Clustering London Boroughs on Real Estate Prices

1. Introduction

Background- London is the capital of United Kingdom and is the largest city in UK. With the largest municipal population in the European Union, London has a diverse range of people and cultures, and more than 300 languages are spoken in the region. Its estimated population is 9.30 million according to the UN's World Urbanization Prospects, the most populous city in the European Union and accounting for 13.4% of the UK population. London's urban area is the second most populous in the EU, after Paris, with 9,787,426 inhabitants at the 2011 census. The population density is 14,500/sq mi.

Problem- London is a city with a high population and population density. As from Real Estate investor point of view we want to invest in such places were the housing prices are low and the facilities (shops, restaurants, parks, Hotels, etc.) and social venues are nearby. Keeping above things in mind it is very difficult for an individual to find such place in such big city and gather this much information. 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 district is clustered according to the venues.

2. Data Collection and Pre-processing

Data Collection- Data was collected from several website which are-

List of areas of London and London Boroughs

Website-https://en.wikipedia.org/wiki/List_of_areas_of_London

List of latest housing prices in London

Website-https://propertydata.co.uk/cities/london

Foursquare API

Website-https://developer.foursquare.com

GeoJson file of London Boroughs

Website-https://joshuaboyd1.carto.com/tables/london_boroughs_pro per/public

London Boroughs Outline

Website-https://skgrange.github.io/www/data/london_boroughs.json

Data Pre-processing- Data collected from Wikipedia and data on Real estate was cleaned first.

I removed all the hyperlinks and there are more than one Postal codes for some Locations so I kept only one Postal code. First of all I removed all null values and then get rid of unwanted columns and only kept 'Area' and 'Avg price' columns. Then 'Avg Price' columns contains string so I processed it to make integer by removing pound sign and comma. After cleaning two tables I performed inner join and merge two table and from resulting table I dropped 'Dial Code' and 'OS grid ref' columns as they were of no use. Then by using geocoder library I find the Longitudes and Latitudes of the Location and add a columns of each in my dataframe.

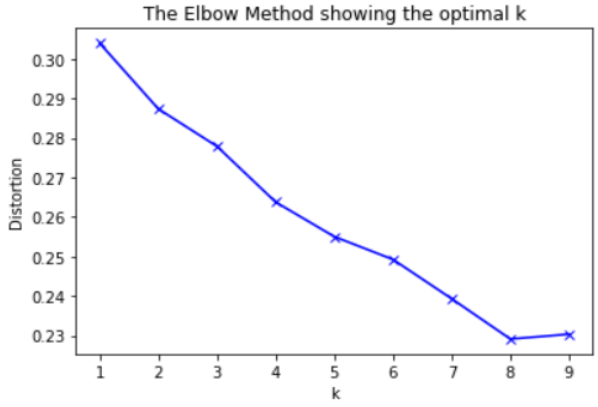
I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as 100 venue and the radius 500 meter for each borough from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API. Finally by using the Foursquare API in conjunction with the created datasets, a table of most common visited venues in London neighbourhoods is generated.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbey Wood	51.49245	0.12127	Lesnes Abbey	51.489526	0.125839	Historic Site
1	Abbey Wood	51.49245	0.12127	Sainsbury's	51.492824	0.120724	Supermarket
2	Abbey Wood	51.49245	0.12127	Lidl	51.496152	0.118417	Supermarket
3	Abbey Wood	51.49245	0.12127	Abbey Wood Railway Station (ABW)	51.490825	0.123432	Train Station
4	Abbey Wood	51.49245	0.12127	Platform 1	51.491023	0.119491	Platform

3. Machine Learning

We have some common venue categories in boroughs. In this reason I used unsupervised learning K-means algorithm to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning.

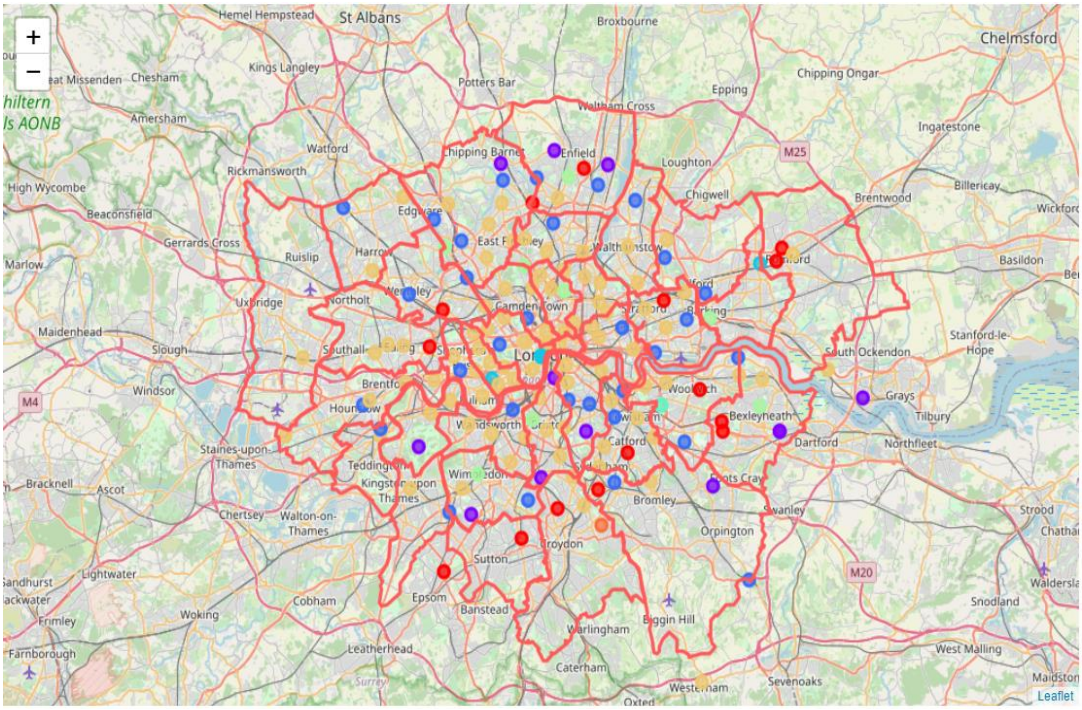
Firstly I ran K-Means algorithm to find out best K using elbow method. From the curve it was evident that best K is 8.



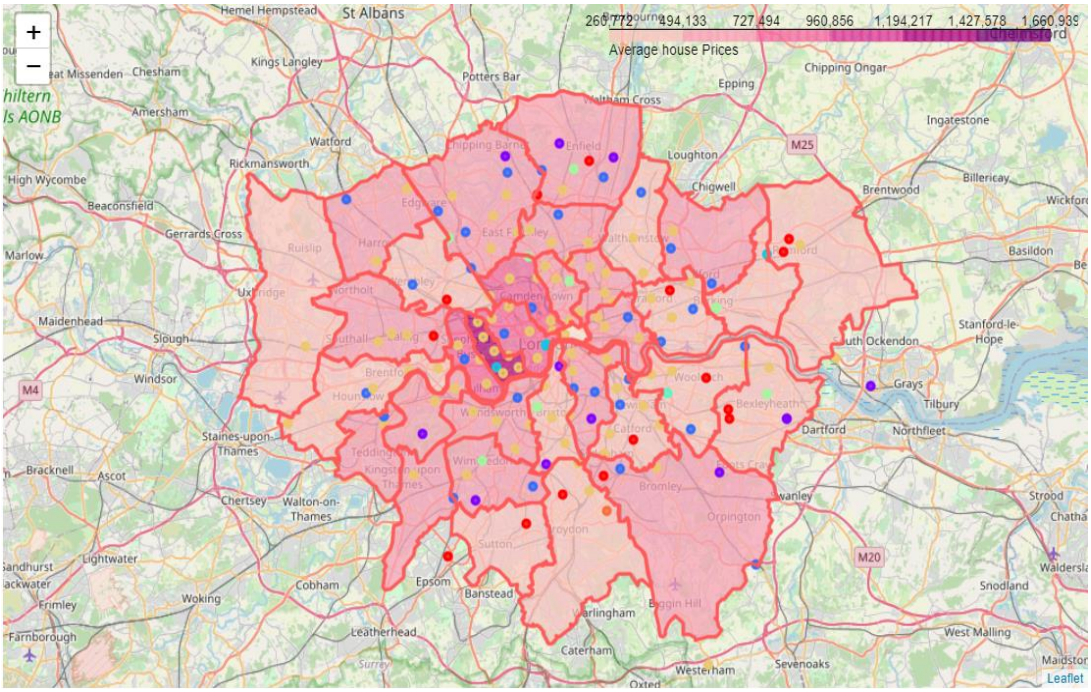
Then I ran K-Means to cluster the boroughs into 8 clusters .Then I merged table with cluster labels for each borough.

4. Visualization

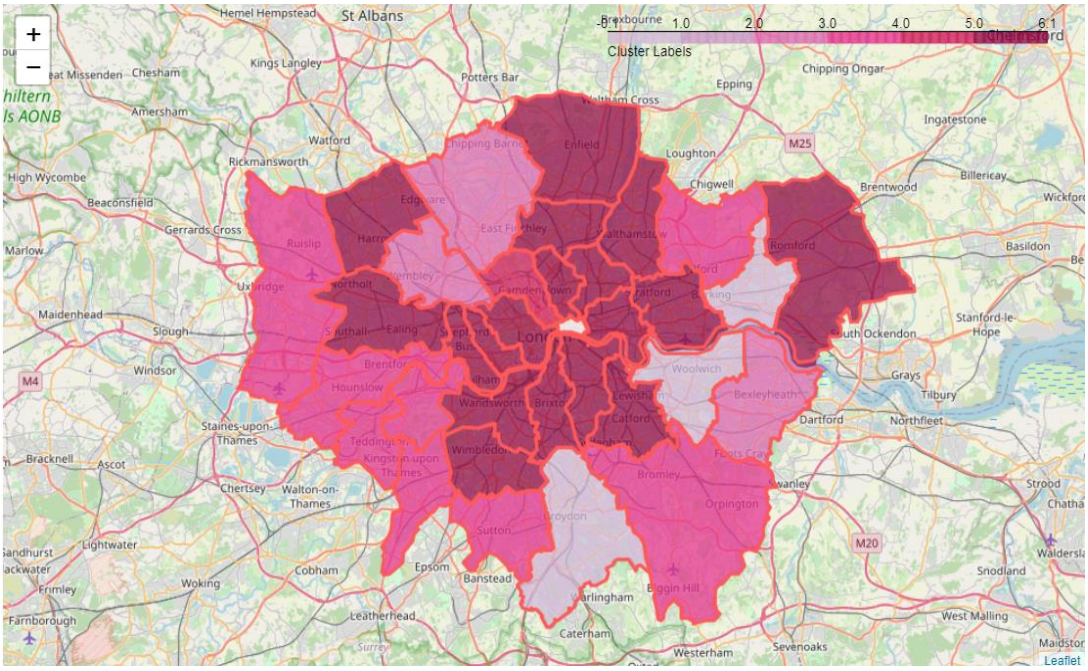
Then I generated a map of London with all clusters imposed on the graph with different colours.



After this I generated a Choropleth Map for visualizing London Boroughs and Avg Price.



After this I generated a Choropleth Map for visualizing London Boroughs and Clusters.



5. Result

I came to the result that the house prices in the downtown and with Hotels and Social venues nearby are very high.

While in the suburbs and the neighbourhoods away from the city centre have low prices but the facilities are also good. Almost all low price neighbourhoods are close to restaurants, pubs, sports facilities etc.

Some Boroughs such as Westminster, and Kensington and Chelsea have very high house prices. Bexley, Croydon, and Sutton Boroughs have very low house prices but have good venues to visit nearby.

6. Conclusion

As people are turning to big cities to start a business or work. For this reason, people can easily interpret where to live with all facilities and cheaply.

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