

Battle of the Neighbourhoods

Eat-In, London!

1. Introduction

1.1. Background

The past decade has witnessed dramatic changes in the operating model of restaurants and eateries all over the world. Until the 2000s, dining-in used to be the default option at a restaurant and takeaway food was restricted to a select few fast food chains. Cut to today - at the touch of a button, one can choose to order food from among a wide range of restaurants and enjoy the ultimate comfort and convenience of eating in. This has been made possible by the emergence of restaurant-aggregating food delivery platforms. Food delivery has become one of fastest growing businesses across the cities of the world.

In a large, diverse city like London, UK, eating-in is quickly becoming more popular and practical than eating out. Also, with the entry of several players in the industry, each food delivery platform is scampering to have that edge over the others. There are several factors that drive the success of these platforms including pricing, frequency of orders, ratings of restaurants and consistent, on-time delivery. This is a really large scale and costly problem to solve. In this study, we specifically look at how we can employ Foursquare data in conjunction with other publicly available data to provide insights that could improve the performance of these platforms. We will specifically look at data pertaining to London and attempt to identify patterns and trends that could aid food delivery services.

1.2. Audience

This study will be of specific interest to people involved in the restaurant/ takeaway business around the world. Although the data is specific to London, the general approach to the analysis can be applied to any city of similar scale and

diversity. Insights gained could be used in refining already existing machine learning techniques being employed by these businesses. The results of the study can also prove useful to a new business (either restaurant or food delivery service) that is in the process of being setup, in order to determine the considerations during setup (i.e. Location, density of operations per area, number of delivery personnel stationed in each area etc.).

2. Data

2.1. Data Sources

London is the capital city of the United Kingdom and is one of the oldest and the most important cities of the world. Occupying an area of 600 square miles, London is home to 8.9 million people.

London is organised into 33 boroughs as shown in the image below:



Most of the data on the internet, pertaining to London, is organised per borough. Hence, it seems logical to use the same approach for this data analysis.

Data for this study was collected from the following sources:

- [Area, Population, Latitude & Longitude data for London boroughs](#)
- Venues across London - Foursquare API
- [Age information for population across London Boroughs](#). This data is not readily available, so it had to be tabulated separately and then used in the analysis, as available [here](#).
- GeoJSON file used for Choropleth mapping from [here](#)


2.2. Approach to data

The data about each London borough from the above sources are available in the form of tables (Wikipedia) and excel sheets. These tables and sheets are first converted into Python dataframes and appropriately merged. Once all the London borough data is organised into one single table, following cleaning and wrangling, the relevant columns of this data (Borough Name, Latitude, Longitude) are passed through to Foursquare through the Foursquare API in order to fetch information about venues in each of the boroughs. The venues are then analysed to categorise them as eateries or otherwise. Following this, further analysis is performed, as will be outlined in the upcoming sections of this report. This includes visualization of the venues, clustering of boroughs based on eateries, age of population etc.

2.3. Data Preparation & Cleaning

The data about London from the aforementioned sources is not readily usable for the task at hand.

-
1. The wiki tables contain a lot of information from which it is required to weed out those columns that are applicable to the problem statement and get rid of all others that are not. For example, columns listing the location of the Headquarters of London Boroughs are not relevant here.
 2. Geographical coordinates are in a format that cannot be parsed by Foursquare/ Folium libraries. This format conversion involves string manipulation and extraction using regular expression matching. The conversion is illustrated below:

Co-ordinates		Lat	Long
51°33'39"N 0°09'21"E / 51.5607°N 0.1557°E		51.5607	0.1557
51°37'31"N 0°09'06"W / 51.6252°N 0.1517°W		51.6252	-0.1517
51°27'18"N 0°09'02"E / 51.4549°N 0.1505°E		51.4549	0.1505
51°33'32"N 0°16'54"W / 51.5588°N 0.2817°W		51.5588	-0.2817
51°24'14"N 0°01'11"E / 51.4039°N 0.0198°E		51.4039	0.0198

The Longitudes that are positive are of those Boroughs that are East of the Greenwich Meridian and the negative ones are of Boroughs that are West of the Greenwich Meridian.

3. Age information collected from the original source is part of an interactive excel sheet with several underlying macros. In order to simplify the acquisition, the age information is copied into a separate csv file which is then loaded into pandas dataframe and manipulated.

3. Exploratory Data Analysis

3.1. Capture venues using Foursquare API

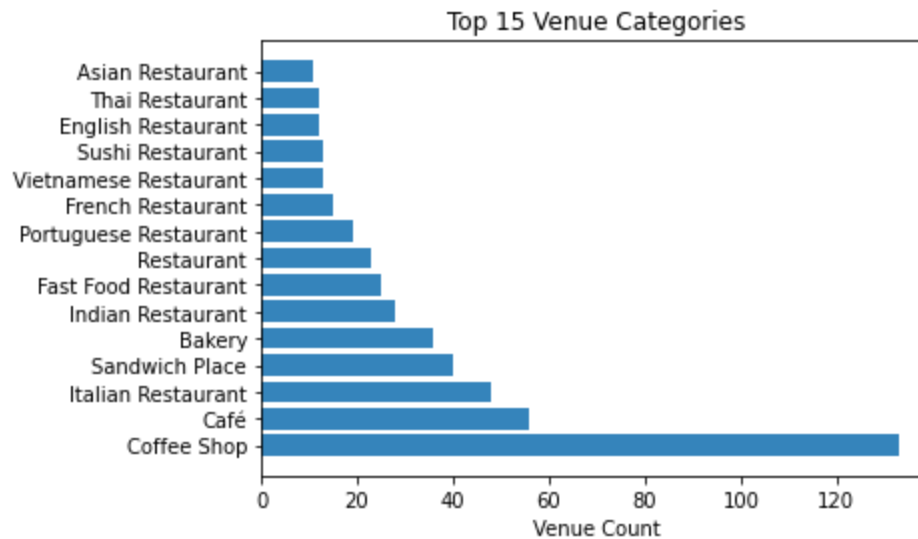
To kick start our analysis, the Foursquare explore call is used to get a list of all venues within a 500m radius from the center of each borough. The request URL includes client credentials and Borough location coordinates. The request call is made repeatedly for each Borough and each time a response is collected. The response provided by Foursquare is in the form of a JSON object containing details about each venue in each Borough. Each response is parsed to extract the names and categories of venues. This is populated into a new dataframe that looks like this:

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barking and Dagenham	51.5607	0.1557	Central Park	51.559560	0.161981	Park
1	Barking and Dagenham	51.5607	0.1557	Crowlands Heath Golf Course	51.562457	0.155818	Golf Course
2	Barking and Dagenham	51.5607	0.1557	Robert Clack Leisure Centre	51.560808	0.152704	Martial Arts Dojo
3	Barking and Dagenham	51.5607	0.1557	Morrisons	51.559774	0.148752	Supermarket
4	Barking and Dagenham	51.5607	0.1557	Becontree Heath Bus Station	51.561065	0.150998	Bus Station
5	Barking and Dagenham	51.5607	0.1557	Beacontree Heath Leisure Centre	51.560997	0.148932	Gym / Fitness Center
6	Barking and Dagenham	51.5607	0.1557	Dagenham Swimming Pool	51.560946	0.150054	Pool
7	Barnet	51.6252	-0.1517	Topsco	51.625717	-0.151247	Kitchen Supply Store
8	Barnet	51.6252	-0.1517	The Atrium	51.624726	-0.151933	Café
9	Barnet	51.6252	-0.1517	Made Curtains	51.623485	-0.153565	Home Service

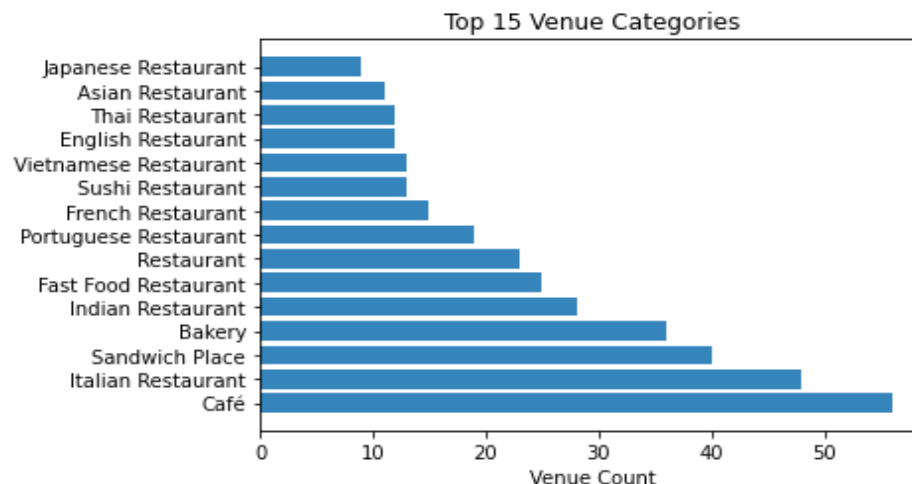
3.2. Extract eateries from among all venues

For the purpose of this study, we are only interested in venues that are eateries. These could be restaurants, cafes, fast food chains etc. Hence the next task is to segregate venues into those that are eateries and those that are not. A manual inspection of the dataframe shown above indicates that the following different venue categories can be tagged as eateries: Restaurant, Bakery, Cafe, Coffee

Shop, Sandwich Place, Fish and Chips Shop. Hence, these venue categories are marked as eateries, giving us a picture about the different types of eateries across London and their frequency:



The above graph tells us that Coffee Shops are the most common type of eatery across London, by a huge margin. But, should we really include coffee shops in our study? First of all, coffee shops don't fall in the typical food delivery umbrella. Secondly, the number of coffee shops across London seem to overshadow all other categories and could potentially skew our analysis of the neighbourhoods in London. Simply put, coffee shops are everywhere in London and including them in this study will not really add any distinguishing features to the Boroughs in question. Here are the Top 15 venue categories having removed Coffee Shops:



It is worth mentioning that delivery from coffee shops itself can be the subject of a separate study, which will be discussed later in this report.

3.3. Find the most common eatery category in each Borough

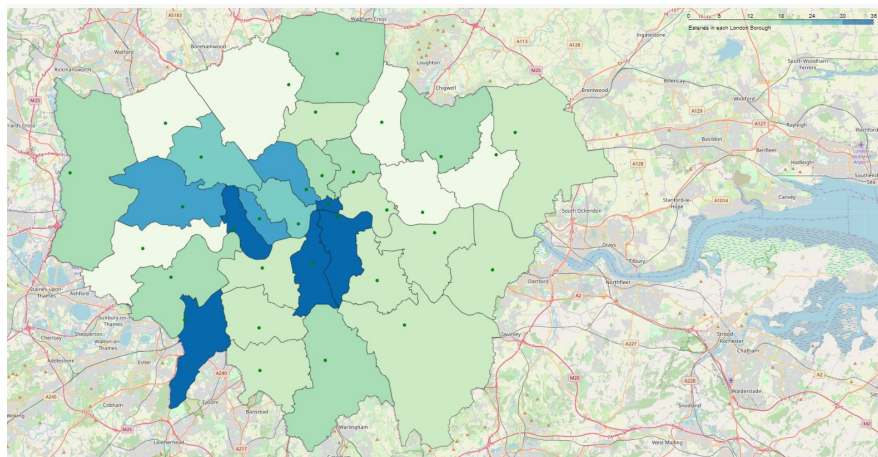
Next step is to determine which venue category occurs most frequently in each borough. This is done by counting the different venue categories in each borough and sorting them by frequency.

3.4. Determine number of eateries in each Borough

This gives us an idea about how busy a borough is with respect to restaurants. If a borough is very busy, food delivery businesses could benefit from operating in such neighbourhoods. Whereas, if an area is under-served with respect to restaurants, a new restaurant opening up could benefit from providing a service that does not exist presently.

3.5. Visualise eatery density in each Borough

The figure below is a Choropleth map illustrating the varying density of eateries across London:



The boroughs marked in darker colours indicate a high density of eateries. This is observed more in the central boroughs, as is expected. The suburban boroughs have a lighter density of eateries and are represented in lighter shades.

4. Borough Clustering

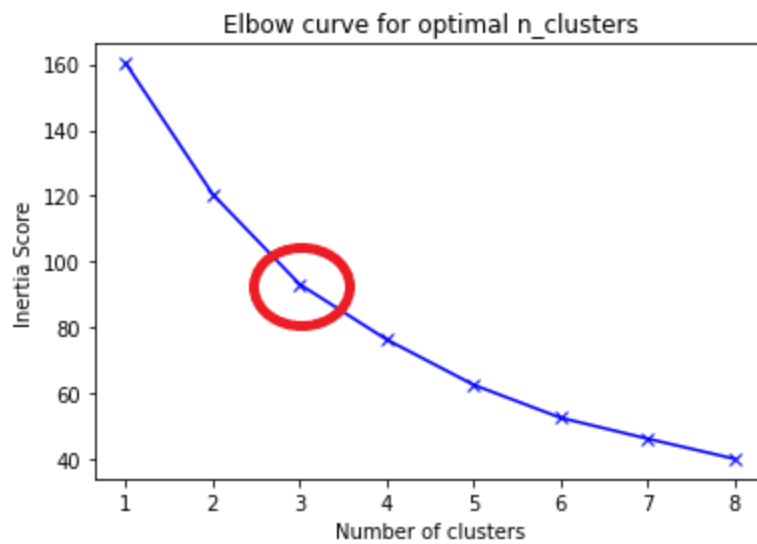
Clustering of the boroughs is performed using the K-Means algorithm.

Following our EDA, the columns of interest are:

```
['Borough', 'Number of eateries', 'Area (sq mi)', 'Population',  
 'Age', 'Most Common Venue Category']
```

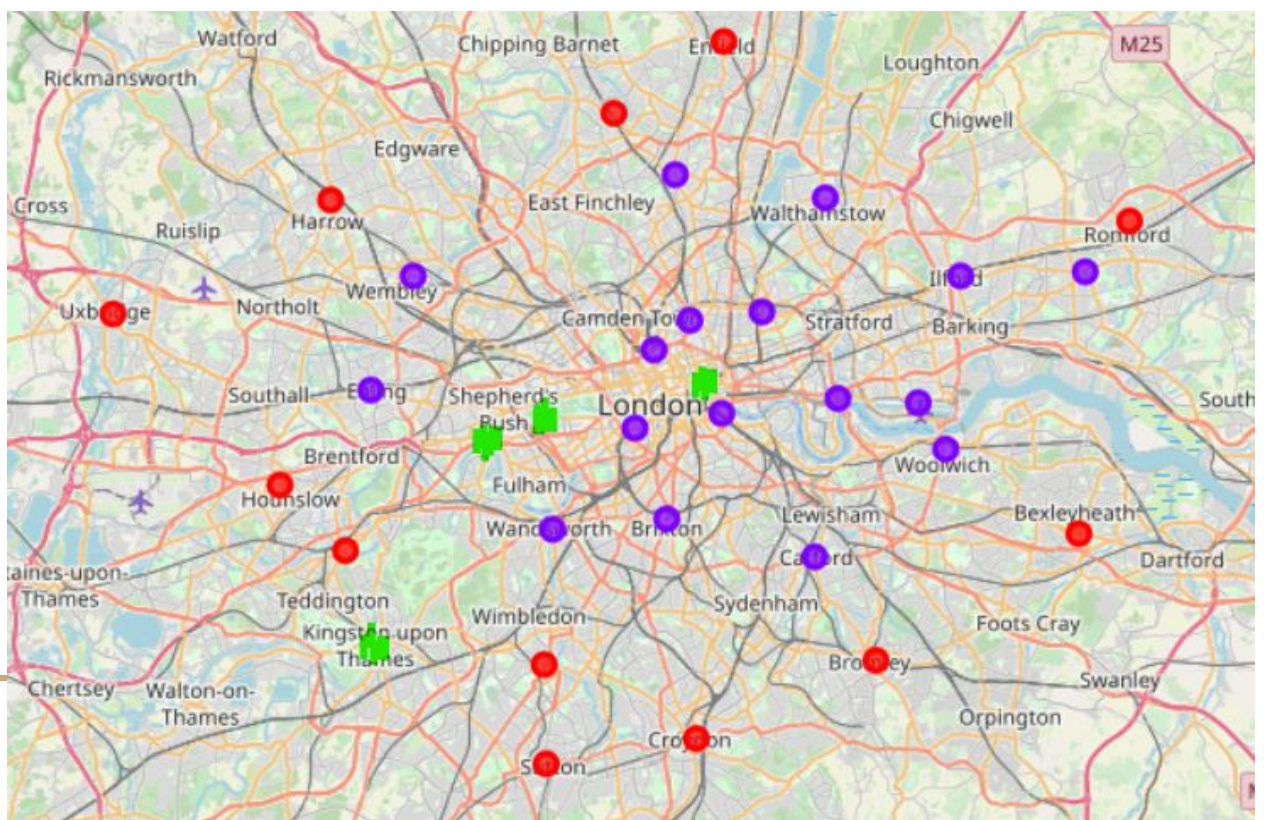
The next step before is to scale values in all the numeric columns in order to give each of them equal importance. The K-Means algorithm is not directly applicable to categorical variables. Hence, the categorical variable columns are one-hot encoded before the algorithm can be applied.

Once the numeric data is standardized and the categorical is one-hot encoded, the dataframe is ready for K-Means clustering. The K-Means wrapper provided by sci-kit learn package will be used here. With K-Means, the number of clusters (`n_clusters`) is to be specified before the model is run. In order to choose the optimal value of `n_clusters`, the algorithm is run in a loop and each time the metrics of performance of the algorithm is computed. There are several metrics that can be evaluated for this algorithm. For the purpose of this study, the `inertia_` attribute of the K-Means class is chosen. The `inertia_` attribute is basically the sum of squared distances of samples to their closest cluster centre. This is plotted for different `n_clusters` values and the elbow method of estimation is employed to determine the best value of `n_clusters`. The figure below illustrates the elbow curve for `n_clusters` ranging from 1 to 8.



From the plot, the rate of decrease of the inertia value drops after `n_clusters = 3`. Hence, the optimum `n_clusters` is 3. Now the K-Means is applied on the data and the three clusters are identified with labels 0, 1 and 2.

The cluster labels for each borough are then appended to the original dataframe and a Folium map of London is plotted, illustrating the different boroughs and their respective clusters encoded in different colours as shown below:



5. Results

The K-Means algorithm classifies the Boroughs of London into 3 clusters which can be broadly described as follows:

1. Cluster 0 - Suburbs (12 Boroughs)

Boroughs with fewer than 10 Foursquare tagged eateries and most of these eateries appear to be Italian Restaurants. Average age in this cluster is 38.

2. Cluster 1 - Somewhere in the middle (17 Boroughs)

Majority of the Boroughs fall under this cluster. These are boroughs with 10-25 Foursquare tagged eateries on average. Some of the Boroughs here are in the center of the city and these tend to have a higher number of eateries and also higher population. The most commonly appearing eatery categories in this cluster are Bakeries, Cafes and Italian Restaurants. Average age in this cluster is 34.

3. Cluster 2 - Busy and Bustling (4 Boroughs)

These are the four boroughs with more than 30 Foursquare tagged eateries. Also, these are boroughs with high density of eateries, both with respect to area and population. The most commonly appearing eatery category here is Cafe. Average age in this cluster is 38.

6. Observations & Recommendations

- In general, the median age of a borough or its population density does not seem to have any major effect on the clustering result. Cluster assignment seems to be influenced mostly by the number of eateries and the dominant eatery category in the borough.

-
- Suburban London (mostly in Cluster 0) is underserved with respect to eateries and could certainly be benefited by the introduction of new food delivery services/ restaurants. Considering Italian restaurants are mostly popular in this cluster, hence businesses can choose to either “ride the wave” and invest more in Italian cuisine, or otherwise take the approach of choosing a cuisine/ eatery category which is rare to find here.
 - Central London is, as expected, swamped with eateries of all kinds. Most of these neighbourhoods fall under Cluster 2. Food delivery businesses will have to deal with high volumes of orders during peak hours here, compounded by the fact that delivery times will be affected by traffic congestion which is synonymous with central districts of all major cities of the world.
 - Cluster 1 represents neighbourhoods with possibly the maximum potential for food delivery expansion. It includes the majority of the boroughs and hence represents a cross section of the city. These boroughs will typically not experience as much traffic congestion as the Central boroughs in Cluster 2, and at the same time, have enough established eateries for food delivery businesses to take advantage of.

7. In conclusion: Future Directions

- Foursquare data in this study was collected using a Foursquare personal account. However, if one has a paid developer account, more information about venues, visitor profiles, ratings etc will be easily available and this will enhance the breadth of the study.
- Data pertaining to current food delivery statistics, listings of restaurants with tie-ups with these platforms etc are available online, but all of these come at a price. With this additional data, one could better estimate the likelihood of a person in a particular neighbourhood of London ordering in.

-
- Foursquare data tells us that Coffee Shops are the most dominant eatery venue across London. We have excluded coffee shops from our analysis here because, traditionally, coffee shops do not fall under the umbrella of eateries from where food is ordered-in. However, in recent times, coffee shops are known to have started tying up with food delivery services. This in itself could be a major disrupter for food delivery and coffee shops alike.