**IBM APPLIED DATA SCIENCE CAPSTONE PROJECT**

**Table of Contents**

This report is presented in the following parts as shown below:

1. Introduction
2. Data
3. Methodology
4. Result
5. Discussion and Conclusion

# 1. Introduction

**1.1 Description of the Problem**

London's population has grown considerably in recent decades. London is very diverse. It represents what is called the reflection of the old British Empire. In London, you can get fresh food from Africa. One begins to wonder how efficient the delivery mechanism is.

The real problem is that even though there are many good restaurants in London - Asian, Middle Eastern, Latin and American restaurants - you may have a hard time finding a good place to dine on the best West African cuisine.

**1.2 Discussion of the Background**

a successful restaurant chain in Africa, is looking to expand its operations in Europe through London. They want to create a high-end restaurant that comes with an organic and healthy mix. Their target is not only West Africans, but they are also pro-organic and have a healthy diet. For them, every meal counts and counts like royalty when you eat.

With London's demographics so large, my client needs a deeper look at the data available in others to decide where to set up Europe's first "palace" restaurant.

**1.3 Target Audience**

London is a place where different shades live. As such, when looking for a high-end restaurant with an African bent, there is a great shortage. The target audience is wide, ranging from Londoners, tourists and passionate about organic food.

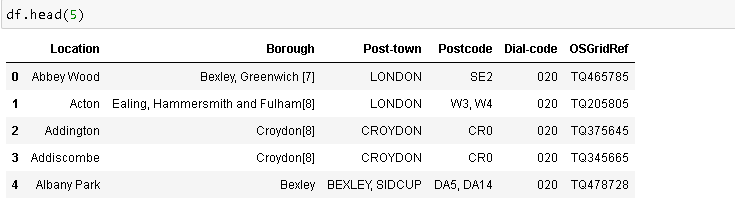
# 2. Data

**2.1 Description of Data**

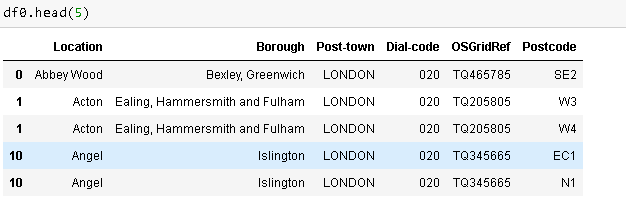
This project will rely on public data from Wikipedia and Foursquare.

**2.1.1 Dataset 1:**

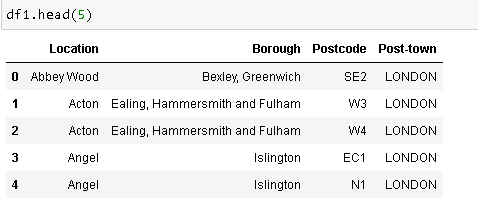
In this project, London will be used as synonymous to the "Greater London Area" in this project. Within the Greater London Area, there are areas that are within the London Area Postcode. The London Area consists of 32 Boroughs and the "City of London". Our data will be from the link - Greater London Area <https://en.wikipedia.org/wiki/List_of_areas_of_London>



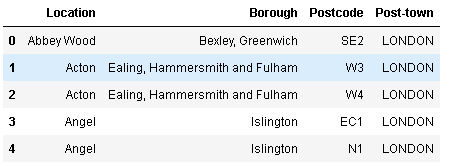
**Assumption 1:**Where the Postcode are more than one, (for example, in Acton, there are 2 postcodes - W3 and W4), the postcodes are spread to multi-rows and assigned the same values from the other columns



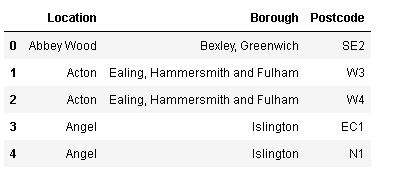
**Assumption 2:** From the data, only the 'Location', 'Borough', 'Postcode', 'Post-town' will be used for this project. So they are extracted into a new data frame.



**Assumption 3:** Now, only the Boroughs with London Post-town will be used for our search of location. Therefore, all the non-post-town are dropped.

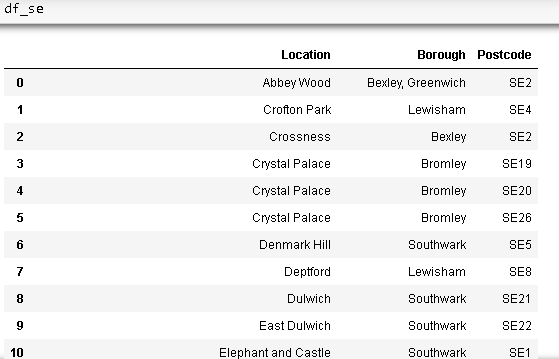


**Assumption 4:** Due to its more diverse outlook, proximity to afro-caribbean markets and accessible facilities, only the South East areas of London will be considered for our analysis. The South East areas has postcodes starting with SE.



**Assumption 5:** 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 Black Africans or Caribbeans.

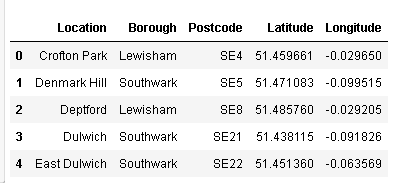
**Assumption 6:** Our next assumption will be based on the top 5 areas will significantly high "Black", "Mixed" and other races. These leaves us with Lewisham, Southwark, Lambeth, Hackney and Croydon.



**2.1.2 Dataset 2:**

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 dataframe that will be used subsequently for the South East London areas.



**2.1.3 Dataset 3:**

The Foursquare API will be used to obtain the South East London Area venues for the geographical location data. These will be used to explore the neighbourhoods of London acordingly.

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

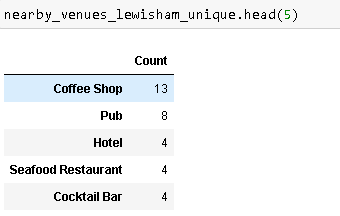
# 3. Methodology

**3.1 Data Exploration**

3.1.1 Single Neighbourhood

An initial exploration of a single Neighbourhood within the London area was done to examine the Foursquare workability. The Lewisham Borough postcode SE13 and Location - Lewisham is used for this. Let's explore the top 100 venues that are within a 2000 metres radius of Lewisham.

From the results, the necessary information needs to be obtained from items key. To do this, the get\_category\_type function is used from the Foursquare lab.

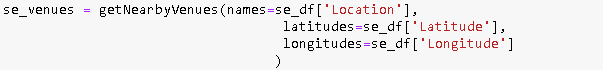


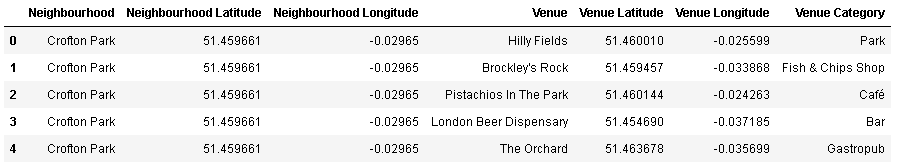
Interestingly, even though there are restaurants are the Lewisham area, they are not even in the top 5 venues. It should be noted that since we are limited by data availability, our perspectives will be on what we have.

# 3.1.2 Multiple Neighbourhoods

Now let's explore (Multiple) Neighborhoods in the South East London area.

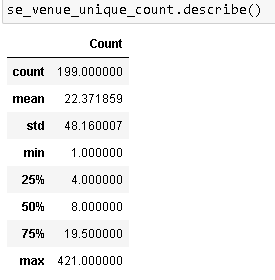
To do this, the function getNearbyVenues is used and it's created to repeat the same process for all neighborhoods. The created function - getNearbyVenues is then used on each neighbourhood. And creates a new dataframe called london\_venues.





The number of venues returned for each neighbourhoods is then explored as follows:

The next step is to check how many unique categories can be returned for the venues. See as follows.



**3.2 Clustering**

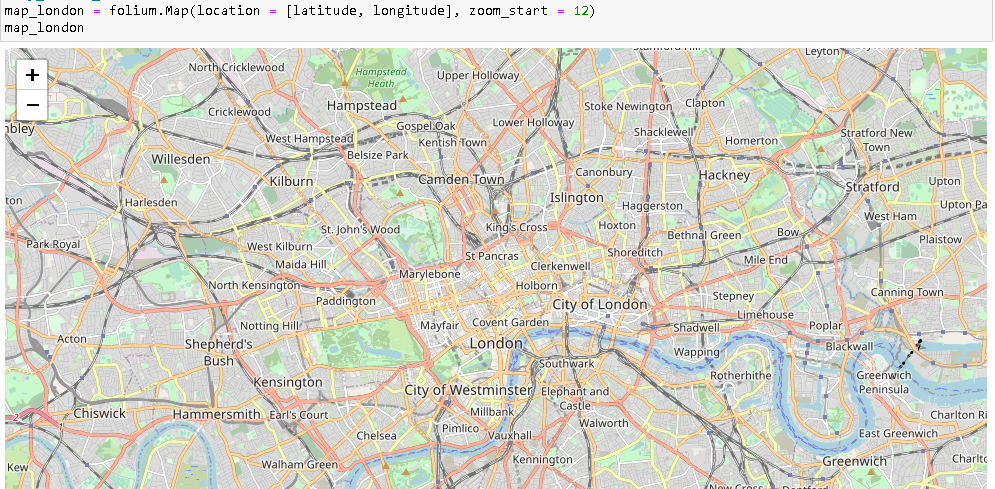
For this section, the neighbourhoods in South East London will be clustered based on the processed data obtained above.

**3.2.1 Libraries**

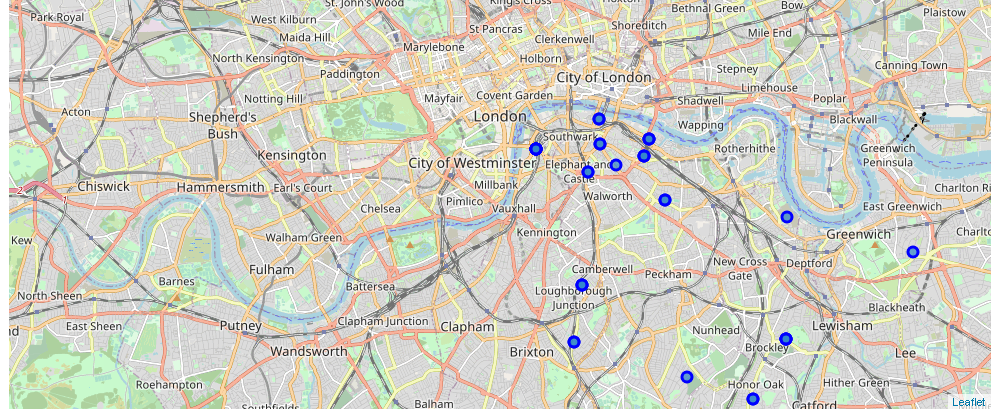
To get started, all the necessary libraries have been called in the libraries section above.

**3.2.2 Map Visualization**

Using the geopy library, the latitude and longitude values of London is obtained.



Adding markers to map.



**3.2.3 Analysing Each Neighborhood**

In this section, the objective is to check and explore the venues in each neighbourhood. As can be seen from above, Lewisham with its demography has no African restaurants within the top spots.



Then we create a new panda dataframe with 10 most common venues as shown below:



# 3.2.4 Clustering of Neighbourhoods

# Now creating a new dataframe that includes the clusters as well as the top 10 venues for each neighbourhoods.

# 

### 3.2.5 Optimal Number of Clusters for K-mean

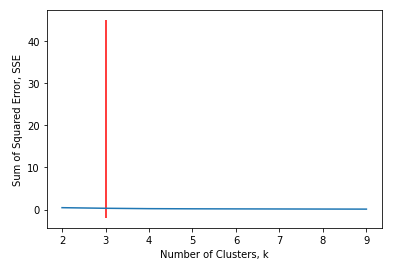
To get the optimal number of clusters to be used for the K-mean, there are a number ways possible for the evaluation. Therefore, in this task, the following are used:

#### 1. Elbow (Criterion) Method 2. Silhouette Coefficient

**1. Elbow Method**

The elbow method is used to solve the problem of selecting k. Interestingly, the elbow method is not perfect either but it gives significant insight that is perhaps not top optimal but sub-optimal to choosing the optimal number of clusters by fitting the model with a range of values for k.

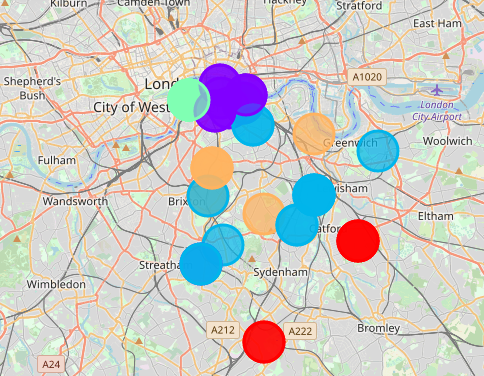
The approach for this is to run the k-means clustering for a range of value k and for each value of k, the Sum of the Squared Errors (SSE) is calculated., calculate sum of squared errors (SSE). When this is done, a plot of k and the corresponding SSEs are then made. At the elbow (just like arm), that 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.



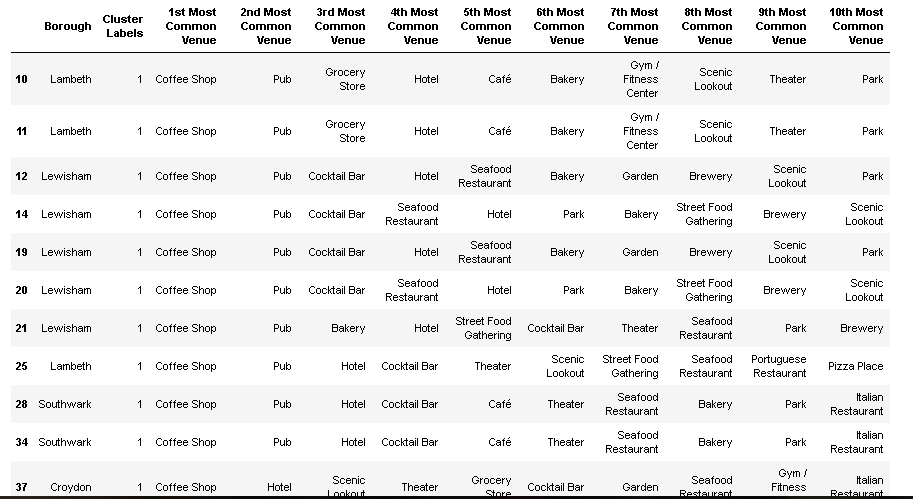
**2. Silhouette Coefficient**

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

**Clusters 1**



**Clusters 2**



**Cluster 3**



# Cluster 4

# 

# Cluster 5

# 

# 4. Result

**The following are the highlights of the 5 clusters above:**

1. Pubs, Cafe, Coffee Shops are popular in the South East London.
2. As for restaurants, the Italian Restaurants are very popular in the South East London area. Especially in Southwark and Lambeth areas.
3. 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.
4. Although, the Clusters have variations, a very visible presence is the predominance of pubs.

# 5. Discussion and Conclusion

It is very important to note that Clusters 2 and 3 are the most viable clusters to create a brand African Restaurant. Their proximity to other amenities and accessibility to station are paramount. These 2 clusters do not have top restaurants that could rival their standards if they are created. And the proximity to resources needed is paramount as Lewisham and Lambeth are not far out from Peckham (under Southwark).

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 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.