Business Problem:

London is the capital of and largest city in England and the United Kingdom, 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 mid-2016 municipal population (corresponding to Greater London) was 8,787,892, the most populous of any 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.

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 venue density.

Data Collection

To consider the above problem the data is collected as following:

1. I found the List of areas of London with its boroughs and postcodes from Wikipedia.

(https://en.wikipedia.org/wiki/List_of_areas_of_London)

- For housing prices, I searched and found a great website where latest London house prices were available with postal codes.

  [https://propertydata.co.uk/cities/london]

- I used **Forsquare** API to get the most common venues of given Borough of London.

  [https://developer.foursquare.com/]

- For choropleth maps I used .geojson file of London.

  [https://joshuaboyd1.carto.com/tables/london_boroughs_proper/public]

Data Preprocessing

First of all the data scraped from Wikipedia has to be clean.

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 1400 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 neighborhoods is generated.

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. First, I run K-Means to cluster the boroughs into 6 clusters because when I analyze the K-Means with elbow method it ensured me the 6 degree for optimum k of the K-Means. Then I merged table with cluster labels for each borough. After examining each cluster I label each cluster as follows:

1. Mixed Social Venues
2. Hotels and Social Venues
3. Stores and seafood restaurants
4. Pubs and Historic places
5. Sports and Athletics
6. Restaurants and Bars

After examining Average Prices I label each price as follows:

- >500000 : “Low Level 1”
- 500000–750000 : “Low Level 2”
- 750000–1000000 : “Average Level 1”
- 1000000–1250000 : “Average Level 2”
- 1250000–1750000 : “High Level 1”
- <1750000 : “High Level 2”

Result

I came to the result that the house prices in the downtown and with Hotels and Social venues nearby are very high you can clearly visualize in the map above while in the suburbs and the neighborhoods away from the city center have low prices but the facilities are also good. Almost all low price neighborhoods 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.

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.

References

1. https://en.wikipedia.org/wiki/London
2. https://en.wikipedia.org/wiki/List_of_areas_of_London
3. https://propertydata.co.uk/cities/london
4. https://developer.foursquare.com/
5. https://joshuaboyd1.carto.com/tables/london_boroughs_proper/public