The Battle of Neighborhood

2020

JULY 28

CAPSTONE PPROJECT REPORT

Authored by: Vrinda Sharda

Business Problem Section

Background

London is the largest city and the capital of United Kingdom. It is considered as one of the most important global cities. Its estimated mid-2018 municipal population (corresponding to Greater London) was 8,908,081, the third most populous of any city in Europe and accounts for 13.4% of the UK population. The city covers a total area of 607 sq. miles. Clearly, London is a city with high population density. Real estate investment in such cities thus becomes a big investment. And before making such an investment, one should have a proper investment strategy.

Business Problem

In a country with such a huge area and high real estate prices, it becomes very tough to find a property with its appropriate value, keeping in mind all the amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores. As a result, the business problem we are currently posing is: How could we recommend profitable venues to support homebuyers according to amenities and essential facilities surrounding such venues?

Interest

Obviously, the real estate investors would be very interested in improving their business and investing at a place which provides them maximum profit.

Other than them, people looking for a house to live in with a good neighborhood with all the essential facilities nearby shall be interested in this project.

Data Section

Data Sources

Data on London properties and the relative price paid data were extracted from the HM Land Registry (http://landregistry.data.gov.uk/). The following fields comprise the address data included in Price Paid Data: Postcode; PAON Primary Addressable Object Name, typically the house number or name; SAON Secondary Addressable Object Name. If there is a sub-building, for example, the building is divided into flats, there will be a SAON; Street; Locality; Town/City; District; County.

To explore and target recommended locations across different venues according to the presence of amenities and essential facilities, we will access data through FourSquare API interface and arrange them as a dataframe for visualization.

By merging data on London properties and the relative price paid data from the HM Land Registry and data on amenities and essential facilities surrounding such properties from FourSquare API interface, we will be able to recommend profitable real estate investments.

Data Description

The first table that we extracted from the HM Land Registry consisted of 16 columns. They were: Postcode; PAON Primary Addressable Object Name, typically the house number or name; SAON Secondary Addressable Object Name. If there is a sub-building, for example, the building is divided into flats, there will be a SAON; Street; Locality; Town/City; District; County.

The locational information (longitude and latitude) was extracted using the geocoders. The neighbourhood was then explored using FourSqaure API interface. This helped to target the recommended locations like parks, supermarket, schools etc. across different venues.

By merging all these data that we have (in the form of a table), it was convenient to recommend a profitable location to the investors.

Methodology

Data Science Methodology consists of the following steps:
Business understanding, Analytic approach, Data Requirement, Data Collection, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment and Feedback.
The first five steps are already explained in the sections above. So here after, I will Provide detailed description of the next steps in the project.

Data Preparation

At this stage, we convert our data from raw initial form into another pre processed from in order to prepare the data for further analysis. Therefore we perform the following steps:

1. **Renaming the columns:** In this step, we assigned meaningful names to the columns of the table we extracted from the HM Land Registry. The names given were: 'TUID', 'Price', 'Date_Transfer', 'Postcode', 'Prop_Type', 'Old_New', 'Duration', 'PAON', 'SAON', 'Street', 'Locality', 'Town_City', 'District', 'County', 'PPD_Cat_Type', 'Record_Status'. This made understanding the data better.

2. **Formatting the date column and sorting the data by the date of sale:** While studying the data, I found it necessary to format the date into an easy readable form. The properties are transferred from one person to another quite often. So the data also had a large no of properties that had very old ‘Date of Transfer’. What we require here is are the properties with their current state, Hence, the data was reduced to the one where ‘Date of Transfer’ less than 2018 were dropped to have a cleaner data.

3. **Making a list of street names in London**: London is the capital and largest city of England and the United Kingdom. It has a total area of 607 sq. miles, out if which the city of London consists of 1.12 sq. miles and the rest 606 sq. miles are the parts of greater London. So here, to see what areas come under the required criteria, we make a list of streets in the City of London.

4. **Calculating the streetwise average price and the property:** One of the main key points of the business problem was to get the property in budget. So Calculate the mean of property rate of a street and place it in our table. At this point we would like our buyer to tell us his/her budget. Say one wats to buy in the range 2,200,000 to 2,500,000. So we simply filter our data to get in the range. We now have 162 streets that fit our client's budget. We get data structure similar to the following:

	Street	Avg_Price
196	ALBION SQUARE	2.450000e+06
390	ANHALT ROAD	2.435000e+06
405	ANSDELL TERRACE	2.250000e+06
422	APPLEGARTH ROAD	2.400000e+06
855	BARONSMEAD ROAD	2.375000e+06
981	BEAUCLERC ROAD	2.480000e+06
1102	BELVEDERE DRIVE	2.340000e+06
1215	BICKENHALL STREET	2.208500e+06
1253	BIRCHLANDS AVENUE	2.217000e+06
1553	BRAMPTON GROVE	2.456875e+06
1632	BRIARDALE GARDENS	2.397132e+06
1797	BROOKWAY	2.400000e+06
1914	BURBAGE ROAD	2.445000e+06
1980	BURY WALK	2.492500e+06

FIG.1: Table showing the name of street and its mean property price.

5. **Mapping the locations:** To map the locations, we first need its coordinates. Therefore we get coordinates of all the 162 streets using geocoders and then map these locations on the Map of London to get a visual of what we have acquired so far. This was done using folium. Now lets take a look at our map.

   We have located the affordable coordinates on the map. One can clearly see that area surrounding Notting Hill, Kensington and Chelsa can be the places of our interests.

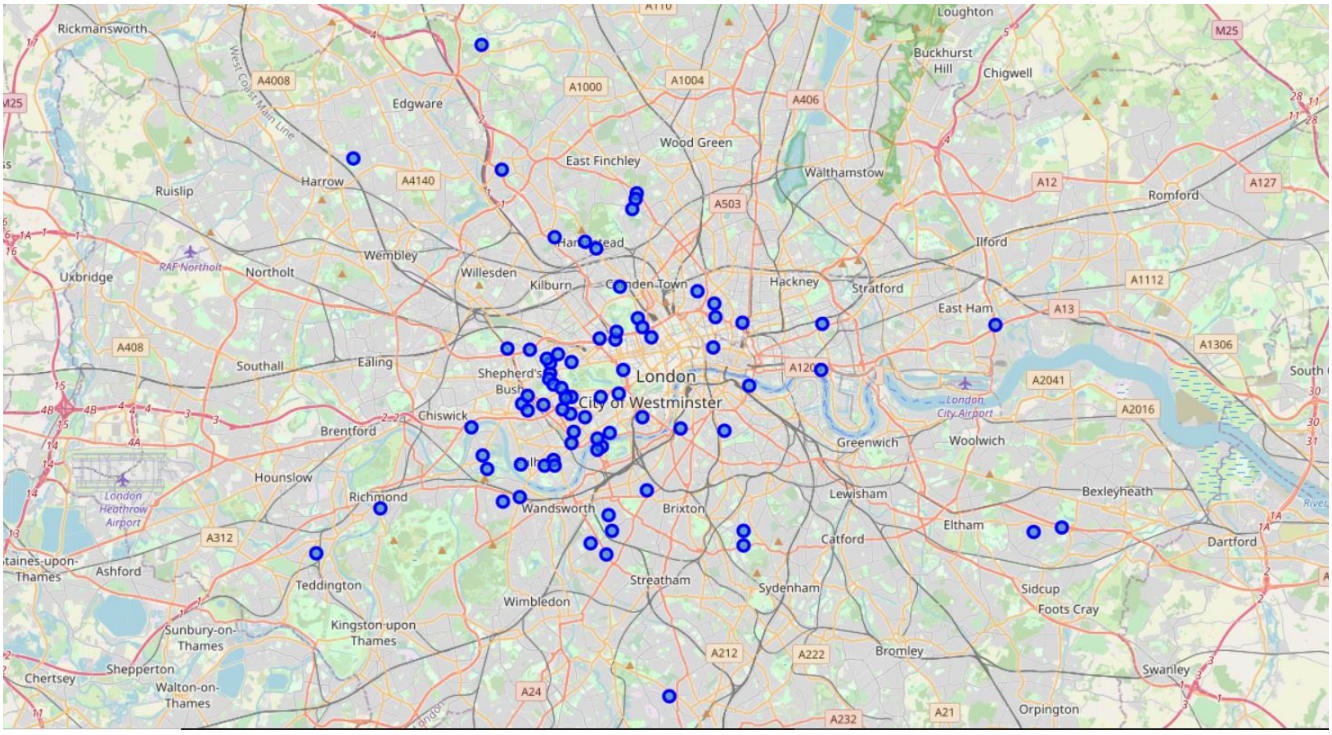

FIG. 2: Map showing the affordable location in London.

Modeling

We have a list of affordable streets in London. Now, out of these 162 available options, which one should the investor invest in?
Some are situated in green areas, some have supermarkets and grocery stores nearby, some have fun places like clubs and pubs while sone have theatres in their neighborhood. We need to decide what is the best as an investor to invest in. Or what would be the best area to live in.

To help our client, we shall now use a machine learning algorithm.
Since we need to group the object into similar objects and dissimilar objects to data points in other groups, we will use clustering. We will divide all the streets into medium or large sized clusters, thereby making it easy to see what places have the required neighborhoods.

We will use the K-Means Clustering technique as it is the fastest and efficient in terms of computational cost, is highly flexible to account for mutations in real estate market in London.

We start with the FourSquare API and collect all the venues int the neighborhood of our filtered streets. We use their coordinates for the same. We now have the list of venues nearby. We take all the unique categories into account. There are 348 unique categories. Since we need decimal values in our algorithm, we use one hot encoding and then get its mean. After sorting them according to the most visited venues, we make a data frame having the street name and top 10 most common venues. We now have a table of shape(149, 11)

After our inspection of venues, facilities and amenities nearby the most profitable real estate in London, we can now start clustering properties accordingly.
We will be distributing the area into 10 clusters and then we merge our two tables to get a final one which includes street, average price, latitude, longitude, cluster labels, and 10 most common venues.

	Street	Avg_Price	Latitude	Longitude	Cluster Labels	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
196	ALBION SQUARE	2.450000e+06	-41.273758	173.289393	7	Café	Indian Restaurant	Pub	Coffee Shop	Restaurant	Bar	Burger Joint	New American Restaurant	Seafood Restaurant	Museum
390	ANHALT ROAD	2.435000e+06	51.480316	-0.166801	1	Pub	French Restaurant	Grocery Store	Diner	Plaza	Japanese Restaurant	Gym / Fitness Center	Cocktail Bar	English Restaurant	Garden
405	ANSDELL TERRACE	2.250000e+06	51.499890	-0.189103	8	Juice Bar	Hotel	Pub	Restaurant	Indian Restaurant	Clothing Store	Italian Restaurant	Café	Supermarket	Mediterranean Restaurant
422	APPLEGARTH ROAD	2.400000e+06	53.748654	-0.326670	9	Pub	Nightclub	Bar	Casino	Food & Drink Shop	Food	Food Service	Exhibit	Factory	Falafel Restaurant
855	BARONSMEAD ROAD	2.375000e+06	51.477315	-0.239457	3	Pub	Thai Restaurant	Pizza Place	Community Center	Restaurant	Coffee Shop	Park	Farmers Market	Café	Nature Preserve
981	BEAUCLERC ROAD	2.480000e+06	30.211452	-81.617981	4	Spa	Automotive Shop	Harbor / Marina	Pizza Place	Zoo Exhibit	Filipino Restaurant	Event Space	Exhibit	Factory	Falafel Restaurant
1102	BELVEDERE DRIVE	2.340000e+06	44.707562	-63.545599	5	Campground	Zoo Exhibit	Film Studio	Exhibit	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Filipino Restaurant
1215	BICKENHALL STREET	2.208500e+06	51.521201	-0.158908	2	Gastropub	Restaurant	Coffee Shop	Pizza Place	Hotel	Italian Restaurant	Garden	Bakery	Greek Restaurant	Bar
1253	BIRCHLANDS AVENUE	2.217000e+06	51.448394	-0.160468	2	Pub	Breakfast Spot	Coffee Shop	Brewery	French Restaurant	Chinese Restaurant	Lake	Train Station	Bakery	Pizza Place
1553	BRAMPTON GROVE	2.456875e+06	51.589961	-0.318525	7	Food Service	Home Service	Zoo Exhibit	Film Studio	Event Space	Exhibit	Factory	Falafel Restaurant	Farm	Farmers Market
1632	BRIARDALE GARDENS	2.397132e+06	51.560175	-0.195431	9	Indian Restaurant	Chinese Restaurant	Health & Beauty Service	Coffee Shop	Grocery Store	Fast Food Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Factory
1797	BROOKWAY	2.400000e+06	45.432185	-122.802812	9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1914	BURBAGE ROAD	2.445000e+06	52.538507	-1.353674	7	Construction & Landscaping	Bar	Grocery Store	Dance Studio	Athletics & Sports	Film Studio	Exhibit	Factory	Falafel Restaurant	Farm
1980	BURY WALK	2.492500e+06	52.145529	-0.423593	4	Supermarket	English Restaurant	Rental Car Location	Gym	Hardware Store	Fast Food Restaurant	Coffee Shop	Café	Dry Cleaner	Park

FIG.3: The merged table

We now head towards our final step, where we map the cluster into the map of London and compare. We have streets with similar neighborhoods in similar clusters. Therefore, we can now come to our result and suggest our client the best areas of his interest which will provide him/her maximum profit.

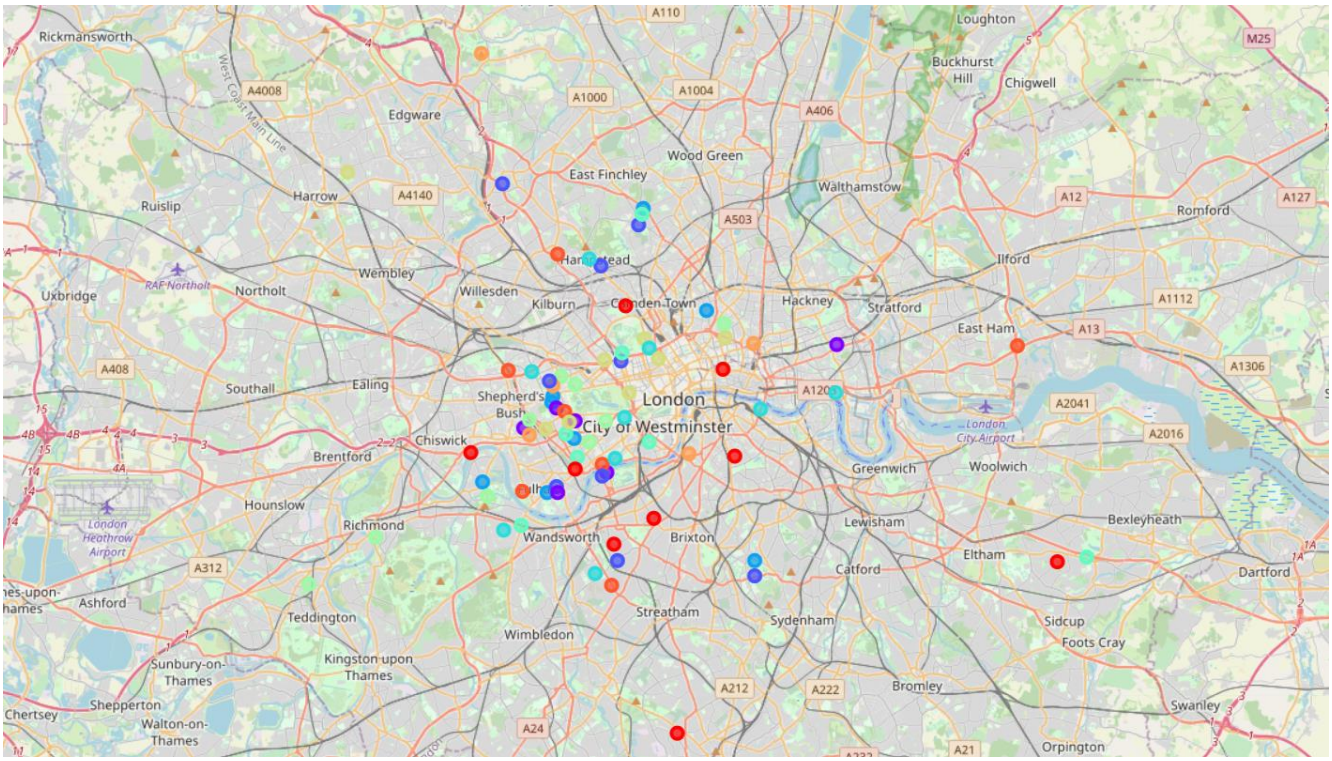

FIG. 2: Map showing the affordable locations divided into suitable clusters.

Results

First of all, The London Housing market is a great investment to make, but its good to make an investment at such area that boosts up the profit.

We have divided our streets into 10 clusters. Lets have a closer look into each of them and see what we can conclude out of these.

a. **Cluster 0:** Pubs and bars are one of the most visited places here. It also has some restaurants and factories in its neighborhoods. People can also find some green areas and river front here.

b. **Cluster 1:** Pubs and restaurants are also common here. But unlike cluster 0, it has some food facilities like departmental store, weight loss center, grocery store and pharmacies.

c. **Cluster 2:** This cluster las a lots of coffee shops and restaurant and hotels nearby. Some pubs and bars can also be seen around.

d. **Cluster 3:** this cluster is for Pub lovers. Hotels, hostels, bakeries and restaurants are also common here.

e. **Cluster 4:** This is a highly profitable area with many facilities like supermarket, spa, automotive shops, gym, convenience stores, dry cleaners etc. nearby. You can also find a decent number of restaurants here.

f. **Cluster 5:** This are comes under green zones. It has parks and gardens as commonly visited places. Film studio and restaurants are also common here.

g. **Cluster 6:** Another green areas with farms, parks and gardens. Film studios, hotels and restaurants are common. Unlike cluster 5, you may find facilities like supermarket, stationary and fitness center here.

h. **Cluster 7:** This place is particularly for people who are particular about their fitness. You may find dance studio, playground, athletics and

sports and fitness centers. Apart from falling under green area, it also some facilities nearby like grocery store, home.

i. **Cluster 8:** It a profitable place with facilities such as cosmetics store, business services, convenience store, bookstore, supermarket, cloth store, bank, pharmacy etc. It also has a decent number of restaurants.

j. **Cluster 9:** This area is a combination of pubs, casinos and night clubs along with nearby facilities like health and beauty center, spa, grocery store, dry cleaner, food service etc. It also has a decent number of restaurants. Hence making it a great place to invest in.

Discussion

We may analyse our results according to the clustered that we produced. Even though, all the clusters could pose an optimal range of facilities and amenities, we have found certain patterns.

First, we examined them according to neighbourhoods/London areas. although West London (Notting Hill, Kensington, Chelsea, Marylebone) and North-West London (Hampsted) might be considered highly profitable venues to purchase a real estate according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores, South-West London (Wandsworth, Balham) and North-West London (Islington) are arising as next future elite venues with a wide range of amenities and facilities. Accordingly, one might target under-priced real estates in these areas of London in order to make a business affair.

Second, we divided them into 10 clusters. Clusters 0, 1 and 3 are for the ones who enjoy spending weekends in clubs and pubs, while cluster 5 and 6 are for theatre lovers. Investing in clusters 4, 8 and 9 would be the best as

they provide the maximum facilities in their neighbourhoods. For the home buyers who are prone to live in 'green' areas should go for areas in cluster 2, 5, 6 and 7 as they have parks, gardens, farms and waterfronts.

Conclusion

To sum up, according to Bloomberg News, the London Housing Market is in a rut. It is now facing a number of different headwinds, including the prospect of higher taxes and a warning from the Bank of England that U.K. home values could fall as much as 30 percent in the event of a disorderly exit from the European Union. In this scenario, it is urgent to adopt machine learning tools in order to assist homebuyers clientele in London to make wise and effective decisions. As a result, the business problem we were posing was: how could we provide support to homebuyers clientele in to purchase a suitable real estate in London in this uncertain economic and financial scenario?

To solve this business problem, we clustered London neighbourhoods in order to recommend venues and the current average price of real estate where homebuyers can make a real estate investment. We recommended profitable venues according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores. First, we gathered data on London properties and the relative price paid data were extracted from the HM Land Registry (http://landregistry.data.gov.uk/). Moreover, to explore and target recommended locations across different venues according to the presence of amenities and essential facilities, we accessed data through FourSquare API interface and arranged them as a data frame for visualization. By merging data on London properties and the relative price paid data from the HM Land Registry and data on amenities and essential facilities surrounding such properties from FourSquare API interface, we were able to recommend profitable real estate investments.

Second, The Methodology section comprised four stages: 1. Collect Inspection Data; 2. Explore and Understand Data; 3. Data preparation and pre-processing; 4. Modelling. In particular, in the modelling section, we used the k-means clustering technique as it is fast and efficient in terms of computational cost, is highly flexible to account for mutations in real estate market in London and is accurate.

Finally, we drew the conclusion that even though the London Housing Market may be in a rut, it is still an "ever-green" for business affairs. We discussed our results under two main perspectives. First, we examined them according to neighbourhoods/London areas. although West London (Notting Hill, Kensington, Chelsea, Marylebone) and North-West London (Hampsted) might be considered highly profitable venues to purchase a real estate according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores, South-West London (Wandsworth, Balham) and North-West London (Islington) are arising as next future elite venues with a wide range of amenities and facilities. Accordingly, one might target under-priced real estates in these areas of London in order to make a business affair. Second, we analysed our results according to the ten clusters we produced. Clusters 0, 1 and 3 are for the ones who enjoy spending weekends in clubs and pubs, while cluster 5 and 6 are for theatre lovers. Investing in clusters 4, 8 and 9 would be the best as they provide the maximum facilities in their neighbourhoods. For the home buyers who are prone to live in 'green' areas should go for areas in cluster 2, 5, 6 and 7 as they have parks, gardens, farms and waterfronts.