Tastes of London

MAY-2020

Coursera Capstone Project
Authored by: Natalia Pleczynska

Introduction and Problem Statement

Background

With 9 million residents and over 40,000 restaurants, London is a large and diverse marketplace. Every year, thousands of business owners open new venues across the city, including restaurants, shops, entertainment centers and others.

Gathering information on all neighborhoods to determine ones with suitable characteristics to open a business can be very time consuming and inefficient. This analysis helps to simplify this process by using clustering techniques to efficiently segment neighborhoods to clusters of similar food tastes.

Problem Statement

The aim of this project is to improve understanding of customer tastes in different parts of London to help business owners determine suitable locations for their business, primarily focusing on restaurants and food-related retail.

"Where to open restaurant or food-related retail business in London?"

Data

List of areas in London with their geographical coordinates

We extracted list **of 119 London district postcodes** and their geographical coordinates from external websites linked below. They have been combined into 'Postcode districts' csv file used for this project.

List of London district postcodes:
https://www.maps.thehunthouse.com/Streets/London_Postal_District _and_Area_Name_finding_aid.htm

Coordinates of London district postcodes:
https://www.doogal.co.uk/UKPostcodesCSV.ashx?area=London

Data on restaurants located in each area of London

We have used Foursquare API to obtain data on restaurants in each of postcode district.

Foursquare API in a location data provider that stores data on variety of venues, including parks, entertainment venues, shopping malls, restaurants etc.

We have structured our Foursquare API query in such a way that only relevant venues are extracted, i.e. venues that are within **Category 'Food'.**

Methodology

Data Preparation: Areas of London

Data with areas of London was relatively clean.

We just needed to perform one transformation to make it fit for analysis: combine all rows with different areas within same district postcode to form one row. **Groupby** function was used to do it.

Data Preparation: Foursquare API

Our Foursquare API query had the following characteristics:

Category ID:	**'Food'**
Radius:	**1,000 meters**
Limit:	**100 venues**

After receiving the results, we look at their main characteristics. We note that there are **133 unique restaurant types** in the dataset. For almost all (109 out of 119) postcode districts there are more than 10 restaurants, which indicates that there is **enough volume of data to proceed** with the analysis.

Methodology

Neighborhood Analysis

We use **'one-hot encoding'** function from pandas library in Python to convert categorical variables on restaurants into the format that can be used by machine learning algorithm.

Further, we calculate mean frequency of occurrence of each restaurant type. We arrange the data to generate dataframe with **10 most common restaurant types** for each area of London.

	Neighborhood	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
0	E1	Indian Restaurant	Café	Pizza Place	Sandwich Place	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bakery	Italian Restaurant
1	E10	Café	Chinese Restaurant	Asian Restaurant	Fish & Chips Shop	Restaurant	Fried Chicken Joint	Indian Restaurant	Sandwich Place
2	E11	Café	Fast Food Restaurant	Pizza Place	Chinese Restaurant	Sandwich Place	Thai Restaurant	Italian Restaurant	Bakery
3	E12	Indian Restaurant	Restaurant	Donut Shop	Falafel Restaurant	Ethiopian Restaurant	English Restaurant	Empanada Restaurant	Eastern European Restaurant
4	E13	Café	Indian Restaurant	Bakery	Vietnamese Restaurant	African Restaurant	Asian Restaurant	Fried Chicken Joint	Falafel Restaurant

Methodology

K-Means Clustering

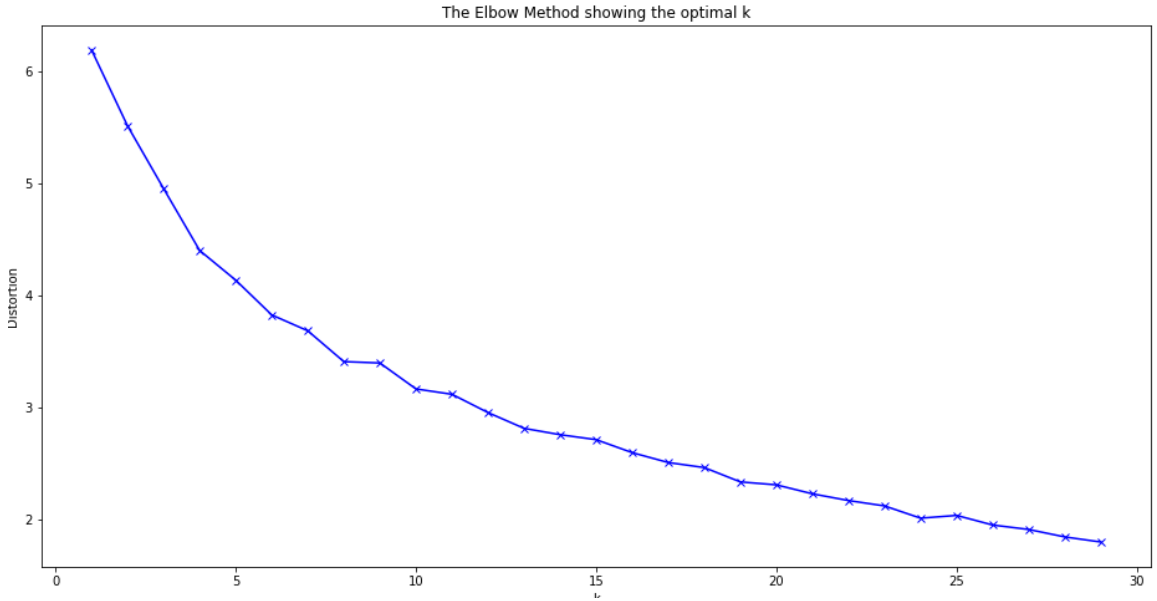
We use **K-means algorithm** to cluster the data. It is unsupervised machine learning method that creates clusters of data points aggregated together because of certain similarities.

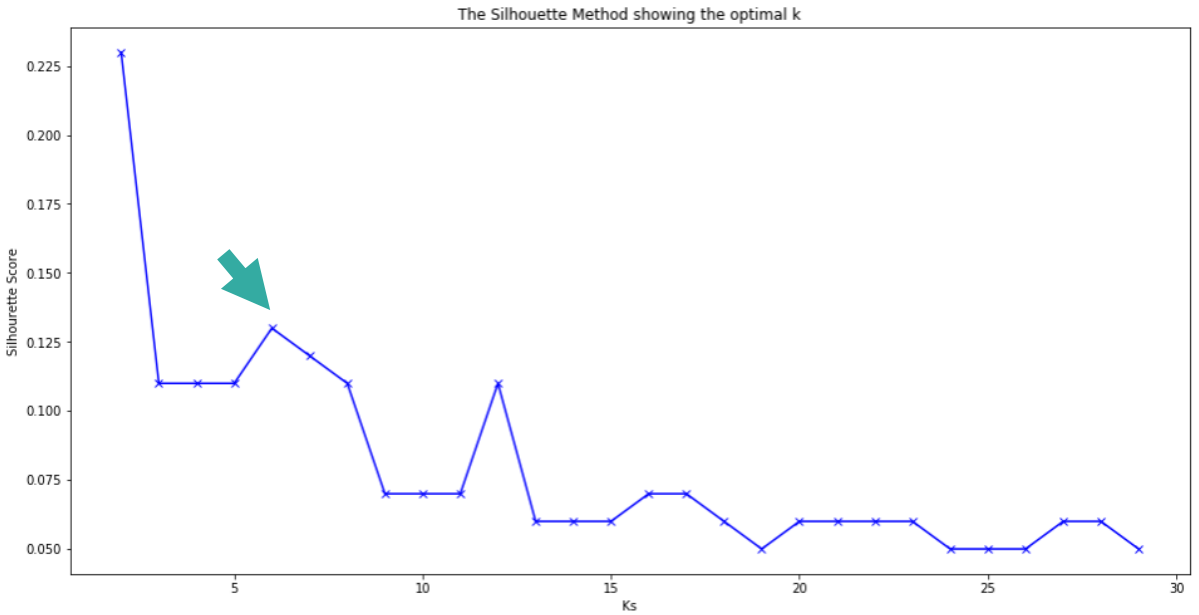
Clustering is performed with use of K-means method from sklearn.cluster library, with use of **6 clusters.** Number of 6 clusters has been determined based on **Elbow and Silhouette methods** (see next slide).

Dataframe with postcode districts and cluster information is shown to the right.

	Postcode	Area Name	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
0	E1	Aldgate, Bethnal Green, City of London, Mile E...	51.5163	-0.060423	1	Indian Restaurant	Café	Pizza Place	Sandwich Place	Vegetarian / Vegan Restaurant	Fast Food Restaurant
1	E10	Leyton	51.5687	-0.012773	0	Café	Chinese Restaurant	Asian Restaurant	Fish & Chips Shop	Restaurant	Fried Chicken Joint
2	E11	Leyton, Leytonstone, Wanstead	51.5674	0.011669	0	Café	Fast Food Restaurant	Pizza Place	Chinese Restaurant	Sandwich Place	Thai Restaurant
3	E12	East Ham, Manor Park, Wanstead	51.5510	0.050806	4	Indian Restaurant	Restaurant	Donut Shop	Falafel Restaurant	Ethiopian Restaurant	English Restaurant
4	E13	Plaistow, West Ham	51.5282	0.025794	3	Café	Indian Restaurant	Bakery	Vietnamese Restaurant	African Restaurant	Asian Restaurant

Methodology





Elbow Method

Plot does not have obvious ‘elbow’ and therefore method does not clearly point to optimal number of clusters.

Silhouette Method

The higher the silhouette score, the higher quality of clustering. Based on that method, we determine that optimal number of clusters for our analysis will be 6 *.

* Technically, 2 clusters have higher score, but considering size and variety of London, dividing it into 2 clusters is unlikely to provide enough differentiation.

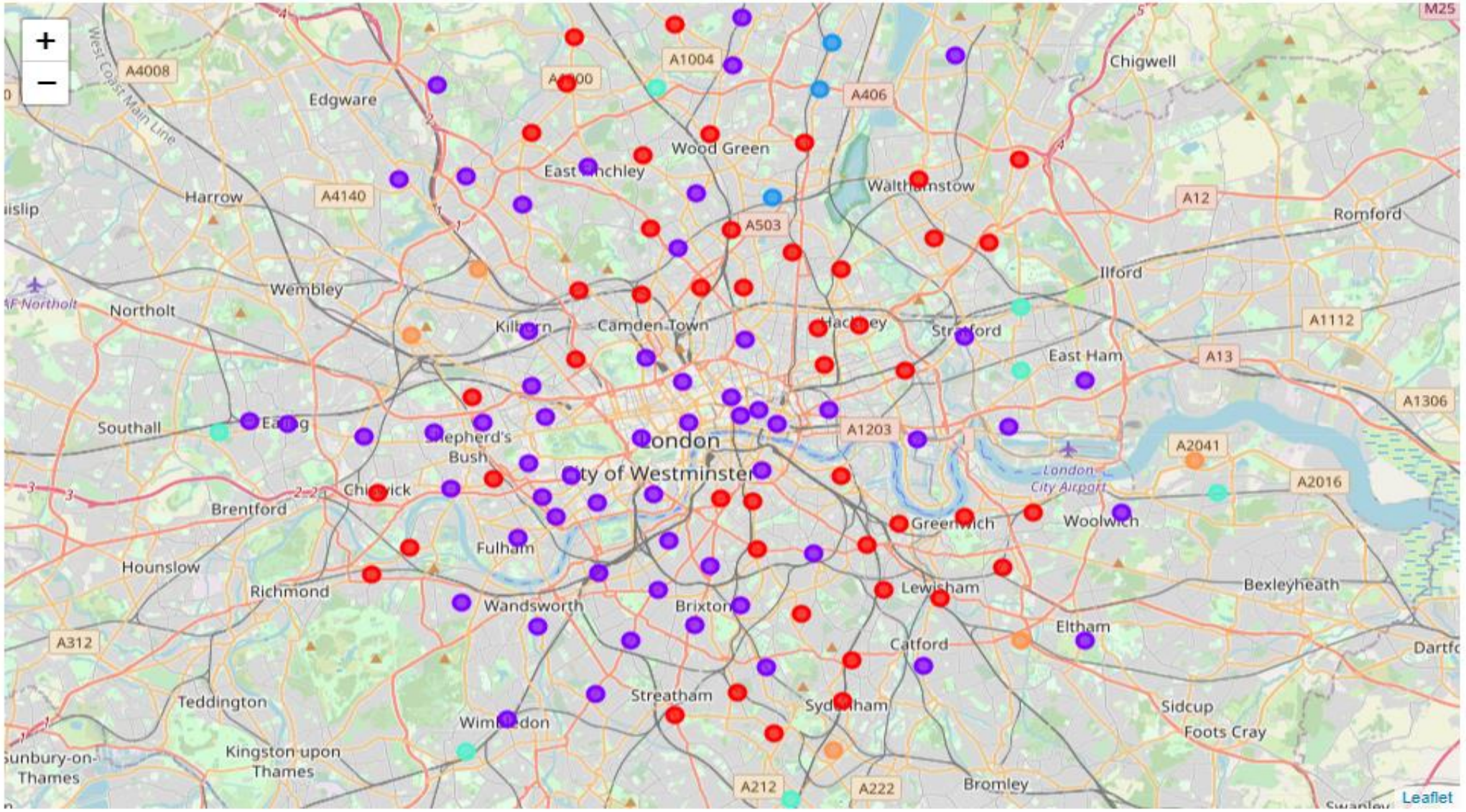
Data Visualization - Clusters

K-means Visualization

Results of K-means clustering are shown to the right.

Each dot represents one neighborhood and dots with same colours are part of same cluster.

Cluster No.	Color
0	Red
1	Purple
2	Blue
3	Turquoise
4	Light Green
5	Orange



Cluster Analysis

Cluster 0 'Affordable restaurants in residential areas' - Red

Cluster zero has 46 neighborhoods. As it can be seen from the map above, it primarily includes locations in East London. Cafes dominate in all postcodes as most common venue. Pizza places, fast food restaurants and bakeries are also high on the list of most common venues. Location and most common venues indicate that:

- those are primarily residential areas
- price range of venues is likely to be relatively low due to their nature

Those could be good locations for business owners looking to set up businesses targeting day-to-day spending done close to home.

Cluster 1 'Higher-end restaurants' – Purple

Cluster one has 57 neighborhoods. As it can be seen from the map above, it primarily includes locations in West London. While cafes are still frequent as most common venue, a significant proportion of neighborhoods has a restaurant as their most common venue. The proportion of restaurants versus bakeries/sandwich places/pizza places is greater than in cluster 1. This indicates that:

- those are primarily in the city center and more wealthy residential areas
- price range of venues is likely to be relatively high due to their nature

Those could be good locations for business owners looking to set up businesses targeting higher end of the market. Any modern/fusion kitchen establishments are also most likely to find customer base in those areas.

Cluster Analysis

Cluster 2 ‘Turkish restaurants’ – Blue

Cluster two has 3 neighborhoods. They are all in relative proximity to each other, north of the city center. As it can be seen below, Turkish restaurants are most common venue in all 3 neighborhoods. While it is likely that the Turkish restaurant market may be saturated in those locations, any businesses that think their products would appeal to same customer base that like Turkish restaurants, should consider those neighborhoods as potentially attractive locations.

Cluster 3 ‘Ethnic cuisine’ – Turquoise

Cluster three has 7 neighborhoods. While located in different parts if the city, they are all clearly in the outskirts. As it can be seen in the list below, those locations have large variety of exotic cuisines from different parts of the world. Those locations can be interesting for business owners looking to set up restaurants that serve ethnic cuisine, as clearly demand for such products is present in those areas.

Cluster 4 *Not named* - Light Green

Cluster four has 1 neighborhood. This is likely to be an outlier and it is difficult to draw any conclusions based on sample of 1. Thus, we do not draw any definitive conclusions.

Cluster 5 ‘Day-to-day, quick purchases’ - Orange

Cluster five has 5 neighborhoods. Similar to cluster 3, while located in different parts if the city, they are all clearly on the outskirts. Fast food restaurants and cafes dominate those areas. These seem to be residential areas, with venues that service day-to-day consumption of residents. Price range of venues is likely to be relatively low.

Conclusions

Analysis of clusters can form a useful insight for businesses looking for locations with particular customer base characteristics.

Our analysis has led us to the distinction of the clusters outlined to the right.

Application of K-means algorithm allowed us to perform analysis of large volume of multi-dimensional data in an effective way and simplify it into easily digestible form.

Cluster 0 ‘Affordable restaurants in residential areas’

Cluster 1 ‘Higher-end restaurants’

Cluster 2 ‘Turkish restaurants’

Cluster 3 ‘Ethnic cuisine’

Cluster 4 *Not named*

Cluster 5 ‘Day-to-day, quick purchases’