1.Introduction

1.1 Background

There are hundreds of new and different restaurants opening across London, from burger places like Five Guys and Gourmet Burger Kitchen to Dessert places like Kaspas and Heavenly Desserts. There are also many different cultural restaurants such as; Preto (*Brazilian*) and Ask Italia (*Italian*). The reason these types of places are so successful is because they provide Quality food in a nice setting and take away all the hassle and mess of cooking at home.

Across London there are an abundance of restaurants. Being able to place yours in a thriving area with minimum competition is vital to the survival of the business. In most areas around the centre of London there is an overwhelming choice of restaurants and it is for this reason this report is being done. It aims to guide potential owners towards making the best decision for locating their new premise based on competition and population.

1.2 Business Problem

In this report the aim is to provide detailed and valuable predictions to potential businesses on where to open a new restaurant based on the amount of restaurants in a neighbourhood.

More and more restaurants are opening up around the different counties of England, particularly in Cities, however a lot of planning goes into where a restaurant should open as this will determine the success of the business.

Property developers are realising there is a lot of money to be made from rental prices as more places open up, therefore being able to determine good areas to open up a restaurant is key to the success of the business. By the end of the report the client should be able to answer the question; “Where is the best place to open up my Restaurant in London .“

1.3 Target Audience

This project will be of interest to anyone who is looking to open up a restaurant in an area. It is also not limited to just London as using different areas you can easily implement this project into different towns, cities and even countries. This may also be of interest to property developers so they can see what areas already have a large sum of restaurants and may choose to build in an area with less competition.

2. Data Acquisition & Cleaning

2.1 Data Acquisition

The data we require will be:

- List of neighbourhoods which will be limited the area of London, UK
- Latitude and Longitude of each area, which will be used as part of the foursquare venue locations and plotting the areas on a map
- Unique Venue categories focusing on restaurants
- Population for each Borough

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_areas_of_London) contains all the Suburbs within the London region. Using data mining techniques such as BeautifulSoup and Pandas we shall extract the relevant information and place it into a table. Using Geolocator we will also get the latitude and longitude of each town to be used with Foursquare API in order to find out how many Restaurants are in a particular area and where is the best place to open up depending on competition.

The reason for using Foursquare is it has over 100million places and is used by many developers in order to get relevant information about certain areas. For the population we shall be using a dataset found on: https://www.statista.com/statistics/381055/london-population-by-borough/

Which was a downloadable xls file that was converted into a .csv file for streamline use with Pandas DataFrames.

2.2 Data Cleaning

As Foursquare returns a variety of different venue categories we shall be focusing on the ones containing the word ‘Restaurant’. This will make clustering the data using K-means more accurate to the industry we are focused on. Also as part of the data cleaning process any rows or columns containing missing non-numerical values will be removed as these may interfere with our results. As part of the Data cleaning process I made sure all the relevant information was correct and in the form of a data frame.

When scraping the main data set from Wikipedia I used pandas inbuilt function as this was the simplest method. As with most wikipedia tables there were bits of the cell strings I had to remove. As part of the cleaning process i also had to alter some of the columns headers as they were not in the correct format.

3. EDA

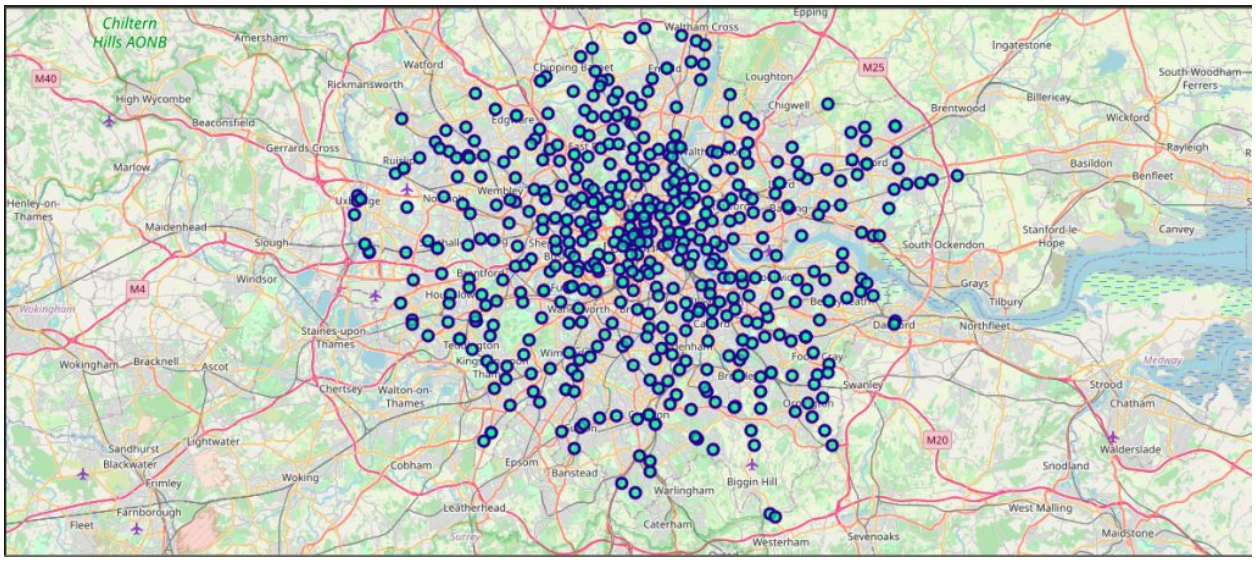
3.1 Overview of the Data

Once the data was cleaned I combined both the Wiki table and the Population dataset for each borough. Some of the rows contained Null values which I just filled with the mean of the population for ease. I then proceeded to add the latitude & longitude of each neighborhood to the dataframe using GeoPy and Geocoder. This will allow me to utilise both the Folium map plotting feature and Foursquare API's search function.

Neighborhood	London borough	Post town	Postcode district	lat	lon	Population
Aldgate	City	LONDON	EC3	51.514248	-0.075719	8.71
Aldwych	Westminster	LONDON	WC2	51.513103	-0.114920	255.32
Angel	Islington	LONDON	EC1, N1	51.531946	-0.106106	239.14
Archway	Islington	LONDON	N19	51.565437	-0.134998	239.14
Balham	Wandsworth	LONDON	SW12	51.445645	-0.150364	326.47
...	...	...	...	...	...	...
Walworth	Southwark	LONDON	SE17	51.490114	-0.090660	317.26
Wapping	Tower Hamlets	LONDON	E1	51.505436	-0.058729	317.71
West Brompton	Kensington and Chelsea	LONDON	SW10	51.486977	-0.195185	156.20
Westminster	Westminster	LONDON	SW1	51.501356	-0.124930	255.32
Whitechapel	Tower Hamlets	LONDON	E1	51.518623	-0.062081	317.71

Combined DataFrame with population

The next was to create a Folium map with data points indicating all the different Neighborhoods within London.



Arial view of london neighborhoods

As you can see from the map above there are a lot of Neighborhoods in and around London. However I do not need all of them.

3.2 Feature selection

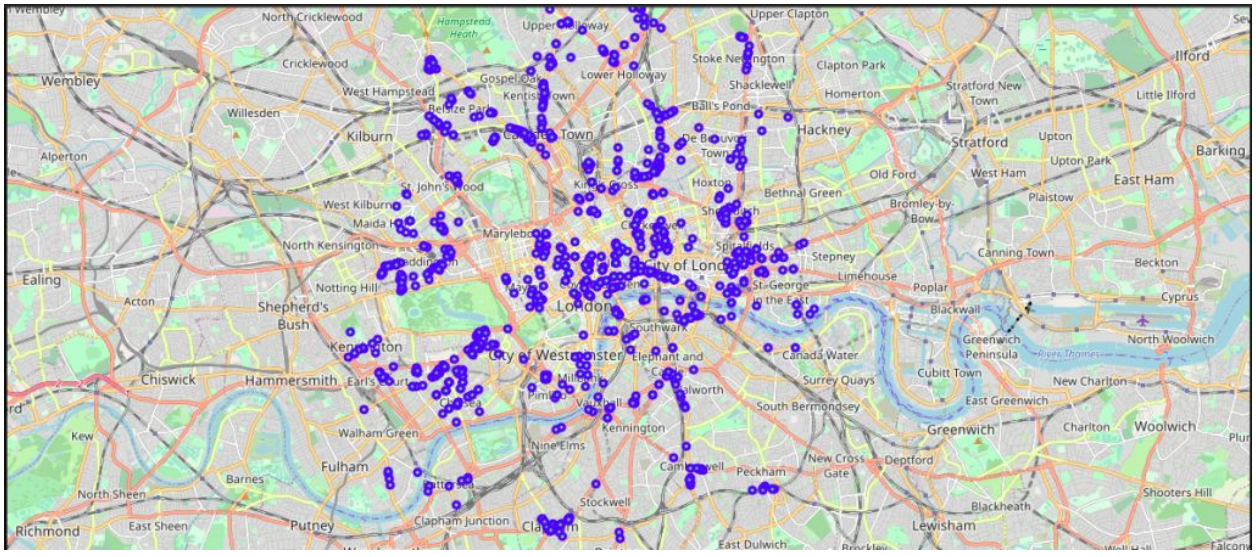
What I want to do is get all the Neighborhoods within a 5 Mile radius of the centre of London. To do this I used the Haversian formula to calculate the distance between each from the center of London and then filter the data into ones that are less than or equal to 5 Miles.

My reason for this is because a lot of the neighborhoods on the outer edges of London are not within the scope of this research project, which is to find the best location within the inner of London.

Before I could apply this formula I first needed to tweak a couple of incorrect coordinates and outliers which I did using GoogleMaps ‘Search Nearby’ feature to obtain the correct latitude and longitude.

The next step was to get the nearby venues using Foursquare. I limited the search to 100 venues within 500m radius of each point. Once I had these venues I filtered out the ones that contained ‘Restaurant’ which did include Fast Food such as KFC and McDonalds’s, these were excluded. There are a total of 1428 unique venues in each Borough of London, that’s a lot! But upon analysis it is clear that the majority belong to the center of London which makes sense as this is a massive attraction not only for tourists but also residents who live around the UK.

Here is the map showing all of the unique restaurants:



Aerial view of restaurants London

4. Modelling

The first step for getting the data ready for use with Kmeans was to encode all the Venue Categories using Pandas get dummies. This converts the categories into 0 or 1 depending if the neighborhood contains that type of Restaurant. I added a Total to the end of the dataframe to the amount of restaurants for each neighborhood and used this data for KMeans clustering.

Neighborhood	Total
Aldgate	30
Aldwych	12
Angel	17
Archway	7
Balham	9
...	...
Walworth	5
Wapping	4
West Brompton	1
Westminster	4
Whitechapel	8

DataFrame for clustering

For KMeans number of clusters I used 3 as I am clustering each neighborhood into either Low, Medium or High density of restaurants per 500m. I used the `*labels_*` function to obtain the cluster labels and then added this information plus the original data to create a final dataset for plotting onto a Folium map.

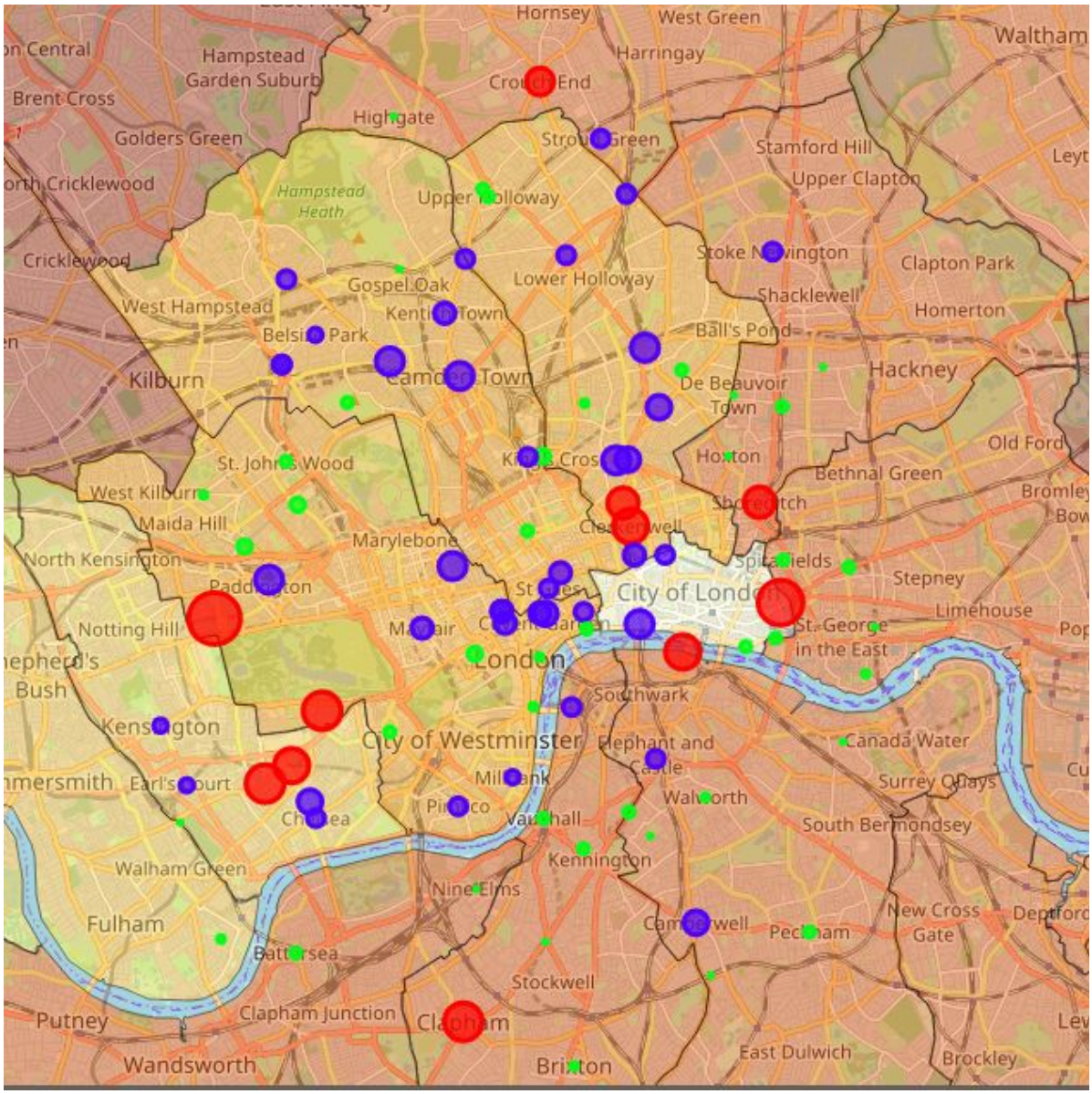
Neighborhood	London borough	Post town	Postcode district	lat	lon	Population	Cluster Labels	Total
Aldgate	City	LONDON	EC3	51.514248	-0.075719	8.71	2	30
Aldwych	Westminster	LONDON	WC2	51.513103	-0.114920	255.32	1	12
Angel	Islington	LONDON	EC1, N1	51.531946	-0.106106	239.14	1	17
Archway	Islington	LONDON	N19	51.565437	-0.134998	239.14	0	7
Balham	Wandsworth	LONDON	SW12	51.445645	-0.150364	326.47	0	9
...	...	...	...	...	...	...	...	...
Walworth	Southwark	LONDON	SE17	51.490114	-0.090660	317.26	0	5
Wapping	Tower Hamlets	LONDON	E1	51.505436	-0.058729	317.71	0	4
West Brompton	Kensington and Chelsea	LONDON	SW10	51.486977	-0.195185	156.20	0	1
Westminster	Westminster	LONDON	SW1	51.501356	-0.124930	255.32	0	4
Whitechapel	Tower Hamlets	LONDON	E1	51.518623	-0.062081	317.71	0	8

Final DataFrame

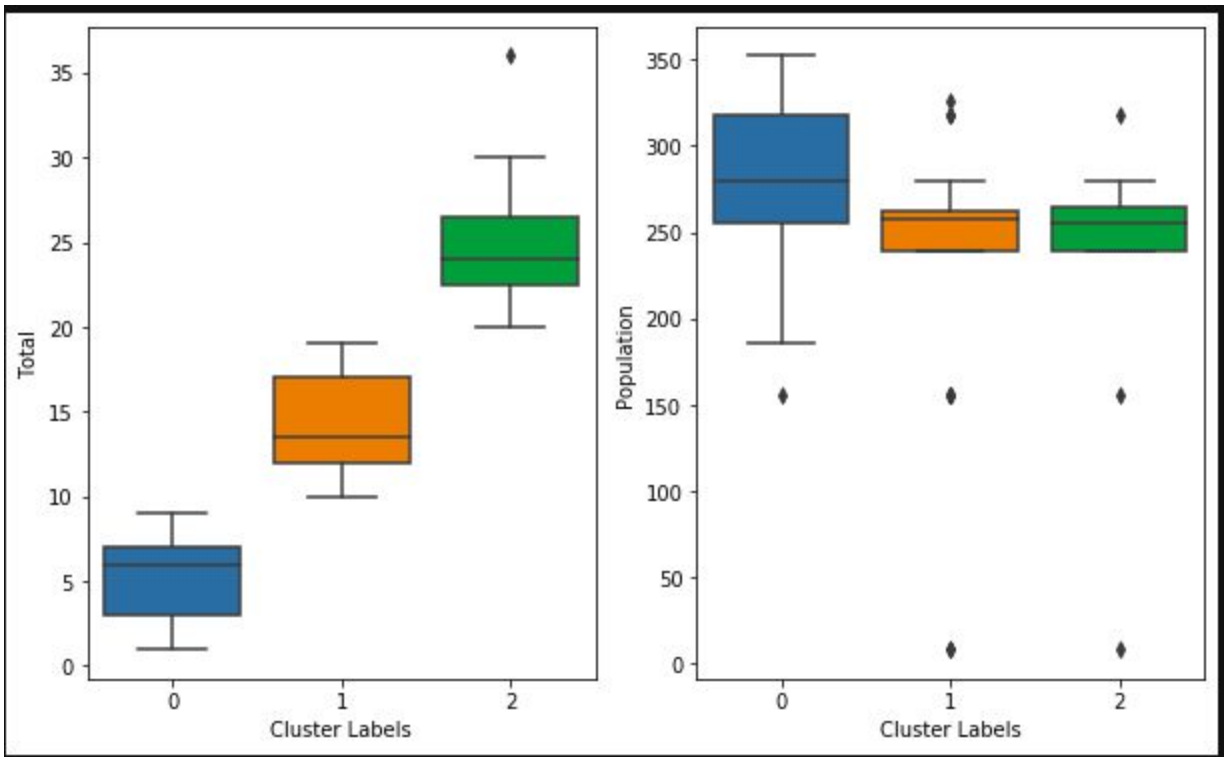
5. Results

The final map shows a colour for each cluster point depending on its restaurant density and also the size depicts the total volume of restaurants for that neighborhood. As the basemap I added a Choropleth map to show the population for each borough in London. Each colour represents a different cluster label:

- Green = Low
- Blue = Medium
- Red = High



Aerial View of London clusters



Cluster Boxplots

The left chart shows the amount of restaurants per cluster label and the right shows the population per cluster.

7. Discussion

As you can see from the map. The majority of the neighborhoods are located near the center of London and most of these have a medium amount of competition, with a total number of restaurants ranging from 10-20 in each neighborhood. Surprisingly the areas that have the most amount of restaurants are not in the center, they are in fact a little bit outside of the center. My main thought for this is that the center contains a lot of offices and commercial businesses as opposed to the outer edges.

I can infer from this that the best places to open a restaurant would be in either cluster 0 or 1 with little competition, however one factor for these results could be the cost of rent in that area.

Looking at population density, it has little impact on the amount of restaurants in an area. In fact the areas with the highest population seem to have the smallest amount of restaurants, mostly due to these being residential areas. So as my final suggestion I would advise an area in cluster 0 but not too near the centre to avoid high rental costs.

7.1 limitations

In this project I used Foursquare's free API account which limits the amount of calls and results given per request. To improve the results one could use a paid account to obtain more results yielding better evaluations. Furthermore I was unable to obtain rental prices for each neighborhood and as such this was not included in the clustering process, in a future project I aim to use this sort of data to make better evaluations.

8. Conclusion

In this project I analysed the amount of restaurants per neighborhood in London. Using data from online sources to produce a report that provides useful insights into the best location to set up a new restaurant.

Using the Kmeans Machine Learning model I clustered the data into 3 groups. Each depending on the total number of restaurants and used this information to produce a map which shows the different neighborhoods, colour coded by amounts.

The final part of this project was to answer the business question; 'where would be the best location to open a new restaurant in London?' . This report has done so despite the limitations incurred so the client can use this information to make a decision.