**New Restaurant in London**

**(Data Science Capstone Project)**

1. **INTRODUCTION**

A new investor who living in London has contacted me looking to open a restaurant in her place. Nevertheless, she doesn’t know in which neighborhood the restaurant would gain more focus from the customers and face lesser competition from other restaurants. Since it is very important to choose one of the most suitable neighborhood and this is an essential step for the success of the business in the future, I would like to use the knowledge in this course to help her by conducting some useful analysis.

1. **DATA GATHERING**

I will collect the data from three major sources, and use the following steps to solve the problem. First, from Wikipedia webpage (<https://en.m.wikipedia.org/wiki/List_of_areas_of_London>), which contains the information of postal codes, boroughs and neighborhoods of London, so that I can extract geography information of London. Second, obtain coordinates data from Geocoder Python package, which provide the longitude and latitude according to London, so that postal code and borough can be converted into their equivalent latitude and longitude values. Lastly, connecting to Foursquare API (<https://api.foursquare.com>), which contains most common venues according to the neighborhood, to take data and gather information about venues in each neighborhoods in London city and get the most common venue categories in each neighborhood. Finally, find the best recommendation neighborhoods for the investor.

1. **METHODOLOGY SECTION**

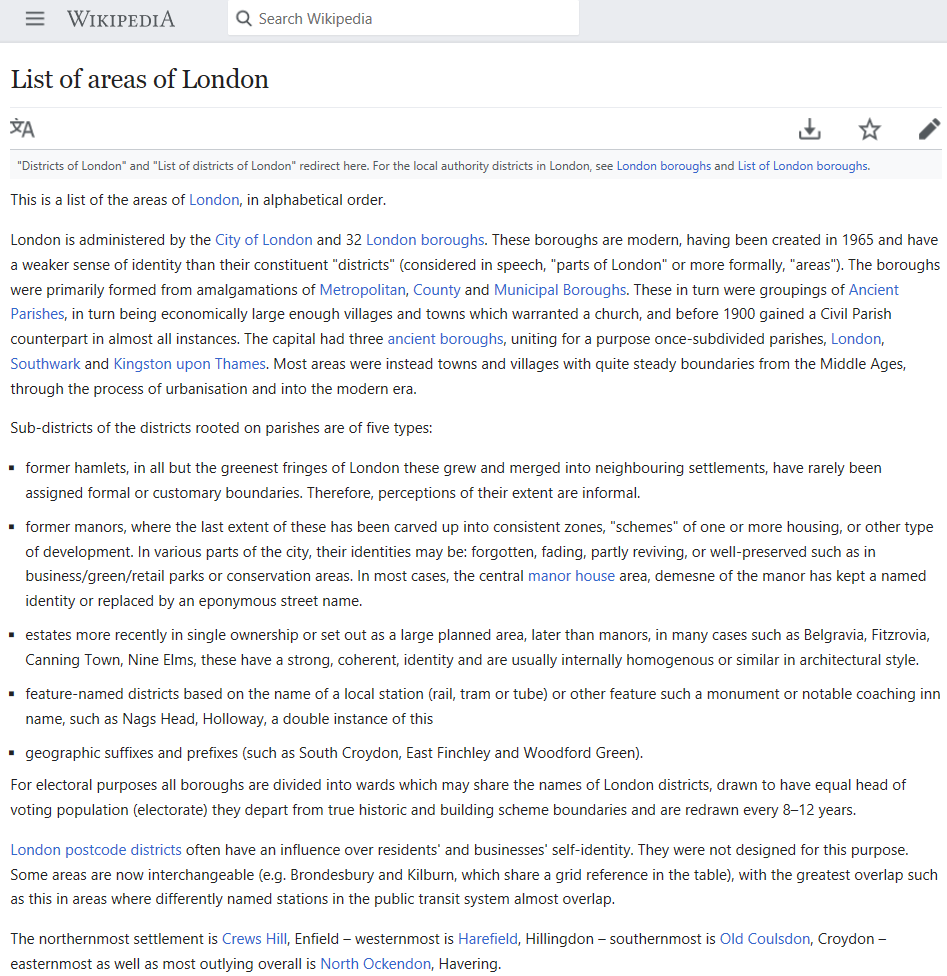
Python packages used:

Below are listed the Python packages used and the corresponding usages for the later work.

|  |  |
| --- | --- |
| Python package | Usage |
| Pandas | For data analysis |
| Numpy | For handle data in a vectorized manner |
| JSON | To handle JSON files |
| Requests | To handle HTML requests |
| Matplotlib | Python plotting module |
| Sklearn | Python machine learning library |
| Geopy | To retrieve location data |
| Folium | Map rendering library |

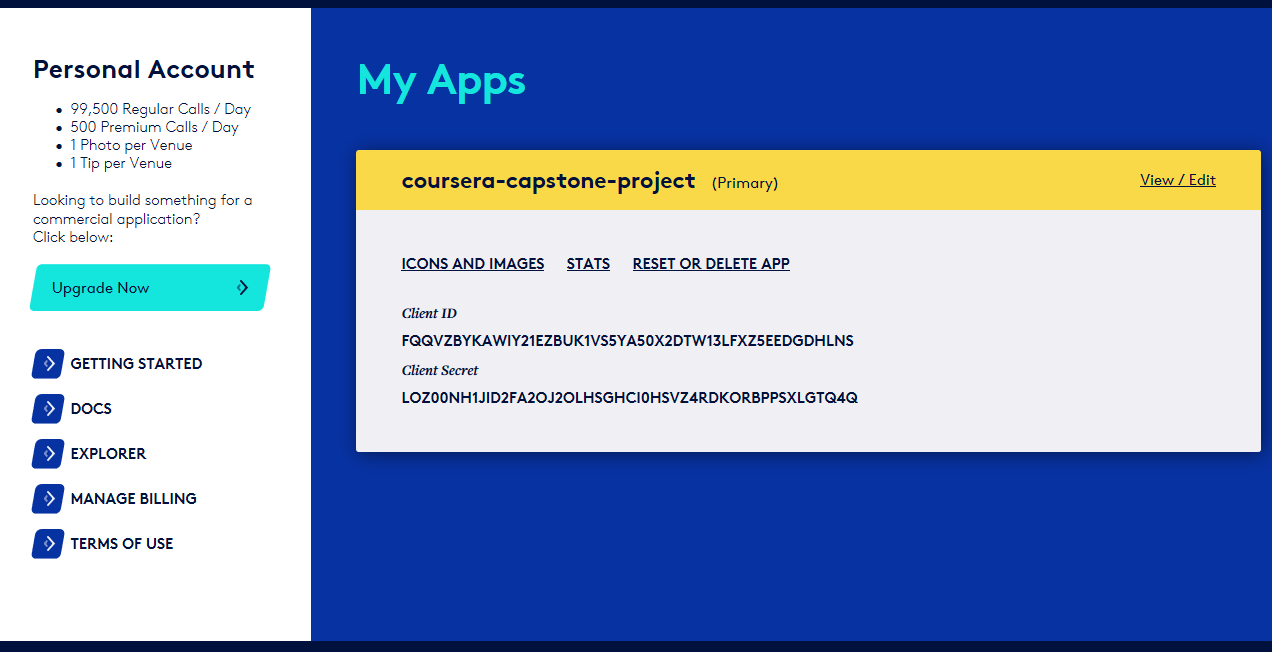
Flow of work:

1. I identified neighborhoods in London City. A List of London district names with Postal codes was used to find the list of neighborhoods and to match with Foursquare API . Webpage of Wikipedia (<https://en.m.wikipedia.org/wiki/List_of_areas_of_London>) was used to obtain the information. After that, I used the webpage to processed the table.



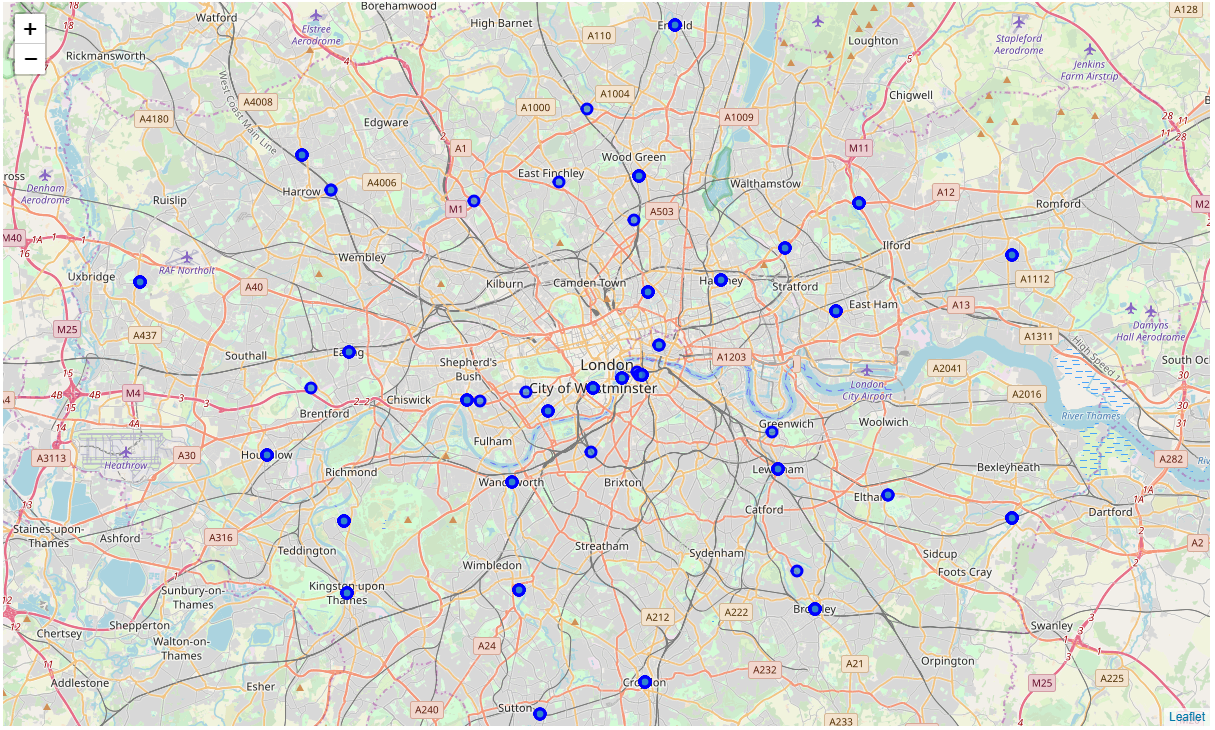
*Wikipedia page*

1. I entered my Client ID and Client Secret to connect to Foursquare and retrieved data of location for each venue in every neighborhood by using a limited regular account. After the list of neighborhoods were found, I connected to the Foursquare API (<https://api.foursquare.com>) and used its search feature to collect information about the nearby places of each of the neighborhood. I set the radius for each of them was equal to 500 m.



*Foursquare page*

1. I used Python visualization library to visualize the neighborhoods cluster distribution of London city over an interactive leaflet map.

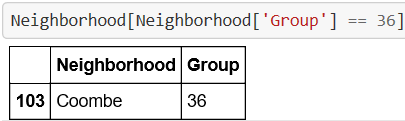


*Map of London*

1. I processed the retrieved data and the generated dataset was created to be a data frame for all the venues within London, with the postal code, names of boroughs, neighborhoods and their location coordinates (venue latitudes and venue longitudes), by using a python function getNearbyVenues(). I performed processing on the raw data and for each venue, I have found the desirable features after the step of data gathering was finished. Our main feature is the category of that venue. I included everything except for the competitors, which are the restaurants. After that, venue's category was to be the area’s hot spot and each venue would have their own featured columns. After that I have combined all stores columns to be total stores and have combined all shops columns to be total shops, with the simple assumption of equal amount of customer in all stores and all shops. For the purpose of analyzing in statistical way and the use of machine learning, the dataset was well prepared.
2. For the gathered data, I used K means clustering, which is one of the unsupervised machine learning algorithms to form the clusters of different categories of places residing in and around the neighborhoods. In this case, I segmented and clustered neighborhoods by using K means clustering method. Since London has 509 neighborhoods, I have tried different numbers of clusters and finally choose that 40 clusters are essential for us to handle the problem, so that to make sure that the number of neighborhoods in the best group becomes one only. I updated the dataset and then created a column for each neighborhood to represent the group when clustering was finished. After that, I present the result.
3. **RESULT AND DISCUSSION SECTION**

The centres of clusters have been used and their total stores and also total shops were compared. I recommended the investor for which the total sum of the centre of the group is the highest.

After applied the K means clustering method to the resultant dataframe to segment the data, and according to the calculation of the 40 group cluster (total stores, total shops and other related venues), there are 7 groups in total that the total sum is not less than 70. For the best three, the total sum of the 3rd best group, i.e. Group 23, is 75, while the total sum of the 2nd best group, i.e. Group 9, is 76 and the total sum of the 1st best group, i.e. Group 36, is 89. For the 1st best group, there is only one neighborhood called Coombe, which is the best location to open a restaurant in London.



1. **CONCLUSION SECTION**

In conclusion, based on the biggest volume of customers, which is positively correlated to the demand of eating out, the investor is recommended to consider Coombe as the optimal neighborhood to invest the new restaurant. Nevertheless, there may be some additional factors that the investor needed to be considered in the decision process, such as rental cost, availability and transportation, etc. Final decision on location of restaurant will be made by the investor based on the biggest volume of customers, budgets and other reasonable factors.

1. **LINK**
2. <https://en.m.wikipedia.org/wiki/List_of_areas_of_London>
3. <https://api.foursquare.com>
4. <https://github.com/matthewyst/final-capstone-project/blob/master/final_code.ipynb>