Capstone Project: The Battle of Neighbourhoods

By Vibha Sharma

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

The population of London has grown considerably over the last decades. London is very diverse. It represents what is called the reflection of the old British Empire. In London, you can get fresh food supplies from Africa. One begins to wonder how efficient the supply mechanism is.

The real deal is that as much as there are many fine restaurants in London – Asian, Middle Eastern, Latin and American restaurants, you can struggle to find good place to dine in the finest of West African cuisine that has combination of Nigerian, Ghanaian, Cameroonian, Senegalese and more.
Eating in a cosy environment with a blend of multicultural background and finely made West African dishes, on time and on point in a London location accessible to tourists, within central London and not far from the "unofficial" capital African market place - Peckham.

1.2 Discussion of the Background

My client, a successful restaurant chain in Africa is looking to expand operation into Europe through London. They want to create a high-end restaurant that comes with organic mix and healthy. Their target is not only West Africans, but they are pro-organic and healthy eating. To them every meal counts and counts as a royal when you eat.

Since the London demography is so big, my client needs deeper insight from available data in other to decide where to establish the first Europe “palace” restaurant. This company spends a lot on research and provides customers with data insight into the ingredients used at restaurants.

1.3 Target Audience

Considering the diversity of London, there is a high multicultural sense. London is a place where different shades live. As such, in the search for a high-end African-inclined restaurant, there is a high shortage. The target audience is broad, it ranges from Londoners, tourists and those who are 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 focus of this project will be the neighbourhoods 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 >

London is big and due to the limitations in the number of calls for the Foursquare API, the following assumptions are made to confine this project to only South East 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`.
So, first, we remove the white-spaces at the start of some of the postcodes and then drop the other non-SE postcodes.

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,

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.
So, we are certain that the geocoder works fine. So we proceed to applying it to our dataframe.
Then we proceed to store the location data - latitude and longitude. The obtained coordinates are then joined to `df_se_top` to create new data frame.

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 accordingly. The venues within the neighbourhoods 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

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

Now, let's use the `Lewisham` with the index location 20

```
The latitude and longitude values of Lewisham with postcode SE13, are 5
1.46196000000003, -0.007539999999949032.
```

Let's explore the top 100 venues that are within a 2000 metres radius of Lewisham.
And then, let's create the `GET` request `URL`, and then the `url` is named.
Then, send the `GET` request and examine the results.
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.
The result is then cleaned up from `json` to a structured **pandas** dataframe
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) Neighbourhoods 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 neighbourhoods. And creates a new dataframe called **`london_venues`**.

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.

```
The geograpical coordinate of London are 51.5073219, -0.1276474.
```

The South East London neighbourhoods are then superimposed on top as shown below, still using the `folium` library. Please note due to the location of the South East London, you might need to zoom to see the superimposed areas.

3.2.3 Analysing Each Neighbourhood

In this section, the objective is to check and explore the venues in each neighbourhood.

One Hot Encoding
The `Neighbourhood` column is added back to the dataframe.
Some re-arrangement - move the new Neighbourhood column to the first column

Creating new dataframe:
Putting the common venues into pandas dataframe, the
following `return_most_common_venues`is used to sort the venues in descending order

3.2.4 Clustering of Neighbourhoods

The next thing to do now, is to create clusters of the neighbourhood using the `k-means` to cluster the neighbourhood into 5 clusters.

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 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 Coefficient` is calculated for each of the k-values used. The highest Silhouette Coefficient gives the best match to its own cluster.

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.