**Capstone Project - The Battle of Neighborhoods (Week 1)**

**Find the best location to open a Multi-Cuisine Restaurant in London Area**

**1. Introduction**

**1.1 Background**

The aim of this artefact is to propose find the best possible location or a spot to begin a new restaurant business in the London city by exploring several top areas around the city. Selecting the right location for a business is one of the first and very important decisions in running a business. Starting a new business in a metropolitan city such as London can be challenging. London has over 150 areas within a total of 30 plus boroughs with various places that attract tourists and people to visit. The zones and neighborhoods of London are different in terms of different factors that can directly affect the success chance of business.

**1.2 Problem**

It is important to evaluate different neighborhoods based on the factors that are important for running a successful business such as the number of competitor restaurants, footprint, vicinity of a food center, recreation & office spaces and the potential demand in that neighborhood. The challenge is to find a suitable neighborhood which is close enough to some amenities and venues, has fewer competitors, also being affordable. Budget plays a critical role in renting or buying a place to make decisions in rounding off a location decision for a small - medium scale businesses. Hence, we need the right data set and the right approach to predict the best venue. Since the London demography is so big, my client needs deeper insight from available data in other to decide where to establish the restaurant.

**1.3 Interest**

This project will be very useful for any startups or an individual who wants to start a Small to Medium scale business in one of the most expensive and busiest cities in the world. The target audience is broad, it ranges from Londoners, foreign tourists and those who are passionate about spicy food.

### 2. Data

### 2.1 Data source / Data requirement

This project will rely mainly on public data from Wikipedia and Foursquare data.

### 2.1.1 Dataset 1 - Greater London area neighborhood

Within the Greater London Area, there are areas that are within the London Area Postcode. The focus of this project will be the neighborhoods are that are within the London Post Code area.

The London Area consists of 32 Boroughs and the "City of London". Our data will be from the link - Greater London Area <https://en.wikipedia.org/wiki/List_of_areas_of_London>.

### 2.1.2 Dataset 2 - Foursquare data for exploring venues

The Foursquare API will be used to obtain the London Area venues especially using case by case analysis considering more than 1 neighborhood for the geographical location data. These will be used to explore the neighborhoods of London accordingly.

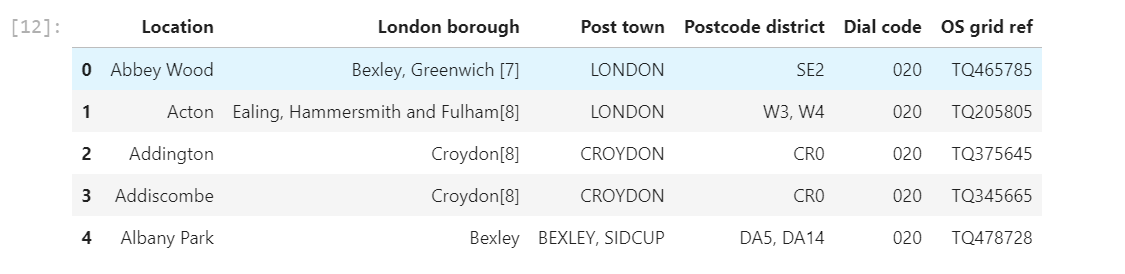
The venues within the neighborhoods of South East London like the area’s restaurants and proximity to amenities would be correlated. Also, accessibility and ease of supplies would be considered as it relates to venues.

To use the Foursquare API, we define the Foursquare Credentials and Version.

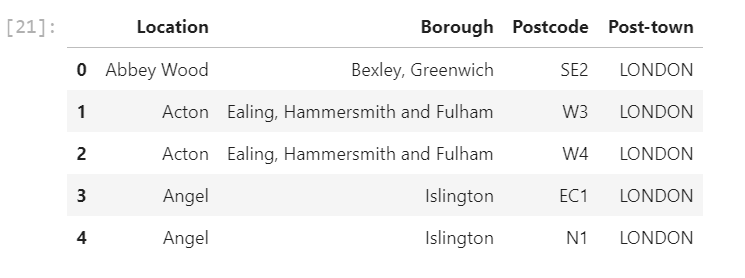
**Greater London areas and within London area and borough details are scrapped off Wikipedia website using Beautiful Soup technique.**

**2.2 Data Cleansing & Data Preparation**

The web scrapped data of the Wikipedia page for the Greater London Area data is provided below; Table available on website is then converted into Data Frame in Python.

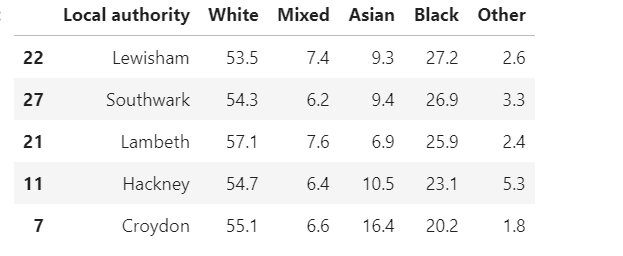


##### Post several assumptions and that only the 'Location', 'Borough', 'Postcode', 'Post-town' will be used for this project;

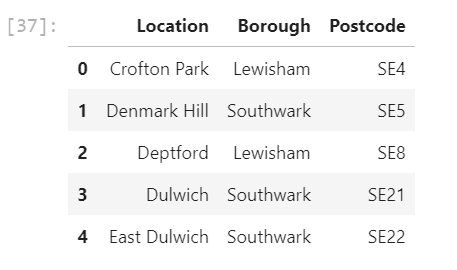


##### Critical Assumption (Based on demography):

##### This assumption will focus on the demography of London where there are predominantly more multicultural groups. According to the proportion of races by London borough as seen in Demography of London, the top 5 races or ethnics are shown below:

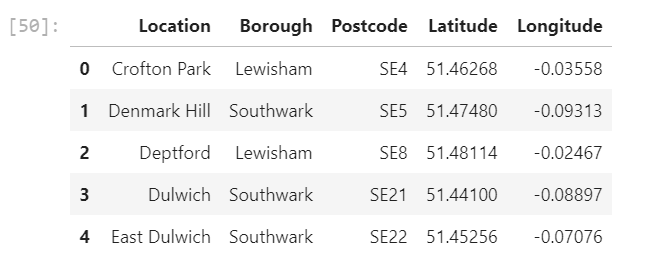


##### Based on the data analysis so far, this will be the top 5 areas significantly high "Black", "Asian", "Mixed" and other races. These leaves us with Lewisham, Southwark, Lambeth, Hackney and Croydon.



In obtaining the location data of the locations, the Geocoder package is used with the ArcGIS geocoder to obtain the latitude and longitude of the needed locations.

These will help to create a new data frame that will be used subsequently for the South East London areas.



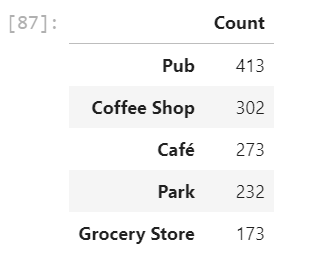
#### **3. Exploratory Data Analysis**

Perform data analysis for single and multiple neighborhoods;

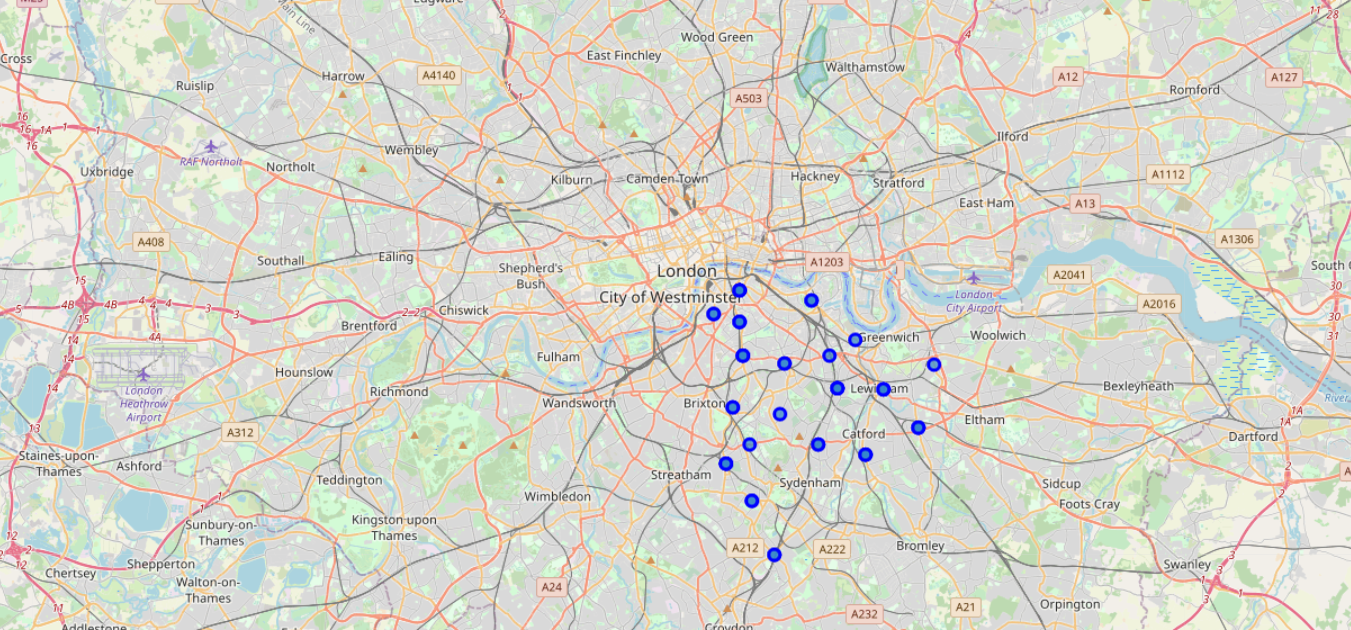
Using Four square API data, prepare URL to do API call and obtain data to explore venues of the above said boroughs.

All required Neighborhood, Venues Latitude and Longitude along with Venue category was extracted;

By performing count of venues, it was known the below;



Using Visualization, neighborhoods in South East London will be clustered based on the processed data obtained above.



**3.1 One-hot Encoding**

This technique was used to explore the Venue Categories for given neighborhoods.



**3.2 Grouping of each neighborhood with common venues**

Few examples are seen below; This step helped further analyzing every neighborhood in question by most common venues.



**4.0 Clustering of Neighborhoods**

K-Means clustering technique is used in the project to accomplish goal.

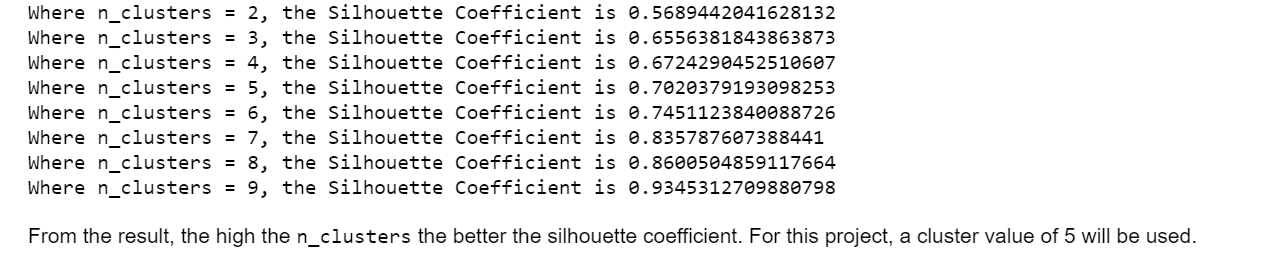
Generate and checking Cluster labels is necessary to identify what area belongs to which cluster label.

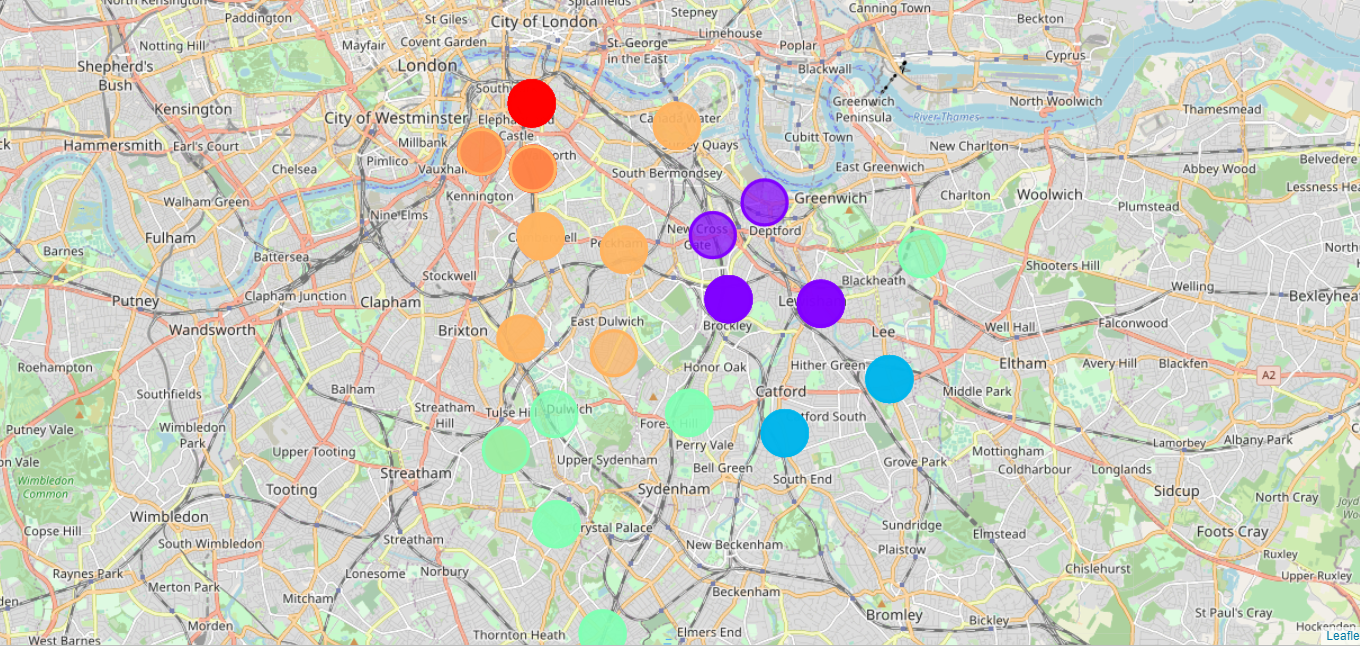
Once labels are assigned, evaluation must be done to validate the optimum no. of clusters to be used for K-means algorithm.

#### **4.1 Evaluation using ways to get the optimum number of clusters to be used for k-means**

The approach for this is to run the k-means clustering for a range of value k and for each value of k, the **Sum of the Squared Errors (SSE)** is calculated., calculate sum of squared errors (SSE). When this is done, a plot of k and the corresponding SSEs are then made. At the elbow (just like arm), that is where the optimal value of k is. And that will be the number of clusters to be used. The whole idea is to have minimum SSE.

To find the optimal value of the number of clusters, k, the number of clusters is iterated corresponding Silhouette Coefficient is calculated for each of the k-values used. The highest Silhouette Coefficient gives the best match to its own cluster. Please see below:





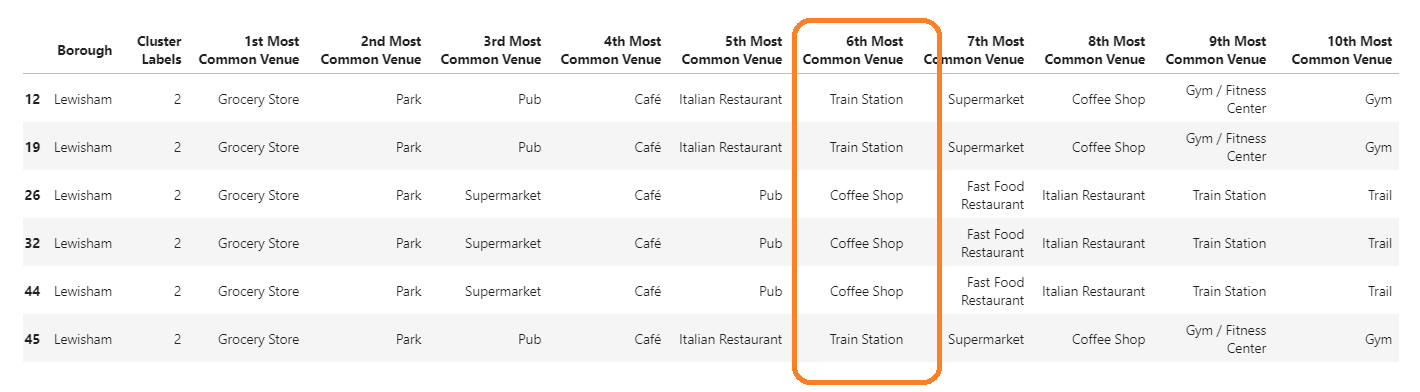
From above, we will choose cluster value of 5. This means that k-means will have 5 clusters and this data will be considered for training.

Of the 5 cluster’s data analyzed, using cluster labels assigned in previous steps, it is determined that cluster 2 and 3 are more suitable in terms of many aspects that reinforces opening a new restaurant / food joint.

Cluster 2:



Cluster 3:



#### **5. Result from Analysis in above steps**

The following are the highlights of the 5 clusters above:

* Pubs, Cafe, Coffee Shops are popular in the South East London.
* As for restaurants, the Italian Restaurants are very popular in the South East London area. Especially in Southwark and Lambeth areas.
* With the Lewisham area being the most condensed area of Africans in the South East Area, it is surprising to see how in the top 10 venues, you can barely see restaurants in the top 5 venues.
* Proximity to other amenities and accessibility to Train station
* Clusters 2 & 3 do not have top restaurants that could rival their standards if they are created and their accessibility, limited number of restaurants of class multi-cuisine
* Although, the Clusters have variations, a very visible presence is the predominance of pubs.

#### **6. Conclusion**

We can conclude with the results above that Cluster 2 and Cluster 3 are the most viable clusters to create a brand multi-cuisine Restaurant. Their proximity to other amenities and accessibility to Train station should be noticed.

These 2 clusters do not have top restaurants that could rival their standards if they are created.

In conclusion, this project would have had better results if there were more data in terms of crime data within the area, traffic access and allowance of more venues exploration with the Foursquare (limited venues for free API calls).

Also, getting the ratings and feedbacks of the current restaurants within the clusters would have helped in providing more insight into the best location.