Applied Data Science Capstone

Finding the best place to open a new restaurant in London

by

Pegah Bangiantabrizi

Aug. 2020

1- Introduction:

1-1- Background:

London is the capital and largest city of the United Kingdom. The city stands on the River Thames in the south-east of England. The restaurant and leisure industry in London is growing exponentially, every street is filled with every variety of restaurant, fast food place, pub, bar, etc; every type of food is available from the classic European cuisines, primarily Italian and French, to more exotic foods originating from Asia or South America. The demand in the culinary industry has become very high, and as a result, has the extent of competition, in order to open a restaurant in a trendy area of this city, like the Shoreditch area.

1-2- Business Problem:

Owning and running your own restaurant business is many people's dream; at the same time, restaurants are a difficult business to own or operate. Many factors will contribute to where you decide your premises to be. It can be challenging to find a venue that will factor in all of your conditions, so you will likely have to compromise on a few things. However, the primary factor that you should consider is to find an appropriate location for this new restaurant and, as well as to be aware of the possible nearby and local competitors.

In this work, we will implement the fundamental analysis, in order to find the optimal London Borough, in which the restaurant can be opened. This is conducted according to the criteria. That there are many additional factors regarding this, but the analysis can be done after choosing the Borough, and thus will not be done within the scope of this project.

1-3- Interest:

The target audience of this work would be anyone with an interest to get involved in the leisure and restaurant business and industry within the London districts.

2- Data

2-1- Data description:

To find a solution to the problem, and to be able to build a recommender model, first we consider the following:

1) Its geographical coordinates (i.e., latitude and longitude) to find out where exactly the venues are located.

and

2) In order to access the location of a restaurant, its latitude and longitude should be known, so that we can point to its coordinates, and further create a map displaying all the restaurants with its labels, respectively.

2-2- Data Cleaning:

Based on the criteria listed above, the following data is utilized in this analysis:

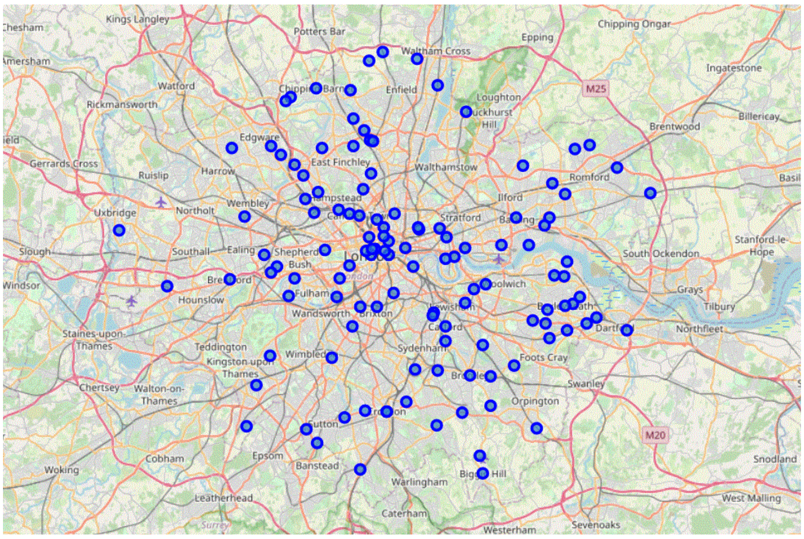
- The deployment of BeautifulSoup, in order to extract the data from Wikipedia, and to further provide the relevant information on the London boroughs, i.e., also known as the local authority districts. Besides, the local areas or neighbourhoods are considered for each borough for the detailed analysis.

- The Foursquare API https://foursquare.com/, to extract the relevant information on the available restaurants, for a given neighbourhood and borough in London. This API also provided information about the restaurant styles based on cuisine.
- The utilized data provided by the UK Government available at data.london.gov.uk to get detailed insights on the London boroughs.

3- Methodology:

After the data has been collected, it is processed into pandas data frames and explored. Folium Python visualisation library has been used to visualize the neighbourhood's distribution over the maps of London. Extensive comparative analysis of the neighbourhoods of London has been performed. Finally, unsupervised machine learning algorithm k-means clustering applied to form the clusters of different categories of places in the above neighbourhoods and visualize the data.

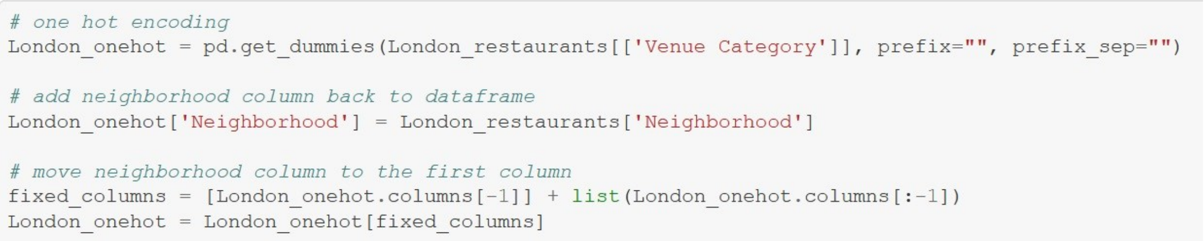
By the use of Geopy and Folium libraries coordinates of all the locations is achieved and mapped geospatial data on the London map.



The blue markers on the map above show the neighbourhoods and indicate that the city is more congested at the centre and widespread in the outskirts.

3-1- Feature Extraction:

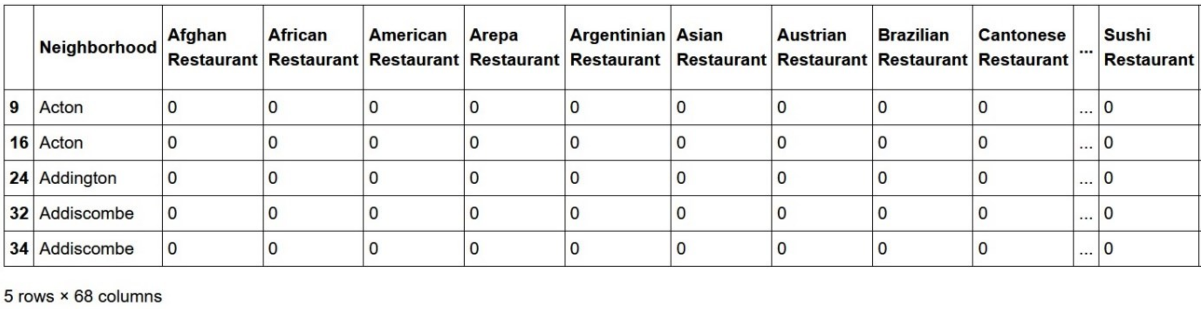
For feature extraction, One Hot Encoding is used in terms of categories. Therefore, each feature is a category that belongs to a venue. Each feature becomes binary. Then, all the venues are grouped by the neighbourhoods. This give us a venue for each row and each column will contain the frequency of occurrence of that particular category.



```
# one hot encoding
London_onehot = pd.get_dummies(London_restaurants[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
London_onehot['Neighborhood'] = London_restaurants['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [London_onehot.columns[-1]] + list(London_onehot.columns[:-1])
London_onehot = London_onehot[fixed_columns]
```

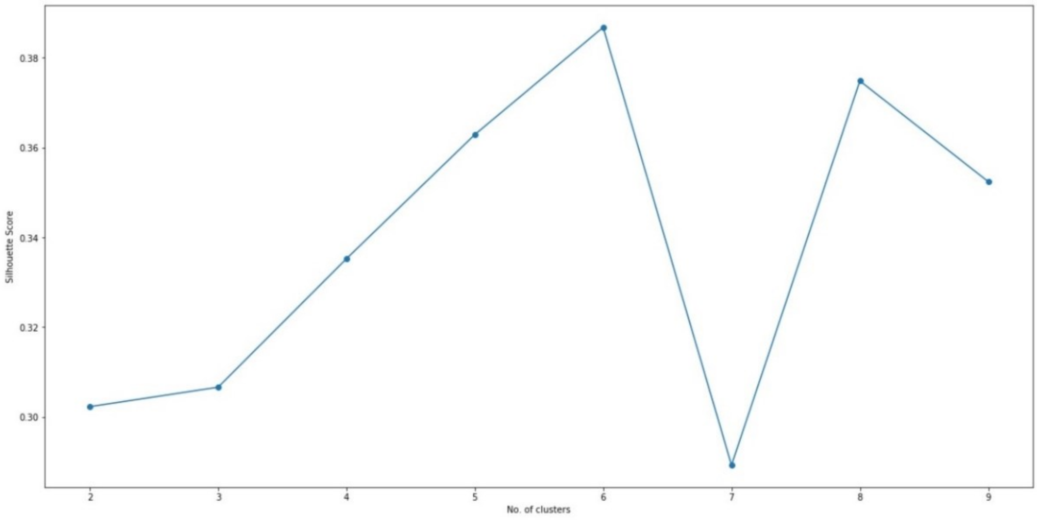


	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Austrian Restaurant	Brazilian Restaurant	Cantonese Restaurant	...	Sushi Restaurant
9	Acton	0	0	0	0	0	0	0	0	0	...	0
16	Acton	0	0	0	0	0	0	0	0	0	...	0
24	Addington	0	0	0	0	0	0	0	0	0	...	0
32	Addiscombe	0	0	0	0	0	0	0	0	0	...	0
34	Addiscombe	0	0	0	0	0	0	0	0	0	...	0

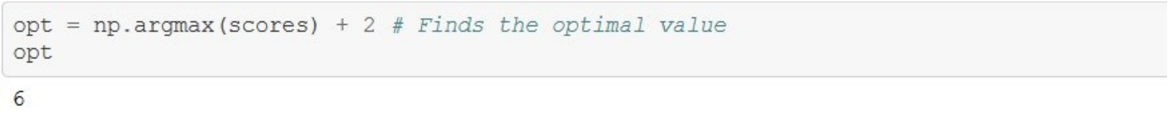
5 rows × 68 columns

3-2- Unsupervised Learning:

The K-Means is a clustering algorithm which search clusters within the data and its main objective function are to minimize the data dispersion for each cluster. Thus, in this work, K-Means is implemented in order to cluster restaurants and analyse the top most common restaurants in each cluster.

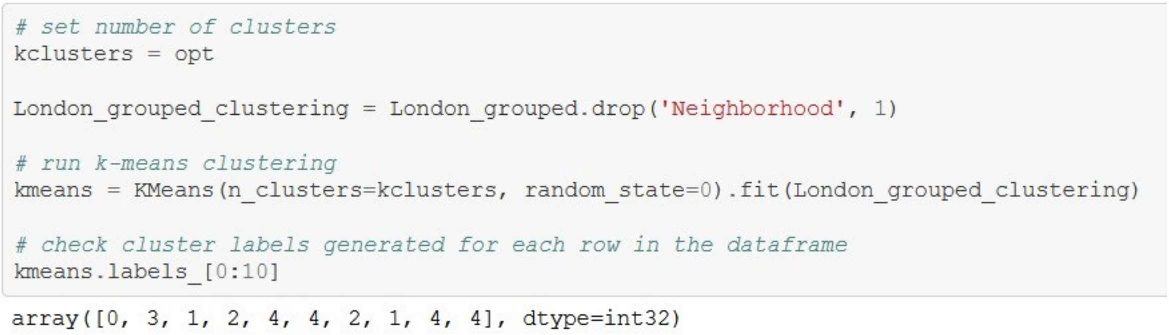


Also, the *elbow method* is implemented in order to reach the number of clusters (6 Clusters) which is used in this analysis prior to clustering the data.



```
opt = np.argmax(scores) + 2 # Finds the optimal value
opt
```

```
6
```



```
# set number of clusters
kclusters = opt

London_grouped_clustering = London_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(London_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([0, 3, 1, 2, 4, 4, 2, 1, 4, 4], dtype=int32)
```

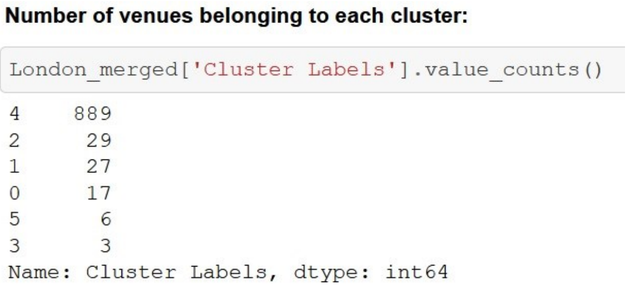
4- Results:

By the use of K-Means neighbourhood as clustered based on the mean occurrence of venue category, the results indicate all 6 clusters follow a unique pattern for the top ten common restaurants for a particular neighbourhood:

- ➢ *The results show that there are 971 restaurants in London with 67 different style of cuisines.*
- ➢ *The detail shows the number of neighbourhoods assigned to each cluster.*

- *Cluster 4 indicates neighbourhoods with the highest concentration of restaurants with the amount of 889*

- *Cluster 3 indicates neighbourhoods with the least number of restaurants with the amount of 3.*

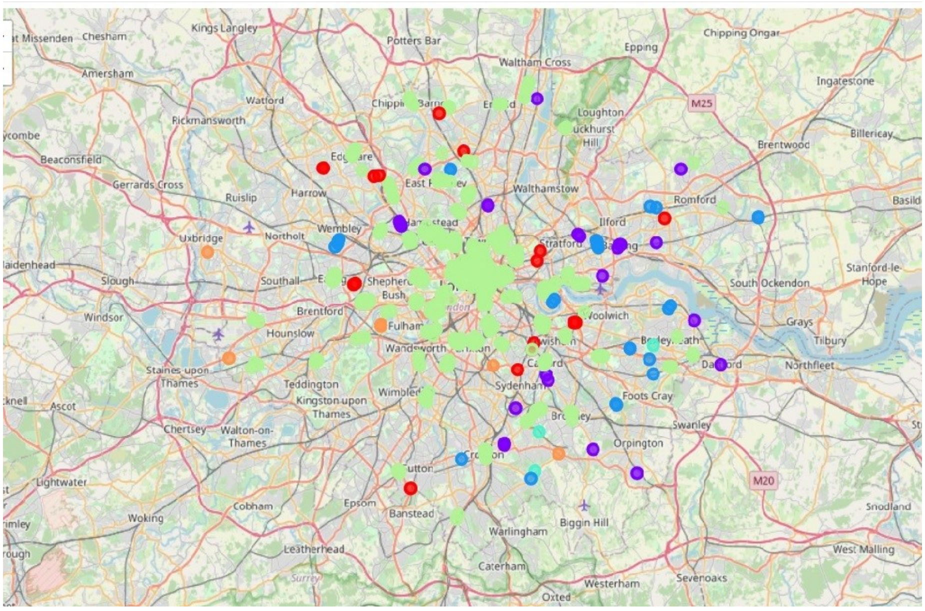


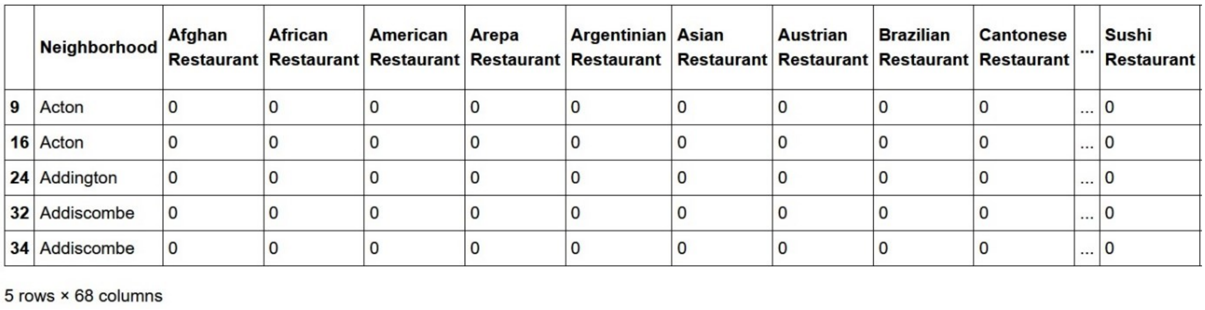
Number of venues belonging to each cluster:

```
London_merged['Cluster Labels'].value_counts()
```

```
4    889
2     29
1     27
0     17
5      6
3      3
Name: Cluster Labels, dtype: int64
```

The clusters are visualized on the map below, thus showing the best locations for opening the chosen restaurant.





	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Austrian Restaurant	Brazilian Restaurant	Cantonese Restaurant	...	Sushi Restaurant
9	Acton	0	0	0	0	0	0	0	0	0	...	0
16	Acton	0	0	0	0	0	0	0	0	0	...	0
24	Addington	0	0	0	0	0	0	0	0	0	...	0
32	Addiscombe	0	0	0	0	0	0	0	0	0	...	0
34	Addiscombe	0	0	0	0	0	0	0	0	0	...	0

5 rows × 68 columns

5- Discussion:

The analysis of the results shows that London overall has a high frequency of restaurants within its neighbourhoods, In terms of the clustered data, cluster 3 has the lowest overall frequency of restaurants by comparison to other groups. Cluster 4 is shown to have the highest frequency of grocery stores by a large margin when compared to other clusters, making it a possible outlier cluster group for a restaurant business in the mentioned neighbourhood.

By further looking at the frequency of restaurants based on locations, the most common restaurants in each cluster have been located and as a result, The best boroughs to open a restaurant with a specific cuisine (i.e., most favourite cuisine of the borough) can be chosen.

The analysis recommends neighbourhoods in a specific cluster can be chosen to open a restaurant which is the least common in that specific region, which would result in profitability and less competition.

6- Conclusion:

The presented work has been conducted based on a small sample cluster, for the sole purpose of depicting the program's feasibility potential. It can further be extended into larger data sets, and as a result, the program would result in more output results and findings. The outcomes would also provide the relevant information on the number and intensity of the restaurants in certain boroughs, which can be interpreted as which areas would be of higher demands in terms of the popular regions for the intended cuisine and leisure industry. Moreover, a number of other factors regarding these places can be included in order to distinguish or narrow down the results, such as the restaurants' type (e.g., Italian, Asian, etc.) or hygiene rate.