IBM Data Science Professional Certificate
Applied Data Science Capstone Project
Final Report
Finding My New home in London

by Panjapol Ampornratana

1.Introduction/Business Problem

1.1 Background and Description

What is the first thing that come to you mind when you considering buying a new house? When searching for a new house, we all might want to concentrate on a Neighborhood with low crime rate, average density of population, and access to nearby venues. All those criteria are usually considered with a certain budget in mind. This report is trying to find borough in London that cover all the necessary properties that we need for a new house.

1.2 Problem

Greater London is organized into 33 government districts:
- 32 boroughs
- The City of London

It's a huge region covers about 1600 square kilometer and has population about 9 million people, all boroughs are very different in many ways. In order to determine borough that satisfied our criteria, these boroughs need to segment based on their safety scores, population density, access to nearby venues. After all of this is done, I'll be able to find borough that suitable for my new house.

1.3 Interest

Obviously, who are interested in buying a new house in London would be very interest in our analyze result.

2. Data

2.1 The London Datastore

https://data.london.gov.uk

The London Datastore is a free and open data-sharing portal where anyone can access data relating to the capital. Whether you're a citizen, business owner, researcher or developer, the site provides over 700 datasets to help you understand the city and develop solutions to London's problems.

The datasets that are useful for us to solve the problem are a list of borough names and borough population for year 2019 and a list of borough level crime for most recent 24 months:

https://data.london.gov.uk/dataset/land-area-and-population-density-ward-and-borough
https://data.london.gov.uk/dataset/recorded_crime_summary

2.2 The Foursquare Places API

https://foursquare.com

The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag.

We will use this dataset to understand better how venues in each borough look like and it also help us know location of each borough so that we can create map with folium library.

2.3 Data Cleaning

After data downloaded from the London data store, we selected only data from 2019 and drop all the rest of it.

2.4 Feature Selection

I extract only useful columns such as population and borough size, then start to create a new feature to help us clustering such as crime rate per square kilometer and population density in each borough

3. Methodology

3.1 Exploratory Data Analysis:

1. Download csv file provided from data section and convert into dataframe by pandas library.

2. Choose borough data in year 2019 only and extract only useful columns.

	Code	Borough	Population	Square_km
0	E09000001	City of London	7953	2.9
1	E09000002	Barking and Dagenham	214858	36.1
2	E09000003	Barnet	402363	86.7
3	E09000004	Bexley	252885	60.6
4	E09000005	Brent	340710	43.2
5	E09000006	Bromley	334292	150.1
6	E09000007	Camden	255526	21.8
7	E09000008	Croydon	396548	86.5
8	E09000009	Ealing	354184	55.5
9	E09000010	Enfield	339480	80.8
10	E09000011	Greenwich	289650	47.3
11	E09000012	Hackney	286425	19.0
12	E09000013	Hammersmith and Fulham	186075	16.4
13	E09000014	Haringey	285949	29.6

Figure 1. A part of clear borough list from The London Datastore

3. Merged with crime data in year 2019, create new features for clustering, and use Python geocode library to find geolocation then add to data to dataframe.

	Borough	Code	Population	Square_km	Crimerate_sq	Population_density	Latitude	Longitude
0	Barking and Dagenham	E09000002	214858	36.1	0.559280	5.951745	51.543932	0.133157
1	Barnet	E09000003	402363	86.7	0.364302	4.640865	51.527095	-0.066826
2	Bexley	E09000004	252885	60.6	0.298218	4.173020	51.452078	0.069931
3	Brent	E09000005	340710	43.2	0.705023	7.886806	51.609783	-0.194672
4	Bromley	E09000006	334292	150.1	0.167908	2.227129	51.601511	-0.066365
5	Camden	E09000007	255526	21.8	1.826193	11.721376	51.532360	-0.127960
6	Croydon	E09000008	396548	86.5	0.391491	4.584370	51.593480	-0.083420
7	Ealing	E09000009	354184	55.5	0.573694	6.381694	51.514060	-0.300730
8	Enfield	E09000010	339480	80.8	0.378465	4.201485	51.540024	-0.077502
9	Greenwich	E09000011	289650	47.3	0.610677	6.123679	51.484540	0.002750
10	Hackney	E09000012	286425	19.0	1.814368	15.075000	51.545050	-0.055320
11	Hammersmith and Fulham	E09000013	186075	16.4	1.456585	11.346037	51.482600	-0.212880
12	Haringey	E09000014	285949	29.6	1.096892	9.660439	51.589270	-0.106405
13	Harrow	E09000015	258861	50.5	0.344059	5.125960	51.513180	-0.106980

Figure 2. A part of merged borough and crime dataframe

4. Create a map of boroughs with the size of borough by Folium library.

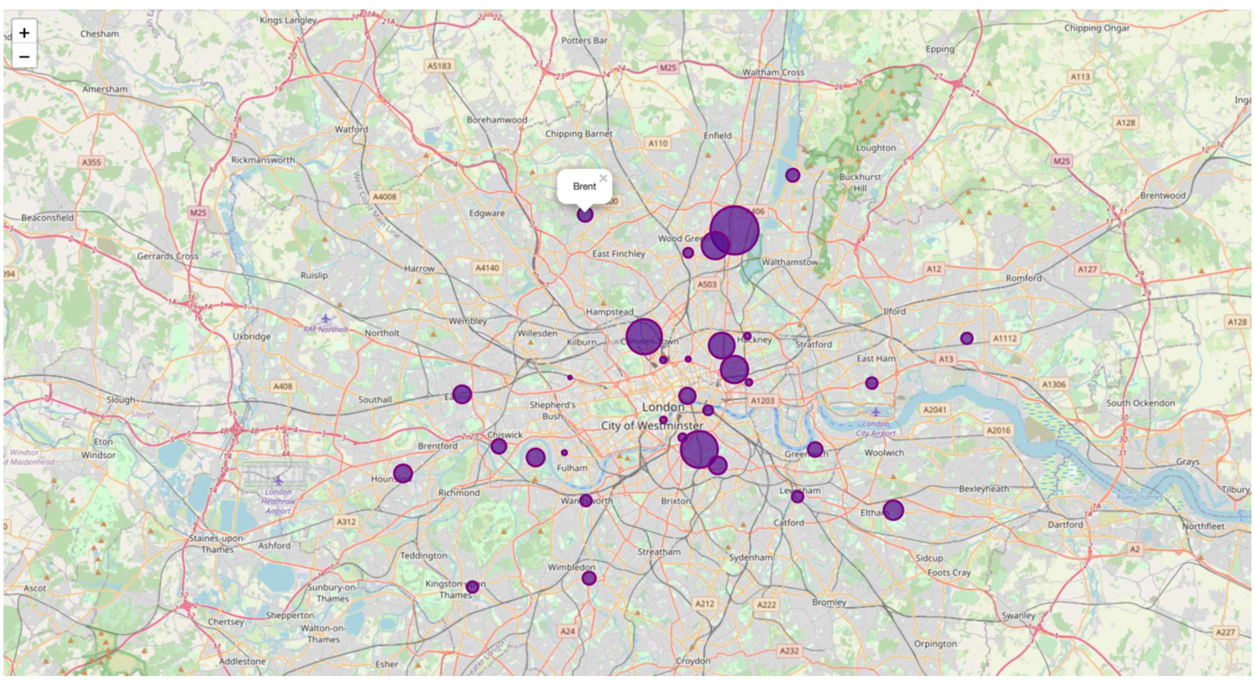

Figure 3. A map of boroughs of London

3.2 Clustering

There are three important factors that I wanted to consider before selecting the borough to live in: crime rate, population density, and access to venues. I have collected measures of crime over 1 year, measures of population density, and 274 venue category features. I would want these three categories to have equal weights when clustering, so I normalized data with StandardScaler module from sklearn and divided by number of measures for each factor.

After that, we use k-means clustering method to segment the borough data. First, the elbow method was used to determined how many clusters would be best to use. The rate of distortion reduction reduced at 3 clusters, so that's my numbers of clusters to use for k-means clustering analysis.

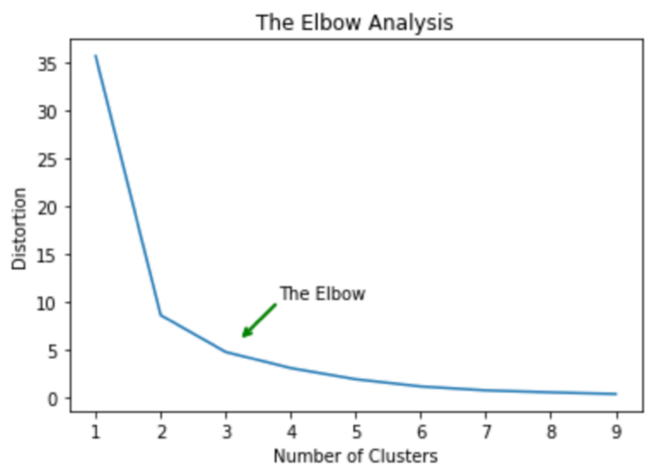

Figure 4. Graph from the elbow method

4. Result

K-means clustering analysis segmented London boroughs into three clusters represented different characteristics, the group of clusters that could be a good choice for buying a new house in London are cluster 1, due to a big size of borough, low crime rate, and good access to nearby venues.

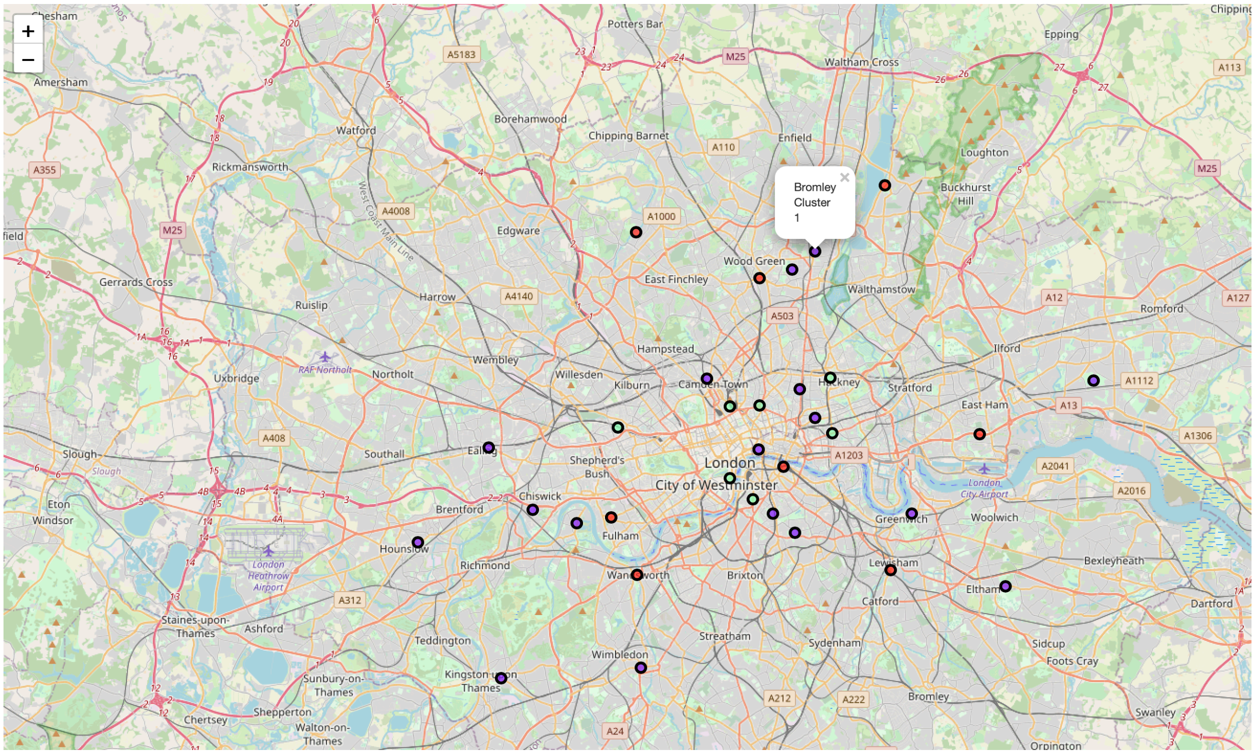

Figure 5. A map of similar borough clusters in London

	Cluster Labels	Population	Square_km	Crimerate_sq	Population_density
0	0	303638.250000	32.812500	0.984018	9.509030
1	1	282634.764706	68.858824	0.370081	4.565676
2	2	266329.428571	19.414286	2.109815	13.894638

Figure 6. Characteristics of each cluster groups

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Barking and Dagenham	Metro Station	Grocery Store	Convenience Store	Park	Plaza
1	Barnet	Coffee Shop	Café	Pub	Italian Restaurant	Beer Bar
2	Bexley	Park	Mediterranean Restaurant	Gas Station	Bakery	Supermarket
4	Bromley	Pub	Train Station	Café	Bar	Supermarket
6	Croydon	Pub	Grocery Store	Park	Bus Stop	Café
7	Ealing	Coffee Shop	Pub	Italian Restaurant	Café	Hotel
8	Enfield	Pub	Café	Restaurant	Coffee Shop	Park
9	Greenwich	Pub	Grocery Store	Café	Garden	Pizza Place
13	Harrow	Coffee Shop	Pub	Theater	Hotel	Art Museum
14	Havering	Coffee Shop	Pub	Café	Yoga Studio	Beer Store
15	Hillingdon	Café	Pub	Coffee Shop	Grocery Store	Park
16	Hounslow	Coffee Shop	Fast Food Restaurant	Indian Restaurant	Sandwich Place	Restaurant
19	Kingston upon Thames	Coffee Shop	Pub	Café	Clothing Store	Indian Restaurant
22	Merton	Coffee Shop	Grocery Store	Pub	Bar	Hotel
24	Redbridge	Pub	Park	Café	Grocery Store	Coffee Shop
25	Richmond upon Thames	Café	Park	Restaurant	Italian Restaurant	Harbor / Marina
27	Sutton	Pub	Café	Coffee Shop	Park	Supermarket

Figure 7. Common venues in each borough of cluster group 1

5. Discussion

We can recommend Bromley, Havering, and Richmon upon Thames to be best borough for considering buying a new house.

These boroughs have a lot of stores, cafes, entertainments, and train stations. At the same time, these boroughs are huge and have a low crime rate. It’s important for consider buying a new house.

6. Conclusions

In this case study, I identified 3 boroughs in London that are likely to be best borough for buying a new house. This analysis will be very useful for someone who in search for a new home in London. This analysis can be used by anyone to find boroughs that are similar to their favorited one, in terms of safety, population density, and access to venues.