Final Report

Capstone Project - The Battle of The Neighbourhoods

Best neighbourhoods to open a new Asian Restaurant in London

Yan Carlomagno

Introduction:

London is one of the most vibrant and diverse cities in the world. Our stakeholders are looking into opening an Asian restaurant in the city, so choosing the best location is crucial to increase the restaurant's likelihood to succeed.

The objective is to help our stakeholders define the best borough, and neighbourhoods within that borough, to establish an Asian restaurant taking into account the Asian population, average income, property prices and competition from other restaurants.

This study is for educational purpose, as part of the IBM Data Science Professional Certificate's Capstone course, so it should not be taken as business advice also because there are several limitations on the depth of the data and analysis.

Data:

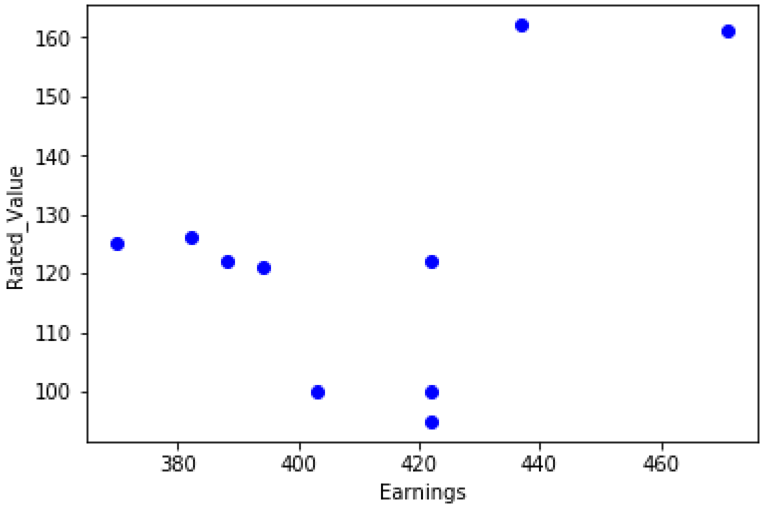
For our restaurant problem, we will focus on the Boroughs of London and work on getting the data from all the Boroughs. There are 32 London Boroughs with a population of around 150,000 to 300,000. To solve our problem of finding a best location to start an Asian restaurant in London, we need datasets based on various parameters such as:

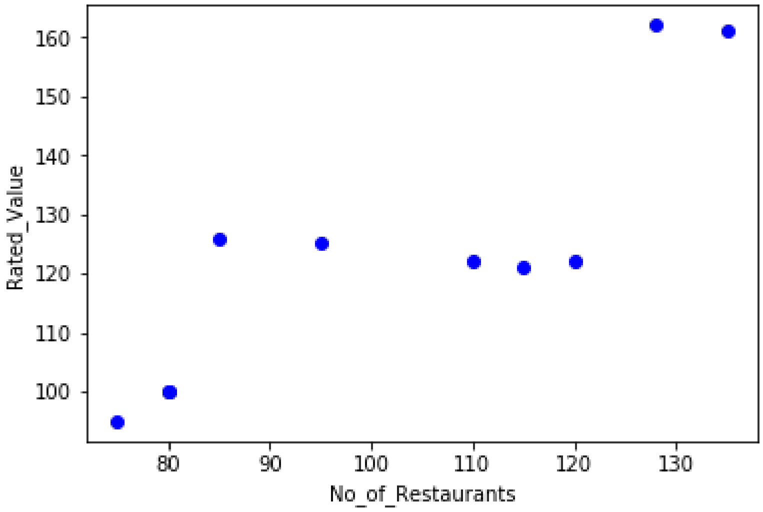
1. Population of target audience in all the boroughs of London based on their ethnicity
2. Earnings data of the working class living in the target location.
3. We need the data about the required business floorspace and rateable value statistics of each borough.
4. Considering the competitors factor, we also need the data of existing Licensed Restaurants in each borough.

All the above required information is available at **London Datastore**, which is a free and open data-sharing portal where anyone can access data relating to the city. The data is available in XLS and CSV format, which we can download and can use as-is for solving our problem.

Methodology:

To work on the solution, I have used Pandas library to read the data in XLS format and convert into pandas dataframe. Extensive data exploration analysis is done, where lot of data is cleaned and presented in a suitable format.
Machine Learning Algorithm **Simple Linear Regression** is used to predict the data for Rated Value for the year 2018 for the selected borough. The dependant variable would be the rated value for year 2018 and the independent variables are the earnings of each borough and the existing restaurants in each borough.

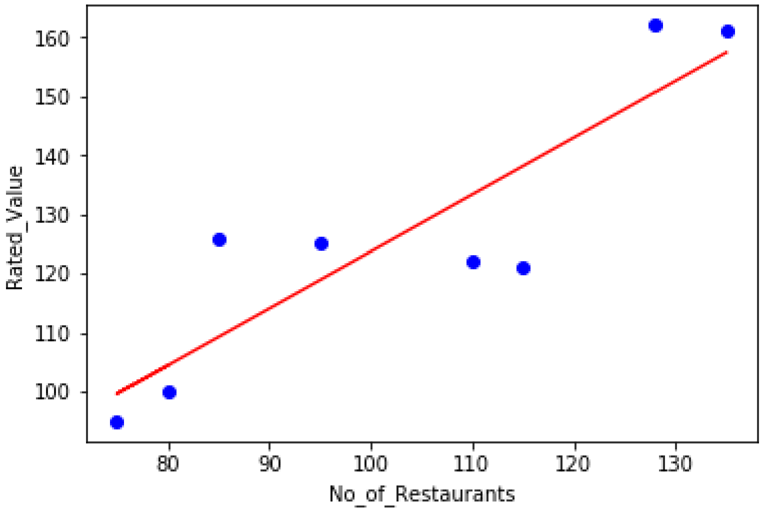




Two scatter plots are plotted between these variables and from these 2 plots, it is observed that the Linear relationship exists between the Restaurants and the Rated Value. A scatter plot clearly shows the relation between variables where changes in one variable explain or possibly cause changes in the other variable. Also, it indicates that these variables are linearly related.

Simple Linear Regression fits a linear model with coefficients Ø = (Ø1, Ø2, …. Øn) to minimize the residual sum of squares between the independent X in the dataset and dependant Y by the linear approximation.

Coefficient and Intercept in the Simple Linear Regression are the parameters of the fit line. Given that it is simple linear regression with only 2 parameters and knowing that the parameters are the intercept and slope of the line, using the python library SciKit Learn, we can estimate them directly from our data. The available data is divided into Train and Test data. The train data is used to train the model and the test data is used to evaluate the model.



Evaluation of the model is performed using the Evaluation Metrics such as **Mean Absolute Error**, **Mean Squared Error** and **R-Squared**. Due to very less available test data, the R-Squared for our model is not that great, but still we can consider our model for the prediction of the Rated Value for the year 2018.

After the prediction of rated value per sqm of a retail space is completed and when we are convinced that a particular borough will be the preferred location for the restaurant, we have to get the necessary data of that borough.

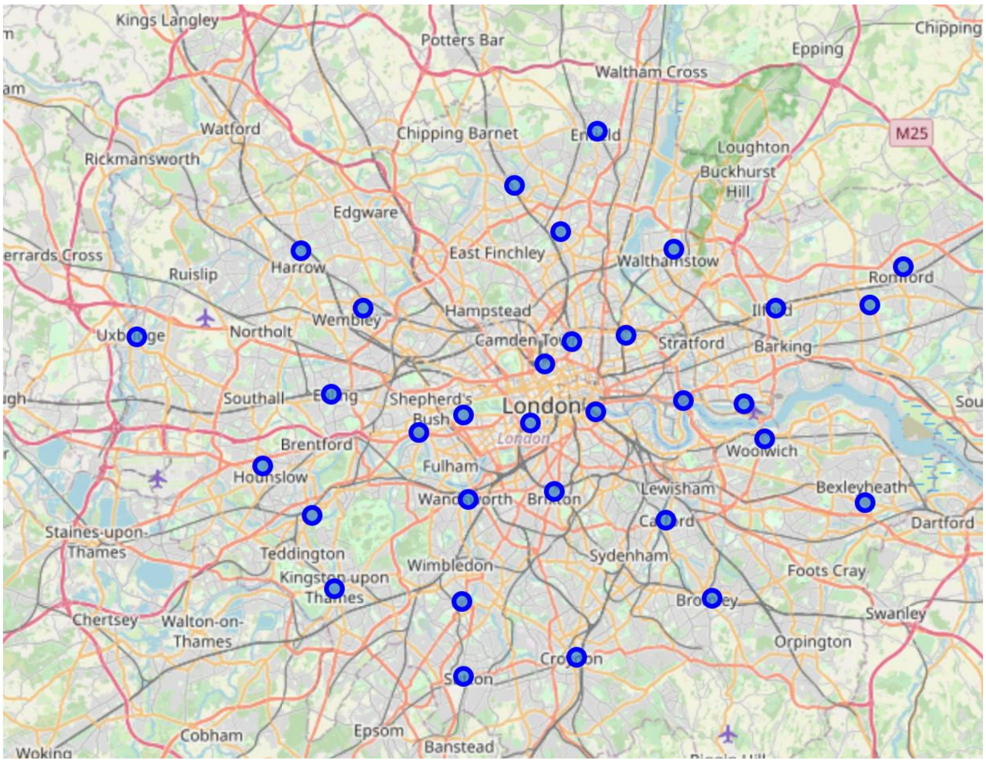
First we need to get the geo-coordinates of the borough and the geo-coordinates of the neighbourhoods of the borough from the web. I have used the Wikipedia pages to get this data.
https://en.wikipedia.org/wiki/List_of_London_boroughs
https://en.wikipedia.org/wiki/List_of_areas_of_London

To read data from these URLs, I have used the **requests**, **urllib** and **BeautifulSoup** libraries of python.

After I have the geo-coordinates information of the borough and its neighbourhoods, I need the other data such as the venues or places of the neighbourhoods, the venue categories, working hours and so on. All this data is called Location data, and to get this data I need a reliable and efficient location data providers and hence I am using Foursquare as the data provider. I have used the Foursquare API to explore the neighbourhoods in London city. I have also used the **Explore** function to get the most common venue categories in each neighbourhood and then use this feature to group the neighbourhoods into clusters. To cluster the neighbourhoods I am using **K-means Clustering** algorithm.

Geopy module and **Nominatim** library is used to convert a given address into the latitude and longitude values.

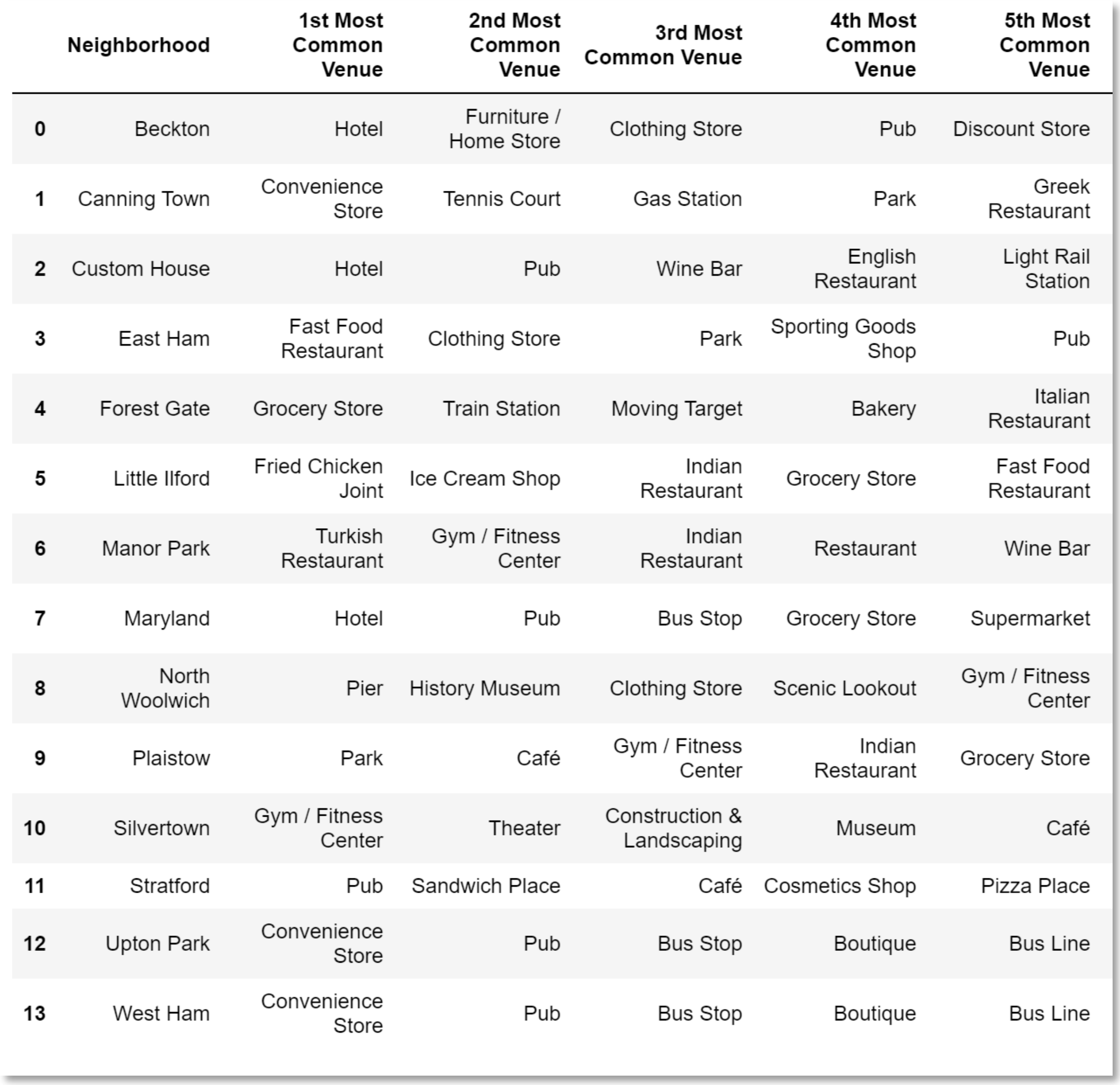
To visualize the neighbourhoods, the library **Folium** is used, to display the maps of London, with the boroughs super imposed on it and to display the map of borough with the neighbourhoods superimposed on it.



A python function **getNearbyVenues()** is created , to give the venue details like venue name, venue latitude, venue longitude, venue category along with neighbourhood name, latitude and longitude for each neighbourhood.

After the venue data for each neighbourhood of the Newham borough is generated , **One-Hot encoding** is applied on the venue category data, so that the analysis of the data will be easy in grouping the neighbourhoods based on the frequency of occurrence of each venue category.

Once the neighbourhoods are grouped based on the frequency of occurrence of the venue category, the top 10 venues of each neighbourhood are displayed as a dataframe.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Beckton	Hotel	Furniture / Home Store	Clothing Store	Pub	Discount Store
1	Canning Town	Convenience Store	Tennis Court	Gas Station	Park	Greek Restaurant
2	Custom House	Hotel	Pub	Wine Bar	English Restaurant	Light Rail Station
3	East Ham	Fast Food Restaurant	Clothing Store	Park	Sporting Goods Shop	Pub
4	Forest Gate	Grocery Store	Train Station	Moving Target	Bakery	Italian Restaurant
5	Little Ilford	Fried Chicken Joint	Ice Cream Shop	Indian Restaurant	Grocery Store	Fast Food Restaurant
6	Manor Park	Turkish Restaurant	Gym / Fitness Center	Indian Restaurant	Restaurant	Wine Bar
7	Maryland	Hotel	Pub	Bus Stop	Grocery Store	Supermarket
8	North Woolwich	Pier	History Museum	Clothing Store	Scenic Lookout	Gym / Fitness Center
9	Plaistow	Park	Café	Gym / Fitness Center	Indian Restaurant	Grocery Store
10	Silvertown	Gym / Fitness Center	Theater	Construction & Landscaping	Museum	Café
11	Stratford	Pub	Sandwich Place	Café	Cosmetics Shop	Pizza Place
12	Upton Park	Convenience Store	Pub	Bus Stop	Boutique	Bus Line
13	West Ham	Convenience Store	Pub	Bus Stop	Boutique	Bus Line

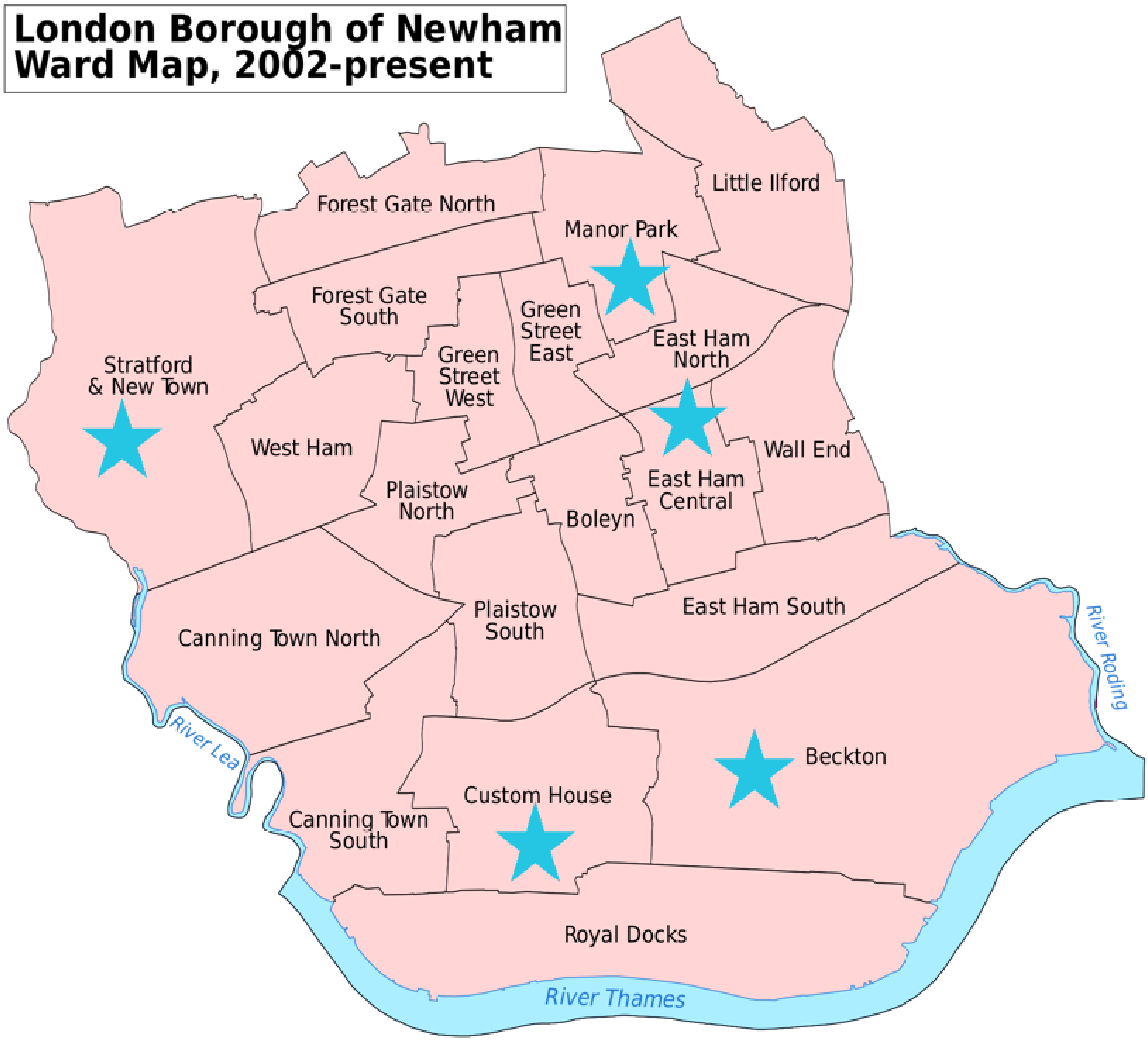
After all the above data exploration and analysis and top 10 venues of each neighbourhood are identified, the K-means Clustering algorithm is applied to the resultant dataframe to segment the data into 5 Clusters and all these 5 clusters are visualised in a map using the Folium library and finally the 5 clusters are examined to determine the discriminating venue categories that distinguish each cluster.

Results:

From the data sets of asian population, we found that **Newham borough has the highest Asian population from the rated value data set, and property costs in this borough are less compared to other boroughs.** These 2 factors influenced more on the decision of choosing Newham borough as the preferred location for our restaurant. The Newham borough has 146 existing restaurants and taking this as a independent variable (X variable) I have **predicted the rated value per sqm(dependant variable) between 160**

to 165 using the Linear Regression model. I have also calculated the MAE and R-Squared with the test data, though we got less values for these metrics due to less available test data.

In the Segmenting and Clustering section, the neighbourhoods of Newham borough are explored, and the top 10 venues of each neighbourhood are listed. The neighbourhoods are Clustered into 5 clusters using K-means algorithm and their most common neighbourhoods are identified. After applying the K-means algorithm the 5 neighbourhoods **Beckton, Custom House, Maryland (in Stratford), Eastham and Manor Park are identified as best neighbourhoods in Newham to open an Asian restaurant.**



Discussion:

My observation after doing this analysis is the model we used could have given better results if we had more data to train and test our model. In spite of that, this model gives us a better insight for our problem and also help us to gain better results. From the clustering results our problem finds a better solution of identifying the best location for the Asian restaurant. We could explore all the neighbourhoods of the borough and could list the most common venues based on their frequency of occurrence. From these results I can strongly recommend the Beckton, Custom house and few other neighbourhoods as a preferred location for our restaurant, as these areas have restaurants as the most common venue.

Conclusion:

There is always room for improvement and hence the above solution I have provided can also be improved and the machine learning models can be trained and tested for best results depending upon the data we have.

This study is for educational purpose, as part of the IBM Data Science Professional Certificate’s Capstone course, so it should not be taken as business advice also because there are several limitations on the depth of the data and analysis.