**Coursera Capstone Project - Battle of Neighbourhoods**

# PROSPECTS OF NEW RESTAURANT BUSINESS IN LONDON

****

Vidya Rani Puliyangodan Kallaiveedu

Table of Contents

**Introduction3**

Background3

Business Problem3

**Data**3

Dataset and Workflow3

Data Cleaning4

Feature Selection5

**Methodology - Exploratory Data Analysis5**

Exploring London Neighborhood5

Exploring London Restaurants6

**Results7**

**Discussion10**

**Conclusion11**

1. **INTRODUCTION**
   1. **BACKGROUND**

Restaurant business is thriving in London due to its diverse culture and London being a hotspot for tourism. Being a Londoner, I have experienced the trouble of finding cuisines of my choice in certain part of the city, leading me to wonder how prosperous the business will be if they carefully plan the location before starting a venture. However rather than going with guess work on the location or cuisines, this project is to apply Machine learning on the data available online to analyse the hotspots for new restaurant locations/cuisines.

* 1. **BUSINESS PROBLEM**

London is the [capital](https://en.wikipedia.org/wiki/Capital_city) and largest city of [England](https://en.wikipedia.org/wiki/England) and the [United Kingdom](https://en.wikipedia.org/wiki/United_Kingdom). It is home to many people from different cultures, and the city has a flourishing restaurant scene. The restaurant industry in London is growing exponentially; every street is filled with every variety of restaurant, fast food place, cafes and pub. In such a competitive market, it is of utmost importance that the entrepreneurs wishing to open up a new restaurant business should tactically plan before choosing the location and cuisine.

Below given are the few factors to consider before starting a restaurant in London:

* Population(/and demography) in the particular location
* Are there any Tourist attractions venues nearby
* Competitors in the location, Menu of the competitors

Etc.

**The business problem we are trying to solve** is – If an investor is interested in opening a new eating spot in the capital city London –the model should be capable of providing insight on what would be the best profitable type of restaurant for a given location? Or what would be the best location for a specific type of cuisine?

**Target audience -** is entrepreneurs or investors considering to start a new restaurant or franchise requiring guidance in picking the locality for a particular gastronomy OR cuisine for definite location.

1. **DATA**
   1. **DATASET AND WORKFLOW**

* London boroughs and locations data from <https://en.wikipedia.org/wiki/List_of_areas_of_London> (Wikipedia). Locations with “Post Town” – London are selected for analysis.
* Use the Geopy and Folium library to get the coordinates of every location and map geospatial data on a London map.
* Foursquare API is used to collect the top 200 restaurants and their categories for each location within a radius 500 meters.

Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters based on categories of restaurants residing in and around the neighbourhoods. And analyse the top 10 most common restaurants in each cluster.

**APIs:**

**• Foursquare API:** This API has a database of more than 105 million places. This project would use Four-square API as its prime data gathering source. Many organizations are using to geo-tag their photos with detailed info about a destination, while also serving up contextually relevant locations for those who are searching for a place to eat, drink or explore. This API provides the ability to perform location search, location sharing and details about a business. Foursquare users can also use photos, tips and reviews in many productive ways to add value to the results.

Folium- Python visualization library would be used to visualize the neighbourhoods cluster distribution of London over an interactive leaflet map.

**Python packages and Dependencies:**

• Pandas - Library for Data Analysis  
• Geopy – To retrieve Location Data  
• Requests – Library to handle http requests

• Wikipedia – to Scrape the data from Wiki  
• Matplotlib – Python Plotting Module  
• Sklearn – Python machine learning Library  
• Folium – Map rendering Library

* 1. **DATA CLEANING**

Data downloaded or scraped from Wikipedia and Geopy sources were combined into one table. The process starts with scrapping the Wikipedia web page containing information about London boroughs and neighbourhoods present in each borough using library Wikipedia, the data is filtered with Post Town column(value – ‘LONDON’) from the dataset. During the process Pandas data frame is constructed with three rows having boroughs, corresponding neighbourhood and post town information.

Two new columns are added to store the coordinates of the location (Longitude and Latitude), co-ordinates are fetched using Geocoder library. Records with no coordinate’s information are removed from the final dataset. (The data frame has **32 boroughs and 299 neighbourhoods**).

****

* 1. **FEATURE SELECTION**

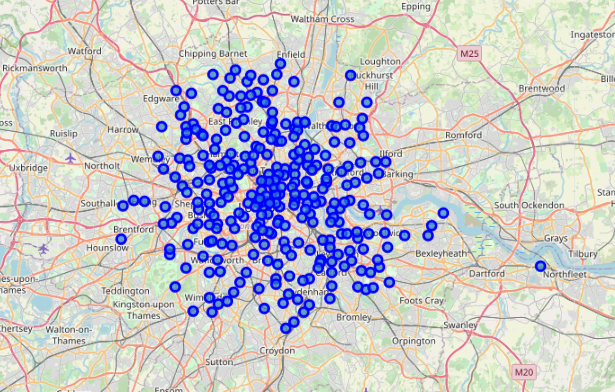
From the location dataset below features are dropped as they don’t add value to the analysis

* Postcode district
* Dial Code
* OS grid ref

Foursquare API is invoked to obtain nearest venues for each of the location in the above dataset. The API is call is limited to restaurants as the analysis is to recommend location or cuisine for the diner business. The dataset returned included Cafes and Bakeries, thus additional criteria is added to limit the characteristic of the venues to Restaurants.

1. **METHODOLOGY- EXPLORATORY DATA ANALYSIS**
   1. **EXPLORING LONDON NEIGHBORHOODS**

All of London’s neighbourhoods (/Locations) and borough information’s are scraped from Wikipedia page by applying filter on Post town to ‘LONDON’ for this project. This data is then explored using Geopy and geospatial libraries to retrieve geographical coordinates of these locations. The data is cleaned by dropping duplicates and records with no coordinate’s information. . Final data frame has **32 boroughs and 299 neighbourhoods**. The folium libraries are used to create a map to visualize neighbourhoods of London.



All the blue indicators on the map above are for neighborhoods. It is apparent that the city is more congested at the center and widespread in the outskirts.

* 1. **EXPLORING LONDON RESTAURANTS**

Foursquare API is used to explore the neighborhoods and segment them by fetching top 200 restaurants in each neighbourhood.

A dataframe is created with all the restaurant venues information provided by foursquare for the given coordinate values. Foursquare API returns 6118 results with 126 distinctive categories. The result set included Cafes and Bakeries. Filter is applied to extract venues with text ‘Restaurant’ in it. I.e. ‘Indian restaurant’, ‘Italian restaurant’, etc. Post mining only restaurants from venue category list, the resulting data frame lists 3342 restaurants88 unique categories of cuisines available in London. One hot encoding is performed on the resulting data frame for each neighbourhood. **The results have 3342 unique restaurants in London with 88 different styles of cuisines.**

Rows are grouped by neighbourhood to determine the frequency of occurrence of each restaurant. A new data frame is created with each row assigned for neighbourhood and its corresponding top ten common restaurants based on cuisine.

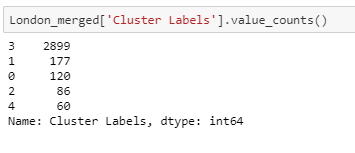


Finally, k-means clustering is performed on the data frame to check the pattern for each neighbourhood and get the information about the top ten common restaurants for each neighbourhood. K-means is an unsupervised machine learning algorithm for clustering unlabelled data. For this project, best value of K is identified using elbow method and silhouette score from Sklearn metrics. The best value returned is K=5 for the model.



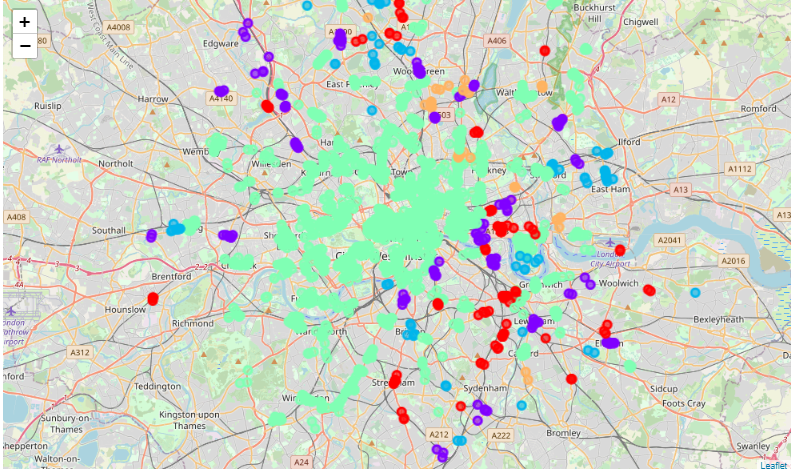
1. **RESULTS**

All 5 clusters follow unique pattern for top ten common restaurants for a particular neighborhood.



Based on the above (Neighborhoods assigned to each cluster) data Cluster 4 has the highest neighborhoods of 2899 and cluster 5 has the least with 60.

The clustered map is shown below:



**Cluster #1:**

The resulting data frame for cluster 1 is shown below:

****

Summary of the cluster 1 dataframe:-



The most common restaurant: Chinese Restaurant

10th most common restaurant: Filipino Restaurant

**Cluster #2:**

The resulting data frame for cluster 2 is shown below:

****

Summary of the cluster 2 dataframe:-



The most common restaurant: Fast Food Restaurant

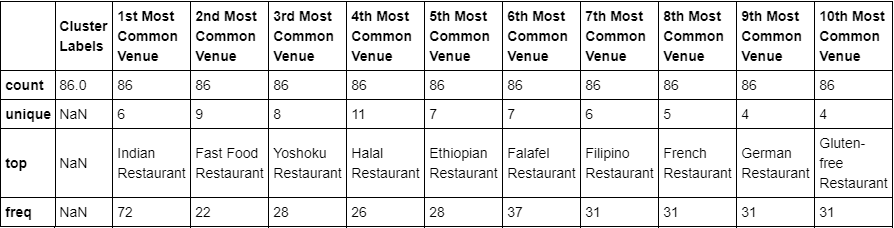
10th most common restaurant: French Restaurant

**Cluster #3:**

The resulting data frame for cluster 3 is shown below:

****

Summary of the cluster 3 dataframe:-



The most common restaurant: Indian Restaurant

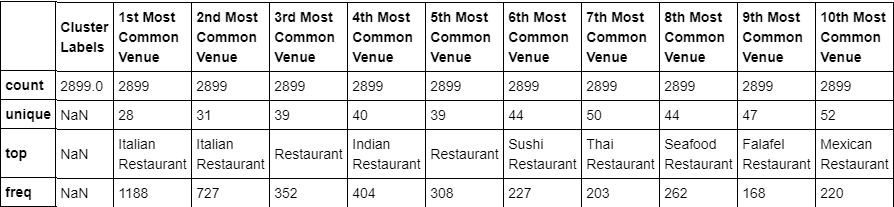
10th most common restaurant: Gluten-free Restaurant

**Cluster #4:**

The resulting data frame for cluster 4 is shown below:

****

Summary of the cluster 4 dataframe:-

****

The most common restaurant: Italian Restaurant

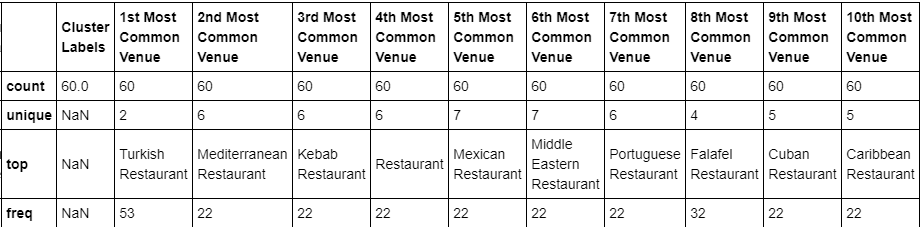
10th most common restaurant: Mexican Restaurant

**Cluster #5:**

The resulting data frame for cluster 4 is shown below:

****

Summary of the cluster 5 dataframe:-

****

The most common restaurant: Turkish Restaurant

10th most common restaurant: Caribbean Restaurant

1. **DISCUSSION**

The analysis can be summarized to convey – the safest option to choose the restaurant type for a particular locality is by considering the least common type. Opening a new Indian restaurant in a street with dozens of Indian dining place makes no sense. Looking at the competition, which will be a risky investment. However, choosing the least popular restaurant is also unreliable, given the demand and popularity for that type is food is less in that area. Hence the analysis can be made superior by including the analysis of population (/demography). The analysis is limited to 10 common venues to eliminate the risk of opening a business with no prospect in a location

Stakeholder can

* Either choose a location from the cluster and this model can advise what would be the best cuisine to opt for in the locality
* Or choose a gastronomy , and we can check the model to find out the best suited cluster/locality for that type

Final result on the analysis for each cluster is:

|  |  |  |
| --- | --- | --- |
| **Cluster** | **Most Common Restaurant** | **Least Common Restaurant** |
| 1 | Chinese Restaurant | Filipino Restaurant |
| 2 | Fast Food Restaurant | French Restaurant |
| 3 | Indian Restaurant | Gluten-free Restaurant |
| 4 | Italian Restaurant | Mexican Restaurant |
| 5 | Turkish Restaurant | Caribbean Restaurant |

Recommendations, based on description of each cluster: 10th most common restaurant type can be considered while considering the gastronomy for a certain locality.

|  |  |
| --- | --- |
| **Cluster** | **Least Common Restaurant** |
| 1 | Filipino Restaurant |
| 2 | French Restaurant |
| 3 | Gluten-free Restaurant |
| 4 | Mexican Restaurant |
| 5 | Caribbean Restaurant |

1. **CONCLUSION**

The analysis is performed to find out the most promising category of restaurant and appropriate location to start the business. The data is collected from Wikipedia, geospatial library and foursquare API. The method used is K-means clustering to group the similar neighbourhood on the basis of frequency of certain cuisine types.

Undoubtedly the model can be improved by collecting further information about the neighbourhood population (/demography), restaurant menu and ratings to come up with improved recommendation.

The nature of this analysis can be applied to any city of your choice that has available geospatial information.

Also, the analysis can be extended to any kind of venue (shopping, clubs, etc.) that is available in

Foursquare database.

.