Predicting the best locations to purchase real estate in London

Alvin Lumumba

November 24, 2020

1. Introduction

1.1 Background

Multifamily is becoming an established sector of the UK property market. From less than 1% in 2014, it now accounts for 7% of total UK real estate investment. 2020 will likely see new entrants to the market, and current investors will continue to build their portfolios. As a result, we expect total multifamily investment in 2020 to significantly exceed 2019.

The UK's decision to vote to leave the EU has resulted in a widening of the differential in pricing between UK yields and the rest of Europe. This means that UK office property will offer relative value to overseas investors in 2020. If EU withdrawal issues are settled during 2020, the conditions for yield compression could emerge as the year progresses. Central London investment volumes should increase in 2020 due to strong occupier fundamentals and c£32bn of overseas equity targeting the region.

1.2 Problem

In this scenario, it is urgent to adopt machine learning tools in order to assist homebuyer's clientele in London to make wise and effective decisions. As a result, the business problem we are currently posing is: how could we provide support to homebuyer's clientele in purchase of suitable real estate in London in this uncertain economic and financial scenario?

To solve this business problem, we are going to cluster London neighbourhoods according to amenities nearby and their real estate prices in order superb locations where homebuyers can make a real estate investment.

1.3 Interest

Real estate buyers would be very interested in accurate predictions of clusters with high growth potential. Undervalued clusters would be a prime opportunity for investment.

2. Data acquisition and cleaning

2.1 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 acquire and explore the amenities and essential facilities in various locations, data was accessed through the FourSquare API interface and formatted in tabular form as a data frame.

2.2 Data cleaning

Below is a snap of our original data frame. It had 1031509 rows and 16 columns.

	{79A74E22-41E2-1289-E053-6B04A8C01627}	60000	2018-06-29 00:00	DH3 1DN	F	N	L	20	Unnamed: 8	BEACONSFIELD TERRACE	BIRTLEY	CHESTER LE STREET	G
0	{79A74E22-41E3-1289-E053-6B04A8C01627}	149950	2018-06-14 00:00	DH4 6NZ	T	Y	F	50	NaN	GLANVILLE DRIVE	NaN	HOUGHTON LE SPRING	SU
1	{79A74E22-41E4-1289-E053-6B04A8C01627}	164950	2018-06-29 00:00	SR2 0FD	S	Y	F	6	NaN	WILSHIRE CLOSE	NaN	SUNDERLAND	SU
2	{79A74E22-41E5-1289-E053-6B04A8C01627}	224950	2018-06-29 00:00	SR2 0FA	D	Y	F	47	NaN	WOODHAM DRIVE	NaN	SUNDERLAND	SU
3	{79A74E22-41E6-1289-E053-6B04A8C01627}	129950	2018-06-28 00:00	DH4 6NY	S	Y	F	65A	NaN	CHALK HILL ROAD	NaN	HOUGHTON LE SPRING	SU
4	{79A74E22-41E7-1289-E053-6B04A8C01627}	144395	2018-02-23 00:00	NE31 2EL	T	Y	F	9	NaN	TURNBERRY DRIVE	NaN	HEBBURN	SC TY

Data downloaded were combined into one table. We assigned meaningful column names such as ‘Price’ and ‘Street’ to the data frame. We formatted the date column into a date type object. We also deleted all obsolete transactions done before Brexit. To understand the data easier, we sorted the data frame by date of sale. We dropped all other initial columns except ‘avg_price’ as they were not required for clustering.

We were now left with a data frame with 159 rows and 2 columns as we can see below.

	Street	Avg_Price
196	ALBION SQUARE	2450000.0
390	ANHALT ROAD	2435000.0
405	ANSDELL TERRACE	2250000.0
422	APPLEGARTH ROAD	2400000.0
857	BARONSMEAD ROAD	2375000.0
...	...	...
13733	WILFRED STREET	2410538.5
13759	WILLOW BRIDGE ROAD	2425000.0
13779	WILSON STREET	2257500.0
13808	WINCHENDON ROAD	2350000.0
13845	WINGATE ROAD	2206400.0

2.3 Feature selection

We created a new data frame containing only the data from London city. We restricted the average price data of real estate to be clustered within a budget range of 2.2m to 2.5m pounds. We also created a new column containing the street feature.

We obtained coordinate data of London and its streets using a python library and used it to obtain the location coordinate columns. Our data frame now looked as follows:

	Street	Avg_Price	Latitude	Longitude
196	ALBION SQUARE	2450000.0	-41.273758	173.289393
390	ANHALT ROAD	2435000.0	29.712770	-98.094806
405	ANSDELL TERRACE	2250000.0	51.500005	-0.189154
422	APPLEGARTH ROAD	2400000.0	53.749244	-0.326780
857	BARONSMEAD ROAD	2375000.0	51.477315	-0.239457

We got a list of venues near each street from the foursquare api. We used these venues to create a new alternate data frame. The data frame had 4518 rows and 344 unique venue categories.

	Street	Street Latitude	Street Longitude	Venue	Venue Latitude	Venue Longitude
0	ALBION SQUARE	-41.273758	173.289393	The Free House	-41.273340	173.287364
1	ALBION SQUARE	-41.273758	173.289393	Queen's Gardens	-41.273671	173.291383
2	ALBION SQUARE	-41.273758	173.289393	The Indian Cafe	-41.273308	173.286530
3	ALBION SQUARE	-41.273758	173.289393	Urban	-41.274355	173.286317
4	ALBION SQUARE	-41.273758	173.289393	Fish Stop	-41.276010	173.289592

We applied one-hot encoding on the venue categories as they will be our model independent variables. We then grouped the data frame by street to get the following data frame.

	Street	ATM	Acai House	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	...	Vietnamese Restaurant
0	ALBION SQUARE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
1	ANHALT ROAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
2	ANSDELL TERRACE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
3	APPLEGARTH ROAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
4	BARONSMEAD ROAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0

We then sorted the data frame to obtain the top five venues near each profitable real estate. Below is a snap of our outputted list.

```
----ALBION SQUARE----
                venue  freq
0                Café  0.22
1                 Pub  0.07
2  Indian Restaurant  0.07
3                 Bar  0.07
4         Coffee Shop  0.07


----ANHALT ROAD----
           venue  freq
0  Intersection  0.17
1  Dance Studio  0.17
2   Coffee Shop  0.17
3         Hotel  0.17
4           Gym  0.17
```

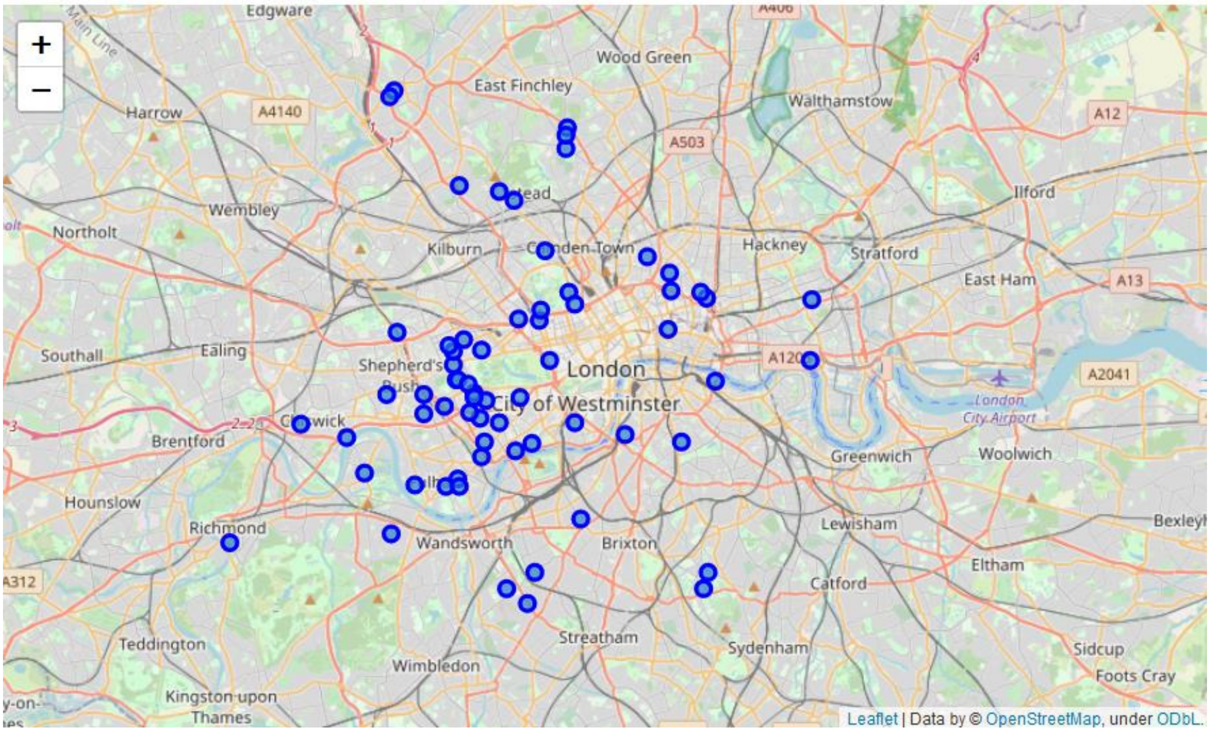
We also sorted the data frame to obtain the most common venue near each street as we can see below.

	Street	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	
0	ALBION SQUARE	Café	Indian Restaurant	Pub	Restaurant	Coffee Shop	Bar	Beer Garden	Paper / Office Supplies Store	Fish & Chips Shop	S
1	ANHALT ROAD	Movie Theater	Coffee Shop	Hotel	Intersection	Gym	Dance Studio	English Restaurant	Escape Room	Ethiopian Restaurant	E
2	ANSDELL TERRACE	Hotel	Indian Restaurant	Café	Pub	Juice Bar	Italian Restaurant	Restaurant	Clothing Store	French Restaurant	(
3	APPLEGARTH ROAD	Sandwich Place	Nightclub	Auto Dealership	Casino	Bar	Flea Market	Fish Market	English Restaurant	Escape Room	
4	BARONSMEAD ROAD	Food & Drink Shop	Breakfast Spot	Nature Preserve	Pizza Place	Movie Theater	Community Center	Indie Movie Theater	Pub	Thai Restaurant	

3. Exploratory Data Analysis

3.1 Visualizations

We created a map of London with our price data as a marker using our current data frame.



4. **Predictive Modeling**

There are two types of models, clustering and classification, that can be used to predict good locations to invest in real estate. Clustering models can group similar data points together, while classification models like decision trees create a decision tree based on the features. Due to limitations in predictive features, we carried out clustering in this study.

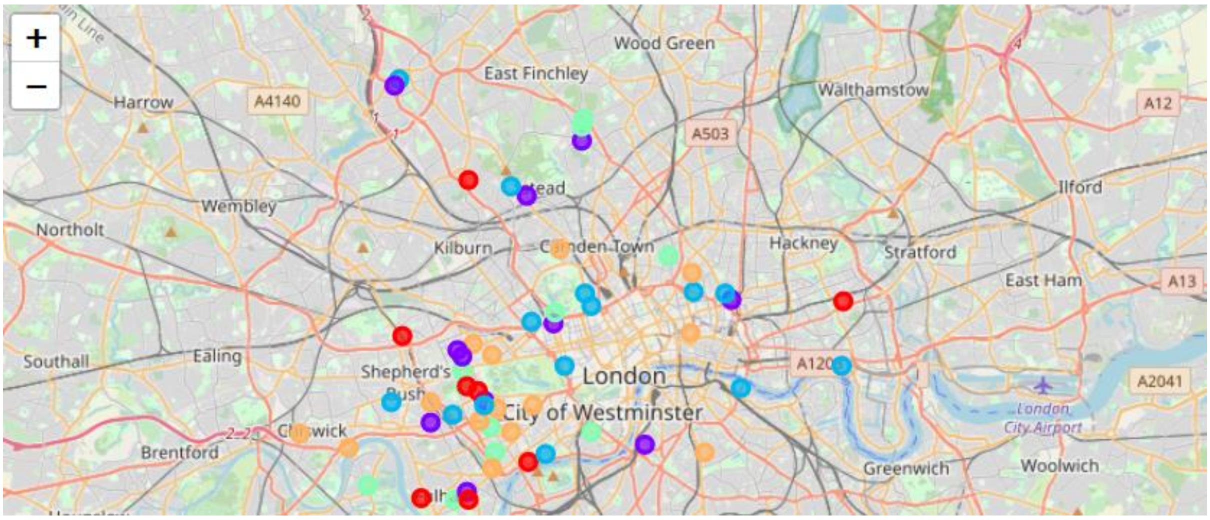
4.1 Clustering models

4.1.1 Applying standard algorithms

We fitted a k-means clustering algorithm to the dataset, using 5 clusters. We added the clustering labels from our data frame to our grouped data frame and got the following data frame.

	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
196	ALBION SQUARE	2.450000e+06	-41.273758	173.289393	2	Café	Indian Restaurant	Pub	Restaurant	Coffee Shop
390	ANHALT ROAD	2.435000e+06	29.712770	-98.094806	0	Movie Theater	Coffee Shop	Hotel	Intersection	Gym
405	ANSDELL TERRACE	2.250000e+06	51.500005	-0.189154	1	Hotel	Indian Restaurant	Café	Pub	Juice Bar
422	APPLEGARTH ROAD	2.400000e+06	53.749244	-0.326780	0	Sandwich Place	Nightclub	Auto Dealership	Casino	Bar
857	BARONSMEAD ROAD	2.375000e+06	51.477315	-0.239457	3	Food & Drink Shop	Breakfast Spot	Nature Preserve	Pizza Place	Movie Theater
983	BEAUCLERC ROAD	2.480000e+06	51.499577	-0.229033	2	Pub	Coffee Shop	Hotel	Bed & Breakfast	Chinese Restaurant
1105	BELVEDERE DRIVE	2.340000e+06	38.072439	-78.459970	3	Pool	Playground	Athletics & Sports	Zoo	Farm

We visualized the predicted clusters as shown below.



5. Results

We see that although West London (Notting Hill, Kensington, Chelsea, Marylebone) and North-West London (Hampsted) might be considered highly profitable venues to purchase 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 (Isliington) are rising 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 profit in the near future.

We have found two main patterns. The first pattern refers to Clusters 0, 2 and 4; here we may target home buyers prone to live in 'green' areas with parks, waterfronts. The second pattern refers to Clusters 1 and 3; here we may target individuals who love pubs, theatres and soccer.

6. Conclusion

We drew the conclusion that even though the London Housing Market may be in disarray, it is still an "ever-green" for business affairs. We discussed our results under two main perspectives. First, we examined them according to neighborhoods/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 (Isliington) 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 analyzed our results according to the five clusters we produced. While Clusters 0, 2 and 4 may target home buyers prone to live in 'green' areas with parks, waterfronts, Clusters 1 and 3 may target individuals who love pubs, theatres and soccer.