Exploring Food Venues in London

IBM Data Science Project

Paolo Giorgiutti

Introduction

London is the most populous city in the UK and its home of people coming from all around the world (it's the second largest immigration city in the world after NY). Being home of such a variety of culture it led to the developing of a huge variety of traditional restaurant bringing the food culture of different countries in the city.

This project wants to segment the boroughs and neighbourhoods in London into different clusters and examine the distribution of the various ethnic restaurant around the city (it might also be a representation of how the different nationality are spread around the city)

To do so we will leverage on Foursquare API and k-means clustering algorithm.

Data

London data

This London neighbourhood was created to have a starting point with the coordinates for London neighbourhoods. This is a simple .csv file easy to load with pandas library.

Foursquare API

The Foursquare Places API provides location-based data with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag.

Methodology

London Data Download

In order to get the London Neighbourhoods and Boroughs list and their coordinates a .csv file with the coordinate information per neighbourhood was used. This file is downloaded with pandas library that allows to read .csv file from different source and transform them in python dataframes, resulting in the following dataframe:

	Neighborhood	Borough	Post Town	Postcode	Latitude	Longitude
0	Abbey Wood	Bexley, Greenwich	LONDON	SE2	51.49245	0.12127
1	Acton	Hammersmith and Fulham	LONDON	W3	51.51324	-0.26746
2	Aldgate	City	LONDON	EC3	51.51200	-0.08058
3	Aldwych	Westminster	LONDON	WC2	51.51651	-0.11968
4	Anerley	Bromley	LONDON	SE20	51.41009	-0.05683

Fig. 1 (1st 5 rows from London Dataframe)

This dataframe is composed of 46 Boroughs and 297 Neighbourhoods

Then GeoPy library allows us to convert an Address into its Latitude and Longitude in order for us to centre the map, London coordinates: Latitude 51.5073, Longitude -0.1276

Next with the help of folium library we render the map of London showing the location of each neighborhood:

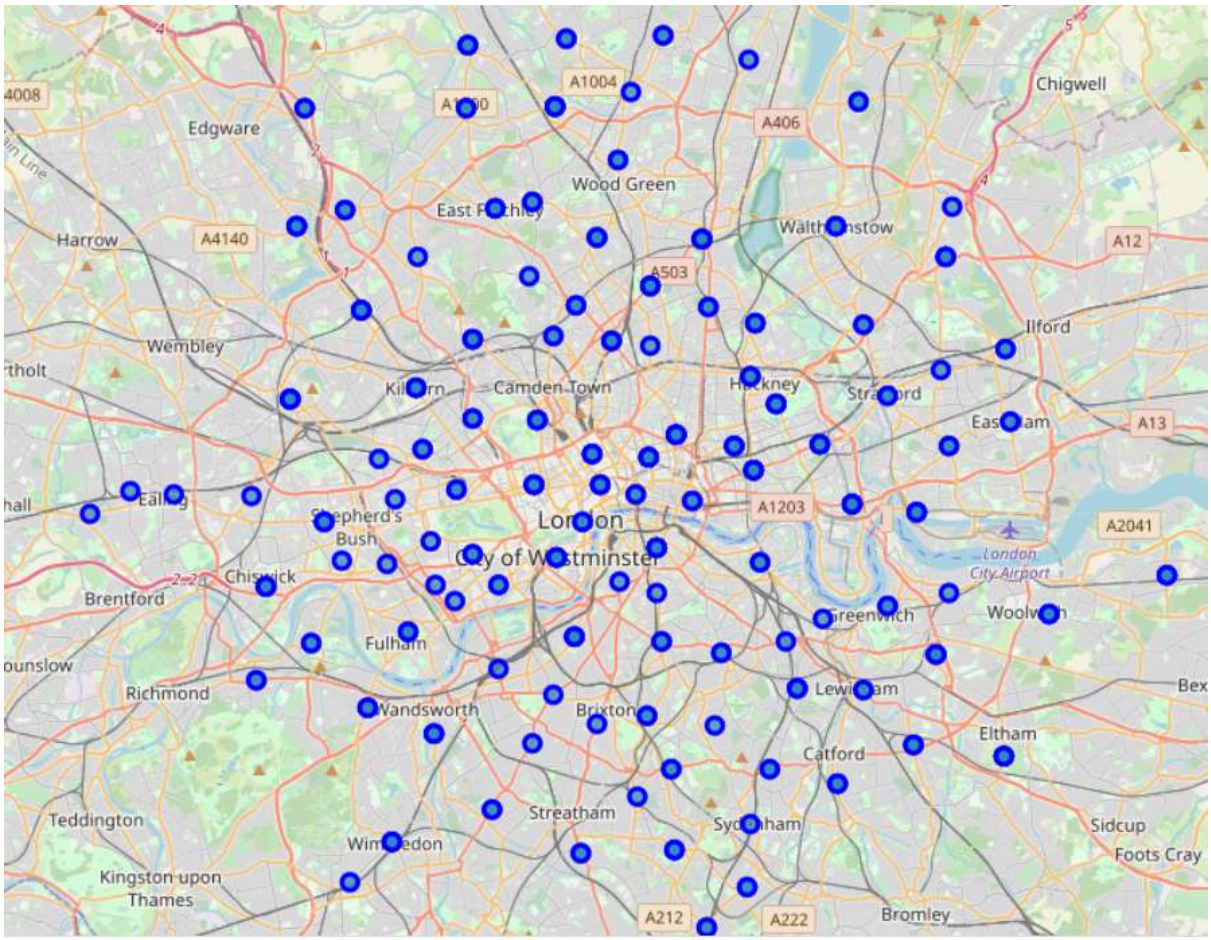

Fig. 2 (London Map)

API calls to Foursquare

The Foursquare API through the developer platform let us explore the neighbourhood and get various information on nearby venues. To access the API, CLIENT_ID and CLIENT_SECRET are needed from a Developer account plus the data VERSION we want to use.

There are many information available in the website but in order to focus the search on food categories we need to look through Foursquare website to find category codes list that are used for each kind of venue (https://developer.foursquare.com/docs/build-with-foursquare/categories/). For this analysis is needed the code for food venue:

Based on CLIENT, SECRET, VERSION and CATEGORY an url is created:

"https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&ll={},{}&v={}&categoryId={}&radius={}&limit={}"

Additional information needed are:

- Latitude
- Longitude
- Radius (radius from the coordinates to look into)
- Limit (max number of venues per coordinates)

For this project the limit of venues returned was set at 200 with a radius of 1500 meters.

In order to avoid repeating the API request manually for all the neighbourhood a function **getNearbyVenues** is created. This function loops through all London neighbourhood in the starting dataset and create an API request based on the information above. The data is then appended to a list and finally is generated a pandas dataframe:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbey Wood	51.49245	0.12127	Greggs	51.490164	0.121305	Bakery
1	Abbey Wood	51.49245	0.12127	Taj Mahal Indian Restaurant	51.491146	0.120691	Indian Restaurant
2	Abbey Wood	51.49245	0.12127	Abbey Cafe	51.489754	0.120822	Café
3	Abbey Wood	51.49245	0.12127	Nom Nom Noms	51.493540	0.109896	Fish & Chips Shop
4	Abbey Wood	51.49245	0.12127	The Crafty Cafe by Sharon	51.487449	0.112696	Café

Fig. 3 (London Dataset with food venues)

Data Cleaning

Since our object is to analyse the food cultural difference around the city, the next step is to remove all the information to generalized food categories that don't represent different culture or food habits (e.g. generic restaurant, café, …)

To manually remove those first we need to find all the unique categorise than remove the generic categorise (with a subtraction between unique list and generic list)

Clustering

In order to cluster the data through machine learning technique some adjustment on the dataframe are required.

The extracted venues are then coded using the one-hot encoding and afterwards grouped by neighbourhood for clustering

As for the clustering method it was decided to use "k-means" clustering, an unsupervised machine learning algorithm that aims to partition the data in k clusters in which observation belong to the cluster with the nearest mean

In order to understand what the optimal number of k is there are two methods, "Elbow method" and "Silhouette Method". These methods are not alternative but rather they are tools to be used together for a more confident decision.

Elbow Method:

This method calculates the sum of the squared distances of the observations to their closest cluster centre considering different values of k. The optimal value is the one after which there is not a significant decrees (elbow shaped curve).

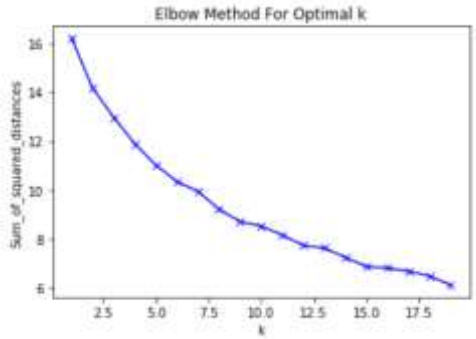

Fig. 4 (k-means "Elbow method" curve)

Unfortunately, not always there are clearly clustered data (such as in our case), that means the elbow method is not sufficient alone to find the optimal k, when there is an ambiguous case we may use the "Silhouette Method".

Silhouette Method:

This method measures the similarity between a point to its own cluster comparing it to other clusters. The range is between -1 and 1 with an high value as the most desirable

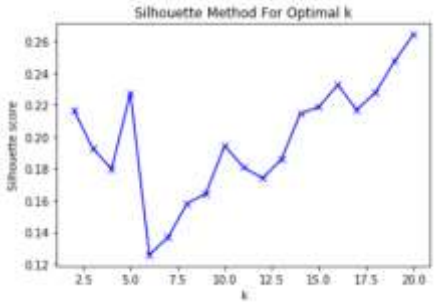

Fig. 5 (k-means "Silhoutte method" curve)

In this case we have spike at k = 5 and while the next highest value is after 15. Since having more the 10 clusters would fragment the data in cluster too small the best choice is k = 5.

Once the optimal k was defined we run the k-means algorithm and split the data in 5 different clusters. The cluster labels are then added together at the food venue and neighbourhood dataframe and rendered on a folium map:

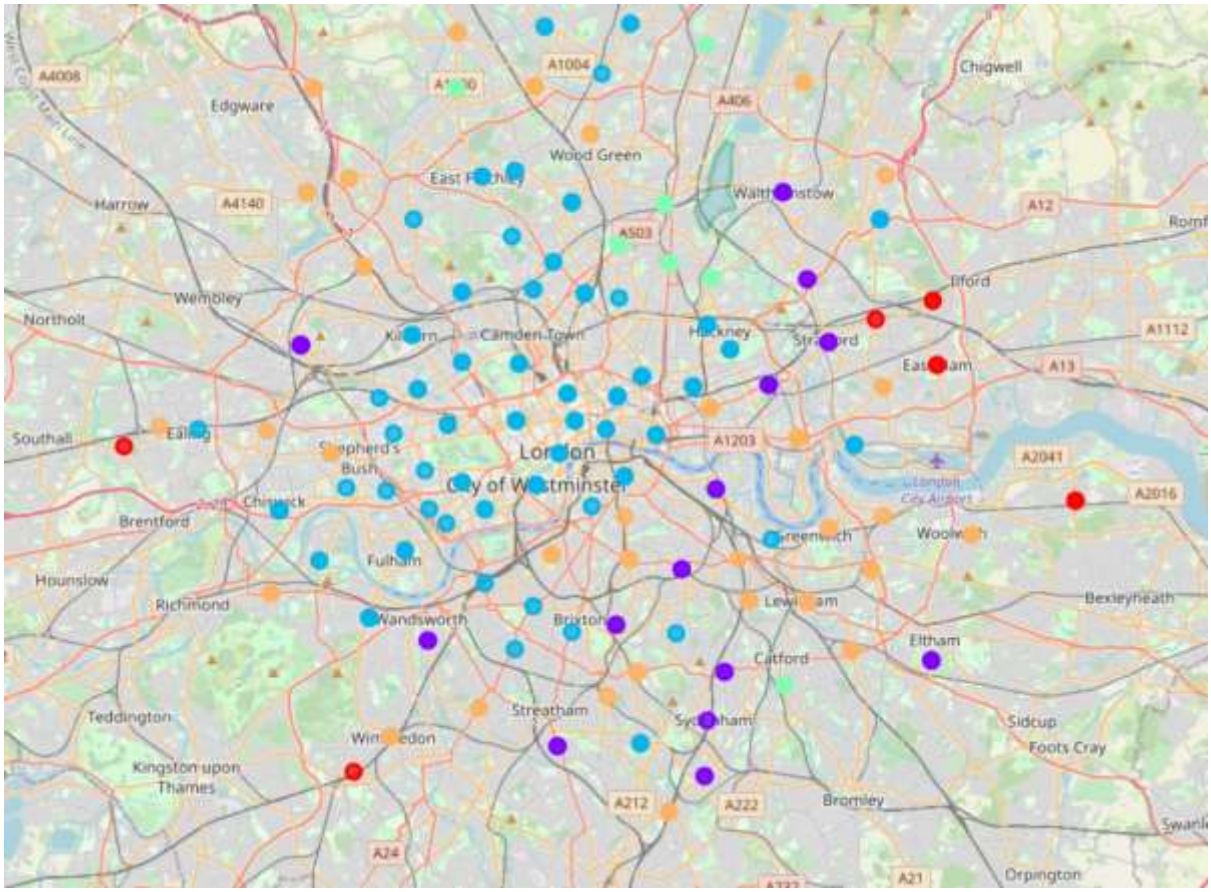

Fig. 5 (London clustering: ***Cluster 0 - Indian/Fish & Chips***, ***Cluster 1 – Pizza***, ***Cluster 2 – Italian***, ***Cluster 3 – Turkish***, ***Cluster 4 - Indian***)

Results

Cluster – 0

The 1st place as most common is taken by Indian restaurant (7 venues) followed by Fish & Chips store. In this cluster there many other restaurant type.

```
Indian Restaurant    7
Fish & Chips Shop    3
Name: 1st Most Common Restaurant, dtype: int64
----------------------------------------
Chinese Restaurant    4
Fish & Chips Shop     2
Turkish Restaurant    2
Italian Restaurant    1
Asian Restaurant      1
Name: 2nd Most Common Restaurant, dtype: int64
----------------------------------------
Newham                5
Bexley                2
Ealing                1
Bexley,  Greenwich    1
Merton                1
Name: Borough, dtype: int64
----------------------------------------
```

Cluster 0 is the Indian restaurant and Fish & Chips store

Cluster – 1

Pizza place are the 1st most common restaurant (24 venues) followed by Romanian restaurant (8 venues). Indian and Chinese restaurant take the 1st and 2nd spot as the 2nd most common but with only 17 venues combined.

```
Pizza Place                24
Romanian Restaurant         8
Mediterranean Restaurant    5
Chinese Restaurant          2
Name: 1st Most Common Restaurant, dtype: int64
----------------------------------------
Indian Restaurant       10
Chinese Restaurant       7
Italian Restaurant       6
Thai Restaurant          5
Portuguese Restaurant    5
Caribbean Restaurant     2
Steakhouse               2
Fish & Chips Shop        2
Name: 2nd Most Common Restaurant, dtype: int64
----------------------------------------
Brent                    7
Southwark                4
Waltham Forest           4
Greenwich                4
Tower Hamlets            3
Wandsworth               3
Lambeth                  3
Bromley                  3
Lewisham                 2
Newham                   2
Croydon                  1
Lewisham, Bromley        1
Hackney                  1
Hammersmith and Fulham   1
Name: Borough, dtype: int64
----------------------------------------
```

Cluster 1 is the Pizza place cluster

Cluster – 2

This cluster is by fare the biggest and take the whole city centre and the majority of the west. It's dominate by Italian cuisine with Italian restaurant and Pizza place topping both 1st and 2nd most common.

Cluster 2 is the Italian Food cluster

Cluster – 3

This cluster most common restaurant is the Turkish with 17 venues as 1st most common. But looking at the total number of venues we see that Italian style places are the most abundant in this cluster

Cluster 3 is the Turkish/Italian

Cluster – 4

This cluster 1st place is taken by Indian restaurants, which top both 1st most common and 2nd most common categories.

Cluster 4 is the Indian cluster.

Discussion

The city of London was clustered through the k-means method and segmented in 6 clusters. With further qualitative analysis we were able to rename the clusters based on the type of restaurant/food store. From the data it appears that the predominant cuisines in London are Italian and Indian.

Cluster	1st Most Common	2nd Most Common
0	Indian	Fish & Chips
1	Pizza	Indian
2	Italian	Italian
3	Turkish	Pizza
4	Indian	Indian

Cluster 0 - Indian/Fish & Chips

Cluster 1 - Pizza

Cluster 2 - Italian

Cluster 3 - Turkish

Cluster 4 - Indian

Conclusion

The neighbourhood of London food categories were clustered and segmented into 5 clusters and renamed based on the most common type of cuisine each of the cluster contained. The most predominant cuisines in London are Indian, Italian and Chinese, but there are present many other types of ethnic restaurants.

This project might be the starting for a deeper analysis of the cultural distribution in the city or to analyse what's the best area to open a certain type of restaurant