Online-Food-Ordering

IBM Data Science Capstone Project

Masoud Ayoubi-Najafabadi

1. TABLE OF CONTENTS

1. INTRODUCTION

A. Business Problem:

Used the Foursquare API and London data to cluster neighborhoods and determine the best location for the delivery workers of an online food ordering company.

B. Background:

Online food ordering is a system to order the restaurants' food through the website or mobile app. Based on the type of this system, a customer can choose a restaurant, scan the menu items, select the desired items and pay it. These websites give customers information about the food, price, duration of the food preparation and so on.

Based on the reports, the online food ordering market have significantly increased all over the world. For instance, 40 percent of Americans having ordered their food online at least once.

There are different types of online food ordering. One of the main ones is managed by online food delivery companies. In this method, these companies receive the costumer order on their website or mobile app and connect the costumer to the restaurant. After that the costumer ordered and payed the food, they can get the food in door. This is done with some delivery workers (normally with motorcycle or bicycle) who working for the online food delivery companies and are supposed to transfer the food from the restaurants to the costumer's house or office. Naturally, these delivery companies work with lots of restaurants.

What is seriously important for the delivery companies is that they should deliver the food as soon as possible, in order to reduce the delivery time, increase customer's satisfaction, and get more orders. Therefore, it is noticeably vital for these companies to arrange their delivery workers to save the time. To do so, the location of the delivery workers should be optimized in order to reduce the distance they have to drive each time.

The idea behind this project is to cluster the neighborhoods of the city (here London is chosen as the case study) to ensure that the workers have to drive less distance in each area. In this project, the data of the restaurants' location and the population of each area of the city is required.

This project is useful for all the online food delivery companies who would like to increase their performance and profit.

2. Data

To answer the business problem, the following data are extracted from the following sources:

- Population of Each Neighborhood
- Number of Restaurants in Each Neighborhood (Foursquare API)
- Coordinates of the Neighborhoods of London

Following data sources will be needed to extract/generate the required information:

- Population of each area will be obtained using **London Borough Profiles Database** (https://data.london.gov.uk/download/london-borough-profiles/c1693b82-68b1-44ee-beb2-3decf17dc1f8/london-borough-profiles.csv)
- Coordinate of London center will be obtained by **parsing the Wikipedia web page: List_of_London_boroughs** (https://en.wikipedia.org/wiki/List_of_London_boroughs)
- Number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**

To start the project, first, it is necessary to download and import the libraries. Then, the data of London borough is downloaded and explored. Here the encoding should "ISO-8859-1" be chosen.

Table 1 Data of London borough

Code	Area_name	Inner/_Outer_London	GLA_Population_Estimate_2017	GLA_Household_Estimate_2017	Inland_Area_(Hectares)	Population_density_(per_hectare)_2017
E09000001	City of London	Inner London	8800	5326	290	30.3
E09000002	Barking and Dagenham	Outer London	209000	78188	3,611	57.9
E09000003	Barnet	Outer London	389600	151423	8,675	44.9
E09000004	Bexley	Outer London	244300	97736	6,058	40.3
E09000005	Brent	Outer London	332100	121048	4,323	76.8
E09000006	Bromley	Outer London	327900	140602	15,013	21.8
E09000007	Camden	Inner London	242500	107654	2,179	111.3
E09000008	Croydon	Outer London	386500	159010	8,650	44.7
E09000009	Ealing	Outer London	351600	132663	5,554	63.3
E09000010	Enfield	Outer London	333000	130328	8,083	41.2
E09000011	Greenwich	Outer London	280100	113964	4,733	59.2
E09000012	Hackney	Inner London	274300	115417	1,905	144
E09000013	Hammersmith and Fulham	Inner London	185300	83552	1,640	113
E09000014	Haringey	Inner London	278000	115608	2,960	93.9

E09000015	Harrow	Outer London	252300	92557	5,046	50
E09000016	Havering	Outer London	254300	104098	11,235	22.6
E09000017	Hillingdon	Outer London	301000	110827	11,570	26
E09000018	Hounslow	Outer London	274200	105887	5,598	49
E09000019	Islington	Inner London	231200	105038	1,486	155.6
E09000020	Kensington and Chelsea	Inner London	159000	80200	1,212	131.1
E09000021	Kingston upon Thames	Outer London	175400	69849	3,726	47.1
E09000022	Lambeth	Inner London	328900	144400	2,681	122.7
E09000023	Lewisham	Inner London	303400	131076	3,515	86.3
E09000024	Merton	Outer London	208100	84201	3,762	55.3
E09000025	Newham	Inner London	342900	119172	3,620	94.7
E09000026	Redbridge	Outer London	304200	110708	5,642	53.9
E09000027	Richmond upon Thames	Outer London	197300	85108	5,741	34.4
E09000028	Southwark	Inner London	314300	134254	2,886	108.9
E09000029	Sutton	Outer London	202600	85243	4,385	46.2
E09000030	Tower Hamlets	Inner London	304000	123720	1,978	153.7
E09000031	Waltham Forest	Outer London	276200	105981	3,881	71.2
E09000032	Wandsworth	Inner London	321000	138149	3,426	93.7
E09000033	Westminster	Inner London	242100	118975	2,149	112.7
E13000001	Inner London		3535700	1522541	31,929	110.7
E13000002	Outer London		5299800	2079422	125,424	42.3
E12000007	London		8835500	3601963	157,215	56.2
E92000001	England		55609600	.	########	.
K02000001	United Kingdom		65999100	.	.	.

As we can see, there are some rows which are not borough name and should be dropped. After cleaning data, it is clear that London has 33 boroughs. After that, it is necessary to extract the required data including Borough Name and Population from the dataset and create a new DataFrame called **Borough** containing each borough's population.

Table 2 Data of Population

	Borough	Population
0	City of London	8800
1	Barking and Dagenham	209000
2	Barnet	389600
3	Bexley	244300
4	Brent	332100
5	Bromley	327900
6	Camden	242500
7	Croydon	386500
8	Ealing	351600
9	Enfield	333000
10	Greenwich	280100
11	Hackney	274300
12	Hammersmith and Fulham	185300
13	Haringey	278000
14	Harrow	252300
15	Havering	254300
16	Hillingdon	301000
17	Hounslow	274200
18	Islington	231200
19	Kensington and Chelsea	159000
20	Kingston upon Thames	175400
21	Lambeth	328900

22	Lewisham	303400
23	Merton	208100
24	Newham	342900
25	Redbridge	304200
26	Richmond upon Thames	197300
27	Southwark	314300
28	Sutton	202600
29	Tower Hamlets	304000
30	Waltham Forest	276200
31	Wandsworth	321000
32	Westminster	242100

Now it's time to get the latitude and the longitude coordinates of each neighborhood. To do so, the list of London boroughs is read from the Wikipedia web page containing coordinates of each borough.

To extract the data of the web page, first, the HTML is parsed using BeautifulSoup. Second, the comments and references of the page are removed from the text. Finally, as the data of this page is provided in two tables, the data is collected from the two tables in this web page, and then cleared.

In the next step, the regular expressions library is applied to extract the values of Latitude and Longitude for each borough. By using the extracted data, the DataFrame of coordinates (called **Coordinates**) is defined and created.

To create the final database called **df**, the data from both dataframes (**Brough** and **Coordinates**) should be merged.

Table 3 Final DataFrame

	Borough	Population	Latitude	Longitude
0	City of London	8800	51.5155	-0.0922
1	Barking and Dagenham	209000	51.5607	0.1557
2	Barnet	389600	51.6252	-0.1517
3	Bexley	244300	51.4549	0.1505
4	Brent	332100	51.5588	-0.2817
5	Bromley	327900	51.4039	0.0198
6	Camden	242500	51.5290	-0.1255
7	Croydon	386500	51.3714	-0.0977
8	Ealing	351600	51.5130	-0.3089
9	Enfield	333000	51.6538	-0.0799
10	Greenwich	280100	51.4892	0.0648
11	Hackney	274300	51.5450	-0.0553
12	Hammersmith and Fulham	185300	51.4927	-0.2339
13	Haringey	278000	51.6000	-0.1119
14	Harrow	252300	51.5898	-0.3346
15	Havering	254300	51.5812	0.1837
16	Hillingdon	301000	51.5441	-0.4760
17	Hounslow	274200	51.4746	-0.3680
18	Islington	231200	51.5416	-0.1022
19	Kensington and Chelsea	159000	51.5020	-0.1947
20	Kingston upon Thames	175400	51.4085	-0.3064
21	Lambeth	328900	51.4607	-0.1163
22	Lewisham	303400	51.4452	-0.0209
23	Merton	208100	51.4014	-0.1958
24	Newham	342900	51.5077	0.0469
25	Redbridge	304200	51.5590	0.0741
26	Richmond upon Thames	197300	51.4479	-0.3260

27	Southwark	314300	51.5035	-0.0804
28	Sutton	202600	51.3618	-0.1945
29	Tower Hamlets	304000	51.5099	-0.0059
30	Waltham Forest	276200	51.5908	-0.0134
31	Wandsworth	321000	51.4567	-0.1910
32	Westminster	242100	51.4973	-0.1372

Now, the geographical coordinates of London should be gotten using **geolocator** which converts an address into latitude and longitude values. After that, the folium library is applied to Visualize the neighborhoods of London.

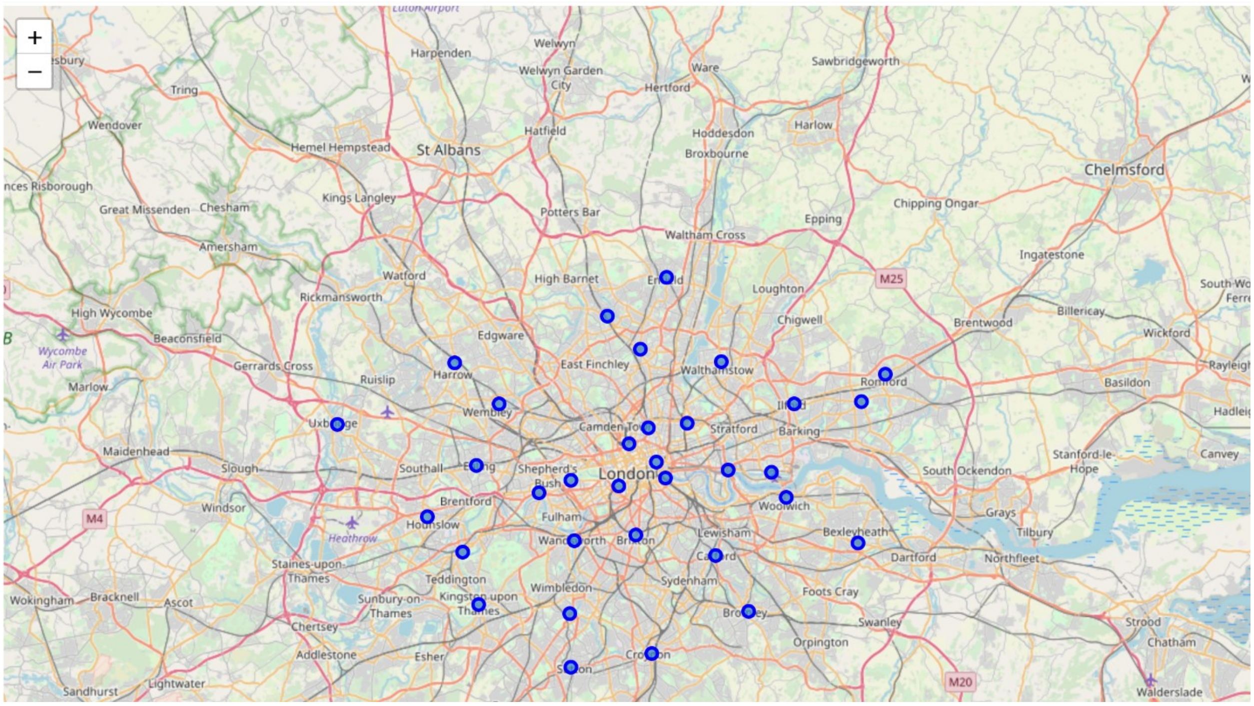

Figure 1 Geographical coordinates of London

The third part of the required data is the data should be extracted from Foursquare by defining a Foursquare Credentials and Version. By exploring the neighborhoods in London, all the restaurants data is extracted by defining a function to look up top 200 places within 2,000 meters of coordinates. Then, the function is applied on each neighborhood and a new dataframe called **London_venues** containing food category venues is created.

Table 4 London food venues

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	City of London	51.5155	-0.0922	Goodman Steak House Restaurant	51.514398	-0.090745	Steakhouse
1	City of London	51.5155	-0.0922	Hawksmoor Guildhall	51.515647	-0.090997	Steakhouse
2	City of London	51.5155	-0.0922	City Càphê	51.514750	-0.091545	Vietnamese Restaurant
3	City of London	51.5155	-0.0922	Pilpel	51.515195	-0.098462	Falafel Restaurant
4	City of London	51.5155	-0.0922	Burger & Lobster	51.513687	-0.094643	Seafood Restaurant

3. Methodology

The scope of this project is to determine the areas of London with most food category venues (restaurant, fast food, café ,...) density, particularly those in crowded areas with high population.

In first step we have collected the required data: location and type (category) of food category venues (according to Foursquare categorization) and population of each area.

Second step in our analysis will be calculation and exploration of 'venue density' across different areas of London.

In the third step, we will focus on most promising areas and within those create clusters of locations. We will take into consideration locations in radius of 2000 meters. We will present map of all such locations but also create clusters (using k-means clustering) of those locations to identify neighborhoods. Then, the scatter plot of population and number of each venue for each cluster is presented.

4. Analysis

At first, one hot encode the unique categories is done to convert categorical variables into a form that could be provided to ML algorithms to do a better job in prediction. Next, we group rows by neighborhood and by taking the mean of the frequency of occurrence of each called **neighborhoods_venues_sorted** and display the category.

In order to put data into a dataframe, first, a function is written to sort the venues in descending order. Then, we create the new dataframe top 10 venues for each neighborhood.

Table 5 Neighborhoods venues sorted

	Neighborhood	Population	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	City of London	8800	51.5155	-0.0922	5	Italian Restaurant	Falafel Restaurant	Food Truck	French Restaurant	Restaurant	Vietnamese Restaurant	Salad Place	Steakhouse	Sushi Restaurant	Seafood Restaurant
1	Barking and Dagenham	209000	51.5607	0.1557	0	Restaurant	Turkish Restaurant	Pizza Place	Chinese Restaurant	Café	Fast Food Restaurant	Breakfast Spot	Fried Chicken Joint	English Restaurant	Doner Restaurant
2	Barnet	389600	51.6252	-0.1517	3	Café	Italian Restaurant	Fish & Chips Shop	Turkish Restaurant	Indian Restaurant	Sushi Restaurant	Fast Food Restaurant	Pizza Place	Chinese Restaurant	Mediterranean Restaurant
3	Bexley	244300	51.4549	0.1505	4	Fast Food Restaurant	Chinese Restaurant	Italian Restaurant	Indian Restaurant	American Restaurant	Bakery	Greek Restaurant	Restaurant	Pizza Place	Sandwich Place
4	Brent	332100	51.5588	-0.2817	2	Indian Restaurant	Fast Food Restaurant	Sandwich Place	Café	Pizza Place	Restaurant	Bakery	Italian Restaurant	Portuguese Restaurant	American Restaurant

In the next step, k-means is executed to cluster the neighborhood into 8 clusters. We create a new dataframe (called **London_merged**) that includes the cluster as well as the top 10 venues for each neighborhood. Finally, let's visualize the resulting clusters using folium.

Now we make a comparison among different clusters based on their population and number of food venues. To do so, we create a dataframe (called **London_venues_count**) to show the total number of food venues in each neighborhood. Also, let's create a new dataframe showing each neighborhood's cluster label. Then, we create the final dataframe of London (called **London_Data**) by merging the previous dataframes.

Table 6 London Data

	Neighborhood	Population	Latitude	Longitude	Cluster Labels	Number of venues
0	City of London	8800	51.5155	-0.0922	5	100
1	Barking and Dagenham	209000	51.5607	0.1557	0	7
2	Barnet	389600	51.6252	-0.1517	3	55
3	Bexley	244300	51.4549	0.1505	4	40
4	Brent	332100	51.5588	-0.2817	2	74
5	Bromley	327900	51.4039	0.0198	1	43
6	Camden	242500	51.5290	-0.1255	5	100
7	Croydon	386500	51.3714	-0.0977	2	79
8	Ealing	351600	51.5130	-0.3089	1	100
9	Enfield	333000	51.6538	-0.0799	2	37
10	Greenwich	280100	51.4892	0.0648	1	45
11	Hackney	274300	51.5450	-0.0553	3	100
12	Hammersmith and Fulham	185300	51.4927	-0.2339	1	100
13	Haringey	278000	51.6000	-0.1119	3	61
14	Harrow	252300	51.5898	-0.3346	2	46
15	Havering	254300	51.5812	0.1837	6	37
16	Hillingdon	301000	51.5441	-0.4760	4	44
17	Hounslow	274200	51.4746	-0.3680	7	51
18	Islington	231200	51.5416	-0.1022	5	100
19	Kensington and Chelsea	159000	51.5020	-0.1947	1	100
20	Kingston upon Thames	175400	51.4085	-0.3064	1	100
21	Lambeth	328900	51.4607	-0.1163	5	100
22	Lewisham	303400	51.4452	-0.0209	3	27
23	Merton	208100	51.4014	-0.1958	1	61
24	Newham	342900	51.5077	0.0469	1	48
25	Redbridge	304200	51.5590	0.0741	2	31
26	Richmond upon Thames	197300	51.4479	-0.3260	1	100
27	Southwark	314300	51.5035	-0.0804	5	100
28	Sutton	202600	51.3618	-0.1945	1	43
29	Tower Hamlets	304000	51.5099	-0.0059	1	100
30	Waltham Forest	276200	51.5908	-0.0134	3	54
31	Wandsworth	321000	51.4567	-0.1910	1	100
32	Westminster	242100	51.4973	-0.1372	5	100

Furthermore, let's see which neighborhoods are in each neighborhoods and create a new dataframe called **London_Data_clustered** grouped by cluster labels.

Table 7 London neighborhoods

Cluster Labels	Population	Number of venues	Neighborhood
0	209000	7	Barking and Dagenham
1	3055200	940	Bromley, Ealing, Greenwich, Hammersmith and Fu...
2	1608100	267	Brent, Croydon, Enfield, Harrow, Redbridge
3	1521500	297	Barnet, Hackney, Haringey, Lewisham, Waltham F...
4	545300	84	Bexley, Hillingdon
5	1367800	600	City of London, Camden, Islington, Lambeth, So...
6	254300	37	Havering
7	274200	51	Hounslow

Finally, we visualize the data provided in this database using bar chart and scatter plot.

5. Results and Discussion

The analysis indicates that although there are 33 boroughs in London having different food venues and population, some of these areas are more crowded in comparison to the other ones. Moreover, the density of the food venues in these areas are completely different. This analysis demonstrates the importance of this survey for the stakeholders. Also, it clarified that this diversity among the areas can widely influence the density and location of delivery workers in the city.

In order to get the best results, the k-means method is used to clustering the neighborhoods into 8 clusters. The areas were illustrated in the map categorizing in the groups with various colors.

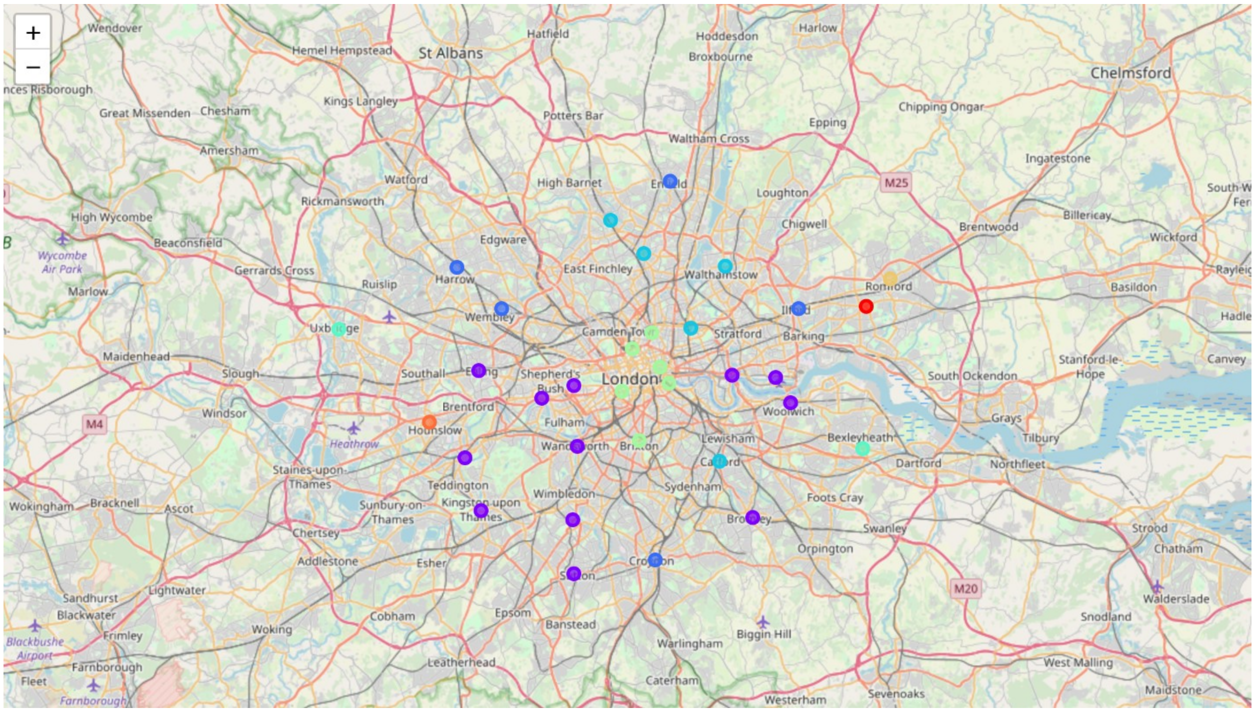

Figure 2 London neighborhoods

As indicated in the "Population" figure, the crowded areas are clusters labeled 3 and 2, while the clusters labeled 4, 6 and 7 have the least population. Obviously, the other clusters are in the middle. If we consider only the population of each area, the results show that the crowded areas could be potentially good places to arrange the most delivery workers, but it is not the only case we should consider.

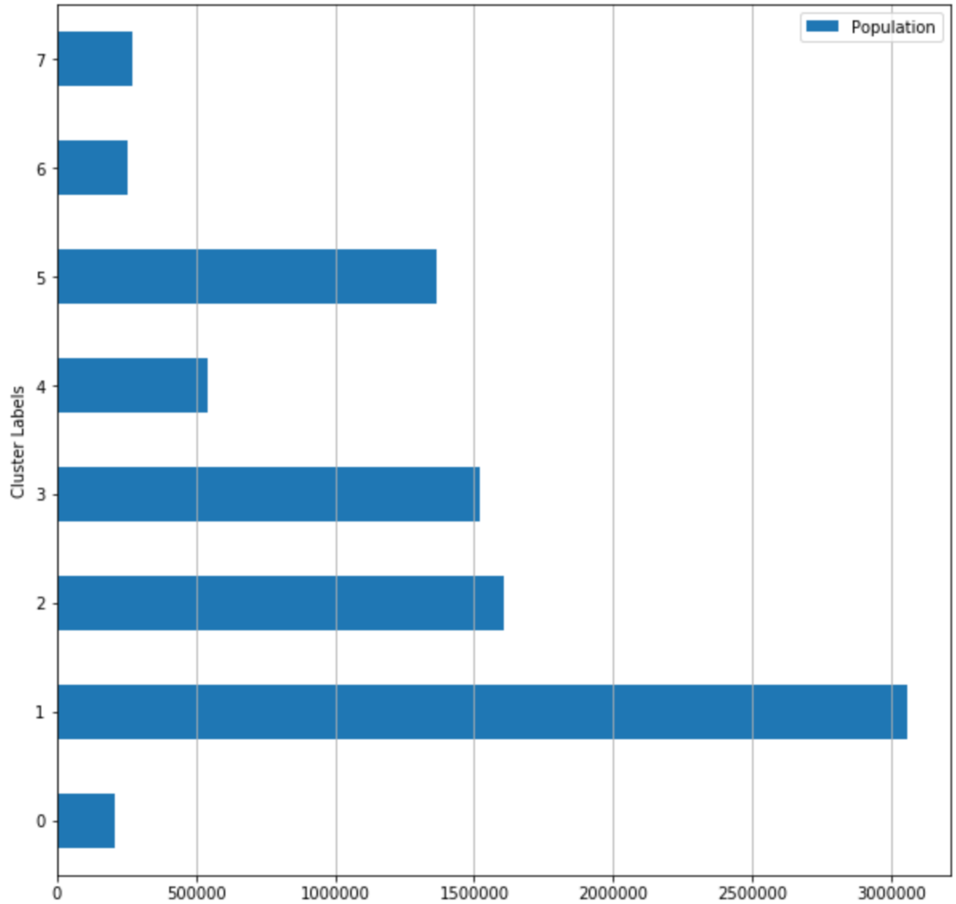

Figure 3 Population of clusters

On the other hand, there is another factor that could be considered that is the number of food venues (restaurants, cafes, fast foods, ...). Generally, we expect that those areas with more venues need more online services. As illustrated in the "Number of Venues" figure, cluster 0 and cluster 3 have the most food venues, while clusters 4, 6 and 7 have the least ones.

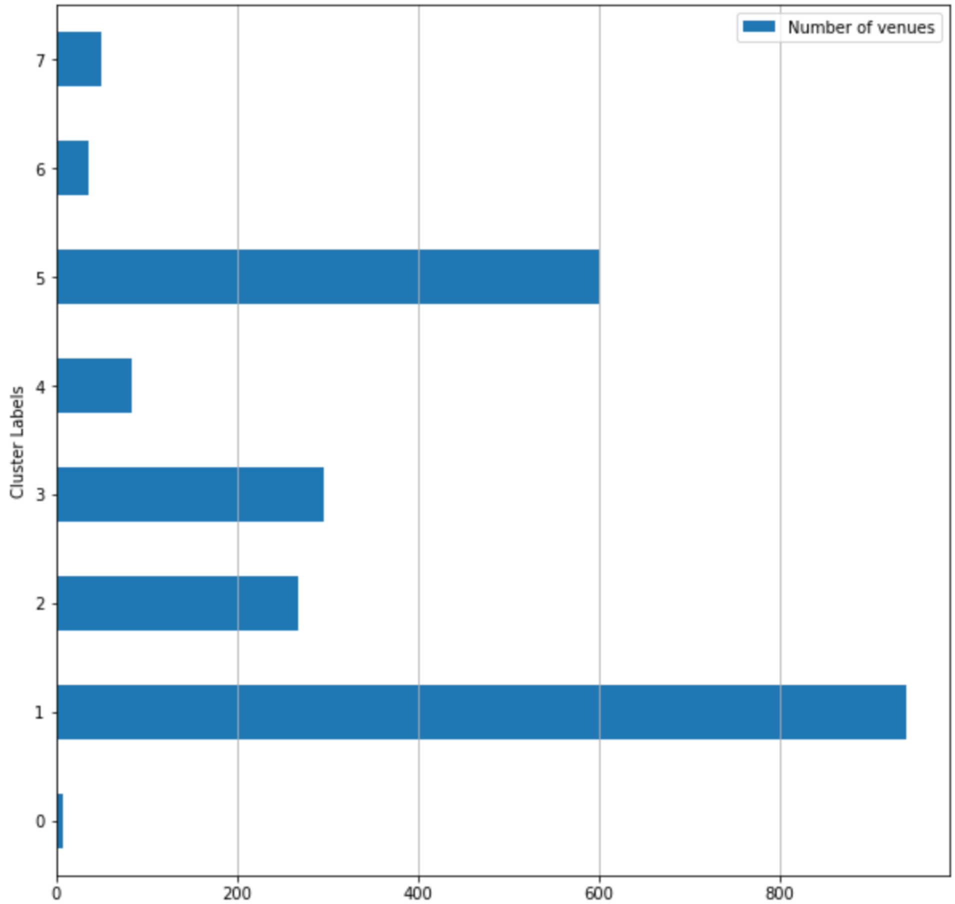

Figure 4 Number of venues in clusters

As we can see, if we consider the two variables ("population" and "number of food venues") one by one, it could give us a good estimation, but not a perfect result. Because, the high number of food venues in an area cannot guarantee that the more people would like use online services, number of food venues itself could not be a variable with a hundred percent accuracy. Moreover, if an area has more population, it does not ensure that always more restaurants and fast foods are there. Therefore, to estimate the tendency of the customers for ordering online food, it is necessary to evaluate both variables in each area together.

The final results provided in the scatter plot reveals the relation between "population" and "number of food venues" in each neighborhood. As we can see, the best place to integrate the workers is cluster 3 (neighborhoods: Bromley, Ealing, Enfield, Hammersmith and Fulham, Kensington and Chelsea, Kingston upon Thames, Merton, Richmond upon Thames, Sutton, Tower Hamlets, Wandsworth). After that, we recommend that the company focuses on cluster 2 (neighborhoods: Bexley, Brent, Croydon, Greenwich, Harrow, Hillingdon, Redbridge), cluster 0 (neighborhoods: City of London, Camden, Islington, Lambeth, Southwark, Westminster) and cluster 5 (neighborhoods: Barnet, Hackney, Haringey, Newham). The other areas are less important.

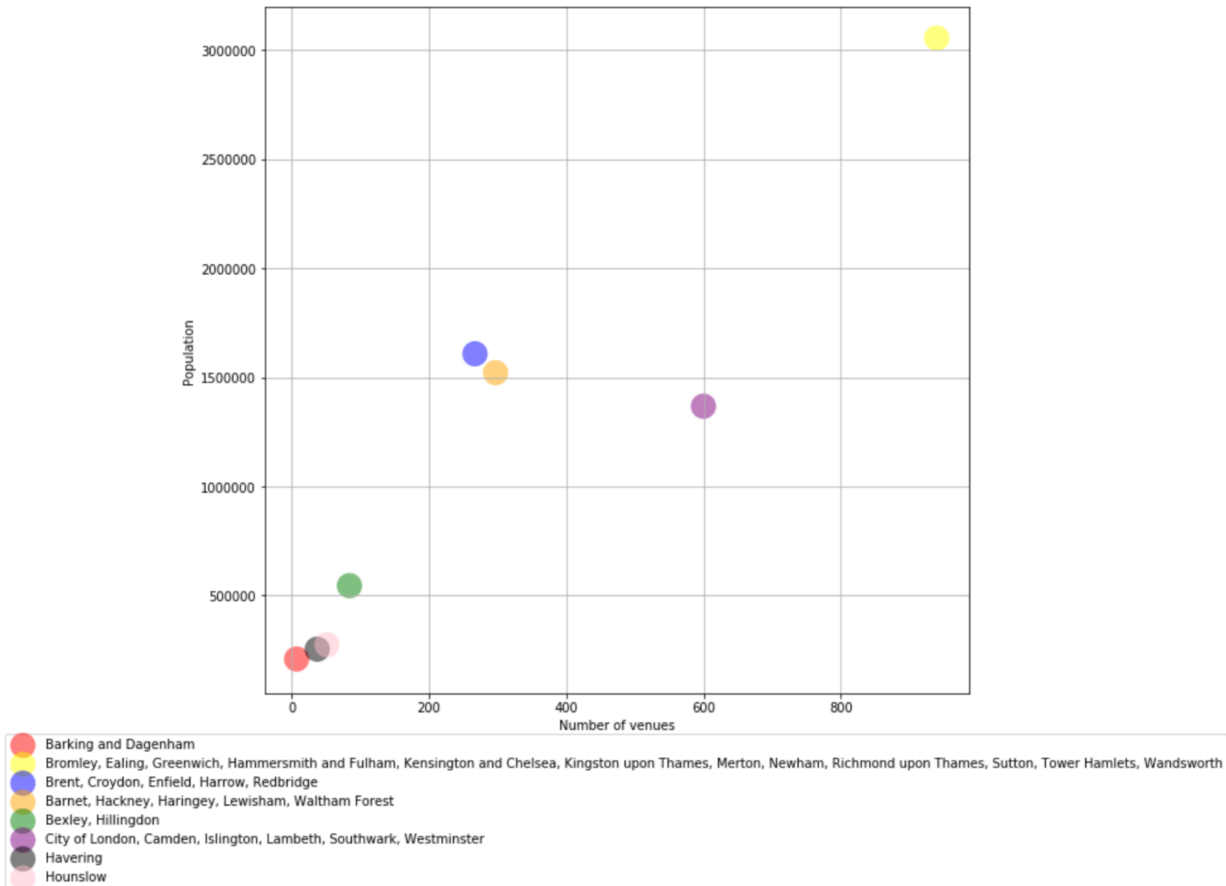

Figure 5 Scatter plot (Population vs Number of venues)

6. CONCLUSION

The scope of this project was to identify London areas with high number of food venues in order to assist online food ordering stakeholders in narrowing down the search for optimal location for arranging delivery workers. By calculating restaurant density distribution from Foursquare data we have first identified general boroughs that justify further analysis, then generated extensive collection of locations considering the population of each area. Clustering of those locations was then performed in order to create major zones of interest.