Looking for accommodation in London

Capstone Project for IBM Data Science Professional Certificate

David Andrés

Introduction

The business problem that this report will try to solve is: Which is the most recommended location for living if you want to study at Middlesex University in London?

Some considerations have to be taken into account:

- The accommodation needs to be close to the university, but not too far from the city centre.
- Supermarkets and cafes in the neighbourhood will also be taken into consideration.
- Being a student the rent is a key factor to bear in mind. A one bedroom house or flat is preferred.

London is a large city with plenty of opportunities. There are many interesting spots in this city, touristic attractions, monuments and even business areas. All these factors are important when looking for accommodation. Rental prices in London are high, the area must be chosen in a way that minimises it while still enjoying the proximity to the key locations mentioned before.

The target audience is university students (in particular studying in Middlesex University) who are looking for accommodation in London, and don't want to give up on all the beauties of London.

Data

The different boroughs of London will be checked, searching for venues of interest, average rental prices and distances to university and city centre.

The data sources used were:

- Foursquare
- https://data.london.gov.uk/dataset/average-private-rents-borough
- https://www.freemaptools.com/download/outcode-postcodes/postcode-outcodes.csv
- https://www.doogal.co.uk/PostcodeDistrictsCSV.ashx

The accommodation must be located close to Middlesex University. Foursquare data will be used to search for venues around this location. The postcode directory of the city will be used to group the venues information and average rent per area by borough.

Methodology

The first step was to select the possible boroughs of London. The postcodes data were imported and cleaned. Only the boroughs close to the university and to city centre would be considered. Therefore the data were filtered by the following borough codes: NW, N, W and E (see Figure 1).

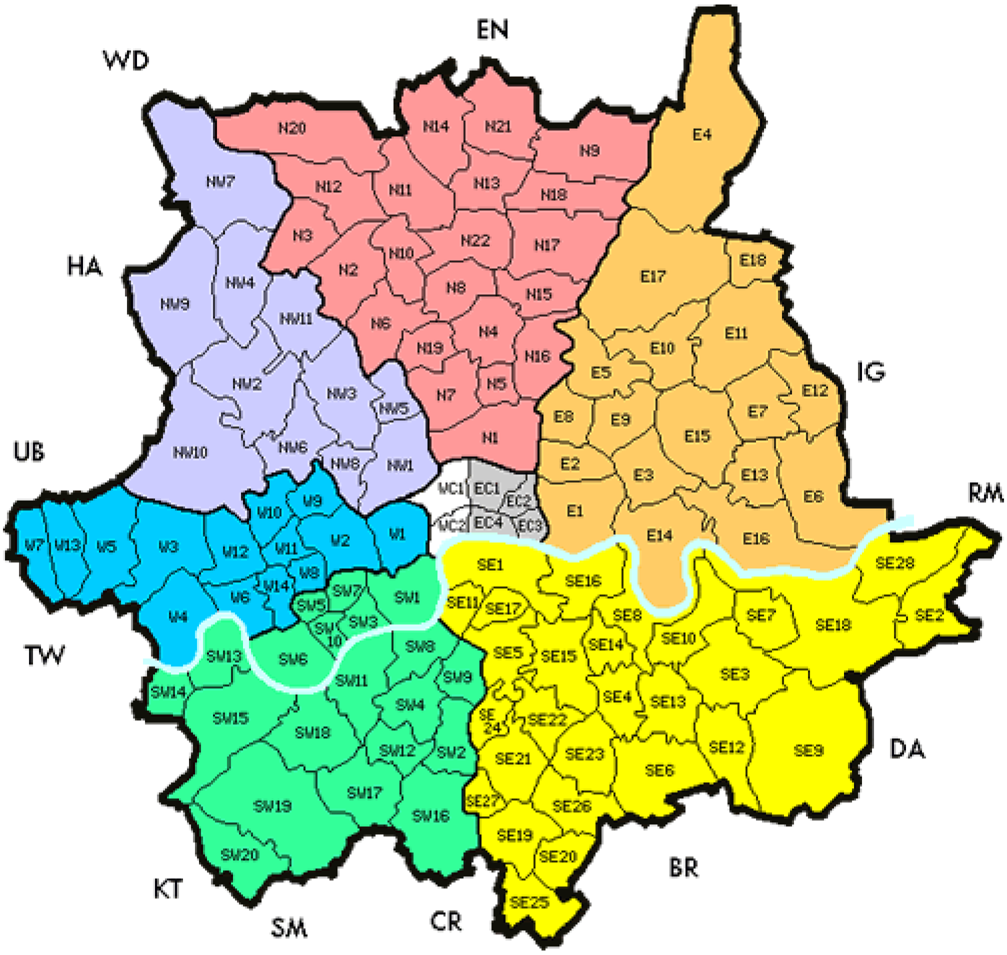


Figure 1 Greater London map

The neighbourhoods considered are shown in the map below (Figure 2).

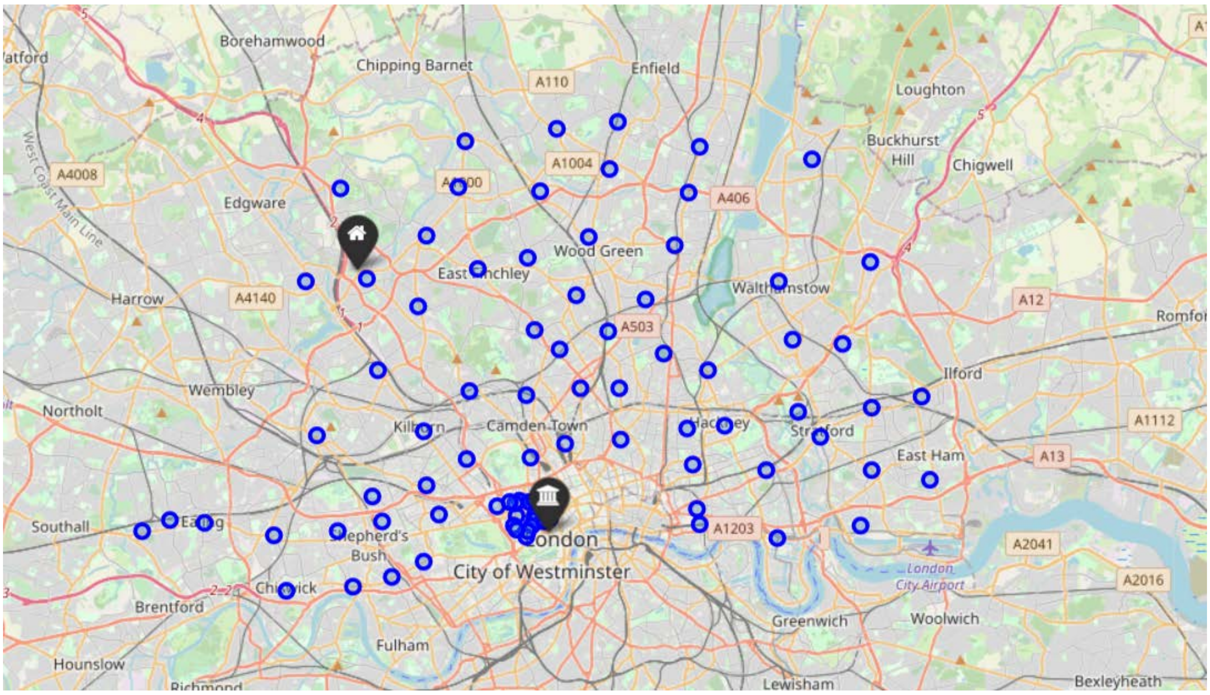


Figure 2 Selected neighbourhoods

The next step was to import the data about rents. It was filtered by the last year (2019) and one bedroom house/flat to simplify the analysis. The data were then grouped by postcode and the rents were aggregated by mean. Only average rent was considered. The following histogram shows the rent by neighbourhood (Figure 3).

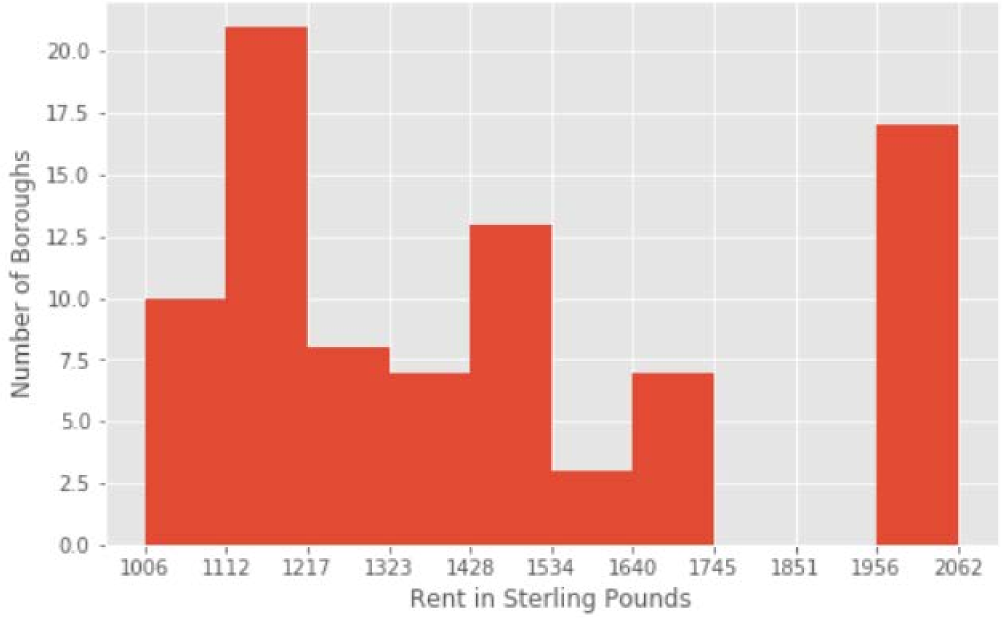


Figure 3 Histogram of rent by borough

After making sure the region fields in both tables were under the same capitalising rules, both tables were merged, having now information about the postcodes and rent, in addition to latitudes and longitudes.

Distances between each neighbourhood and university and city centre were calculated. The city centre location was assumed the same as Piccadilly Circus. Also total distance was calculated as the sum of both distances. The following histogram shows the total distance by borough (Figure 4).

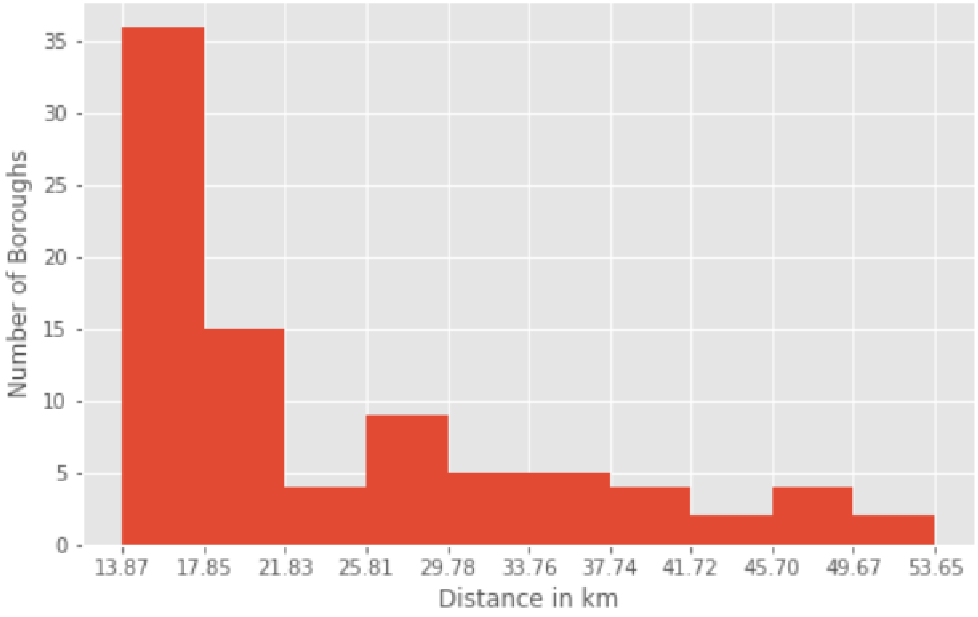


Figure 4 Histogram of total distance to university and city centre by borough

The Foursquare API was then used to extract the cafes and supermarkets in a region of 20km around the university. This information was grouped by postcode and merged to the previous table, aggregating by number of cafes and number of supermarkets. Figure 5 shows the map displaying the previous information.

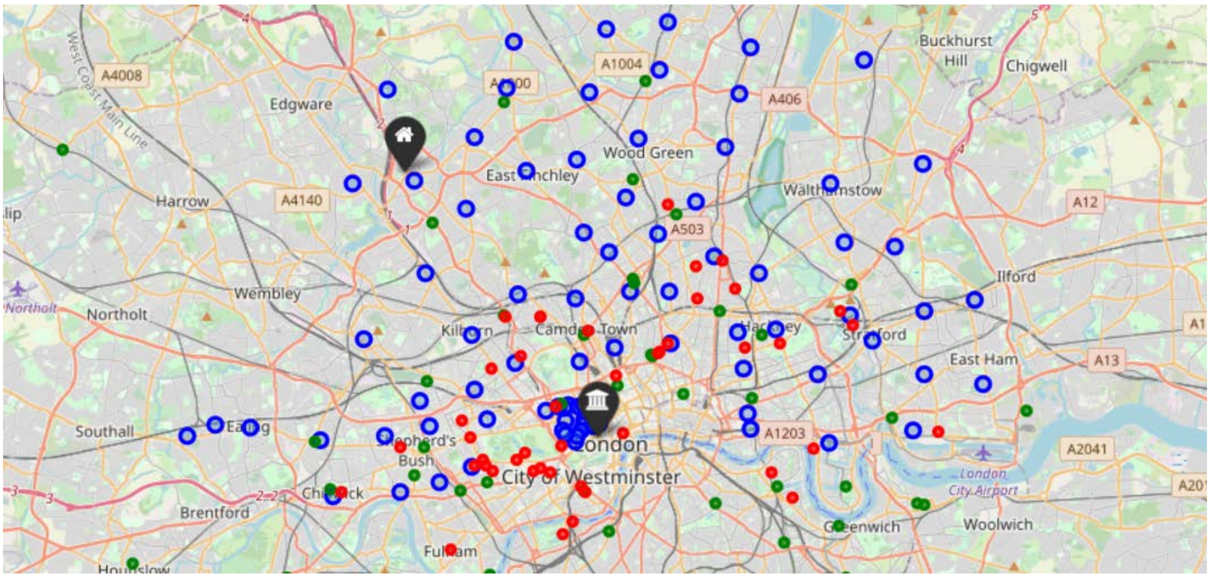


Figure 5 Selected neighbourhoods (blue) with cafes (red) and supermarkets (green)

With all this information collected and gathered in a table, a machine model was utilised to cluster the neighbourhoods in areas. A K-Means algorithms was selected due to its simplicity and good results. Only the following features were kept for the analysis: Rent, Distance to University, Total Distance, Number of Cafes and Number of Supermarkets. A standard scaler was used to avoid the distance governance of large features. Finally the “Elbow Criterion Method” was used to determine the most suitable parameter K (Figure 6).

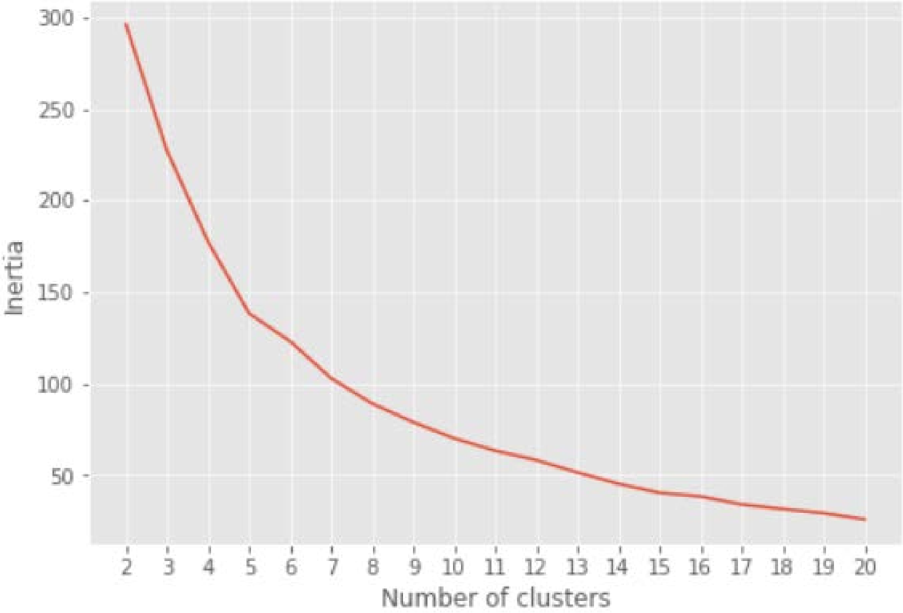


Figure 6 Elbow Criterion Method for selection of K

This analysis suggested the use of 5 or 6 clusters. Therefore the parameter K was chosen to be 6.

Results

The K-Means model yielded the following clusters (Figure 7). Data were then grouped by label and the features aggregated by mean in order to determine the characteristics of each cluster (Table 1).

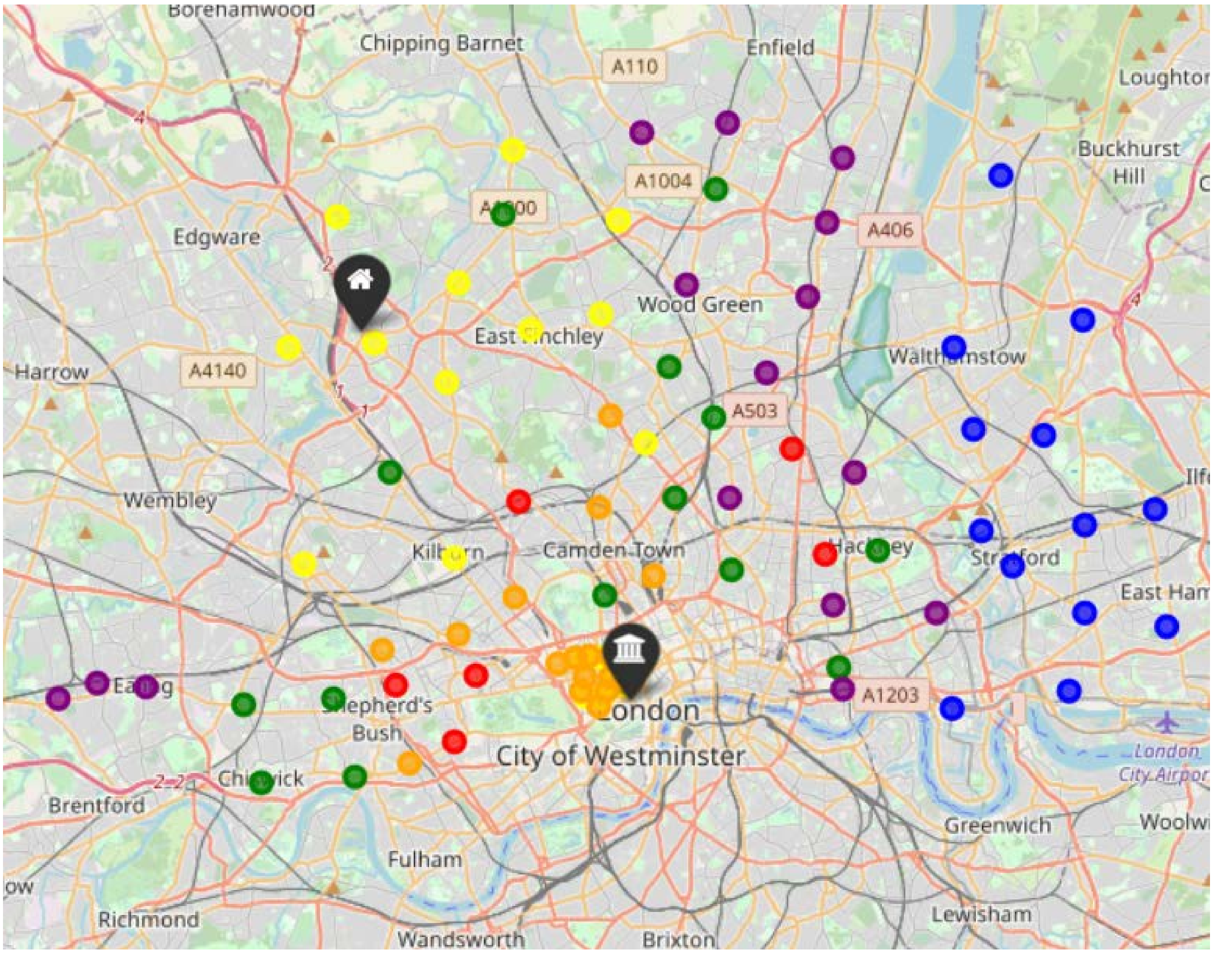
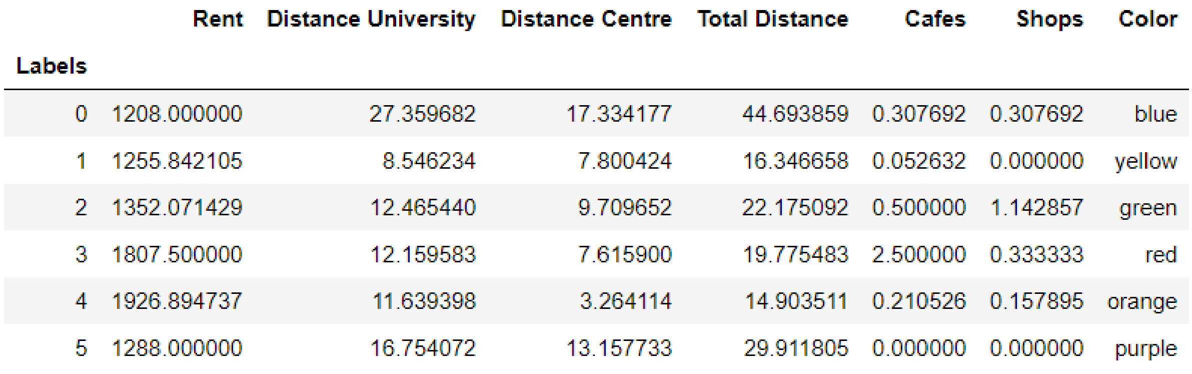


Figure 7 Clustered neighbourhoods



	Rent	Distance University	Distance Centre	Total Distance	Cafes	Shops	Color
Labels							
0	1208.000000	27.359682	17.334177	44.693859	0.307692	0.307692	blue
1	1255.842105	8.546234	7.800424	16.346658	0.052632	0.000000	yellow
2	1352.071429	12.465440	9.709652	22.175092	0.500000	1.142857	green
3	1807.500000	12.159583	7.615900	19.775483	2.500000	0.333333	red
4	1926.894737	11.639398	3.264114	14.903511	0.210526	0.157895	orange
5	1288.000000	16.754072	13.157733	29.911805	0.000000	0.000000	purple

Table 1 Data grouped by clusters aggregating features by mean

Discussion

The different clusters can be interpreted as follows:

- Label 0 (blue): lowest rent, but far from university and city centre. Also a medium number of cafes and shops.
- Label 1 (yellow): second lowest rent, shortest distance to university and short distance to city centre. No supermarkets in the area and very low number of cafes.
- Label 2 (green): affordable rents, medium distance to university and city centre. Good number of cafes and supermarkets.
- Label 3 (red): expensive area, medium distance to university but close to city centre. A great number of cafes and medium number of supermarkets.
- Label 4 (orange): very expensive area, located in city centre, medium distance to university. A medium number of cafes and medium-low number of supermarkets.
- Label 5 (purple): cheap rents, far from both university and city centre. No cafes or shops.

In order to make a more conclusive analysis, more variables should be taken into consideration. For example, take-away venues may be of importance to the user, or small food stores. However, for the sake of simplicity only cafes and supermarkets have been considered in this study. Also, the distance to the closest metro station would be relevant to be added, since it is key in a city like London.

Conclusion

The final choice has to be made by the person who is going to live and study in London. However, based on the analysis, the most suitable location would be either regions in yellow (1) or in blue (green), depending on the preference of the user. If the main priority is being close to university, then the user should choose the yellow regions (1). If on the contrary the number of cafes and supermarkets is relevant, the user should go for the green regions (2). This conclusion is assuming the rent is a key factor to take into account due to the student condition.