**Capstone Project - The Battle of Neighborhood**

**Applied data science capstone**

**Table of contents**

1. Introduction: Business Problem
2. Data
3. Methodology
4. Analysis
5. Results and Discussion
6. Conclusion

**1. Introduction: Business Problem**

In this project, we will try to find an optimal location for a new Asian restaurant in London. Specifically, this report will be targeted to stakeholders interested in starting an \*\*business for Asian restaurant in London, UK.

As Asians are increasing more and more in London, there are not enough stores for Asian restaurants. Since there are lots of restaurants for different countries' people in London, we will try to detect \*\*areas with no Asian restaurants in vicinity\*\*. We would also prefer locations \*\*as close to city center as possible\*\*, assuming that first two conditions are met.

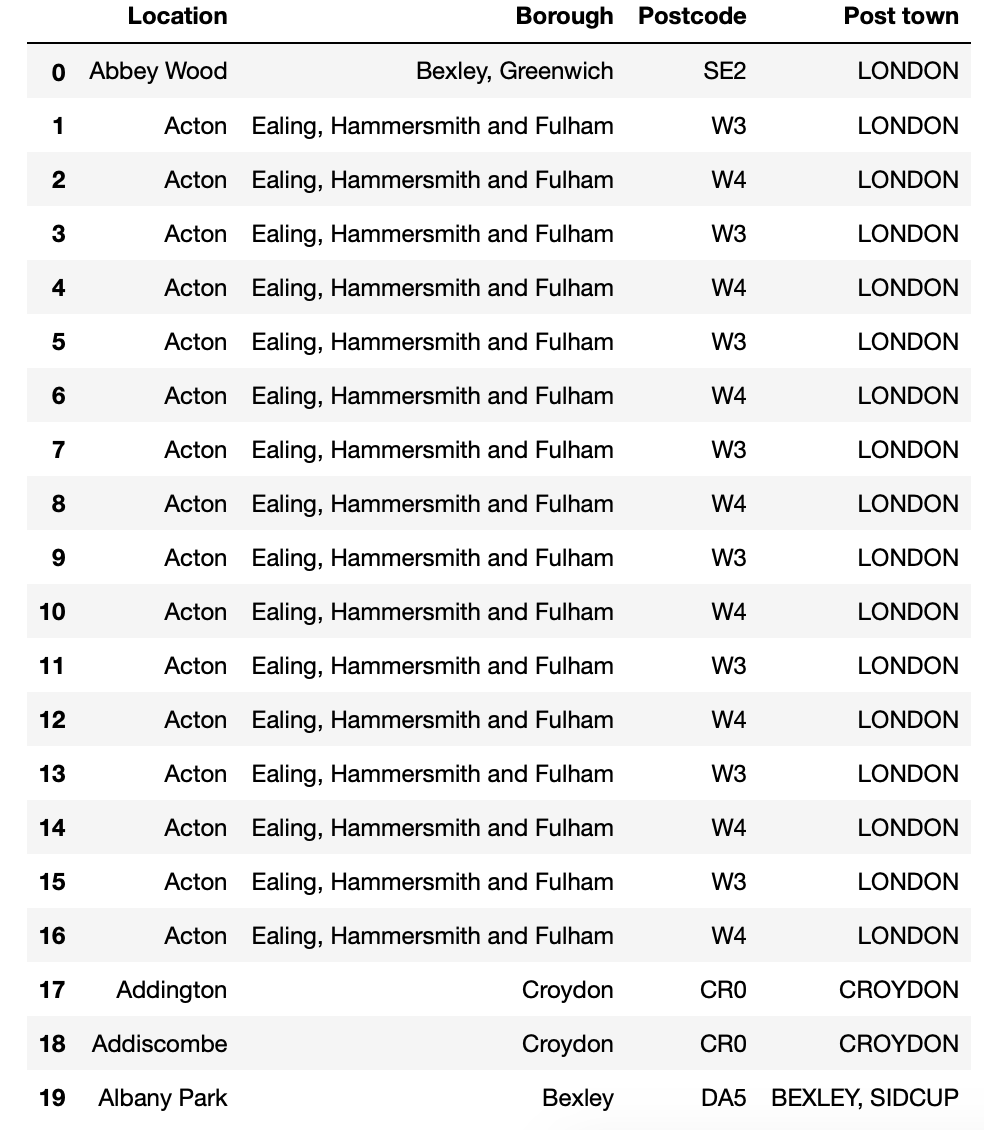
We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

**2. DATA Section**

I analyze the data based on the area data from Wikipedia page and the map data from google maps API and Foursquare API.

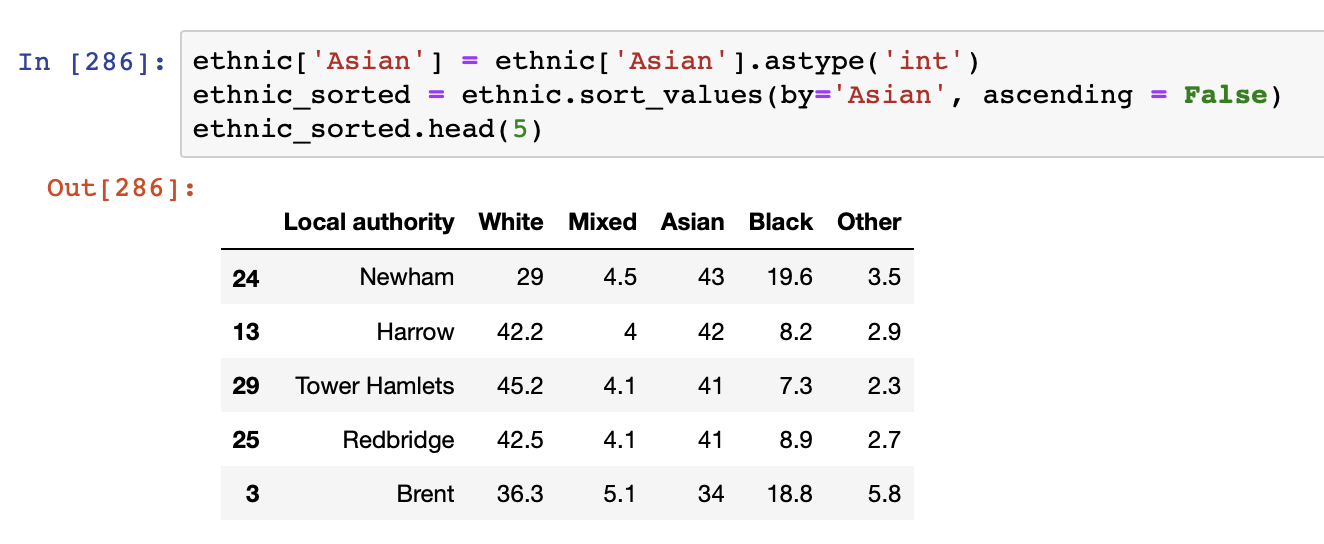
**London Area data:** Use the Notebook to build the code to scrape the following Wikipedia page, https://en.wikipedia.org/wiki/List\_of\_areas\_of\_London, in order to obtain the data that is in the table of postal codes and to transform the data into a pandas dataframe.

* There are different website scraping libraries and packages in Python. One of the most common packages is BeautifulSoup. Here is the package's main documentation page: http://beautiful-soup-4.readthedocs.io/en/latest/
* The package is so popular that there is a plethora of tutorials and examples of how to use it. Here is a very good Youtube video on how to use the BeautifulSoup package: https://www.youtube.com/watch?v=ng2o98k983k
* Use the BeautifulSoup package or any other way you are comfortable with to transform the data in the table on the Wikipedia page into the above pandas dataframe



**Ethnicity data:** Use the Notebook to build the code to scrape the following Wikipedia page, https://en.wikipedia.org/wiki/Demography\_of\_London, in order to obtain the data that is in the table of postal codes and to transform the data into a pandas dataframe.

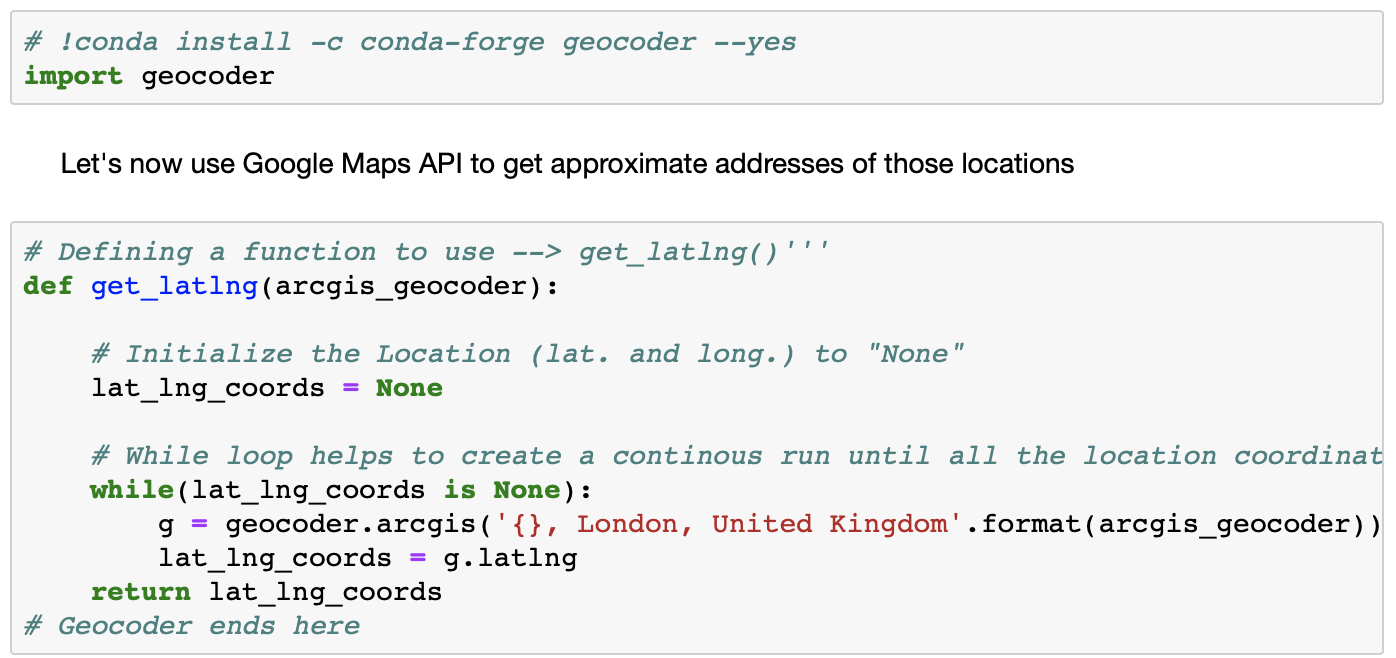
Using the proportions of ethnicity data in London, we can detect the areas where a lot of Asians live. And, we can find that the area with the highest Asian population in London is Newham, Harrow, Redbride, Tower Hamlets, and Hounslow.

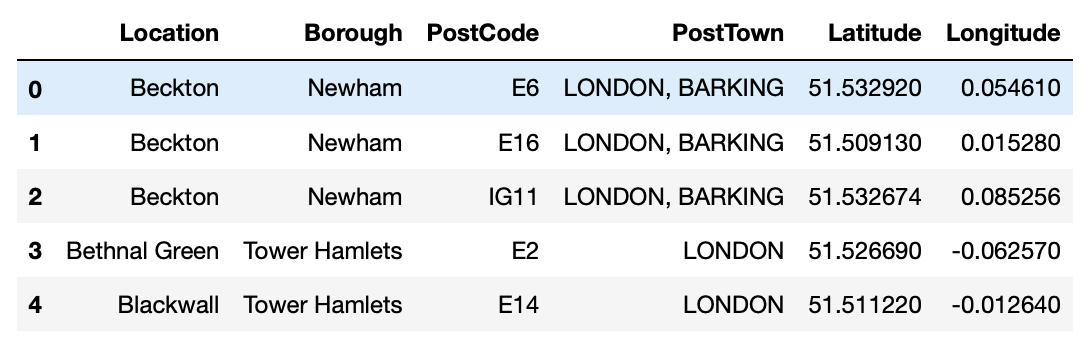


And, we can focus on the specific area and analyze that area in more detailed.

**Geographical Data:** Use the Geocoder package or the csv file to get the geographical coordinates of the neighborhood

* centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using **Google Maps API reverse geocoding**





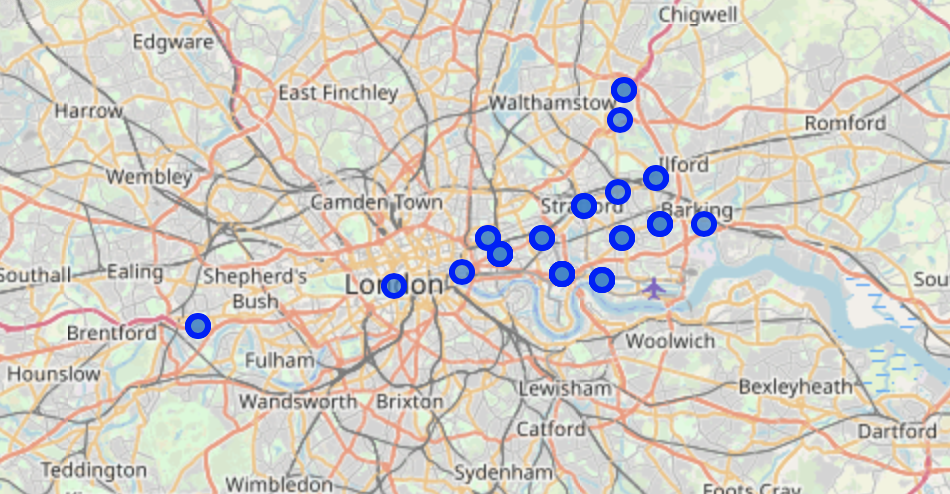
* number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**



* coordinate of Berlin center will be obtained using **Google Maps API geocoding** of London location

**Data visualization:** After the data preprocessing, we can create map of London using latitude and longitude values using the folium function. Folium is a great visualization library.

Feel free to zoom into the above map, and click on each circle mark to reveal the name of the neighborhood and its respective borough.



**3. Methodology**

In this project, we will direct our efforts on detecting areas of London. We will limit our analysis to area around city center. In first step we have collected the required data: location and type (category) of every restaurant near London center. And, we will focus on most promising areas and within those create clusters of locations. We will present map of all such locations but also create clusters (using \*\*k-means clustering) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

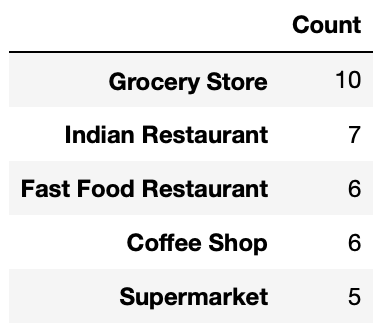
**4. Analysis**

Let's perform some basic explanatory data analysis and derive some additional info from our raw data. First let's count the **number of restaurants in every area candidate**:

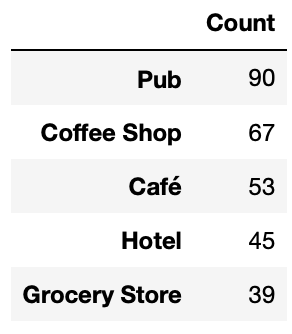
**Analyze one Neighborhood**

A single Neighborhood within the London area where a lot of Asians lived are examined by using the Foursquare API.

When we check the top 5 venues at the Beckton area, there are Indian restaurants and fast food restaurant there.



**Analyze multiple Neighborhoods in London**

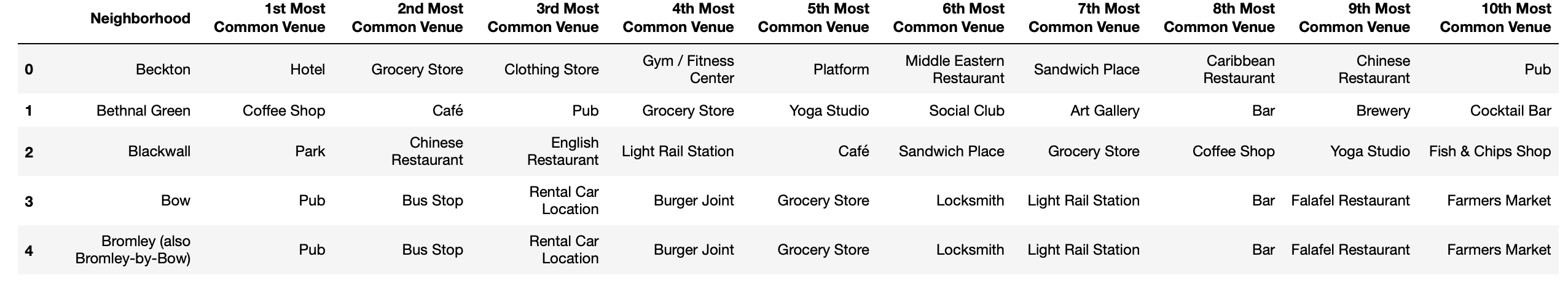


**Analyze Each Neighborhood**

And, let’s analyze each neighborhoods

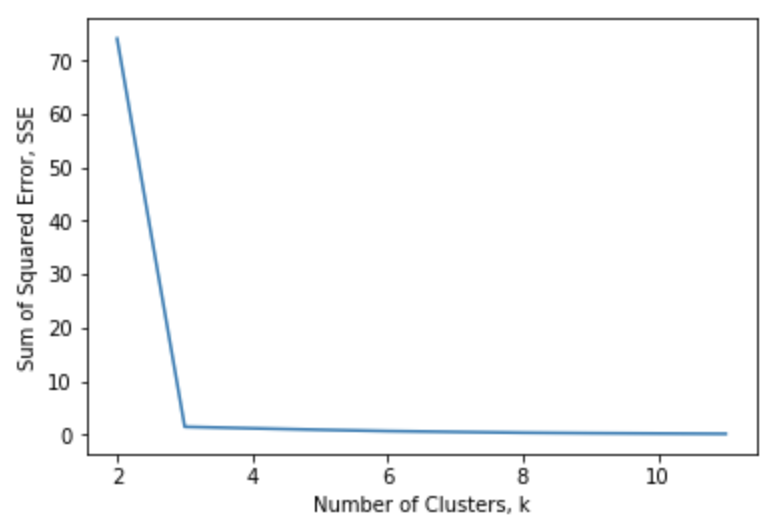
One hot encoding function is used to check the venues in each neighborhood.

There are 10 most common venues in each neighborhood.



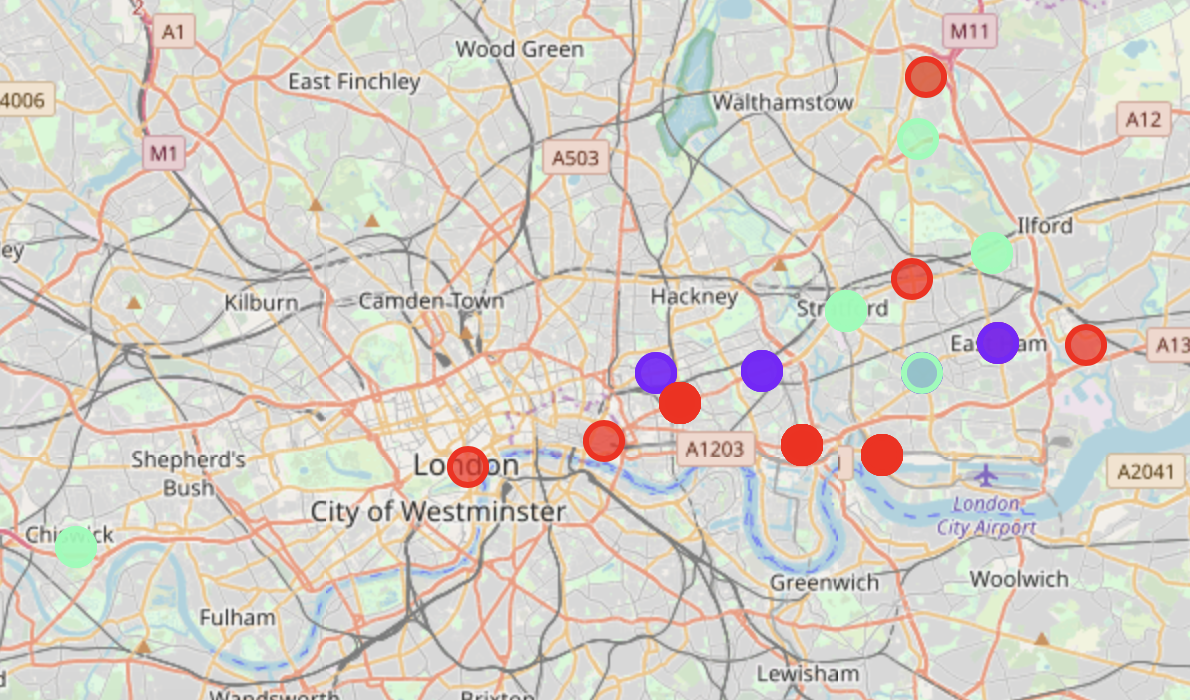
**Cluster Neighborhoods**

Before creating K-means clustering, we are needed to check the elbow method and get the best K.



Using the elbow method, we can select the best number of k. So, k is 3.

Visualize the clusters by using the folium function (Cluster 1: Red, Cluster 2: Violet, Cluster 3: Green)



**5. Results and Discussion**

After the cluster analysis, Cluster 1 is the Cultural & Tourist area and Hub (Hotel and Park are popular), cluster 2 falls within more of a main residential district with Pubs and cluster 3 is much more residential area and there are pubs and café there.

In Newham area where many Asians lived, Indian restaurants are popular.

In conclusion, pubs and hotels are popular in the Asian district in London. And, Most popular venue type in Downtown London is Pubs.

London is the capital and the largest city with a high population density in a small area. There are 32 boroughs in London. As London is considered as one of the world's most global cities, there are many creative approaches in clustering and classification studies.

The K-means algorithm are used for clustering study. Through using the Elbow method, the best number of k is 3. For future study, more data set can be added and more details of the neighborhood can also be included.

When we recommend the zones where stakeholders can consider for opening their restaurants, we need to check if there is no nearby competition. and the conditions in that area are related with the restaurants in which they try to open.

**6. Conclusion**

|  |  |
| --- | --- |
| **Cluster** | **Description** |
| **Cluster 1** | **The most common venue type is Hotel and Park.** |
| **Cluster 2** | **The most common venue type is Pub.** |
| **Cluster 3** | **The most common venue type is Pub and Café.** |

The purpose of the project is to find out areas in London center with low number of Asian restaurants in order to help stakeholders narrowing down the candidate of best location for new Asian restaurants.

So, after the analysis of the venues in London, we can recommend our stakeholders that if the Asian restaurants are open in the cluster 1 area, it will be successful. This is because there are a lots of asian restaurant (Especially Chinese restaurant) and china restaurants (3rd Most Common Venue) are already popular in cluster 2 area. Also, the cluster 3 is a tourist area, it is not good for asian restaurants. In the areas which belongs to the cluster 1, there is no competitor for Asian restaurants and a lot of asian lived there so these areas are good place ofr new Asian restaurants.

And, if there were more data about the housing price, traffic access, ratings or so on, we can get better understanding of the results and give the good insights for stakeholders.

**7. Recommendation**

When stakeholders start a business in a big city, it is needed to understand what the district needed. If they understand this clearly, people can achieve better outcomes through their analysis. There are different approaches to analyze the big cities.

The techniques which I used in London, can be used to analyze the different big cities' cases. And, these skills are useful to analyze the other cities.

And, we need to remember that "There is no free lunch", which means that there are a lot of methods to analyze the dataset and according to these techniques, the results are very different so we need to try diverse different methods to figure out what we need and we can find the best outcomes.