Tastes of London

MAY-2020

Coursera Capstone Project

Authored by: Natalia Pleczynska

Introduction

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. When considering new business one of critical variables to make it successful is clear definition of target customer base and setting up the business in a way that allows to effectively capture it.

For restaurants and shops, location is key. There are multiple venues of those types across the city, thus to effectively reach target customers, venue needs to be in proximity of their home, work or location that they can conveniently visit.

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.

London is a large city with many neighborhoods, so gathering information on all neighborhoods to determine ones with suitable characteristics would 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. This is followed by analysis of clusters of neighborhoods to get a better understanding of food habits and tastes in each cluster.

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

Stakeholders

While key users of this information are considered to be business owners looking to set up restaurant or food-related retail business, the analysis can be also used by other stakeholders interested in segmenting London into areas with similar food tastes, such as data analysts performing studies of London market or governments looking to improve understanding of food industry in different areas of London.

Data

Key data used for this problem include:

- List of areas in London with their geographical coordinates
- Data on restaurants located in each area of London

List of areas in London with their geographical coordinates

We have downloaded list of London district postcodes from external website:
https://www.maps.thehunthouse.com/Streets/London_Postal_District _and_Area_ Name_finding_aid.htm

We have obtained coordinates of London district postcodes from external website:
https://www.doogal.co.uk/UKPostcodesCSV.ashx?area=London

We combined list of London district postcodes with coordinates in excel and saved down as csv file that was used for analysis:
'Postcode districts'

This list of district postcodes divides London into 119 areas. This was deemed to be appropriate way to divide London into neighborhoods. Use of boroughs was considered but rejected as it would only divide London into 32 areas. Considering large size and diversity of London, this would result in aggregation of potentially very different

neighborhoods into 1 area (borough) and thus reduce usefulness of analysis. Use of individual postocodes was also considered, however it would result in thousands of areas, which would inhibit the ability to perform meaningful analysis, with potentially few restaurants in each cluster, not reflecting the proportion of restaurants in the wider neighborhood. District postcodes were considered a measure that provides a good balance between those trade-offs.

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

Before we start analysis, we need to ensure that format of data on areas of London is fit for analysis. We note that they are stored in a table in csv file, with column names. That makes it straightforward to download them as dataframe in pandas in Python. Data has information on district postcode, area name and latitude and longitude, which are all data points that we need. There are no NaN values in any of the columns. The only modification that we need to do relates to duplicates. Some district postcodes have multiple rows, which outline different areas of district postcode. As our analysis is done at district postcode level, those rows will be effectively duplicates. We use .groupby method to combine multiple areas with same district postcode into one row and change content of 'Area Name' column to include names of all areas located in this postcode, separated by commas.

Data Preparation: Foursquare API

We create a function that generates API request, makes GET request and saves down information on venues as .json file for each postcode district. Obtained data is then transformed into dataframe.

API request specifies that only venues within 'Food' category should be provided (Food Category ID: 4d4b7105d754a06374d8125). Food Category ID can be obtained from documentation on Endpoints on Foursquare website. Further, we specified radius of our query as 1,000, resulting venues within 1,000 metres of coordinates for each London postcode district being returned. API query has a limit of 100 venues, which means that for each postcode district maximum of 100 venues will be downloaded.

After receiving the results, we look at their main characteristics. We note that there are 133 unique restaurant types in the dataset. We check count number of venues in each postcode district. 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.

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.

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.

K-means algorithm requires user to specify number of clusters that algorithm will use. We use Elbow and Silhouette methods to help us determine optimal number of clusters. Elbow method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. In our case, plot does not have obvious 'elbow' and therefore method does not clearly point to optimal number of clusters. Silhouette method is a way to measure how similar an object is to its own cluster compared to other clusters. 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.

Clustering is performed with use of K-means method from sklearn.cluster library, with use of 6 clusters. Resulting clusters are added to the dataframe with postcode districts, geographical coordinates and most common restaurants types.

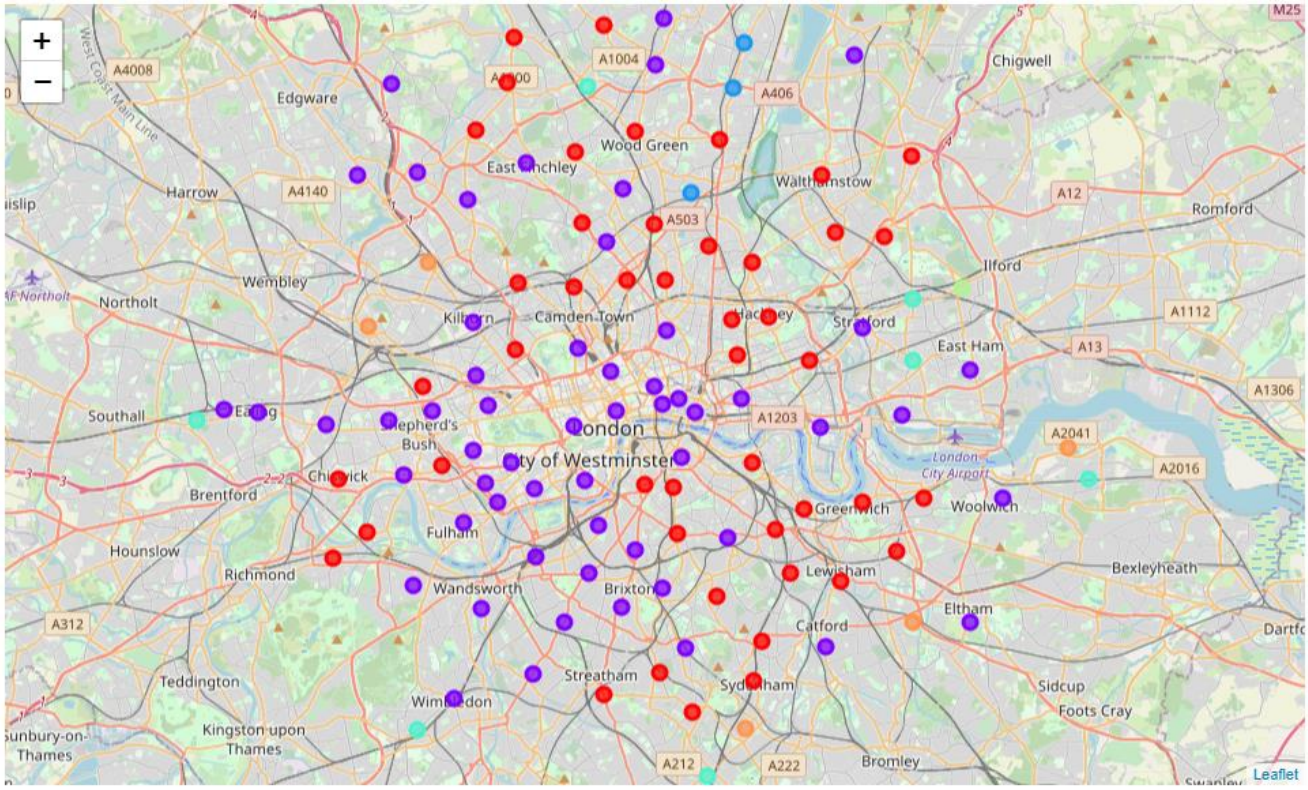
Data Visualization

We use folium library to create visualization of clusters on a map of London. Each postcode district is represented by a dot on a map, with postcode districts belonging to the same cluster having the same colour.

Results and Discussion

Clusters created by K-means algorithm are shown on the map below. Our analysis is performed below.

Cluster Number	Cluster Color on the map	No. of areas in the cluster
0	Red	46
1	Purple	57
2	Blue	3
3	Turquoise	7
4	Light Green	1
5	Orange	5



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 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 based on that cluster.

Cluster 5 'Day-to-day, quick purchases' - Orange

Cluster five has 5 neighborhoods. Similarly to cluster 3, while located in different parts if the city, they are all clearly on the outskirts. Fast food restaurants and cafes seem to 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 due to their nature.

Conclusions and Further Work

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 below. 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'

Information on tastes in neighborhoods could be further enhanced by enriching it with other informative data, specifically:

- Average income of residents in the area
- Demographic characteristics of residents in the area
- Growth of the area (measured, for example, by number of openings of new venues)