Capstone Project - The Battle of the Neighborhoods

Jyoti Kumar

May 7, 2020

1. Introduction

London is the capital city of United Kingdom with dense population. It has a huge variety of restaurants for every taste and, thus, to start a restaurant business in this area is not an easy task. Our stakeholder is willing to open an Indian restaurant in the London city with middle-high level prices.

Of course, choosing a location for business is one of the stressful and controversial tasks, since there are a lot of criteria that should be satisfied in order to achieve the highest revenue.
Here are some of them:

- The density of other restaurants
- The density of specifically Indian restaurants
- Population density around the location
- Population density of specifically Asian community around the location
- Solvency of the population around the location

In this project, we will implement the basic analysis and try to find the most optimal Borough to open the Indian restaurant according to those criteria. It's obvious, that there are many additional factors, such as distance from parking places or distance from the main streets, but this analysis can be done after choosing the Borough, and thus will not be performed within the scope of this project.

2. Data description

Based on criteria listed above the following data will be utilized in our analysis:

• The number of restaurants/Indian restaurant within the certain radius of each borough (Foresquare API)

• The net income per person in each borough. Since the restaurant will have middle-high prices, it is important to consider the solvency of population. Source: London datstore, Mayor of London (https://data.london.gov.uk/dataset/earnings-place-residence-borough)

• The population and the population density of the borough. Source: Wikipedia and Mayor of London (https://data.london.gov.uk/dataset/land-area-and-population-density-ward-and-borough and https://en.wikipedia.org/wiki/List_of_London_boroughs)

• The population of Asian community in each borough. Source: London datstore, Mayor of London (https://data.london.gov.uk/dataset/ethnic-groups-borough)

• The coordinates of the borough. Source: Wikipedia (https://data.london.gov.uk/dataset/land-area-and-population-density-ward-and-borough)

After cleaning and preparing the data, we defined the following master data frame:

	Borough	Latitude	Longitude	Population	Population Density	Asian Population	Asian Density	Net income per person	Number of restaurants	Number of Indian restaurants
0	Barking and Dagenham	51.5607	0.1557	212773	5892	54000	1495	479.1	1.0	0.0
1	Barnet	51.6252	-0.1517	397049	4577	57000	657	536.6	0.0	0.0
2	Bexley	51.4549	0.1505	249999	4126	17000	280	513.8	12.0	0.0
3	Brent	51.5588	-0.2817	336859	7791	107000	2474	480.0	18.0	4.0
4	Bromley	51.4039	0.0198	332733	2216	15000	99	632.5	11.0	2.0
5	Camden	51.5290	-0.1255	252637	11594	39000	1789	634.7	13.0	1.0
6	Croydon	51.3714	-0.0977	391296	4523	70000	809	552.0	26.0	3.0
7	Ealing	51.5130	-0.3089	350784	6315	96000	1728	523.0	28.0	3.0
8	Enfield	51.6538	-0.0799	337697	4177	37000	457	479.1	9.0	2.0
9	Greenwich	51.4892	0.0648	286322	6048	39000	823	573.7	10.0	2.0
10	Hackney	51.5450	-0.0553	281740	14790	32000	1679	555.6	8.0	0.0
11	Hammersmith and Fulham	51.4927	-0.2339	184050	11224	20000	1219	681.3	31.0	7.0
12	Haringey	51.6000	-0.1119	284288	9604	18000	608	548.1	19.0	3.0
13	Harrow	51.5898	-0.3346	255369	5060	98000	1942	538.3	9.0	3.0
14	Havering	51.5812	0.1837	257511	2292	13000	115	544.2	12.0	0.0
15	Hillingdon	51.5441	-0.4760	309926	2678	100000	864	531.9	16.0	2.0
16	Hounslow	51.4746	-0.3680	278264	4970	86000	1536	535.9	16.0	7.0
17	Islington	51.5416	-0.1022	238267	16037	17000	1144	687.6	22.0	1.0
18	Kensington and Chelsea	51.5020	-0.1947	159301	13139	18000	1484	669.3	28.0	3.0
19	Kingston upon Thames	51.4085	-0.3064	179581	4819	30000	805	627.1	26.0	2.0
20	Lambeth	51.4607	-0.1163	334724	12485	28000	1044	620.4	31.0	2.0
21	Lewisham	51.4452	-0.0209	310324	8828	23000	654	551.4	7.0	0.0
22	Merton	51.4014	-0.1958	209421	5566	35000	930	594.8	6.0	1.0
23	Newham	51.5077	0.0469	353245	9758	166000	4585	479.1	4.0	0.0
24	Redbridge	51.5590	0.0741	305910	5422	126000	2233	554.7	5.0	1.0
25	Richmond upon Thames	51.4479	-0.3260	199419	3473	11000	191	678.2	15.0	2.0
26	Southwark	51.5035	-0.0804	322302	11166	17000	589	589.4	16.0	0.0
27	Sutton	51.3618	-0.1945	207378	4729	36000	821	541.2	6.0	1.0
28	Tower Hamlets	51.5099	-0.0059	317203	16035	128000	6470	627.9	13.0	1.0
29	Waltham Forest	51.5908	-0.0134	283524	7305	44000	1133	529.2	7.0	0.0
30	Wandsworth	51.4567	-0.1910	324400	9467	19000	554	689.9	10.0	1.0

Fig. 2.1 Main data frame

3. Methodology and Analysis

After cleaning and preparing the data, let us identify the steps that have to be performed in order to find the most optimal boroughs. Firstly, we will apply some basic exploratory analysis to our data. For that let's find the location of each borough on the map. Then we can visually inspect some values in our data with the help of bar charts.

Secondly, we have the possibility to reduce the number features in data frame by replacing them with more reasonable data. Finally, we will perform cluster analysis to find the best cluster of boroughs with meaningful features.

3.1 Exploratory Data Analysis

First, it's is quite useful to visualize the center locations of each borough. For that, the map of London was created with boroughs superimposed on top.

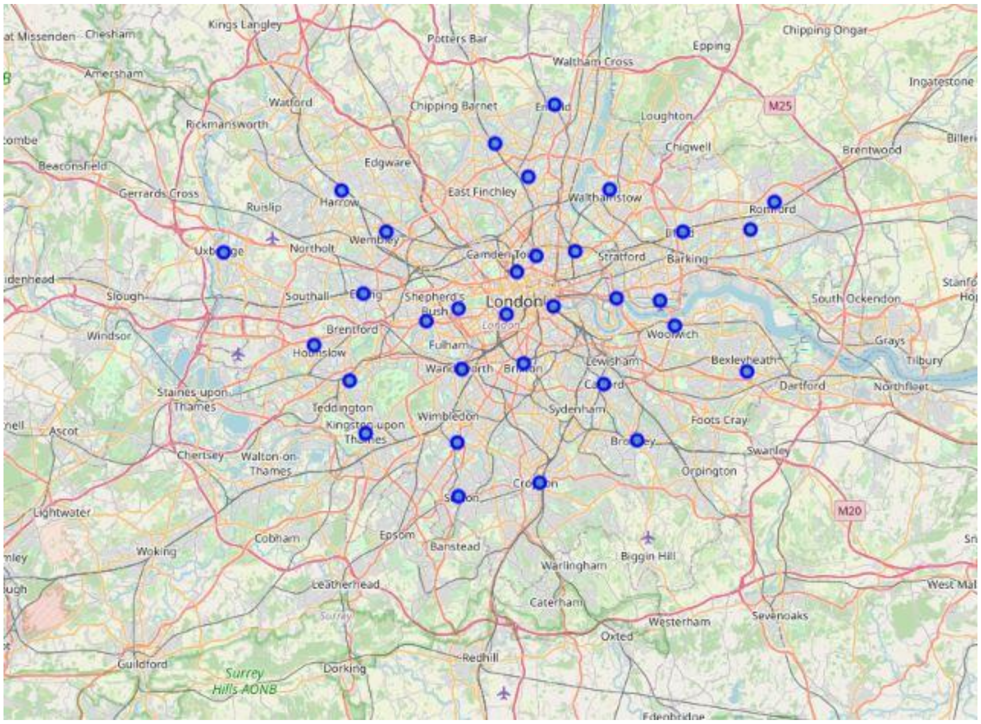

Fig. 3.1 London map with centers of the boroughs

Additionally, the several features, that influence our choice, were visualized with respect to the Boroughs.

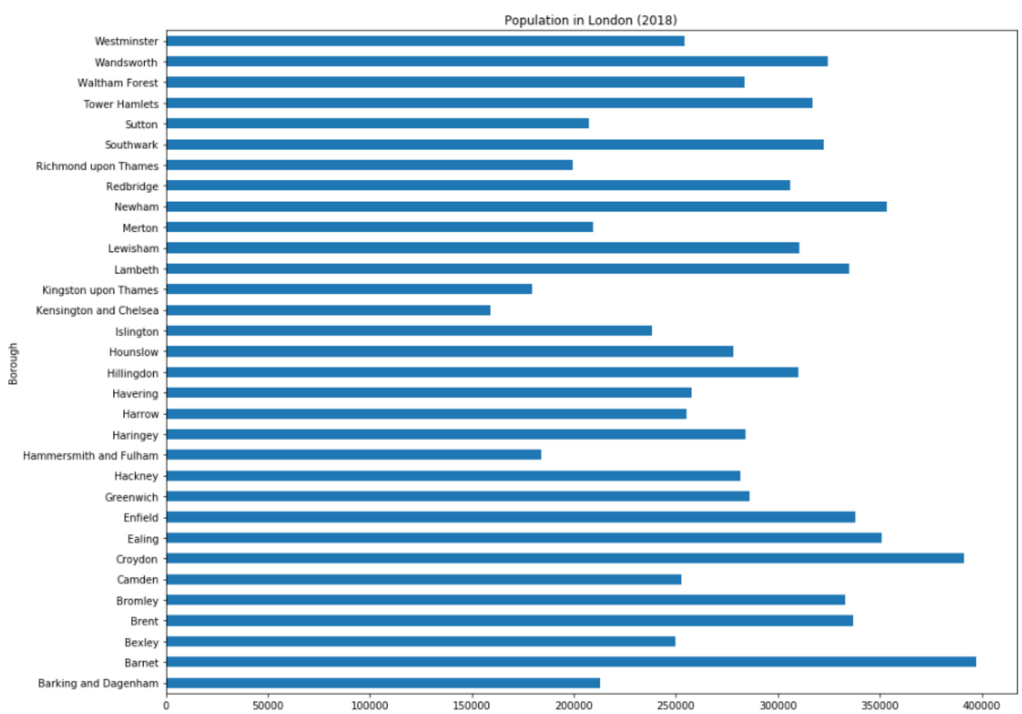

Fig. 3.2 Population in London (2018)

Fig. 3.2 shows that Barnet is most populous borough in the city. However, since it has one of the largest areas, the population density barely exceeds the average value (Fig. 3.3). From the figure, the most densely populated borough is Tower Hamlets.

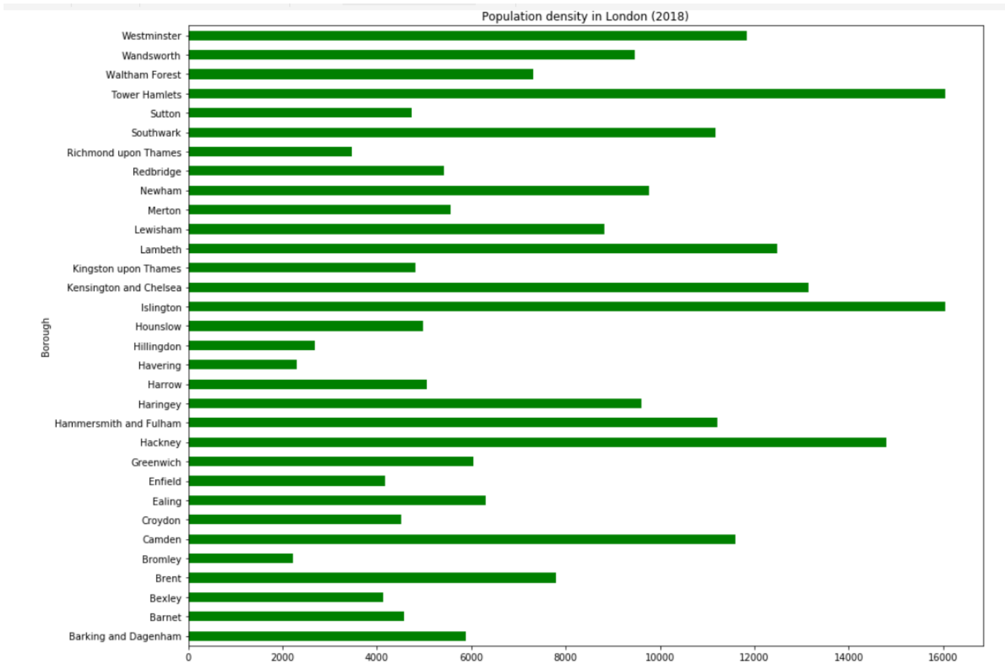

Fig. 3.3 Population Density in London (2018)

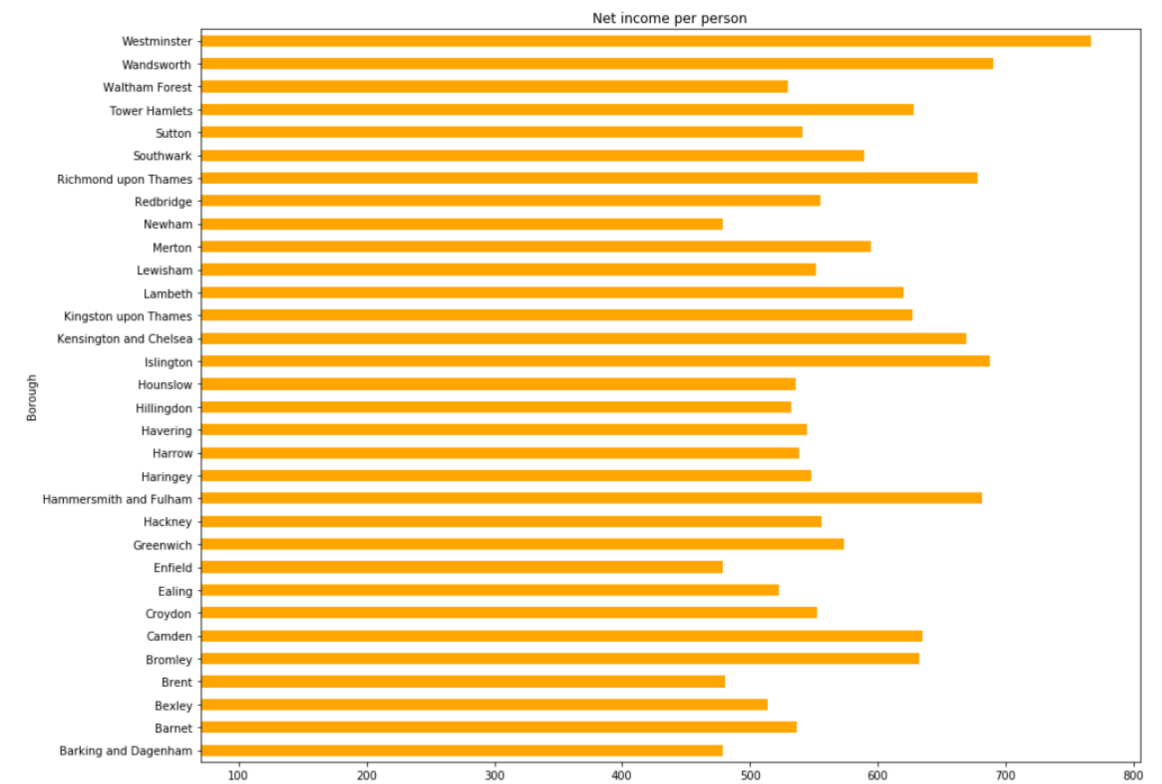

Fig. 3.4 Net weekly Income per Person in £

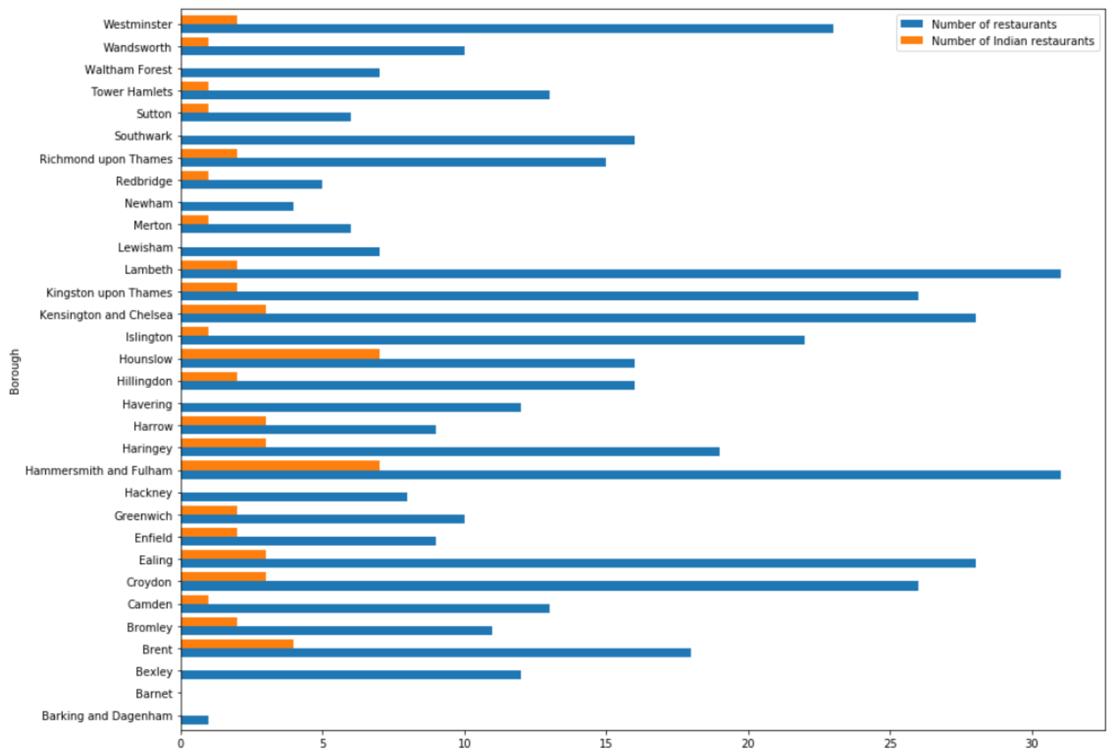

Fig. 3.5 Number of restaurants

3.2 Cluster Analysis

To identify groups (clusters) with similar characteristics, the unsupervised learning method to our data, namely K-Means algorithm, was applied to our data. I performed the analysis twice, one using the total population and another using only Asian population. The intersection of both these analysis can give more optimum borough for the restaurant location.

3.2.1 Using total population

Before performing K-mean algorithm, to reduce dimensionality of the problem the columns "Population", "Number of restaurants" and "Number of Indian restaurants" were removed. These three columns were replaced with two new ones, namely, "Number of restaurants per 10 thousand people" and "Number of Indian restaurants per 10 thousand people".

	Net income per person	Number of restaurants per 10000 people	Number of Indian restaurants per 10000 people
0	479.1	0.046998	0.000000
1	536.6	0.000000	0.000000
2	513.8	0.480002	0.000000
3	480.0	0.534348	0.118744
4	632.5	0.330595	0.060108

Fig. 3.6 Modified dataframe

To identify the optimal number of clusters, the Elbow method is used:

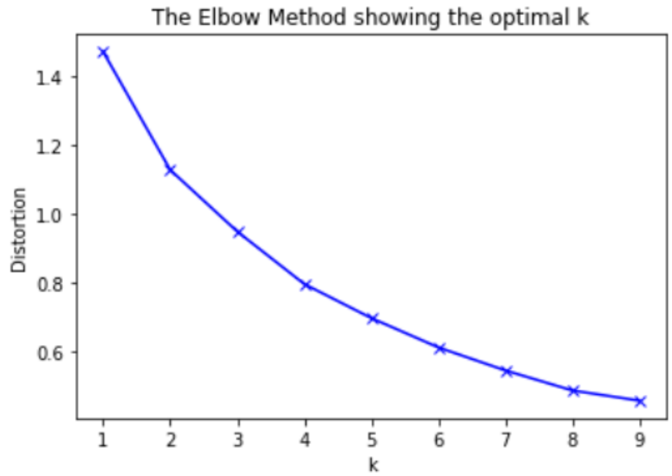

Fig. 3.7 The Elbow method showing the optimal number of clusters

It is not very clear from the graph where the exact elbow is however it seems to be somewhere between 2 and 4, so three clusters are the best choice.

Based on clustering results, two maps are created. The first map illustrates the clusters where the colors are associated with clusters and radius of the Circle marker is proportional to a number of restaurants per 10000 people in each borough. The second map shows the clusters where the radius of the Circle marker is proportional to a Net income per person in each borough.

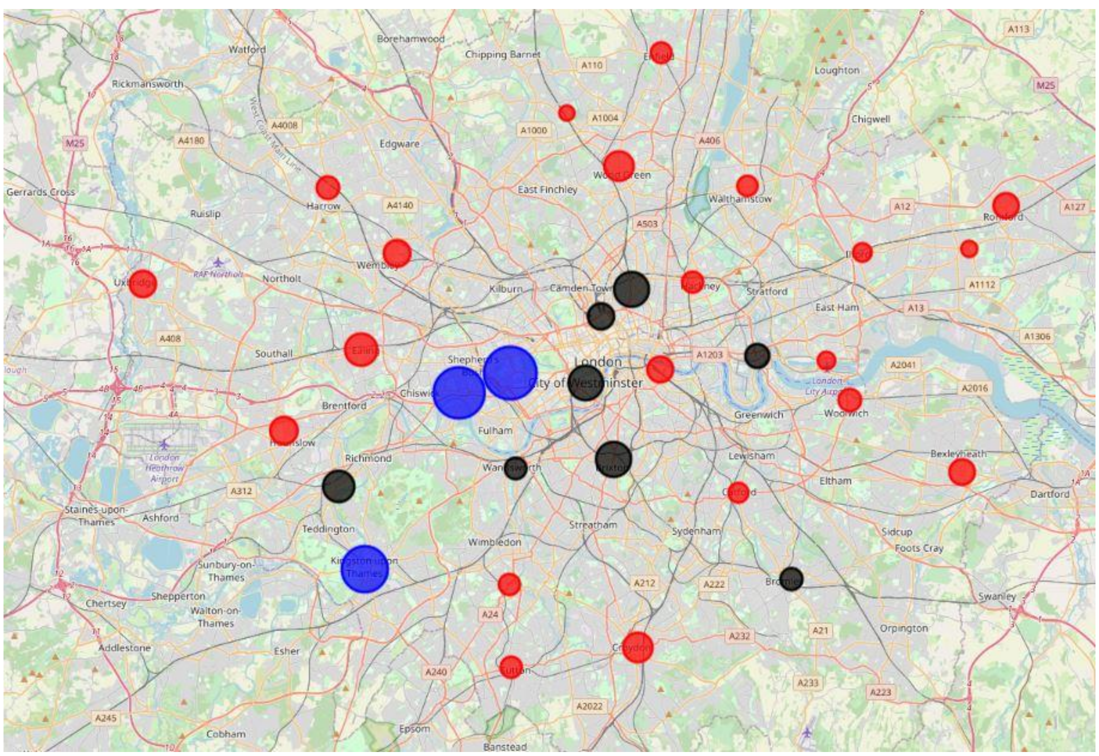

Fig. 3.8 London Map with clustered boroughs (radius of the circle marker is proportional to a number of restaurants per 10000 people in each borough)

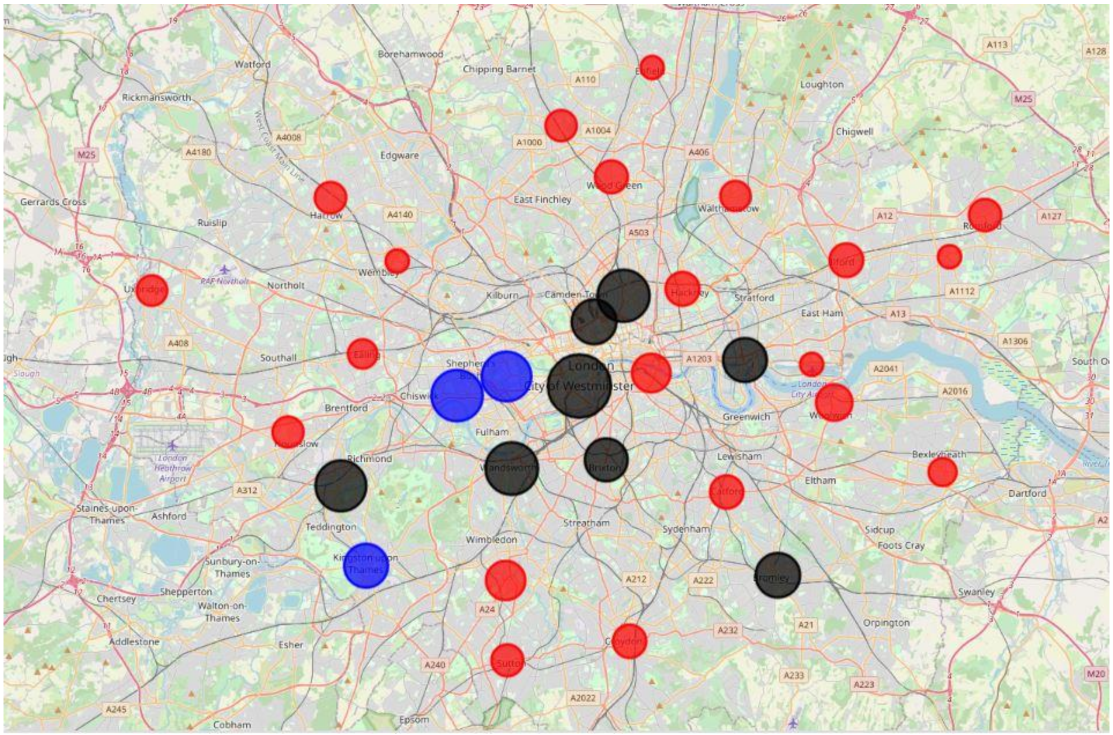

Fig. 3.9 London Map with clustered boroughs (radius of the circle marker is proportional to a net income per person in each borough.)

Let's look at the scatter plots of our data and define our clusters with colors. The grey circle marker is representing the centroid of each cluster. Don't forget, that our data is normalized, so the axes do not deliver real values.

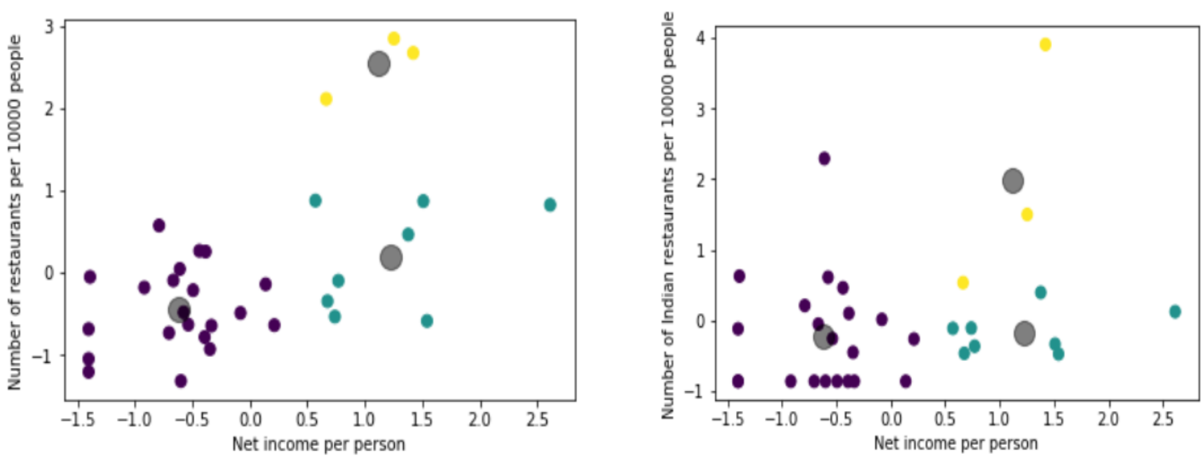

Fig. 3.10 Scatter plots of clustered boroughs with no of restaurants (left) and no of Indian restaurant (right) for every 10000 people

One can observe obvious outlier cluster here. The three boroughs, in this cluster, have too high concentration of restaurants in general and also have medium to high number of Indian restaurants.

Two other clusters were defined according to a net income per person. It’s interesting to see that there is one borough in low income area where number of Indian restaurants are quite high in proportion.

3.2.2 Using Asian population only

Before performing K-mean algorithm, to reduce dimensionality of the problem the columns "Population", "Number of restaurants" and "Number of Indian restaurants" were removed. These three columns were replaced with two new ones, namely, "Number of restaurants per 1000 Asian" and "Number of Indian restaurants per 1000 Asian ".

	Net income per person	Number of restaurants per 1000 Asian	Number of Indian restaurants per 1000 Asian
0	479.1	0.037037	0.000000
1	536.6	0.000000	0.000000
2	513.8	0.647059	0.000000
3	480.0	0.130841	0.037383
4	632.5	0.733333	0.133333

Fig. 3.11 Modified dataframe using Asian population

To identify the optimal number of clusters, the Elbow method is used:

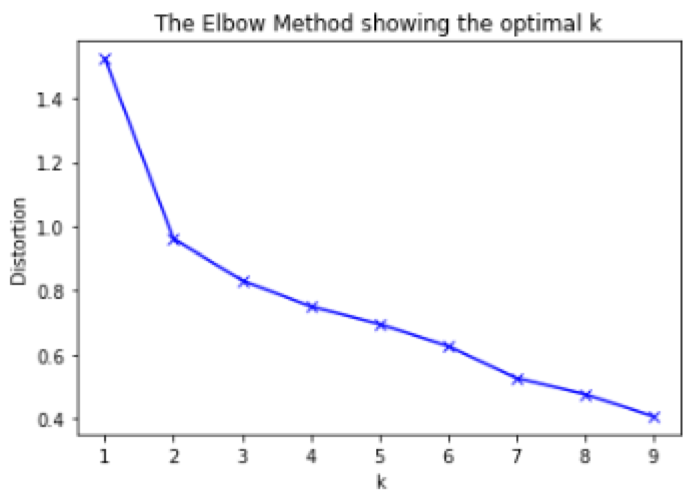

Fig. 3.12 The Elbow method showing the optimal number of clusters

This time unlike previously it is very clear from the graph that two clusters are the best choice.

Based on clustering results, two maps are created. The first map illustrates the clusters where the colors are associated with clusters and radius of the Circle marker is proportional to a number of restaurants per 1000 Asian in each borough. The second map shows the clusters where the radius of the Circle marker is proportional to a Net income per person in each borough.

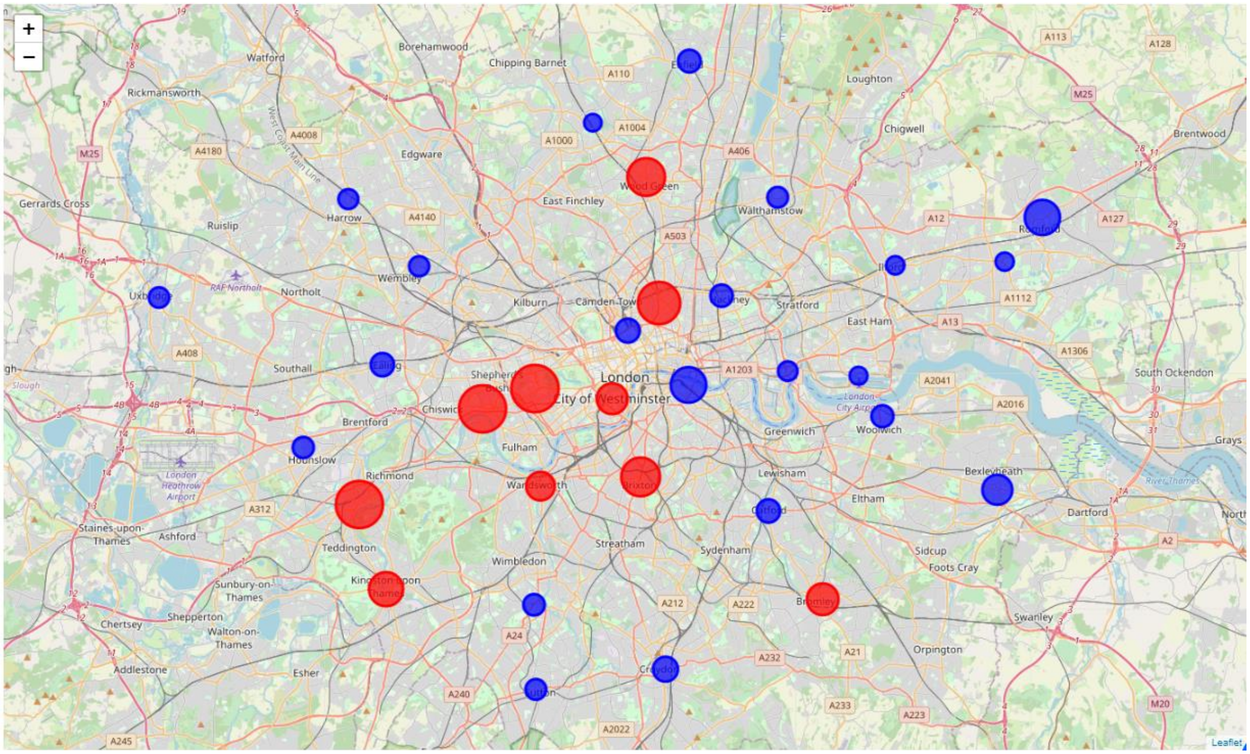

Fig. 3.13 London Map with clustered boroughs (radius of the circle marker is proportional to a number of restaurants per 1000 Asian in each borough)

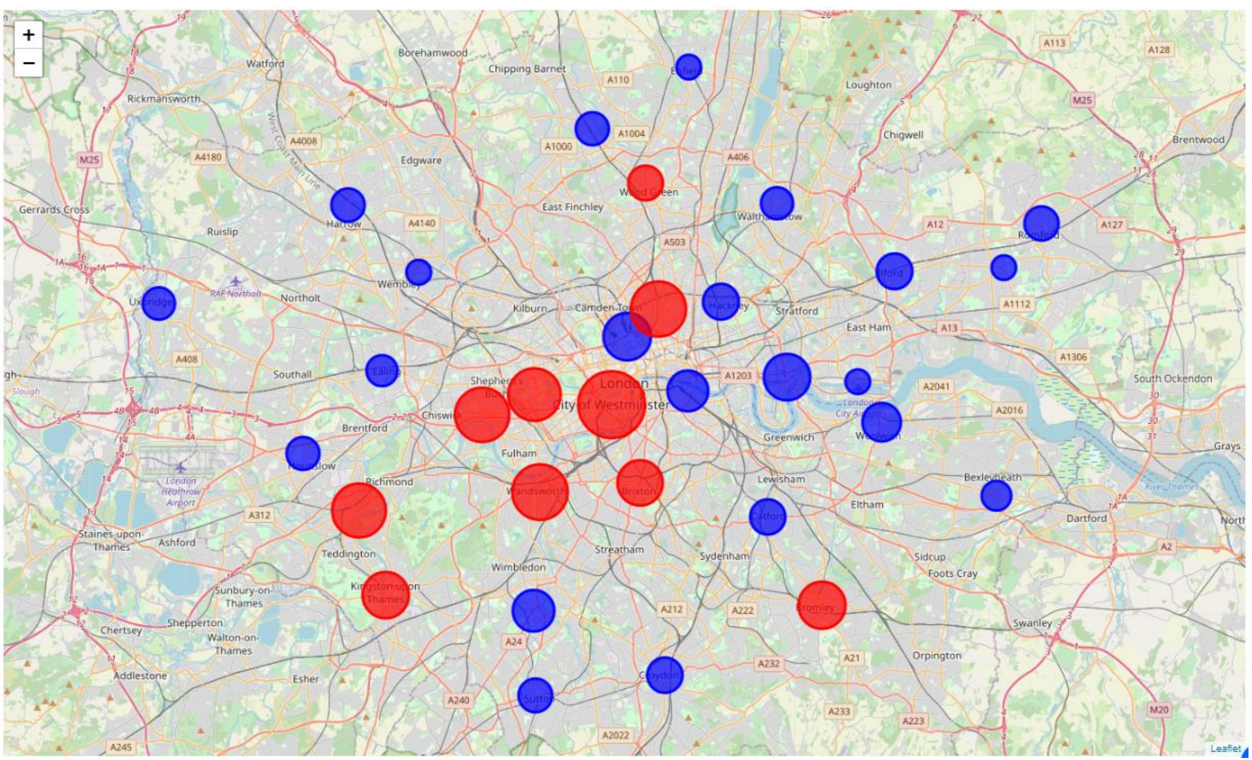

Fig. 3.14 London Map with clustered boroughs (radius of the circle marker is proportional to a net income per person in each borough.)

Let's look at the scatter plots of our data and define our clusters with colors. The grey circle marker is representing the centroid of each cluster. Don't forget, that our data is normalized, so the axes do not deliver real values.

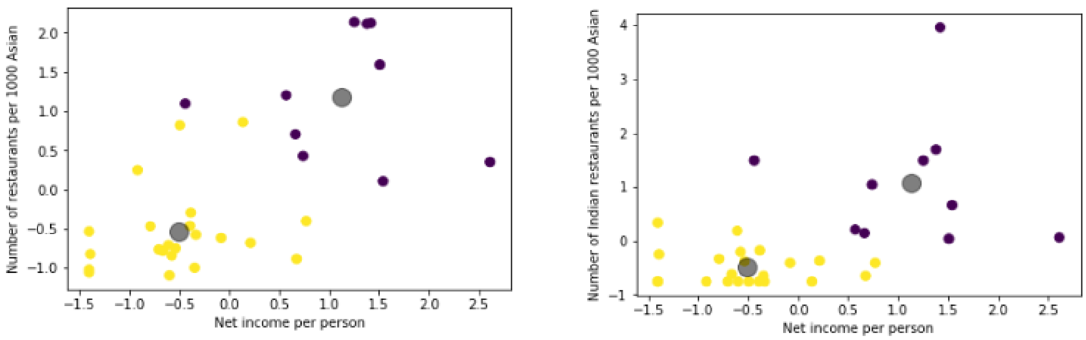

Fig. 3.15 Scatter plots of clustered boroughs with no of restaurants (left) and no of Indian restaurant (right) for every 1000 Asians

One can observe obvious outlier in one of the cluster here especially with respect to Indian restaurant. The outlier borough is Hammersmith and Fulham and definitely not a good choice for Restaurant.

However other Borough in this cluster shows that this can have various potential borough as they have high number of restaurants but not enough Indian restaurants

4. Results and discussion

During the analysis using total population, three clusters were defined. One cluster, which consists of 3 boroughs, has been defined as the outlier, due to the high number of competitors, which means that the placement of Indian restaurant in that area is too risky venture. Two other groups were clustered according to the amount income per person. It is obvious, that the cluster with highest average income per person has the highest priority for us (Cluster 2).

On the other hand during the analysis using only Asian population, two clusters were identified. There is an obvious outlier borough in one of the cluster which definitely will not be good choice for Indian restaurant. However other boroughs in that cluster show good potential as even if there are good amount of restaurants in these boroughs there are not enough Indian restaurant to serve Asian population.

Westminster and Wandsworth are the most attractive options as they are present in favorable clusters in both analysis. They also have relatively high value of income per person. However, one can perform further analysis of this particular cluster with additional features, such as distance to the center of city or to the center of cluster. After defining a borough, one can perform deeper analysis to find the best exact location of the restaurant taking into account factors such as number of parking places in the vicinity of the spot or distances to the main streets.

What could be done better?

Foursquare doesn't represent the full picture, since many venues are not on the list. For that reason, another map could be utilized such as Google map or Openstreet map. Boroughs have too complex geometry, thus defining the closest venues within the certain radius brings additional error to our analysis.

5. Conclusion

To conclude, the basic data analysis was performed to identify the most optimal boroughs for the placement of the Indian restaurant in the London city. During the analysis, several important statistical features of the boroughs were explored and visualized. Furthermore, clustering helped to highlight the group of optimal areas. Finally, Westminster & Wandsworth were chosen as the most attractive options for the further analysis.