Capstone Project - The Battle of Boroughs

Moving and buying a house in London, UK

Created by

Kwan Chan

05-02-2020

Introduction:

Background:

London is the capital and largest city of the UK, as well as being one of the world's most important financial centers. Aside from being the world's most popular for work, its arts, commerce, education and media have also attracted a lot of individuals/families to move there for a living. Different aspects need to be considered when it comes to buying real estates: cost, safety and services of the neighborhood, services, convenience etc.

Business Problem:

As a foreigner thinking of moving to London, they most likely do not have an idea which area to look and how to start the house hunting. Therefore, it is crucial to provide the location data to the customer displayed on a map showing aspects of different areas ie. Crime rate, point of interest, house price.

Data:

Data Acquisition:

Necessary libraries and packages, they will be used to build the model for the analysis.

A list of boroughs and areas in London (https://en.wikipedia.org/wiki/London_boroughs), it will be used to request the foursquare data about the venues in these areas.

Geographical coordinates of the neighbourhoods in London, these will become the markers of the different areas on the map.

Crime rate in London by borough (https://www.finder.com/uk/london-crime-statistics)

London Map from folium, this will be the base map where the areas will be superimposed onto.

London house prices by borough (https://www.theweek.co.uk/99093/london-house-prices-which-boroughs-are-falling-and-which-are-on-the-rise)

Methodology

Data Exploratory:

Necessary libraries and packages are first imported, such as numpy, pandas, geocoder for mapping, sklearn for clustering, folium for mapping and BeautifulSoup for web scraping.

Next is the web scrapping, information on the London boroughs such as coordinates and list of boroughs are extracted from a Wikipedia page, it is then put into a pandas dataframe, techniques such stripping, replacing, extracting are used to form the dataframe. Crime data and average house price of each borough are then extracted from different websites, forming another 2 dataframes using similar exploratory data techniques. The 3 dataframes produced are then combined together as london_data.

Geographical coordinates of London are then obtained and as a preview, a London map with the boroughs marked on top is visualized using folium. Next, data of the top 100 venues in each borough are obtained from foursquare (using defined credentials), they include the names and categories of the venues and their coordinates. Each borough is then analysed ie. One-hot encoding, grouping

venues by borough and by taking the mean of the frequency of occurrence of each category. These are put into a pandas dataframe that displays the top 10 venues for each borough.

Data Analysis:

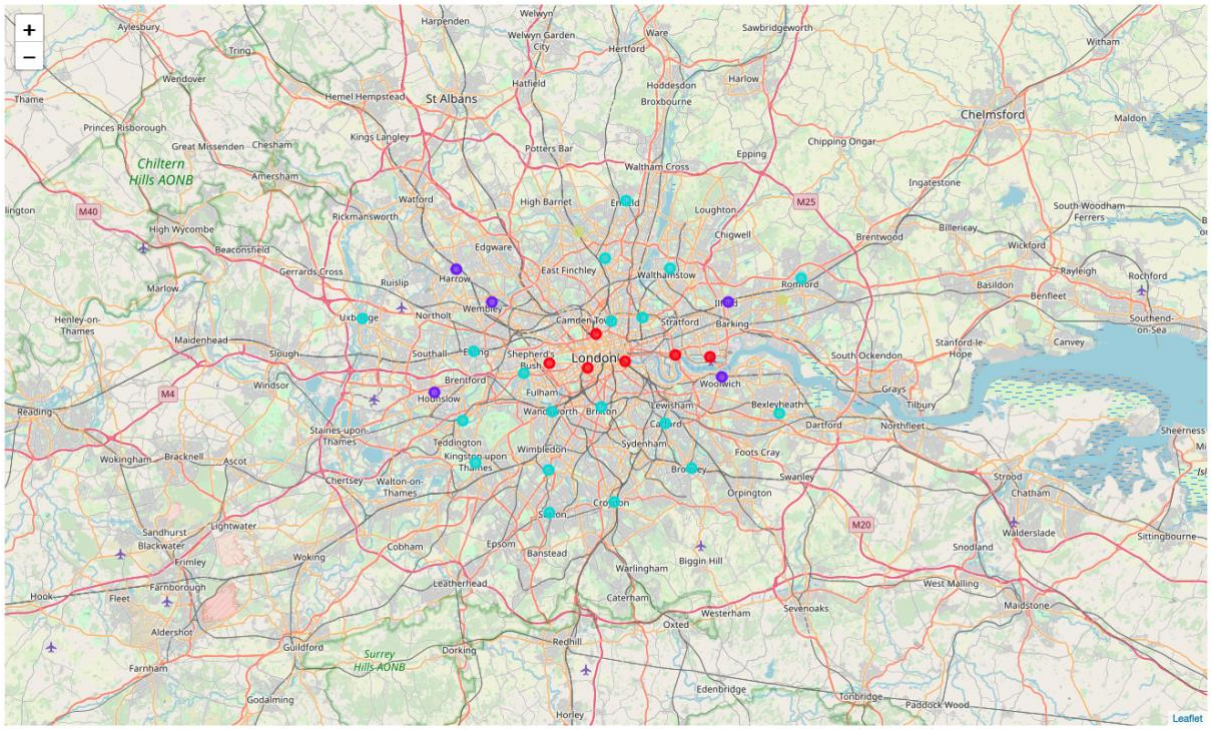
K-means Clustering is used to group the borough as the first analysis. They are grouped based on their most common venues with a set number of 4 clusters. Clusters are then mapped on the London map as markers.

Next, the house prices are classified into 'Very Cheap', 'Cheap', 'Average', 'Expensive', 'Very Expensive' of equal classes, based on the average house value by borough. It is visualized on a map with color-coded markers representing each borough.

Lastly, the safety is classified into 'Very Safe', 'Safe', 'Average', 'Dangerous', 'Very Dangerous' of equal classes based on the crime count per year by borough. It is visualized on a map with color-coded markers representing each borough.

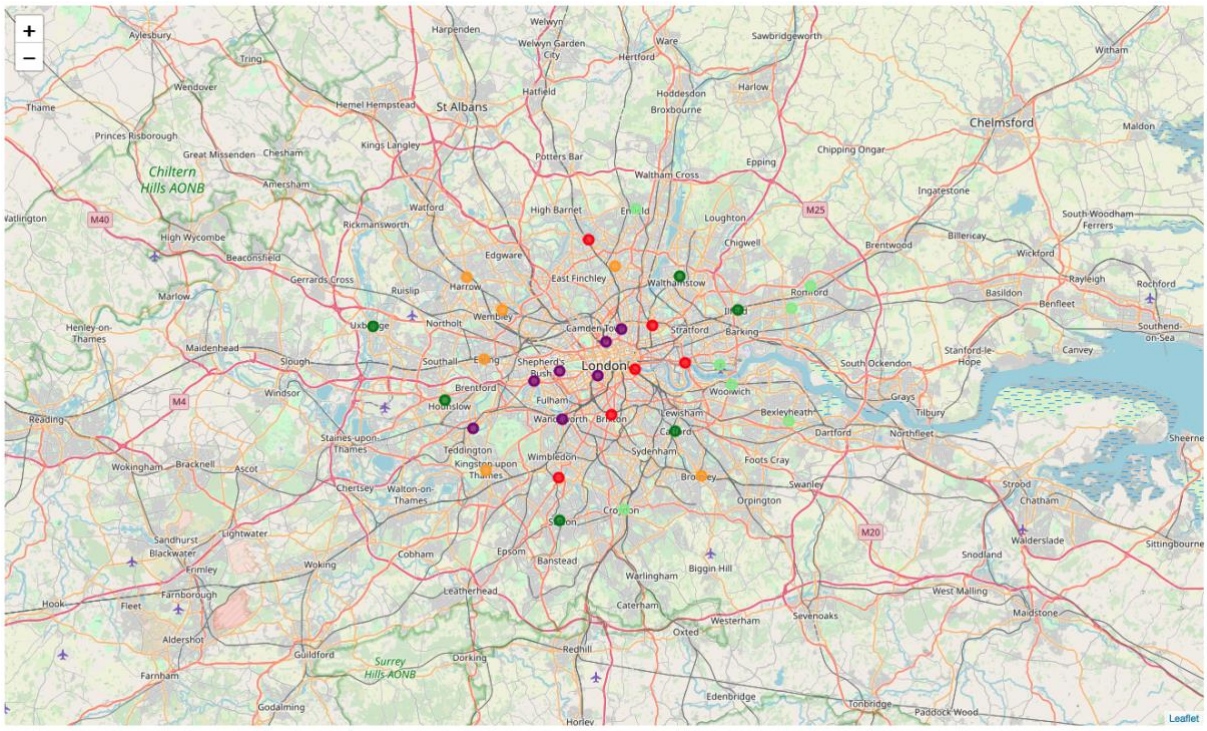
Results

K-means Clustering:



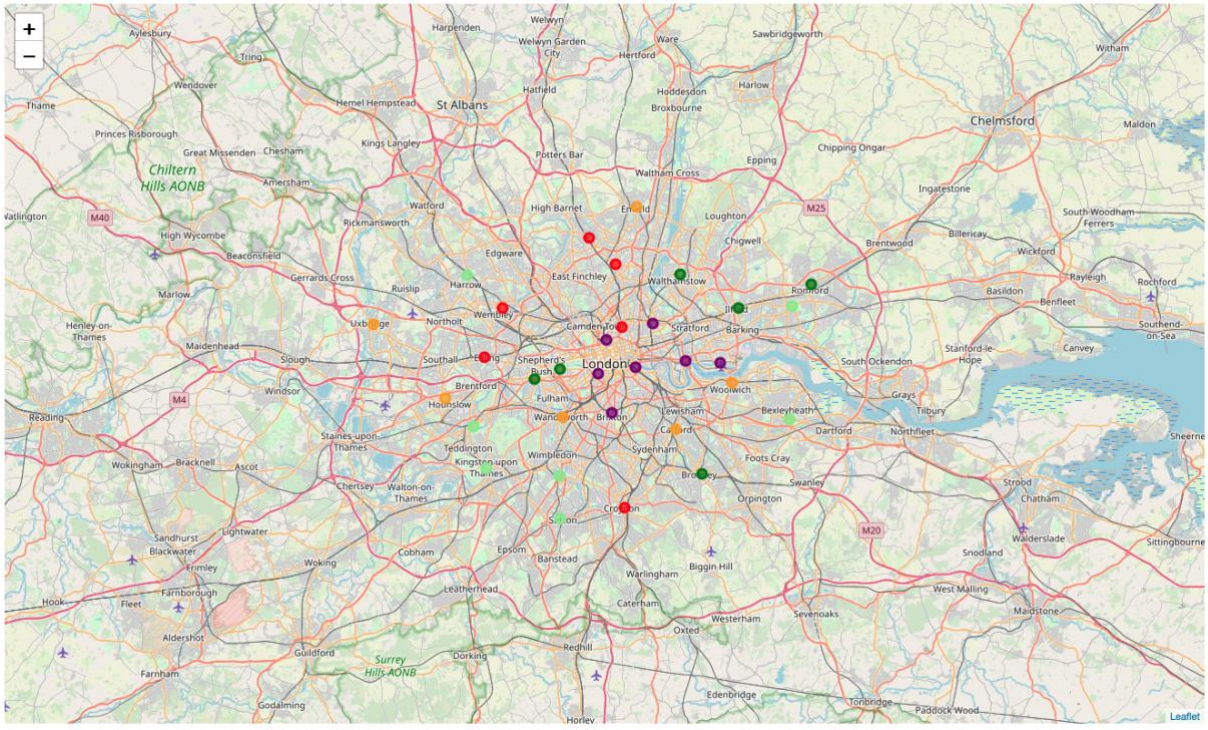
The 4 clusters, namely: Cluster 1 (red) - Commercial+Tourist Boroughs, Cluster 2 (purple)- Suburban Boroughs with plenty of Shops, Cluster 3 (blue) - Boroughs with lots of Pub/Coffee shops & Cluster 4 (yellow) - Quiet Boroughs, are mapped onto the London map.

House Price Visualised:



The 5 classes are namely: 'Very Expansive' (purple), 'Expansive' (red), ‘Average’ (orange), ‘Cheap’ (green), ‘Very Cheap’ (light green)

Borough Safety Visualised:



The 5 classes are namely: 'Very Dangerous' (purple), 'Dangerous' (red), ‘Average’ (orange), ‘Safe’ (green), ‘Very Safe’ (light green)

Discussion

K-means Clustering:

Cluster 1 - It is clear that the central boroughs are mainly for commercial/tourism with top venues being hotels and restaurants. Ideal for young individuals/couples for a fast-paced life style.

Cluster 2 – The outer suburban boroughs are full of shops, grocery stores etc. Ideally for young families.

Cluster 3 – These boroughs have lots of pubs and coffee shops in the area, ideally for people who do not prefer the congested inner city life.

Cluster 4 – Quiet boroughs ideal for elderly and the retired.

House Prices Visualised:

It can be seen that the central/west central areas are the most expensive, reasons being the ease of commute from central to other parts of the country and the number of famous tourist spots in the area. Unexpected, just like any other country, the further from the central the cheaper the real estates are. The cheapest boroughs are located at the eastern end of the city.

Borough Safety Visualised:

The south/east central have the highest crime count per year, it is historically known that those are the unsafe parts of the city where there are more immigrants and drug trafficking. Also due to the much less human flow in the outer city, the crime counts are significantly less in those residential boroughs.

Conclusion

Overall it is not a bad representation of a London map for any potential immigrant to catch a glimpse of different aspects of each borough. They now have an idea of the life style living at any borough from the k-means clustering analysis. Depending on their budget, they can find their feasible options using the price map while checking their safety level on the safety map.