# Comparative analysis of London Boroughs using k-means clustering

Final assignment for IBM Applied Data Science Capstone

Published: January 2021

Prepared by: Anna W.

# Table of contents

## 1. Introduction

London constitutes a home to around 9 million people[1] and is one of the top global financial centres, as well as a popular tourist destination. It is also one of the most expensive European cities to live in according to Mercer 2020 Cost of Living Survey.

On the one hand, London residents can greatly benefit from the world-class facilities such as theatres, museums and sport amenities. London also provides opportunities to study at the top universities as well as is famous for its nightlife. On the other hand, life satisfaction from living in the capital of the UK may be greatly undermined by high property and rent prices, which in recent years have been driven up because of the role London plays as a tourist destination and a financial centre and as a consequence a high demand for residential properties for investment purposes. In this regard London, has one of the worst scores in terms of finding good housing at a reasonable price.[2]

Nevertheless, according to a survey by the European Commission conducted in 2019, 93 per cent of Londoners were satisfied to live in the city, a score which was above the average for European cities. At the same time 88 per cent stated that London is a good place to live for people in general, a score slightly below the European average.[3]

Quality of life in the city has been of interest to a wide range of stakeholders. Policy makers conduct regular reviews on this matter (e.g European Commission), but it has also been subject to analysis prepared by consulting firms (e.g. CBRE Borough by Borough: London Living report).

This report will **aim** for a more in-depth view on the heterogeneity of living in London by providing a comparative analysis of its 32 boroughs. It will look at some of the indicators which are often used in surveys and reports measuring life quality in the city (such as personal well-being indicators or average income and rent).

The contribution of this report stems from the use of machine learning methods to cluster and segment London boroughs. It will also use spatial data on venues such as museums and restaurants, to explore characteristics of a chosen borough.

[1] Eurostat, as of 2019.
[2] https://urban.jrc.ec.europa.eu/thefutureofcities/affordable-housing#the-chapter
[3] https://ec.europa.eu/regional_policy/en/information/maps/quality_of_life/

This report should be of interest to policy makers as a valuable source of information when designing a strategy for urban development and public services, for instance housing policies. Hopefully, private sector companies, such as consulting businesses, will also benefit from its findings. Last but not least, it should be of interest to Londoners as well as other people considering to move within or to London.

## 2. Data

Data used as variables in the clustering model come from London datastore, available publicly at: https://data.london.gov.uk/. A general summary of data gathered and used is presented in the table below.

Table 1 Summary of variables used

| Variable | Description |
|---|---|
| green_space | Borough Green Space Surface (%) as of 2016 |
| mean_income | Personal income by tax year in GBP as of 2018 |
| median_income | Personal income by tax year in GBP as of 2018 |
| total_area | Total Area (Hectares) |
| dwellings_p_hectare | Dwellings per hectare as of 2019 |
| life_satisfaction | Mean score to a question "Overall, how satisfied are you with your life nowadays?", with a possible score from 0 to 10, as of 2018/2019 |
| worhtwile | Mean score to a question "Overall, to what extent do you feel the things you do in your life are worthwile?", with a possible score from 0 to 10, as of 2018/2019 |
| happiness | Mean score to a question "Overall, how happy did you feel yesterday?", with a possible score from 0 to 10, as of 2018/2019 |
| anxiety | Mean score to a question "Overall, how anxious did you feel yesterday?", with a possible score from 0 to 10, as of 2018/2019 |
| e_to_p | Ratio of House Prices to Earnings (full-time workers by place of work) |
| own_outright | Households owned outright (%) as of 2018 |
| buying_w._mortg. | Households bought with mortgage or loan (%) as of 2018 |
| rented | Households rented from local authority (%) as of 2018 |
| rented_from_private | Households rented from private landlord (%) as of 2018 |
| average_rent | Mean gross monthly rent paid in GBP as of 2019 Q1 ( for all categories of dwellings) |

London borough coordinates were extracted from the Wikipedia page available at: https://en.wikipedia.org/wiki/List_of_London_boroughs

Data on venues in a chosen borough were extracted using Foursquare API.

## 3. Methodology

During data preparation process, data on all London districts were gathered from various sources and put together in a dataframe (**df**). Names of the boroughs had to be standarized so that they would be exactly the same both in the dataframe and the json file with the location data.

There are 33 local authority districts in London. Many indicators for the district *City of London* were missing, therefore this area was later excluded from the dataframe. However, City of London is the only district that does not constitute a London Borough. What is more it is primarily a business rather than residential district, therefore its exclusion was considered reasonable. Its non-borough status may also explain why some data for this district are not being collected.

As part of the analysis, firstly, some visual analysis was conducted to obtain a better understanding and intuition behind the data. This was done using the choropleth maps. A graph of inter alia median income and average rent in particular boroughs gave a justification for the number of three clusters used during the later stages of analysis (i.e. segmentation using k-means method).

For spatial data analysis required by this assignment one borough was used (Hounslow), since an in-depth analysis of venues in all London boroughs would be problematic due to a vast amount of data to be downloaded with the use of free Foursquare account. The venues in the Hounslow Borough were explored in the radius of 500 from its headquaters.

Top 5 of the venues found are presented in the table below.

Table 2 Venues in the Hounslow area

| Venue no. | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| name | Hyatt Place London Heathrow Airport | Costa Coffee | Runway 09L / 27R | Leonardo Hotel | Easirent |
| categories | Hotel | Coffee Shop | Airport Service | Hotel | Rental Car Location |
| address | The Grove Bath Road | NaN | LHR Airport | NaN | NaN |
| crossStreet | Bath Rd | NaN | NaN | NaN | NaN |
| lat | 51.481709 | 51.480961 | 51.477534 | 51.481277 | 51.483394 |
| lng | -0.468062 | -0.468985 | -0.460181 | -0.457622 | -0.460111 |
| labeledLatLngs | [{'label': 'display', 'lat': 51.48170877103277... | [{'label': 'display', 'lat': 51.48096072959825... | [{'label': 'display', 'lat': 51.47753414052395... | [{'label': 'display', 'lat': 51.48127725201596... | [{'label': 'display', 'lat': 51.48339449907593... |
| distance | 316 | 373 | 463 | 414 | 350 |
| postalCode | UB7 0DG | NaN | NaN | NaN | UB7 |
| cc | GB | GB | GB | GB | GB |
| city | West Drayton | NaN | Hounslow | NaN | West Drayton |
| country | United Kingdom | United Kingdom | United Kingdom | United Kingdom | United Kingdom |
| formattedAddress | [The Grove Bath Road (Bath Rd), West Drayton, ... | [United Kingdom] | [LHR Airport, Hounslow, Greater London, United... | [United Kingdom] | [West Drayton, Greater London, UB7, United Kin... |
| state | NaN | NaN | Greater London | NaN | Greater London |
| id | 5856d8af2b04f83e93db3ab8 | 5476e91c498e323e46ccc9a8 | 56cc4100cd104e0453c632c2 | 550d5dc3498ed2931e90f431 | 591feffc6bdee630c352c020 |

In order to segment boroughs based on similar characteristics a k-means clustering method was used. K-means clustering is easy to implement and allows to arbitrarily choose the number of centroids/clusters (k). The chosen value of k is 3, which was supported by choropleth maps analysis and furthermore was considered to be analogous to a segmentation into low, medium and high income boroughs (such categories are often used for countries classification – see e.g. IMF).

Next, centroid mean values were obtained for all the variables analysed in order to have a better understanding of what differentiates each cluster. These are discussed more thoroughly in the *Results* section.

## 4. Results

Table 3 presents the results of k-means clustering. Based on the mean and medium income (see Table 4) Cluster 0 is considered to be the "low-income" cluster, Cluster 1 a "high-income" cluster, whereas cluster 2 is a "medium-income" cluster. Nevertheless the difference between mean and median income of Cluster 0 and Cluster 2 is much narrower than between Cluster 2 and Cluster 1. Considering the map (Figure 1 in Annex) the level of income seems to be correlated with how "central of London" a particular borough is. Thus Cluster 1 constitutes the very centre of the city, Cluster 0 represents outer boroughs and Cluster 2 boroughs in between.

Table 3 Members of clusters

| Cluster 0 | Cluster 1 | Cluster 2 |
|---|---|---|
| Barking and Dagenham | Camden | Hackney |
| Barnet | Westminster | Hammersmith and Fulham |
| Bexley | Kensington and Chelsea | Haringey |
| Brent | | Islington |
| Bromley | | Lambeth |
| Croydon | | Lewisham |
| Ealing | | Newham |
| Enfield | | Southwark |
| Greenwich | | Tower Hamlets |
| Harrow | | Wandsworth |
| Havering | | |
| Hillingdon | | |
| Hounslow | | |
| Kingston upon Thames | | |
| Merton | | |
| Redbridge | | |
| Richmond upon Thames | | |
| Sutton | | |
| Waltham Forest | | |

Table 4 Mean characteristics of a cluster

| Cluster | 0 (low-income) | 1 (high-income) | 2 (middle-income) |
|---|---|---|---|
| **green_space** | 39.094737 | 20.466667 | 21.09 |
| **mean_income** | 40636.84211 | 131200 | 49770 |
| **median_income** | 28089.47368 | 39133.33333 | 29850 |
| **total_area** | 6667.578321 | 1873.443633 | 2685.13805 |
| **dwellings_p_hectare** | 18.179714 | 58.727989 | 47.975812 |
| **life_satisfaction** | 7.64 | 7.456667 | 7.514 |
| **worhtwile** | 7.836842 | 7.613333 | 7.695 |
| **happiness** | 7.572105 | 7.38 | 7.44 |
| **anxiety** | 3.008421 | 3.553333 | 3.28 |
| **e_to_p** | 12.478947 | 23.02 | 14.272 |
| **own_outright** | 28.510526 | 24.366667 | 17.65 |
| **buying_w._mortg.** | 33.126316 | 12.8 | 23.57 |
| **rented** | 15.915789 | 33.733333 | 30.45 |
| **rented_from_private** | 22.457895 | 29.133333 | 28.32 |
| **average_rent** | 1358.263158 | 2822.333333 | 1708 |

## 5. Discussion

Based on the mean characteristics of clusters, Cluster 0 boroughs seem to be a good choice for first-time home owners. This is indicated by lower earnings to price ratio, lower income ratio and a high rate of properties bought with a mortgage (**buying_w._mortg.**). The last factor may hint to the fact that the property demand driven by investment motives is lower in these borrows, therefore it does not generate additional price pressures.

There are several factors that point to Cluster 1 boroughs as being the most attractive for property investment (i.e. buying residential property with the aim of gaining profit from renting it and not using it as the primary residence). Earnings to price ratio (**e_to_p**) is much higher than in the rest of the city, indicating a higher propensity and ability to invest spare income. A ratio of residential properties owned (**own_outright**) is similar to Cluster 0, but the share of

those bought with a mortgage is much lower, supporting the hypothesis of higher ability (spare disposable income) and propensity to invest.

Cluster 2 boroughs are characterised by the lowest share of properties owned (**own_outright).** These are not as far away from the centre of the city as Cluster 0 boroughs, therefore the commuting times are shorter. This is a good alternative for young and mobile workers, who have no need to settle down yet. What is more, the **average_rent** in these cluster is not much higher than in Cluster 0 (at least in comparison to the centre of London).

It is interesting to note that in terms of well-being indicators (**life_satisfaction, worthwile, happiness, anxiety**) there is not much difference between the clusters. This may point to the fact that material status (including the level of income and possessions such as property) does not have a direct impact on personal well-being. Another interpretation is that each type of cluster may offer some type of benefits that compensate for its deficiencies. For instance, cluster 0 (low-income) boroughs are less densely built-up **(dwellings_p_hectare)** and have a higher share of green space **(green_space)**, while cluster 1 (high-income) boroughs may appeal to aiming for high material status.

## 6. Conclusion.

To conclude, London boroughs can be characterised by a significant degree of heterogeneity. In this report, three borough clusters were distinguished, with each one of them having specific characteristics that may appeal to potential and current residents. Cluster 0 boroughs are characterised by relatively low rent and higher share of green space. Cluster 1 features point to its attractiveness in terms of property investment. In turn, young and mobile workers should benefit from medium rent prices and location of Cluster 2 boroughs. In terms of well-being of its residents, London boroughs were found to be very homogenous.

**Annex**

Figure 1 London borough maps


Source: Londonist.com