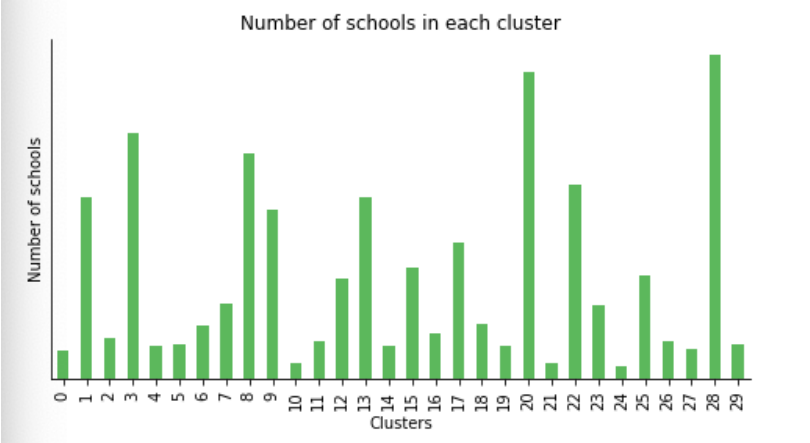

Fig 6: Number of schools in each cluster

4. Results

From the above analysis, we have concluded that from a geo-location perspective it is most optimal to cluster Greater London postal areas into 30 clusters. Although not all postal areas in London have schools within the area, it has been verified that there is at least one nearby, i.e. within the same geographic cluster.

Among the 30 clusters, I have focused on only the availabilities of relevant pre-college education institutions. As a result, schools identified in the target clusters can be plotted on the map (Fig 7):

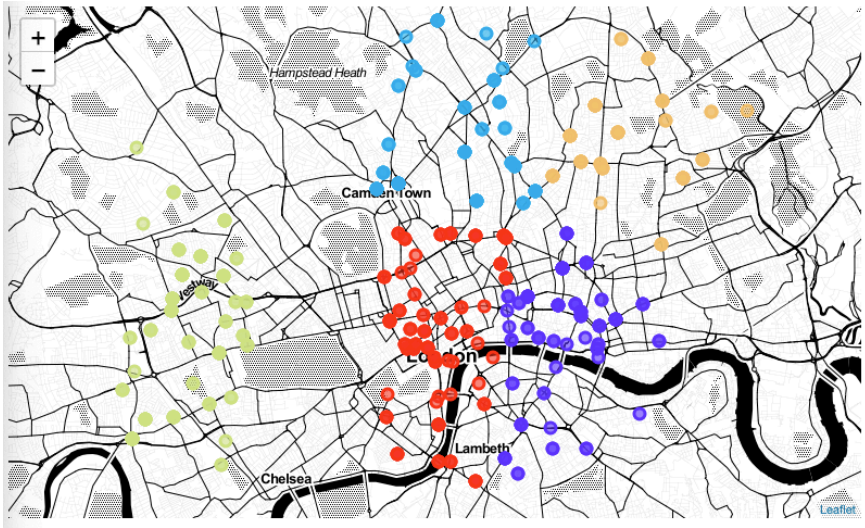

Fig 7: Schools in target neighbourhoods (clusters)

We can also identify the different types of schools in these clusters (Fig 8).

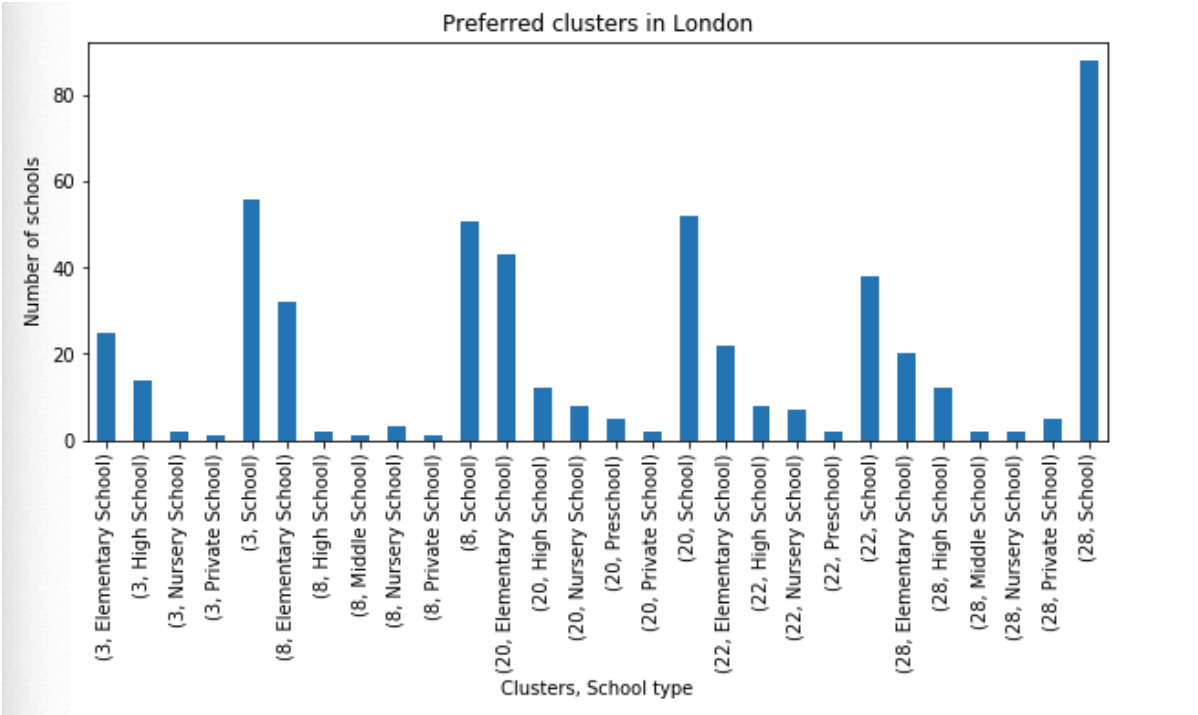

Fig 8: School types by clusters

5. Discussions

According to Fig 8, we can see that clusters 8 and 28 have greater variety, and more equally distributed types of schools compare to the other clusters. While all clusters have high number of primary schools, clusters 8 and 28 also offer a descent number choices for middle and high schools which enable families to live in the same area for longer as their kids grow older. Also this would be favoured by families with both young and older kids.

Given these neighbourhoods are more likely to be preferred by families considering medium to long term settlements, we recommend that the real estate development project to take place in areas within cluster 8 and 28.

As such this specifies the recommendation list down further to 80 postal areas in London. Fig 9 shows a snapshot of this final recommendation list.

	PostCode	Areas
0	WC2	Aldwych
1	N22	Alexandra Park
2	E14	All Saints
3	EC1	Angel
4	N19	Archway
...	...	...
75	W9	Warwick Park
76	SE1	Waterloo
77	SW1	Westminster
78	SW1A	Westminster Abbey
79	SE1	Westminster Bridge

80 rows × 2 columns

Fig 9: A snapshot of the final recommendation list.