**Clustering the neighbourhoods of London**

Harshil Desai

15th May,2021

# **1.Introduction**

* 1. **Background**

London is the capital and largest city of England and the United Kingdom. London has 32 borough. London is one of the world's most important global cities. London has a diverse range of people and cultures, and more than 300 languages are spoken in the region. London covers an area of 1,579 square kilometres (610 sq mi). The population density is 5,177 inhabitants per square kilometre (13,410/sq mi), more than ten times that of any other British region. In terms of population, London is the 19th largest city and the 18th largest metropolitan region.

* 1. **Problem**

As you can see from the above figures, London is a city with a high population and population density. Being such a crowded city leads the owners of shops and social sharing places in the city where the population is dense. When we think of it by the investor, we expect from them to prefer the districts where there is a lower real estate cost and the type of business, they want to install is less intense. If we think of the city residents, they may want to choose the regions where real estate values are lower. At the same time, they may want to choose the district according to the social places’ density. However, it is difficult to obtain information that will guide investors in this direction.

* 1. **Interest**

When we consider all these problems, we can create a map and information chart where the real estate index is placed on London and each neighbourhoods is clustered according to the venue diversity. This will simplify our process for opening a venue in given a particular location.

# **2. Data acquisition and cleaning**

**2.1 Data sources:**

The data that I have used are as follows:

1. [The neighbourhood data from Wikipedia page.](https://en.wikipedia.org/wiki/List_of_areas_of_London)
2. [The house price data is taken from united kingdom government website.](https://data.london.gov.uk/dataset/average-house-prices)
3. [I used Forsquare API to get the most common venues of given neighbourhood of london](https://developer.foursquare.com/).

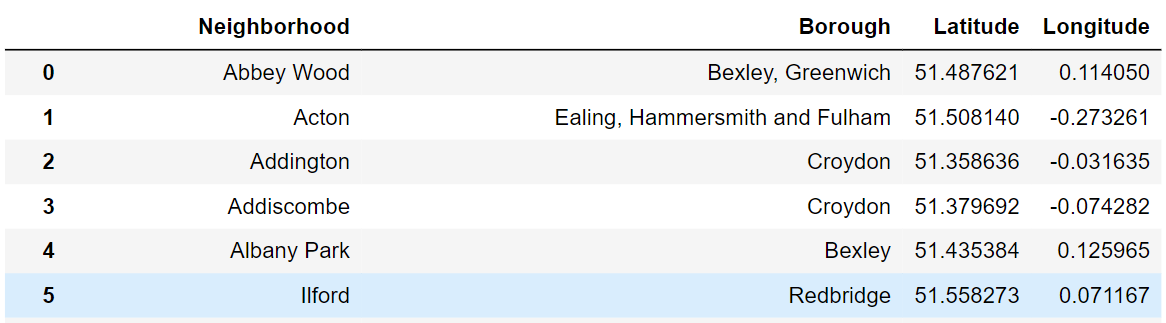
**2.2 Data Cleaning:**

From the neighbourhood data I have to extract the neighbourhood name and the borough name that I have done by using web scrapping. after that from geopy I import geocoders to locate each neighbourhoods centers geographic co-ordinates. which I append in a list after that I append the list in the dataframe.

From the United Kingdom Government website I download houses price by Borough excel sheet. From that excel sheet I have taken the avg data for house price and I have selected the Year ending Dec 2017. I have append that data into a list and append the list into the dataframe.

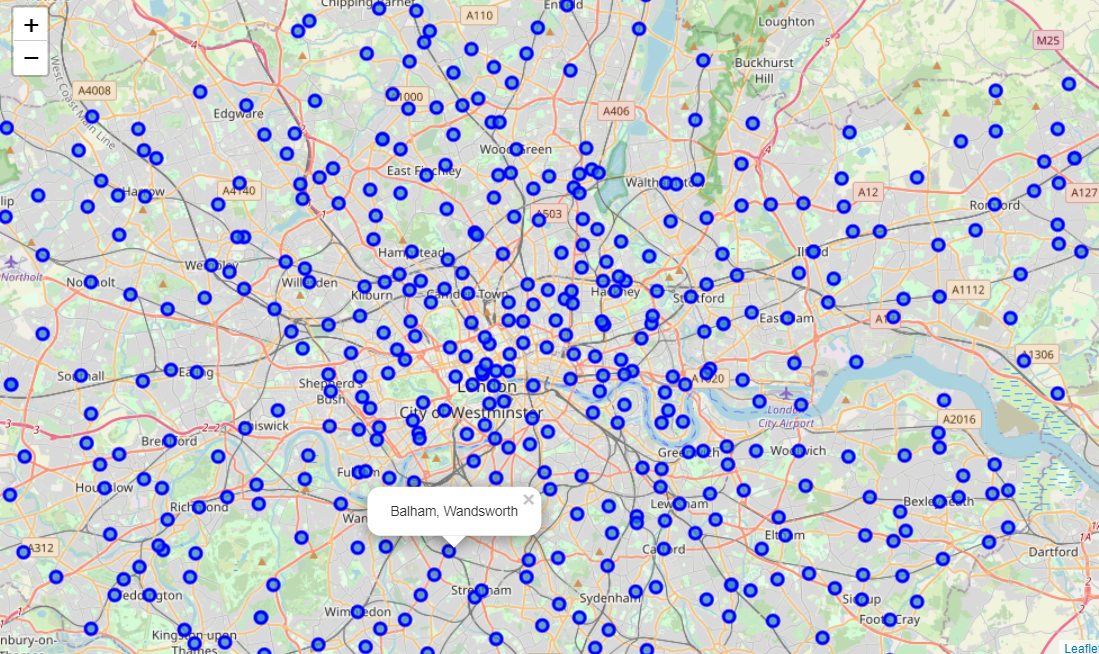
# **3.Data Methodology Analysis**

Here we have dataframe that contain neighbourhood name , borough name, latitude and longitude of neighbourhood.



**Figure 1 dataframe**

First, we have to visualize the data. So, I have created geographical map of London and I have added label which contain the name of the neighbourhood and the name of borough. I used the folium library in python for visualizing data.

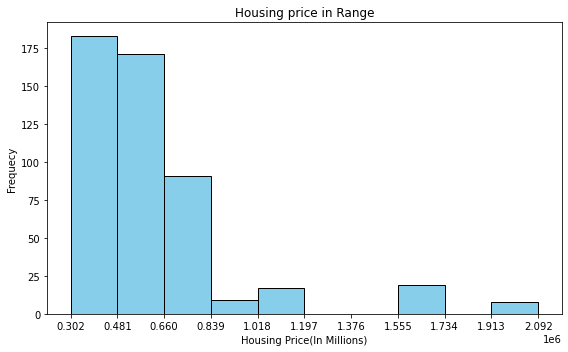


**Figure 2 map of London with neighbourhoods superimposed on top.**

As you can see here there are around 498 neighbourhood and 32 boroughs are in London. we want to cluster the neighbourhood by venues and by the price of neighbourhood.

First we have to categorized the neighbourhoods by their price. For that I have plot a histogram which divide the housing price by category.

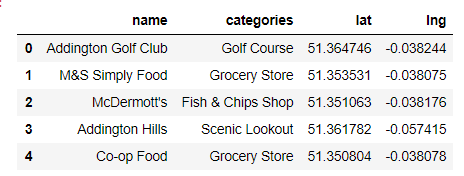
|  |  |
| --- | --- |
| **Price(In GBP)** | **Category** |
| Less than 481000 | Low\_Level\_HSP |
| Between 481000 - 839000 | Mid-1\_Level\_HSP |
| Between 839000 - 1197000 | Mid-2\_Level\_HSP |
| Between 1197000 - 1734000 | High-1\_Level\_HSP |
| Greater than 1734000 | High2\_Level\_HSP |



**Figure 3 Range for housing price**

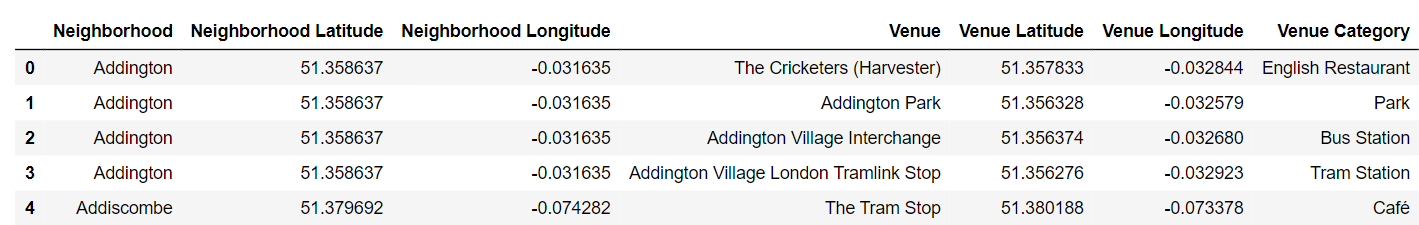
We append the list in our dataframe by creating a list. now we have five columns in our dataframe.

I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as 100 venues and the radius of 2000m for each neighbourhood from their given latitude and longitude information’s. Here is a head of the list Venues name, category, latitude and longitude information’s from Forsquare API.



**Figure 4 Head of venues returned by Forsquare**

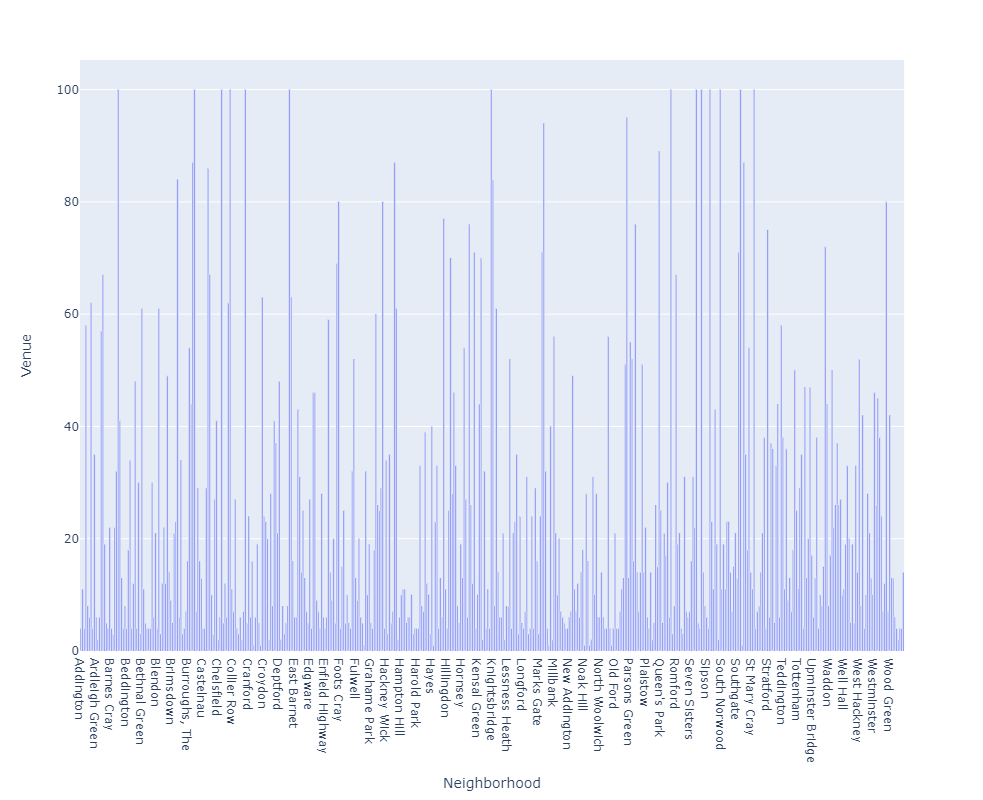
In summary of this data 21 venues were returned by Foursquare. Here is a merged table of boroughs and venues.

****

**Figure 5 Merged table of venues and neighbourhood**

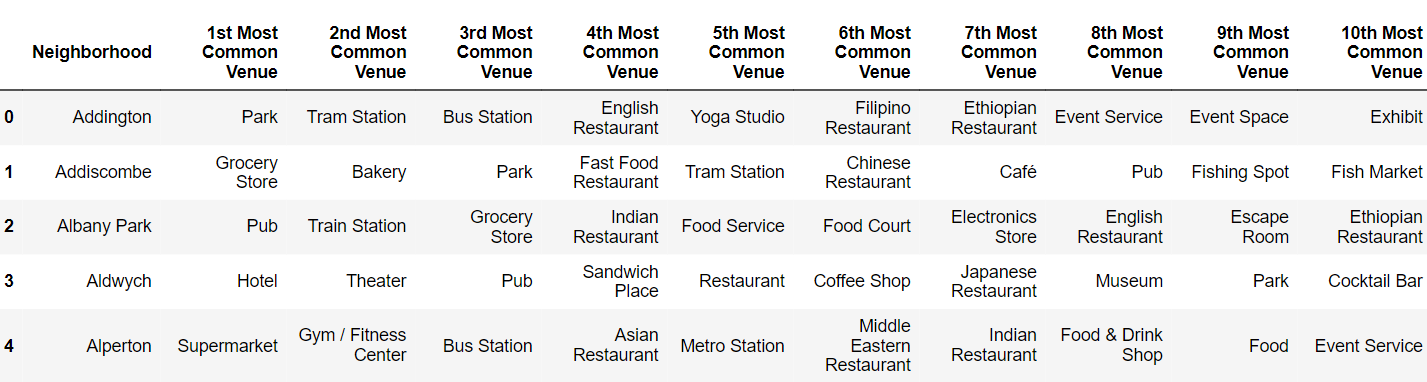
we can see here for some neighbourhoods Forsquare has returned 100 venues like Kingston, Ealing, Chinatown, Bayswater and for some neighbourhood venues it is less than 20 like Bickley, Arkley, Hampton.

The result doesn’t mean that inquiry run all the possible results in neighbourhoods. Actually, it depends on given Latitude and Longitude information’s and here is we just run single Latitude and Longitude pair for each neighbourhood. We can increase the possibilities with areas information’s with more Latitude and Longitude information’s.



**Figure 6 Venues in each neighbourhood**

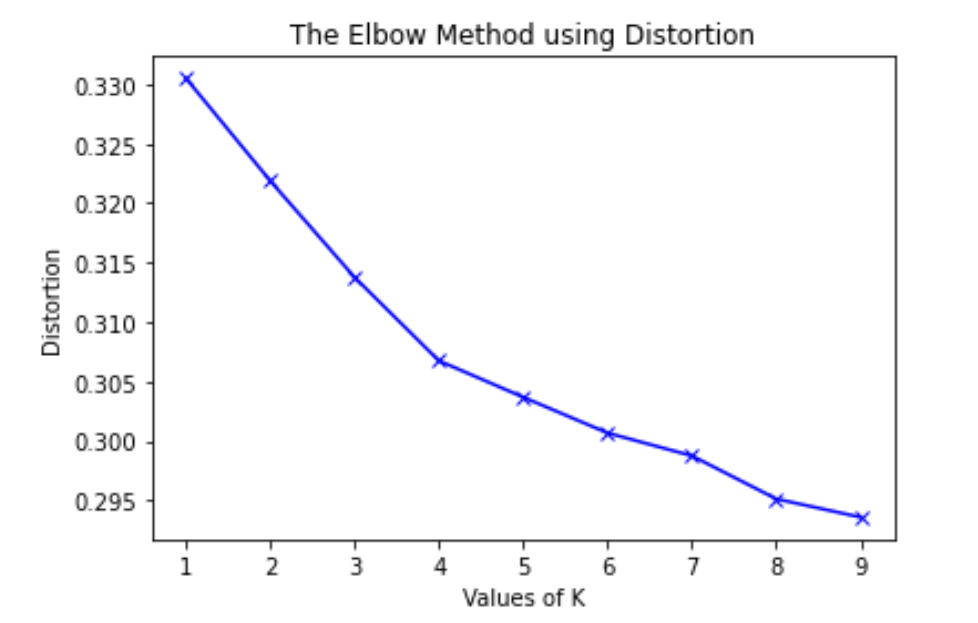
In summary of this graph 410 unique categories were returned by Foursquare. I have created a dataframe which shows top 10 venue category for each neighbourhood.



**Figure 7 Top 10 venue category for each neighbourhood**

Now for clustering the neighbourhood we are going to use unsupervised learning **K-means algorithm** to cluster the neighbourhoods. K-Means algorithm is one of the most common cluster method of unsupervised learning.

First I will use random number for clusters then by using the elbow method I am going to determine the number of cluster for optimum k of k-means clustering.



**Figure 8 Elbow method for optimum k**

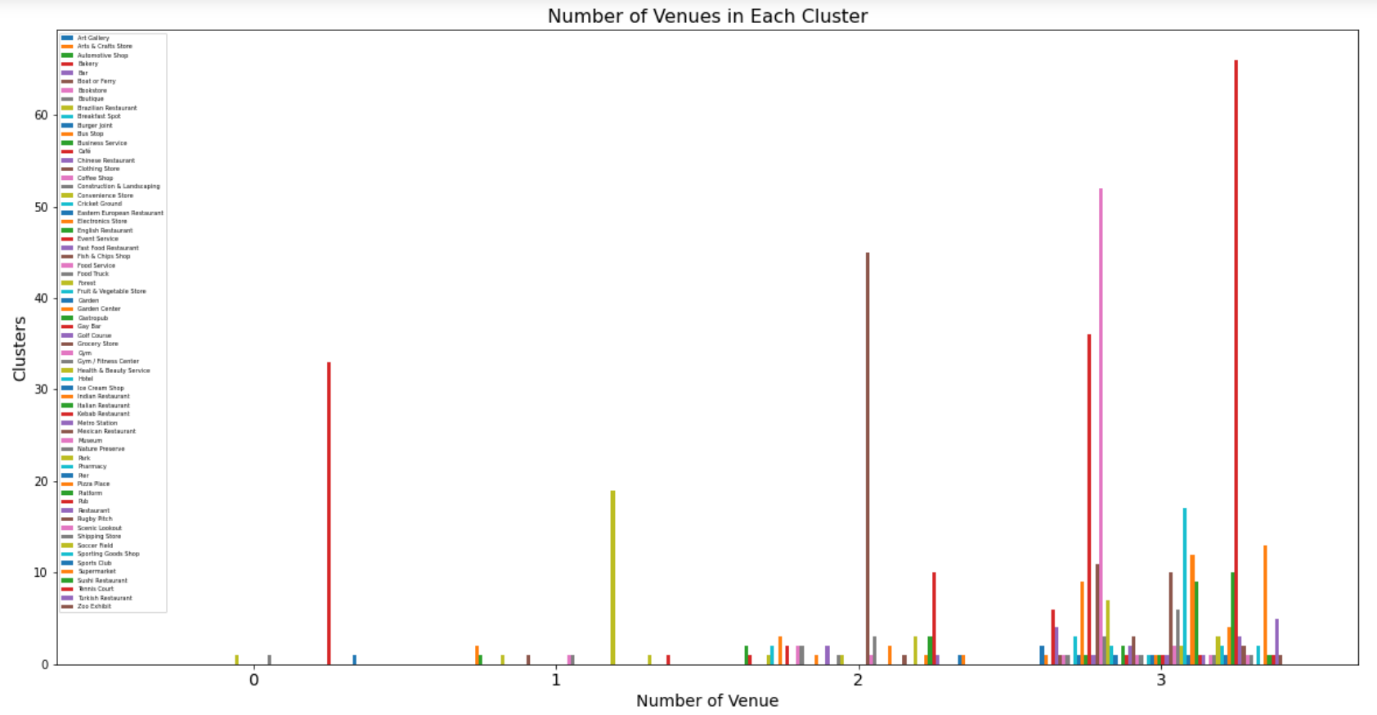
As you can see here the efficient number of clusters are 4. So I am using k-means clustering method to cluster neighbourhood.

Here is my merged table with cluster labels for each neighbourhood.



**Figure 9 merged table with cluster number**

We can also estimate the number of **1st Most Common Venue** in each cluster. Thus, we can create a bar chart which may help us to find proper labels for each cluster.

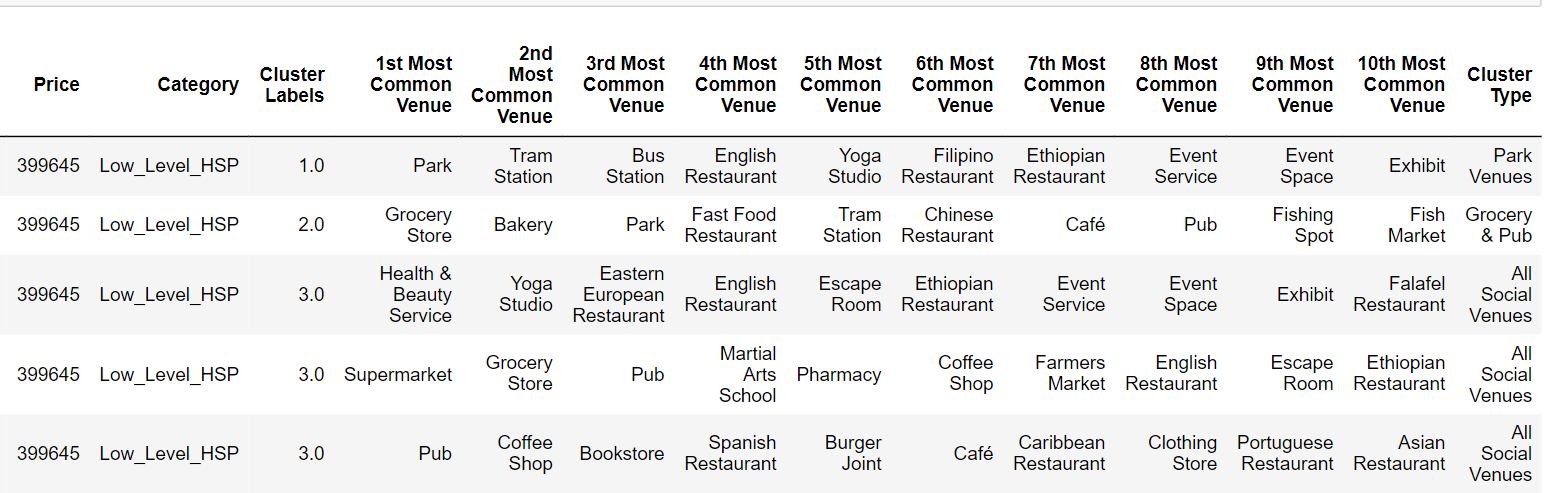


**Figure 10 1st Most Common Venue in each cluster**

By this chart we can categorized the cluster in following cluster type.

|  |  |
| --- | --- |
| **Cluster number** | **Cluster Type** |
| cluster 0 | Pub venues |
| cluster 1 | park venues |
| cluster 2 | grocery & pub |
| cluster 3 | all social venues |

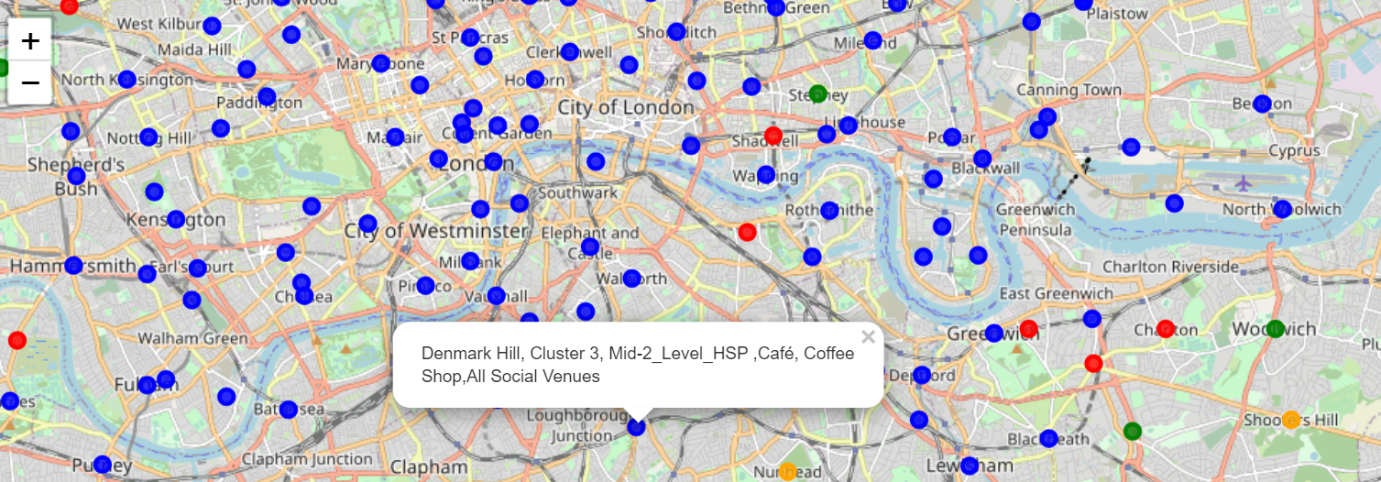
Now we can add this in our dataframe

****

**Figure 11 Added cluster label in table**

Now we can use all this information to visualize data very easily.

I have created map of London using folium and labelled all the neighbourhood by their name, price category, , 1st most common venue, 2nd common venue, cluster type.



**Figure 12 Clustered map with label**

# 4.Discussion

As I mentioned before, London is a big city with a high population density . As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for this busy city.

I used the K-means algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 4. For more detailed and accurate guidance, the data set can be expanded and the details of the neighbourhoods can also be scrutinized.

I ended the study by visualizing the data and clustering information on the London map. In future studies, web or telephone applications can be carried out to direct investors.

# 5.Conclusion

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.