Location, Location, Location – Finding the best London borough

Carlos Noble Jesus

February 27, 2020

1 Introduction

London is the largest city in the UK, with almost 9 million people currently residing there (Office for National Statistics, 2019). Moving to or within the city can be a difficult task as the 32 main boroughs that make up London are vastly different. Choosing the right borough is in important decision due to the financial, safety and lifestyle considerations that an individual has to make. As such, the ability of classify the London boroughs and identify the best borough for an individual may be an invaluable tool for newcomers to the city.

Furthermore, being able to sort boroughs by desirability can also be useful in a business or policy context. Understanding why particular boroughs are less desirable than others may allow for changes to be made to improve the area. For instance, identifying a lack of a certain amenity or type of restaurant may open up many business opportunities. Stakeholders may want a company to expand in more affluent areas, or may wish to find a location where the market remains largely unfulfilled. Similarly, pinpointing the areas with the highest crime rates may allow for local government to more efficiently focus their resources.

2 Data

For this project, many data sources were utilised to produce a large dataset of features for each borough. Crime data from February 2018-January 2020 was obtained from the Metropolitan Police. This was used to calculate total number of crimes over the previous two years for each borough. Several other datasets were accessed from the London Datastore, an open data-sharing portal managed by the London government. NO2 emissions data was obtained and an additional column was created to identify the boroughs that were exceeding the EU limit. Data for London schools was cleaned to remove any irrelevant columns and schools that had been closed, before being grouped to provide a school count for

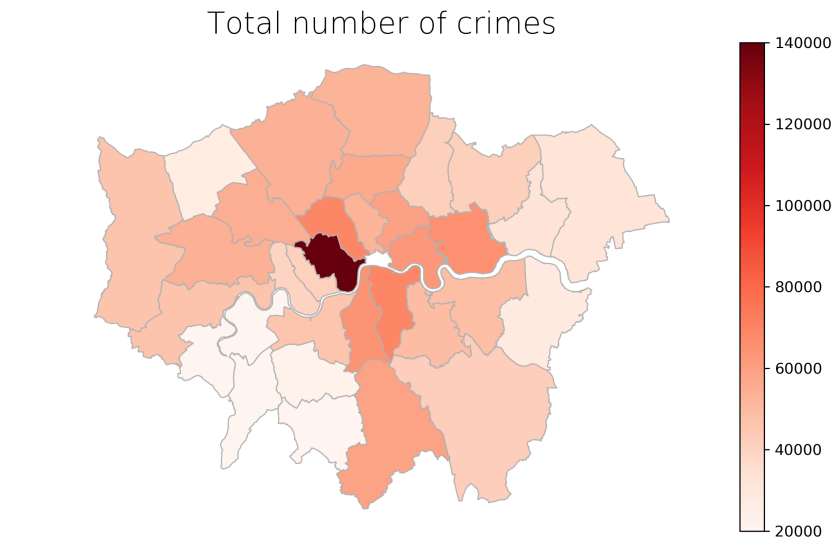
each borough. House price data for each borough was used to find the median house price for the most recent time period available.

A Wikipedia table was parsed to get further information for each borough, such as coordinate data, area in square miles, population numbers and political control. The table was cleaned up to remove any extraneous text (such as footnotes or HTML formatting) and convert numerical data that was formatted as strings into integers or floats, as appropriate. The population data was updated with more recent data from a dataset accessed via the London Datastore. GeoJSON and shape data were also obtained to produce borough boundaries and create choropleth maps. Finally, FourSquare API calls were made to access the location data for parks, supermarkets and hospitals for each borough. Combining this data allowed for a comprehensive set of features for each borough, which were used for data visualisation and machine learning, to identify certain strengths and weaknesses.

Some of the data was normalised to allow for a better representation of many features. For instance crime and school data was normalised by population, and population data was normalised by the area of each borough to provide population density values. The most recently available public data was used where possible. However, some datasets were from different years – e.g. school data is from 2016 whereas crime data is from 2020. While there is unlikely to be any major distinctions in the relationships between features because of these discrepancies, it is worth noting for the sake of completeness.

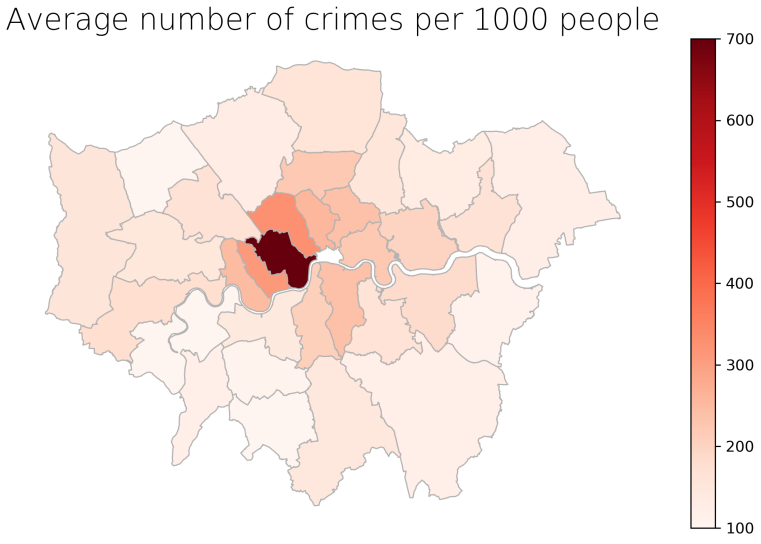
3 Methodology

Exploratory data analysis was carried out to visualise how many of the features differed by population. For instance, choropleth maps of crime data showed the most central boroughs had the highest number of crimes (Figure 3.1). It also showed that a fairly high number of crimes had been committed in some of the outer boroughs. However, normalising the data by population revealed that the crime rate, measured by number of crimes per 1000 people, was directly related to how central the borough was (Figure 3.2). This demonstrates the importance of



Source: Metropolitan Police, 2020

Figure 3.1: Total number of crimes in each borough



Source: Metropolitan Police, 2020

Figure 3.2: Number of crimes per 1000 people in each borough

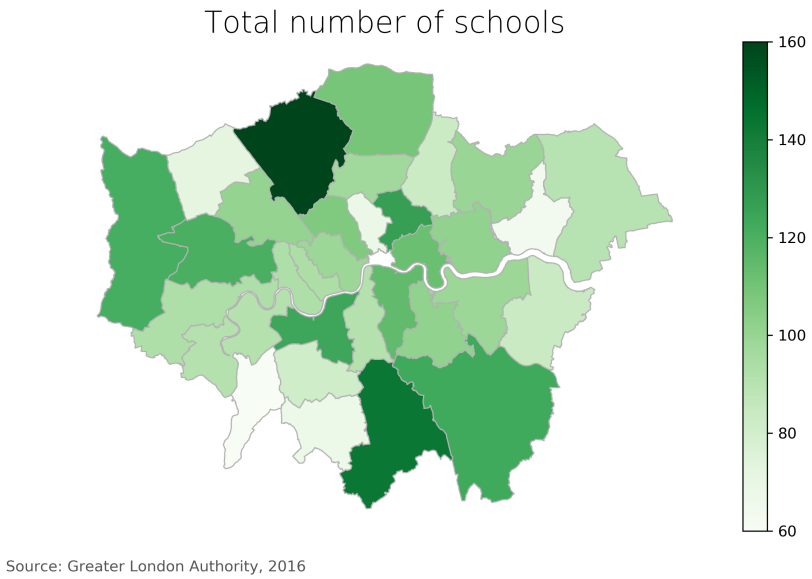
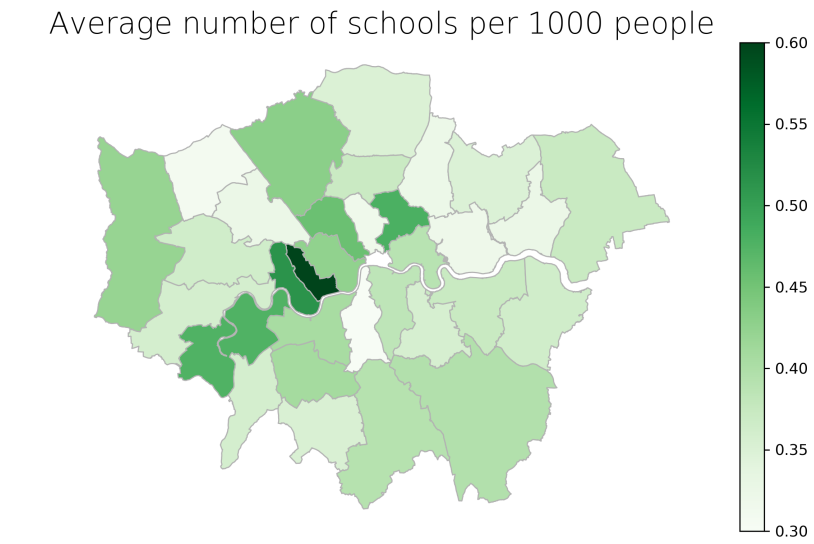


Figure 3.3: Total number of schools in each borough

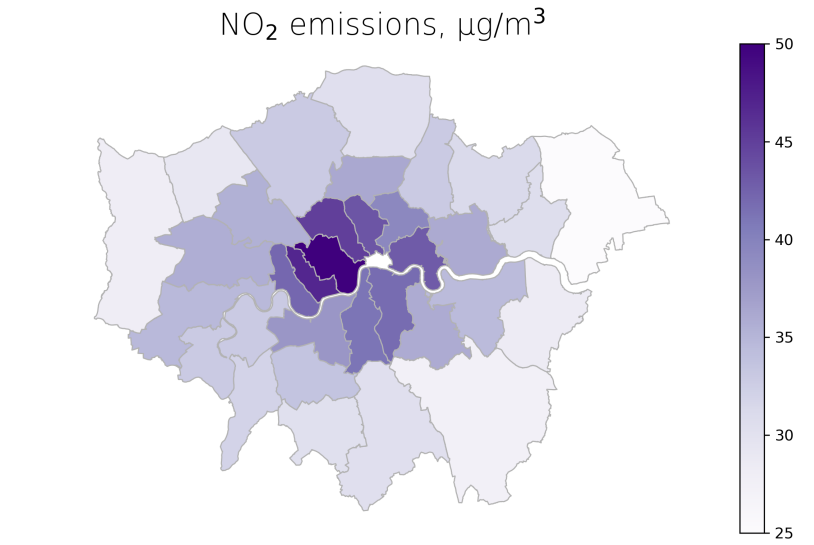
normalising certain parameters to highlight the true underlying relationship. A similar relationship was observed for the school count data, where the total number of schools (Figure 3.3) does not accurately reflect the relative number of schools per population (Figure 3.4).

Several other features showed a significant difference between the most central boroughs and the outer boroughs. For instance, air pollution, as measured by NO_2 emissions, was highest in the central boroughs (Figure 3.5). In fact, eight of the most central boroughs had NO_2 emissions exceeding the EU limit (Figure 3.6). The median house price was perhaps unsurprisingly also highest in the inner boroughs (Figure 3.7). Many of these features show a very similar trend, having the highest values in the inner boroughs. Therefore, the Pearson correlation coefficients were calculated to look for any correlations between these values. This demonstrated, for instance, that the normalised crime rate had positive correlations with population density (0.5729), median house price (0.6738) and NO_2 emissions (0.7833). Interestingly, the crime rate was also strongly correlated with positive features such as the number of hospitals (0.8284). As such, these relationships are most likely a result of the higher population density and more advanced infrastructure of inner city boroughs.



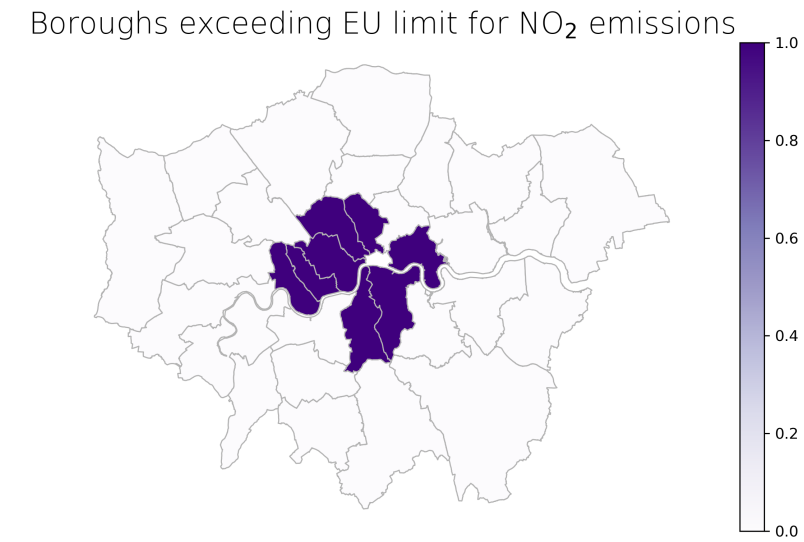
Source: Greater London Authority, 2016

Figure 3.4: Number of schools per 1000 people in each borough



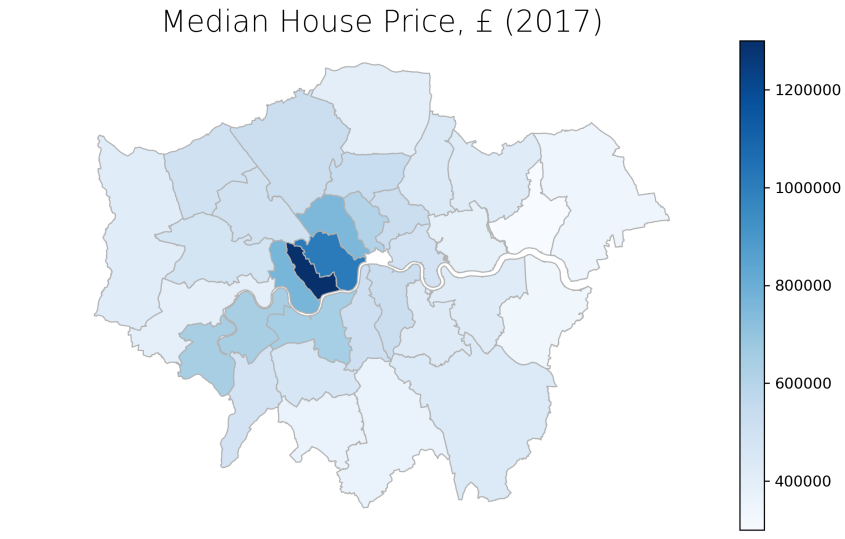
Source: London Datastore, 2018

Figure 3.5: NO_2 emissions in each borough



Source: London Datastore, 2018

Figure 3.6: Boroughs that exceed NO_2 emission limits



Source: London Datastore, 2018

Figure 3.7: Median house prices in each borough

	Latitude	Longitude	Population	Area (sq mi)	Pop Density	Average Age	Median House Price	Total Crimes	Crimes per 1000	School Count	Schools per 1000	NO2_value	exceed_NO2	Parks	Supermarkets	Hospitals
Latitude	1.000000	0.120888	0.204037	0.004139	0.094228	-0.221280	0.013501	0.175793	0.097107	0.120497	-0.112047	0.003650	-0.008704	-0.131622	-0.111270	-0.092257
Longitude	0.120888	1.000000	0.113751	0.044846	0.064431	-0.165191	-0.279085	0.070163	0.007810	-0.037383	-0.228689	-0.117592	-0.019259	-0.188692	0.036576	-0.086534
Population	0.204037	0.113751	1.000000	0.404083	-0.114601	-0.271189	-0.402813	0.348898	-0.157015	0.757189	-0.360456	-0.178855	-0.242426	0.165877	0.089707	-0.123569
Area (sq mi)	0.004139	0.044846	0.404083	1.000000	-0.786447	0.523293	-0.450305	-0.234060	-0.431678	0.371763	-0.112789	-0.753088	-0.524079	-0.219343	-0.226385	-0.307213
Pop Density	0.094228	0.064431	-0.114601	-0.786447	1.000000	-0.572999	0.532429	0.493735	0.572910	-0.003660	0.227589	0.884903	0.752482	0.302969	0.284012	0.426764
Average Age	-0.221280	-0.165191	-0.271189	0.523293	-0.572999	1.000000	0.197557	-0.289689	-0.112565	-0.078178	0.272802	-0.360322	-0.169980	-0.102812	0.107821	0.077466
Median House Price	0.013501	-0.279085	-0.402813	-0.450305	0.532429	0.197557	1.000000	0.383191	0.673753	0.026013	0.722004	0.759405	0.646326	0.297399	0.251051	0.615719
Total Crimes	0.175793	0.070163	0.348898	-0.234060	0.493735	-0.289689	0.383191	1.000000	0.863667	0.381517	0.058448	0.649158	0.496317	0.368072	0.444338	0.721642
Crimes per 1000	0.097107	0.007810	-0.157015	-0.431678	0.572910	-0.112565	0.673753	0.863667	1.000000	0.040857	0.312129	0.783268	0.657832	0.283730	0.404572	0.828366
School Count	0.120497	-0.037383	0.757189	0.371763	-0.003660	-0.078178	0.026013	0.381517	0.040857	1.000000	0.312105	0.047531	-0.064984	0.266998	0.134916	0.045083
Schools per 1000	-0.112047	-0.228689	-0.360456	-0.112789	0.227589	0.272802	0.722004	0.058448	0.312129	0.312105	1.000000	0.403146	0.331912	0.187500	0.023670	0.269559
NO2_value	0.003650	-0.117592	-0.178855	-0.753088	0.884903	-0.360322	0.759405	0.649158	0.783268	0.047531	0.403146	1.000000	0.833110	0.390038	0.404430	0.659196
exceed_NO2	-0.008704	-0.019259	-0.242426	-0.524079	0.752482	-0.169980	0.646326	0.496317	0.657832	-0.064984	0.331912	0.833110	1.000000	0.168932	0.309886	0.617634
Parks	-0.131622	-0.188692	0.165877	-0.219343	0.302969	-0.102812	0.297399	0.368072	0.283730	0.266998	0.187500	0.390038	0.168932	1.000000	0.319292	0.290840
Supermarkets	-0.111270	0.036576	0.089707	-0.226385	0.284012	0.107821	0.251051	0.444338	0.404572	0.134916	0.023670	0.404430	0.309886	0.319292	1.000000	0.535254
Hospitals	-0.092257	-0.086534	-0.123569	-0.307213	0.426764	0.077466	0.615719	0.721642	0.828366	0.045083	0.269559	0.659196	0.617634	0.290840	0.535254	1.000000

Figure 3.8: Correlation between the variables with numerical values

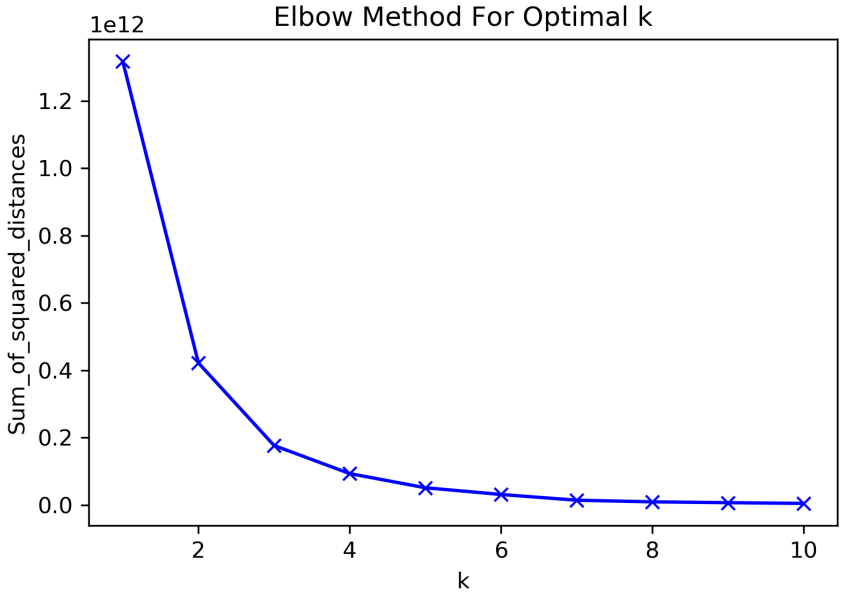


Figure 3.9: Elbow method to obtain the optimal value for k

After exploring the data, an unsupervised machine learning technique called K-Means clustering was employed to group similar boroughs into clusters. A separate dataframe was created with only the features of interest to be used in the model. To optimise the number of clusters used for the model, the elbow method was implemented. This graphs the error vs the number of clusters (k). The 'elbow point' of this graph shows the optimal number of clusters to use. For my data, the elbow point was found at k=3 (Figure 3.9). The K-Means model was then fitted with the data using k=3 to group the boroughs into 3 clusters.

4 Results

The three clusters were plotted over a Folium map to visualise their distribution (Figure 4.1). At first glance, it appears that the smallest cluster is made up of the two most central boroughs and the majority of the largest cluster is made up of the outer boroughs. To analyse the composition of each cluster, the correlation between the cluster label (0, 1, 2) and the features used in the model was calculated (Figure 4.2). The correlation suggested that median house price may be one of the most influential features in clustering the boroughs.

Box plots of certain features for each cluster were produced to understand the distribution of that feature in each cluster. Indeed, the median house price was observed to be significantly different between clusters (Figure 4.3). Other features, such as normalised crime and school data, exhibited clear differences between each cluster, but not to the same extent. These findings suggested that the median house price was the main feature by which the boroughs were clustered.

To demonstrate this finding, the clusters for each borough were colour-coded – green for the cluster with the lowest median house price, red for the cluster with the highest median house price, and yellow for the cluster in between. This was plotted over a choropleth map of median house prices in each borough to confirm

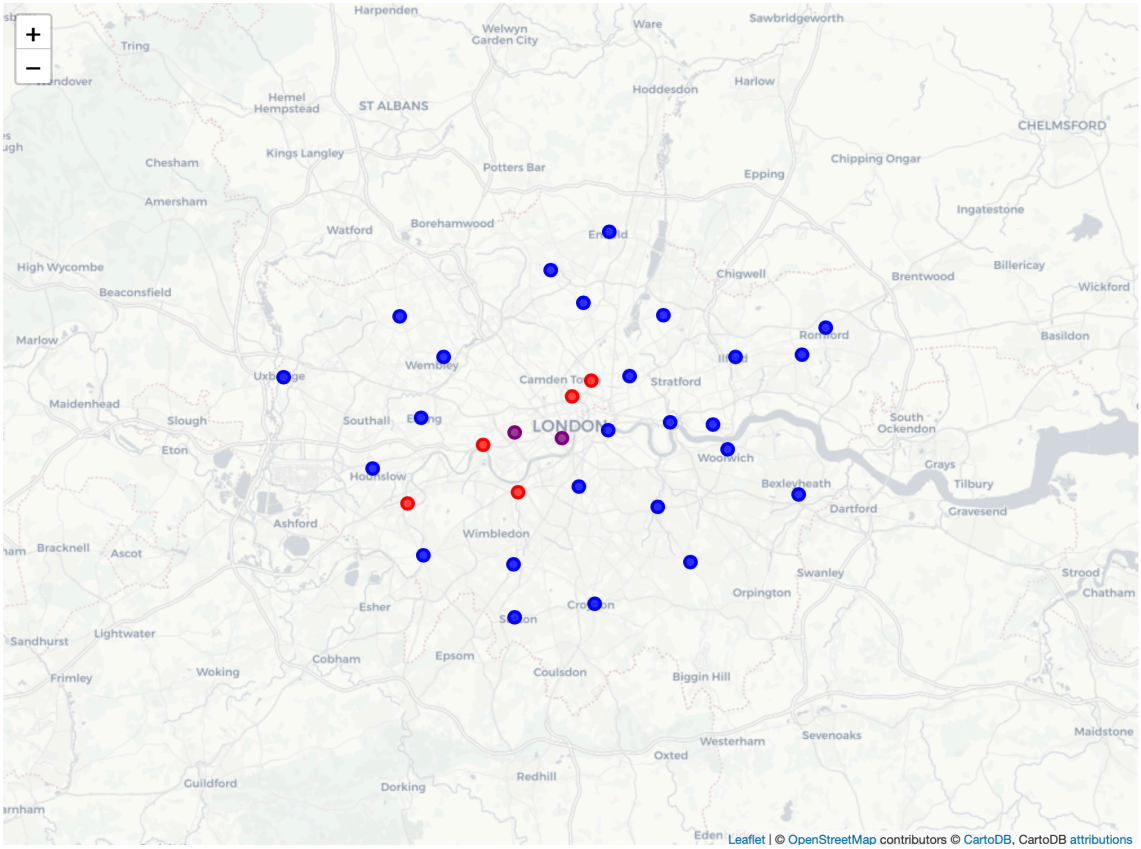


Figure 4.1: Initial clustering of boroughs

Median House Price	0.610855
exceed_NO2	0.487950
Schools per 1000	0.479676
Pop Density	0.382992
Supermarkets	0.365960
Crimes per 1000	0.341027
Hospitals	0.295030
Parks	0.278380

Figure 4.2: Correlation between cluster label and model features

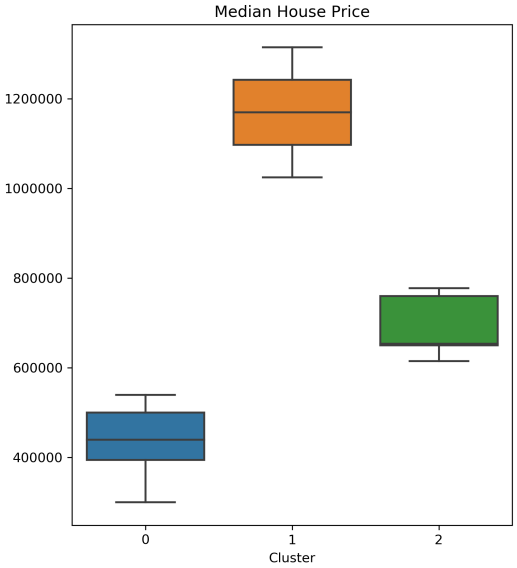
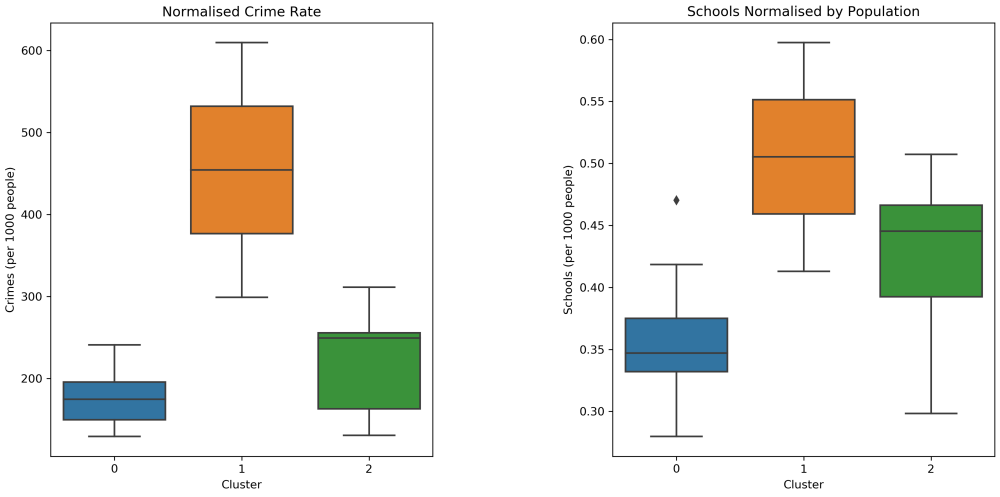


Figure 4.3: Box plot of median house price for each cluster



(a) Crime rate per 1000 people

(b) School count per 1000 people

Figure 4.4: Box plots of model features for each cluster

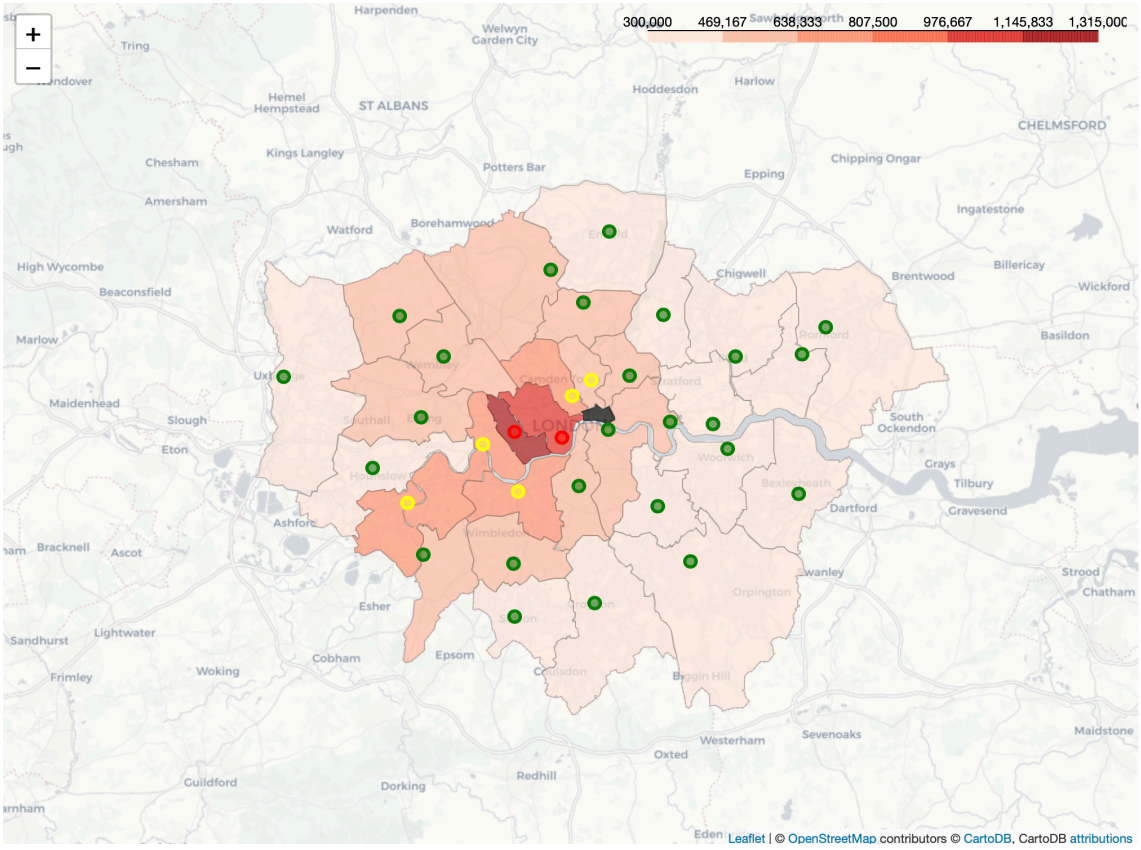


Figure 4.5: Median house price choropleth map overlayed with borough clusters

the relationship (Figure 4.5).

5 Conclusions

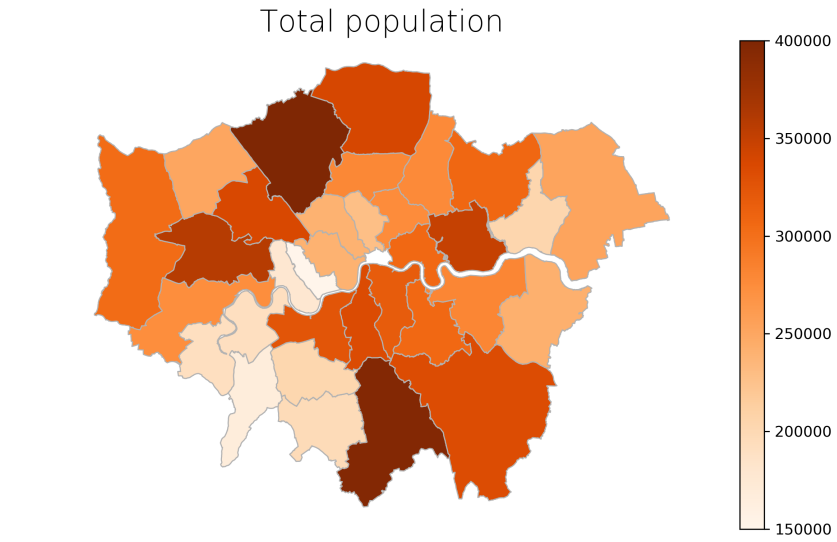
In this project, several features of each London borough have been visualised and analysed to determine the best areas to live. K-Means clustering allowed for the boroughs to be grouped into clusters of desirability. The main criteria for the clusters were found to be median house price, air pollution and normalised school count. However, there was also a slight correlation with other features such as crime rate and the number of several amenities such as supermarkets and hospitals. As such, it can be said that:

- Green cluster (cluster 0) – low house prices, low air pollution, low crime rate, low number schools per population, lowest number of amenities
- Yellow cluster (cluster 2) – medium house prices, medium/high air pollution, low/medium crime rate, medium number of schools per population, medium number of amenities
- Red cluster (cluster 1) – high house prices, high air pollution, high crime rate, high number of schools per population, highest number of amenities

Therefore, depending on an individual's priorities, each cluster may appear less or more desirable. Many may choose a borough in the green cluster due to the lower house prices and crime rates, but some may prefer having more amenities and schools in a borough in the yellow or red clusters. It should be noted that the red cluster only contains two boroughs – 'Kensington and Chelsea' and 'Westminster' – thus suggesting these boroughs vary greatly from the others as evidenced by their crime rate and house prices. However, this data could be valuable for the government to try and tackle crime and implement more affordable housing in these central boroughs. Furthermore, companies could utilise this data to better understand where many services are lacking and capitalise on this (e.g. opening new supermarkets), or to implement eco-friendly policies to reduce pollution in the capital.

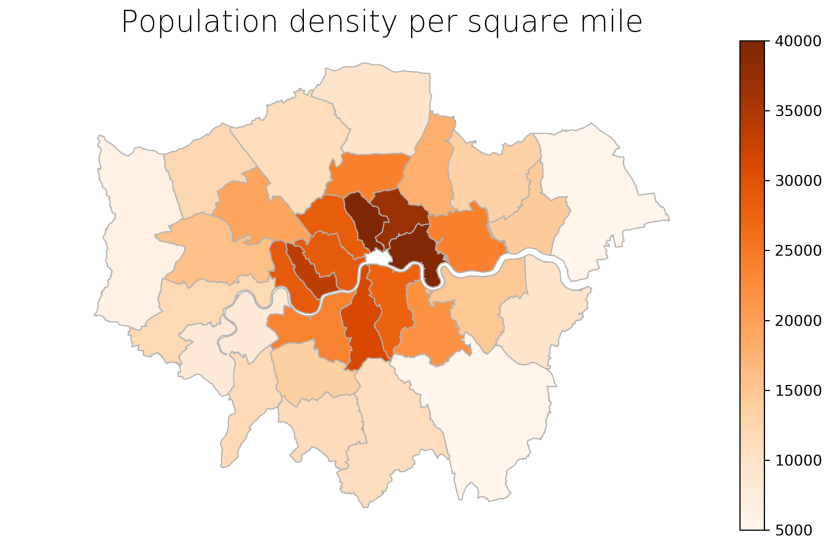
Further work could be carried out to expand this project. For instance, each borough encompasses many smaller areas that may significantly differ from one another. Analysing the features of each London postcode may allow for a better representation of the London areas. Providing such information for each postcode would allow those looking to move to London, for living or business purposes, to have a more detailed idea about suitable locations. The features used could also be further tailored towards parents and children by including more kid-friendly venues and differentiating between primary and secondary schools. Additionally, future work could investigate the use of other clustering techniques that may uncover other 'categories' that can group the boroughs/postcodes.

Appendix



Source: Greater London Authority, 2017

Figure 5.1: Total population for each borough



Source: London Datastore, 2017

Figure 5.2: Population density for each borough

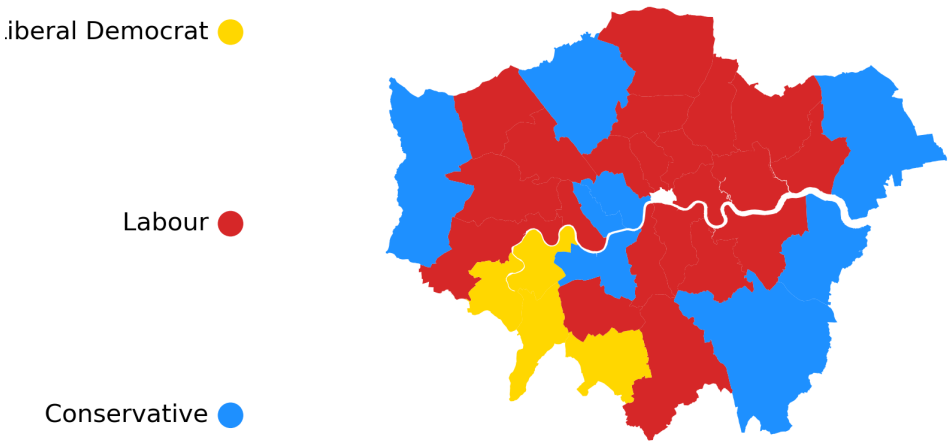


Figure 5.3: Political control of each borough

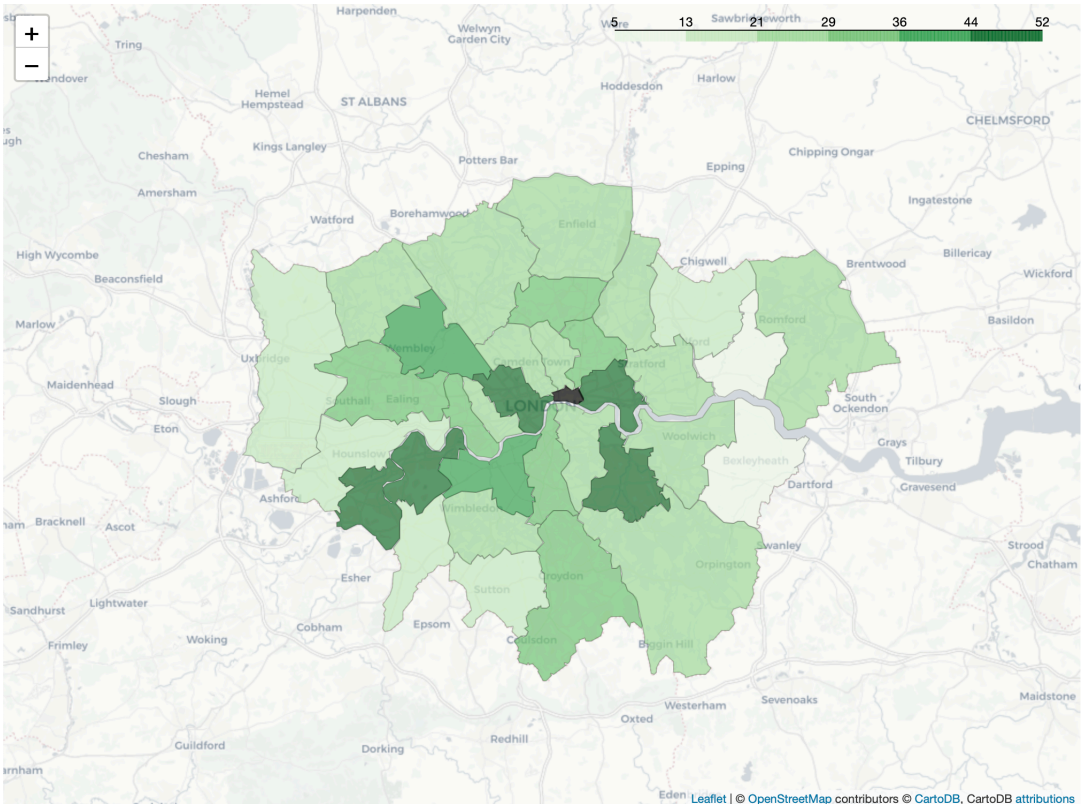


Figure 5.4: Number of parks for each borough

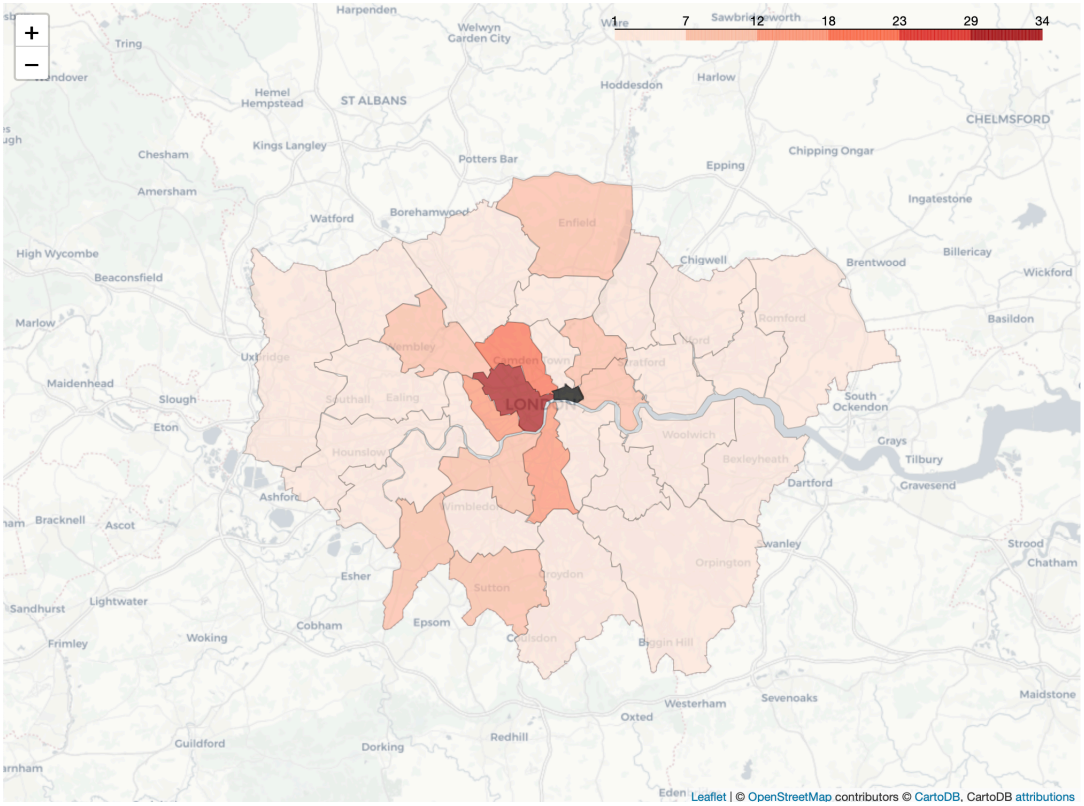


Figure 5.5: Number of hospitals for each borough

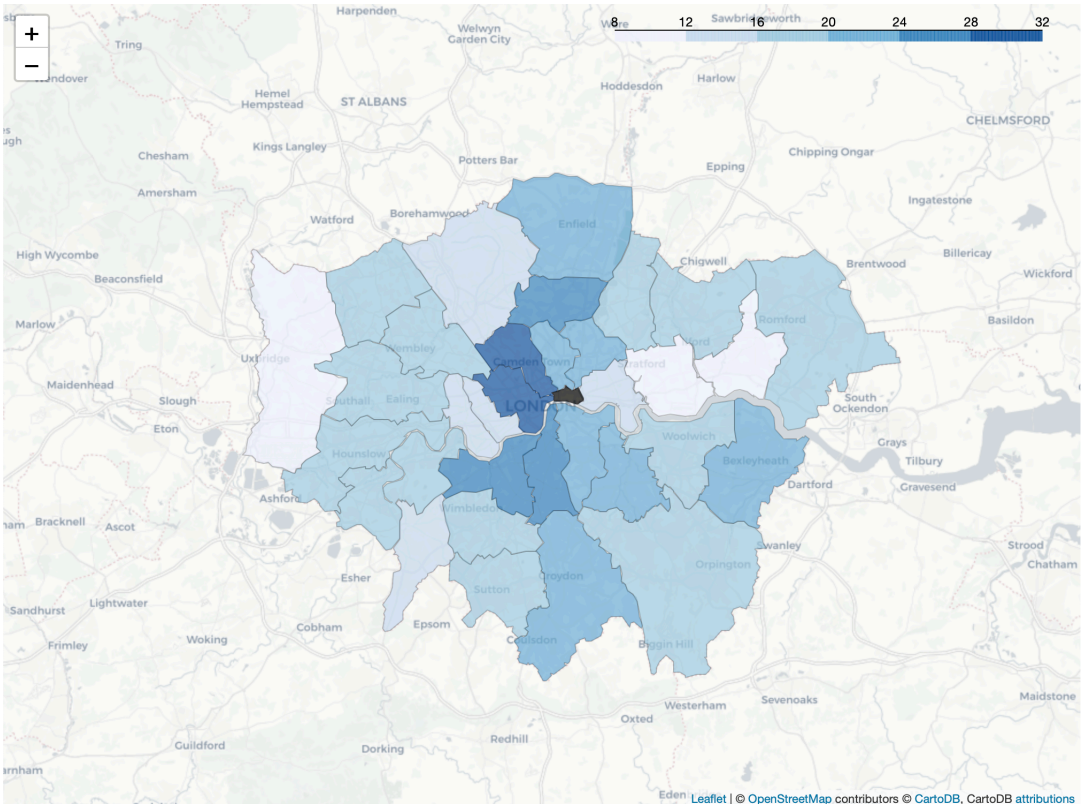


Figure 5.6: Number of supermarkets for each borough