Coursera Capstone

The Battle of the Neighborhoods: LONDON

Victoria Ahmadi
6 April 2020

I. INTRODUCTION

Business Problem

As one of the major cultural and financial capitals of Europe, London is a dynamic and transient city that attracts expatriates across all nationalities. A major task of relocation that these individuals must tackle includes deciding where one should buy or rent a property. However, London's real estate market can be daunting due to its sheer size, diversity, and competitiveness. This report aims to predict the best neighborhood based on certain factors.

Determining the best neighborhood varies for everyone. For the purposes of this report, we will consider the following factors: affordability, accessibility, and nearby facilities, such as parks, pharmacies, pubs, etc. The data will mainly be taken with Foursquare location data.

Interest

The target audience for this report include: those who are interested in buying or renting a property in London, such as those who are relocating to London, those who would like to invest in property for their financial portfolio, or those who are looking to settle down. As a young professional myself trying to move to London, I would personally benefit from this report as it would help me determine best areas based on safety, affordability, and vibe.

II. DATA

Sources

i. A listing of all London boroughs taken from Wikipedia. LINK.
ii. Crime records in the London boroughs. LINK.
iii. Affordability of houses using official data. LINK.
iv. Happiness index in the London boroughs. LINK.
v. Districts in Sutton. LINK.

Data Cleaning

i. First, we obtained a listing of all the boroughs in London. To do this, we used BeautifulSoup to scrap the London Wikipedia page. Throughout the project, we will need to ensure the exact borough names are used, so we created a data frame with borough names and the index.
ii. Crime Rate: a data set from Kaggle was downloaded. The variables consisted of the following:
    a. lsoa_code: lower super output area in Greater London
    b. borough: common name for London boroughs
    c. major_category: top crimes in the area
    d. minor_category: low level categorization of crimes within major category
    e. value: monthly count of crime in given area
    f. year: reported year of crime
    g. month: reported month of crime

The csv file is read using pandas. As 2016 was the most recent year reported in the data file, we will use 2016 data for the purposes of this report. If the 'value column' is 0, then the row is deleted. Finally, the dataset was grouped into borough name to obtain the overall count/crime rate.

	lsoa_code	borough	major_category	minor_category	value	year	month
0	E01004177	Sutton	Theft and Handling	Theft/Taking of Pedal Cycle	1	2016	8
1	E01000733	Bromley	Criminal Damage	Criminal Damage To Motor Vehicle	1	2016	4
2	E01003989	Southwark	Theft and Handling	Theft From Shops	4	2016	8
3	E01002276	Havering	Burglary	Burglary in a Dwelling	1	2016	8
4	E01003674	Redbridge	Drugs	Possession Of Drugs	2	2016	11

Figure 1: London Crime Data in 2016

iii. Price per Square Meter: an official dataset was downloaded from the UK Office of National Statistics. The variables in the data included:

a. local authority code
b. local authority name
c. year
d. price per meter squared

Out of these columns, we only need local authority name and price per meter squared. The other two columns were dropped. Next, the dataset was sorted in the ascending order, with the most affordable areas first.

	London_Borough	price per m2
0	Hartlepool	987
1	Middlesbrough	1120
2	Redcar and Cleveland	1182
3	Stockton-on-Tees	1254
4	Darlington	1260
5	Halton	1339

Figure 2: House price per meter squared comparison - London

III. METHODOLOGY

The methodology in this project consists of two parts: exploratory data analysis and modeling.

Exploratory Data Analysis

Using the data frames from the section above, we will conduct an exploratory data analysis. First, we have to visualize the safest areas in London. Next, we have to visualize the most affordable places in London. Lastly, we will need to analyze the happiness index.

Crime data

First, we need to look at an overview of our data frame. In Figure 3, we can see Lambeth has the highest crime rate and that out of 292042 crimes reported, Theft and Handling comprised of most crimes.

	lsoa_code	borough	major_category	minor_category	value	year	month
count	392042	392042	392042	392042	392042.000000	392042.0	392042.
unique	4835	33	7	28	NaN	NaN	NaN
top	E01033583	Lambeth	Theft and Handling	Harassment	NaN	NaN	NaN
freq	256	17605	129159	36213	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	1.877659	2016.0	6.54307
std	NaN	NaN	NaN	NaN	2.650033	0.0	3.42346
min	NaN	NaN	NaN	NaN	1.000000	2016.0	1.00000
25%	NaN	NaN	NaN	NaN	1.000000	2016.0	4.00000
50%	NaN	NaN	NaN	NaN	1.000000	2016.0	7.00000
75%	NaN	NaN	NaN	NaN	2.000000	2016.0	10.0000
max	NaN	NaN	NaN	NaN	149.000000	2016.0	12.0000

Figure 3: Crime data analysis

Next, we plotted a bar graph of the boroughs with the 15 lowest crime rates. In this graph, we can see that Kingston Upon Thames had the lowest crime rate. Sutton, Richmond Upon Thames, and Merton followed.

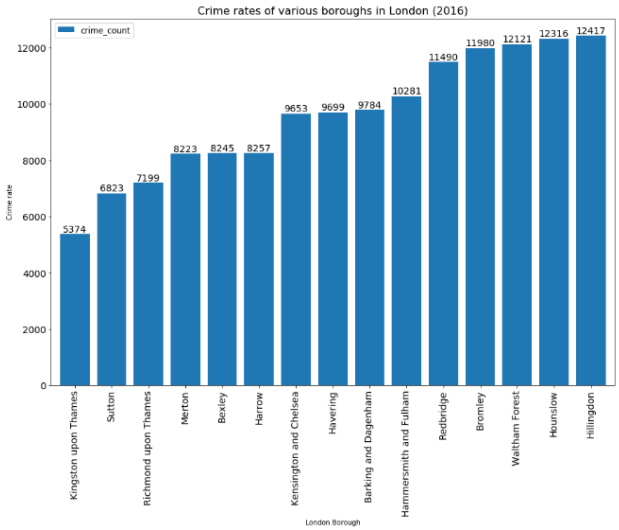

Figure 4: Crime data analysis – bar graph

Housing Price

In this section, we need to determine the most affordable boroughs. First, let's describe the dataset. The mean value of price per square meter is 7473, with a minimum value of 3994 and a maximum value of 19439.

	price per m2
count	32.000000
mean	7473.187500
std	3423.874446
min	3994.000000
25%	5262.500000
50%	6510.000000
75%	8549.750000
max	19439.000000

Figure 5: Housing Price Analysis

Next, we plotted a bar graph to get a visual. Based on this result, we can see that Baking and Dagenham have the most affordable price per square meter, followed by Havering, Bexley, and Croydon.

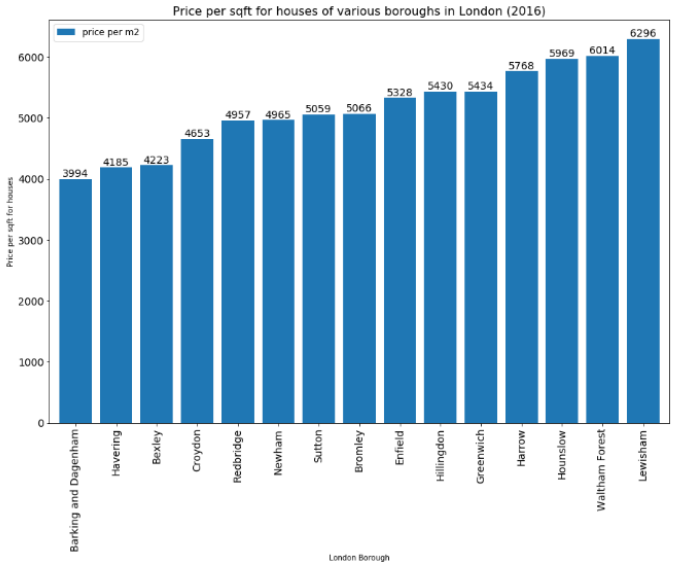

Figure 6: Housing Price Analysis – Bar Graph

Happiness Index:

The boroughs with the highest happiness index is as follows:

	London_Borough
0	Richmond Upon Thames
1	Kingston upon Thames
2	Bromley
3	Sutton
4	Wandsworth
5	Camden
6	Barnet
7	Ealing
8	Greenwich
9	Havering
10	Hackney
11	Waltham Forest
12	Merton
13	Kensington & Chelsea
14	Hammersmith & Fulham

Figure 7: Happiness Index

Based on these three analyses, we can identify that the best boroughs in London for safe and affordable housing are: Sutton, Bromley, and Havering. For this report, I decided to select Sutton as the best option, as it ranked 2nd in safest, 7th in affordability, which is still much less than the mean value in London, and 4th in happiness index.

Modeling

Now that Sutton was selected as the borough we want to analyze, we need to build a dataset with the districts within Sutton. The data used was scrapped from the Sutton Wikipedia page. Using Foursquare Location data, we obtained the 100 most popular venues in a radius of 500m for each district. Next, we performed one hot encoding to convert the categorical variables to binarization. The final dataset contained the following variables:

- District
- District Latitude / Longitude
- Venue
- Venue Latitude
- Venue Longitude
- Venue Category

Next, we calculate the top 10 venues in each district based on their mean.

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Bandon Hill	Park	Pub	Convenience Store	Indian Restaurant	Veterinarian	Creperie	Hardware Store
1	Beddington	Indian Restaurant	Hardware Store	Park	Pub	Veterinarian	Gym / Fitness Center	Gym
2	Beddington Corner	Business Service	Racetrack	Veterinarian	Cosmetics Shop	Hardware Store	Gym / Fitness Center	Gym
3	Belmont	Train Station	Asian Restaurant	Bus Stop	Playground	Pub	Veterinarian	Creperie
4	Benhilton	Indian Restaurant	Gym / Fitness Center	Coffee Shop	Park	Supermarket	Grocery Store	Clothing Store

Figure 8: Top 10 Venues by District

Finally, k-means clustering is performed. K-means clustering is an unsupervised machine learning algorithm that groups similar data points together to a fixed number of clusters (k). The neighborhoods are classified into 3 categories based on the results of the elbow method.

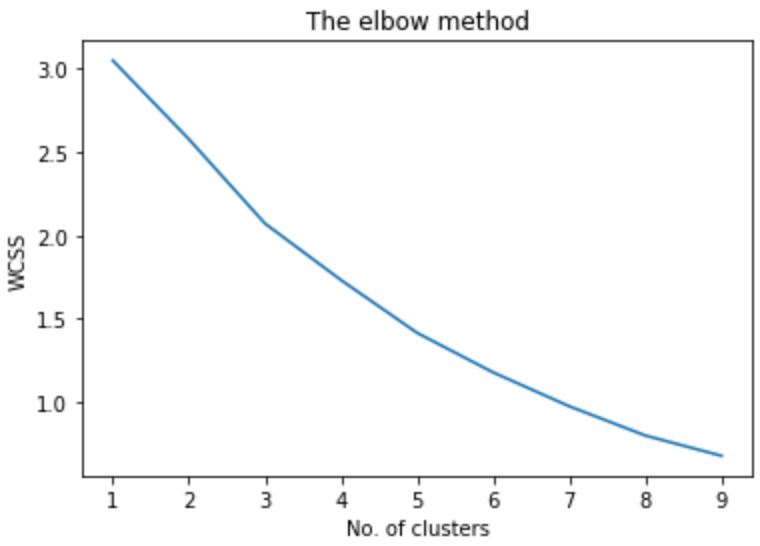

Figure 9: Elbow Method

IV. RESULTS

Cluster 1:

	District	Borough	Latitude	Longitude	Cluster_FINAL	Cluster_Label	Cluster Label	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Co Ver
0	Bandon Hill	Sutton	51.364777	-0.134833	0	1	0	1	Park	Pub	Co Sto
1	Beddington	Sutton	51.371988	-0.132393	0	1	0	1	Indian Restaurant	Hardware Store	Par
14	South Beddington	Sutton	51.371988	-0.132393	0	1	0	1	Indian Restaurant	Hardware Store	Par

Figure 10: Cluster 1

The first cluster consists of the Bandon Hill, Beddington, and South Beddington districts. This cluster mainly contains parks, Indian restaurants, hardware stores, and pubs. Based on this data, this district seems to be a more residential area.

Cluster 2:

	District	Borough	Latitude	Longitude	Cluster_FINAL	Cluster_Label	Cluster Label	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
3	Belmont	Sutton	51.343785	-0.201152	1	2	1	2	Train Station	Asian Restaurant
4	Benhilton	Sutton	51.371642	-0.191571	1	2	1	2	Indian Restaurant	Gym / Fitness Center
5	Carshalton	Sutton	51.365788	-0.161086	1	2	1	2	Grocery Store	Park
6	Carshalton Beeches	Sutton	51.357196	-0.169351	1	0	1	0	Train Station	Italian Restaurant
8	Cheam	Sutton	51.357616	-0.216241	1	2	1	2	Pub	Coffee Shop
9	Hackbridge	Sutton	51.379613	-0.156754	1	2	1	2	River	Train Station
10	Little Woodcote	Sutton	51.346076	-0.145932	1	5	3	5	Park	Sports Club
11	North Cheam	Sutton	51.371578	-0.220225	1	2	1	2	Coffee Shop	Gym / Fitness Center
12	Rosehill	Sutton	51.012505	-0.140639	1	2	1	2	Restaurant	Grocery Store
13	St. Helier	Sutton	51.386695	-0.180057	1	3	1	3	Pharmacy	Breakfast Spot
15	Sutton	Sutton	51.357511	-0.173640	1	0	1	0	Train Station	Italian Restaurant
16	Sutton Common	Sutton	51.375373	-0.196032	1	2	1	2	Train Station	Gym / Fitness Center

Figure 11: Cluster 2

The second cluster is by far the largest cluster in this analysis. The top common venues in this cluster include: train stations, pubs, grocery stores, coffee shops, gym/fitness center, cafe. Based on this data, this cluster seems to be a busy, happening, and perhaps central area of Sutton.

Cluster 3:

	District	Borough	Latitude	Longitude	Cluster_FINAL	Cluster_Label	Cluster Label	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
2	Beddington Corner	Sutton	51.386942	-0.149532	2	4	2	4	Business Service	Racetrack

Figure 12: Cluster 3

The last cluster in this analysis solely consists of the Beddington Corner district, which consists of business services, racetrack, and vet. Based on this data, this area seems to be quieter and perhaps more of a business area, rather than a residential area.

V. DISCUSSION

Three clusters were generated using the k-means method. As a young professional looking to move to London, I have various requirements regarding possible neighborhoods to move into: afforadability, crime rates, and general vibe (happiness). Based on this analysis, I have come to the conclusion that the Sutton borough would be the best location for my preferences. Within Sutton, the best neighborhoods would be any in the second cluster because those areas had numerous and diverse venues that fits my lifestyle - active, accessible, and social.

VI. CONCLUSION

The Battle of the Neighborhoods assignment was extremely useful in helping me understand the application of k-means clustering in determining the best neighborhoods based on certain characteristics. However, most importantly, this project provided me the opportunity to maneuver through the data science methodology using real-life data.