The Battle of Neighborhoods

Author Tzuhsien, Yang
Date 6th July, 2021

1. Introduction

Real estate buyers searching for a new home always face big decisions. There are lots of factors effecting where to buy, such as the trend of price growth, local amenities and other factors which might make negative effects on the property value. London, the capital and largest city of England and the United Kingdom, is one of the world's most important global cities. London's population is about 9 million which accounts for 13.4% of the U.K. population. Also, London has a diverse range of people and full of different cultures. London attracts people from all over the world.

Grocery stores, restaurants, shopping, entertainment, are some of the top amenities people are looking for when buying a home. In this project we will focus on grocery stores and restaurants as factors to be included in the features being targeted.

Problem Description

How to decide which borough in London the property to buy? For those who want to buy real estate in London, it's hard to know where to start.

Target Audience

This project aims to make an analysis of features for buyers who want to find and purchase property in London. The buyers can target at the features including local restaurants and grocery stores in each borough to decide where is the best borough to look for.

2. Data

Data was obtained from Wikipedia and Foursquare. Wikipedia provides a list of 32 London boroughs within Greater London. Foursquare consisted a comprehensive location data for understanding local amenities nearby targeted places. All of the above data was based on London borough which gave an easier way for the linkage because the variable in each datasets was consistent.

The datasets will include the following data:

1. London boroughs
   Source: https://en.wikipedia.org/wiki/List_of_London_boroughs
   Description: The data will be scraped from web url.
   Selected columns: borough, population, coordinates

2. Foursquare Location Data
   Source: https://foursquare.com
   Description: Retrieve the venues in each borough and filter grocery stores and restaurants.

3. Methodology

1. Data Analysis

Local amenities data with the categories of restaurant and grocery stores were linked with borough geospatial data by each boroughs. An unsupervised machine learning technique was utilized to identify clusters of boroughs, due to the reason that the dataset was unlabelled. The type of clustering methods we used was k-means clustering. K-means clustering models were built to perform the clustering with variable of top 10 common type of restaurants and grocery stores.

The K was determined by measuring silhouette coefficient. Silhouette coefficient quantifies how well a data point fits into its assigned cluster. The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

2. Data Cleaning

The borough data which was scraped from the web were consisted of Degrees minutes seconds (DMS) coordinates initially. Therefore, the longitude and latitude information was obtained from geopy library. However, after plotting London borough, it was found that there were 6 boroughs wrongly located, since the name of these boroughs were vague and commonly used. The coordinates data was corrected by manually updating the real information.

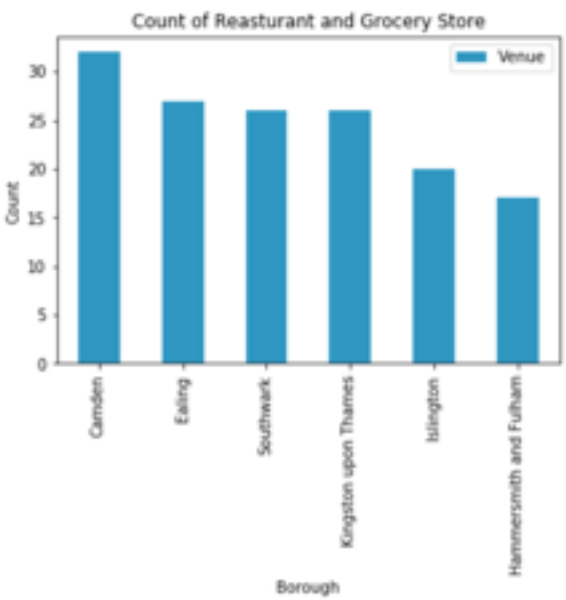
3. Library List

Library	Description
pandas	for easy and fast manipulation
numpy	provides multiple ways on dataframe and matrix manipulation
requests	http library for web data scraping
BeautifulSoup	for pulling data from web and navigating
Nominatim	geocoder to identify geometries
folium	mapping
KMeans	performing machine learning technique of K-means
silhouette_score	evaluating the K clusters
matplotlib	creating static, animated, and interactive visualizations

4. Results

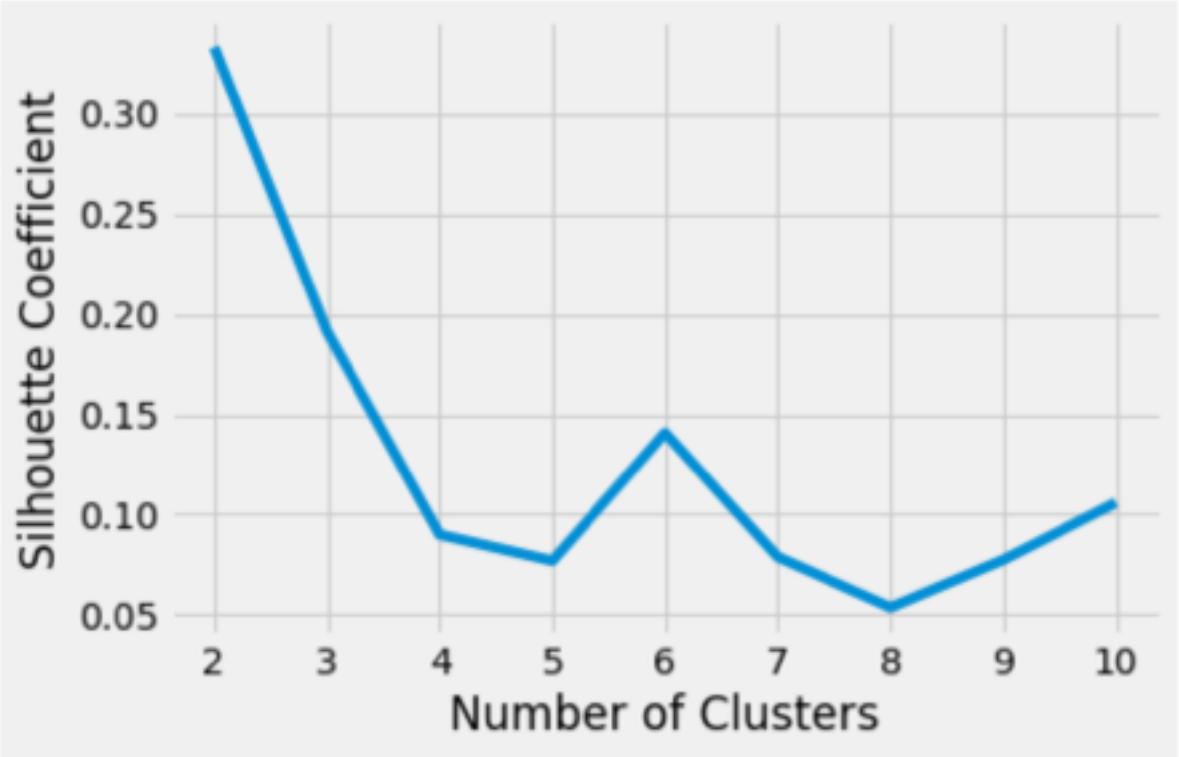
Firstly, London was plotted geospatially for understanding its location of each borough. The map gave an overview of how borough was distributed in London.

To further explore the city of London, local amenities data which was pulled from Foursquare were utilized for advanced analysis. The venue of restaurant and grocery stores were extracted, because these features were considered one of the most common factors to be considered before buying a house. However, after extracted data only included grocery stores and restaurants, it was found that these amenities were not in Bexley, Brent or Waltham Forest.



The data was aggregated and it found that Camden, Ealing, Southwark, Kingston upon Thames, Islington, Hammersmith and Fulham were the boroughs with the most quantities of the restaurants and grocery stores. Especially, there were more than 30 restaurants and grocery stores in Camden, which is an important place with rich retail, tourism and entertainment. The second borough which owns a number of amenities is Ealing. Ealing, there were 3 grocery stores, 3 Italian Restaurants, 3 Vietnamese Restaurant, while in there were 3 Chinese Restaurant, 2 Ramen Restaurant and 2 English Restaurant etc..

After data exploration, the dataset was modeled by the technique of machine learning to see how these boroughs would cluster. Before model fitted, the number of K should be determined in advance. We used silhouette coefficient to determine how many clusters gave a best fit for the data. The silhouette coefficient is a measure of cluster cohesion and separation. It quantifies how well a data point fits into its assigned cluster based on two factors: 1) How close the data point is to other points in the cluster? 2) How far away the data point is from points in other clusters? The result was plotted into a line chart for quick understanding. Silhouette coefficient score suggested that 2 clusters was best fit for the model.

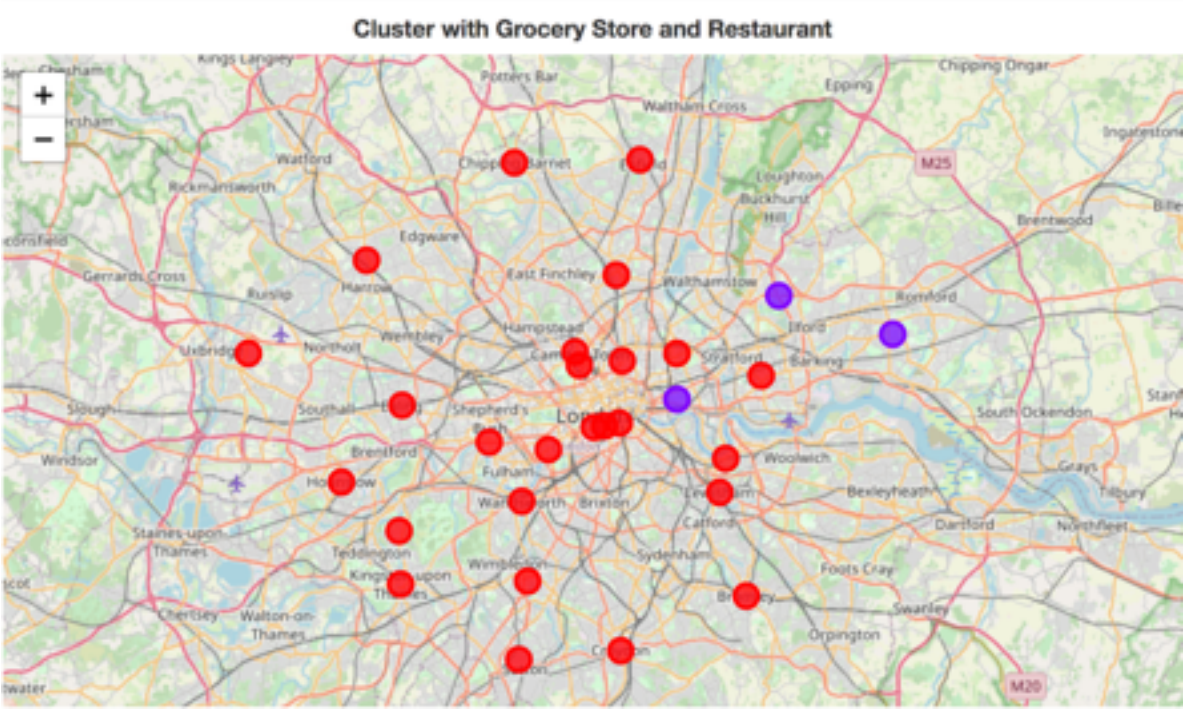


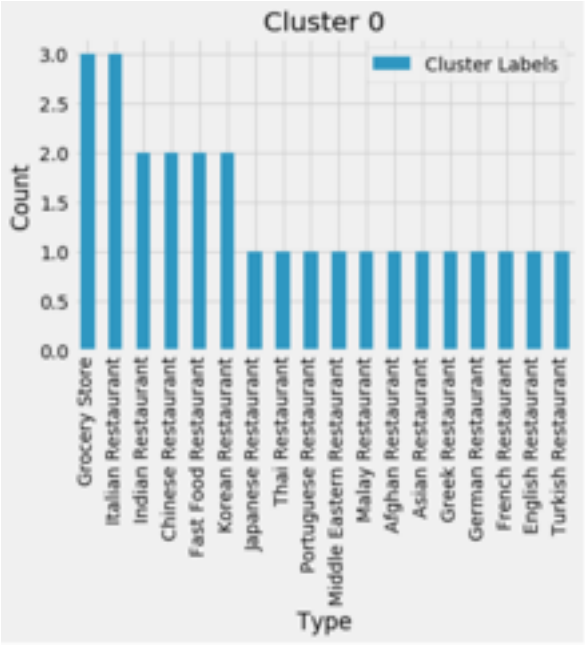
The dataset would be segmented into 2 clusters and the result was geographically plotted.
Red: Cluster 0 | Purple: Cluster 1

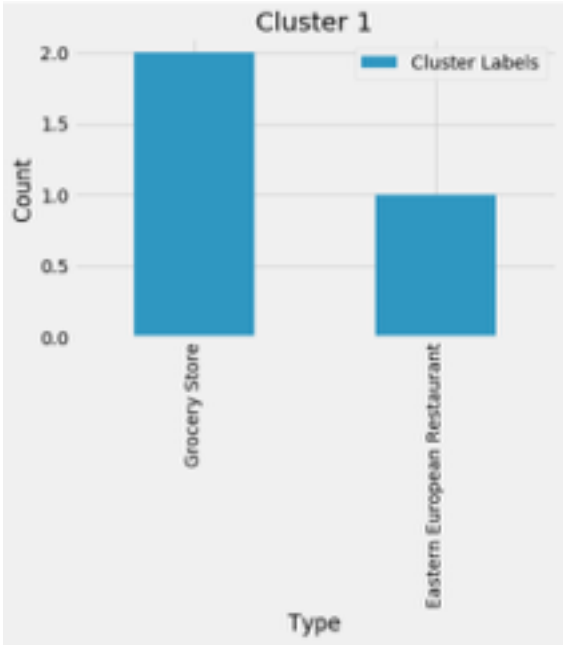
In cluster 0, color in red, we could see that there were 26 boroughs in cluster 0, and there are three boroughs with grocery store as the most common type of amenities, and other kinds of restaurants could also be easily found here. In cluster 1, consisted of 3 boroughs, most common venue found here included grocery store and Eastern European Restaurant. However, there was just a few restaurants Asian style.

5. Discussion

This project was conducted to support those who wants to buy a property in London for living using the technique of machine learning. The features of the borough we mainly focused on were grocery stores and restaurants. In the first stage of the project, the London borough data was initialized and was merged with local grocery and restaurant data. At this phase, since we focus on these two amenities, it was found that there were no grocery stores and restaurants found in Bexley, Brent or Waltham Forest. If the property buyers value the importance of grocery stores and restaurants, these three boroughs were suggested not to be considered in the first stage.







According to the cluster results, there were more options of different type of restaurants in cluster 0, while in cluster 1, the style of restaurants was limited. However, for grocery store, it was found that it should be easy to go grocery shopping in anywhere in London. This result could be concluded that cluster 1, Barking and Dagenham,Redbridge and Tower Hamlets, was suitable for those buyers who do not go out for dining often, and for cluster 0 could be considered for those who enjoy trying various type of restaurants. Especially, boroughs of Camden, Ealing, Southwark, Kingston upon Thames, Islington, Hammersmith and Fulham own as many as the restaurants and grocery stores.

6. Conclusion

Although there are some limitations and improvements on this project, it still gives an insight for those property buyers to discover which borough is filled with diverse restaurants and which borough has limited restaurants but does not lack of grocery stores. This is supportive to the decision making at a very first stage of borough screening.

Future recommendation: The features can consider more aspects, such as the distance to city center, the development of transportation, historical sale price, etc for more advanced discovery.