IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

Table of Contents

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

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

```
df.head(5)
```

	Location	Borough	Post-town	Postcode	Dial-code	OSGridRef
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728

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

```
df0.head(5)
```

	Location	Borough	Post-town	Dial-code	OSGridRef	Postcode
0	Abbey Wood	Bexley, Greenwich	LONDON	020	TQ465785	SE2
1	Acton	Ealing, Hammersmith and Fulham	LONDON	020	TQ205805	W3
1	Acton	Ealing, Hammersmith and Fulham	LONDON	020	TQ205805	W4
10	Angel	Islington	LONDON	020	TQ345665	EC1
10	Angel	Islington	LONDON	020	TQ345665	N1

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.

```
df1.head(5)
```

	Location	Borough	Postcode	Post-town
0	Abbey Wood	Bexley, Greenwich	SE2	LONDON
1	Acton	Ealing, Hammersmith and Fulham	W3	LONDON
2	Acton	Ealing, Hammersmith and Fulham	W4	LONDON
3	Angel	Islington	EC1	LONDON
4	Angel	Islington	N1	LONDON

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.

	Location	Borough	Postcode	Post-town
0	Abbey Wood	Bexley, Greenwich	SE2	LONDON
1	Acton	Ealing, Hammersmith and Fulham	W3	LONDON
2	Acton	Ealing, Hammersmith and Fulham	W4	LONDON
3	Angel	Islington	EC1	LONDON
4	Angel	Islington	N1	LONDON

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.

	Location	Borough	Postcode
0	Abbey Wood	Bexley, Greenwich	SE2
1	Acton	Ealing, Hammersmith and Fulham	W3
2	Acton	Ealing, Hammersmith and Fulham	W4
3	Angel	Islington	EC1
4	Angel	Islington	N1

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.

```
df_se
```

	Location	Borough	Postcode
0	Abbey Wood	Bexley, Greenwich	SE2
1	Crofton Park	Lewisham	SE4
2	Crossness	Bexley	SE2
3	Crystal Palace	Bromley	SE19
4	Crystal Palace	Bromley	SE20
5	Crystal Palace	Bromley	SE26
6	Denmark Hill	Southwark	SE5
7	Deptford	Lewisham	SE8
8	Dulwich	Southwark	SE21
9	East Dulwich	Southwark	SE22
10	Elephant and Castle	Southwark	SE1

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.

	Location	Borough	Postcode	Latitude	Longitude
0	Crofton Park	Lewisham	SE4	51.459661	-0.029650
1	Denmark Hill	Southwark	SE5	51.471083	-0.099515
2	Deptford	Lewisham	SE8	51.485760	-0.029205
3	Dulwich	Southwark	SE21	51.438115	-0.091826
4	East Dulwich	Southwark	SE22	51.451360	-0.063569

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.

```
nearby_venues_lewisham_unique.head(5)
```

	Count
Coffee Shop	13
Pub	8
Hotel	4
Seafood Restaurant	4
Cocktail Bar	4

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.

```
se_venues = getNearbyVenues(names=se_df['Location'],
                            latitudes=se_df['Latitude'],
                            longitudes=se_df['Longitude']
                           )
```

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Crofton Park	51.459661	-0.02965	Hilly Fields	51.460010	-0.025599	Park
1	Crofton Park	51.459661	-0.02965	Brockley's Rock	51.459457	-0.033868	Fish & Chips Shop
2	Crofton Park	51.459661	-0.02965	Pistachios In The Park	51.460144	-0.024263	Café
3	Crofton Park	51.459661	-0.02965	London Beer Dispensary	51.454690	-0.037185	Bar
4	Crofton Park	51.459661	-0.02965	The Orchard	51.463678	-0.035699	Gastropub

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.

```
se_venue_unique_count.describe()
```

	Count
count	199.000000
mean	22.371859
std	48.160007
min	1.000000
25%	4.000000
50%	8.000000
75%	19.500000
max	421.000000

3.2 Clustering

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

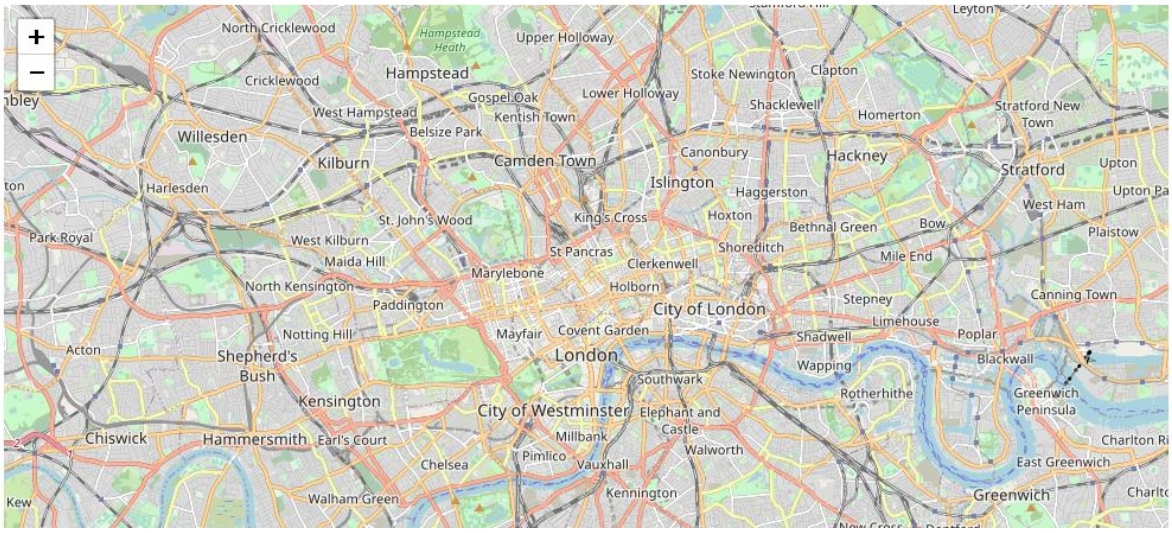
3.2.1 Libraries

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

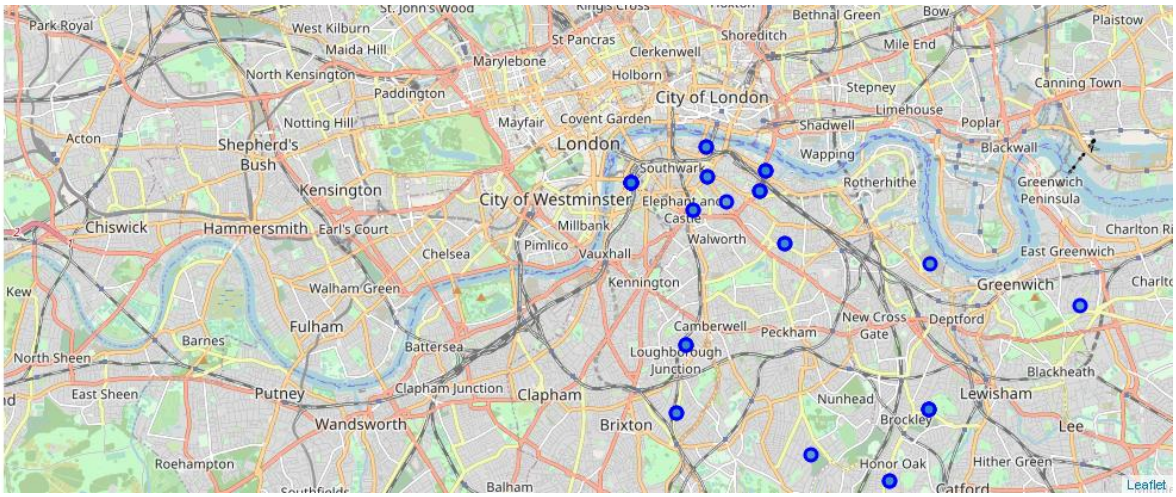
3.2.2 Map Visualization

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

```
map_london = folium.Map(location = [latitude, longitude], zoom_start = 12)
map_london
```



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.

	Neighbourhood	African Restaurant	R
1999	Lewisham	0	
2000	Lewisham	0	
2001	Lewisham	0	
2002	Lewisham	0	
2003	Lewisham	0	
2004	Lewisham	0	
2005	Lewisham	0	
2006	Lewisham	0	

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

	Neighbourhood	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	9th Most Common Venue	10th Most Common Venue
0	Bankside	Hotel	Plaza	Bookstore	Theater	Garden	Coffee Shop	Art Museum	Bakery	Dessert Shop	Park
1	Bellingham	Grocery Store	Park	Pub	Supermarket	Coffee Shop	Café	Italian Restaurant	Train Station	Pharmacy	Gym / Fitness Center
2	Bermondsey	Hotel	Plaza	Bookstore	Theater	Garden	Coffee Shop	Art Museum	Bakery	Dessert Shop	Park
3	Blackheath	Pub	Café	Park	Garden	Coffee Shop	Grocery Store	Bakery	Historic Site	History Museum	Brewery
4	Brixton	Coffee Shop	Café	Pub	Park	Beer Bar	Market	Cocktail Bar	Pizza Place	Italian Restaurant	Brewery

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.

	Location	Borough	Postcode	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	7th Most Common Venue	8th Most Common Venue	
0	Crofton Park	Lewisham	SE4	51.459661	-0.029650	2	Pub	Café	Coffee Shop	Park	Gastropub	Food Truck	Turkish Restaurant	Restaurant	
1	Denmark Hill	Southwark	SE5	51.471083	-0.099515	4	Coffee Shop	Café	Pub	Park	Beer Bar	Market	Cocktail Bar	Pizza Place	Re
2	Deptford	Lewisham	SE8	51.485760	-0.029205	4	Pub	Coffee Shop	Café	Bar	Park	Cocktail Bar	Brewery	Market	S
3	Dulwich	Southwark	SE21	51.438115	-0.091826	2	Pub	Grocery Store	Café	Park	Coffee Shop	Bakery	Pizza Place	Brewery	
4	East Dulwich	Southwark	SE22	51.451360	-0.063569	4	Pub	Coffee Shop	Café	Pizza Place	Italian Restaurant	Gastropub	Indian Restaurant	Bar	

3.2.5 Optimal Number of Clusters for K-mean

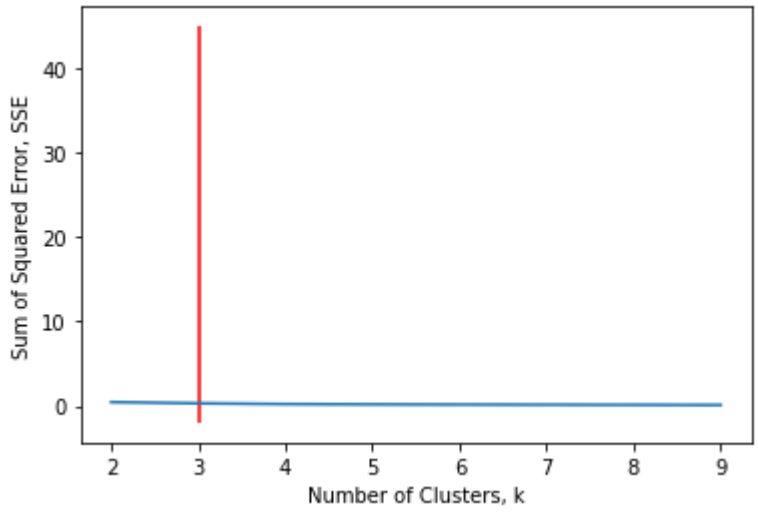
To get the optimal number of clusters to be used for the K-mean, there are a number ways possible for the evaluation. Therefore, in this task, the following are used:

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.

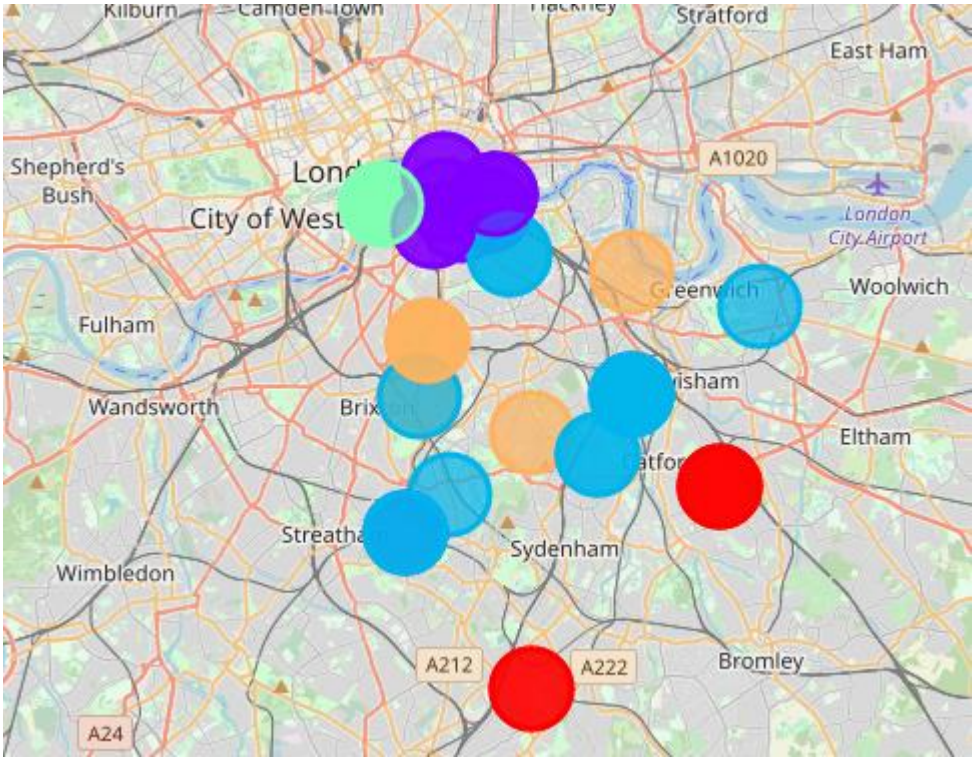
The approach for this is to run the k-means clustering for a range of value k and for each value of k, the Sum of the Squared Errors (SSE) is calculated., calculate sum of squared errors (SSE). When this is done, a plot of k and the corresponding SSEs are then made. At the elbow (just like arm), that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

	Borough	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
10	Lambeth	1	Coffee Shop	Pub	Grocery Store	Hotel	Café	Bakery	Gym / Fitness Center	Scenic Lookout	Theater	Park
11	Lambeth	1	Coffee Shop	Pub	Grocery Store	Hotel	Café	Bakery	Gym / Fitness Center	Scenic Lookout	Theater	Park
12	Lewisham	1	Coffee Shop	Pub	Cocktail Bar	Hotel	Seafood Restaurant	Bakery	Garden	Brewery	Scenic Lookout	Park
14	Lewisham	1	Coffee Shop	Pub	Cocktail Bar	Seafood Restaurant	Hotel	Park	Bakery	Street Food Gathering	Brewery	Scenic Lookout
19	Lewisham	1	Coffee Shop	Pub	Cocktail Bar	Hotel	Seafood Restaurant	Bakery	Garden	Brewery	Scenic Lookout	Park
20	Lewisham	1	Coffee Shop	Pub	Cocktail Bar	Seafood Restaurant	Hotel	Park	Bakery	Street Food Gathering	Brewery	Scenic Lookout
21	Lewisham	1	Coffee Shop	Pub	Bakery	Hotel	Street Food Gathering	Cocktail Bar	Theater	Seafood Restaurant	Park	Brewery
25	Lambeth	1	Coffee Shop	Pub	Hotel	Cocktail Bar	Theater	Scenic Lookout	Street Food Gathering	Seafood Restaurant	Portuguese Restaurant	Pizza Place
28	Southwark	1	Coffee Shop	Pub	Hotel	Cocktail Bar	Café	Theater	Seafood Restaurant	Bakery	Park	Italian Restaurant
34	Southwark	1	Coffee Shop	Pub	Hotel	Cocktail Bar	Café	Theater	Seafood Restaurant	Bakery	Park	Italian Restaurant
37	Croydon	1	Coffee Shop	Hotel	Scenic Lookout	Theater	Grocery Store	Cocktail Bar	Garden	Seafood Restaurant	Gym / Fitness	Italian Restaurant

Cluster 3

	Borough	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Lewisham	2	Pub	Café	Coffee Shop	Park	Gastropub	Food Truck	Turkish Restaurant	Restaurant	Gym	Grocery Store
3	Southwark	2	Pub	Grocery Store	Café	Park	Coffee Shop	Bakery	Pizza Place	Brewery	Forest	Italian Restaurant
9	Lewisham	2	Pub	Coffee Shop	Café	Grocery Store	Park	Fish & Chips Shop	Supermarket	Gastropub	Gym / Fitness Center	Italian Restaurant
15	Lewisham	2	Pub	Coffee Shop	Café	Grocery Store	Park	Fish & Chips Shop	Supermarket	Gastropub	Gym / Fitness Center	Italian Restaurant
16	Lewisham	2	Pub	Coffee Shop	Café	Park	Hotel	Cocktail Bar	Gastropub	Bakery	Seafood Restaurant	Brewery
17	Lewisham	2	Pub	Coffee Shop	Café	Park	Hotel	Cocktail Bar	Gastropub	Bakery	Seafood Restaurant	Brewery
24	Southwark	2	Coffee Shop	Pub	Café	Park	Supermarket	Brewery	Bakery	Grocery Store	Beer Bar	Pharmacy
26	Southwark	2	Coffee Shop	Pub	Café	Park	Supermarket	Brewery	Bakery	Grocery Store	Beer Bar	Pharmacy
33	Lewisham	2	Pub	Café	Coffee Shop	Park	Gastropub	Food Truck	Turkish Restaurant	Restaurant	Gym	Grocery Store

Cluster 4

	Borough	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	Southwark	3	Hotel	Coffee Shop	Theater	Pub	Bookstore	Scenic Lookout	Garden	Plaza	Art Museum	Park
6	Southwark	3	Hotel	Coffee Shop	Theater	Pub	Bookstore	Scenic Lookout	Garden	Plaza	Art Museum	Park
7	Southwark	3	Hotel	Coffee Shop	Theater	Pub	Bookstore	Scenic Lookout	Garden	Plaza	Art Museum	Park
8	Southwark	3	Hotel	Plaza	Bookstore	Theater	Garden	Coffee Shop	Art Museum	Bakery	Dessert Shop	Park
18	Lambeth	3	Hotel	Plaza	Bookstore	Theater	Garden	Coffee Shop	Art Museum	Bakery	Dessert Shop	Park
22	Southwark	3	Hotel	Plaza	Bookstore	Theater	Garden	Coffee Shop	Art Museum	Bakery	Dessert Shop	Park
23	Southwark	3	Hotel	Plaza	Bookstore	Theater	Garden	Coffee Shop	Art Museum	Bakery	Dessert Shop	Park
30	Southwark	3	Hotel	Plaza	Bookstore	Theater	Garden	Coffee Shop	Art Museum	Bakery	Dessert Shop	Park
38	Southwark	3	Hotel	Plaza	Bookstore	Theater	Garden	Coffee Shop	Art Museum	Bakery	Dessert Shop	Park

Cluster 5

	Borough	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Southwark	4	Coffee Shop	Café	Pub	Park	Beer Bar	Market	Cocktail Bar	Pizza Place	Italian Restaurant	Brewery
2	Lewisham	4	Pub	Coffee Shop	Café	Bar	Park	Cocktail Bar	Brewery	Market	Sandwich Place	Farm
4	Southwark	4	Pub	Coffee Shop	Café	Pizza Place	Italian Restaurant	Gastropub	Indian Restaurant	Bar	Park	Gym / Fitness Center
13	Lambeth	4	Coffee Shop	Pub	Café	Pizza Place	Cocktail Bar	Brewery	Market	Caribbean Restaurant	Park	Restaurant
41	Lambeth	4	Coffee Shop	Café	Pub	Park	Beer Bar	Market	Cocktail Bar	Pizza Place	Italian Restaurant	Brewery
43	Southwark	4	Coffee Shop	Café	Pub	Park	Beer Bar	Market	Cocktail Bar	Pizza Place	Italian Restaurant	Brewery

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.