IBM Applied Data Science Capstone Project, 13.09.2019

Introduction

London has attracted peoples from all corners of the world due to its current status as one of the world’s financial capitals and its previous status as the centre of a large international empire. Anybody who has spent some time in different parts of London will find that different immigrant communities have tended to clump together in different parts of the city, giving many of the neighbourhoods a unique cultural vibe.

Therefore, when opening for example a supermarket or restaurant in a particular neighbourhood, tailoring products to local demographics could be highly beneficial to business performance. The results of this study are therefore of interest to any number of businesses which have a product to sell: this could be retail products in supermarkets perhaps targeted towards a Middle-Eastern community, or food products sold in a restaurant made to be more appealing to an Indian community, or even travel agency products tailored towards a Chinese community.

This project proposes the following question: can we use location data to gain an understanding of local demographics in different areas of Greater London? There have of course been very detailed studies of local demographics in London in the past (i.e. the census); however, such studies require far greater amounts of time and investment than what the current project proposes. Furthermore, the existence of the census provides a good measure for model verification. In addition, the methods proposed may be extendable to another large urban area for which such detailed data may not be present, or data is not up to date.

Data

Location data for Greater London is readily available on many webpages: in this project the data from the following webpage is downloaded in the .csv format: <https://www.doogal.co.uk/Counties.php?county=E11000009>. Doogal.co.uk contains a UK wide repository of postcodes and their longitudes and latitudes, which is ideal for use in location APIs such as Foursquare. In the case of London, additional information such as ‘boroughs’ and ‘wards’ are also provided for each postcode. Greater London can be divided into 32 ‘boroughs’ or local government districts which are further divided into electoral ‘wards’ or electoral areas. In this project, London neighbourhoods will be classified as different electoral wards.

Methodology

Data manipulation was carried out using the Pandas library in Python 3. The data was first filtered down to a table containing Greater London postcodes, latitudes and longitudes for each postcode, and the associated London ward. The data was then grouped by ward, and location coordinates for each ward were obtained by averaging over the coordinates of all postcodes belonging to it. The calculated ward locations can be visualized in Figure 1.

Next, the coordinates for each ward were used for ‘explore’ calls using the Foursquare API. No search query was applied, instead the categorical term ‘food’ was applied and the top 50 venues within 500 m of each location were determined. Here the assumption was made that local restaurants cater strongly to local demographics. The first 5 rows of the resulting dataset are shown in Table 1 below.



**Figure 1:** Map of Greater London with blue pins positioned at the locations of different electoral wards. The positions have been derived by averaging over the coordinates of every postcode which belongs to a particular ward.



**Table 1:** First five rows of dataset created from using the Foursquare API to search for the top 50 food venues within 500 m of each London ward.

The following step consisted of first one-hot encoding the ‘Venue Category’ column of the dataset. In the process, all new columns without the term ‘restaurant’ in their name were dropped after inspection of the data set: it was found that in almost all such cases the venue category does not indicate the origin of the food served at the establishment. Next, any remaining venue categories where the geographic origin of the food served at the restaurant cannot be easily determined such as ‘Fast-food Restaurant’ were also dropped. The additional categories dropped are detailed in the attached jupyter notebook. The first 5 rows and columns of the resulting dataset are shown in Table 2.

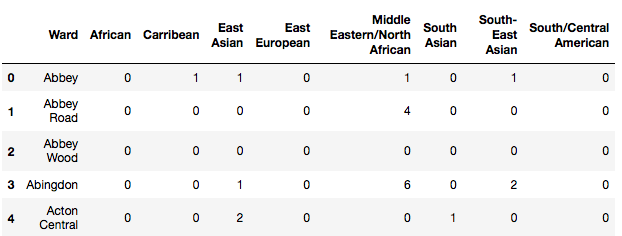


**Table 2:** Final result of one-hot encoding Table 1 (first 5 rows and columns are shown). For each London ward, a ‘1’ or greater is displayed in the corresponding restaurant column when such venues are present within the ward.

The 90 remaining categories were classified into the 9 following categories based on general cultural similarity: West European/North American; East European; Middle Eastern/North African; South/Central American; East Asian; South Asian; South-East Asian; African; Caribbean. The exact details of the classification process can be found in the attached jupyter notebook. The process of course introduces some error either through unintentional misclassification or loss of detail. On the other hand, this greatly simplifies the interpretation process, and from a retail or restaurant point of view, many of the most popular products available in these 9 regions are broadly similar, albeit with local variations. Possible examples could be the types of spices used in South Asian cuisine, or pop music in East Asia.

Finally, initial further exploratory analysis resulted in the West European/North American establishments generally outnumbering the others. This is to be expected since the UK itself belongs to this category: the decision was therefore made to drop this category from the dataset and focus on the others. The first 5 rows of the dataset at this stage are shown in Table 3.

The Greater London wards were then clustered based on their associated categories. The partition-based k-means clustering method was employed as implemented in the Scikit-learn package due to the ease in determining input parameters. The ‘elbow’ method was employed to limit the number of clusters to 5 as described in the attached jupyter notebook. In contrast, hierarchical clustering is not intuitive for such a problem, and density-based clustering, for example through the DBSCAN algorithm, requires input parameters which cannot be well determined: in particular the ‘eps’ parameter.



**Table 3:** Table 2 after reclassifying the different categories into 9 more general groups and removing the ‘Western Europe/North American’ category. Only the first 5 rows are shown.

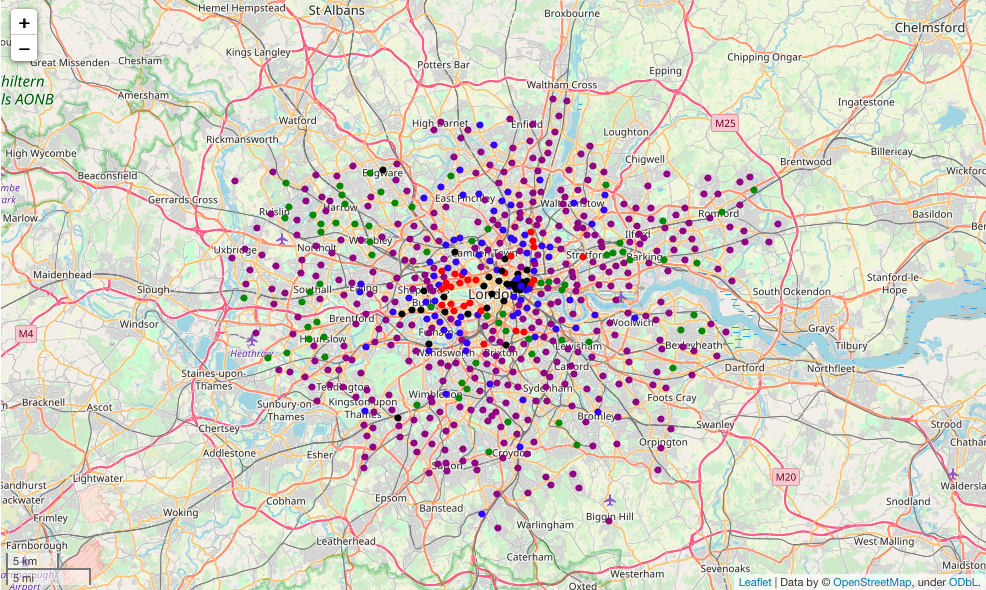
Results

The results of running the k-means clustering algorithm can be visualised in Figure 2 below. The positions of the pins are the same as those in Figure 1 but the pins are now colour-coded to represent different clusters. The 5 clusters are coloured purple, black, blue, red and green. Neighbourhoods belonging to black, blue and red clusters are concentrated towards the centre of the Greater London, whereas neighbourhoods belonging to purple and green clusters are located mostly away from the centre. Red cluster neighbourhoods tend to belong to west central London, whereas black cluster neighbourhoods tend to belong to east central London. Furthermore, blue cluster neighbourhoods tend to be concentrated in outer central London.

The clusters are analysed in more detail in Figure 3, where pie charts showing the most popular venues for each cluster can be observed. It should be noted that some classes in the legends have a 0.00 % contribution. If present, these classes form part of the 2nd and 3rd most popular venues in the cluster, and not the 1st. Further details for the 2nd and 3rd most popular venues can be found in the attached jupyter notebook.

From Figure 3, in three cases (b, d, and e) there appears to be a clear preference for one category: black cluster neighbourhoods tend to prefer East Asian cuisine, red cluster neighbourhoods tend to prefer Middle Eastern/North African cuisine, and green cluster neighbourhoods tend to prefer South Asian cuisine.

In the two remaining cases (a, c) there appears to be more balance across the categories. This is particularly true in Figure 3 (a) (purple cluster), where the most popular cuisine is Central/South American at 29.2 %, which is followed closely by South Asian cuisine at 21.6 %. Middle Eastern/North African, East Asian, and South-East Asian cuisines also form sizeable components. Also of note is the presence of African, Caribbean and East European cuisines in the purple cluster, which are not observed be the most popular venues in the remaining clusters. In comparison, blue cluster neighbourhoods tend to prefer South-East Asian cuisine at 39.5 %, with Middle Eastern/North African and East Asian cuisines also forming a sizeable component.

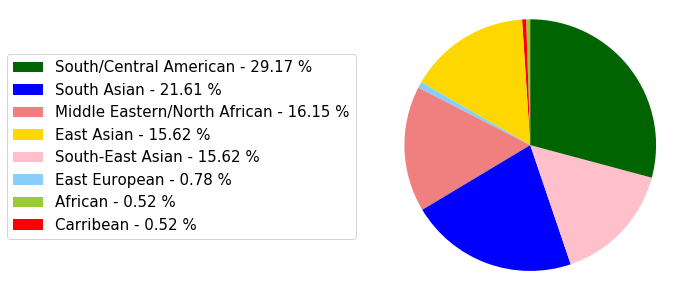


**Figure 2:** Map of Greater London with pins positioned at the locations of different electoral wards. The pins are colour coded purple, black, blue, red or green depending on their associated cluster determined by the k-means clustering algorithm.

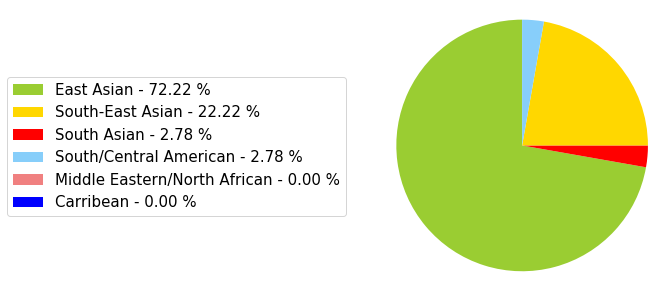
Discussion

The use of the k-means clustering algorithm has firstly identified 3 groups of neighbourhoods where local cuisine preferences are strongly either East Asian (black cluster), Middle Eastern/North African (red cluster), or South Asian (green cluster). These can be compared to some well-known example ground truths. London’s Chinatown is situated in the St James’ ward, which belongs to the black cluster. Edware road is well known for a large Middle Easter/North African community, and its associated ward (Bryanston and Dorset Square) belongs to the red cluster. Finally, the Whitechapel ward (well known for ‘Brick Lane’) holds a large Bangladeshi community and belongs to the green cluster. In these cases, given that only location data has been leveraged to come to the respective conclusions, the clustering algorithm appears to perform rather well for identifying local demographics.

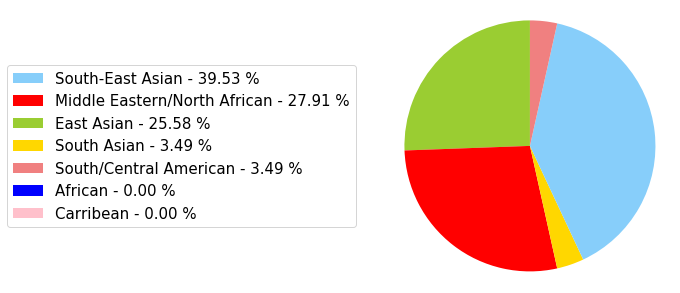
On the other hand, the remaining 2 groups of neighbourhoods (purple and blue) hold more balanced local cuisine preferences, and in these cases a closer look into the dataset on which the clustering was performed (see ‘wards\_venues\_sorted’ dataset in attached jupyter notebook) is required to determine the exact local preference in a neighbourhood. For example, Childs Hill is identified as having East European restaurants as the most popular kind of venue. The 2011 UK census reveals that Polish was the second most spoken language in the ward after English in 2011, therefore indicating a large East European population.



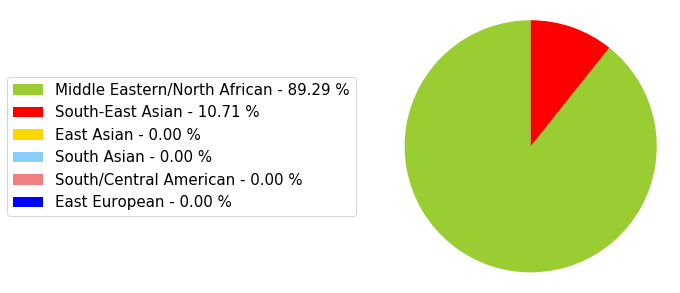
**(a)**



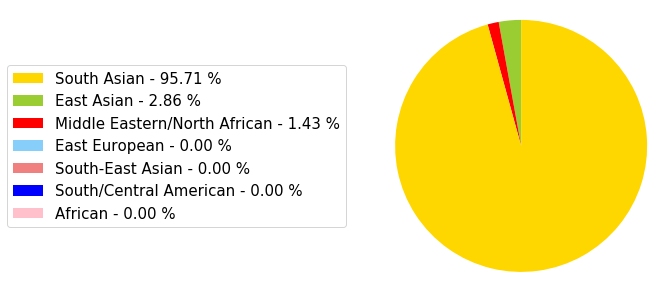
**(b)**



**(c)**



**(d)**



**(e)**

**Figure 3:** Pie charts showing the most popular venues for each cluster from Figure 2. (a) Black cluster, (b) Blue cluster, (c) Red cluster, (d) Green cluster, (e) Purple cluster.

Norbury Park is identified as having Caribbean restaurants as the most popular venue. Census data from 2011 [2] reveals that the Caribbean community is the largest after the ‘White British’ community. Finally, Caledonian is identified as having African restaurants as the most popular venue, which is consistent with a relatively large‘Black African’ population in the ward [3]. These examples support the initial assumption that local restaurants serve as a good indication of local demographics in an area.

The most subjective part of the data wrangling process is the classification of different kinds of restaurant into one of the 9 more broad categories. This requires some intuition of cultural similarities between different regions. Indeed, a different data scientist may not naturally classify Greek cuisine as ‘Middle Eastern/North African’, or those who have visited Poland more recently may feel it is more of a ‘West European’ country than an ‘East European’ country. In this respect, the clustering model is perhaps best used as a means to gain an initial understanding of local demographics before carrying out further more targeted research.

Conclusion

This project proposed the following question: can we use location data to gain an understanding of local demographics in different areas of Greater London?

In order to answer this question, the Foursquare API was used to identify the most popular restaurant venues within 500 m of each London ward. A simple k-means clustering model was then used to cluster the wards in terms of local cuisine popularity. Three of the five clusters were successful in identifying a non-native majority demographic in the wards. These results suggest such a method would be useful tool for a business to gain an initial understanding of consumer demographics in a particular neighbourhood.

References

[1] <http://www.ukcensusdata.com/childs-hill-e05000045#sthash.PfmyRpJ2.dpbs>, accessed 13/09/2019

[2] <http://www.ukcensusdata.com/norbury-e05000158#sthash.tAq7L4mT.dpbs>, accessed 13/09/2019

[3] <http://www.ukcensusdata.com/caledonian-e05000368#sthash.SuX7sKr3.dpbs>, accessed 13/09/2019