**Analyzing the Neighborhoods in London for Starting a Café**

Dimitrios Christodoulou

Coursera Capstone Project

## Introduction

London is the capital city of the United Kingdom and it is one of the most densely populated cities in the globe. It is located at the south part of the country and attracts numerous tourists for all over the world. Also, many clothing stores are located in the city center. To be more precise, there many shops from small stores to luxurious fashion designers stores. And as a result a coffee shop that serve tai is an essential part of the city culture. Starting a new business can be a really demanding process and requires a vast amount of money to invest in advance, so the selection of the store location can have a significant impact on the prosperity of the store. This project can be useful for people who are interesting in opening a new coffee store in London since this project focus to determine possible locations for starting a café.

## Data Collection

The following data is required for the project:

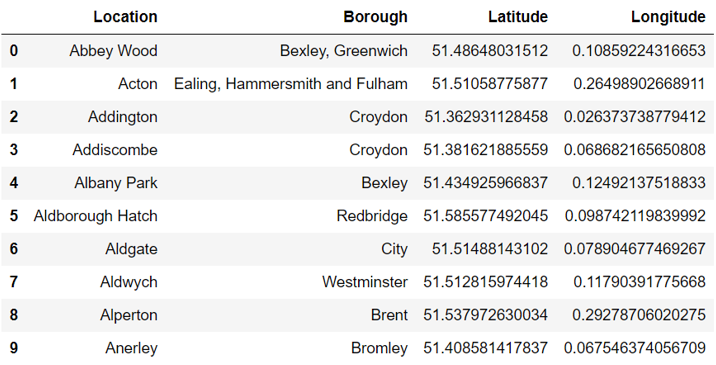
1) Location and Boroughs of London.

2)Geographical coordinates of all locations in London.

3)Venue data for each location.

### Locations Data

The data of the location Boroughs in London was scraped from the Wikipedia (URL: <https://en.wikipedia.org/wiki/List_of_areas_of_London>). To read the data the read\_html() method was used to read into a pandas data. The Wikipedia page provides a comprehensive and detailed table which includes all the required information for this project.



### Geographical Coordinates

The geographical coordinates for London location has been obtained from Wikipedia and more especially from GeoHack which is used from many large companies. This data will be used to plot the locations on the map using Folium library and get the venues tips.

## Venue Data

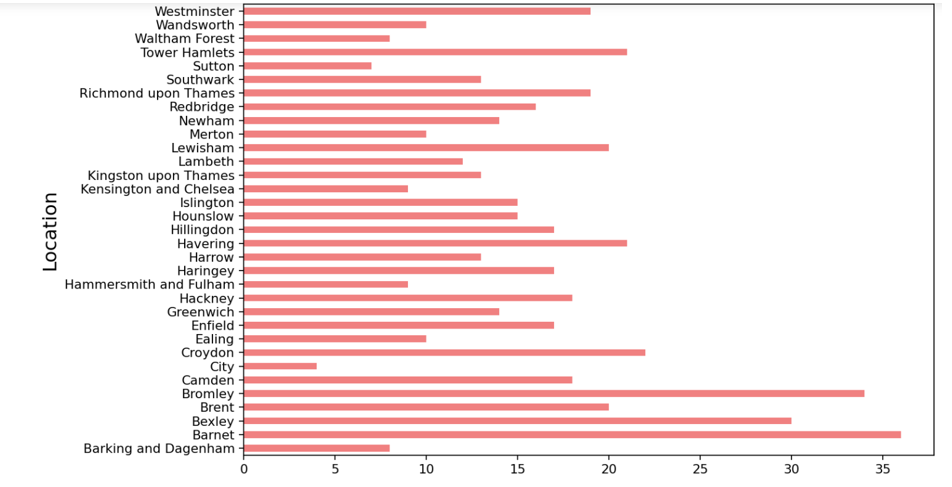
The venue data has been extracted using the Foursquare API. This data contains venue recommendations for each location in London and is used to study the popular venues of different location by building an unsupervised learning model to cluster locations. The venue recommendations of all locations were obtained with a limit of 200, that is, maximum of 200 venue recommendations per location and a radius of 1000m around the location’s coordinates.

## Methodology

This section provides details for the methodology used in the project.

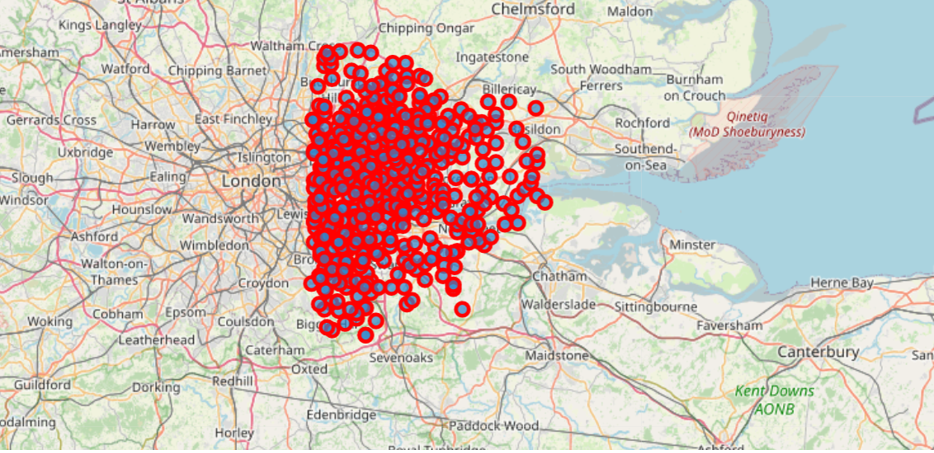
### Data Visualization

In order to understand the data obtained for east London boroughs, basic visualization was carried out.



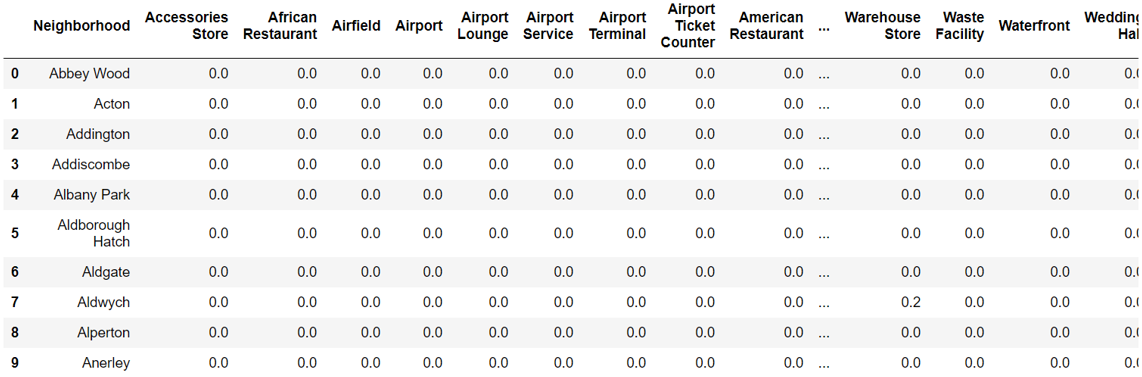
We can conclude that Barnet, Bexley and Bromley have the most number of Neighborhoods.

Using folium, locations was plotted to a map, as shown in the following figure.



## Feature Extraction

Feature extraction was carried out to obtain features from the Foursquare API data which was used for building the unsupervised learning model. In order to achieve this, the Venue Category column had to be converted to some form of numeric value to be used for building the model. This was achieved by the One-hot Encoding method which takes all the unique categories and creates a column for each category. Then, if a location venue belongs to that category get a value of 1 for that row and column and if a locations venue does not belong to the particular category, the value is 0.

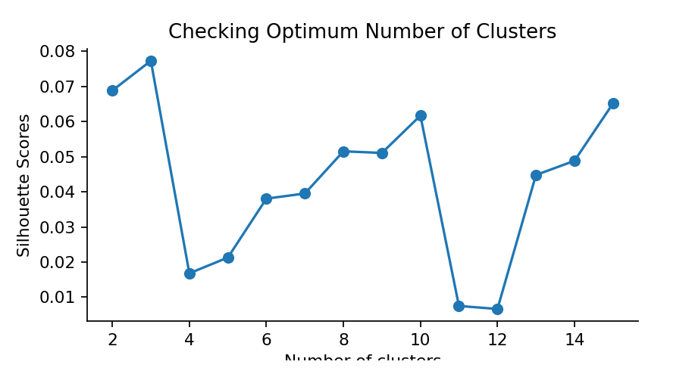


Also, a dataframe containing the top 10 most common venues of all locations was created. It is important to provide a glimpse into what this dataframe looks like as it will be used later to combine the results from the unsupervised learning model. The following figure shown the 10 top venues of some locations.



## Unsupervised Learning

K-means unsupervised learning technique was used to cluster the locations based on the category of venues near the locations. One important aspect of the k-means model is to determine the number of clusters to use in model development. This was determined by the Silhouette score which was calculated for a range of clusters from 2 to 15. The resulting number of clusters are shown in the following figure.

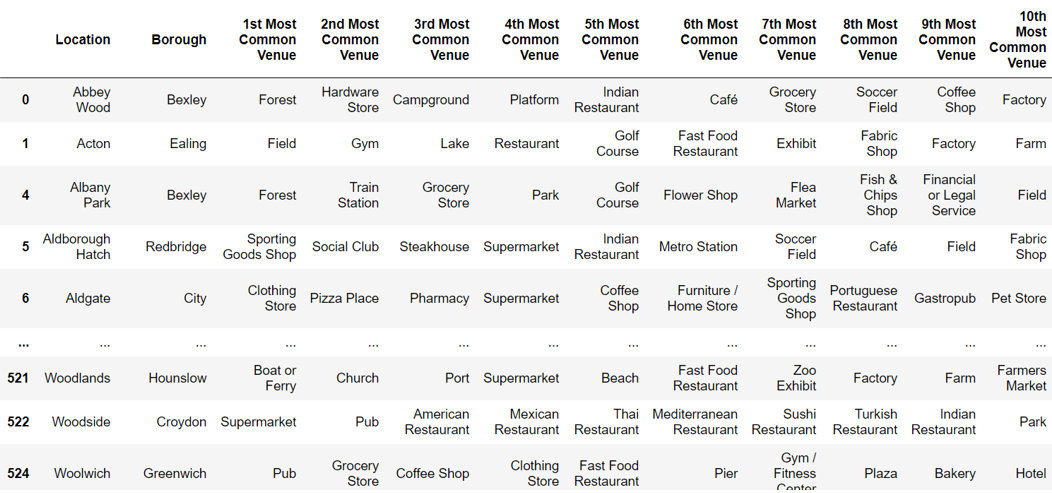


We can conclude that K=3 was the best value to achieve the best possible clustering using the k-means.

## Results

The clustering model then clusters the Locations in London and provides a label for each location which is representative of the cluster it belongs to. Locations in each individual cluster can be extracted using cluster labels and thus the most common venues of specific clusters can be extracted. This is done below for the 3 clusters by printing the 10 rows for each clusters.

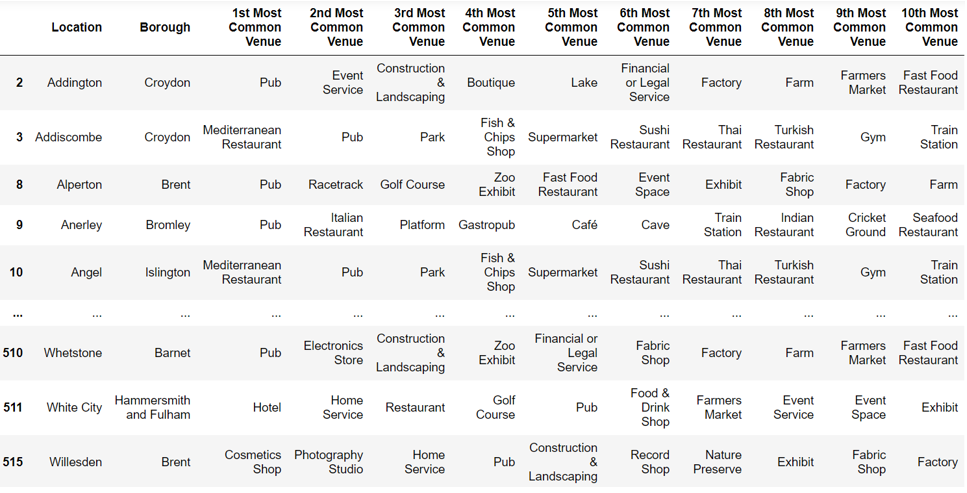
Cluster 1:



Cluster 2:

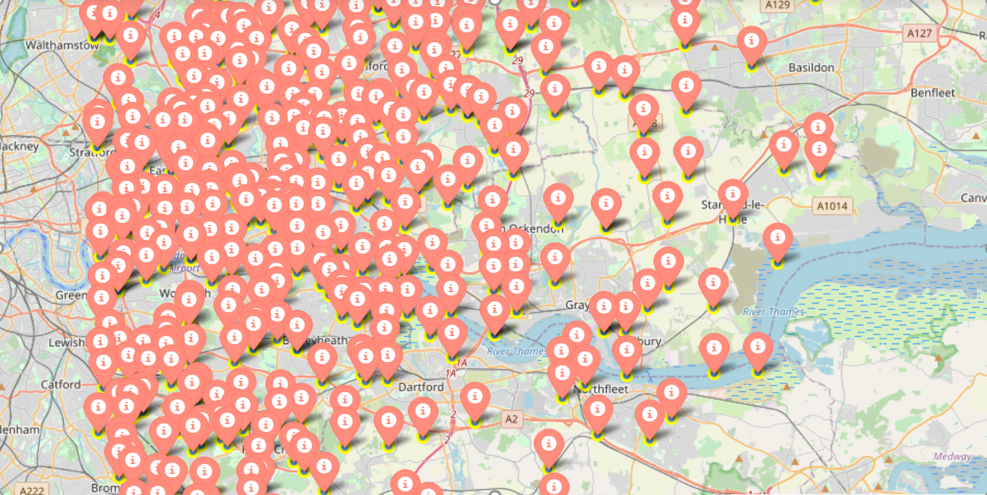


Cluster 3:



## Discussion

By analysing the 3 clusters, we can see that one cluster is more suitable for opening a new café or coffee shop. Location in cluster 2 has a small percentage of this type of stores in their top 10 common venues. Cluster 3 contain a higher degree of this type of shops comparing to the previous cluster (Cluster 2). Location to cluster 1 seem to be the most appropriate location to open a café or a coffee shop. Thus, it is recommended that the new restaurant can be opened in the neighborhoods belonging to cluster 1. These locations can be plotted on a map as shown below.



## Conclusion

In this project, the location in London, UK have been successfully analyzed for determining which would be the best locations for opening a new café. Based on the analysis carried out, locations in cluster 1 are recommended as locations for launching a new café or coffee shop.