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.

[12]:		Location	London borough	Post town	Postcode district	Dial code	OS grid ref
	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

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

[21]:		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

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:

	Local authority	White	Mixed	Asian	Black	Other
22	Lewisham	53.5	7.4	9.3	27.2	2.6
27	Southwark	54.3	6.2	9.4	26.9	3.3
21	Lambeth	57.1	7.6	6.9	25.9	2.4
11	Hackney	54.7	6.4	10.5	23.1	5.3
7	Croydon	55.1	6.6	16.4	20.2	1.8

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.

[37]:		Location	Borough	Postcode
	0	Crofton Park	Lewisham	SE4
	1	Denmark Hill	Southwark	SE5
	2	Deptford	Lewisham	SE8
	3	Dulwich	Southwark	SE21
	4	East Dulwich	Southwark	SE22

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.

[50]:

	Location	Borough	Postcode	Latitude	Longitude
0	Crofton Park	Lewisham	SE4	51.46268	-0.03558
1	Denmark Hill	Southwark	SE5	51.47480	-0.09313
2	Deptford	Lewisham	SE8	51.48114	-0.02467
3	Dulwich	Southwark	SE21	51.44100	-0.08897
4	East Dulwich	Southwark	SE22	51.45256	-0.07076

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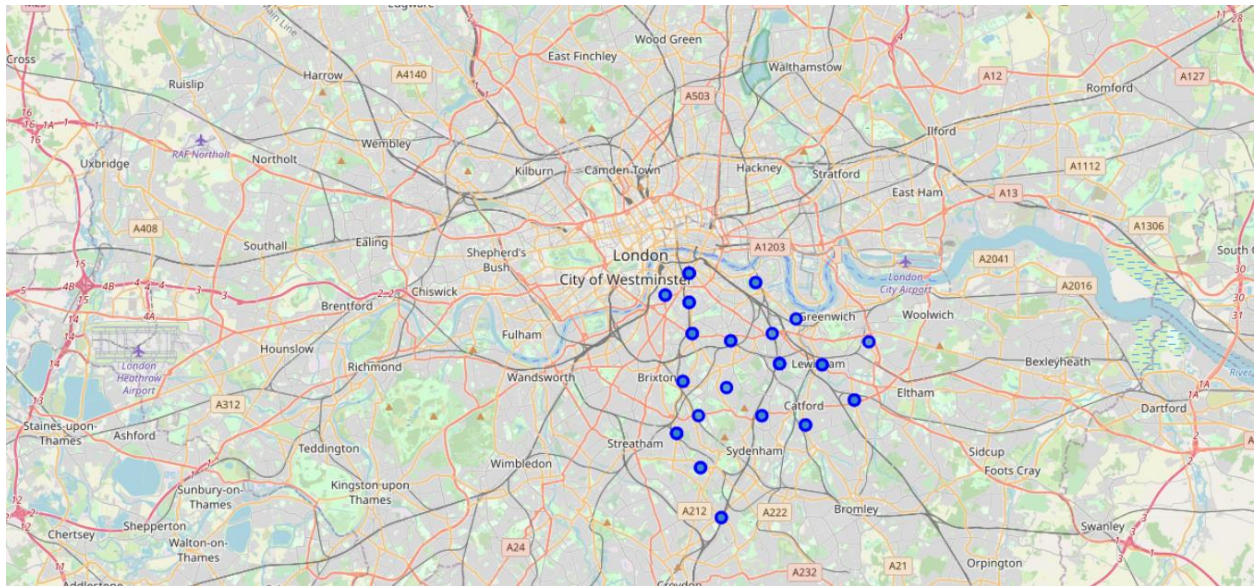
All required Neighborhood, Venues Latitude and Longitude along with Venue category was extracted;

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

[87]:

	Count
Pub	413
Coffee Shop	302
Café	273
Park	232
Grocery Store	173

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.

	Neighbourhood	African Restaurant	American Restaurant	Antique Shop	Aquarium	Argentinian Restaurant	Art Gallery	Art Museum	& Crafts Store	Asian Restaurant	Athletics & Sports	Australian Restaurant	BBQ Joint	Bakery	Bar	Beach	Beer Bar	Beer Garden	Beer Store	Bike Shop	Bistro
0	Crofton Park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Crofton Park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Crofton Park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Crofton Park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Crofton Park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

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.

	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	Coffee Shop	Theater	Scenic Lookout	Italian Restaurant	Grocery Store	Cocktail Bar	Pub	Street Food Gathering	Art Museum
1	Bellingham	Grocery Store	Park	Supermarket	Café	Pub	Coffee Shop	Fast Food Restaurant	Italian Restaurant	Train Station	Trail
2	Bermondsey	Hotel	Coffee Shop	Theater	Scenic Lookout	Italian Restaurant	Grocery Store	Cocktail Bar	Pub	Street Food Gathering	Art Museum
3	Blackheath	Pub	Grocery Store	Park	Coffee Shop	Café	Italian Restaurant	Garden	Supermarket	Bakery	Bus Stop
4	Brixton	Café	Coffee Shop	Pub	Park	Italian Restaurant	Middle Eastern Restaurant	Grocery Store	Pizza Place	Cocktail Bar	Bar

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:

```
Where n_clusters = 2, the Silhouette Coefficient is 0.5689442041628132
Where n_clusters = 3, the Silhouette Coefficient is 0.6556381843863873
Where n_clusters = 4, the Silhouette Coefficient is 0.6724290452510607
Where n_clusters = 5, the Silhouette Coefficient is 0.7020379193098253
Where n_clusters = 6, the Silhouette Coefficient is 0.7451123840088726
Where n_clusters = 7, the Silhouette Coefficient is 0.835787607388441
Where n_clusters = 8, the Silhouette Coefficient is 0.8600504859117664
Where n_clusters = 9, the Silhouette Coefficient is 0.9345312709880798
```

From the result, the high the `n_clusters` the better the silhouette coefficient. For this project, a cluster value of 5 will be used.
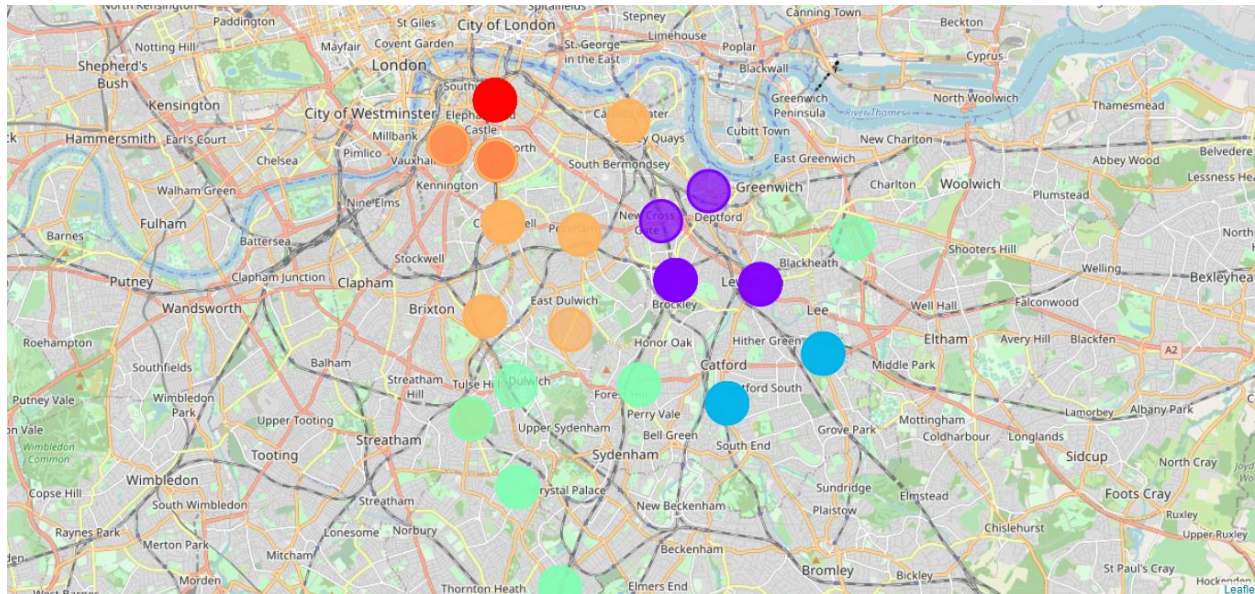
From above, we will choose cluster value of 5. This means that k-means will have 5 clusters and this data will be considered for training.

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:

	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	1	Pub	Coffee Shop	Café	Park	Indian Restaurant	Italian Restaurant	Cocktail Bar	Bar	Gastropub	Food Truck
2	Lewisham	1	Pub	Coffee Shop	Café	Bar	Park	Gastropub	Garden	Vietnamese Restaurant	Italian Restaurant	History Museum
14	Lewisham	1	Pub	Café	Park	Gastropub	Garden	Coffee Shop	Gym / Fitness Center	Fish & Chips Shop	Restaurant	Food Truck
16	Lewisham	1	Pub	Coffee Shop	Café	Park	Gastropub	Supermarket	Italian Restaurant	Food Truck	Bar	Fish & Chips Shop
17	Lewisham	1	Pub	Coffee Shop	Café	Park	Gastropub	Supermarket	Italian Restaurant	Food Truck	Bar	Fish & Chips Shop
20	Lewisham	1	Pub	Café	Park	Gastropub	Garden	Coffee Shop	Gym / Fitness Center	Fish & Chips Shop	Restaurant	Food Truck
21	Lewisham	1	Pub	Coffee Shop	Café	Bar	Italian Restaurant	Gastropub	Park	Brewery	Indian Restaurant	Vegetarian / Vegan Restaurant
33	Lewisham	1	Pub	Coffee Shop	Café	Park	Indian Restaurant	Italian Restaurant	Cocktail Bar	Bar	Gastropub	Food Truck
42	Lewisham	1	Pub	Coffee Shop	Café	Park	Indian Restaurant	Italian Restaurant	Cocktail Bar	Bar	Gastropub	Food Truck

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
12	Lewisham	2	Grocery Store	Park	Pub	Café	Italian Restaurant	Train Station	Supermarket	Coffee Shop	Gym / Fitness Center	Gym
19	Lewisham	2	Grocery Store	Park	Pub	Café	Italian Restaurant	Train Station	Supermarket	Coffee Shop	Gym / Fitness Center	Gym
26	Lewisham	2	Grocery Store	Park	Supermarket	Café	Pub	Coffee Shop	Fast Food Restaurant	Italian Restaurant	Train Station	Trail
32	Lewisham	2	Grocery Store	Park	Supermarket	Café	Pub	Coffee Shop	Fast Food Restaurant	Italian Restaurant	Train Station	Trail
44	Lewisham	2	Grocery Store	Park	Supermarket	Café	Pub	Coffee Shop	Fast Food Restaurant	Italian Restaurant	Train Station	Trail
45	Lewisham	2	Grocery Store	Park	Pub	Café	Italian Restaurant	Train Station	Supermarket	Coffee Shop	Gym / Fitness Center	Gym

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.