The Battle of the Neighbourhoods Capstone Project

Identifying the best ward to purchase a property in within Greater London

Author: Charlotte Fettes
Date: 17/08/2019

Contents

1 Introduction

1.1 The Background

Despite property market fluctuations, people continue to buy and sell property throughout Greater London. According to data published by HM Land Registry (2018), total property sales for the year ending March 2018 was approximately 94,000. And this is considered low; in 2006, property sales volume was approximately 172,000.

People spend a lot of time searching for their ideal property. Rightmove, a property website, reported households in the UK spend an average of almost 1 billion minutes a month searching for properties on their website.

1.2 The Problem

London is a diverse city, with different locations being highly unique. For someone looking to buy a property in London, it is very difficult to know which area would be most suited to their personal needs and preferences. Although property price is a major deciding factor, there are numerous other variables that help to determine where someone decides to concentrate their property search. These variables include, for example, property types, the quality of primary and secondary education if they have children, public transport accessibility, crime levels, and the general characteristics of the area. Furthermore, every property buyer is different; no two buyers will be the same, and so an area that suits one may not be suited to another. Trying to find an appropriate area within London can take days, weeks or even months of research. And this is before the search for an actual property can even begin.

1.3 The Idea

Utilising Government data disaggregated by London wards[1], and Foursquare location data to assist in characterising those wards, we can use Data Science techniques to build a clearer picture of what defines these various locations throughout London. We can then apply this information to any property buyer and their unique explicitly stated requirements to generate a shortlist of locations in London most suited to that particular buyer.

1.4 The Target Audience

Multi-dimensional characterisation of London wards will be particularly useful for anyone looking to purchase property within Greater London, especially those that are not familiar with the different locations within London. By being able to match their personal needs and preferences to London locations, and selecting a property from these locations, they will have a greater sense of confidence in their final purchase. Furthermore, being able to narrow down

[1] Wards are electoral districts that make up Boroughs (local authority districts that make up Greater London). Boroughs are quite large areas, so going down to ward level reduces variability, but retains the size necessary for a property buyer to have a realistic chance of finding a property for sale in the shortlisted locations. To further reduce variability, we could take the postcode level, but then there will be significantly fewer properties for sale within each location.

location options in this way will greatly reduce the time and effort spent on researching locations before reaching a decision.

This type of characterisation will also assist estate agents in providing a highly personalised service. With this tool, estate agents are better equipped to find suitable options for their clients, increase their efficiency, and help them secure repeat business.

People looking to rent property may also be interested in this project. Although the property price category would not be relevant, all other categories may be of interest for potential tenants to determine where in London would be most suitable for them to look for a property to rent based on their needs and preferences.

1.5 The Case Study

Once London wards have been characterised, a case study will be applied in order to demonstrate how this project will provide house buyers with a shortlist of recommended London locations to centre their property search based on their personal criteria.

Our case study is a family of 5 – husband and wife, and their 3 children aged 12, 9, and 5 – moving into London from a more rural location due to the wife accepting a job position in central London, the husband wanting to build a personal training business and access more clients, and wanting to provide their children with more to do.

The family have little knowledge of areas within London. With their busy schedules and preference to move as soon as possible, they have limited time to conduct research to find the ideal location within London to buy in. Ideally, a location would fulfil the following requirements:

- The property price to be around their budget of £800,000.
- The husband is a personal trainer looking to build up and expand his business, and so access to gyms and outdoor areas is important.
- The wife will be working in central London, so good access to the public transport network would be beneficial.
- With one child now in secondary school, and GCSE's on the horizon, the quality of secondary education is particularly important. With another two children in primary school, the quality of primary education is also important.
- They are looking for a house – ideally semi-detached or detached, with a terraced house being an option in the right area. They do not want a flat.
- With children, safety is always a concern. They want to be sure that they are in a safe area with low crime levels.
- An area with parks and open space would be great, but not so much that facilities and amenities of London are just as inaccessible as in their current location. One of the reasons they want to be in London so there is more to do around them at a closer distance.
- They would like a sociable area with cafes, food stores, and things for the children to do (e.g. libraries, swimming pools/leisure centres), especially outdoors activities as that is what they are used to coming from a rural location. Not too busy in terms of nightlife but some options available for dining out. And not a tourist area.

2. Data

2.1 Data Sources

In this project, data was used from three sources: London DATASTORE, Foursquare API, and LocationIQ.

2.1.1 London DATASTORE datasets

London DATASTORE is a free and open data-sharing portal to access data related to London. The London DATASTORE has numerous data resources. From this source, the following datasets were utilised:

1. Ward profiles and atlas dataset: this dataset presents 63 key summary measures on London wards and the populations within them. The profiles were created using the most up-to-date information available at the time. The most recent version was published in 2015. Some measures may be outdated, but should still provide an adequate overview of the wards. For example, the GCSE data available is related to the old system, however more recent data is only available at the Borough level and given the relative importance families often place on the quality of secondary education, it is important to get a picture of performance at the ward level.

	Ward name	Old code	New code	Population - 2015	Children aged 0-15 - 2015	Working-age (16-64) - 2015	Older people aged 65+ - 2015	% All Children aged 0-15 - 2015	% All Working-age (16-64) - 2015	% All Older people aged 65+ - 2015	Mean Age - 2013	Median Age - 2013	Area - Square Kilometres	Population density (persons per sq km) - 2013	% BAME - 2011	% Not Born in UK - 2011
0	City of London	00AA	E09000001	8100	650	6250	1250	7.96206	76.8868	15.1512	41.3039	39	3.15148	2538.06	21.4	36.7
1	Barking and Dagenham - Abbey	00ABFX	E05000026	14750	3850	10150	750	25.9583	69.0142	5.02748	29.5	29	1.3	10500	71.9	57.3
2	Barking and Dagenham - Alibon	00ABFY	E05000027	10600	2700	6800	1050	25.7013	64.2769	10.0217	33.8	33	1.4	7428.6	29.9	24.7

Fig 2.1 Head of ward profiles and atlas dataset

2. Key stage 1 and Key stage 2 results by borough for 2018: these datasets provide percentage measures of the children achieving the expected levels in reading, writing, maths and science at the Borough level. This is considered an adequate proxy for primary education level quality within Boroughs. There was no data available for primary achievement at the ward level.

	Unnamed: 0	Unnamed: 1	All	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unna
0	NaN	NaN	Number of eligible pupils	NaN	NaN	NaN	Percentage reaching the expected standard	NaN	NaN	NaN	Percentage reaching the higher standard	NaN	
1		NaN	English Reading	English Writing	Mathematics	All (reading,writing and maths)	English Reading	English Writing	Mathematics	All (reading,writing and maths)	English Reading	English Writing	Mather
2	Code	Area	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	E09000001	City of London	29	29	29	29	76	86	90	72	45	62	
4	E09000002	Barking and Dagenham	3411	3411	3411	3410	75	83	80	67	27	43	

Fig 2.2 Head of key stage 2 results by Borough dataset

3. Land registry average house prices by ward dataset: this provided average (mean) house prices at the ward level for years up to and including 2017.

	New code	Ward name	Borough name	Year ending Dec 1995	Year ending Mar 1996	Year ending Jun 1996	Year ending Sep 1996	Year ending Dec 1996	Year ending Mar 1997	Year ending Jun 1997	Year ending Sep 1997	Year ending Dec 1997	Year ending Mar 1998	Year ending Jun 1998	Year ending Sep 1998	Yea endin De 199
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nai
1	E09000001	City of London	City of London	-	-	-	-	-	-	-	-	-	-	-	-	
2	E05000026	Abbey	Barking and Dagenham	51077.6	49868.8	49901.6	51935.1	50766.5	50189.5	47807.5	48004.6	51318.4	53046	56650.8	62273.5	62804.
3	E05000027	Alibon	Barking and Dagenham	45490.4	44701.5	44486	45894.1	46145.1	47019.2	48608.8	50142.5	51232.4	51968.8	52586.1	54027.2	56402.

Fig 2.3 Head of land registry mean house prices by ward dataset

2.1.2 Geocoder and LocationIQ

To obtain latitude and longitude coordinates of all London wards, the provider LocationIQ was used alongside geocoder. LocationIQ provides geocoding based on OpenStreetMap's Nominatim.

	Ward loc	Latitude	Longitude
0	City of London, City of London	51.515618	-0.091998
1	Abbey, Barking and Dagenham	51.535688	0.075530
2	Alibon, Barking and Dagenham	51.547549	0.153114
3	Becontree, Barking and Dagenham	51.549265	0.127538
4	Chadwell Heath, Barking and Dagenham	51.567904	0.128041

Fig 2.4 Head of ward and coordinates dataframe generated by geocoder and LocationIQ

2.1.3 Foursquare

Using the latitude and longitude coordinates, Foursquare located the wards and returned Foursquare location venue data for all wards.

The explore endpoint was used to ensure that the Foursquare API explored and returned popular spots around each location: by learning about the popular spots we can get a sense of the character of each area – what there is to do, the facilities available and what the current residents and visitors enjoy in each area.

With the average area of London wards being 2.6km, the radius for the search within each ward was set to 1300m, and venues limited to 100 to ensure a sufficient number of venues for each ward that would assist in drawing conclusions on the character of the area. Foursquare returned 429 unique venue categories for the 571 London wards.

	Ward loc	Ward Latitude	Ward Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	City of London, City of London	51.515618	-0.091998	Goodman Steak House Restaurant	51.514398	-0.090745	Steakhouse
1	City of London, City of London	51.515618	-0.091998	The Ned Hotel	51.513565	-0.090166	Hotel
2	City of London, City of London	51.515618	-0.091998	The Merchant House	51.513264	-0.093039	Cocktail Bar
3	City of London, City of London	51.515618	-0.091998	Daunt Books	51.513982	-0.092995	Bookstore
4	City of London, City of London	51.515618	-0.091998	Hawksmoor Guildhall	51.515647	-0.090997	Steakhouse

Fig 2.5 Head of dataframe containing ward venue data generated by Foursquare API

2.2 Data cleaning and feature selection

2.2.1 Government datasets

On examining the individual datasets, there was a lot of data that was not relevant to the problem at hand.

The ward profiles and atlas dataset contained many measures relating to the health of the existing population and other such measures, which would not be particularly relevant to a family looking to move into the area. The focus when selecting measures to retain was on those variables that property buyers find important when looking for a property. The measures retained for the purposes of this project were:

- New ward code (for initial merging with other datasets);
- Crime rate - 2014/15: Index scores of overall notifiable offences per 1,000 daytime population. Daytime population used as denominator because areas with high crime rates tend to be areas with high daytime populations (ie non-resident) and therefore don't directly relate to well-being of the resident population. Consequently, a lower proportion of crimes in town centres would be against a resident. Source: MPS, Home Office, and ONS Workday population 2011 Census (modelled for other years).
- Average GCSE capped point scores – 2014: the top 8 GCSE grades of each student are converted into a point score (A*=8, A=7, etc.) and added together to give a student's point score
- % area that is open space – 2014;
- Average Public Transport Accessibility score – 2014: - generated by TfL and GLA, they are an average Public Transport Accessibility Level (PTAL) score. PTALs are a measure of the accessibility of a point to the public transport network, taking into account walk access time and service availability. Population weighted average scores were calculated using output area data. There are 9 levels of access, 0 to 9. Each area was given an average score out of 8, where 8 is the highest level of accessibility. Open space was removed from the data as no population lives there.
- % detached houses – 2011; % semi-detached houses – 2011; % terraced houses – 2011; % Flat, maisonette or apartment – 2011.

The Key stage 1 and Key stage 2 datasets contained multiple years. The data for the most recent year (2018) was selected, and the average percentage across all subjects (reading, writing, maths and science) was retained. The overall average of Key stage 1 and Key stage 2 combined was calculated as to provide a single overall measure of the quality of primary education.

The house price dataset contained data from 1995. Given the fluctuating nature of property prices in London, and the need to know prices as close to current reality as possible to make an informed decision, only the most recent data was retained – year ending 2017.

The Ward name format on ward profile dataset was not conducive to a latitude-longitude coordinate search, and so the ward name combined with Borough name from the house prices dataset was used instead.

Broader level summary data, such as UK level Key stage 1 data, was removed from all datasets to ensure they did not skew the data and results.

2.2.2 Geocoder and Location IQ

Once the datasets were merged, the wards were converted into a list and run through the geocoder to obtain latitude and longitude coordinates.

Running the geocoder revealed some of the ward names as being ambiguous to the provider and so producing incorrect coordinates. To combat this issue, the problematic ward names were changed to a location central to the ward e.g. Queen's Park, Westminster (the geocoder was returning a similar named location in New Zealand) was changed to Queen's Park Library. Some continued to be returned as incorrect, and so were corrected

The resulting coordinates were added to the ward dataframe. The resulting dataframe contained a total of 571 wards.

3. Methodology

3.1 Exploratory Analysis

K-Means clustering was conducted in order to group wards according to similarities within each feature – price, property type, primary and secondary education, crime rate, public transport accessibility, open space, and venues. Clustering could have been carried out manually for some features, e.g. price could have been separated into equal price ranges; however, the use of K-Means helped ensure that there was minimal variability within clusters and maximal variability between clusters. Each cluster was then summarised in terms of what the wards within each cluster represented.

Descriptive statistics were also obtained for each feature overall. This was to promote understanding of the feature in question. For example, descriptive statistics for secondary education revealed the mean ward average GCSE point score, which was then used to inform what clusters could be considered to have an acceptable GCSE point score range, and which suggested poor secondary education performance.

3.1.1 Property Price

K-Means clustering grouped wards according to the average property price for the year ending December 2017. In order to make this project potentially applicable to a broad range of property buyers, and given the large average price range of properties within London wards (£259,486.62 - £4,416,659.40), wards were grouped into 8 clusters. Figure 3.1 shows how these clusters map across London.

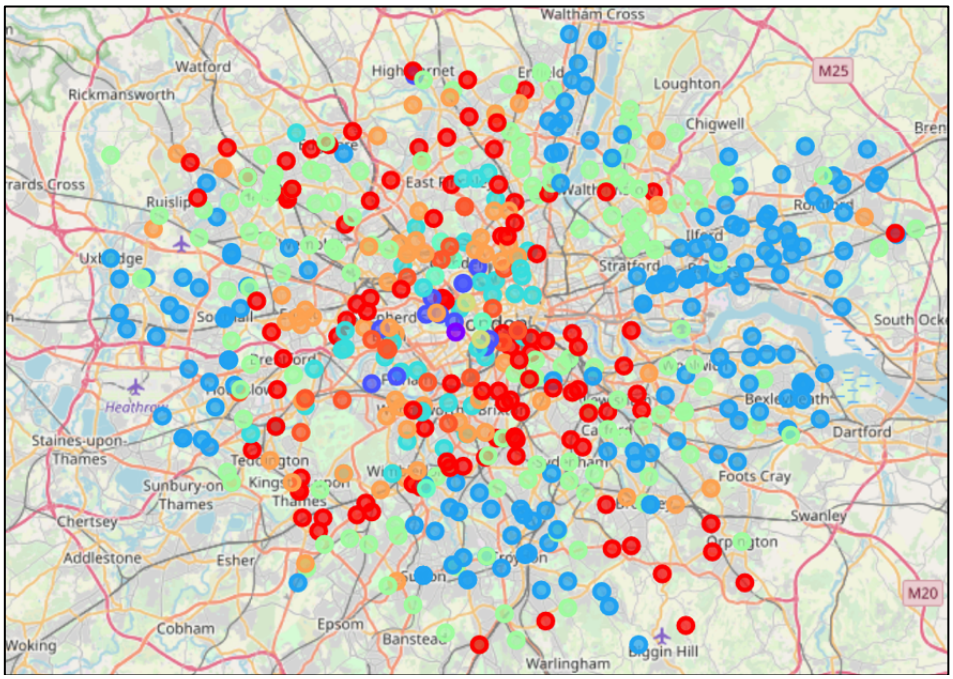

Fig 3.1 London wards clustered according to average property price 2017

The cluster mean price ranges were as follows (colours are for corresponding map clusters):

- Cluster0 (red): £524,100.61 - £648,806.41
- Cluster1 (purple): £4,416,659.40
- Cluster2 (dark blue): £1,296,800.86 - £1,585,562.76
- Cluster3 (light blue): £259,486.62 - £401,320.31
- Cluster4 (turquoise): £823,462.88 - £1,008,300.62
- Cluster5 (green): £1,891,715.94 - £2,186,666.67
- Cluster6 (lime green): £403,917.61 - £518,567.23
- Cluster7 (yellow): £2,856,435.71 - £3,033,756.54
- Cluster8 (light orange): £649,877.67 - £817,646.99
- Cluster9 (dark orange): £1,021,486.54 - 1,238,295.89

3.1.2 Property Type

K-means clustering grouped wards into clusters according to the prevalence of each property type within the wards – detached houses, semi-detached houses, terraced houses, and flats, apartments and maisonettes. An elbow curve was devised to deduce the optimal number of clusters - 5 (see fig 3.2). Figure 3.3 shows how these clusters map across London.

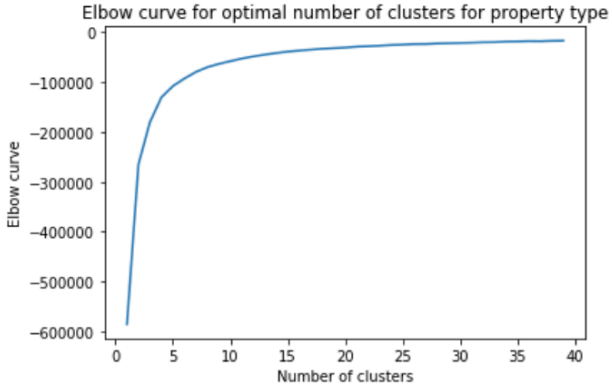

Fig 3.2 Elbow curve to show optimal number of clusters for property type

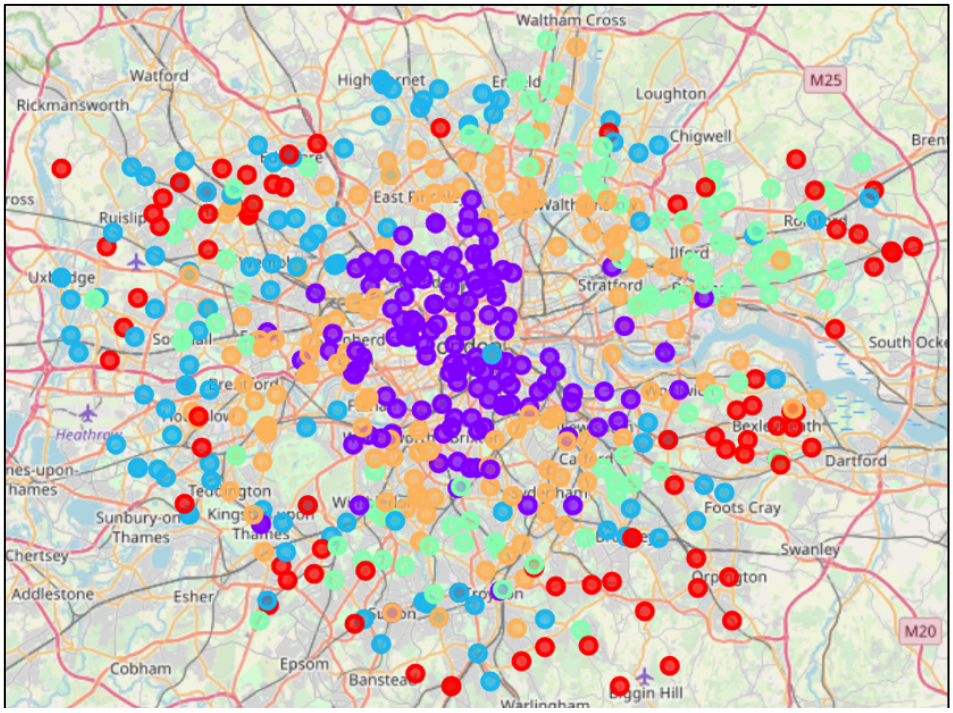

Fig 3.3 London wards clustered according to prevalence of property type

In order to better understand these clusters, Fig 3.4 below shows a bar chart of the means of the property type percentages per cluster.

As certain property types will always be more prevalent in London (e.g. flats) than others (e.g. detached), when summarising the clusters each property type was given a number 1-5 such that for example detached (1) indicates that this cluster has the highest percentage of detached houses out of all the clusters. The clusters were as follows (colours are for corresponding map clusters):

- Cluster0 (red): detached (1), semi-detached (1), terraced (4), flats(5)
- Cluster1 (purple): detached (5), semi-detached (5), terraced (5), flats(1)
- Cluster2 (blue): detached (2), semi-detached (2), terraced (3), flats(3)
- Cluster3 (green): detached (3), semi-detached (3), terraced (1), flats(4)
- Cluster4 (orange): detached (4), semi-detached (4), terraced (2), flats(2)

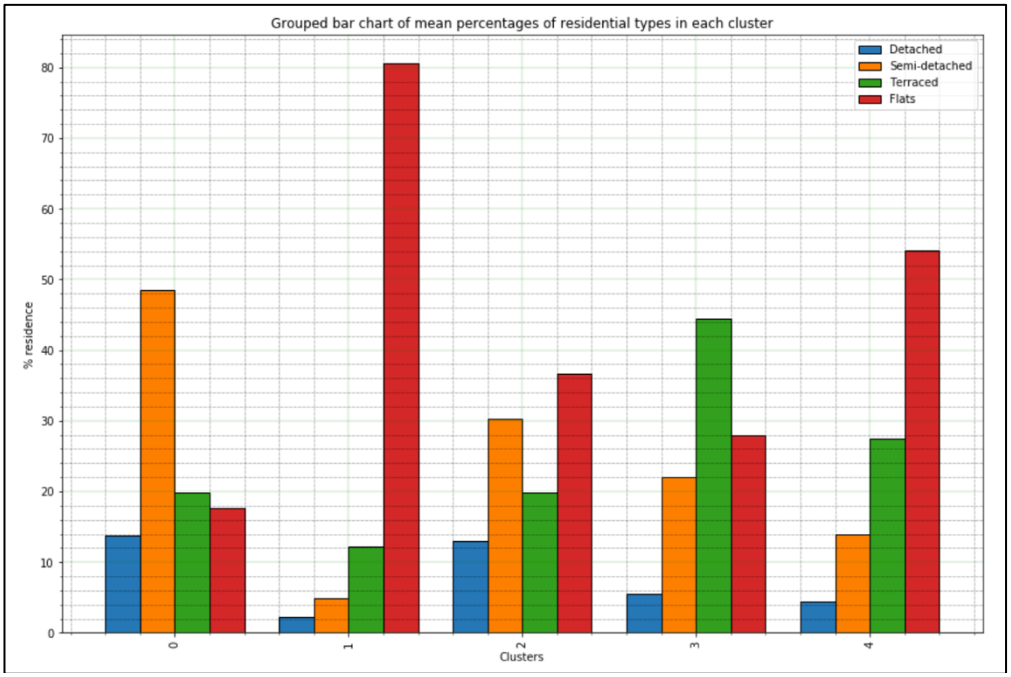

Fig 3.4 grouped bar chart of mean percentages of residential types in each cluster

3.1.3 Education

Primary and secondary education were kept separate rather than combined into a single education measure for a number of reasons: (1) for property buyers who have children only in one or the other, it would be useful to only consult the feature that is of relevance to them; (2) primary education was measured at the Borough level, whereas secondary was at the ward level; (3) primary measured the percentage of children as a whole who reached the expected level or above, whereas secondary education measured the average point score per student (actual attainment). Summary descriptive statistics for both primary and secondary are shown in figure 3.5.

	Average GCSE capped point scores - 2014	% children achieving expected level at primary	Latitude	Longitude
count	571.0	571.0	571.0	571.0
mean	327.0	75.0	52.0	-0.0
std	22.0	3.0	0.0	0.0
min	276.0	69.0	51.0	-0.0
25%	311.0	72.0	51.0	-0.0
50%	324.0	75.0	52.0	-0.0
75%	341.0	76.0	52.0	-0.0
max	396.0	83.0	52.0	0.0

Fig 3.5 Summary statistics for primary and secondary performance for all London wards

3.1.3.1 Primary Education

In order to understand what could be reasonably considered a good and poor primary performance, summary statistics were used to generate a box plot to visualise the distribution

of attainment (see fig 3.6). The mean percentage of children achieving the expected level at primary level (75%) was taken as an acceptable level from which to distinguish between good, average and poor performance of primary schools.

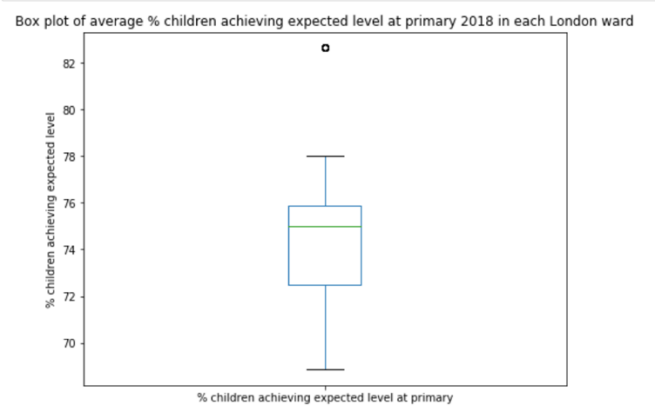

Fig 3.6 Box plot of average % children per ward achieving expected level at primary 2018

K-Means clustering was used to cluster wards into 5 clusters – determined by devising an elbow curve (see fig 3.7) – according to primary school performance (see fig 3.8).

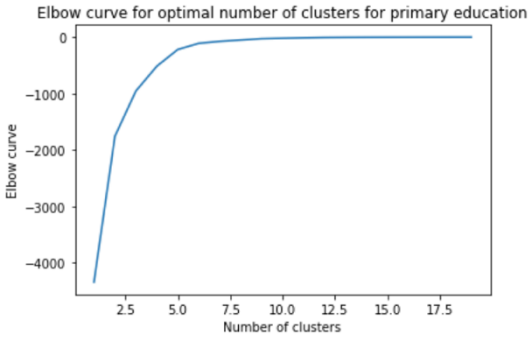

Fig 3.7 Elbow curve to show optimal number of clusters for primary education

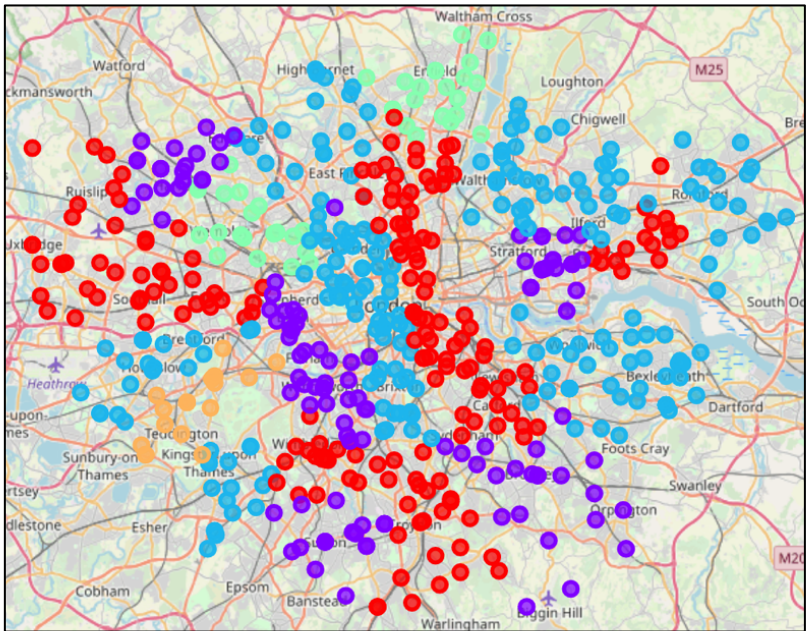

Fig 3.8 London wards clustered according to primary education performance

Comparison box plots were generated to show how each of the 5 clusters compared to on another (see fig 3.9). The summary of primary education clusters are as follows (colours are for corresponding map clusters):

- Cluster0 (red): 71.38-73.62% - below average
- Cluster1 (purple): 76.38-78% - above average
- Cluster2 (blue): 74.38-75.88% - low average
- Cluster3 (green): 68.88-69.12% - lowest level
- Cluster4 (orange): 82.62% - highest level

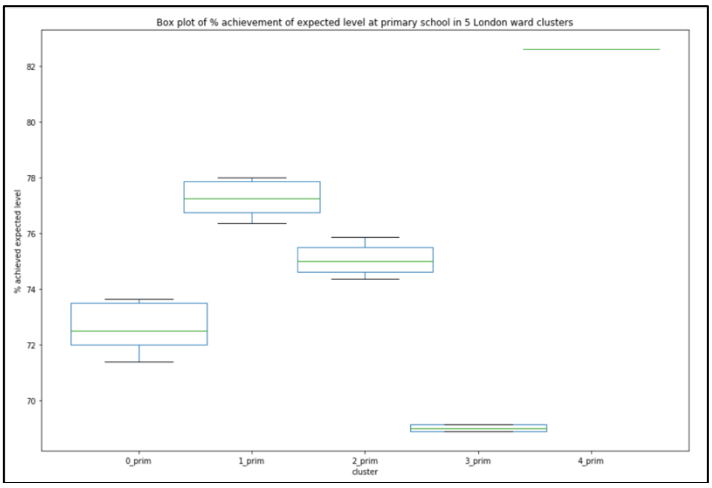

Fig 3.9 Box plot comparison of percentage of children achieving expected level at primary school in 5 London ward clusters

3.1.3.2 Secondary Education

As for primary, in order to understand what could be reasonably considered a good and poor secondary performance, summary statistics were used to generate a box plot to visualise the distribution of attainment (see fig 3.10). The average GCSE point score per student for all of London (327) was taken as an acceptable level from which to distinguish between good, average and poor performance of secondary schools.

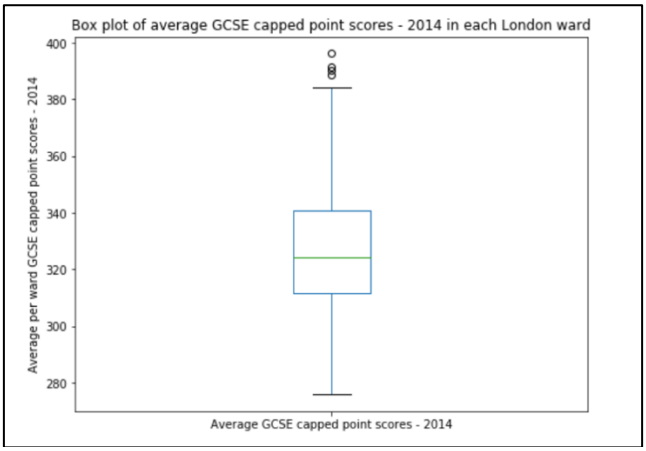

Fig 3.10 Box plot of average GCSE point scores per student for London wards

K-Means clustering was used to cluster wards into 5 – determined by devising an elbow curve (see fig 3.11) – clusters according to GCSE point score attainment (see fig 3.12).

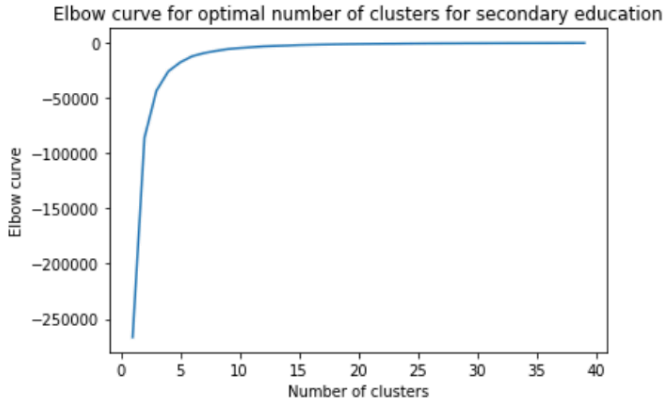

Fig 3.11 Elbow curve for the optimal number of clusters for secondary education

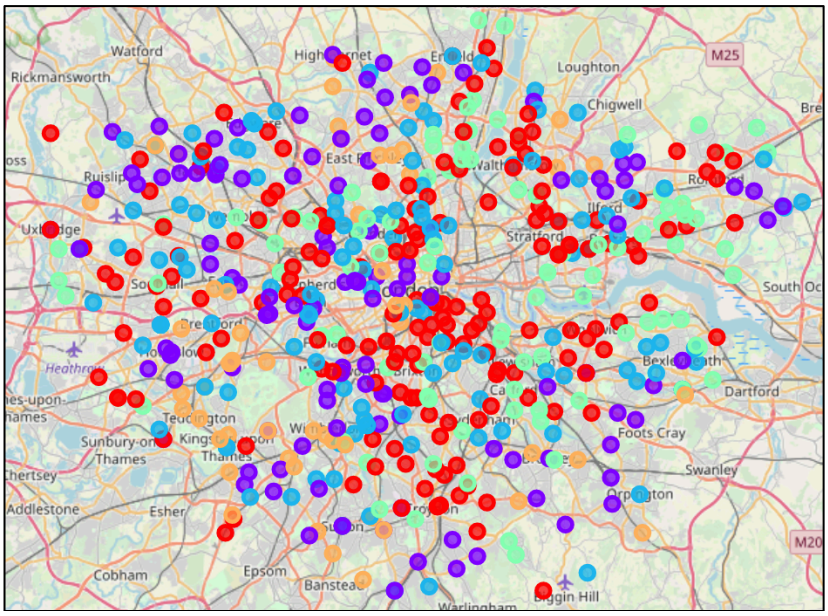

Fig 3.12 London wards clustered according to GCSE point score per student

Comparison box plots were generated to show how each of the 5 clusters compared to one another (see fig 3.13). The summary of secondary education clusters are as follows (colours are for corresponding map clusters):

- Cluster0 (red): 306.94 - 322.14 – below average
- Cluster1 (purple): 338.09 - 357.23 – above average
- Cluster2 (blue): 322.44 - 337.78 – low to average
- Cluster3 (green): 275.95 - 306.49 – lowest level
- Cluster4 (orange): 358.47 - 396.21 – highest level

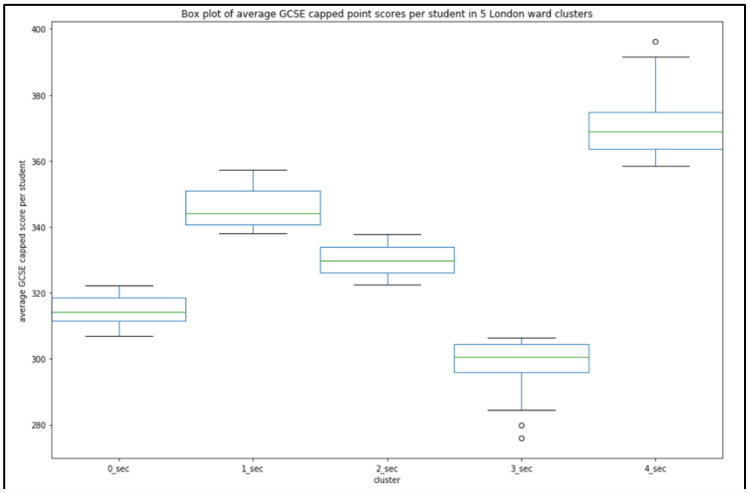

Fig 3.13 Box plot comparison of GCSE average point score per student in 5 London ward clusters

3.1.4 Crime

In order to understand what could be reasonably considered a low, average and high crime rate, summary statistics were used to generate a box plot to visualise the distribution of crime rate (see fig 3.14). The mean crime rate (80.87) was taken as an acceptable level from which to distinguish between low (less than 80.87), average and high (greater than 80.87) crime rate.

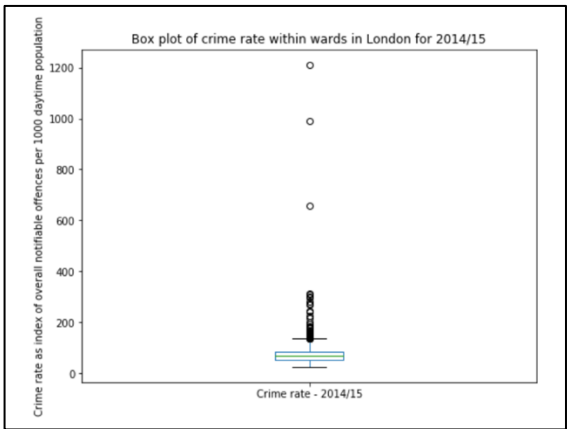

Fig 3.14 Box plot of crime rates within London wards for 2014/15

K-Means clustering was used to cluster wards into 5 clusters – determined by devising an elbow curve (see fig 3.15) – according to crime rate (see fig 3.16).

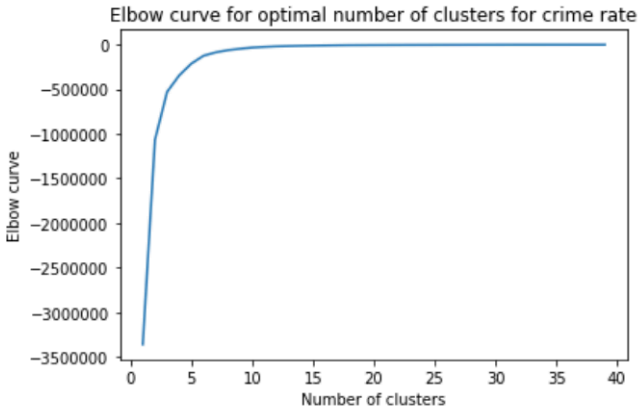

Fig 3.15 Elbow curve for optimal number of clusters for crime rate

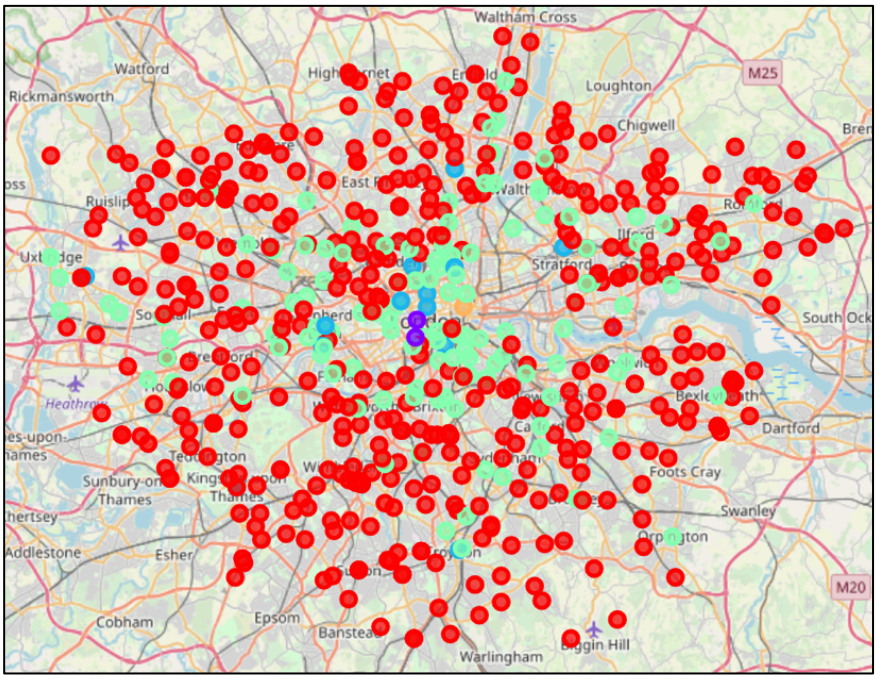

Fig 3.16 London wards clustered according to crime rate

Comparison box plots were generated to show how each of the 5 clusters compared to one another (see fig 3.17). The summary of crime rate clusters are as follows (colours are for corresponding map clusters):

- Cluster0 (red): 2.82 - 3.74 – lowest to average
- Cluster1 (purple): 4.89 - 6.30 – very high
- Cluster2 (blue): 3.77 - 4.86 – high
- Cluster3 (green): 1.32 - 2.81 – above average
- Cluster4 (orange): 6.48 - 8.00 – highest

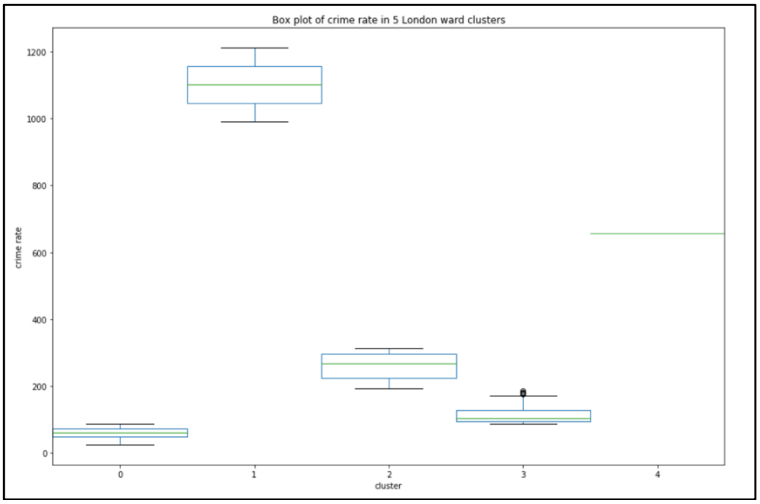

Fig 3.17 Box plot comparison of crime rates in 5 London ward clusters

3.1.5 Public Transport Accessibility

Once again, in order to understand what could be reasonably considered a good, average and bad level of public transport accessibility within London, summary statistics were used. The mean public transport accessibility (3.63) was taken as an acceptable level from which to distinguish between good (higher than 3.63), average and poor (less than 3.63) accessibility. K-Means clustering was used to cluster wards into 5 clusters – determined by devising an elbow curve (see fig 3.18) according to crime rate (see fig 3.19).

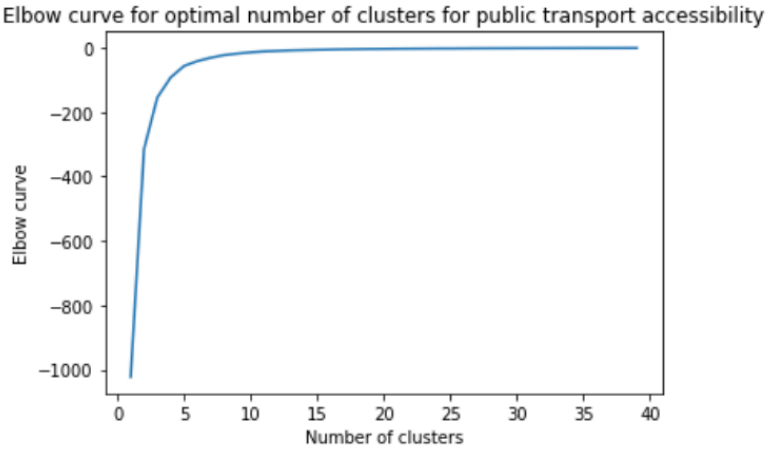

Fig 3.18 Elbow curve for optimal number of clusters for public transport accessibility

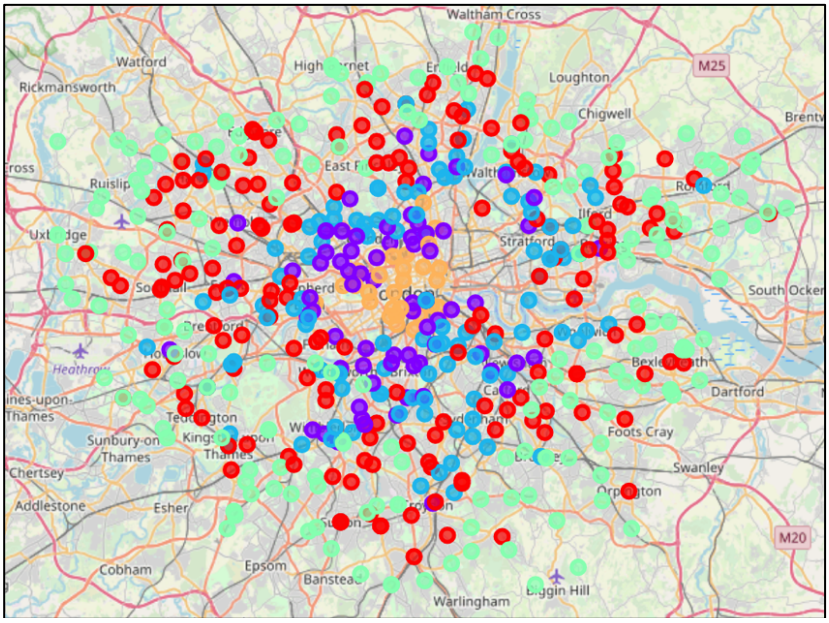

Fig 3.19 London wards clustered according to public transport accessibility

Comparison box plots were generated to show how each of the 5 clusters compared to one another (see fig 3.20). The summary of public transport accessibility clusters are as follows (colours are for corresponding map clusters):
Cluster0 (red): 2.82 - 3.74 – low to average
Cluster1 (purple): 4.89 - 6.30 – high
Cluster2 (blue): 3.77 - 4.86 – above average
Cluster3 (green): 1.32 - 2.81 – lowest
Cluster4 (orange): 6.48 - 8.00 – highest

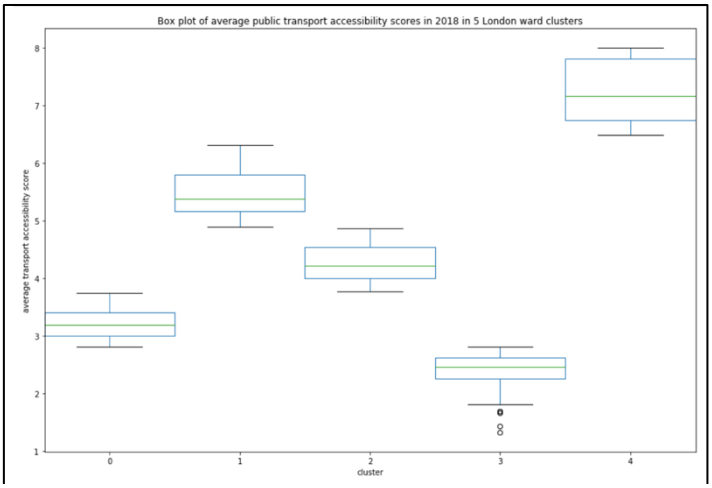

Fig 3.20 Box plot comparison of public transport accessibility in 5 London ward clusters

3.1.6 Open space

K-Means clustering was used to cluster wards into 5 clusters – determined by devising an elbow curve (see fig 3.21) – according to percentage of open space within wards (see fig 3.22).

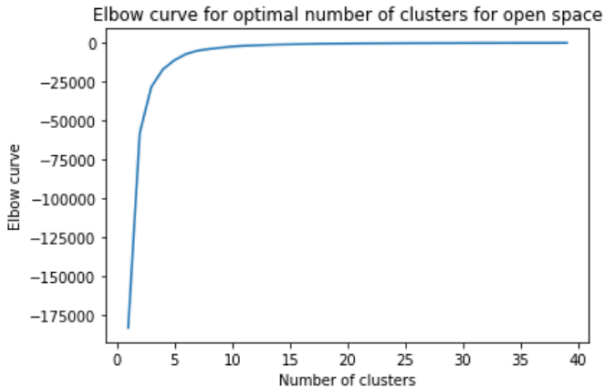

Fig 3.21 Elbow curve for optimal number of clusters for open space

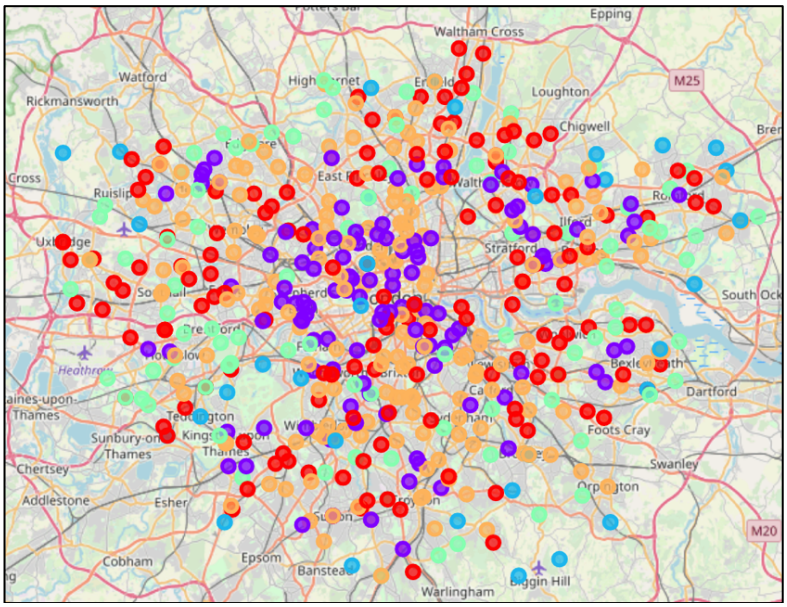

Fig 3.22 London wards clustered according to percentage open space

Comparison box plots were generated to show how each of the 5 clusters compared to one another (see fig 3.23). The summary of open space clusters are as follows (colours are for corresponding map clusters):

- Cluster0 (red): 25.62 - 40.61 – average to above average
- Cluster1 (purple): 0.00 - 12.87 – lowest
- Cluster2 (blue): 60.63 - 88.53 – highest
- Cluster3 (green): 40.95 - 59.08 – above average
- Cluster4 (orange): 12.99 - 25.17 – below average

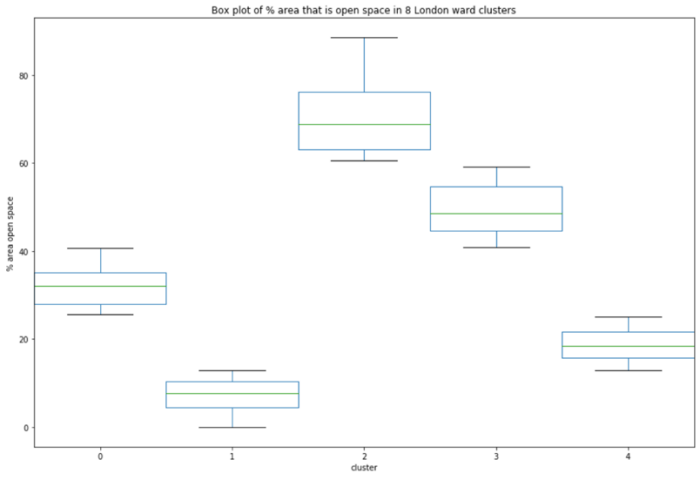

Fig 3.23 Box plot comparison of open space in 5 London ward clusters

3.1.7 Venues

Using venue data generated through Foursquare, one hot encoding was first conducted to convert venue categories into binary values (see fig 3.24).

	Ward loc	Accessories Store	Acupuncturist	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Animal Shelter	Antiqu Sho
0	City of London, City of London	0	0	0	0	0	0	0	0	0	0	
1	City of London, City of London	0	0	0	0	0	0	0	0	0	0	
2	City of London, City of London	0	0	0	0	0	0	0	0	0	0	

Fig 3.24 wards and associated venues in binary form

Then, the data was grouped by wards, and venue categories for each ward were counted from the 100 per ward. The category counts were provided as frequencies of how often the category appears within the wards. Based on these frequencies a dataframe of the 10 most common venue categories in each ward was created (see fig 3.25).

	Ward loc	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abbey Road, Westminster	Pub	Coffee Shop	Deli / Bodega	Italian Restaurant	Café	Thai Restaurant	Bakery	French Restaurant	Middle Eastern Restaurant	Gym / Fitness Center
1	Abbey Wood, Greenwich	Supermarket	Convenience Store	Train Station	Grocery Store	Historic Site	Campground	Coffee Shop	Forest	Furniture / Home Store	Food Court
2	Abbey, Barking and Dagenham	Grocery Store	Supermarket	Hotel	Pub	Sandwich Place	Park	Coffee Shop	Gas Station	Performing Arts Venue	Gym
3	Abbey, Merton	Park	Coffee Shop	Bar	Pub	Burger Joint	Italian Restaurant	Supermarket	Café	Sushi Restaurant	Bookstore

Fig 3.25 dataframe of the top 10 venue categories per ward

An elbow curve was devised to determine the optimal number of clusters for venues. Figure 3.26 shows the elbow curve; unlike the other features, there is no clear 'elbow'. However, the curve appears to flatten from 5 onwards, therefore 5 clusters was selected.

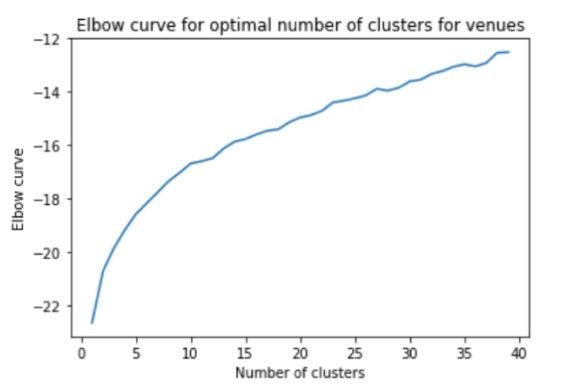

Fig 3.26 elbow curve for optimal number of clusters for venues

K-Means clustering was then used to cluster wards into 5 clusters according to venue type frequency within wards (see fig 3.27).

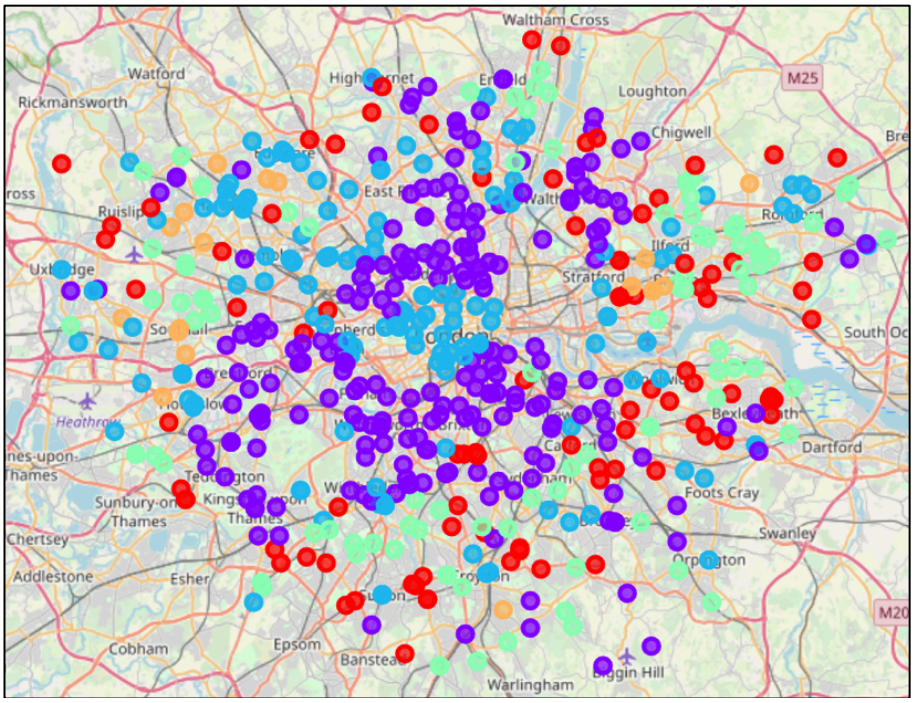

Fig 3.22 London wards clustered according to venues

Given the large number of wards per cluster, clusters were analysed using value count to provide a list of venue categories and the frequency of them in each column. All the clusters contained commonalities (e.g. supermarkets, parks and cafes), so relative frequencies of these, level (1st to 10th) that they occurred, as well as less frequency but distinguishing venue categories were used to characterise clusters. The clusters are as follows (colours are for corresponding map clusters):

Cluster 0 (red): active/busy family cluster - a lot of places families are likely to be found, such as grocery stores/supermarkets, parks, fitness centres, garden centres, as well as family- or sports-type activities (e.g. Go Kart Track, soccer field, athletics and sports, ferry, pools, golf)

Ward loc	Latitude	Longitude	Venue cluster labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mc Comm Ven
Becontree, Barking and Dagenham	51.549265	0.127538	0	Grocery Store	Convenience Store	Bus Stop	Park	History Museum	Soccer Field	Café	Supermarket	Home Service	Cosmet Sh
Eastbury, Barking and Dagenham	51.534474	0.099394	0	Pub	Grocery Store	Chinese Restaurant	History Museum	Food & Drink Shop	Fish & Chips Shop	Martial Arts Dojo	Auto Garage	Supermarket	Ri
Mayesbrook, Barking and Dagenham	51.542299	0.108920	0	Pub	Grocery Store	Park	History Museum	Soccer Field	Metro Station	River	Supermarket	Auto Garage	Chine Restaur

Cluster 1 (purple): young/business professional cluster - a plethora of places to eat, drink and socialise as well as a lot of gyms/fitness centres

Ward loc	Latitude	Longitude	Venue cluster labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10t Co
ast Barnet, Barnet	51.641681	-0.162824	1	Pub	Coffee Shop	Italian Restaurant	Café	Park	Grocery Store	Playground	Chinese Restaurant	Convenience Store	Fas Res
t Finchley, Barnet	51.587188	-0.164814	1	Café	Coffee Shop	Park	Grocery Store	Bakery	Indian Restaurant	Italian Restaurant	Pub	Pizza Place	
n Suburb, Barnet	51.650054	-0.198560	1	Coffee Shop	Grocery Store	Pub	Bookstore	Italian Restaurant	Pizza Place	Park	Café	Fast Food Restaurant	Sa

Cluster 2 (blue): tourist/traveller/visitor cluster - a concentration of hotels, restaurants, cafes, shops, attractions (e.g. zoo, theatre, theme parks, galleries), and transportation options (e.g. airport lounge, bus stations, rail stations)

Ward loc	Latitude	Longitude	Venue cluster labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mc Comm Ven
y of London, ty of London	51.515618	-0.091998	2	Coffee Shop	Gym / Fitness Center	Hotel	Cocktail Bar	Salad Place	Steakhouse	Garden	Food Truck	French Restaurant	Wine
Burnt Oak, Barnet	51.604988	-0.264450	2	Coffee Shop	Supermarket	Grocery Store	Sandwich Place	Bakery	Clothing Store	Climbing Gym	Park	Salon / Barbershop	G
Childs Hill, Barnet	51.562982	-0.197249	2	Coffee Shop	Café	Park	Grocery Store	Korean Restaurant	Turkish Restaurant	Pub	Chinese Restaurant	Sushi Restaurant	Mid East Restaur

Cluster 3 (green): urban family cluster - a lot of places families are likely to be found, such as supermarkets/grocery stores, parks, family-type shops (e.g. furniture and clothing stores) and activities (e.g. multiplex, movie theatres, indoor play areas)

Ward loc	Latitude	Longitude	Venue cluster labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mc Comm Ven
Abbey, arking and Dagenham	51.535688	0.075530	3	Grocery Store	Supermarket	Hotel	Pub	Sandwich Place	Park	Coffee Shop	Gas Station	Performing Arts Venue	Gy
Alibon, arking and Dagenham	51.547549	0.153114	3	Grocery Store	Platform	Bus Stop	Supermarket	Soccer Field	Pub	Soccer Stadium	Furniture / Home Store	Bar	Me Stati
Chadwell Heath, arking and Dagenham	51.567904	0.128041	3	Supermarket	Grocery Store	Electronics Store	Train Station	Pizza Place	Fast Food Restaurant	Shopping Plaza	English Restaurant	Gym	Gym Fitne Cen

Cluster 4 (orange): dining out and eating in cluster - a concentration of restaurants and take away establishments

	Ward loc	Latitude	Longitude	Venue cluster labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mos Commo Venu
	Whalebone, Barking and Dagenham	51.593602	0.142199	4	Chinese Restaurant	Indian Restaurant	Gym	Bowling Alley	Grocery Store	Fast Food Restaurant	Event Space	Zoo Exhibit	Fish & Chips Shop	Factor
	Queensbury, Brent	51.594190	-0.286181	4	Indian Restaurant	Grocery Store	Coffee Shop	Park	Gym / Fitness Center	Supermarket	Portuguese Restaurant	Mobile Phone Shop	Sandwich Place	Metr Statio
	Croham, Croydon	51.352181	-0.080840	4	Hotel	Indian Restaurant	Grocery Store	Beer Garden	Wine Bar	Park	Fish & Chips Shop	Exhibit	Factory	Falafe Restaurar

4 Results – applying the case study to the clusters

Now that the data has been analysed and clustered, we can use the case study information and property buyer's preferences to score clusters within each feature, and generate a weighted total score to determine which wards are most suited to the property buyers.

First, we provide each cluster within features a score based on the house buyer's preferences for features. Then we note the relative weighting the house buyer's give to each feature.

4.1 Scores for clusters within features

We then assign a score to each cluster within a feature based on how closely it approximates the buyer's preferences, with the following guidelines:

- Scores are not exclusive - two clusters may be equally preferable and so can have the same score
- The maximum score is 5, with 5 indicating a perfect match
- A large negative value (-1000) is assigned when an option needs to be removed from the running e.g. a price that the buyer would not be able to afford
- A score of 0 indicates the option is not a preference but is not a sufficient condition to need it to be removed from the running with a large negative value

4.1.1. Property Price

The house buyer has a budget of up to £800,000. With clusters based on the mean property price within wards, with expected variability within wards, and with the potential for price negotiation, all clusters containing wards with a mean property price of up to £1m will be considered as a feasible option. All clusters containing wards with a mean property price exclusively above £1m will be excluded from the dataset by assigning them a large negative value (-1000). Clusters containing average property prices slightly below the mean will be awarded a high score as average price indicates just that, with property prices likely to vary around this, and so the buyer will be able to afford a property at the upper end of the market in these locations.

The clusters, and scores assigned, are as follows:

Cluster	Mean price range	Score
0	£524,100.61 - £648,806.41	5
1	£4,416,659.40	-1000
2	£1,296,800.86 - £1,585,562.76	-1000
3	£259,486.62 - £401,320.31	1
4	£823,462.88 - £1,008,300.62	2
5	£1,891,715.94 - £2,186,666.67	-1000
6	£403,917.61 - £518,567.23	4
7	£2,856,435.71 - £3,033,756.54	-1000
8	£649,877.67 - £817,646.99	5
9	£1,021,486.54 - 1,238,295.89	-1000

4.1.2 Property Type

The house buyer expressed a preference for a house over a flat, apartment or maisonette. They are indifferent between detached and semi-detached as their first choice, with terraced houses being the second choice.

The clusters, and scores assigned, are as follows:

Cluster	Property types	Scores
0	detached (1), semi-detached (1), terraced (4), flats(5)	5
1	detached (5), semi-detached (5), terraced (5), flats(1)	0
2	detached (2), semi-detached (2), terraced (3), flats(3)	4
3	detached (3), semi-detached (3), terraced (1), flats(4)	2
4	detached (4), semi-detached (4), terraced (2), flats(2)	1

4.1.3 Primary Education

With two children in primary school, the quality of primary education is important to the property buyer. The higher the percentage of children achieving the expected level the better. The mean percentage (75%) and above is taken as indication of the primary education being of an acceptable standard; any clusters with wards below the mean level are assigned a value of 0.

The clusters, and scores assigned, are as follows:

Cluster	Range of % of children achieving expected primary level	Score
0	71.38-73.62% - below average	0
1	76.38-78% - above average	4
2	74.38-75.88% - low average	0
3	68.88-69.12% - lowest level	0
4	82.62% - highest level	5

4.1.4 Secondary Education

With one child in secondary school, the quality of secondary education is important to the case study property buyers. The higher the average GCSE point score of a ward the better. The mean point score (327) and above is taken as indication of the secondary education being of an acceptable standard; any clusters with wards below the mean level are assigned a value of 0.

The clusters, and scores assigned, are as follows:

Cluster	Range of average GCSE point scores	Score
0	306.94 - 322.14 – below average	0
1	338.09 - 357.23 – above average	4
2	322.44 - 337.78 – low to average	2
3	275.95 - 306.49 – lowest level	0
4	358.47 - 396.21 – highest level	5

4.1.5 Crime Rate

With children, safety is important. Crime rate has been used as a proxy for safety - the higher the crime rate, the less safe the ward. A crime rate of at or below the mean is considered acceptable and will generate a score for the cluster. Clusters that contain wards predominantly above the mean crime rate (80.87) will be assigned a value of 0

The clusters, and scores assigned, are as follows:

Cluster	Range of crime rates	Score
0	24.50 - 86.42 – lowest to average crime rate	5
1	990.00 - 1212.13 – very high crime rate	0
2	192.22 - 314.13 – high crime rate	0
3	86.64 - 184.91 – above average crime rate	0
4	656.38 – highest crime rate	0

4.1.6 Public Transport Accessibility

The wife has her main work office in central London. The requirement to commute into central London for work means access to public transport is necessary.

The public transport accessibility scores calculated by Transport for London and GLA is considered an adequate measure of access to public transport within a ward. Each area has been given an average score out of 8, with 8 being the highest level of accessibility. Clusters containing wards with the mean (3.63) or above is considered an acceptable level, with clusters below this being assigned a score of 0.

The clusters, and scores assigned, are as follows:

Cluster	Range of public transport accessibility scores	Score
0	2.82 - 3.74 – low to average	2
1	4.89 - 6.30 – high accessibility	4
2	3.77 - 4.86 – above average accessibility	3
3	1.32 - 2.81 – lowest accessibility	0
4	6.48 - 8.00 – highest accessibility	5

4.1.7 Open Space

With three children and the husband being a personal trainer (offering the option of training clients outdoors), the property buyer expressed an interest in areas with open space. As they are moving from a rural area to London because they want to be in a more accessible area with options for the children, they want enough open space to meet their needs but not so much that they are just moving to yet another rural area. The mean percentage open space of London wards (27.6%) will be considered optimal.

The clusters, and scores assigned, are as follows:

Cluster	Range of % open space of wards	Score
0	25.62 - 40.61	5
1	0.00 - 12.87	1
2	60.63 - 88.53	1

3	40.95 - 59.08	3
4	12.99 - 25.17	4

4.1.8 Venues

The property buyer is looking for a family area. They want an area with things to do with the children. With the husband being a personal trainer, an active area with plenty of gym and park access would be ideal. They want more of a community-type area with a fairly stable population as opposed to a tourist area.

The clusters, and scores assigned, are as follows:

Cluster	Venues/area character	Score
0	active/busy family cluster	5
1	young/business professional cluster	0
2	tourist/traveller/visitor cluster	0
3	urban family cluster	5
4	eating in and dining out cluster	0

4.2 Weight of Features

We will assume that the case study property buyers, when providing a list of criteria that are important to them when choosing a location to buy a property in, these were accompanied with a score of 1-10 to indicate the relative importance of each feature to them - the higher the score, the more important the criteria. These scores assist in calculating a weighted score for each ward. The scores for features are as follows:

- Price - for those within budget 0.7
- Property type – 0.7
- Secondary education – 0.9
- Primary education – 0.9
- Safety/lack of crime – 0.8
- Transport accessibility – 0.8
- Open space – 0.5
- Area type (venues) – 1.0

4.3 Calculating ward scores based on buyer preferences

To calculate the score for each ward, the score for individual features was multiplied by the weight of that feature and weighted feature scores were added together.

5 Discussion

From the dataframe containing each ward's overall score, the top ten scoring wards were extracted (see fig 5.1). These top ten wards are the recommendations that would be made to the case study property buyer. In first place is Carshalton Central, Sutton.

Ward loc	Score
Carshalton Central, Sutton	26.8
West Wickham, Bromley	26.6
Headstone North, Harrow	25.6
Bickley, Bromley	25.4
Headstone South, Harrow	25.4
Cheam, Sutton	25.4
Hatch End, Harrow	25.0
Farnborough and Crofton, Bromley	24.7
Chelsfield and Pratts Bottom, Bromley	24.6
Stonecot, Sutton	24.5

Fig 5.1 top ten scoring wards

To understand these recommendations, their locations were plotted on a map of London (see fig 5.2, and the breakdown of scores along with the total was put into a dataframe (see fig 5.3).

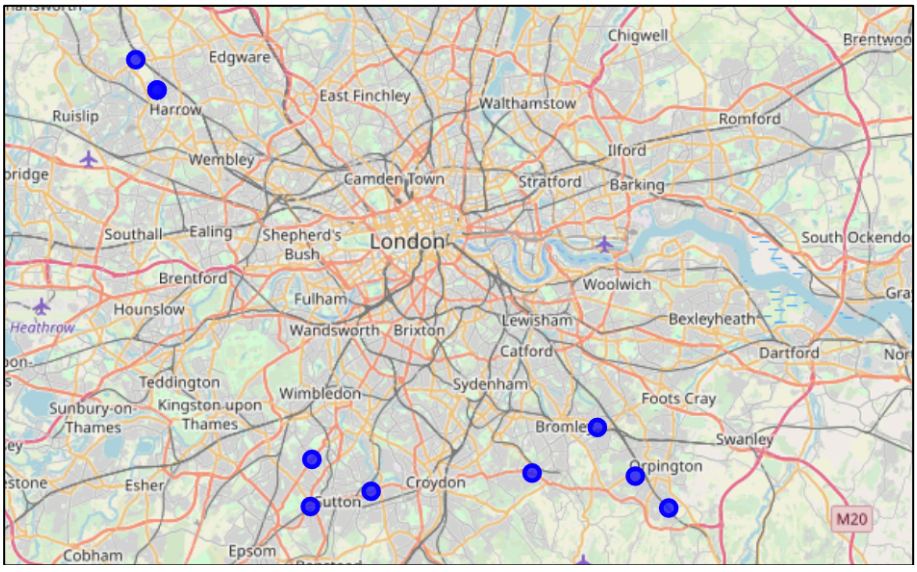

Fig 5.2 London map showing the location of the top ten recommended wards

	Ward loc	Score	Latitude	Longitude	Price score	Property type score	Primary score	Secondary score	Crime score	Transport score	Space score	Venue score
0	Carshalton Central, Sutton	26.8	51.365788	-0.161086	4	4	4	5	5	2	5	5
1	West Wickham, Bromley	26.6	51.375804	-0.014684	5	5	4	5	5	0	5	5
2	Headstone North, Harrow	25.6	51.591883	-0.354217	5	5	4	5	5	0	3	5
3	Bickley, Bromley	25.4	51.401740	0.043712	5	4	4	5	5	0	4	5
4	Headstone South, Harrow	25.4	51.591883	-0.354217	4	4	4	4	5	2	4	5
5	Cheam, Sutton	25.4	51.357616	-0.216241	5	4	4	5	5	0	4	5
6	Hatch End, Harrow	25.0	51.608440	-0.373548	5	4	4	4	5	0	5	5
7	Farnborough and Crofton, Bromley	24.7	51.374000	0.077803	5	5	4	4	5	0	3	5
8	Chelsfield and Pratts Bottom, Bromley	24.6	51.356611	0.108328	5	5	4	5	5	0	1	5
9	Stonecot, Sutton	24.5	51.383526	-0.214193	4	5	4	4	5	0	4	5

Fig 5.3 dataframe of the top ten scoring wards and the feature score breakdown

The map shows that all of the locations recommended are on the outskirts of Greater London. As distance from central London was not a feature specified by the buyer, this still means that the recommendations would be satisfactory to the buyer. As indicated by the cluster maps for price and venues, the more central to London the location is, the more expensive and the more it caters to the young or the business professional, or the tourist, visitor or traveller. Furthermore, the south and west of London are consistent with higher average GCSE point scores and primary level achievement.

As shown in the table of the top ten score breakdown, the majority score highly for all features, except for accessibility to public transport. As shown on the public transport accessibility cluster map, the further the location from central London, the lower the public transport accessibility, thus the low scores achieved by the top ten, which are all in outer Greater London, are consistent with this. This points to a trade-off for those living in London, especially those with children – areas more suited to families, with better educational performances and lower crime rates have lower levels of accessibility to the public transport network. With the case study buyer requiring public transport access to travel to work, the low scores for public transport accessibility may make these recommendations unsuitable. However, poor scoring does not mean that there is no access, and there will be different travel times depending on destinations – each location would need to be reviewed to determine actual travel times to work and other locations to decide suitability.

Open space was not as consistent as most other features. This is due to it being of a lower priority to the buyer. Given that the recommendations are all in outer Greater London, there is more open space, rather than less, than declared as optimal.

6 Conclusion

6.1 Final Result

This project identified some key variables that are important features people consider when deciding on a location within which to purchase a property – property price, property type, quality of education, crime rate, public transport accessibility, open space and the character of the area.

Utilising London DATASTORE data on London wards, geocoder and location IQ for latitude and longitude coordinates, and Foursquare for popular venues within each ward, and exploring the features using K-Means clustering, each feature was clustered according to commonality. The resulting clusters were applied to a case study property buyer and their preferences to generate a top ten list of recommended London wards to search for a property in. The top ten recommended wards were all characterised as family areas, with good primary and secondary education performance, low to average crime rate, within price range, and containing property types to the buyer's preference. Open space scores were more variable as it was considered less important by the buyer, and public transport accessibility consistently scored low for the recommendations.

4.2 Limitations and Recommendations for Future Study

As noted previously, all the recommendations are located in the outskirts of Greater London. This is not necessarily an issue for the case study buyer – it was not a feature explicitly stated. However, to further refine the project, distance from central London could be an additional variable to consider and for buyers to rate.

With public transport accessibility consistently scoring low for the recommendations, and with the buyers weighting it relatively highly, scores of zero may not be acceptable. To further refine the rating system, a limit to prevent those rated zero could be placed on the calculation. This would, however, likely lead to a trade-off with lower scores for the other features. This requires feedback from buyers as to their preferences.

Further testing and refining of the model could improve its performance. For example, additional cases with different needs and preferences could be tested. Furthermore, additional variables could be added that may be important to other buyers; for example, some buyers may think healthcare services are an important consideration.

One issue with this project is that the data used, although the most up-to-date that could be obtained, was still dated (most was 2014/15). For certain features, changes will have occurred between recording of the data and the use of it in this project. For example, the GCSE system has changed recently, which may impact quality of secondary education throughout London, and works are constantly being conducted on the public transport network, which could alter the accessibility within wards. More recent data would provide more accurate results.

References

Foursquare Developer, https://developer.foursquare.com/

Lesson notes from IBM Professional Data Science Certificate Specialization, Coursera

LocationIQ, https://locationiq.com/

London DATASTORE, 'ward-profiles-excel-version.xls' https://data.london.gov.uk/download/ward-profiles-and-atlas/a187b63e-bf4f-4449-b644-ab86a0a8569d/ward-profiles-excel-version.xls

London DATASTORE, 'ks1-results.xls' https://data.london.gov.uk/download/key-stage-1-results-by-borough/ab9fc44c-a6d7-4649-8f22-5726b7bc47ff/ks1-results.xls

London DATASTORE, 'ks2-results.xls' https://data.london.gov.uk/download/key-stage-2-results-by-borough/46a5a04b-35bf-4152-8816-8242a57c12b7/ks2-results.xls

London DATASTORE, 'land-registry-house-prices-ward.xls' https://data.london.gov.uk/download/average-house-prices/fb8116f5-06f8-42e0-aa6c-b0b1bd69cdba/land-registry-house-prices-ward.xls

London property sales volume, HM Land Registry Data (2018) reported on Plumplot, https://www.plumplot.co.uk/London-property-sales.html

Strategic Report (2019), Rightmove