Property Price Analysis/Visualization of London

Business Problem

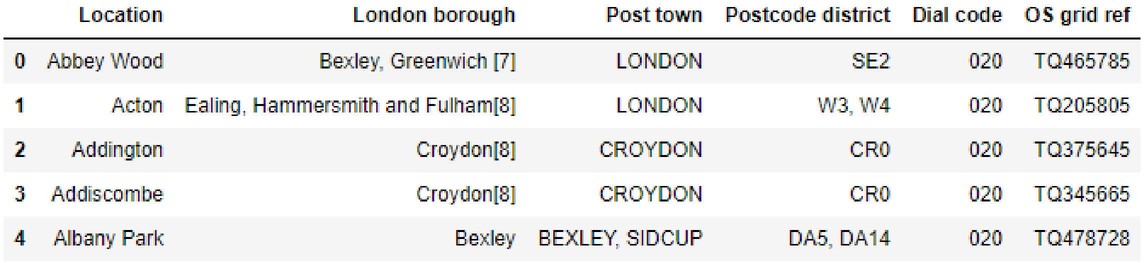
When someone looks to move from one neighborhood to another in a major city, it can be assumed that they will usually look for neighborhoods similar/dissimilar to the one they have stayed in earlier. This can be a factor when deciding which neighborhood to move to. So, after retrieving property prices and relevant neighborhood data, a system can be designed to cluster/classify neighborhoods and help the stakeholders decide on the best fit according to their preferences.

London being the largest/most populated city in the UK, this can be a very valid use case for people looking to move in London, from any other city or London itself.

Data

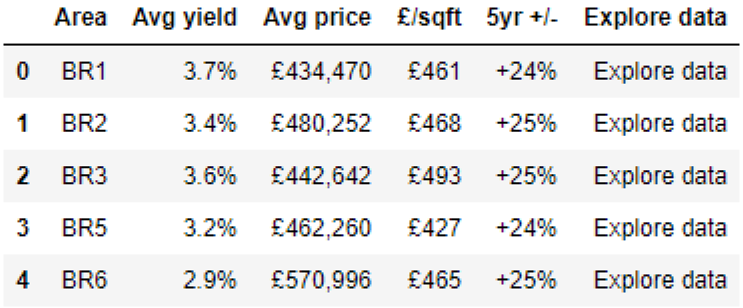
To build this service, we make use of property prices data from propertydata.co.uk and Foursquare API for the other venue/amenity details. Combining both, we can make a dataset to compare the amenities/prices of individual neighborhoods and then cluster them based on similarity. The neighborhood/area codes are retrieved from Wikipedia.

1. Neighborhood codes/Area Data: https://en.wikipedia.org/wiki/List_of_areas_of_London
    a. Sample:



	Location	London borough	Post town	Postcode district	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728

2. Property Data for London: https://propertydata.co.uk/cities/london
   a. Sample:



	Area	Avg yield	Avg price	£/sqft	5yr +/-	Explore data
0	BR1	3.7%	£434,470	£461	+24%	Explore data
1	BR2	3.4%	£480,252	£468	+25%	Explore data
2	BR3	3.6%	£442,642	£493	+25%	Explore data
3	BR5	3.2%	£462,260	£427	+24%	Explore data
4	BR6	2.9%	£570,996	£465	+25%	Explore data

3. Neighborhood Information/Venues/Amenities: https://api.foursquare.com

Methodology

Data Preprocessing

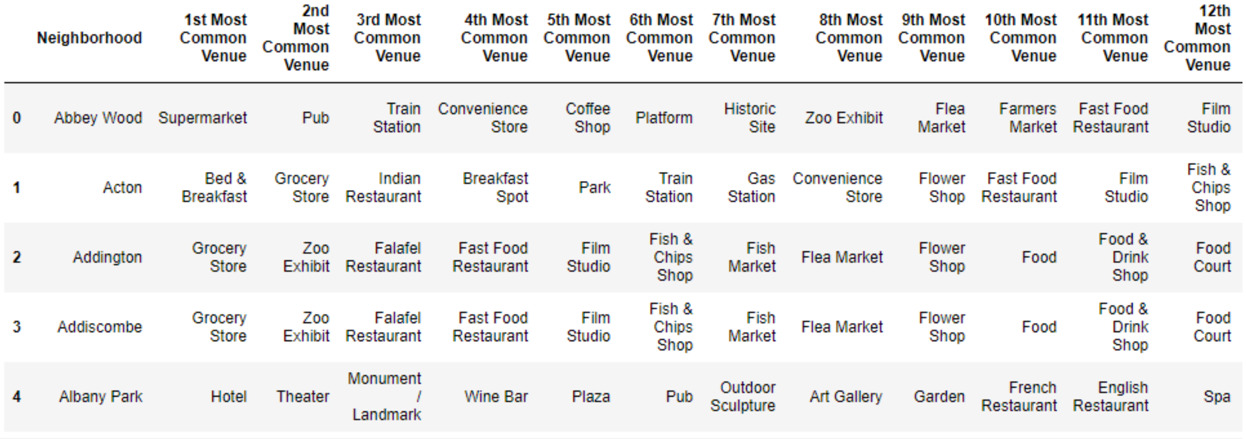
There were certain preprocessing steps followed, to clean up data from the various sources.

Hyperlinks and numbers were removed from the Area data, and some rows having multiple data sets were split into individual data cells. Columns which were not relevant to the problem at hand such as trends etc. were dropped from all data sets.

After Foursquare data was retrieved, only the relevant venue details were used. One hot encoding was used to convert the categorical data into numerical data to use in clustering.

For the final clustering input, only the top 15 types of venues for each neighborhood were considered and the rest were ignored. This was to minimize noise as much as possible.

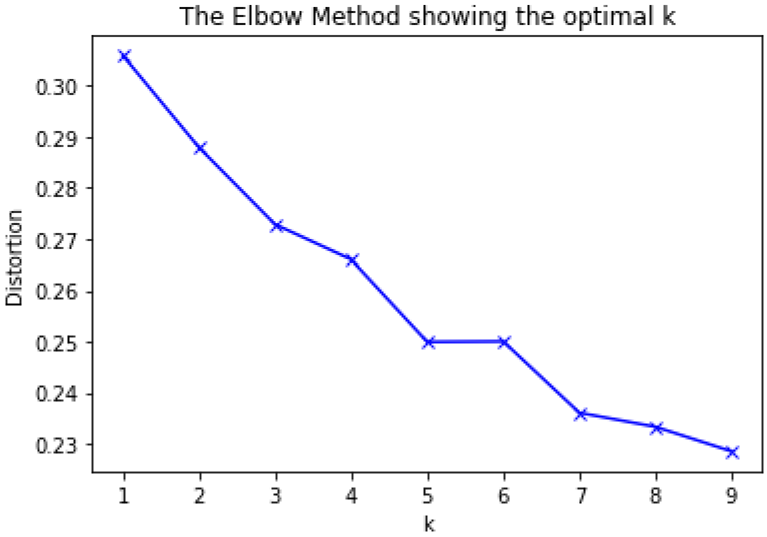
The price of properties was normalized, to nullify any disproportionate influence on the clustering algorithm.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue
0	Abbey Wood	Supermarket	Pub	Train Station	Convenience Store	Coffee Shop	Platform	Historic Site	Zoo Exhibit	Flea Market	Farmers Market	Fast Food Restaurant	Film Studio
1	Acton	Bed & Breakfast	Grocery Store	Indian Restaurant	Breakfast Spot	Park	Train Station	Gas Station	Convenience Store	Flower Shop	Fast Food Restaurant	Film Studio	Fish & Chips Shop
2	Addington	Grocery Store	Zoo Exhibit	Falafel Restaurant	Fast Food Restaurant	Film Studio	Fish & Chips Shop	Fish Market	Flea Market	Flower Shop	Food	Food & Drink Shop	Food Court
3	Addiscombe	Grocery Store	Zoo Exhibit	Falafel Restaurant	Fast Food Restaurant	Film Studio	Fish & Chips Shop	Fish Market	Flea Market	Flower Shop	Food	Food & Drink Shop	Food Court
4	Albany Park	Hotel	Theater	Monument / Landmark	Wine Bar	Plaza	Pub	Outdoor Sculpture	Art Gallery	Garden	French Restaurant	English Restaurant	Spa

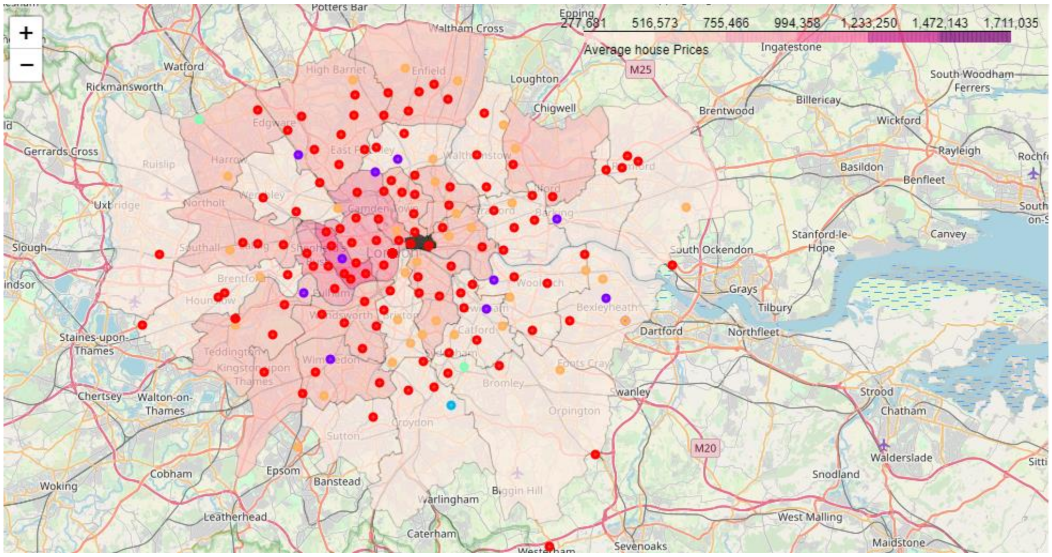
Algorithm Selection and Evaluation

K Means clustering was used as it is a robust unsupervised algorithm to cluster groups of similar items. After plotting for the optimal number of the number of clusters, it was decided to use 5 clusters as based on the elbow methodology for optimal number of clusters.



Results

Based on the above methodology, we were able to generate a heat map to show the clustering of neighborhoods as well as the average prices of suburbs. This can be of valuable information to people looking to move in/around London. However, there is no clear correlation between prices and clusters.



Discussion

Further work/analysis can be done on this in order to get more insights on significant drivers of property prices. As observed in the above map, there is no clear correlation between prices and clusters.

Conclusion

Therefore, using the above map and results obtained, anyone desiring to move into some other neighborhood in London can view similar neighborhoods and make an informed choice. The multiple clusters associated with the areas have also been elaborated on in the Notebook.