Identifying affordable London homes

Tagkoulis Dimitrios

April 29, 2019

## 1. Introduction

### 1.1 Background

The UK is the second most popular study destination worldwide. According to official international enrollment statistics, 458,490 foreign students are attending university in UK. As a result, London is a really popular destination for students. Even though students can be accommodated by the universities most of the time they will need to find a home in order to keep living in the UK. When they initiate their search, they will quickly realize that the Housing Market is not in a good shape.

### 1.2 Business Problem

In this project we will try to assist home buyers located or moving in London to make optimal decisions using machine learning tools. As a result, the business problem we are currently posing is: how could we provide support to home buyers clientele in purchasing a suitable piece of real estate in London in this uncertain economic and financial scenario? To solve this business problem, we are going to cluster London neighborhoods in order to recommend venues and the current average price of real estate where homebuyers can make a real estate investment. We will recommend affordable venues according to amenities and essential facilities surrounding such venues. Since our target group is students either coming in or graduating from Universities so we assume that their budget is limited from 100.000 to 300.000 pounds.

## 2. Data section

Data on London properties was extracted from the HM Land Registry (<http://landregistry.data.gov.uk/)>.

The following fields comprise the address data included in Price Paid Data: Postcode; PAON Primary Addressable Object Name. Typically, the house number or name; SAON Secondary Addressable Object Name. In order to find and recommend locations based on the presence of essential facilities, we will access data through Foursquare API interface and arrange them as a data frame for visualization. By merging all the data gathered on London we will be able to recommend good buying opportunities for the homebuyers.

## 

## 3. Methodology section

The Methodology section will describe the main components of our analysis and prediction system. The Methodology section comprises four stages:

3.1 Collection of Needed Data

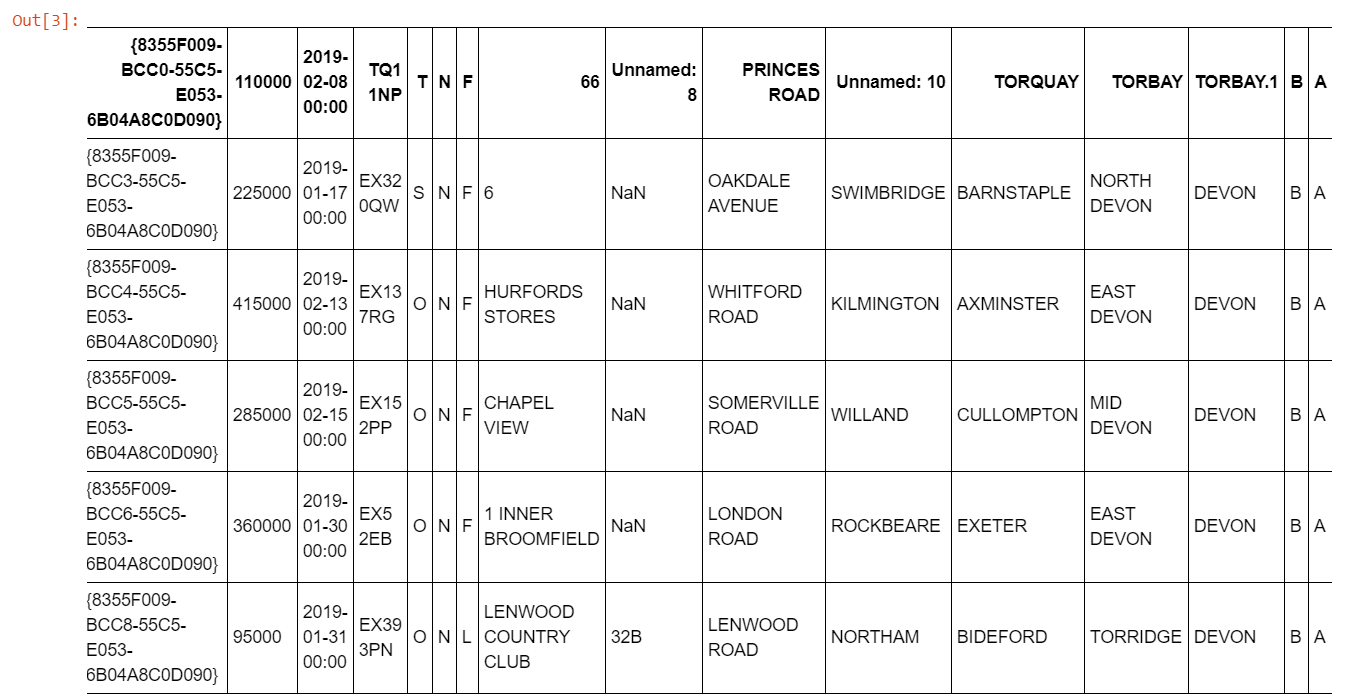
3.2 Explore and Understand Data

3.3 Data Wrangling

3.4 Modeling

### 3.1 Collection of Needed Data

After importing the necessary libraries, we download the data from the HM Land Registry:



First 5 rows of the Data extracted from the Registry

### 3.2 Explore and Understand Data

After we saw the format of our data using the .head method from Pandas library we also asked for the shape of our data frame using the .shape method and we saw that our first dataset consists of 78004 rows and 16 columns. The current format is not ideal to process so we must proceed to the data wrangling phase to transform it in a useful manner.

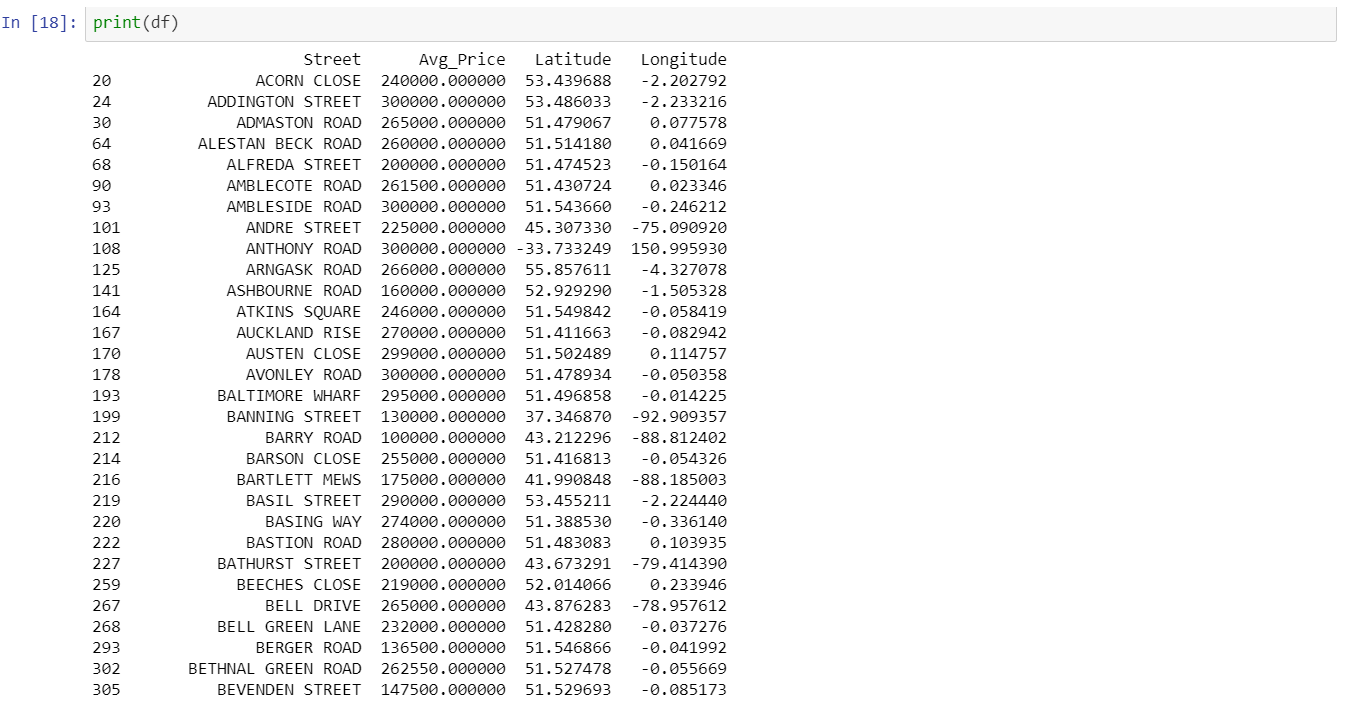
### 3.3 Data Wrangling

At this stage, we reshape our dataset for the modeling phase. Firstly, we rename the column names to something useful.



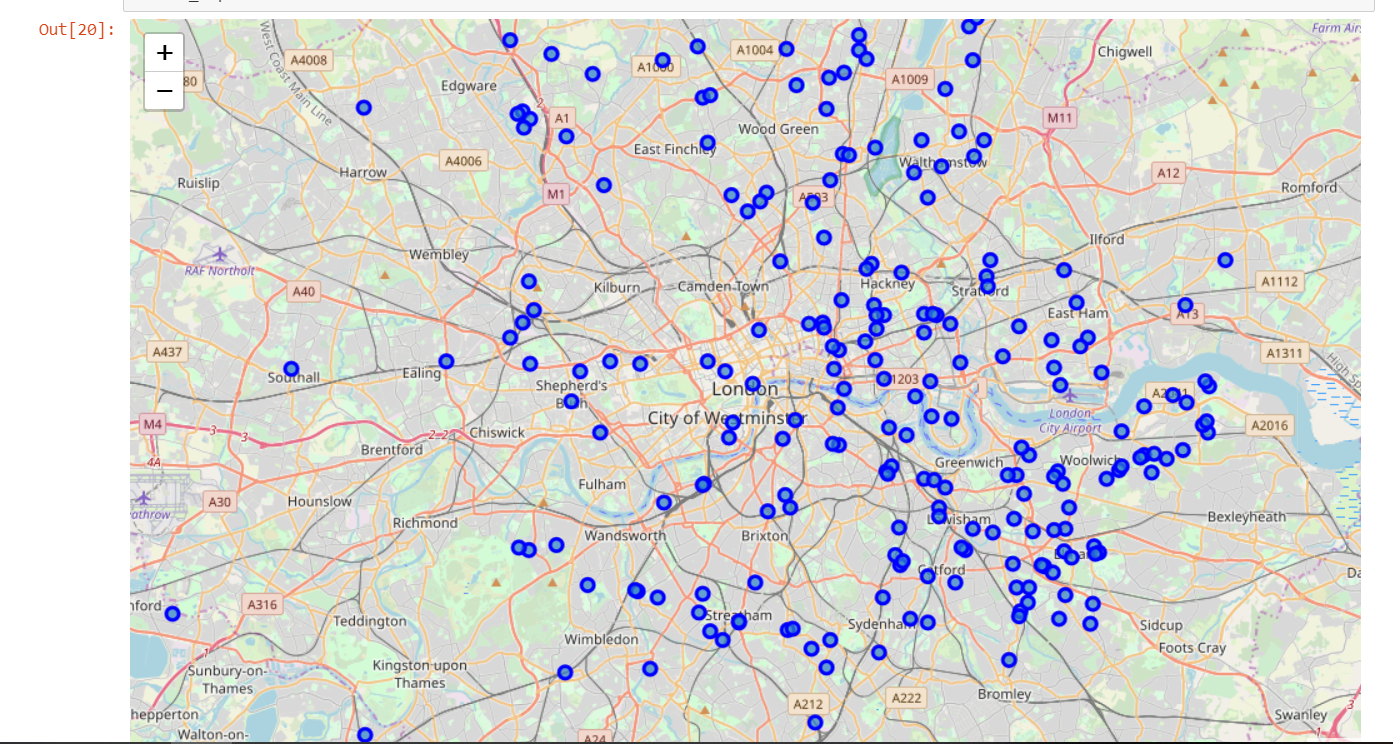
Registry Data After our Change in Names

Then to keep only the useful data we format the date column; we sort data by date of sale, and we delete data that is not from the city of London. To create the values, we will examine we create a list of street names in London, calculate the street-wise average price of the property and read the street-wise coordinates into a data frame, while also eliminating recurring word London from individual names. Then we define our budget and join the data to find the coordinates of locations which fit into client's budget. Finally, we plot recommended locations on London map along with current market prices.



Our Data Frame is ready for the modeling phase.

Our new data frame ready for the modelling phase (above) and the plot of the identified real estate (below)



Map Plot of our Data to help us visualize how our real estate is distributed.

We are now ready to proceed to the modeling phase.

### 

### 3.4 Modeling

First, we scrape data from the Foursquare API, getting nearby Venues of interest for every street. Using the .head method we can see the first five rows of our new data to get a feel of our new data frame.



Our nearby Venues data

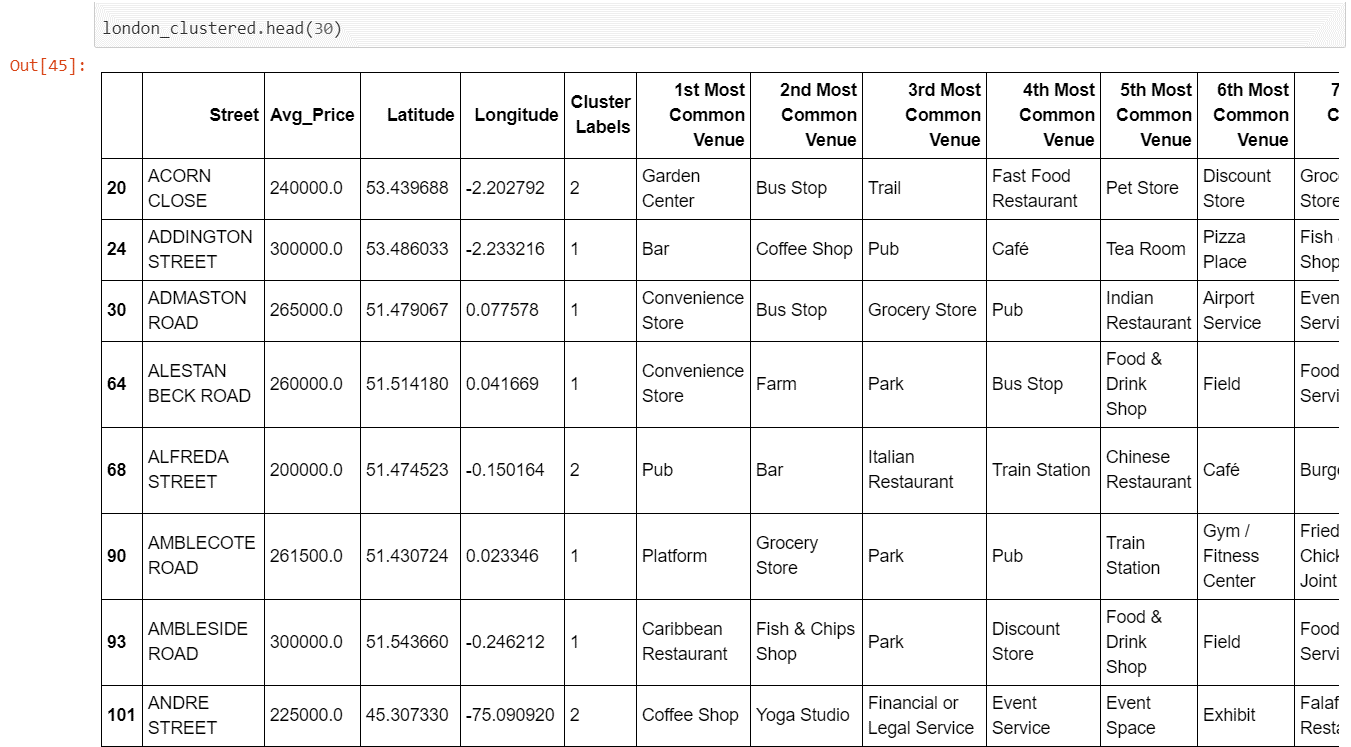
We then group those Venues by street and get the number of venues located near every street and reshape that data frame so that it shows the top 10 venues per street.



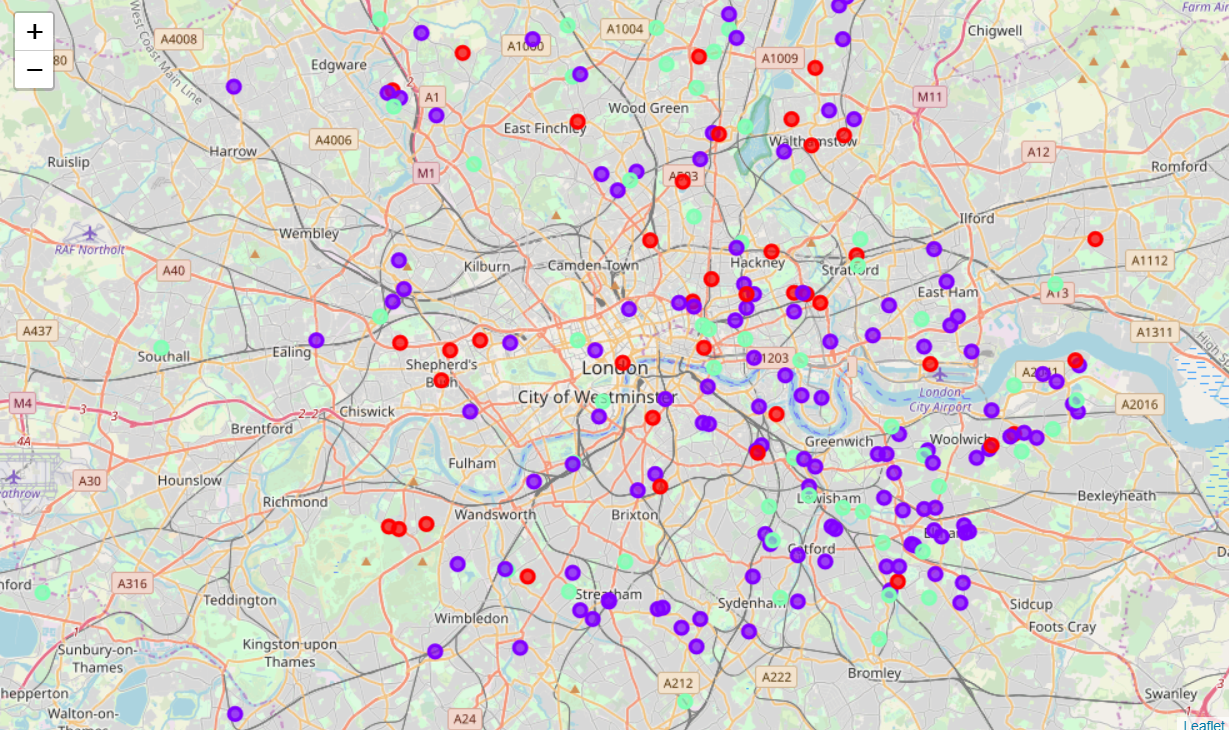
Our Venues Data now shows us the 10 most common Venues per Street

Using the .shape method again we see now that in this format our data only has 337 rows and is manageable. This data frame is ready for clustering using the k-means clustering method. With this method we attempt to categorize our real estate into 3 different groups based on price.

Here is a look at the first 30 rows of the now clustered data frame.



Finally, we plot the Clusters into the map to see how they are distributed. Each color is a cluster label.



Clustered Map of London. Red: Cluster 0, Purple: Cluster 1, Blue: Cluster 2

We see that there is no clear geographical distinction between the clusters since they are somewhat randomly distributed on the map. But if we take a look at each clusters nearby venues, the reasons of the different pricing can be identified almost instantly.



Cluster 0



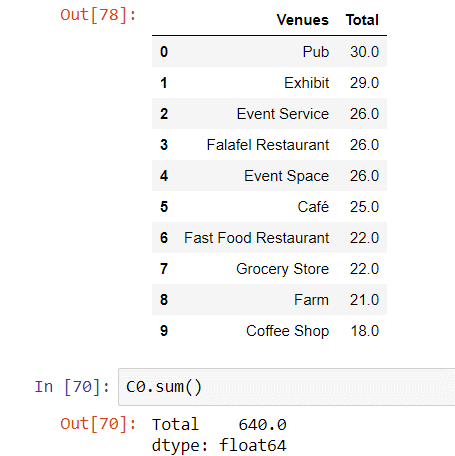
Cluster 1



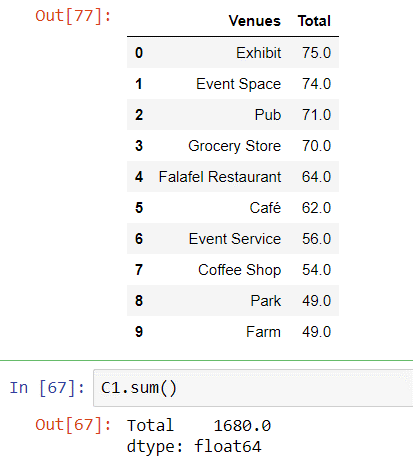
Cluster 2:

Now that we have created and viewed our clusters, we can clearly see that Cluster 0 (Lowest Prices) is the smallest having only 69 Venue Categories while Cluster 1 (Highest Prices) is the largest having 183 Venue Categories and our mid-price Cluster 2 is sitting in between with 110 Venue Categories. We can easily conclude that as the price rises, we can expect increasingly more variance in Venues.

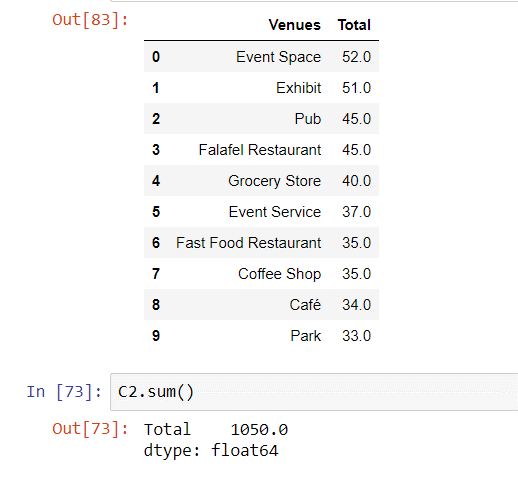
For our next step we will see how our venues are distributed among the three clusters. To do so our Clusters need some reformatting. First, we drop column Avg\_Price, then we count how many times each Venue appears per row, we replace NaN values with 0, sum each row to get the number of times each Venue appeared ,put the index into a column so we can plot more easily, rearrange the data frame and finally we drop the old index. Let’s have a look at our new clusters and also calculate how many venues each one has:



Cluster 0 Venue totals



Cluster 1 Venue totals

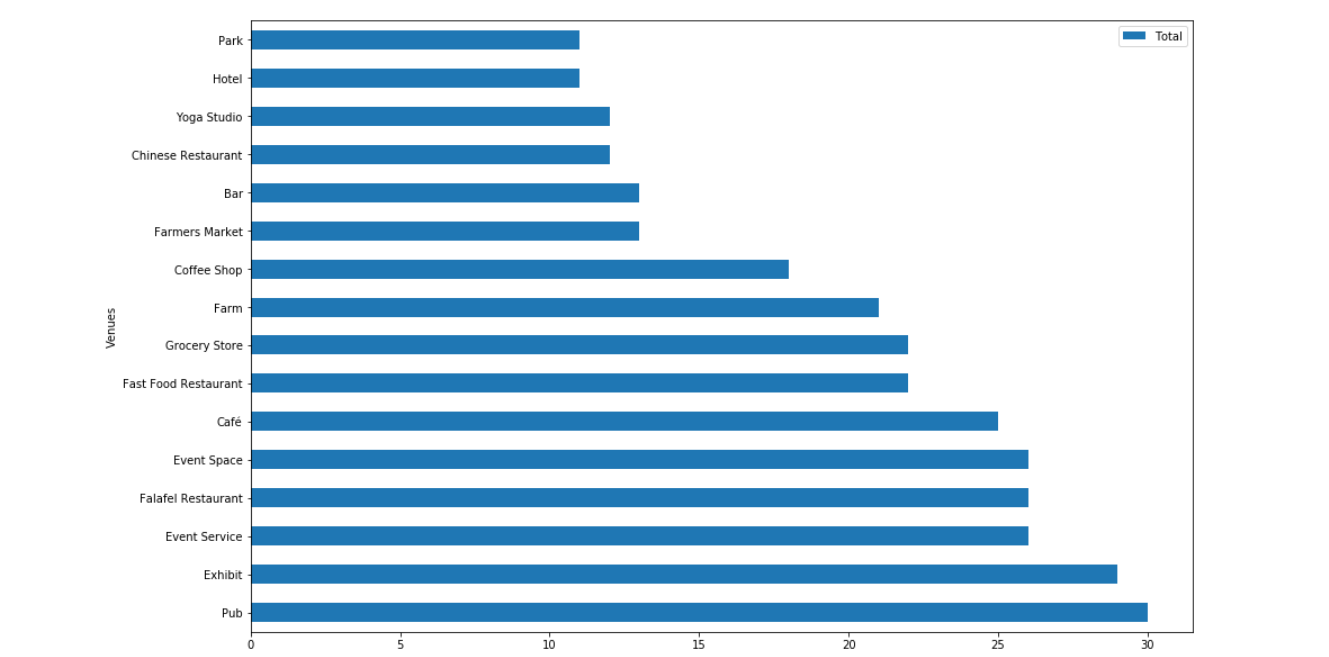


Cluster 2 Venue totals

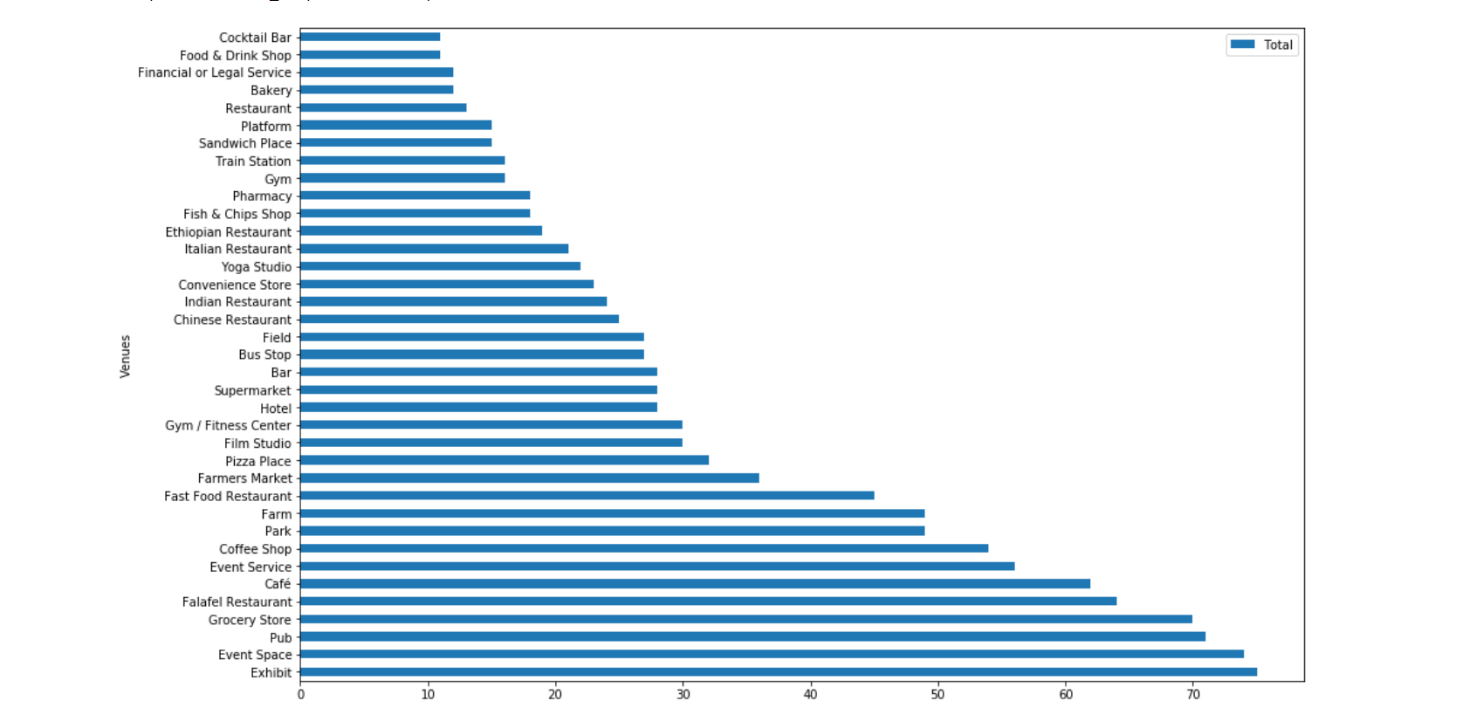
We can calculate how dense in terms of venues is each Cluster by dividing the number of venues within it by the total real estate. Following our calculations, we see that there is not much difference in density between each cluster since Cluster 0 has 9.28 venues per property Cluster 1 has 9.18 and Cluster 2 has 9.55 venues per property.

All three Clusters have almost the same venue density regardless of price, so we must look elsewhere to identify the reasons for the price differences.

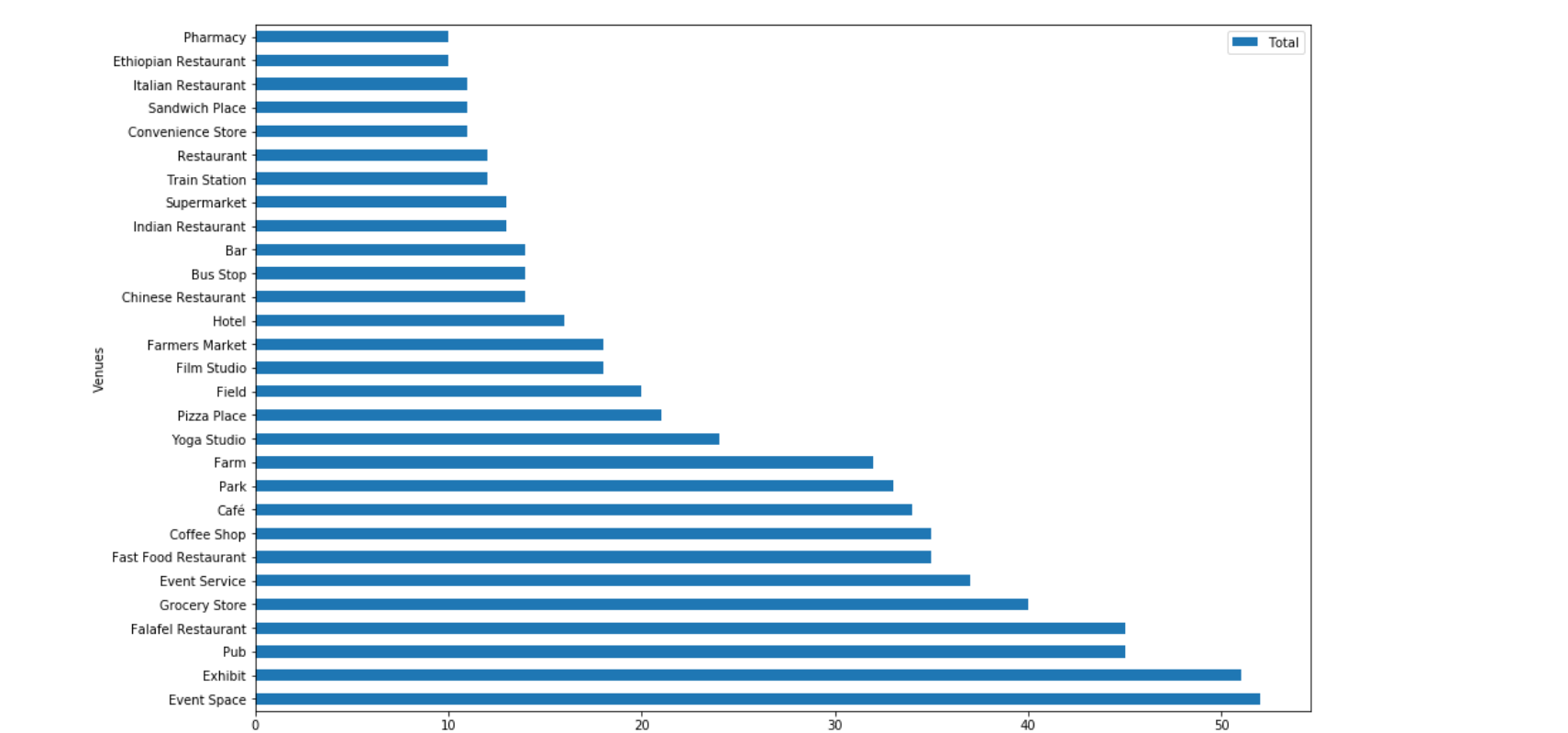
For our final analysis we plot each Cluster showing the frequency of which venue appears within it (we chose the only Venues that appear more than 10 times per Cluster to make the plot easier to understand:



Bar graph of the Venue Distribution for Cluster 0



Bar graph of the Venue Distribution for Cluster 1



Bar graph of the Venue Distribution for Cluster 2

By comparing the graphs, we can immediately see that cluster 1 has all the Venue categories the others have while also adding bakeries, cocktail bars, financial or legal services, fish & chips shops, gyms, Italian restaurants, platforms, supermarkets and train stations. Also, Clusters 1 and two share bus stops, convenience stores, Ethiopian restaurants, fields, film studios, Indian restaurants, parks, pharmacies, pizza places and sandwich places. The Venues not mentioned can be found on all three Clusters.

## 4. Results and Discussion section

First, even though the London Housing Market has relatively high prices, finding affordable homes is still possible. We may discuss our results for every Cluster.

#### High: Cluster 1 (Purple)

As the Higher end Tier this Cluster is the most complete having both vital venues such as Convenience Stores and Food and Drink stores, while also having Higher end Venues such as Airport Services and all kinds of cuisines and services. This cluster has the most and the most variable venues. Ideal investments for anyone that can afford the steep price.

#### MId: Cluster 2 (Cyan)

As the mid-Tier those Cluster seem more balanced in terms of Venues having both lower end (i.e. coffee shops, Grocery stores) and some unique and higher end ones (airport and train stations, foreign cuisine, legal services). This Cluster sits perfectly in between the other two both in price, variety and amount of Venues and thanks to that balance this Cluster is our default choice.

#### Low: Cluster 0 (Red)

The final Cluster has the lowest Avg prices. It doesn't have higher end Venues (ex. foreign cuisine or financial services) which plays a role on the reduced prices. We believe such facilities are not essential for a student so this cluster should be prioritized. Also, the big amount of Pubs, Exhibits and Cafés are certain to welcome new students to the University lifestyle.

## **5. Conclusion**

To solve our business problem, we clustered London neighborhoods in order to identify affordable homes depending on nearby venues and the current average price of real estate where homebuyers can live. We recommended profitable venues according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores.

First, we gathered data on London properties and the relative price paid data were extracted from the HM Land Registry (http://landregistry.data.gov.uk/). Moreover, to explore and target recommended locations across different venues according to the presence of essential facilities, we accessed data through Foursquare API interface and cross referenced them with the Registry data. By cross referencing the data on London properties and the data from Foursquare API interface, we were able to identify real estate investments that both fall into our budget and have enough nearby venues to support different lifestyles.

Second, The Methodology section comprised four stages: 1. Collect Needed Data; 2. Explore and Understand Data; 3. Data Wrangling; 4. Modeling. We chose the k-means clustering technique as it is the most efficient in terms of computational cost, and it is highly flexible and accurate.

Finally, we identified 3 Cluster based on price range to tried to identify what kind of homebuyer is best served by them.