A clustering approach to understanding business popularity in London.

Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods for the IBM Data Science Professional Certification.

Temitope Roland Alademehin

Table of Contents

1. Introduction: Business Problem

London is one of the world's leading tourism destinations, and the city is home to an array of famous tourist attractions. This city is widely known for its vast selection of wonderful places, not only on the weekends but also during the weekdays. Even with hundreds of such venues all across the city, there are abundant opportunities for investors to choose from. Understanding the attributes of these popular places might give investors an idea of some business opportunities in a certain area of London.

In other to select the best business opportunity, optimal location coupled with the understanding of the environment is required. Hence, this research employed the used a clustering method in categorising all the boroughs in London and going further in analysing top common places in each cluster based on their popularity, hence giving investors a glance at the wide variety of top business selections for different areas and less popular ones with possibilities of doing well in all the areas. Also, tourist can use this analysis in understanding areas to visit in London based on their interest.

2. Data

Based on the definition of our research problem, we have scraped and filtered the London Borough's data from Wikipedia [1], the extracted columns are the names, population and land area of all London's borough. Also, the package Google Maps API reverse geocoding was used in matching both latitude and longitude of each borough to the data and lastly the Foursquare API was employed in obtaining top common venues for all the boroughs.

3. Methodology

Note: head denotes the first five rows of a dataframe.

3.1. Dataframe

Initial data extraction from Wikipedia page was done using the BeautifulSoup and lxlm packages in python and the resulting data head is shown in Fig 1. Only three columns were extracted from the resulting table - the Borough, Area and Population which was filter into a new dataframe.

	Borough	Inner	Status	Local authority	Political control	Headquarters	Area (sq mi)	Population (2019 est)[1]
1	Barnet	None	None	Barnet London Borough Council	Conservative	Barnet House, 2 Bristol Avenue, Colindale	33.49	395896
2	Bexley	None	None	Bexley London Borough Council	Conservative	Civic Offices, 2 Watling Street	23.38	248287
3	Brent	None	None	Brent London Borough Council	Labour	Brent Civic Centre, Engineers Way	16.70	329771
4	Bromley	None	None	Bromley London Borough Council	Conservative	Civic Centre, Stockwell Close	57.97	332336
5	Camden	None	None	Camden London Borough Council	Labour	Camden Town Hall, Judd Street	8.40	270029

Fig 1. Extracted and filtered data head from the Wikipedia page.

In matching latitude and longitude of corresponding boroughs, the geocoder API was used. The resulting dataframe head is shown in Fig 2. Fig 3 presents the merging of the two data frames and an addition of a new column, population density which was calculated using below.

$$Population\ density = \frac{\text{Population}}{\text{Land Area}}$$

	Borough	Latitude	Longitude
0	(Chipping Barnet, London, Greater London, Engl...	51.65309	-0.2002261
1	(Bexley, London Borough of Bexley, London, Gre...	51.4416793	0.150488
2	(The Brent, Dartford, Kent, South East, Englan...	51.4420262	0.2315227
3	(Bromley, London, Greater London, England, BR1...	51.4028046	0.0148142
4	(Camden Town, London, Greater London, England,...	51.5423045	-0.1395604

Fig 2. Data head extracted for the coordination of the boroughs

	Borough	Area	Population	Density	Latitude	Longitude
1	Barnet	33.49	395896	11821.319797	51.653090	-0.200226
2	Bexley	23.38	248287	10619.632164	51.441679	0.150488
3	Brent	16.70	329771	19746.766467	51.442026	0.231523
4	Bromley	57.97	332336	5732.896326	51.402805	0.014814
5	Camden	8.40	270029	32146.309524	51.542305	-0.139560

Figure 3. The merged data of both Wikipedia and coordination data.

The resulting coordination (latitude and longitude) of each borough was superimposed on the London's map as seen in Fig 4. This was done by using the geocoder API to obtain the latitude and longitude of London and plotting the map using Folium package in python and thereafter adding a circular marker of each borough's coordination on the map. This visualisation shows that the boroughs are well located on the map with an acceptable approximation.

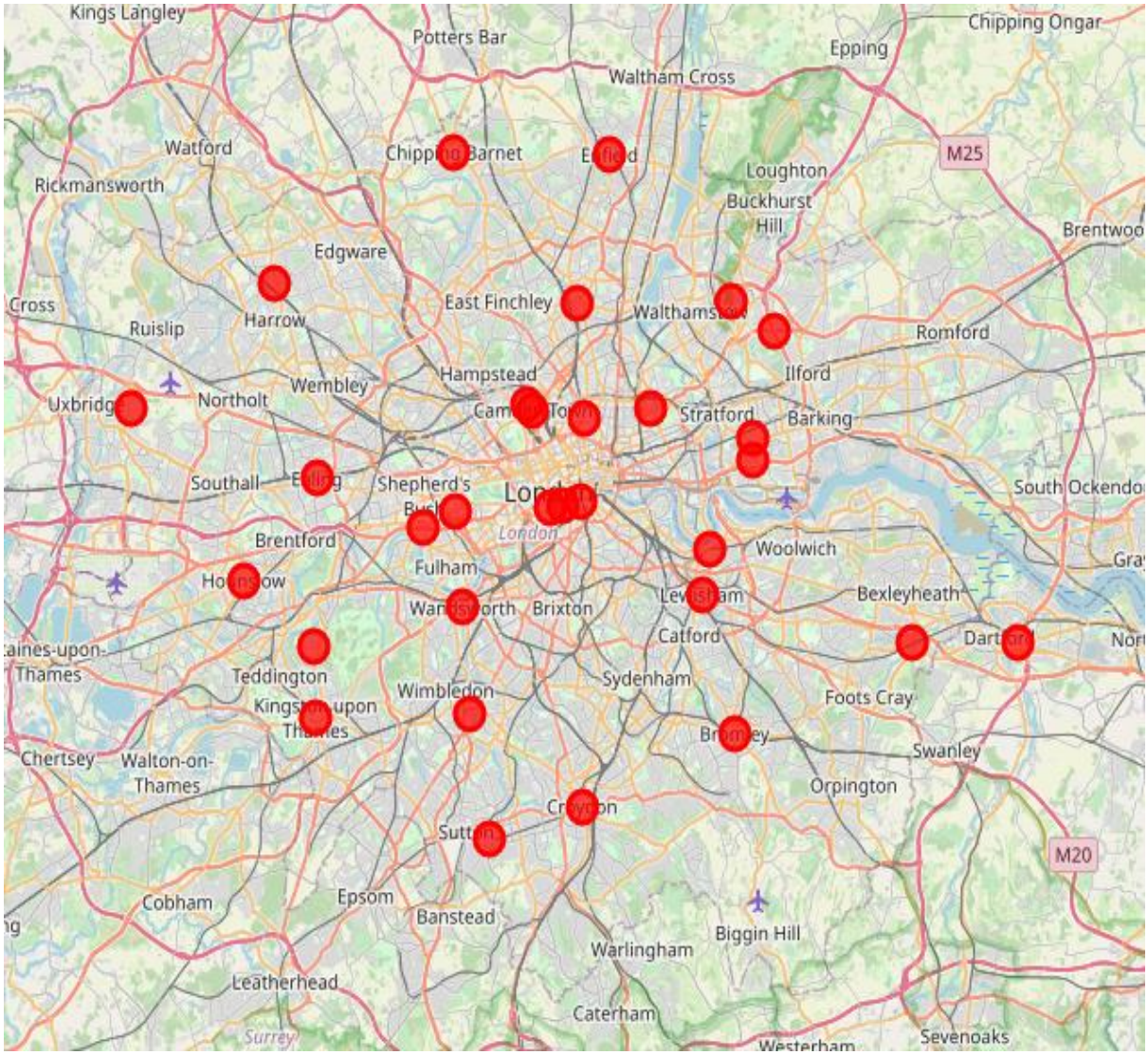

Fig 5. All the boroughs superimposed on London's map.

3.2. Venues extraction

In this section the top venues for all the boroughs were extracted using the Foursquare API, the boroughs from the dataframe in the previous section were fed into the API and the resulting data is a (3100, 7) dataframe presented in Fig 6 denoting that the API returned 100 venues categories for each of the boroughs.

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barnet	51.65309	-0.200226	Ye Old Mitre Inne	51.652940	-0.199507	Pub
1	Barnet	51.65309	-0.200226	Everyman Cinema	51.646793	-0.187675	Movie Theater
2	Barnet	51.65309	-0.200226	The Black Horse	51.653075	-0.206719	Pub
3	Barnet	51.65309	-0.200226	Joie de Vie	51.653659	-0.201288	Bakery
4	Barnet	51.65309	-0.200226	Caffè Nero	51.654861	-0.201743	Coffee Shop

Fig 6. The resulting data head from the foursquare API (3100, 7).

The dataframe in Fig 6 was processed to allow for easy usage, these stages involve one-hot encoding and moving the venues categories to the first column of the dataframe and finally grouping all the venues categories by frequency of occurrence in each borough. The resulting data is shown in Fig 7, with five most common venues but 10 most common places were used in the data analysis to allow for extended robustness.

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Barnet	Pub	Coffee Shop	Café	Park	Supermarket
1	Bexley	Pub	Grocery Store	Coffee Shop	Supermarket	Pharmacy
2	Brent	Clothing Store	Coffee Shop	Pub	Grocery Store	Pharmacy
3	Bromley	Pub	Coffee Shop	Grocery Store	Park	Pizza Place
4	Camden	Hotel	Coffee Shop	Park	Pub	Market

Fig 7. Processed dataframe for the top venues of all the boroughs.

4. Result and Discussion

4.1. Clustering

In this project, the K-means algorithm was used in categorising the boroughs into different clusters. A cluster refers to a collection of data points aggregated together because of certain similarities. Therefore, this algorithm will segment boroughs based on certain detected features in the data. This algorithm required the initialisation of n-number of clusters, several numbers were tested but n = 3 was chosen as it gives a unique clustering when visualised.

Figure 8 presents the resulting clustering obtained using K-means, an interesting segmentation was observed. The boroughs in central London were clustered together – coloured green, and the boroughs to the left (west) and right(east) were also clustered together – coloured red. Similarly, the boroughs to the top (North) and bottom (South) were clustered together – coloured purple.

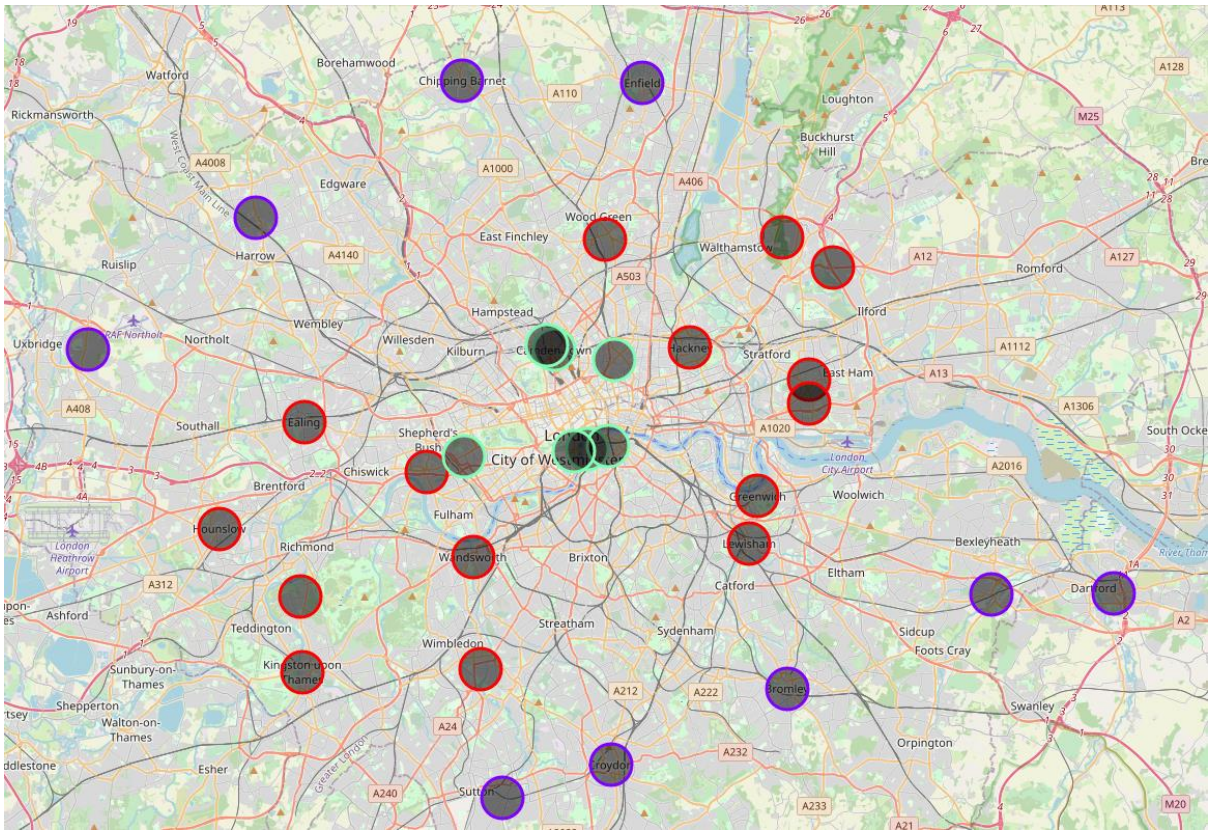

Fig 8. The resulting clusters for all the boroughs in London using K-means (n = 3).

4.2. Population density

In understanding the relationship of clusters in Figure 8, the population density is plotted as seen in Figure 9. This shows that the cluster 1 (left and right of central - red) boroughs are majorly in the mid bound of the population density, the second cluster (top and bottom of central – purple) with population density at the lower band, while the third cluster (central - green) are mostly in the top bound of the population density. This observation might be as a result of central London having lots of economic activities and people tends to live closer to such areas. This such shows a correlation between the population density and the clustering achieved by K-means as it was trained using the data obtained from foursquare and not Wikipedia data.

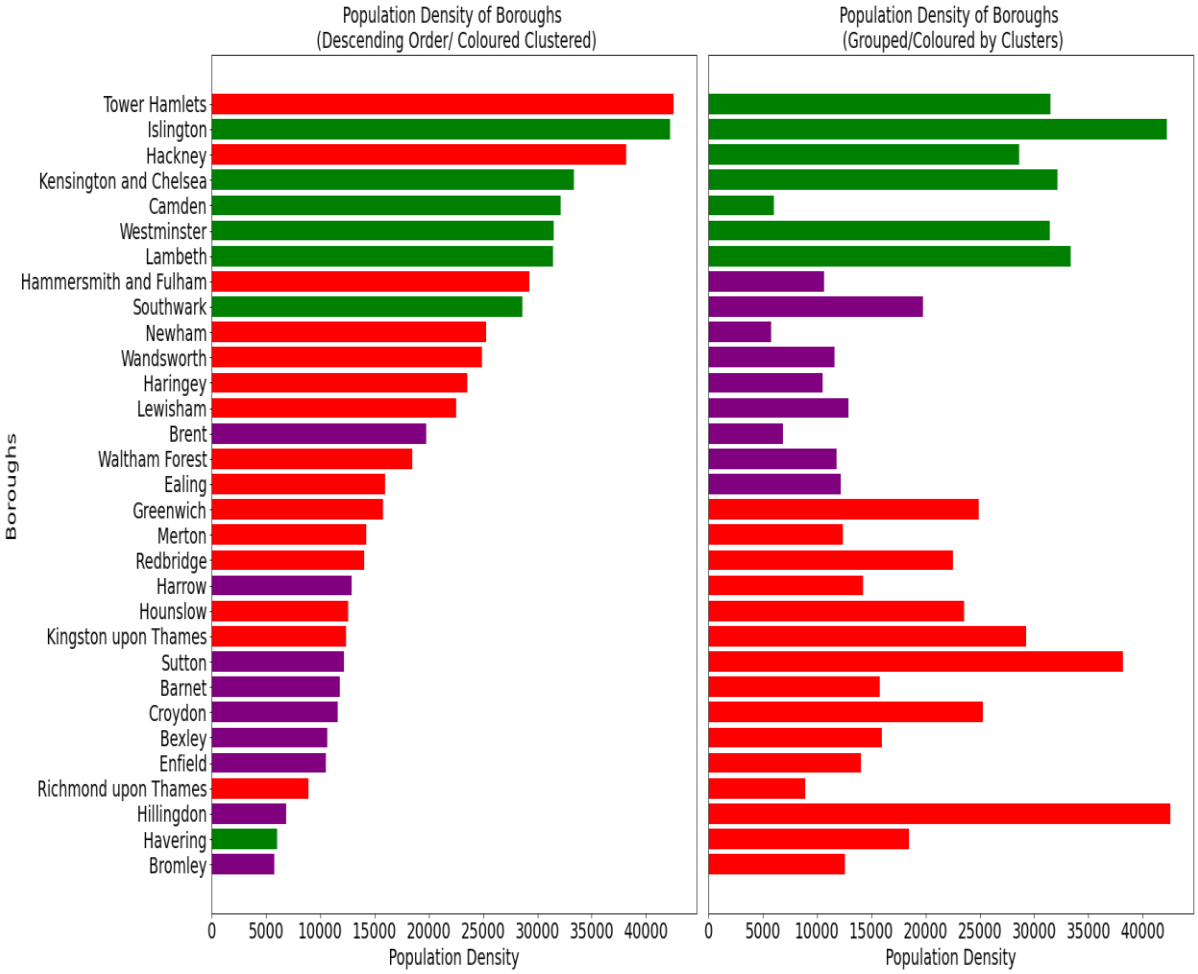

Fig 9. The population density of all the boroughs coloured based on each clustering.

4.3. Venues analysis

The three clusters have some unique attributes relating to the venue categories obtained from the foursquare API. In understanding these unique attributes, we used a visualization method called WordCloud. WordCloud is an image composed of words, in which each word sizes represent its frequency or importance. In this research, all the venues for each cluster were extracted from the dataframe and plotted on the WordCloud using their frequencies, such that the font size of the words is a function of their frequencies i.e. The higher the frequency, the bigger the word and vice versa.

Figure 10 presents the result for cluster 1, it shows that the most population density of the boroughs falls within the mid-range. The frequency of venues is majorly café, pub and park with a vast distribution of fewer others such as restaurants of different kinds, beer bar and gym. These less distributed venues provide a potential opportunity which will require further analyses.

Figure 11 presents the result for cluster 2, the most population density of the boroughs falls within the lower band. Also, Pub is the major common venue at this cluster followed closely by the coffee shop, restaurant, grocery store and supermarket. Less common are bar, department store and bakery. And lastly, in the last clustering, hotel, park, bakery and coffee

shops are the most common while the least in these zones are the gym, bar and restaurants. The observed result might be due to socio-economic factors such as house prices, population distribution, activities in the zones and income distribution. For instance, property rents are usually high in central London resulting in less occurrence of small businesses such as bars, pubs and gym.

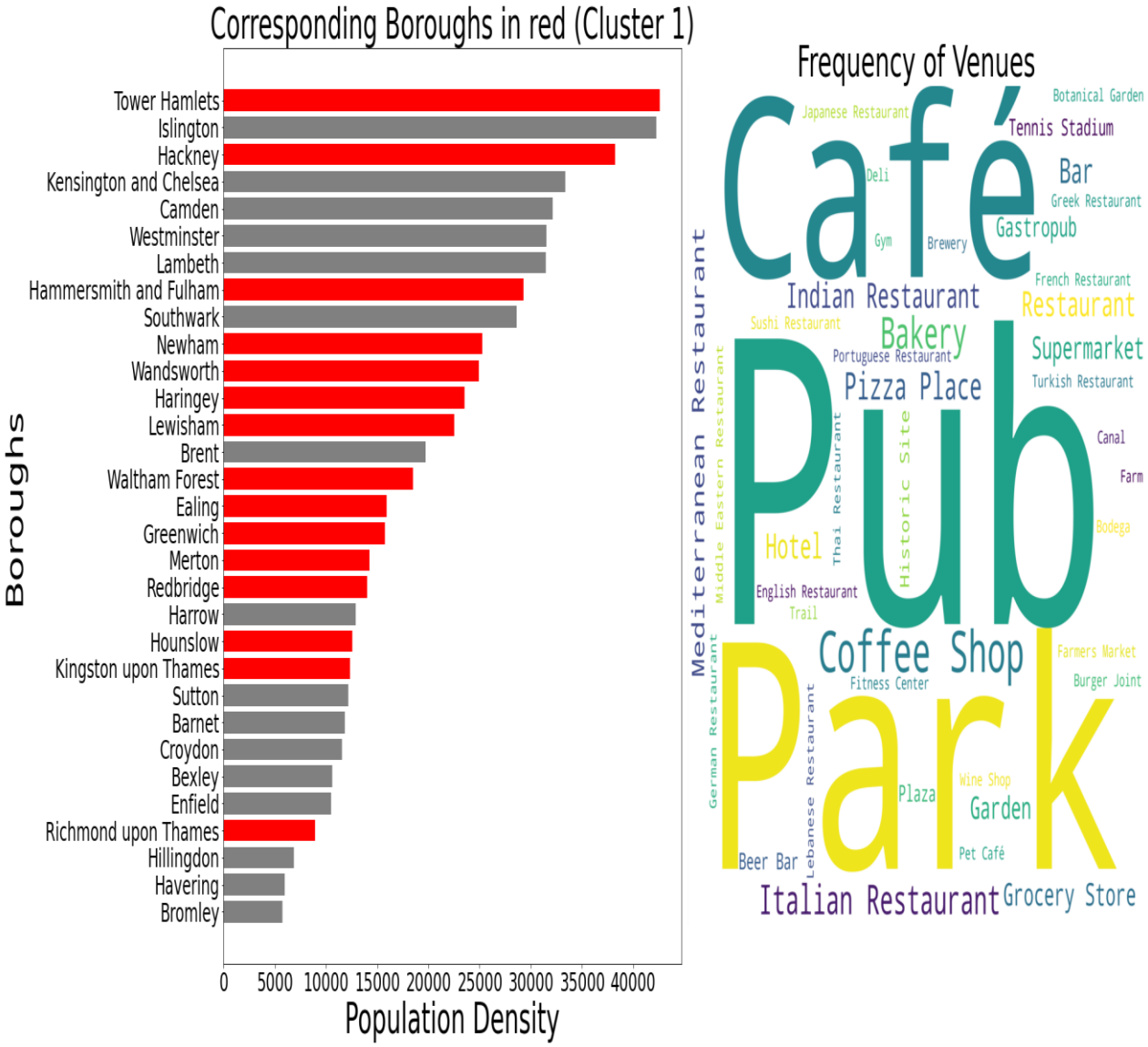

Fig 10. Venue frequency with corresponding boroughs in red for the first cluster.

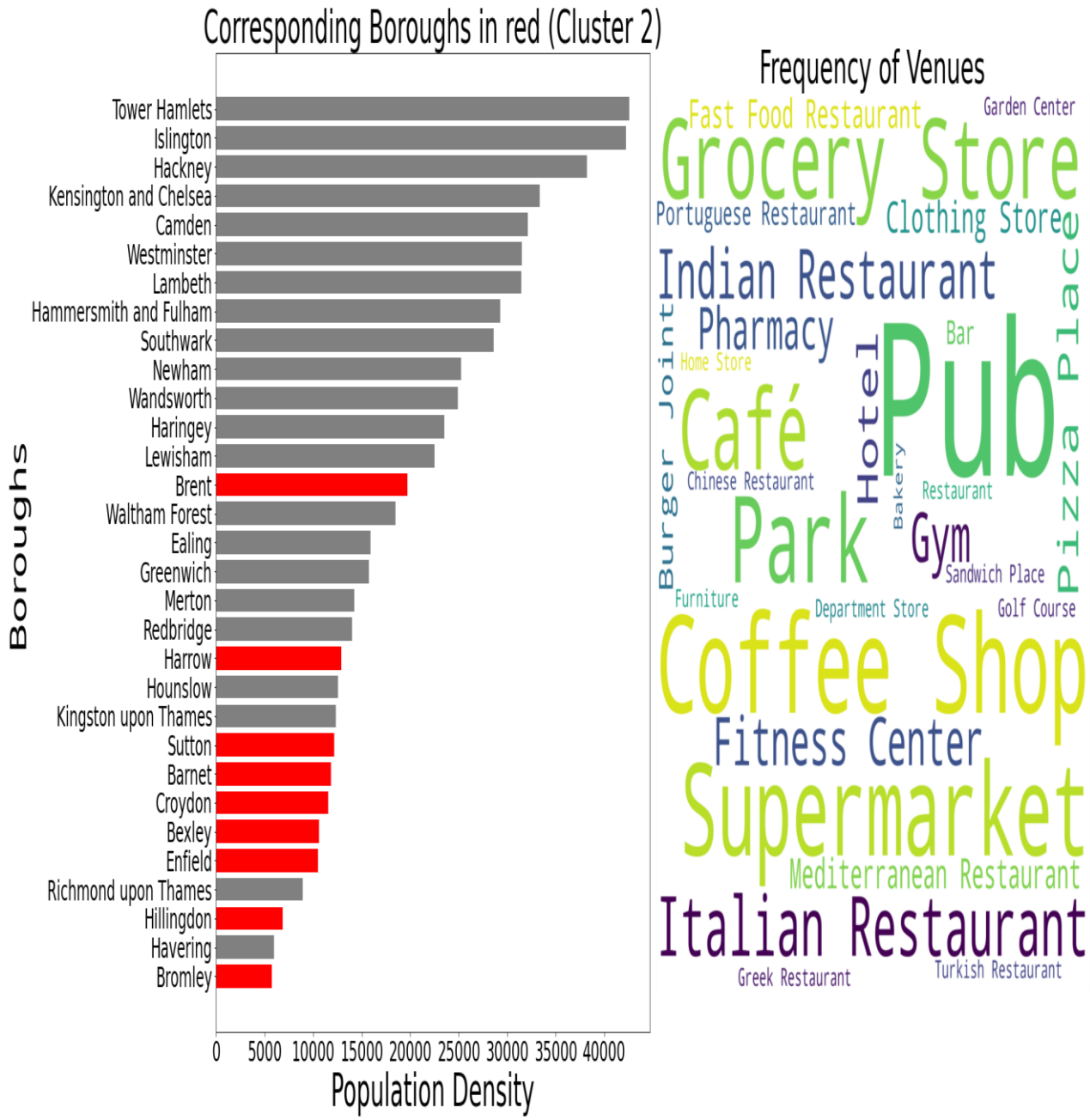

Fig 11. Venue frequency with corresponding boroughs in red for the second cluster.

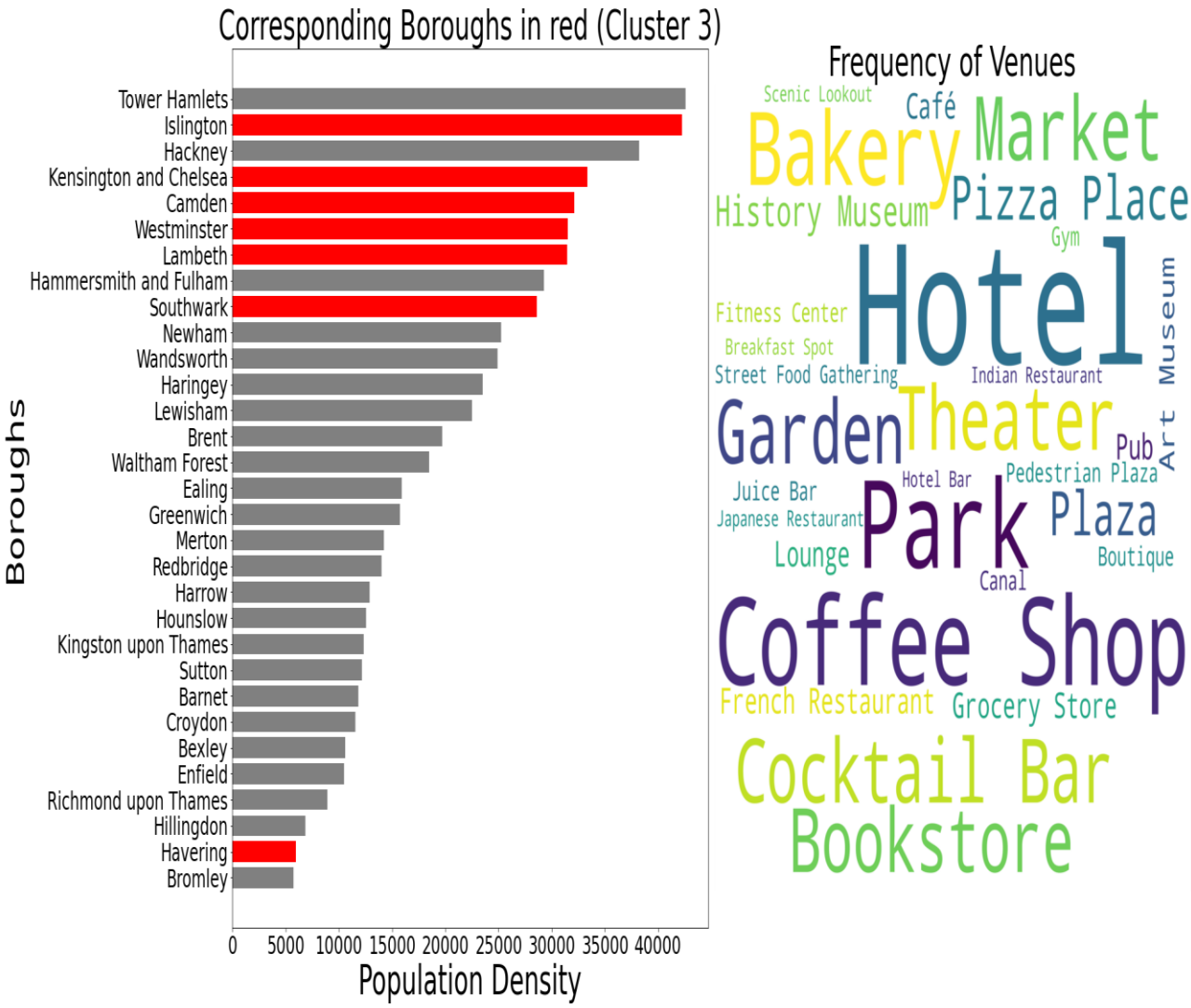

Fig 12. Venue frequency with corresponding boroughs in red for the third cluster.

5. Conclusion

This research has explored common venues and businesses in different clusters of London boroughs, achieved using the k-means clustering of all the boroughs in London. Three clusters were used, resulting in the segmentation of boroughs into three different groups with unique similarities. It was observed that the groups have a relationship with the population density of the region i.e. each of the three clusters mostly belong to low, mid or high population density. Using the data obtained from the foursquare API of the popular place, it was also observed that the popularity of venues in each cluster was uniquely matching to the characteristics observed in the region. This will give both investor and tourists a glance of business opportunities and visiting places for different zones in London. Even though this research has suggested some business idea, it is limited as an optimal location with a very good approximation is needed. Therefore, further research is needed to investigation exact location of setting the suggested businesses in each zone.

Reference

[1] https://en.wikipedia.org/wiki/List_of_London_boroughs