# FINDING THE OPTIMAL LOCATION FOR A BUSINESS

Nataliya Fotina

June, 2020

## 1. Problem Description

In this project, the problem attempted to solve will be to find the best possible location or the most optimal, for an Indian restaurant in the city of London, Great Britain. To achieve this task, an analytical approach will be used, based on advanced machine learning techniques and data analysis, concretely clustering and perhaps some data visualization techniques.

During the process of analysis, several data transformations will be performed, in order the find the best possible data format for the machine learning model to ingest. Once the data is set up and prepared, a modelling process will be carried out, and this statistical analysis will provide the best possible places to locate the Indian restaurant.

## 2. Data Presentation

The presented project will use data obtained from the following sources:

1. Foursquare API: requesting data on this site allows you to get information about the most popular and visited places for each of the districts in the city of London. Using this service is necessary not only to determine the geographical location of the main attractions and objects, popular places of recreation for citizens, but also to get an idea of traffic, people's preferences, photos and reviews of a particular place.

2. Wikipedia web page "ethnic groups in London" (https://en.wikipedia.org/wiki/Ethnic_groups_in_London): this site provides information about the ethnicity of the population of London, which is of great importance for solving this task. Data posted on the Wikipedia web page is cleared using BeautifulSoup4, and a table containing information about the Asian population of London is converted into a data frame. This data contains information about the immigrant population by region and nationality. The data obtained from this open source will be analyzed in order to identify the best location of the restaurant, determine the optimal venues for events, depending on the nationality of people living in each of the city's districts. It is suggested that people's sympathies vary depending on their nationality and that people from one particular country will be more attached to a place that corresponds to the environment and culture of their own countries.

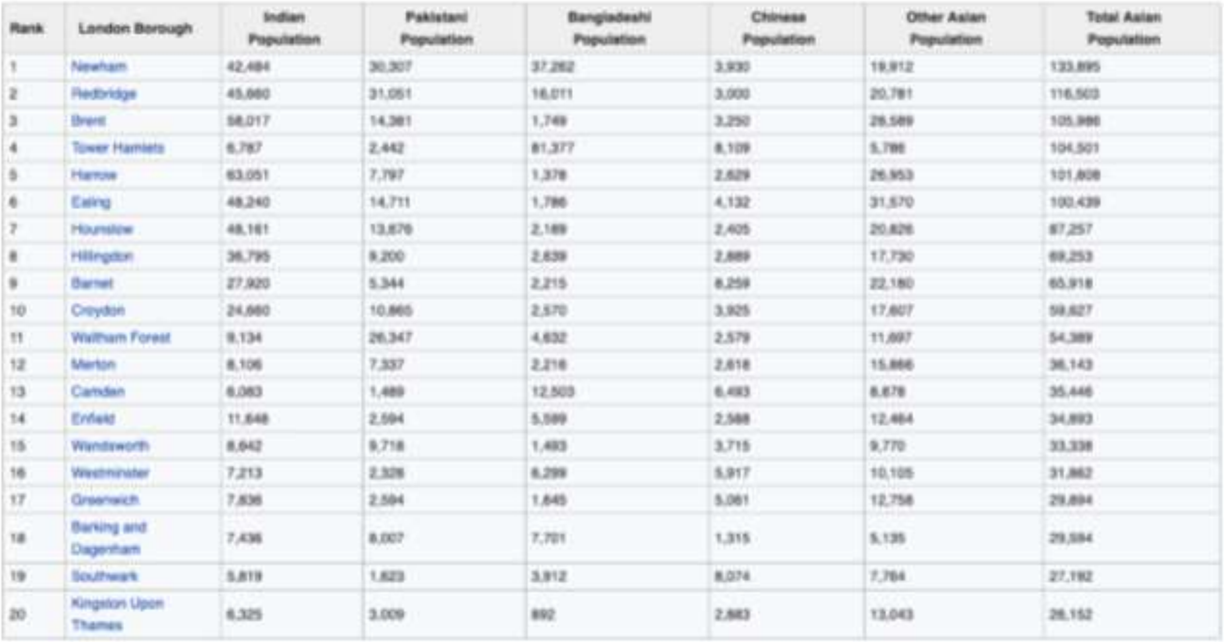
## 3. Methodology

In the development of this project, the clustering method was applied, including the use of the K-Means algorithm, implemented using Python.

The most difficult task in this project was to obtain the necessary data to implement a constructive approach to the problem. Since the project needed to find the optimal location for a restaurant of national cuisine, special attention was paid to finding information about consumers of these services. A study was conducted on information about residents of London, taking into account the national characteristic. It was assumed that representatives of certain nationalities prefer to visit restaurants that provide dishes of the appropriate national cuisine, reflecting the traditions and culture of these countries.

To effectively solve this problem, a study of London boroughs was performed, the number of representatives of the target audience living in each of the city's boroughs was determined, as well as competitor objects that already exist in each of the boroughs were studied.

Here is an example of the data used:



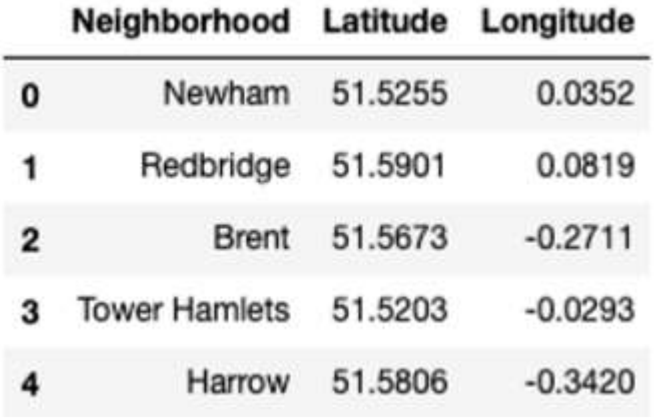
| Rank | London Borough | Indian Population | Pakistani Population | Bangladeshi Population | Chinese Population | Other Asian Population | Total Asian Population |
|---|---|---|---|---|---|---|---|
| 1 | Newham | 42,484 | 30,307 | 37,262 | 3,930 | 19,912 | 133,895 |
| 2 | Redbridge | 45,660 | 31,051 | 16,011 | 3,000 | 20,781 | 116,503 |
| 3 | Brent | 58,017 | 14,381 | 1,749 | 3,250 | 28,589 | 105,986 |
| 4 | Tower Hamlets | 6,787 | 2,442 | 81,377 | 8,109 | 5,786 | 104,501 |
| 5 | Harrow | 63,051 | 7,797 | 1,378 | 2,629 | 26,953 | 101,808 |
| 6 | Ealing | 48,240 | 14,711 | 1,786 | 4,132 | 31,570 | 100,439 |
| 7 | Hounslow | 48,161 | 13,676 | 2,189 | 2,405 | 20,826 | 87,257 |
| 8 | Hillingdon | 36,795 | 9,200 | 2,639 | 2,889 | 17,730 | 69,253 |
| 9 | Barnet | 27,920 | 5,344 | 2,215 | 8,259 | 22,180 | 65,918 |
| 10 | Croydon | 24,660 | 10,865 | 2,570 | 3,925 | 17,607 | 59,627 |
| 11 | Waltham Forest | 9,134 | 26,347 | 4,632 | 2,579 | 11,697 | 54,389 |
| 12 | Merton | 8,106 | 7,337 | 2,216 | 2,618 | 15,866 | 36,143 |
| 13 | Camden | 6,083 | 1,489 | 12,503 | 6,493 | 8,878 | 35,446 |
| 14 | Enfield | 11,648 | 2,594 | 5,599 | 2,588 | 12,464 | 34,893 |
| 15 | Wandsworth | 8,642 | 9,718 | 1,493 | 3,715 | 9,770 | 33,338 |
| 16 | Westminster | 7,213 | 2,328 | 6,299 | 5,917 | 10,105 | 31,862 |
| 17 | Greenwich | 7,836 | 2,594 | 1,645 | 5,061 | 12,758 | 29,894 |
| 18 | Barking and Dagenham | 7,436 | 8,007 | 7,701 | 1,315 | 5,135 | 29,594 |
| 19 | Southwark | 5,819 | 1,623 | 3,912 | 8,074 | 7,764 | 27,192 |
| 20 | Kingston Upon Thames | 6,325 | 3,009 | 892 | 2,883 | 13,043 | 26,152 |

This data contains information about the quantities of Asian immigrant populations in London inside each borough. The main features are the ethnicities, which indicates where the people of that live in those boroughs come from. It contains also the quantities of people by country living in each borough. So, with this, it is already possible to have an idea of where is our target population located.

In this project, the idea is to open an Indian restaurant in the city. With further analysis, this question will be answered. Nevertheless, this task could not be achieved only working with this raw data. It was also needed to obtain information about the most common venues in these boroughs, besides of the population kind that was inhabiting on the different boroughs. It was also needed to determine somehow in what measure these boroughs were different or similar between them.

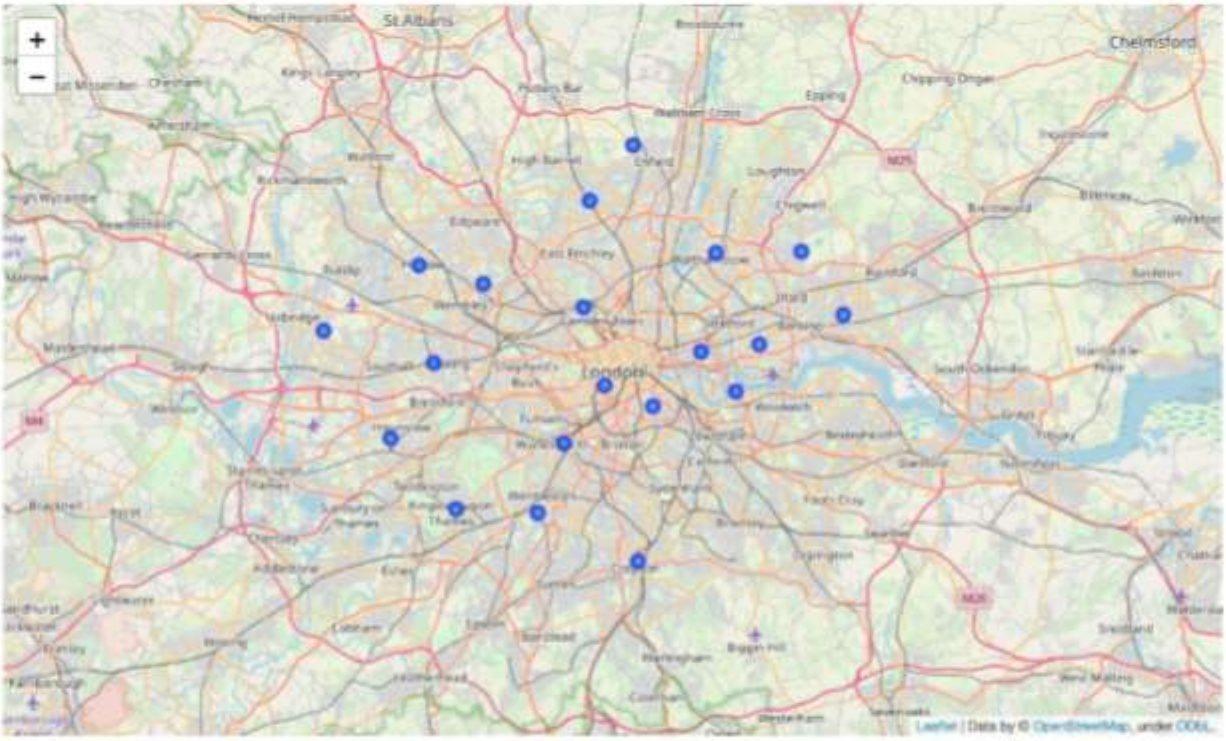
To continue this line, The Foursquare API was used to obtain the needed data about the venues in each boroughs, but to use the Foursquare API, it was first necessary to transform the raw data to something the Foursquare API was capable to handle. Basically, the coordinates of each boroughs were needed.

This is an example of the transformed data:

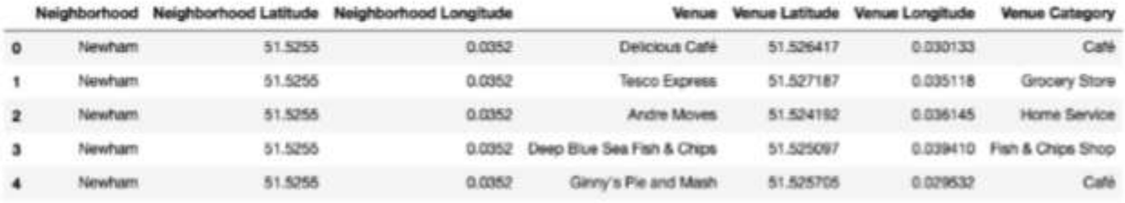


|   | Neighborhood | Latitude | Longitude |
|---|---|---|---|
| 0 | Newham | 51.5255 | 0.0352 |
| 1 | Redbridge | 51.5901 | 0.0819 |
| 2 | Brent | 51.5673 | -0.2711 |
| 3 | Tower Hamlets | 51.5203 | -0.0293 |
| 4 | Harrow | 51.5806 | -0.3420 |

Once the data was transformed into a format ingestible by the Foursquare API, the information about the venues could be obtained. The boroughs were then onto a map of London, so it was possible to have an idea of their geographical situation:



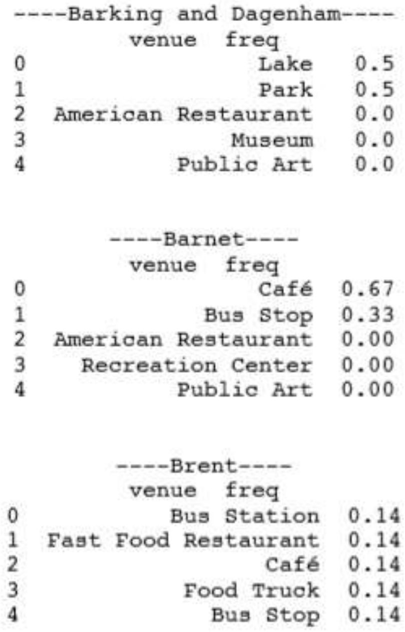
The next step was to obtain the nearby venues by boroughs, together with their respective coordinates:



| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Newham | 51.5255 | 0.0352 | Delicious Café | 51.526417 | 0.030133 | Café |
| 1 | Newham | 51.5255 | 0.0352 | Tesco Express | 51.527187 | 0.035118 | Grocery Store |
| 2 | Newham | 51.5255 | 0.0352 | Andre Moves | 51.524192 | 0.036145 | Home Service |
| 3 | Newham | 51.5255 | 0.0352 | Deep Blue Sea Fish & Chips | 51.525097 | 0.039410 | Fish & Chips Shop |
| 4 | Newham | 51.5255 | 0.0352 | Ginny's Pie and Mash | 51.525705 | 0.029532 | Café |

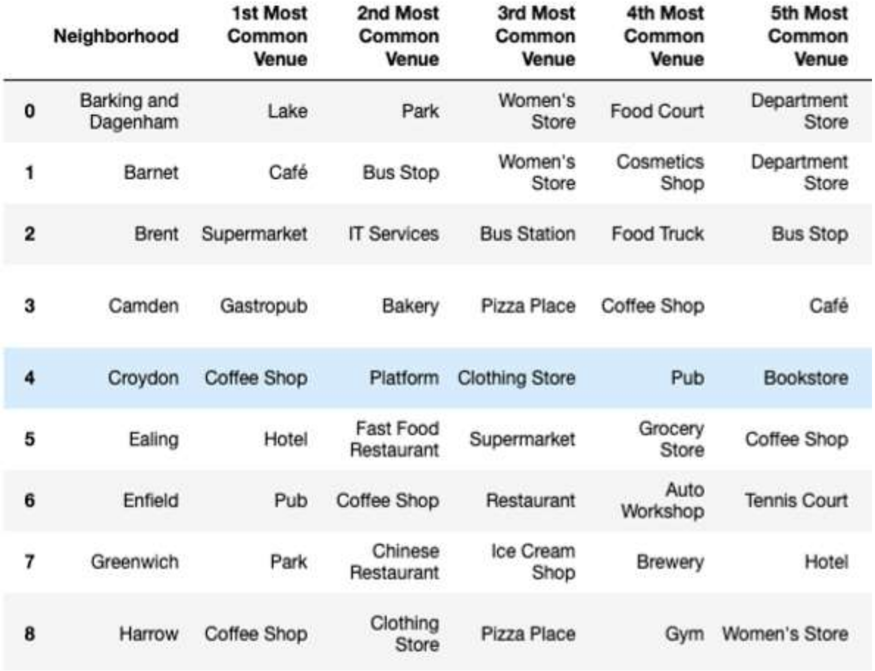
Looking at this sample, it is possible to see the names of the venues, their coordinates, and the category of each venue. The results are ordered by boroughs. This is a vital step in the segmentation process, since all the important data about the venues is obtained from here.

Once the venues per boroughs were obtained, it was then needed to look at the mean occurrence of each venue by neighbourhood:



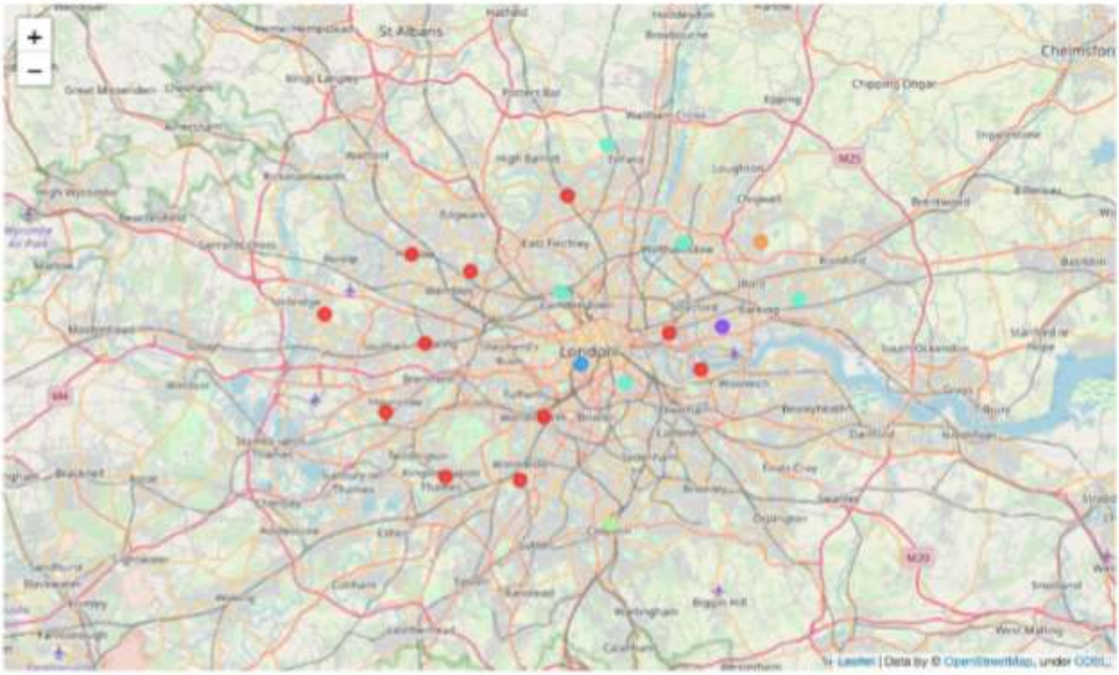
```
----Barking and Dagenham----
         venue  freq
0                  Lake   0.5
1                  Park   0.5
2  American Restaurant   0.0
3               Museum   0.0
4           Public Art   0.0


       ----Barnet----
         venue  freq
0                 Café  0.67
1             Bus Stop  0.33
2  American Restaurant  0.00
3    Recreation Center  0.00
4           Public Art  0.00


        ----Brent----
         venue  freq
0            Bus Station  0.14
1  Fast Food Restaurant  0.14
2                  Café  0.14
3            Food Truck  0.14
4              Bus Stop  0.14
```

This what the frequencies of occurrence looks like. With this data, it is possible to know which the most common venues are:



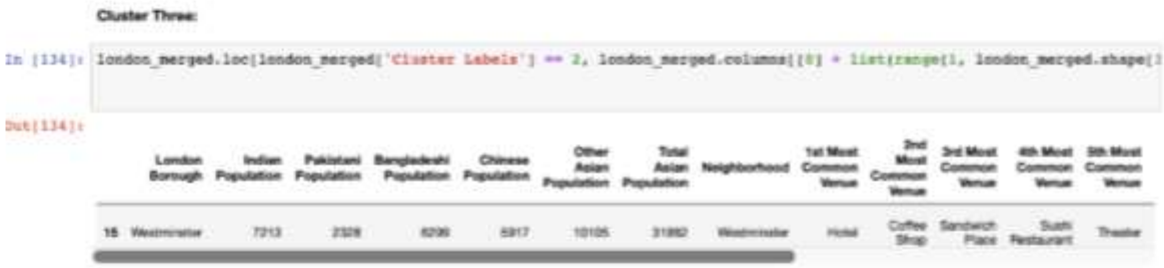
| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|
| 0 | Barking and Dagenham | Lake | Park | Women's Store | Food Court | Department Store |
| 1 | Barnet | Café | Bus Stop | Women's Store | Cosmetics Shop | Department Store |
| 2 | Brent | Supermarket | IT Services | Bus Station | Food Truck | Bus Stop |
| 3 | Camden | Gastropub | Bakery | Pizza Place | Coffee Shop | Café |
| 4 | Croydon | Coffee Shop | Platform | Clothing Store | Pub | Bookstore |
| 5 | Ealing | Hotel | Fast Food Restaurant | Supermarket | Grocery Store | Coffee Shop |
| 6 | Enfield | Pub | Coffee Shop | Restaurant | Auto Workshop | Tennis Court |
| 7 | Greenwich | Park | Chinese Restaurant | Ice Cream Shop | Brewery | Hotel |
| 8 | Harrow | Coffee Shop | Clothing Store | Pizza Place | Gym | Women's Store |

The next step was to perform segmentation and create clusters. But first you need to determine what the appropriate number of clusters is. The elbow method was used to perform this task. This method consists of constructing a hypothetical and usually large number of clusters in your data and constructing a curve representing the squares of the distances between each cluster. At some point, the distance will decrease to a point where there is no need to constantly increase it. This means that creating additional divisions in the data (clusters) is pointless, since the difference between groups begins to be very difficult to assess. In our case, the distance started to decrease significantly only from the 6th cluster. And so, it was determined that the optimal number of clusters for this task is 6.



These are the 6 clusters on the map of London, it is possible to see how many neighbourhoods belong to each cluster, which is also important information.

Now it is possible to examine the data of each cluster:

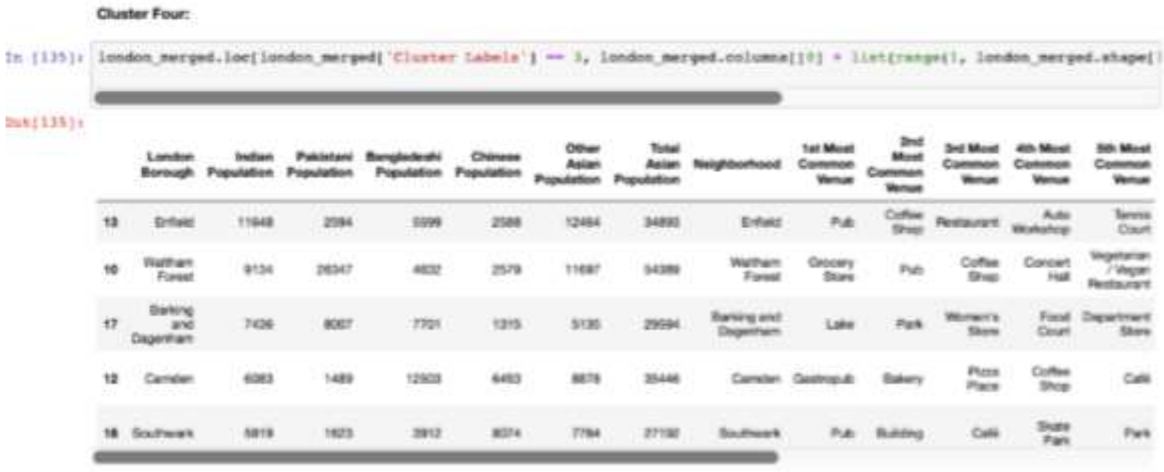


Cluster Three:

```
In [134]: london_merged.loc[london_merged['Cluster Labels'] == 2, london_merged.columns[[0] + list(range(1, london_merged.shape[
```

Out[134]:

| | London Borough | Indian Population | Pakistani Population | Bangladeshi Population | Chinese Population | Other Asian Population | Total Asian Population | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | Westminster | 7213 | 2328 | 6299 | 5917 | 10105 | 31862 | Westminster | Hotel | Coffee Shop | Sandwich Place | Sushi Restaurant | Theater |



Cluster Four:

```
In [135]: london_merged.loc[london_merged['Cluster Labels'] == 3, london_merged.columns[[0] + list(range(1, london_merged.shape[
```

Out[135]:

| | London Borough | Indian Population | Pakistani Population | Bangladeshi Population | Chinese Population | Other Asian Population | Total Asian Population | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13 | Enfield | 11648 | 2594 | 5599 | 2588 | 12464 | 34893 | Enfield | Pub | Coffee Shop | Restaurant | Auto Workshop | Tennis Court |
| 10 | Waltham Forest | 9134 | 26347 | 4632 | 2579 | 11697 | 54389 | Waltham Forest | Grocery Store | Pub | Coffee Shop | Concert Hall | Vegetarian / Vegan Restaurant |
| 17 | Barking and Dagenham | 7436 | 8007 | 7701 | 1315 | 5135 | 29594 | Barking and Dagenham | Lake | Park | Women's Store | Food Court | Department Store |
| 12 | Camden | 6083 | 1489 | 12503 | 6493 | 8878 | 35446 | Camden | Gastropub | Bakery | Pizza Place | Coffee Shop | Café |
| 18 | Southwark | 5819 | 1623 | 3912 | 8074 | 7764 | 27192 | Southwark | Pub | Building | Café | Skate Park | Park |

So, this kind of approach, allow us to perform an analysis of an entire city by looking at its venues and population. With this information, observations and conclusions can be made now.

## 4. Results

The results obtained were six clusters of very different population and venues distribution. The following is a description of the clusters:

| **Cluster One** | **Cluster Two** | **Cluster Three** | **Cluster Four** | **Cluster Five** | **Cluster Six** |
|---|---|---|---|---|---|
| Mostly inhabited by Indians and other Asians. The most common venues are Coffee shops, pizza places and supermarkets, among many others | This cluster is mostly composed of 2 different population kinds: Indian people and Bangladeshi people. The most common venues are Coffee shops, Parks, Women's Store and Department Stores, among others | This cluster is majorly composed of other Asian population. The most common places Hotels, coffee shops and bars | This is a very variate cluster, we see a majority of Pakistani, Bangladeshi and Indian population. The most common venues are Pubs, parks, gyms or fitness centres and electronic stores | This cluster is mostly comprised of Indian population in the lead followed by Pakistani population. The prominent venues here are sushi restaurants, coffee shops and other Asian restaurants | This cluster is similar to cluster five in terms of population diversity with Indian population in the lead followed by Pakistani population. The prominent venues here are Supermarkets, pharmacies and fast food restaurants |

.

## 5. Discussion

This study shows how venues and people from different countries change in different clusters. The factor that most actively influences the choice of the location of a restaurant of national cuisine in this project is the number of people of a certain nationality. However, to determine the location of a given object more accurately, other social and demographic data about the population should also be taken into account, such as income level, children, education level, profession, and others. In addition, one of the most important data for careful study is data related to people's likes and tastes about how they prefer to spend their leisure time, what types of food they like or what their Hobbies are. By collecting all this data, it would be possible to conduct a deeper analysis, and the segmentation would be more accurate. For this project, this data was not available and was also outside the scope of the project.