London foodie map

Magdalena Mitraszewska

June 2021

Contents

1 Introduction 2
1.1 Background 2
1.2 Problem 2
1.3 Interest 2
2 Data 2
2.1 Data sources and acquisition 2
2.2 Data cleaning 2
2.2.1 London neighborhoods 2
2.2.2 Restaurant data 4
3 Methodology 5
3.1 Mapping London neighborhoods 5
3.2 Restaurant data 6
3.2.1 Most common types of restaurants 6
3.3 K-means Clustering 6
4 Results 8
5 Discussion 9
6 Conclusions 10

1 Introduction

1.1 Background

London is one of the most diverse cities in the world – dynamic and multicultural. It's at the same time a financial center of the United Kingdom and one of the most touristic places in Europe. Each year more than 20 million tourists come to London. And everyone needs to eat. The number of restaurants in London is impressing – according to the Office for National Statistics in 2018 there were 11400 restaurants (including takeaways and mobile food stands) in London. This vastness of choice may be confusing, especially for people not living in London. It would be helpful to at least know where to look for a kind of cuisine you like. In which district should I look for Indian restaurants? Where do I find the most choices of Italian cuisine?

1.2 Problem

This project aims to organize London neighborhoods in clusters according to the types of restaurant most common in the area.

1.3 Interest

Such clustering may help people visiting London in choosing the best area to look for a restaurant. It creates a kind of "London foodie map" which divides London according to the most common cuisine. People interested in such map include not only tourists and visiting businessmen, but also people choosing the best area to live. It could be also useful for people interested in opening their own restaurant.

2 Data

2.1 Data sources and acquisition

In order to create restaurant clusters in London I need two types of data:

- restaurant data (mainly the type of cuisine) with their geographical location,
- borough and neighborhood data.

Restaurant data was acquired from Foursquare by using Foursquare API and put into a panda dataframe. The data important to me are mainly type of restaurant and its location.

Borough and neighborhood data is available on the Wikipedia page - https://en.wikipedia.org/wiki/List_of_areas_of_London. I scraped the page, using the BeautifulSoup library and organized the data into another panda dataframe.

2.2 Data cleaning

Both sets of data need preformatting and cleaning.

2.2.1 London neighborhoods

Before getting to Foursquare data I need to prepare a dataframe with desired information about London neighborhoods. Data scraped from Wikipedia is good, but not perfect. Here is the head of the created dataframe:

	Neighborhood	Borough	Post_town
0	Abbey Wood	Bexley, Greenwich [7]	LONDON
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON
2	Addington	Croydon[8]	CROYDON
3	Addiscombe	Croydon[8]	CROYDON
4	Albany Park	Bexley	BEXLEY, SIDCUP

Table 1. London area neighborhoods - raw

As I want to work only with neighborhoods which post town is London, I filtered the data accordingly. I also noticed that in the "Borough" column there are unnecessary numbers and brackets, so I got rid of them, using regular expressions. This is how the head of the dataframe looks like after these steps:

	Neighborhood	Borough	Post_town
0	Abbey Wood	Bexley, Greenwich	LONDON
1	Acton	Ealing, Hammersmith and Fulham	LONDON
2	Aldgate	City	LONDON
3	Aldwych	Westminster	LONDON
4	Anerley	Bromley	LONDON

Table 2. London neighborhoods.

In the dataframe there are 297 rows. It's a lot, so I've decided to work only with the "Inner London" boroughs, which are:

- Camden,
- Greenwich,
- Hackney,
- Hammersmith and Fulham,
- Islington,
- Kensington and Chelsea,
- Lambeth,
- Lewisham,
- Southwark,
- Tower Hamlets,
- Wandsworth,
- Westminster.

After filtering the data accrodingly I get:

	Neighborhood	Borough	Post_town
0	Aldwych	Westminster	LONDON
1	Angel	Islington	LONDON
2	Archway	Islington	LONDON
3	Balham	Wandsworth	LONDON
4	Bankside	Southwark	LONDON

Table 3. Inner London neighborhoods.

Now I have 168 neighbourhoods in the dataframe. That's ok for the further analysis.

2.2.2 Restaurant data

Foursquare offers a great variety of data about venues around the world. What I need is the data about restaurants in London – especially the type of restaurant (cuisine) and their geographical location. To get the data I use Foursquare API and requests library.

First I defined the function that explores the nearby venues of each neighborhood and add it to the dataframe. That's how the head of the dataframe looks like:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Aldwych	51.513103	-0.11492	180 The Strand	51.512671	-0.115009	Art Gallery
1	Aldwych	51.513103	-0.11492	The Delaunay	51.513181	-0.117988	Restaurant
2	Aldwych	51.513103	-0.11492	Twinings	51.513421	-0.112955	Tea Room
3	Aldwych	51.513103	-0.11492	The Pig and Goose	51.513287	-0.113273	French Restaurant
4	Aldwych	51.513103	-0.11492	Temple Brew House	51.512940	-0.113029	Pub

Table 4. London nearby venues.

As I am interested only in restaurants I filtered the data accordingly and I got:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Aldwych	51.513103	-0.11492	The Delaunay	51.513181	-0.117988	Restaurant
3	Aldwych	51.513103	-0.11492	The Pig and Goose	51.513287	-0.113273	French Restaurant
9	Aldwych	51.513103	-0.11492	Roka	51.513312	-0.116194	Japanese Restaurant
24	Aldwych	51.513103	-0.11492	Gaucho	51.514069	-0.111131	Argentinian Restaurant
32	Aldwych	51.513103	-0.11492	Spring	51.510966	-0.118575	Restaurant

Table 5. London nearby restaurants.

I noticed that among venue categories there is a category called "Restaurant". As I am analyzing only restaurants and the type of restaurant is crucial for me this category is non-informative. I've decided to drop it. After gathering and cleaning the data it was ready to start the analysis.

In the final dataset there are 1422 restaurants.

3 Methodology

In this section I am explaining the methods used to cluster the neighborhoods.

3.1 Mapping London neighborhoods

As explained above I scraped the data about geographical areas in London from Wikipedia. I created a dataframe with columns: neighborhood, borough and post town. Next thing I needed was geographical coordinates of each neighborhood. To get that I used geopy library and going row by row added to the dataframe data about longitude and latitude of each neighborhood:

	Neighborhood	Borough	Post_town	Latitude	Longitude
0	Aldwych	Westminster	LONDON	51.513103	-0.11492
1	Angel	Islington	LONDON	51.531842	-0.105714
2	Archway	Islington	LONDON	51.565437	-0.134998
3	Balham	Wandsworth	LONDON	51.444749	-0.151294
4	Bankside	Southwark	LONDON	51.507499	-0.099302

Table 6. Inner London neighborhoods with coordinates.

Next I checked if all the neighborhoods were assigned coordinates and if not, then deleted these neighborhoods from my analysis. I was left with 164 neighborhoods.

To visualize the neighborhoods and check if the coordinates were assigned correctly, I put them on the map using folium library:

3.2 Restaurant data

The data most important for the project is how many restaurants of each type are in each neighborhood. As the data about restaurants is of categorical type I've used one-hot encoding to prepare the data for analysis. The result is put into a dataframe:

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	Brazilian Restaurant	...	R
1	Aldwych	0	0	0	0	0	0	0	0	0	...	
3	Aldwych	0	0	0	0	0	0	0	0	0	...	
9	Aldwych	0	0	0	0	0	0	0	0	0	...	
24	Aldwych	0	0	0	0	1	0	0	0	0	...	
32	Aldwych	0	0	0	0	0	0	0	0	0	...	

Table 7. London restaurants - one-hot encoding.

Next I grouped the results by neighborhood and by the mean of the frequency of occurrence of each category:

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	Brazilian Restaurant	...	S Re
0	Aldwych	0.000000	0.0	0.071429	0.0	0.071429	0.000000	0.000000	0.000000	0.0	...	
1	Angel	0.052632	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.052632	0.0	...	
2	Archway	0.000000	0.0	0.000000	0.0	0.000000	0.142857	0.000000	0.000000	0.0	...	
3	Balham	0.000000	0.0	0.000000	0.0	0.000000	0.090909	0.000000	0.000000	0.0	...	
4	Bankside	0.000000	0.0	0.000000	0.0	0.000000	0.052632	0.052632	0.000000	0.0	...	

Table 8. London restaurants - one-hot encoding grouped.

From exploring the data and previous analysis I know that among restaurant categories there is one called "Restaurant". As I am analyzing only restaurants, this category is non-informative for me, so I have dropped it.

3.2.1 Most common types of restaurants

To explore data further I wrote a function and created a new dataframe to display the top 10 most common types of restaurants for each neighborhood, the head of the dataframe looks like:

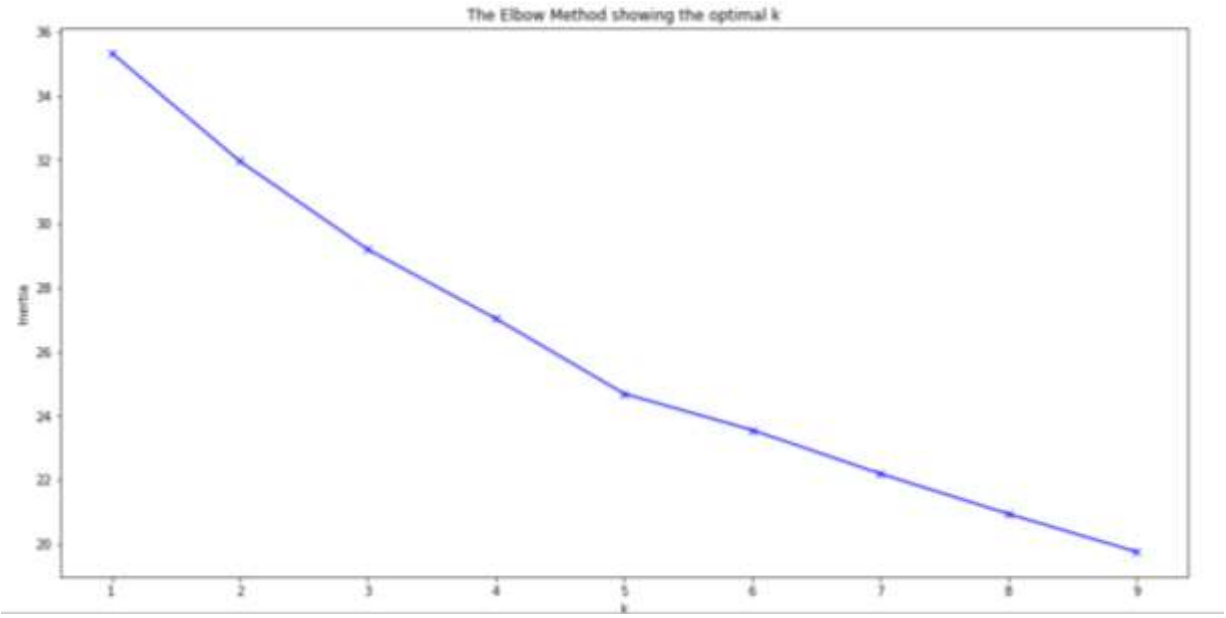
	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
0	Aldwych	Japanese Restaurant	French Restaurant	American Restaurant	Lebanese Restaurant	Argentinian Restaurant	Korean Restaurant	Turkish Restaurant	Italian Restaurant	North Indian Restaurant	Peruvian Restaurant
1	Angel	Mediterranean Restaurant	Vietnamese Restaurant	Indian Restaurant	French Restaurant	Caucasian Restaurant	Mexican Restaurant	Korean Restaurant	Japanese Restaurant	Hunan Restaurant	Sushi Restaurant
2	Archway	Italian Restaurant	Indian Restaurant	Asian Restaurant	Vegetarian / Vegan Restaurant	Kebab Restaurant	Japanese Restaurant	Persian Restaurant	Pakistani Restaurant	Paella Restaurant	Okonomiyaki Restaurant
3	Balham	Italian Restaurant	Indian Restaurant	Moroccan Restaurant	Asian Restaurant	Japanese Restaurant	Portuguese Restaurant	North Indian Restaurant	Peruvian Restaurant	Persian Restaurant	Pakistani Restaurant
4	Bankside	Italian Restaurant	Portuguese Restaurant	Israeli Restaurant	Vietnamese Restaurant	Ramen Restaurant	Asian Restaurant	Australian Restaurant	Modern European Restaurant	Indian Restaurant	Spanish Restaurant

Table 9. London neighborhoods - most common restaurants.

3.3 K-means Clustering

After preparing and exploring the data I have decided to work with K-means Clustering to organize London neighborhoods in clusters according to the types of restaurant most common in the area.

First I needed to decide how many clusters I want to get. In order to do that I used the “elbow method” that shows how the level of inertia changes with different values of k. That’s how it looks like:



In this case the optimal k is not obvious but I have decided to work with 5 clusters – this is the best I can choose basing on the above graph.

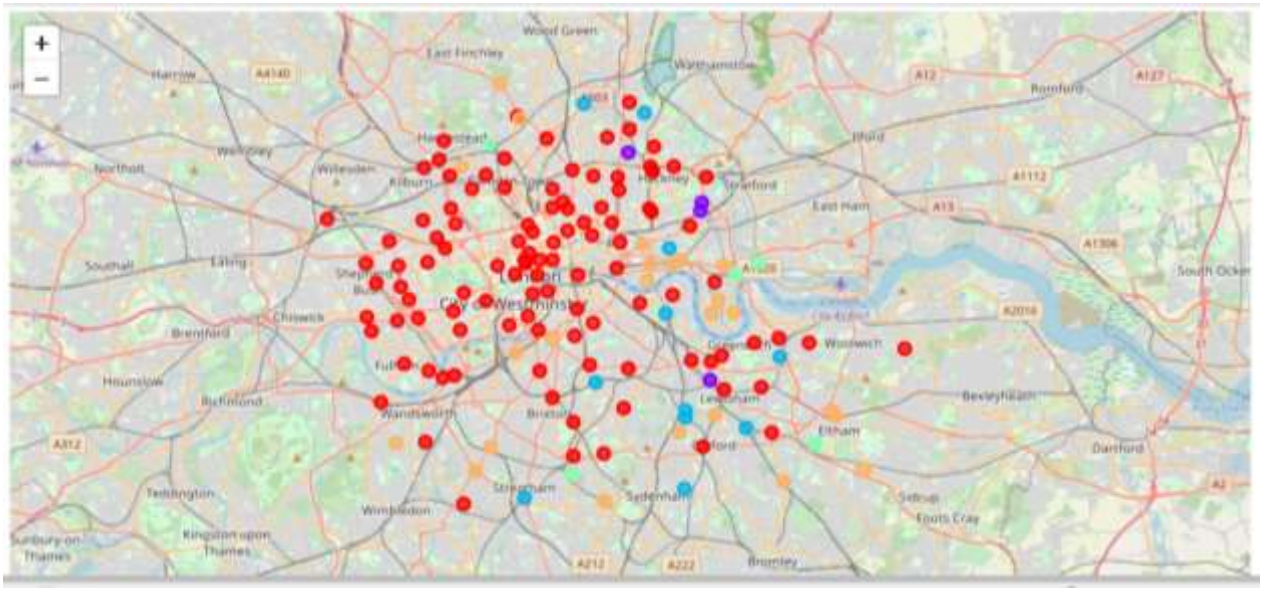
Next step is fitting the model with the data shown in the report in the section 3.2 (Table 8. London restaurants - one-hot encoding grouped).

After fitting the model I have created a new dataframe that includes the cluster as well as the top 10 restaurants for each neighborhood. I took into account only the 152 neighborhoods where restaurants are located. This is what I got at that point (not the whole table):

	Neighborhood	Borough	Post_town	Latitude	Longitude	Cluster_Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant
0	Aldwych	Westminster	LONDON	51.513103	-0.114920	0	Japanese Restaurant	French Restaurant	American Restaurant	Lebanese Restaurant	Argentinian Restaurant	Korean Restaurant	Turkish Restaurant	Italian Restaurant
1	Angel	Islington	LONDON	51.531842	-0.105714	0	Mediterranean Restaurant	Vietnamese Restaurant	Indian Restaurant	French Restaurant	Caucasian Restaurant	Mexican Restaurant	Korean Restaurant	Japanese Restaurant
2	Archway	Islington	LONDON	51.565457	-0.134998	0	Italian Restaurant	Indian Restaurant	Asian Restaurant	Vegetarian / Vegan Restaurant	Kebab Restaurant	Japanese Restaurant	Persian Restaurant	Pakistani Restaurant
3	Balham	Wandsworth	LONDON	51.444749	-0.151294	4	Italian Restaurant	Indian Restaurant	Moroccan Restaurant	Asian Restaurant	Japanese Restaurant	Portuguese Restaurant	North Indian Restaurant	Peruvian Restaurant
4	Bankside	Southwark	LONDON	51.507499	-0.099302	0	Italian Restaurant	Portuguese Restaurant	Israeli Restaurant	Vietnamese Restaurant	Ramen Restaurant	Asian Restaurant	Australian Restaurant	Modern European Restaurant

Table 10. London neighborhoods - most common restaurants with clusters.

Next I visualized the clustered neighborhoods on the map:



4 Results

In this section I am going to analyze the results and explore the created clusters.

First I checked how many neighborhoods are included in each cluster:

Cluster_Labels	Neighborhood
0	108
1	4
2	11
3	5
4	24

In order to analyze the clusters I have added the data about clusters (to which cluster is each neighborhood assigned) to the dataframe with one-hot encoding (joining Table 8 and Table 10). I have dropped all unnecessary columns - I need only Neighborhoods, Cluster Label and restaurant categories. This is what I get:

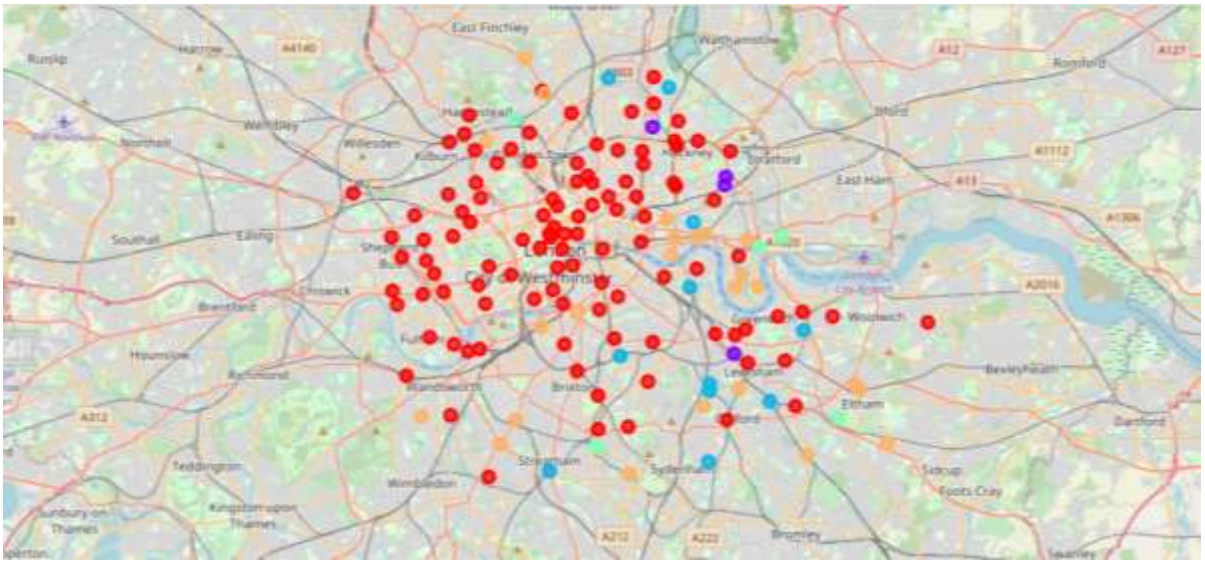
	Neighborhood	Cluster_Labels	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	...	Szechuan Restaurant	Tapas Restaurant	Thai Restaurant	Ti Resta
0	Aldwych	0	0.000000	0.0	0.071429	0.0	0.071429	0.000000	0.000000	0.000000	...	0.0	0.000000	0.000000	0.0
1	Angel	0	0.052632	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.052632	...	0.0	0.000000	0.000000	0.0
2	Archway	0	0.000000	0.0	0.000000	0.0	0.000000	0.142857	0.000000	0.000000	...	0.0	0.000000	0.000000	0.0
3	Balham	4	0.000000	0.0	0.000000	0.0	0.000000	0.090909	0.000000	0.000000	...	0.0	0.000000	0.000000	0.0
4	Bankside	0	0.000000	0.0	0.000000	0.0	0.000000	0.052632	0.052632	0.000000	...	0.0	0.052632	0.052632	0.0

Next I have grouped the data by clusters:

	Cluster_Labels	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	Brazilian Restaurant	...	Szechuan Restaurant	Tapas Restaurant	Thai Restaurant	Turkish Restaurant
0	0	0.000843	0.00154	0.004220	0.004479	0.005164	0.027442	0.0012	0.001116	0.017789	...	0.005479	0.012483	0.054284	0.034658
1	1	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.736111
2	2	0.000000	0.00000	0.022727	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	...	0.000000	0.000000	0.053030	0.060606
3	3	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000
4	4	0.000000	0.00000	0.000000	0.000000	0.000000	0.018371	0.0000	0.000000	0.000000	...	0.000000	0.015625	0.010417	0.029861

To better understand and explore the clusters I have used the function defined before and displayed 10 most common restaurants in each cluster. Together with cluster visualization I've got a good representation of the results:

Cluster_Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
0	Italian Restaurant	Japanese Restaurant	Thai Restaurant	French Restaurant	Indian Restaurant	Middle Eastern Restaurant	Vietnamese Restaurant	Fast Food Restaurant	Turkish Restaurant	English Restaurant
1	Turkish Restaurant	Sri Lankan Restaurant	Cuban Restaurant	Kebab Restaurant	Mediterranean Restaurant	Modern European Restaurant	Afghan Restaurant	Peruvian Restaurant	Persian Restaurant	Pakistani Restaurant
2	Chinese Restaurant	Fast Food Restaurant	Malay Restaurant	Vietnamese Restaurant	Turkish Restaurant	Thai Restaurant	Portuguese Restaurant	American Restaurant	Mamak Restaurant	Okonomiyaki Restaurant
3	Italian Restaurant	Chinese Restaurant	Afghan Restaurant	North Indian Restaurant	Polish Restaurant	Peruvian Restaurant	Persian Restaurant	Pakistani Restaurant	Paella Restaurant	Okonomiyaki Restaurant
4	Indian Restaurant	Italian Restaurant	Chinese Restaurant	Fast Food Restaurant	Turkish Restaurant	Kebab Restaurant	Japanese Restaurant	Korean Restaurant	Asian Restaurant	Tapas Restaurant



5 Discussion

As we can see and read in the Results section above more than 2/3 of the analyzed neighborhoods were assigned to one cluster. This makes me wonder, if K-means is the best algorithm for the task? With neighborhoods dispersed more equally between clusters the "London foodie map" would be more informative. Further analysis and trying other algorithms could be useful.

While analyzing the results one could also notice that it's hard to describe each cluster in a way it could be really informative for regular people. For example, Italian restaurants are no 1 in two clusters, there is some kind of Asian cuisine in almost everyone cluster, etc. Seeing how many different (and particular) types of restaurants there are (77 unique restaurant categories) I'm thinking it would be useful to categorize them into cuisines, for example "Middle Eastern, "Asian", "Mediterranean", etc. Then the difference between clusters would be easier to catch and understand.

6 Conclusions

This project aim was to organize London neighborhoods in clusters according to the types of restaurant most common in the area. This was achieved. Using K-Means Clustering algorithm I created 5 clusters and assigned neighborhoods to them. Each cluster was described by showing 10 most common restaurants in them. Clustering was also visualized on a map.

This tool could still use some work (for example categorizing restaurants into broader terms of cuisine types), but it is already a start for “London foodie map” and could be of interest both for people visiting London as well as businessman thinking about opening a restaurant.