The Battle of Neighbourhoods

What's the Trend?

Introduction Section:

London is the capital and largest city of England and the United Kingdom. The city stands on the River Thames in the south-east of England, at the head of its 50-mile (80 km) estuary leading to the North Sea. London has been a major settlement for two millennia.

London is one of the world's most important global cities. It exerts a considerable impact upon the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research and development, tourism and transportation. It is one of the largest financial centres.

London has a diverse range of people and cultures, and more than 300 languages are spoken in the region. It is estimated mid-2018 municipal population (corresponding to Greater London) was roughly 9 million, which made it the third-most populous city in Europe. London accounts for 13.4% of the U.K. population. Greater London Built-up Area is the fourth-most populous in Europe, after Istanbul, Moscow, and Paris, with 9,787,426 inhabitants at the 2011 census.

This project aims at exploring possible venues in South East London to open a restaurant where are there is a high ethnic population. London is home to a mix of fine dining restaurants, eateries, coffee shops, pubs, street food markets etc. My client is aiming to open a restaurant that caters to Asian and African population in the South East of London. But there are various stages of investigation to zero in on the right area to open a new one according to the current market trends and the possibility of attracting customers. Hence to get a thorough understanding of the venues the following steps will be implemented:

- Explore the South East Area of London to obtain the list of regions with a higher concentration of ethnic population such as Asians and Africans.
- Obtain the required data from various data sources and extract the necessary information required for this project.
- Collect and collate data regarding the different ethnic groups living in South East London.
- Collect data regarding the list of available restaurants and other eateries in the chosen areas which gives a clear picture of the possibility of opening a new one and an insight into the current market trends.
- Group the collected data into different categories by using different data science methodologies and analyse the best possible venue to open a new restaurant which will cater to the ethnic population.

Data Section:

The following data will be required for this project:

1. The geographic divisions of London along with the area postcodes, borough detail etc.
2. Data for this project will be obtained from regions are that are within the London Post Code area. The London Area consists of 32 Boroughs and we will retrieve data from the link - **Greater London Area** https://en.wikipedia.org/wiki/List_of_areas_of_London
3. The retrieved data will then be scrapped and only necessary information will be retained for data processing.
4. The data obtained will be further processed to obtain details of ethnicity, adjacent eateries etc.
5. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.
6. Foursquare API will be used to retrieve information regarding the additional venues of the chosen area to open a new restaurant.
7. After the data collection we can run k-means clustering to cluster the potential regions of interest and visualize them on choropleth maps.

Business Problem:

The objective of this capstone project is to analyse and select the best locations in the city of London to open a new restaurant. Using data science methodologies and machine learning techniques this project aims to provide solutions to answer the business question: In the city of South East area of London, if a client is looking to open a new restaurant, what is the ideal location?

Target Audience of this project:

This project is particularly useful to restaurateurs and investors who are looking to open or invest in new restaurants in the city of London. This project could be implemented to finding ideal locations in other cities too.

Methodology:

A. Creating Datasets:

1. Using the appropriate data extracting methods and packages such as such as beautifulsoup we scrap and download the table containing the details of London boroughs, postcodes etc. We then load this data into a dataframe. We get a resulting table with around 500 rows and then we trim it down to the level as shown below:

	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

2. The next step is to extract only the necessary data (South East London Boroughs) required for this project and hence we drop the unwanted columns, save it in a new dataframe and obtain the table shown below:

	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

3. The Geocoder package is used with the arcgis geocoder to obtain the latitude and longitude of the needed locations. This will help to create a new dataframe that will be used subsequently for the South East London areas. These longitudes and latitudes will be joined with the dataframe to obtain the table below:

	Location	Borough	Postcode	Latitude	Longitude
0	Crofton Park	Lewisham	SE4	51.46268	-0.03558
1	Denmark Hill	Southwark	SE5	51.47478	-0.09312
2	Deptford	Lewisham	SE8	51.48117	-0.02476
3	Dulwich	Southwark	SE21	51.44100	-0.08897
4	East Dulwich	Southwark	SE22	51.45256	-0.07076

4. The Foursquare API is used to obtain the venue details in the South East London Area which will help us to explore the neighbourhoods in depth.

We will be able to obtain valuable information regarding the restaurants in SE London and other places of entertainment, nearby amenities etc.

B. Data Analysis of a Single Area:

1. To zero in on a single neighbourhood, Lewisham was chosen due to its diverse background. The FourSquare API was the used to extract information about the various types of eateries and other venues in the Lewisham area. The latitude and longitude coordinates of the Lewisham area was passed to the Foursquare API which resulted in a JSON file.

2. The JSON file is then processed and structured into a dataframe with name of the eatery, the type of the eatery and the location data. A sample of the resulting dataframe is shown below:

	Name	Categories	Lat	Long
0	Street Feast Model Market	Street Food Gathering	51.460209	-0.012199
1	Maggie's Kitchen	Café	51.465380	-0.011213
2	Gennaro Delicatessan	Deli / Bodega	51.461765	-0.009726
3	Levante restaurant	Restaurant	51.462072	-0.009491
4	Dirty South	Pub	51.458846	-0.002666
5	Levante Pide Restaurant	Turkish Restaurant	51.459848	-0.011476
6	Manor House Gardens	Park	51.456686	0.004684
7	Corte	Coffee Shop	51.459776	-0.011554
8	Everest Curry King	Sri Lankan Restaurant	51.466012	-0.019656
9	Blackheath Farmers' Market	Farmers Market	51.465913	0.007945
10	Côte Brasserie	French Restaurant	51.467378	0.007176
11	Buenos Aires Cafe	Argentinian Restaurant	51.467260	0.007083
12	Hilly Fields	Park	51.460010	-0.025599
13	The Spice Of Life	Indian Restaurant	51.458654	0.002613
14	Brockley Market	Farmers Market	51.467980	-0.024795
15	The Sausage Man	Food Truck	51.462507	-0.010248
16	Ladywell Tavern	Pub	51.456485	-0.021502
17	Pistachios In The Park	Café	51.460144	-0.024263
18	The Point Greenwich	Scenic Lookout	51.473202	-0.009293

3. A further analysis was done on the dataframe to obtain the number of eateries and their types as shown below:

Type	Count
Pub	13
Café	9
Park	6
Gastropub	6
Coffee Shop	5

4. From the data, it was inferred that there are around 100 venues in the Lewisham area.

C. Data Analysis of a Multiple Areas:

1. The FourSquare API is again used to venue details of multiple areas in South East London Area. The same process mentioned implemented in a single are analysis is done and we get a dataframe for each Borough as shown below:

	Neighbourhood	Neighbourhood Lat	Neighbourhood Long	Venue	Venue Lat	Venue Long	Venue Categoy
0	Crofton Park	51.46268	-0.03558	The Orchard	51.463678	-0.035699	Gastropub
1	Crofton Park	51.46268	-0.03558	Browns Of Brockley	51.464513	-0.037346	Coffee Shop
2	Crofton Park	51.46268	-0.03558	Brockley's Rock	51.459457	-0.033868	Fish & Chips Shop
3	Crofton Park	51.46268	-0.03558	Saka Maka	51.464826	-0.036437	Indian Restaurant
4	Crofton Park	51.46268	-0.03558	Salthouse Bottles	51.463916	-0.036618	Beer Store

2. After extracting venue details of all neighbourhoods, we then group all the neighbourhoods to get a comprehensive list of the number of venues and their categories. A sample of the resulting dataframe is shown below:

Neighbourhood	Neighbourhood Lat	Neighbourhood Long	Venue	Venue Lat	Venue Long	Venue Category
Bankside	100	100	100	100	100	100
Bellingham	71	71	71	71	71	71
Bermondsey	100	100	100	100	100	100
Blackheath	84	84	84	84	84	84
Brixton	100	100	100	100	100	100
Brockley	100	100	100	100	100	100
Camberwell	100	100	100	100	100	100
Catford	71	71	71	71	71	71
Chinbrook	57	57	57	57	57	57
Crofton Park	100	100	100	100	100	100

3. We then explore the above dataframe to get the number and type of eateries in the multiple neighbourhoods. We the get the following table:

Type	Count
Pub	423
Coffee Shop	317
Café	268
Park	210
Grocery Store	163

D. Clustering the Multiple Neighbourhoods:

1. We use the folium library to get a superimposed map of South East London Area with the help of latitudes and longitudes obtained from the geopy package.

2. The next step is to use the one-hot coding technique to explore in detail the venues in each neighbourhood based on a single category. A sample of the resulting dataframe is shown below:

	Neighbourhood	African Restaurant	American Restaurant	Antique Shop	Aquarium	Asian Restaurant	Art Gallery
134	Denmark Hill	1	0	0	0	0	0
658	Elephant and Castle	1	0	0	0	0	0

3. A grouping of each Neighbourhood with 10 common venues is done to extract the following results from them. An example is shown below:

```
----Bankside----
                venue  freq
0         Coffee Shop  0.09
1                 Pub  0.07
2               Hotel  0.06
3  Italian Restaurant  0.05
4             Theater  0.05
5        Cocktail Bar  0.03
6          Art Museum  0.03
7  Seafood Restaurant  0.03
8          Restaurant  0.03
9                 Bar  0.03
```

4. A new dataframe is created with results obtained in the previous step. The new dataframe contains the following details:

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
Bankside	Coffee Shop	Pub	Hotel	Italian Restaurant	Theater	Seafood Restaurant	Restaurant	Art Museum	Cocktail Bar	Bar
Bellingham	Grocery Store	Park	Supermarket	Café	Coffee Shop	Pub	Fast Food Restaurant	Train Station	Gym / Fitness Center	Gas Station
Bermondsey	Coffee Shop	Pub	Hotel	Italian Restaurant	Theater	Seafood Restaurant	Restaurant	Art Museum	Cocktail Bar	Bar
Blackheath	Pub	Grocery Store	Coffee Shop	Park	Café	Indian Restaurant	Bakery	Italian Restaurant	Supermarket	Gym
Brixton	Café	Coffee Shop	Park	Pub	Cocktail Bar	Italian Restaurant	Pizza Place	Grocery Store	Bar	Brewer

5. Using the K-means we now group the neighbourhoods into different clusters.

Location	Borough	Postcode	Lat	Long	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
Crofton Park	Lewisham	SE4	51.46268	-0.03558	1	Pub	Coffee Shop	Café	Park	Bar	Gastropub	Pizza Place	Bakery	Italian Restaurant	Turkish Restaurant
Denmark Hill	Southwark	SE5	51.47478	-0.09312	4	Café	Coffee Shop	Park	Pub	Cocktail Bar	Italian Restaurant	Pizza Place	Grocery Store	Bar	Brewery
Deptford	Lewisham	SE8	51.48117	-0.0	1	Pub	Coffee	Café	Bar	Park	Garden	History	Vietnamese	Italian	Historic Site

				2476			Shop					Museum	Restaurant	Restaurant	
Dulwich	Southwark	SE21	51.44100	-0.08897	3	Pub	Café	Park	Coffee Shop	Grocery Store	Bakery	Italian Restaurant	Brewery	Farmers Market	Bookstore
East Dulwich	Southwark	SE22	51.45256	-0.07076	4	Café	Pub	Coffee Shop	Pizza Place	Park	Gastropub	Burger Joint	Italian Restaurant	Restaurant	Platform

6. Using Folium, the clusters can be viewed as a map. Details of each clusters can be viewed as a tabular column stored in a dataframe.

Discussion:

According to the analysis, Lewisham and Lambeth will provide the least competition for an upcoming restaurant as due to the lack of many multi-cultural restaurants. Also these two areas have a number of other amenities close by. Though a number of eateries, pubs etc are available, the absence of Indian and African restaurants are prominent. Hence analysing the cluster information it is easy to recognise that the above mentioned are are ideal locations to open a new restaurant.

However with the availability of more relevant data such as traffic in the area etc we will be able to analyse the neighbourhoods in depth to arrive at an ideal location accurately.

Conclusion:

This project sheds light on a real world application of Data Science and gives a hands-on experience to solve real time problems. With the help of the methodologies learnt and useful libraries/packages it was a quite interesting to perform exploratory data analysis. To obtain better results many limitations of using FourSquare etc could be rectified in future projects. The results obtained from this project will be useful for clients to analyse the locations to open new restaurants based on current market trends. This project could be further enhanced to analyse data based on the variety of customers, their background and preferences. With accurate and details venue locations, we could also analyse the crime rates, population density, congestion etc to get better results.