FINDING THE OPTIMAL LOCATION FOR A BUSINESS

Sergey Fotin

May, 2020

Contents:

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, England. 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 data that will be used to develop this project is based on two sites:

1. The Foursquare API: This data will be accessed via Python and used to obtain the most common venues per neighbourhood in the city of London. This way, it is possible to have a taste of how the city's venues are distributed, what are the most common places for leisure, and in general, it will provide an idea of what people's likes are.

2. Wikipedia's Ethnic groups in London webpage: This site provides information about ethnicity of population in London which is of great utility to solve this problem. The webpage is scraped using BeautifulSoup4, and the table containing Asian population of London is converted into DataFrame. The data contains information about the immigrant population per borough and per nationality. This data will be analysed in such a way that one could determine the best location of venue/restaurant/other based on people's nationalities. For the sake of simplicity, it will be assumed for this exercise that people's likes vary according to their nationality, and that people from one specific country will be more attracted to place that matches the environment and culture of their own countries, rather than the ones from foreign countries.

You can access the data by clicking this link:

https://en.wikipedia.org/wiki/Ethnic_groups_in_London

3. Methodology

The methodology used to approach this problem includes some statistical exploration of the data and some visualizations. The main machine learning technique involved in the development of this project is clustering, in concrete the K-Means algorithm was used, implemented with Python.

At a first moment, the main problem was how to obtain the necessary data to build a constructive approach to the problem to be tackled. Usually, to solve these kinds of optimal business location problems, a lot of consumer’s data are needed, but for this example and for the sake of simplicity, the focus was put mainly on the population’s nationality. A study was carried out over the inhabitants of London, and it was assumed for this example that the national population from a certain country would prefer restaurants based on their national country and food, rather than restaurants from other countries or that have nothing to do with the culture of their countries, especially when it comes to immigrant populations, that are not in their countries, and certainly would like to usually have a taste of their food and original culture. Because in the end, it is not only about the food, it is also about having a piece of the country in question. When a someone enters in an Italian restaurant, or American, or Peruvian restaurant, they are not only consuming the food and culinary specialties of the country in question, but also the culture, the people, the music, the decoration. All of this must make people feel like they were there on the country.

With all this being considered, it was decided that the main goal to efficiently solve this problem, was firstly to define what our target population is, and secondly, find the areas where this population is living, and finally, examine the venues and restaurants in this area to see if our product could work.

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.

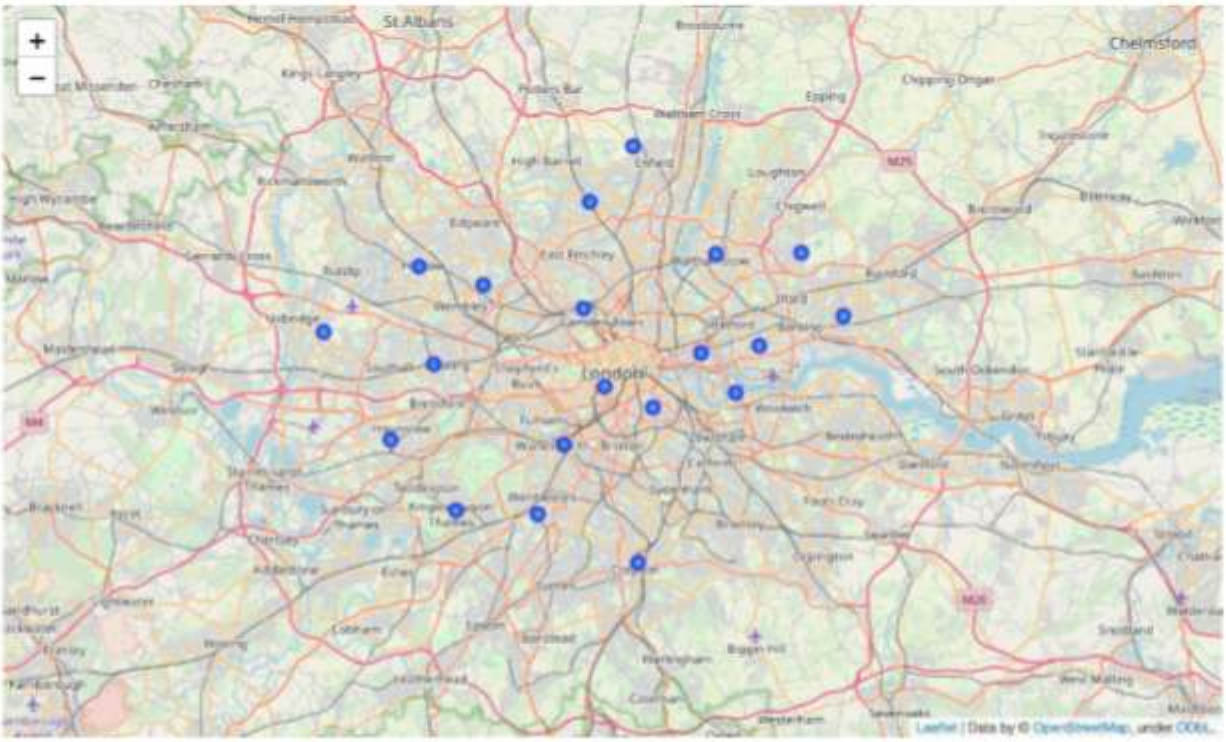
In this project, the idea is to open an Indian restaurant in the city. With further analysis, this question will be answered. Nevertheless, this task could not be achieved only working with this raw data. It was also needed to obtain information about the most common venues in these boroughs, besides 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	0.029632	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.

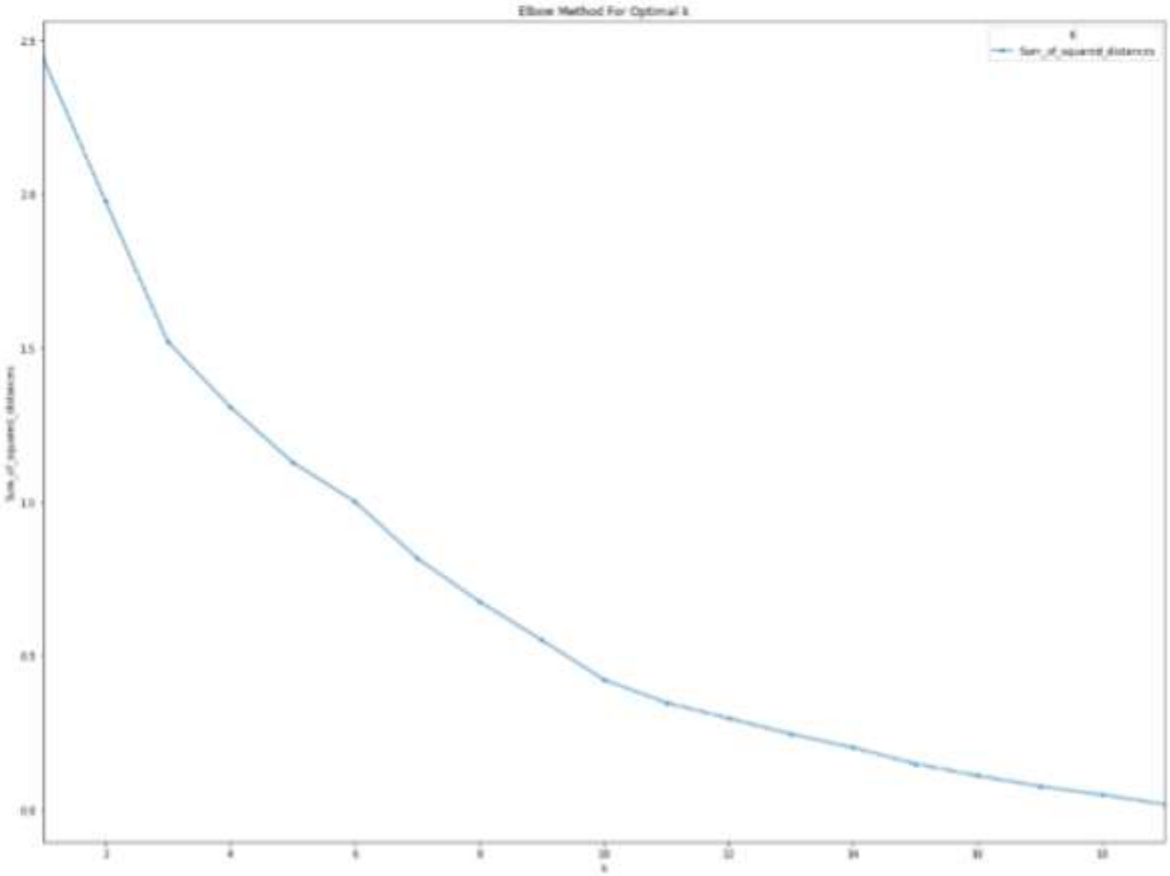
Once the venues per boroughs were obtained, it was then needed to look at the mean occurrence of each venue by neighbourhood:

```
----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
```

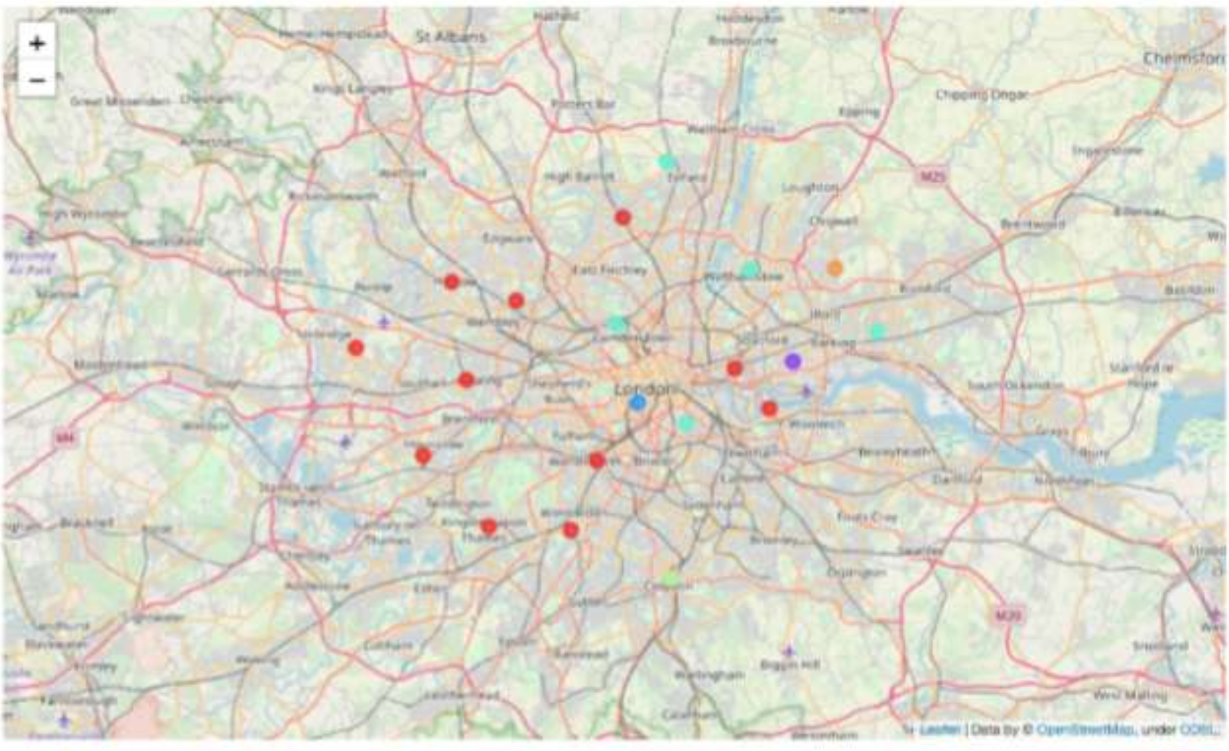
This what the frequencies of occurrence looks like. With this data, it is possible to know which the most common venues are:

	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

This process is progressive, once a piece of information is obtained, it is possible to go for the next one. With this data in hand, now the segmentation can be made, and the clusters created. But first it is necessary to determine somehow, what the appropriate number of clusters is. To perform this task, the elbow method was used. This method consists in plotting a hypothetical and usually large number of clusters in our data, and draw a curve representing the squared distances between each cluster. At some point, the distances will descend to a point where there is no need to keep increasing them. This means that creating more divisions in the data (clusters) is pointless as the difference between groups starts being highly difficult to appreciate:



This is our curve. The distances start reducing importantly from cluster 6 on. So, it was determined that the optimal number of clusters for this problem was 6. With this being done, it is possibly to build the clusters now and have a look at them:



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[1
```

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	31882	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[1
```

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	4832	2579	11687	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:

Mostly inhabited by Indians and other Asians. The most common venues are Coffee shops, pizza places and supermarkets, among many others.

• Cluster Two:

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.

• Cluster Three:

This cluster is majorly composed of other Asian population. The most common places Hotels, coffee shops and bars.

• Cluster Four:

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.

• Cluster Five:

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.

• Cluster Six:

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

It is interesting how the venues and people from different countries varies to one cluster to another. The main differentiation is located on these two variables. Each cluster has its own characteristics, but also common spots with other clusters. If we examine with more detail these results, some conclusions can be made.

As a recommendation, it must be said in a study of this size, to make good predictions about where to open a certain business or shop, more data is needed. For example, socio-demographic data about the population, like their income level, if they have children or not, the education level, what kind of job do they make a living from, etc.… Also, one of the most important data to examine carefully are the data related to the people's likes and tastes about how they prefer to spend their leisure time, what kinds of food do they like, or what are their hobbies. With all these data gathered, a more in- depth analysis could be performed, and the segmentations would be more accurate. For this project, these data weren't available, and was also out of the project's scope.