The Battle of Neighborhoods – Relocation Challenge

Brian Lai
January 5, 2021

1 Introduction

Relocation is always a big problem and there are a lot of factors need to be considered. Key factors such as safety of the area, how far and how fast to commute to work, for family the education (i.e., surrounding schools) and last but not least the living style that preferred or used to have.

1.1 Problem

This case study is about a family which is currently living in an area called "Sutton South" in London and looking for a new place to relocate within London and would like to use data science to locate a new place with certain characteristics including good quality schools, safe and convenient to travel to central London for work, the family also want to have certain number of venues to match their lifestyle, i.e., with nightlife venues and plenty of coffee shops, and there are Japanese restaurants nearby.

Relocation is an important decision as it is not easy to move to different places to try especially for a family with kids, this is also not easy to manually search around given the city is big and each area has its own characteristics, thus using data and machine learning can potentially resolve the problem. This study will be useful for family who are looking for a place in London. The case study makes use of Foursquare API to search for venues around a location as a proxy of nearby living lifestyle as well as other data source / API which will be discussed in follow section.

2 Data

2.1 Data Sources

Data are obtained from different sources, massaged and combined to a single table for fit to the model.

School reports, crime data, population of the area are publicly available under gov.uk website. School reports are provided by Ofsted which is the Office for Standards in Education, Children's Services and Skills. Ward level crime data are provided by Metropolitan Police and Population data are provided from Office for National Statistics (ONS) Population Estimates, Borough and Ward.

Location and travel time from the ward to central London (Using Canary Wharf which is the office to travel on every weekday) is estimated by the geographic location provided by Google Geocoding API using the name of ward and the travel time provided by Google Direction API.

Lifestyle information is obtained from Foursquare API to locate nightlife venues, Japanese restaurants and coffee shops within 500 meters of the location obtained from Google Geocoding API as mentioned above.

2.2 Data Features, cleansing and special consideration

School Data

School information downloaded from Ofsted are pretty much ready to use, samples of data as show below.

	Selected	_	As at date	Deprivation	Phase	Provider type	Provision type	Remit	Sector	Constituency	Deprivation index	Does the inspection relate to the URN of the current school?	Epoch	Local authority area	OE number	Overall effectiveness	Pop filter	Postcode	Provider name	Published date	Region 9 name	URN
0	Select for download	1	31 Aug 20	Least deprived	NaN	Children's Home	Children's home	Children's social care	Local Authority	High Peak	1	NaN	After Dorset Change	Derbyshire	1	Outstanding	True	ZZ99 3CZ	REDACTED	30 Sep 2020	East Midlands	SC065553
1	Select for download	1	31 Aug 20	Least deprived	NaN	Children's Home	Children's home	Children's social care	Local Authority	Newark	1	NaN	After Dorset Change	Nottinghamshire	1	Outstanding	True	ZZ99 3CZ	REDACTED	30 Sep 2020	East Midlands	SC036737
2	Select for download	1	31 Aug 20	Least deprived	NaN	Children's Home	Children's home	Children's social care	Local Authority	Westmorland and Lonsdale	1	NaN	After Dorset Change	Cumbria	1	Outstanding	True	ZZ99 3CZ	REDACTED	30 Sep 2020	North West	SC034158
3	Select for download	1	31 Aug 20	Least deprived	NaN	Children's Home	Children's home	Children's social care	Local Authority	Lancaster and Fleetwood	1	NaN	After Dorset Change	Lancashire	1	Outstanding	True	ZZ99 3CZ	REDACTED	30 Sep 2020	North West	SC035459
4	Select for download	1	31 Aug 20	Least deprived	NaN	Children's Home	Children's home	Children's social care	Local Authority	St Helens North	1	NaN	After Dorset Change	St Helens	1	Outstanding	True	ZZ99 3CZ	REDACTED	30 Sep 2020	North West	SC035971

As the data is for the whole country, I have dropped all records outside London as those locations are not in consideration, records with status "REDACTED" are also dropped as the information may be outdated.

Features I need are 1) Local authority area, 2) Overall effectiveness, 3) Learner numbers (i.e. number of students), remaining not required columns are dropped. Overall effectiveness has values of Outstanding, Good, Requirement improvement and Inadequate, I sum the number of students in schools with “outstanding grade” grouped by the local area, to calculate the % of outstanding in each area.

Limitation of the school data are

1) there is no ward information, and
2) no official unique key provided for the area (for example Borough code).

Thus, I can only use the outstanding school % whole borough as proxy for all wards within the same borough, and borough name is used to match with other datasets.

Crime data and population data are straight forward. For population, I only need the total number of all ages of each ward, the latest year of data is used which is 2015.
On crime data, I need to sum up all crime by ward code of past 2 years (Dec 2018 to Nov 2020). Crime rate is calculated based on the total crime of all types in 2 years divided by the population.

Before merging school dataset to crime dataset, I did a check if all boroughs can be found in both datasets. I noted the borough “City of London” has no crime data, that is due to that area is not managed by Metropolitan Police. For simplicity, I decided to drop the school records of this borough as “City of London” is the heart of London and clearly not the area we are looking for.

In order to obtain a more accurate location by Google Geocoding API, I have applied a filer of postal town = London and Country = GB. Ward and borough names are added together as the query, as below code:

```
gmaps.geocode(ward + ' in ' + borough,
components={'postal_town':'London','country':'GB'})
```

Travel time is estimated from Google Direction API with starting point set as location obtained above, destination at Canary Wharf and limited to use rail (i.e. trains or subway) and set the departure time at 8am of Monday to estimate the time travel for work. As Google API has limited free quota, I saved the result as csv file for future use.

Lifestyle data from Foursquare is more straight forward, I obtained the number of Nightlife venues, Japanese Restaurants and Coffee shops using the Category ID of the

Foursquare API, within 500 meters of the location.

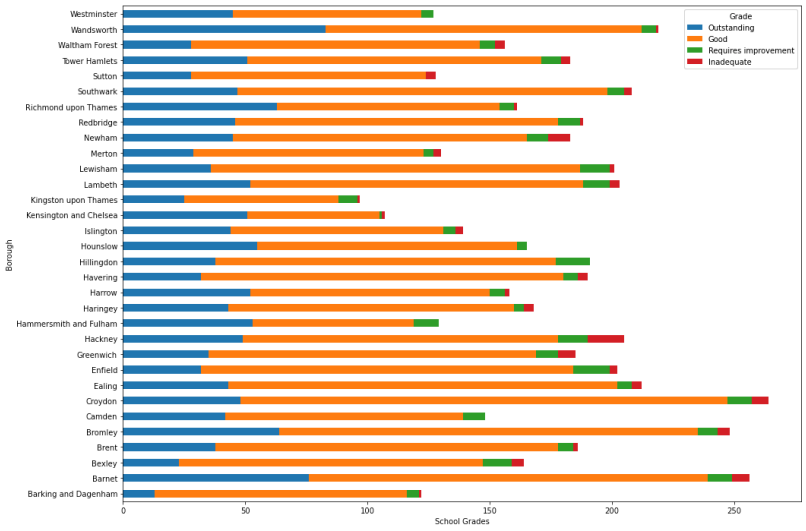
After loading all the data, the dataset is like below:

	WardCode	TotalCrime	WardName	Area	WardPopulation	CrimeRate	Population	Outstanding	Outstanding_percentage	Latitude	Longitude	TravelSecond	Nightlife	Restaurant	Coffee
0	E05000026	5691	Abbey	Barking and Dagenham	14179	40.14	58867.0	5414.0	9.2	51.543674	0.102531	2858	0	0	1
1	E05000027	1739	Alibon	Barking and Dagenham	11007	15.80	58867.0	5414.0	9.2	51.547616	0.151888	3062	1	0	0
2	E05000028	2075	Becontree	Barking and Dagenham	13687	15.16	58867.0	5414.0	9.2	51.562598	0.122430	3471	1	0	2
3	E05000029	1964	Chadwell Heath	Barking and Dagenham	10769	18.24	58867.0	5414.0	9.2	51.572228	0.141995	3500	2	0	0
4	E05000030	1623	Eastbrook	Barking and Dagenham	11255	14.42	58867.0	5414.0	9.2	51.551440	0.163567	3197	1	0	0

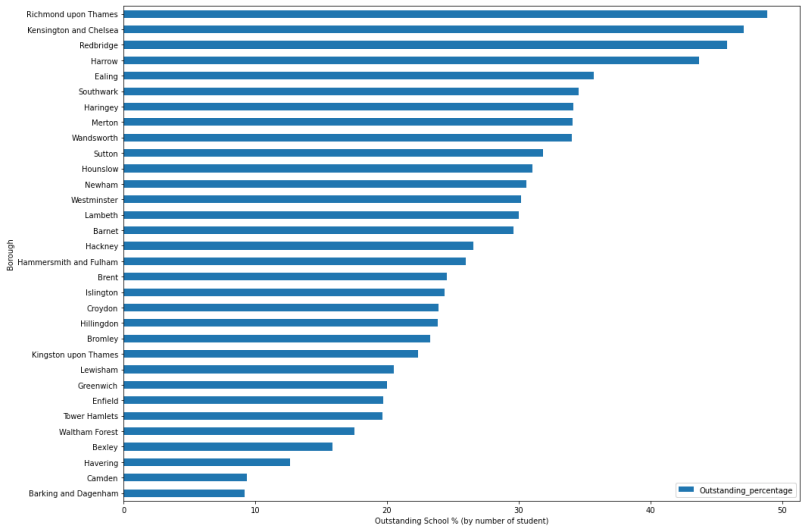
3 Exploratory Data Analysis

3.1 Outstanding Schools distribution

It is commonly understanding that location to live is an important factor for education, I plot the number of schools in each borough by different grades. In fact, it cannot tell a major different from the chart on the distribution by number.

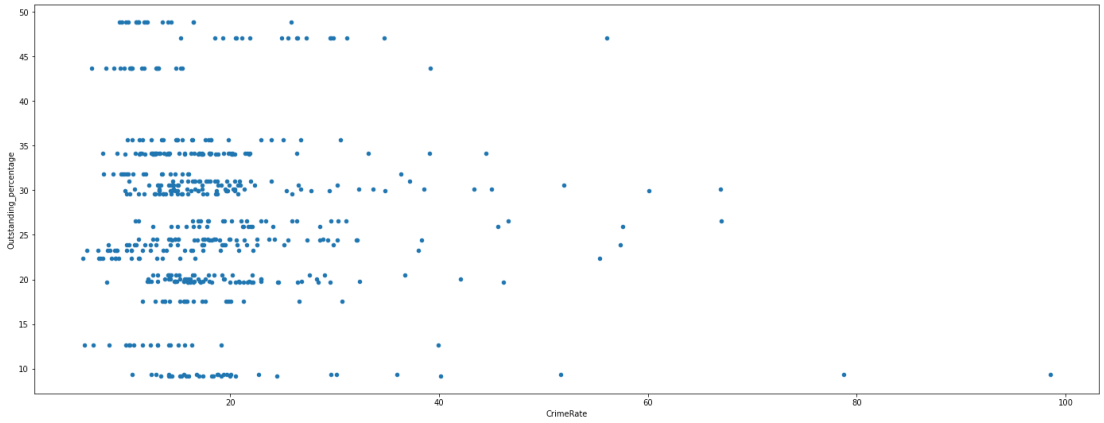


However, that shows a quite signigicant difference if the chart is using % of students in outstanding schools instead of number. Here is the comparison sorted by the % decendingly. We can see some famous names like Richmond, Kensington and Harrow are at the top of the list.



3.2 Relationship between crime rate and school grades

It is always a believe that a safe area should relate to the quality of school nearby. However, the scatter chart does not seems show this relationship. This could be due to the crime data we have is at ward level why school we can only use borough level as proxy as stated in section 2.

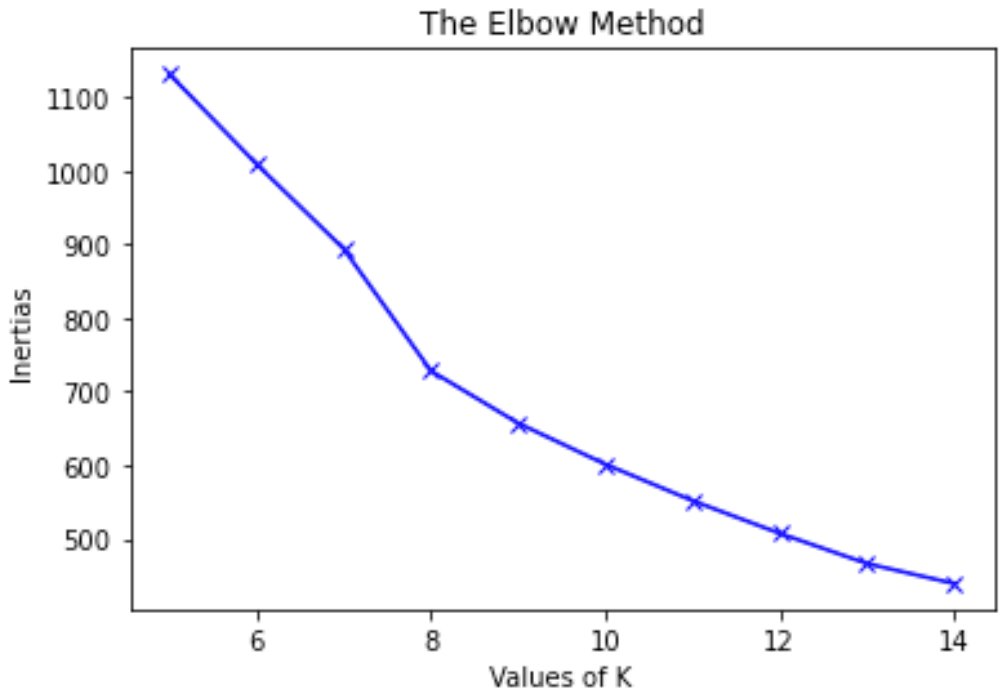


4 Methodology

I use k-Means Clustering in this case study to clustering wards to different groups and aims to find similar wards within the same group of Sutton South where the family currently live. Crime rate, Outstanding School %, Travel Time (in second), Nightlife, Restaurant and Coffee data are extracted for fitting into the model, sample data below.

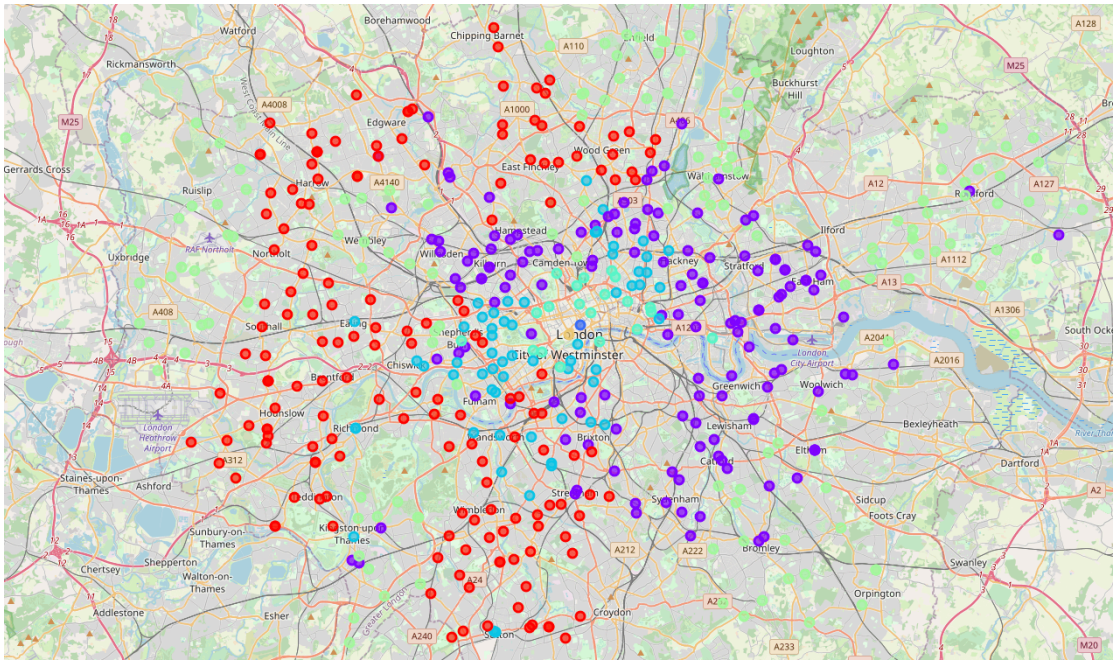
	CrimeRate	Outstanding_percentage	TravelSecond	Nightlife	Restaurant	Coffee
0	40.14	9.2	2858	0	0	1
1	15.80	9.2	3062	1	0	0
2	15.16	9.2	3471	1	0	2
3	18.24	9.2	3500	2	0	0
4	14.42	9.2	3197	1	0	0

Different number of clusters are used to find the best K using Elbow method. Data are preprocessed with StandardScaler before fit into the model. K = 8 is chosen as the result.



5 Result

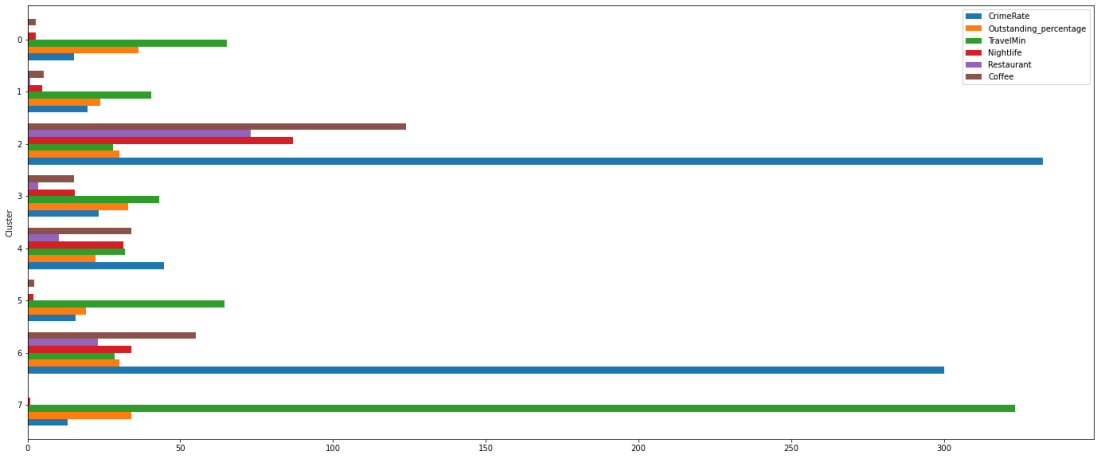
The result of clustering into 8 groups plot on London map is shown below.



We note that roughly the clustering group are spread by the distance from the central

London, with 1 cluster (Purple) at east side while 1 cluster (Red) at left side, the one with Light Green spread at the outer area of London.

I plot the distribution of the factors' mean value we fit into the model to see their relationship.



While I do not see an obvious relationship between distance (i.e. Travel Time) with the cluster, I noted there are 3 clusters with extreme values from the chart and I count of each cluster as below:

```
df_merged.groupby(['Cluster']).count()[['WardCode']]
```

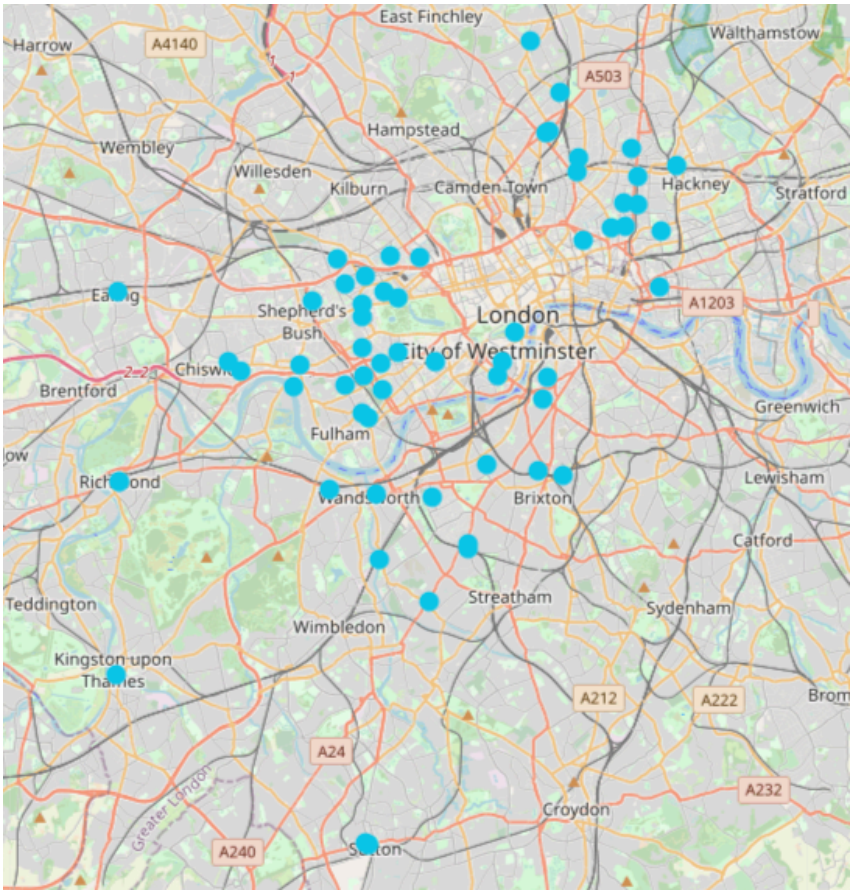
Cluster	WardCode
0	165
1	144
2	1
3	61
4	20
5	149
6	1
7	1

3 outliner clusters (#2,6,7) with only 1 ward, 2 are having extremely high Crime Rate but also high numbers of lifestyle venues decided, the remaining with extremely high Travel Time will be discuss in next section.

	WardCode	TotalCrime	WardName	Area	WardPopulation	CrimeRate	Population	Outstanding	Outstanding_percentage	Latitude	Longitude	TravelSecond	Nightlife	Restaurant	Coffee	Cluster
364	E05000472	1075	Village	Merton	8116	13.25	35598.0	12142.0	34.11	50.889660	-4.094900	19393	1	0	0	7
477	E05000644	33173	St James's	Westminster	11052	300.15	54511.0	16440.0	30.16	51.506993	-0.136060	1712	34	23	55	6
482	E05000649	40266	West End	Westminster	12114	332.39	54511.0	16440.0	30.16	51.511841	-0.127104	1674	87	73	124	2

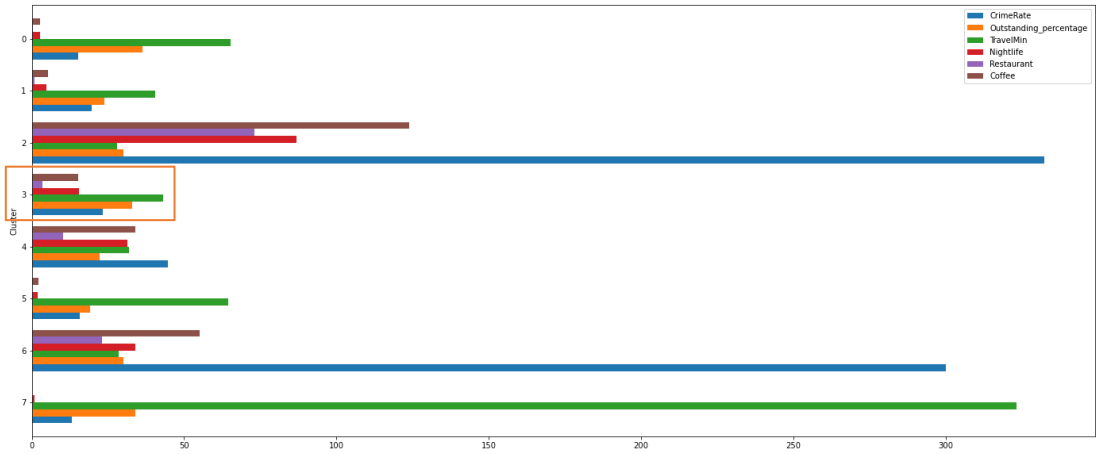
Let zoom in to the purpose of the study. We noted Sutton South is in cluster 3 and if

we plot all wards in this cluster on the map below, we noted majority of wards are near the center while a few of them located at the high street of popular areas (Richmood, Earling, Kingston upon Thames…etc) at the outer part of London.



If we look into the details of this cluster, it is not difficult to see these wards are having a relatively low crime rate, good outstanding school and travel time and some but not a lot lifestyle venues required, which is quite similar to the location the family currently lives.

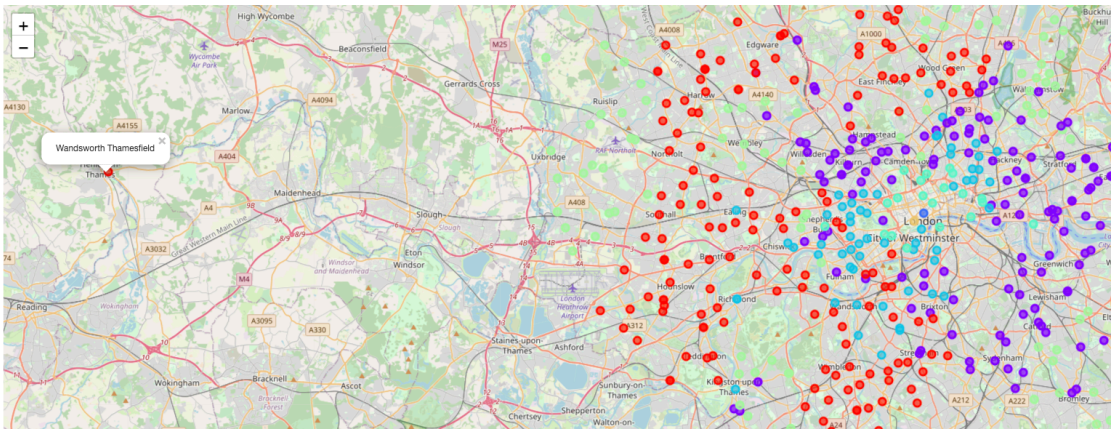
	TotalCrime	WardPopulation	CrimeRate	Population	Outstanding	Outstanding_percentage	Latitude	Longitude	TravelSecond	Nightlife	Restaurant	Coffee	Cluster
count	61.000000	61.000000	61.000000	61.000000	61.000000	61.000000	61.000000	61.000000	61.000000	61.000000	61.000000	61.000000	61.0
mean	2895.278689	12399.131148	23.415410	46933.508197	14597.245902	32.918197	51.494374	-0.164695	2593.196721	15.540984	3.459016	15.327869	3.0
std	1361.904269	2741.010455	9.649576	14508.016882	3013.265486	8.328817	0.047011	0.062649	570.078206	5.858253	1.708545	5.322034	0.0
min	840.000000	5736.000000	9.040000	24788.000000	8598.000000	19.660000	51.361428	-0.304545	1342.000000	6.000000	0.000000	7.000000	3.0
25%	2089.000000	10612.000000	17.260000	41826.000000	11735.000000	26.570000	51.465881	-0.196295	2240.000000	11.000000	2.000000	12.000000	3.0
50%	2639.000000	11919.000000	21.360000	46536.480000	14740.000000	30.160000	51.494197	-0.180248	2536.000000	15.000000	4.000000	15.000000	3.0
75%	3278.000000	14229.000000	27.270000	54511.000000	16440.000000	34.150000	51.523484	-0.114493	2934.000000	19.000000	5.000000	18.000000	3.0
max	8835.000000	20255.000000	67.060000	95269.000000	22216.000000	48.880000	51.581998	-0.057387	4040.000000	28.000000	8.000000	30.000000	3.0



6 Discussion

Based on the numbers we see in the same cluster, the K-mean clustering is doing a good job to help grouping similar wards into same cluster. However, several improvement opportunities are noted on this case study.

First, regarding the outliner of cluster #7 with extremely high travel time, it is due to the wrong location return by the Google Geocoding API, despite I have narrowed the scope of search in London, UK, this location is clearly outside of London area. This issue could probably be resolved by looking another data source which provide the location of different Wards directly instead of lookup by name.



Second, currently we only use the school for the whole borough as proxy for all Wards inside the borough. However, in reality the actual school assigned is usually depending on the distance from the school to the resident address, i.e. some schools within the same borough may not available or some schools in nearby borough could be available to apply. This require a much more accurate location data or even calculating the distance between school and home, which will be costly and timely.

7 Conclusion

In this study, I use K-mean cluster model to group different living areas (Wards) into different clusters and aims to look for a similar cluster that currently living at. Factors to use are depending on individual preference / requirement on their living place and can be fine tune for other person by getting data of other factors by other means or sources. In this study, I used crime rate with population, school rating & number of students, travel time estimated by Google Direction API, and nearby venues of preference located by Foursquare API. The K-mean cluster model is quite simple to use and easy to understand but at the same time produced a relatively good outcome, beside finding a similar place the result can also easy to be used as a high level guide on different characteristics of different areas for recommendation based on individual's requirement.

8 Reference

School Data – Ofsted
(https://reports.ofsted.gov.uk/)
Crime Data – Metropolitan Police
(https://data.london.gov.uk/dataset/recorded_crime_summary)
Population – Office for National Statistic (ONS)
(https://data.london.gov.uk/dataset/office-national-statistics-ons-population-estimates-borough)