**Capstone Project – The Battle of Neighborhoods**

Author: Anastasiya Lyamina

Submission Date: 07/05/2020

1. Introduction
   1. Background

London is the most populated city in the UK. It is very attracted to tourists as well as immigrants looking to move for a living. Considering all the benefits London can bring, it is also highly competitive, no surprise that cost of living is very high. It is always a challenge to find the most suitable place to rent a flat and at the same time to be quite close to all needed venues.

* 1. Business Problem

This Project would help the stakeholders to take a thoughtful decision on choosing the best borough to rent a flat in London based on the distribution of various entertainment facilities in and around that borough. This project analyses all London boroughs: explores it fully, finds the 5 most common venues of Arts & Entertainment category in each borough, clusters London’s boroughs using k-mean clustering and also gives a visualisation on other interesting factors for each borough, including the house price and life satisfaction.

* 1. Target Audience

As the main goal, I would like to provide targeted information near property to let out for real estate agents. Being target audience, real estate agents would benefit the most from this project, as they compete for the same customer, while customers want to find the most appropriate flat for themselves. Deep knowledge of the area and borough brings a competitive advantage.

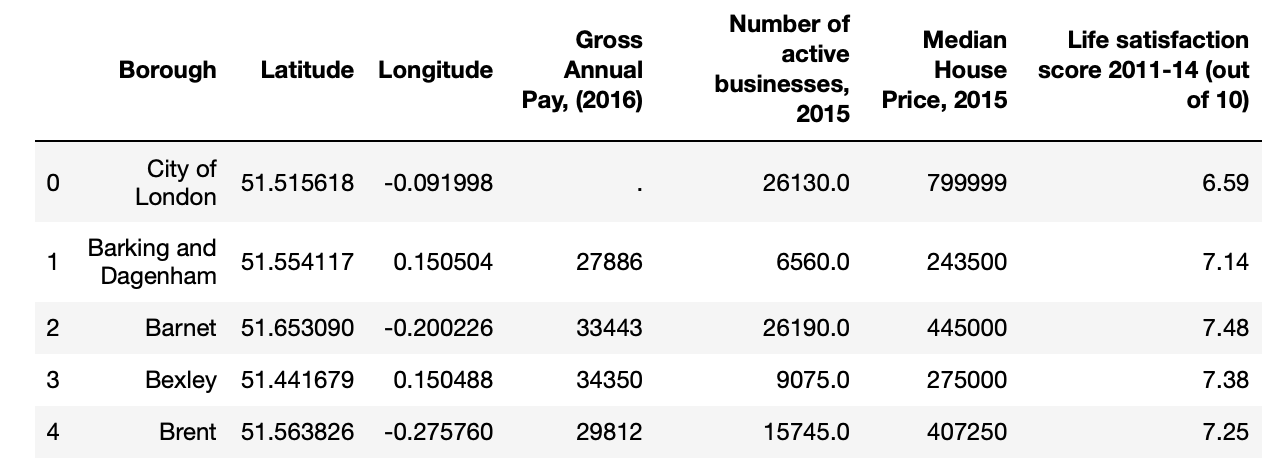
1. Data Acquisition and cleaning
   1. Data Acquisition

In order to build the model, the following datasets and information are considered for analysis:

* London boroughs’ average house prices and life satisfaction score will be taken from: <https://data.london.gov.uk/download/london-borough-profiles/80647ce7-14f3-4e31-b1cd-d5f7ea3553be/london-borough-profiles.xls>
* London boroughs’ geo codes will be obtained using the package GeoPy and Nominatim geocode service. This will allow us to be able to plot the London boroughs on the map
* FourSquare API will be used to collect data about locations of different venues in the boroughs. I will connect to Foursquare API using the Client ID and Client Secret.
* Furthermore, various python libraries will be used as imported below to create maps, machine learning models and graphs:
  + Pandas - Library for Data Analysis
  + NumPy – Library to handle data in a vectorized manner
  + Requests – Library to handle http requests
  + Matplotlib and Seaborn – Python Plotting Module
  + Sklearn – Python machine learning Library
  + Folium – Map rendering Library
  1. Data Cleaning

The data preparation is firstly started by cleaning the downloaded file from the data.london.gov.uk website. This data source has a lot of information which is not needed for this report purposes. By having a glance on the raw dataset, a few steps were performed to get only the dataset required: renaming the columns, dropping unnecessary columns with meaningless information, dropping rows which are not in the category “Inner/Outer London” and extracting 5 columns in a separate data frame: 'Borough', 'Gross Annual Pay, (2016)', 'Number of active businesses, 2015', 'Median House Price, 2015', 'Life satisfaction score 2011-14 (out of 10)'. Out of these columns only the information about house price and life satisfaction will be used in the scope of this project.

Furthermore, the longitude and latitude data was missing from the excel file, so it’s been added in a different data frame (GeoPy package), which was then merged with the data frame we got earlier. That way we had the main data frame we could work on (showing only first 5 rows):



To get the venue information from the Foursquare API, a few filters were put in place to get it correctly. Firstly, the category id for Arts & Entertainment is specified as 4d4b7104d754a06370d81259. Secondly, the borough name was not enough to search the venues in Foursquare, hence I had to add “ , London, United Kingdom” to the query to make the address specific. To put the boundaries on the search polygon, only 1000 meter radius has been specified.

1. Methodology
   1. Exploratory Data Analysis

The first data analysis is performed with the describe() method on the raw dataset from the data.gov.uk website. It gives us quite an interesting and brief insight into all available columns: “Life Satisfaction Score” feature was selected at this point, as it might give a better understanding of the borough. The minimum value available across all 33 boroughs is 7.18, while maximum is 7.61.

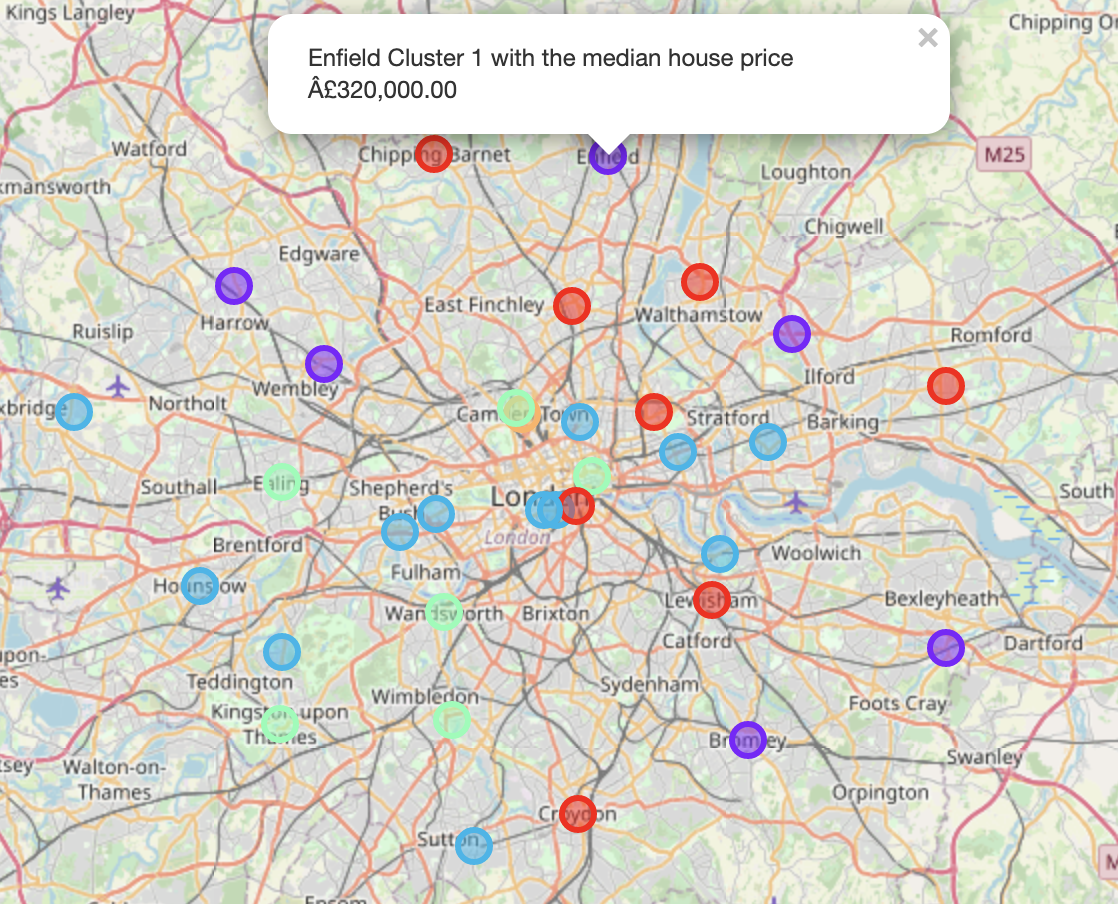
Using the final dataset containing the boroughs in London, its information about the house prices and life satisfaction score, along with the latitude and longitude, we can find all the Arts & Entertainment venues within a 1000 meter radius of each borough by connecting to the Foursquare API. This returns a JSON file containing all the venues in each borough, which is converted to a Pandas data frame. This data frame contains all the venues from the specified category along with their coordinates and sub-category. The size of the file is 760 rows, containing columns of Borough, Venue Name, Venue Latitude, Venue Longitude and Venue Category. There are only 36 unique categories.

One hot encoding is done on the venues data. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. The venues data is then grouped by the borough and the mean of the venues are calculated, finally the 5 common sub-categories are calculated for each of the boroughs. This information is then transformed in a table, so each row shows 5 most common venues for each borough. The first 5 rows in the result table:



* 1. Modelling

To help our target audience to find similar boroughs in London we will be building the model by clustering similar boroughs using K - means clustering which is a form of unsupervised machine learning algorithm, that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster the 33 boroughs into 5 clusters. The reason to conduct a K- means clustering is to cluster boroughs with similar Arts & Entertainment venues together so that our target audience can shortlist the area of their interests based on the venues/amenities around each borough. When visualised, we also add the information of the Median House Price to each point:

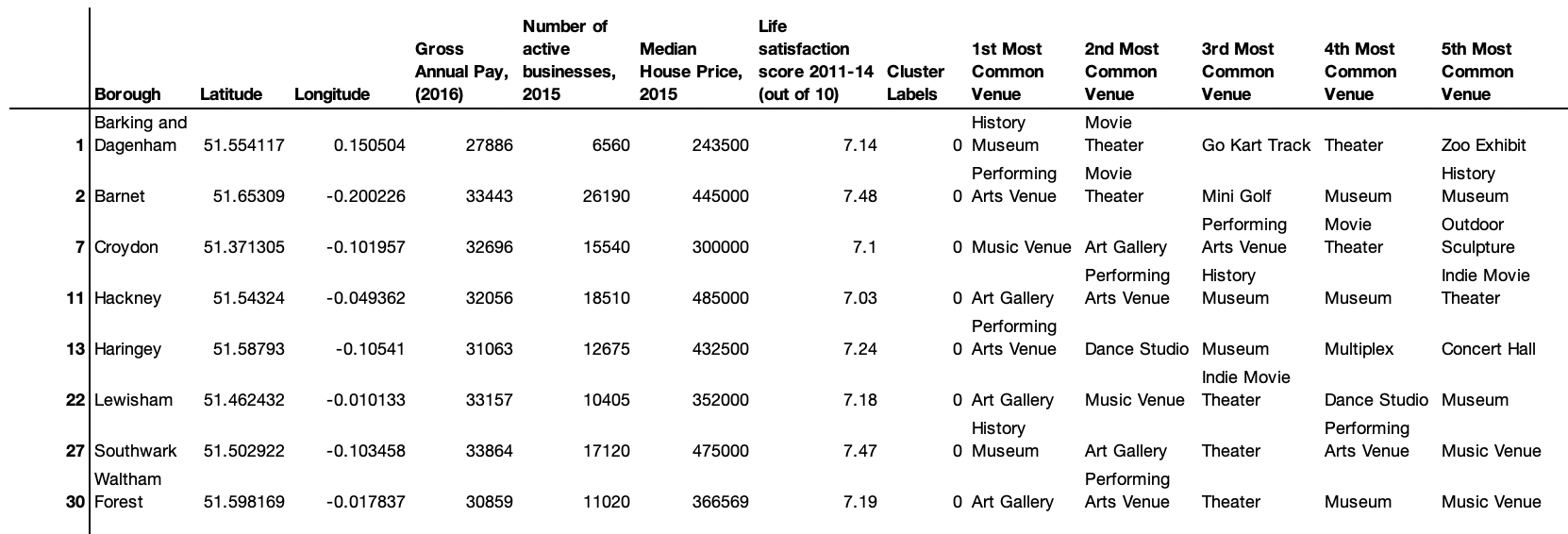


1. Results

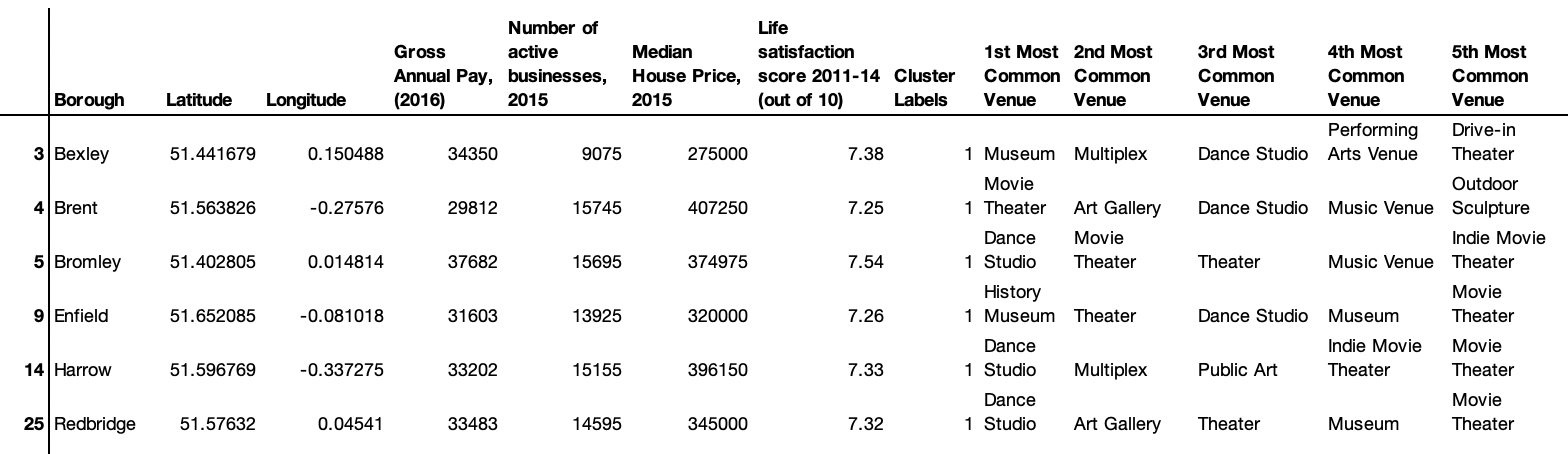
After running the K-means clustering we can access each cluster created to see which boroughs were assigned to each of the five clusters, which are labelled from 0 to 4.

Here is the detailed overview for each cluster:

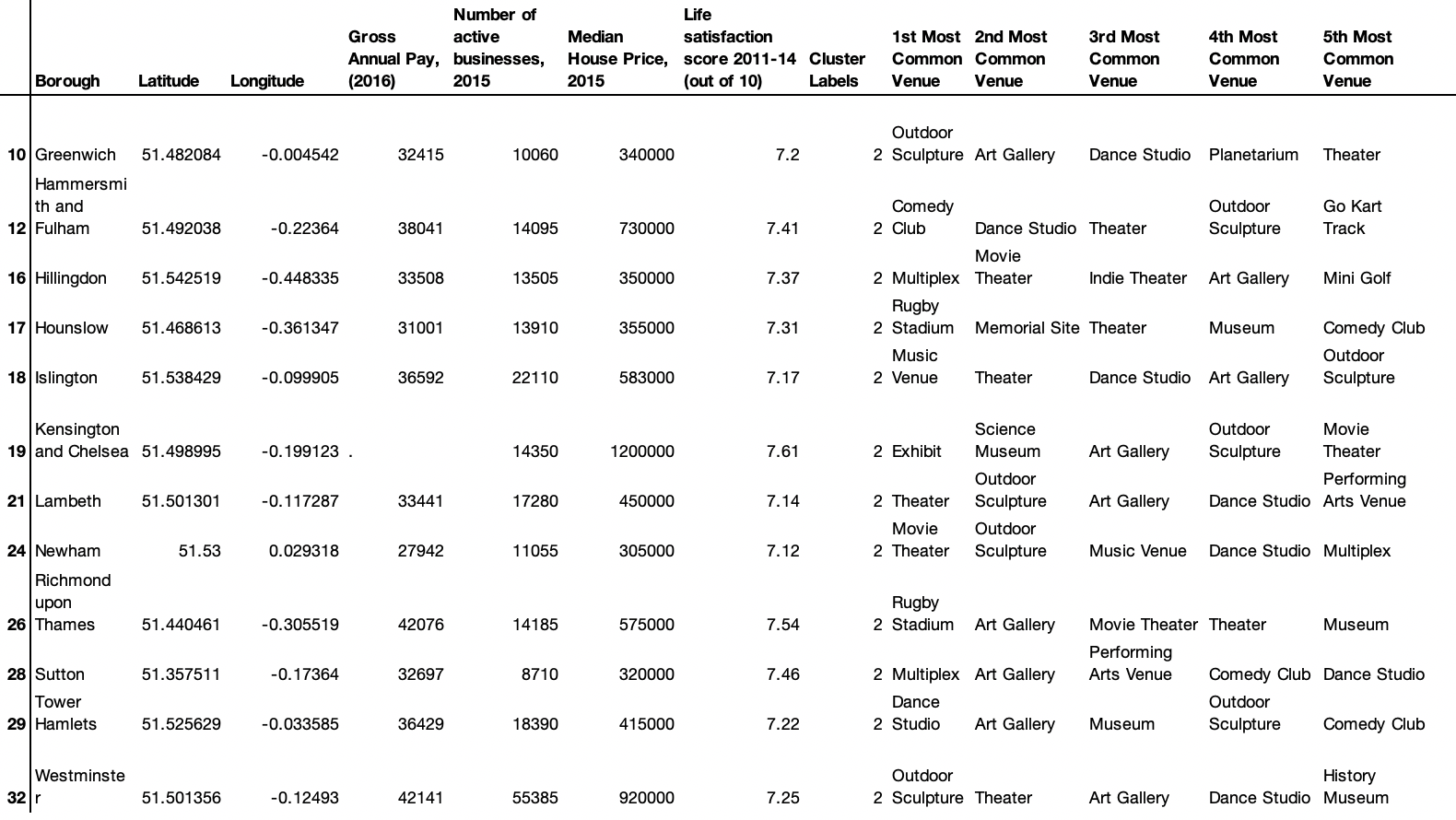
* Cluster 0: Upon close examination of the cluster, we can see that the most common venue among boroughs is Art Gallery. Another observation can be made by looking at the Gross Annual Pay column, which was not used for clustering: coincidentally all selected boroughs are in the range of 30-33k apart from Barking and Dagenham, which is 27k. By looking at the Median House Price we can also establish the range is between 300k-485k apart from the same one borough again.



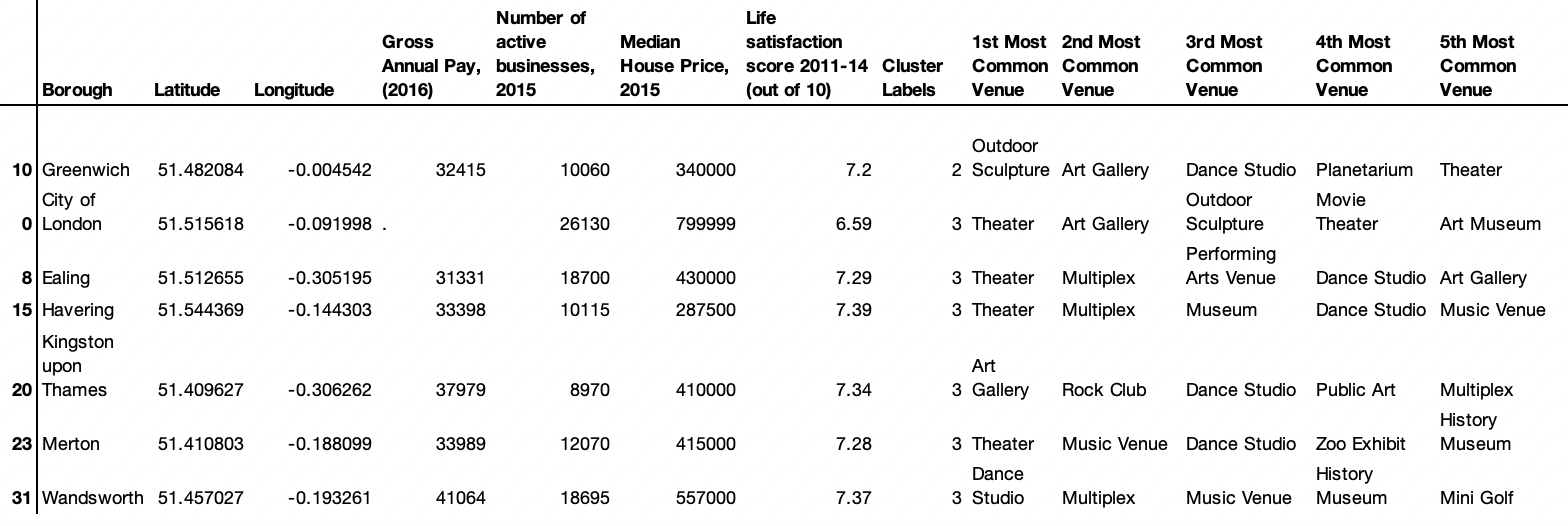
* Cluster 1: The 6 boroughs selected for this cluster have a few things in common: the most common venue seems to be Dance Studio for all boroughs; Number Of Active Business (feature which was not specifically selected for clustering) has a value of >15k for the majority of boroughs.



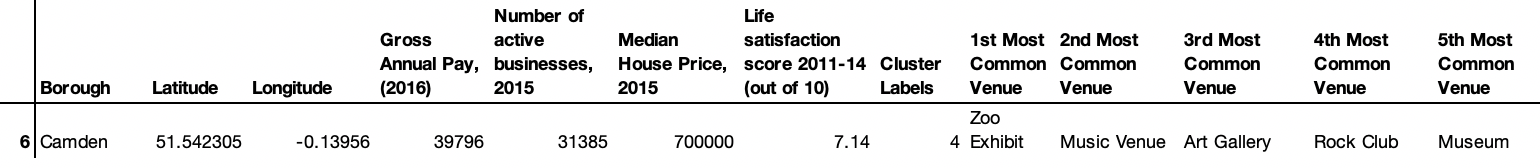
* Cluster 2: the biggest cluster generated – it has 12 boroughs out of 33. By the first glance, there is no one single venue, which was the reason for clustering these particular boroughs. In fact there is a combination of the venues from the table, which are Dance Studio, Movie Theater and Multiplex.



* Cluster 3: With quite a broad range of Median House Price and Number of Active Businesses, all 7 boroughs from this cluster has Theater venue as one of the most common.

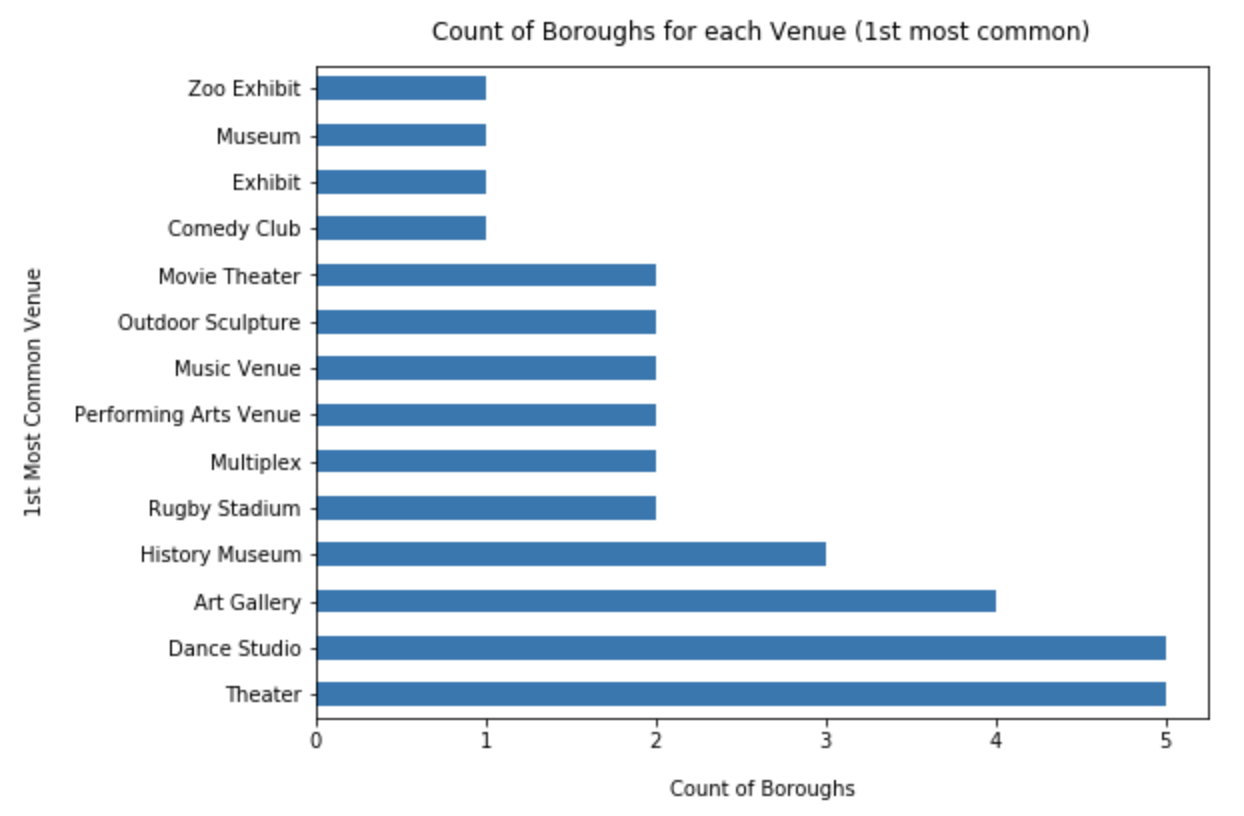


* Cluster 4: this cluster has only one borough inside, which is Camden. It has Zoo Exhibit as the 1st most common venue, which is not listed in any other boroughs as a top 5 venue. It also has a Rock Club, as the 4th most common venue, which is a rare venue across all other boroughs.

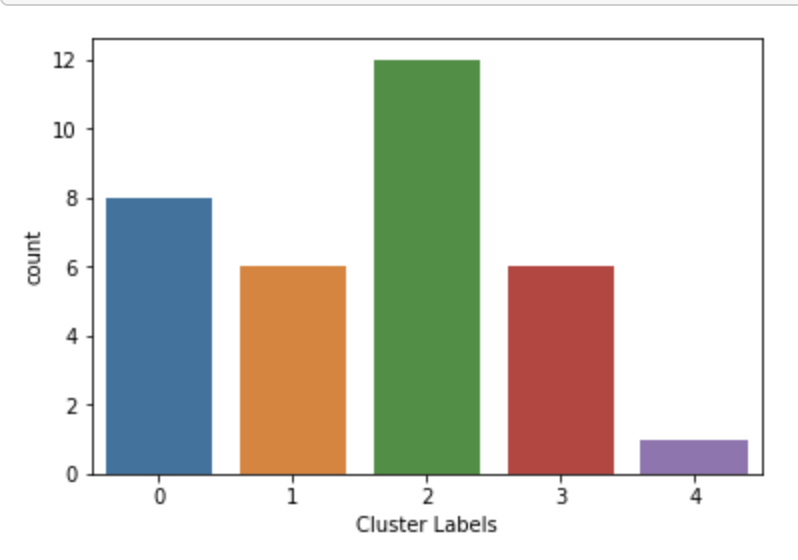


A few of additional analysis were made on a dataset with cluster:

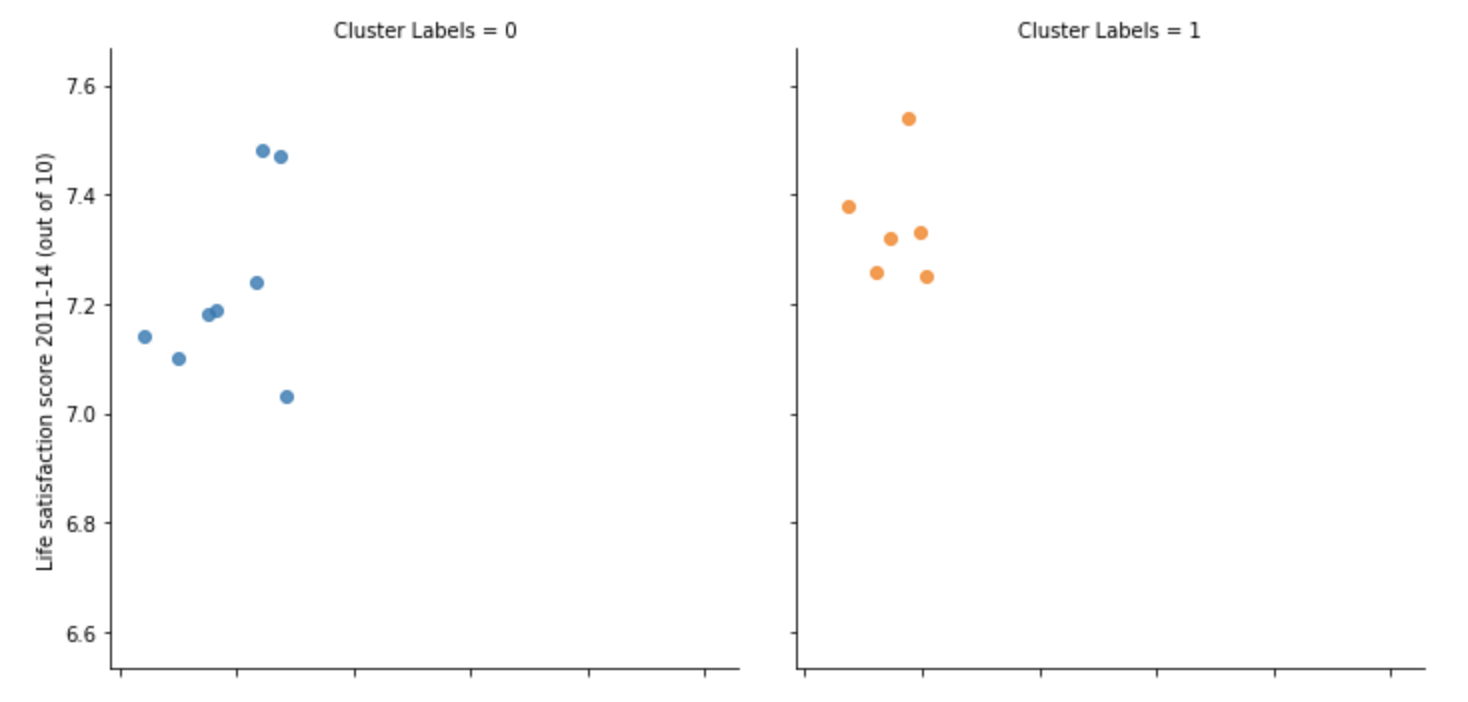
1. The most frequent venues from the 1st Most Common Venue column across all boroughs are Theatre and Dance Studio. Zoo Exhibit, Museum, Exhibit and Comedy Club are the ones mentioned the least times.

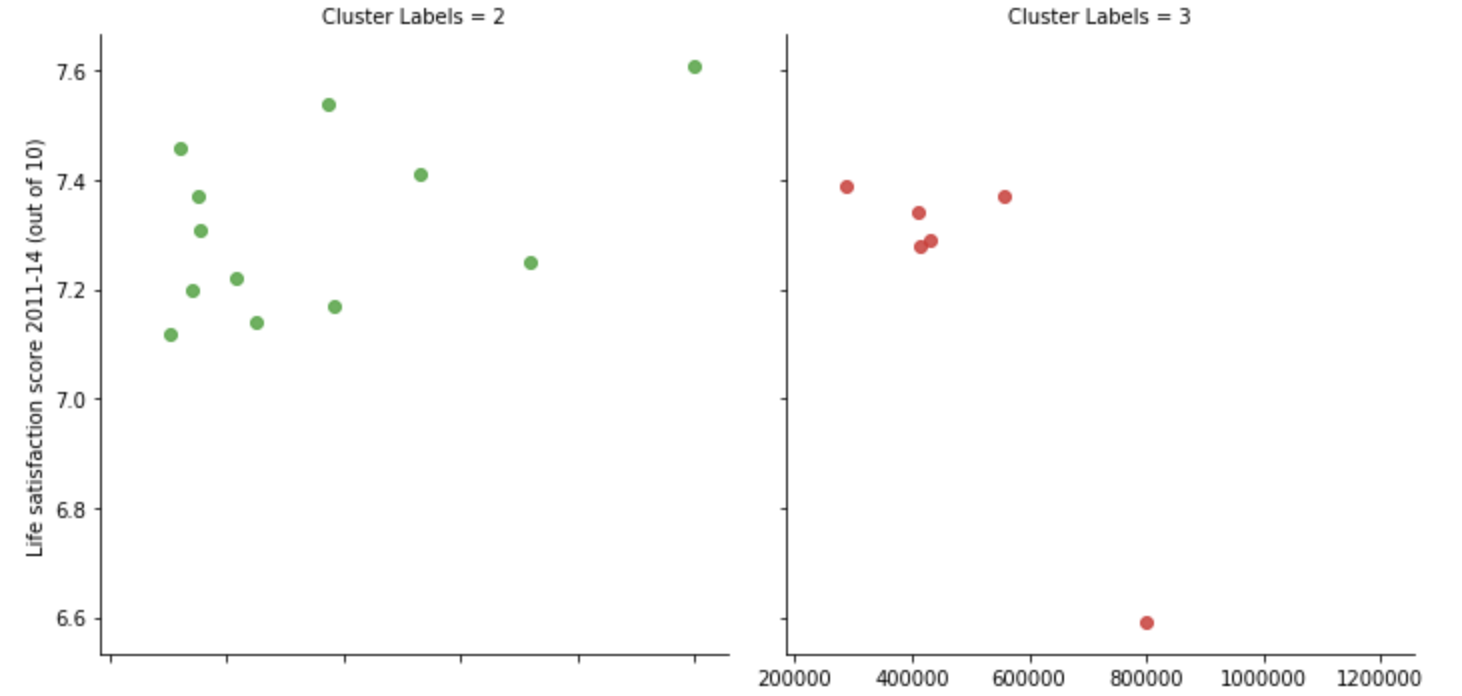


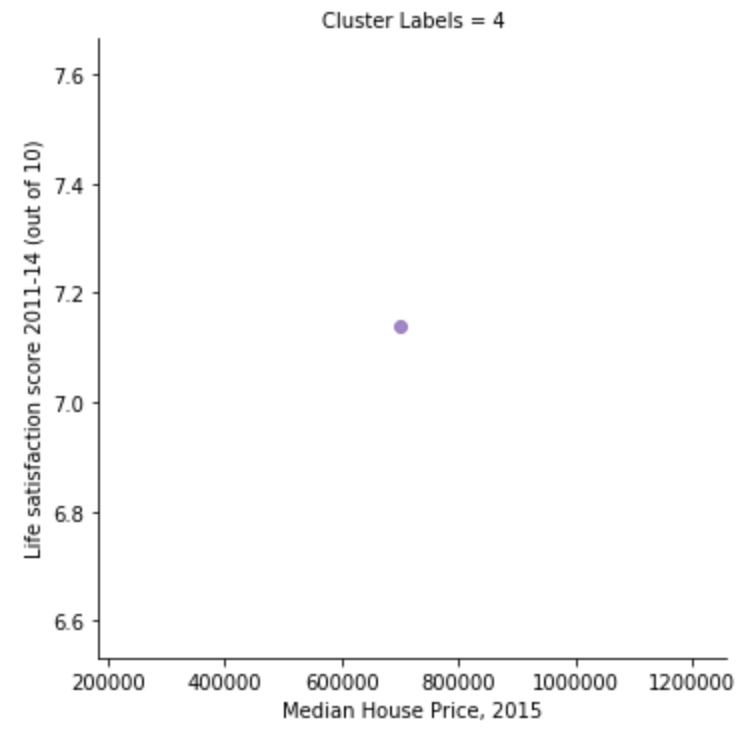
1. Counts of the boroughs inside each cluster: Cluster 2 has the majority of boroughs – 12, while Cluster 4 consists of only 1 borough. Cluster 0, 3 and 4 has 6-8 boroughs each.



1. By plotting the clusters data from Median House Price and Life Satisfaction Score, it’s shocking to see the correlation between these 2 features in some clusters. Ignoring Cluster 4 with just 1 borough and Cluster 2 with no obvious correlation, all other clusters (Cluster 0, 1, 3) have a visible tendency, especially Cluster 3 with range of Life Satisfaction score 7.2-7.4 apart from 1 outlier.







1. Discussion

A vast majority of the results available were already discussed above in the Results section.

In general, dividing all 33 London boroughs with various Median House Price into 5 clusters makes it easy to pick boroughs to look into in more detail for living. If a client is interested in one borough, the retail agency might also recommend other similar boroughs which sit in the same cluster. Furthermore, the Arts & Entertainment interests of the client might be in particular taken into account: for young professionals or students it might worth to look into boroughs from Cluster 1, where Dance Studio is the most common venue; for more conservative type of people and families it might be a good idea to consider boroughs from Cluster 3 with Theatre being the most popular venue. The choices of boroughs may vary from person to person.

Looking at the graphs with Life Satisfaction Score included, it makes it obvious that the majority of boroughs have a score in the range of 7.2 and above, which makes it tough to consider this feature for decision making. If the range was broader, it could distinguish boroughs even further.

1. Conclusion

The aim of this project is to provide the support to both retail agents and customers who want to relocate to the most suitable borough in London, taking into account the house prices and most common entertainment venues in that borough. This project gives the first glance on the process of how it can be done, although the analysis has been performed on the limited data. If the following data limitations could be removed, it would significantly increase the accuracy of the model and hence the results would be more thorough:

* Foursquare filters: the category is limited to Arts & Entertainment, however it can be changed to any other category interested to the clients: Food, Shops, or it could be completely removed by taking into consideration all possible venues in the borough. Additional filter was specified on the radius of the search being 1000 meters. The area could be increased, although it should be done carefully with the concern that it might overcome boroughs nearby. Another point is that we had used the value of LIMIT = 100 # limit of number of venues returned by Foursquare API. If we increase this number to a larger value, then we would have more data and venues to cluster and which would refine our data and analysis. This would result in crisp and much detailed findings. Data Scientists with Premium Foursquare account can overcome this bottleneck.
* Clustering was done primarily on the list of venues available from Foursquare and it didn’t include the Median House Price, Life Satisfaction Score or any other useful features available from the raw dataset from data.gov.uk website. In the future the clustering might include these or any other features, which would make the model more sophisticated, however it might be too complex for the initial purpose.
* London boroughs is a detailed scope enough, however it might be worth looking at street-level and cluster particular streets inside each borough.